Financial Prediction
Using Non-Linear Classification Techniques

George T. Albanis

A Thesis Submitted for the Degree of PhD in Finance

City University Business School
Department of Banking and Finance

September 2001

Paginated blank pages are scanned as found in original thesis

No information is missing

CONTENTS

CHAPTER 3: CLASSIFICATION RULES - SUPERVISED LEARNING

Part One: Parametric, Semi-Parametric, and Non-Parametric Smoothing Methods

Part Two: Computer-Intensive Classification Techniques

CHAPTER 9: DIMENSIONALITY REDUCTION TECHNIQUES BASED ON NEURAL NETWORKS - APPLICATIONS TO STOCK SELECTION AND CREDIT RATINGS

Part One: Predicting High Performing Shares Using Dimensionality Reduction Techniques Based on Neural Networks

Part Two: Assessing the Long-Term Credit Standing of Debt Issuers Using Dimensionality Reduction Techniques Based on Neural Networks - An Alternative to Overfitting

Part Three: Summary and Conclusions

LIST OF TABLES

CHAPTER 4: AN INVESTIGATION OF THE DISTRIBUTIONAL PROPERTIES OF FINANCIAL RATIOS - APPLICATION TO BOND RATINGS

CHAPTER 5: PREDICTING HIGH PERFORMANCE STOCKS USING FIVE STATISTICAL CLASSIFICATION ALGORITHMS

CHAPTER 6: COMBINING HETEROGENEOUS CLASSIFIERS TO PREDICT HIGH PERFORMANCE STOCKS

CHAPTER 7: PREDICTING HIGH PERFORMING SHARES USING ACCOUNTING AND NON-ACCOUNTING INFORMATION

CHAPTER 8: STOCK RETURN PREDICTABILITY IN UK INDUSTRIAL SECTORS

CHAPTER 9: DIMENSIONALITY REDUCTION TECHNIQUES BASED ON NEURAL NETWORKS APPLICATIONS TO STOCK SELECTION AND CREDIT RATINGS

LIST OF FIGURES

CHAPTER 3: CLASSIFICATION RULES - SUPERVISED LEARNING

CHAPTER 4: AN INVESTIGATION OF THE DISTRIBUTIONAL PROPERTIES OF FINANCIAL RATIOS - APPLICATION TO BOND RATINGS

CHAPTER 5: PREDICTING HIGH PERFORMANCE STOCKS USING FIVE STATISTICAL CLASSIFICATION ALGORITHMS

CHAPTER 6: COMBINING HETEROGENEOUS CLASSIFIERS TO PREDICT HIGH PERFORMANCE STOCKS

CHAPTER 7: PREDICTING HIGH PERFORMING SHARES USING ACCOUNTING AND NON-ACCOUNTING INFORMATION

Acknowledgments

I would like to thank my supervisor Professor Roy A. Batchelor who very kindly provided me with comments, corrections, suggestions and insights. My formal expression of thanks is wholly inadequate.

I feel obliged to thank the School of Informatics who provided me with financial assistance to perform my research.

I would like to thank my colleagues at City University Business School for their emotional support.

I am indebted to my family who shared in this burden fully and magnificently.

This study is dedicated to the God who drives me to go ahead.

DECLARATION

"I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement".

ABSTRACT

In this thesis, we explore the ability of statistical classification methods to predict financial events in the bond and stock markets. Our classification methods include conventional Linear Dicriminant Analysis (LDA), and a number of less familiar non-linear techniques such as Probabilistic Neural Network (PNN), Learning Vector Quantization (LVQ), Oblique Classifier (OC1), and Ripper Rule Induction (RRI).

Initially, we investigate the distributional properties of financial ratios and we compare three classifiers namely, LDA, PNN, and RRI in terms of their ability to predict long-term bond ratings. After experimentation, we found that PNN and RRI which are non-linear classifiers not only significantly outperform the LDA, but they are also more robust to different distributional assumptions compared to LDA which is affected from the assumption of multivariate normality. The final outcome of this experiment is a new quantitative system to predict long-term bond ratings using probabilistic neural networks and rule induction techniques.

Considering the empirical evidence but also the "strong" and "wild" capabilities of supervised classification rules, we explore the possibility of identifying outperforming shares by combining our five heterogeneous classifiers namely, LDA, PNN, LVQ, OC1, and RRI through Majority Voting (MV) and Unanimous Voting (UV) schemes, and using accounting information, economic information, past share and index returns information as well as information about the industrial classification of around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. After experimentation, we found that all classification methods produce consistent excess returns for the out-of-sample years 1993-97, whereas greater gains result from UV in terms of classification accuracy, profitability, and trading volume. Using accounting information results in a very accurate and profitable trading system, whereas additional benefits for the individual classifiers and voting methodologies arise after adding non-accounting information. After applying LDA and PNN to homogeneous U.K. industrial sectors using accounting information only, we found that both classifiers produce consistent excess returns after restricting the sample to service companies, whereas additional benefits arise after adding utility, financial and property companies. On the other hand, the high performing portfolios that result after restricting the sample to manufacturing and extractive companies fail to produce consistent excess returns. After restricting our sample to service companies and combining our five classifiers namely LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes, we found that all classification methods produce consistent excess returns. However, greater gains result from UV in terms of classification accuracy, profitability, and trading volume.

To avoid the possibility of overfitting, we investigate a variety of methodologies to reduce the dimensionality of our data including principal component analysis (PCA) as well as linear and non-linear dimensionality reduction techniques based on neural networks. After experimentation, we found that neural network linear PCA (NN-PCA) and neural network non-linear PCA (NN-NLPCA) explain a higher proportion of variation in the original set of variables than the common PCA methodology. On the other hand, we found that the resulting principal components (PCs) from NN-PCA and NN-NLPCA are competitive to other dimensionality reduction techniques in maintaining important discriminating power to identify which shares are likely to have exceptional returns in the future. We verify further the effectiveness of the neural network dimensionality reduction methodologies by applying them to homogeneous subsets of financial ratios and using the derived PCs to assess the long-term credit standing of U.K. debt issuers. We found that NN-PCA and NN-NLPCA architectures can be successfully implemented as a preliminary step to assess the credibility of U.K. debt issuers and at the same time provide an alternative solution to overfitting.

Our results provide evidence for the relatively greater ability of non-linear classification methods over the linear model to predict financial events in the bond and stock markets. In response to previous studies that support combinations of homogeneous component classifiers rather than heterogeneous classifiers, we provide substantial evidence that a combination of heterogeneous classifiers using the UV scheme is also successful and produces impressive results over the individual component classifiers.

CHAPTER 1: INTRODUCTION

1.1 THEORETICAL BACKGROUND

The aim of this thesis is to explore the ability of statistical classification methods to predict events in financial markets. The classification methods include conventional linear discriminant analysis, and a number of less familiar non-linear techniques such as the probabilistic neural network. We look at two applications. In one case we try to predict the credit ratings of bonds. In the other case we try to predict which shares will strongly outperform the stock market as a whole.

This is an interesting exercise for a number of reasons. One is commercial. Predictability in asset prices may be translatable into profits for traders in these markets. Another is more academic. Predictability casts some doubt on the benchmark "efficient markets" theory that has dominated the modern theory of finance for the past fifty years. In addition we have made some technical innovations in forecasting technology that may have implications beyond this study of financial markets. For example, data pre-processing is known to be important for the effective use of non-linear statistical models, and we have carefully investigated this aspect of model implementation. Combining forecasts from a number of diverse models is also known to be effective in forecasting expected values. We have investigated whether this proposition carries over into the domain of event forecasting.

This Chapter sets the scene for these empirical investigations. We first review the background to the study, and then summarise the steps we have taken.

Any study of predictability in financial markets is immediately confronted with the limitations imposed by the efficient market hypothesis, and its close relative the random walk hypothesis. Nobel Laureate Paul Samuelson (1965) claimed that price changes will be wholly unforecastable if they fully incorporate the expectations of the market participants. Changes in prices will reflect only new information, and "news" by definition arrives randomly, so changes in prices will themselves follow a random walk.

Subsequently this extreme view has moderated somewhat. Fama (1970) pointed out that "information" is an ambiguous concept, and defined three degrees of market efficiency, depending on just how much information is impounded in asset prices. In a "weak form" efficient market only the past history of the asset price is used, so excess returns cannot be

made by exploiting past patterns in the asset price, but might conceivably be made using other information. In a "semi-strong form" efficient market, publicly available information outside the asset market is also used, so for example information on economic indicators cannot be used to earn excess returns. In a "strong-form" efficient market, all information, including private "inside" information, is used in trading. Only in this extreme case does the Samuelson result of wholly unforecastable prices emerge.

The capital asset pricing model (CAPM) of Sharpe (1964), Lintner (1965), and Mossin (1966) states that some assets will predictably yield higher returns than others. But the predictable excess return – the "drift" in any random walk in asset prices - will entirely reflect risk as measured by the "beta" of the asset. So while stock returns are to some degree predictable, this predictability cannot be exploited to earn excess, risk-adjusted, profits. This kind of result is also characteristic of more recent asset pricing theories such as the Arbitrage Pricing Theory, which generalises the CAPM by making the risk premium depends on a number of factors in addition to beta.

A substantial number of empirical studies in the literature have examined the predictability of stock returns. Some relevant studies are those by Chen et al. (1986), Campbell (1987), Fama and French (1988, 1989, 1992, 1993), Balvers et al. (1990), Cohrane (1991), Ferson and Harvey (1993b), Glosten et al. (1993), Whitelaw (1994), and Pesaran and Timmermann (1994, 1995). The results of these studies generally support the robustness of the efficient market theory. For example, in cross-sectional data, even though we may able to identify in advance shares that outperform the market index, these shares will be in general riskier. Similarly, from time series data, even though we may be able to identify whether one year will be better than another year for the stock market as a whole, evidence suggests that the high-earning years will be years of high risk as well. Finally, even if we succeed in identifying genuine "anomalies" – violations of the efficient market theory - the financial profits that emerge from econometric models of stock markets tend to be eaten up by transaction costs.

However, some potentially exploitable patterns do seem to be established, beyond what might be expected on the basis of differences in risk. First, empirical studies suggest several fundamental variables that might be able to explain the cross section of expected returns. These include among others firm size (Banz 1981; Reinganum 1981a), earnings yield (Basu, 1983), dividend yield (Fama and French, 1988), leverage (Bhandari, 1988), the ratio of the firm's book-to-market equity (Stattman, 1980; Rosenberg et al. 1985; Chan et al. 1991; Fama and French 1992) or interactions among these variables (Reinganum 1981b; Jaffe et al. 1989; Chan et al. 1991). Second, other empirical studies suggest that stock returns can be predicted by other

means of publicly available information such as time-series data on various economic variables, especially those with an important business cycle component. This conclusion holds true over different investment horizons and across different markets. Economic variables that have been found to be highly correlated with stock returns include among others dividend yields, measures of interest rates, measures of inflation, and growth rates of industrial production. Another important empirical finding is that statistical models and moving average trading rules result in excess profits while there is no consistency in the performance of the fund managers from one year to the next (Brock et al. 1992; Malkiel 1996).

Many of the tests of asset pricing models assume that expected returns, betas, and risk premia are constant through time and independent of any ex-ante known information. However, relevant research on time-series predictability documented evidence that expected returns are not constant through time. In response to these findings, some researchers have attempted to integrate the time-series properties of conditional moments with the cross-sectional implications of asset pricing models in a framework known as conditional asset pricing models. Some commonly used conditional asset pricing models include among others latent variable models, ARCH-type models, and GARCH-type models. These and other similar models allow for time-varying expected returns and assume that the return distribution depends on a set of ex-ante observable variables. Therefore, these models are more consistent with the evidence reporting predictable components in the time series of returns. They provide a useful framework to integrate the time series predictability of stock returns with the cross-sectional implications of asset pricing models. However, conditional asset pricing models have two major drawbacks. The first drawback is that a particular model for the conditional expectations has to be specified. The form of the function depends on the joint probability distribution of the returns and the instrumental variables. Although most studies have adopted a linear relation for the first moments, other relations may also be valid. The second drawback is that it may be difficult to conclude anything about the validity of an asset pricing model given the fact that all information that investors use to set prices is unobservable and that the econometrician only uses a reduced set of information.

All the econometric methods that have been discussed above have been designed to detect either linear structure or strictly defined forms of non-linearities in the financial data. For example, the CAPM and the APT are based on linear models of expected returns, whereas ARCH- and GARCH-type models are based on strictly defined non-linear models. Therefore, all these models might not be the most suitable if the data is fuzzy, chaotic, or exhibits unpredictable non-linearities. A number of recent studies have suggested the existence of non-linear structures in many economic and financial variables (Hinich and Patterson 1985;

Abhyankar et al. 1997). For example, the investors' attitudes toward risk and expected return are non-linear. On the other hand, the strategic interactions among market participants, the process by which information is incorporated into security prices, and the dynamic fluctuations of the economic environment are all inherently non-linear (Campbell et al., 1997). Given the evidence of non-linearities in many financial and economic variables, a number of researchers have examined the predictability of stock returns under the non-linear framework. Pesaran and Timmermann (1994) found significant non-linear structures in quarterly and monthly regressions of excess returns on economic variables and some non-linear terms such as the lagged values of the squared returns. Qi and Maddala (1999) provide evidence that a neural network model can improve upon the linear regression model in terms of predictability but not in terms of profitability.

Further to the above empirical findings, we have to consider that several factors such as human judgements, human emotions, human feelings, human expectations, psychology, politics, and other qualitative factors affect in a high degree the process that drives stock prices. Under these considerations, it is obvious that most of the models that have been used so far to predict stock returns may not be able to deal with the actual process in the financial data that is unpredictable and lacks a well-defined physical content. More powerful models should be applied that will be able to extract the hidden knowledge in the financial data that cannot be detected by either linear or well-defined non-linear models. On the other hand, we have to consider that the ultimate purpose of stock return predictability is profitability. A well-defined theoretical model might not be the most efficient solution to stock return predictability if it is not profitable as well.

In the last decades, various researchers have considered combining forecasts in the hope that a combination of forecasts can be found which yield final classifications superior to those of each individual forecast. There is little doubt that combining individual forecasts improves forecast accuracy. The conclusion holds true in statistical forecasting, judgmental estimates as well as averaging statistical and subjective predictions (Clemen 1989, Makridakis 1989). Empirical results suggest three factors that encourage combinations of forecasts: first, combining individual forecasts improves forecasting accuracy and reduces the variance of forecasting errors; second, simple combination models work as well as more complex combinations; and third, combining can be done with little or no increase in cost. Apart from lower forecast errors, combinations of forecasts allow for greater flexibility in the model switching sense to utilise a broader information set. In this content, a combination of diverse models may be able to utilise a broader information set than a combination of similar models. However, a better utilisation of the information set may lead to an increase in forecasting accuracy. Therefore, forecast

diversity may be an essential technique in improving forecasting accuracy.

In recent years, many researches have applied composite classifier architectures and mixture models to various financial applications (Faria and Souza, 1995; Waterhouse, 1997). However, although there have been a number of successful efforts at combining homogeneous classifiers including multiple decision trees (Breiman et al. 1984), multiple neural networks (Maclin and Shavlik, 1995), and nearest neighbour algorithms (Skalak, 1997), searching the more complex space of sets of heterogeneous component classifiers has not been shown necessarily to achieve high accuracy (Battiti and Colla, 1994; Breiman 1994).

1.2 AIMS AND OBJECTIVES OF THE THESIS

The focus of this thesis is to predict events in financial markets by using individually and combining a "portfolio" of five heterogeneous classifiers namely

- Linear Discriminant Analysis (LDA)
- Probabilistic Neural Network (PNN)
- Learning Vector Quantization (LVQ)
- Oblique classifier (OC1)
- Ripper Rule Induction classifier (RRI)

We use two combining rules:

- Majority voting (MV)
- Unanimous voting (UV)

Applying these models to share price performance, for example, we attempt to answer the following questions:

1) How can classification algorithms from different model families be applied individually to correctly classify and predict which shares are likely to have exceptional returns in the future ?

2) How can classifiers from different model families be combined to correctly classify and predict which shares are likely to have exceptional returns in the future ?

3) How can classifiers from different model classes be combined to create a composite

classifier with higher accuracy than the individual component classifiers to correctly classify and predict which shares are likely to exhibit exceptional returns in the future ?

4) What types of information should we use to increase the accuracy of either an individual classifier or a composite classifier architecture to correctly classify and predict which shares are likely to have exceptional returns in the future ?

5) What data preprocessing techniques should we undertake in order to improve more the classification accuracy and reduce the computational expense of the proposed algorithms ?

The motivation for combining heterogeneous classifiers is the possibility that by combining a set of heterogeneous classifiers, we may be able to perform classification better than the individual classifiers. This motivation is supported by the following ideas:

- A group decision based on "different" experts may be more reliable on average than the decision of the individual expert or the decision of "similar" experts. If the predictions of the component classifiers are exactly the same, then a combination of these predictions will not improve the predictive accuracy of the composite classifier. On the other hand, if the individual components make some different predictions, then there is a hope that combining their predictions using the appropriate mechanisms might improve the predictive accuracy of the composite architecture.

- A composite architecture based on heterogeneous component models will be more flexible to capture the structure of the data set if there are sub-areas with different underlying processes. For example, if the data set is non-linear, then a composite architecture that combines linear components may not perform well. On the other hand, a composite architecture that combines both linear and non-linear components will be more flexible to capture the underlying structure of the data.

- A composite classifier architecture, which combines heterogeneous component classifiers, will be able to hedge the classification bets better than an individual classifier or an architecture which combines homogeneous component classifiers. In this sense, a "portfolio" of classifiers is analogous to a "portfolio" of financial instruments where diversification is used to reduce risk (Skalak, 1997).

- In a classification setting, an algorithm that combines the predictions of a set of heterogeneous components may learn from the fact that a particular classifier misclassifies some particular instances. Even if a unanimous decision has not been reached, something has been learned from the fact that a specific individual component classifier has a different opinion on a given issue (Skalak, 1997).

The general hypothesis examined in this thesis can be stated as follows:

The ability of any classifier to predict high performing shares can be improved or exceeded through composite classifier architectures that combine a small number of heterogeneous classifiers using voting procedures.

The particular types of statistical models studied in this thesis are ones that predict outcomes from observed data. Prediction of continuous variables is known as regression, whereas prediction of categorical variables is known as classification. This thesis will be concerned more with classification models.

1.3 ORIGINAL CONTRIBUTION OF THE THESIS

In this thesis, we extend previous research that examined the predictability of share returns mostly under restricted forms of linear and non-linear models. We propose a variety of new methodologies to address the problems of stock return predictability and stock selection by combining a "portfolio" of heterogeneous classifiers through voting procedures and applying sophisticated data preprocessing techniques to correctly classify and predict stocks that are likely to have exceptional returns in the future. We also demonstrate the applicability of non-linear classification methods to predict long-term bond ratings as well as the long-term credit standing of debt issuers. Our original contribution can be summarised as follows,

1) We provide a very thorough and extensive review of the relevant literature in stock return predictability, supervised classification, and combination of forecasts.

2) We investigate the distributional properties of financial ratios and we compare three heterogeneous classifiers namely, LDA, PNN, and RRI in terms of their ability to predict long-term bond ratings. The final outcome of this experiment is a new quantitative system to predict long-term bond ratings using probabilistic neural networks and rule induction techniques.

3) We propose five new methodologies to address the problems of stock return predictability

and stock selection by applying five existing classification methods namely LDA, PNN, LVQ, OC1, and RRI to correctly classify and predict which shares are likely to have exceptional returns in the future using accounting information.

4) We propose five new methodologies to address the problems of stock return predictability and stock selection by applying five existing classification methods namely LDA, PNN, LVQ, OC1, and RRI to correctly classify and predict which shares are likely to have exceptional returns in the future using economic information, past share and index returns information, as well as information about the industrial classification of the companies included in our sample.

5) We propose five new methodologies to address the problems of stock return predictability and stock selection by applying five existing classification methods namely LDA, PNN, LVQ, OC1, and RRI to correctly classify and predict which shares are likely to have exceptional returns in the future by mixing accounting and non-accounting information.

6) We propose two new methodologies to address the problems of stock return predictability and stock selection by combining five existing heterogeneous classification methods through MV and UV schemes to correctly classify and predict which shares are likely to have exceptional returns in the future using accounting information.

7) We propose two new methodologies to address the problems of stock return predictability and stock selection by combining five existing heterogeneous classification methods through MV and UV schemes to correctly classify and predict which shares are likely to have exceptional returns in the future using economic information, past share and index returns information, as well as information about the industrial classification of the companies included in our sample.

8) We propose two new methodologies to address the problems of stock return predictability and stock selection by combining five existing heterogeneous classification methods through MV and UV schemes to correctly classify and predict which shares are likely to have exceptional returns in the future by mixing accounting and non-accounting information.

9) We propose a new methodology to address the problems of stock return predictability and stock selection by applying the UV methodology over two parallel implementations of the classifiers using accounting and non-accounting information, respectively, and then

assigning a share to the high performing portfolio only if the five classifiers from the first implementation based on accounting information as well as the same classifiers from the second implementation based on non-accounting information agreed unanimously on their decisions.

10) We propose a new data preprocessing methodology for linear and non-linear classification methods that predict high performing shares using existing dimensionality reduction techniques based on neural networks and we examine the applicability of this methodology to correctly classify and predict which shares are likely to have exceptional returns in the future.

11) We propose a new data preprocessing methodology for linear and non-linear classification methods that predict long-term credit ratings using existing dimensionality reduction techniques based on neural networks and we examine the applicability of this methodology to correctly classify and predict long-term bond ratings.

Some parts of this thesis have been presented at international conferences, whereas some other parts either have been submitted or they are going to be submitted for publication at international journals. Relevant information about these papers can be found in the references Section.

1.4 OVERVIEW OF THE THESIS

This remaining of this thesis is divided into nine Chapters as follows:

In Chapter two, we discuss asset pricing models and alternative approaches that have been suggested in the literature to explain the cross-section of stock returns. This Chapter is divided into three parts: In the first part, we discuss theory and empirical evidence on asset pricing models. We discuss unconditional tests as well as conditional versions of asset pricing models. We also present evidence for the ability of firm specific variables to explain the cross-section of stock returns. In the second part, we present evidence on stock return predictability using past returns and macroeconomic variables and we discuss briefly empirical evidence on seasonal patterns in returns. Finally, in the third part, we summarise the discussion and we provide the conclusions.

In Chapter three, we discuss the main approaches to statistical classification focusing on the supervised learning paradigm. This Chapter is divided into four parts: In the first part, we

discuss parametric, semi-parametric, and non-parametric smoothing classification rules. In the second part, we review neural networks, and data mining techniques focusing on decision trees and rule induction algorithms. In the third part, we discuss empirical evidence from comparative studies on supervised learning algorithms. Finally, in the fourth part, we present the summary and conclusions.

In Chapter four, we investigate the distributional properties of financial ratios and we compare three heterogeneous classifiers namely, LDA, PNN, and RRI in terms of their ability to predict long-term bond ratings. The results of this experiment suggest that non-linear models do not depend on distributional assumptions in the same degree as the linear model and they are more flexible to deal with unpredictable non-linearities and other complex processes in the financial data. The outcome of this implementation is a new methodology to predict long-term bond ratings using probabilistic neural networks and rule induction techniques. We discuss the economic implications of these findings and we provide the conclusions.

In Chapter five, we present the first implementation of the central idea in this thesis which is a new methodology to identify outperforming shares using five classification methods namely, LDA, PNN, LVQ, OC1, and RRI classifier. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This data consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our predictor variables are 38 accounting ratios drawn from published accounting statements. After experimentation, we found that all classification methods produce consistent excess returns in ex ante forecasting, whereas there are some inconsistencies in the classification accuracy and profitability of the classifiers from one year to the next. We discuss the economic implications of these findings and we provide the conclusions.

In Chapter six, we investigate the potential to apply composite classifier architectures to predict high performing shares by intelligently combining the five heterogeneous classifiers, namely LDA, PNN, LVQ, OC1 and RRI through MV and UV schemes. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97, whereas our predictor variables are 38 accounting ratios drawn from published accounting statements. This Chapter is divided into three parts: In the first part, we discuss design specifications to construct composite models and we discuss the literature review in combining forecasts. In the second part, we investigate empirically the possibility of combining our five classification methods, namely LDA, PNN, LVQ, OC1 and RRI and we apply the resulting predictions to predict high performing shares. We discuss our proposed composite classifier architecture and we explain in detail the criteria we considered for the selection of the individual component classifiers that we

used to build this architecture. After experimentation, we found that the UV scheme produces significant improvements in both classification and profitability over the individual classifiers and reduces substantially the trading volume. Finally, in the third part, we summarise the discussion and we provide the conclusions.

In Chapter seven, we extend our methodology to predict high performing shares. We perform two different experiments: In the first experiment, we compare and contrast the five classifiers namely, LDA, PNN, LVQ, OC1 and RRI and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume using three different subsets of information: first, using accounting information only; second, using economic information, past share and index returns information, and industrial classification information only; and third, using all the available information by mixing accounting and non-accounting information. In this experiment, we apply both the MV and UV methodologies for each individual implementation, respectively. The findings of this experiment suggest that the UV principle produces significant improvements in classification and profitability if compared to the individual classification methods and reduces substantially the trading volume. On the other hand, using accounting information results in a very accurate and profitable trading system, whereas additional benefits for individual classifiers and voting methodologies arise after adding non-accounting information. In the second experiment, we apply the UV methodology over two parallel implementations of the classifiers using accounting and non-accounting information, respectively. According to this latter implementation, a share is not assigned to the to the high performing portfolio unless the five classifiers from the first implementation based on accounting information as well as the same classifiers from the second implementation based on non-accounting information agree unanimously on their decisions. The results from experiment suggest that there are additional benefits after implementing the classifiers in parallel using accounting and non-accounting subsets of information, respectively. We discuss the economic implications of these findings and we provide the conclusions.

In Chapter eight, we examine the potential of identifying outperforming shares in homogeneous U.K. industrial sectors by combining the five classification methods namely, LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1994-97. Our predictor variables are 38 accounting ratios drawn from published accounting statements. After applying the LDA and the PNN, we found that both classifiers produce consistent excess returns if we restrict the sample to service companies, whereas additional benefits may arise after adding utility, financial, and property companies. On the other hand, the high performing portfolios that result if we restrict the sample to manufacturing and extractive companies are not particularly profitable and fail to

produce consistent excess returns. After restricting the sample to service companies and combining the five classification methods through UV and MV voting schemes, we found that the UV principle produces significant improvements in classification and profitability if compared to the individual classification methods and reduces substantially the trading volume.

In Chapter nine, we investigate alternative approaches to dimensionality reduction including principal component analysis (PCA), neural network linear PCA (NN-PCA), and neural network non-linear PCA (NN-NLPCA). This Chapter is organised as follows: In the first part, we use dimensionality reduction techniques based on neural networks and we apply the five classification methods namely, LDA, PNN, LVQ, OC1, and RRI to test the ability of the resulting principal components (PCs) to predict which shares are likely to have exceptional returns in the future. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This data consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our predictor variables are 38 accounting ratios that were drawn from published accounting statements. After experimentation, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology. On the other hand, we found that the resulting PCs from NN-PCA and NN-NLPCA are competitive to other dimensionality reduction techniques in maintaining important discriminating power to identify which shares are likely to have exceptional returns in the future. In the second part, we apply the same techniques to reduce the dimensionality of a small data set of financial ratios and we apply three classification methods namely, LDA, PNN, and Backpropagation Neural Network (BPNN) to test the ability of the resulting PCs to predict the long-term credit standing of debt issuers. Our target data are 120 rated debt issuers, whereas our predictor variables are thirty financial ratios that were drawn from published accounting statements. The results of this experiment confirm the findings of the first experiment. More specifically, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology. Furthermore, we found that the PCs extracted from NN-PCA and NN-NLPCA are better discriminators than the PCs extracted from PCA and are easier to interpret if extracted from homogeneous groups of financial ratios. Overall, the results of this experiment suggest that linear and non-linear dimensionality reduction techniques based on neural networks can be an efficient tool to assess the long-term credit standing of debt issuers and at the same time provide an efficient solution to overfitting. Finally, in the third part, we present the summary of this Chapter and we provide the overall conclusions.

Finally, in Chapter ten, we conclude the thesis with an overview of the main findings and suggestions for future research.

CHAPTER 2: ASSET PRICING MODELS AND ALTERNATIVE APPROACHES TO STOCK RETURN PREDICTABILITY

In this Chapter, we discuss asset pricing models and alternative approaches that have been suggested in the literature to explain the time series behaviour and cross-section patterns of stock returns.

Models of asset pricing depict the theoretical equilibrium relationship between expected returns and risk. They have direct implications for any exercise that attempts to predict asset returns. If actual asset prices cluster around the theoretical prices, and deviate from them only by a random disturbance, then we say that the market is efficient. This does not mean that asset prices are necessarily unpredictable. Indeed, asset prices will certainly be predictable if changes in risk or risk aversion are predictable. However, this kind of predictability will not be associated with any market inefficiency, in the sense that it cannot be exploited to make excess risk-adjusted, returns.

On the other hand, asset prices may deviate from the theoretical price in some systematic way – for example, if actual prices converge slowly on equilibrium prices, or if certain types of asset are persistently under- or over-valued. In this case, when the market is not fully efficient, research on asset pricing theory can be used to identify factors that can help to predict potential excess returns. This is the main motivation for this Chapter.

Several well-specified asset pricing models have been developed and tested empirically over the past decades. The best known is the Capital Asset Pricing Model (CAPM) which predicts that the expected return of an asset will be linearly related to the covariance of its return with the return on the market portfolio. Empirical research has provided evidence that is inconsistent with the predictions of the CAPM. Evidence suggests instead that the cross-section of stock returns is better described by means of publicly available information such as firm specific variables, past returns, and macroeconomic variables. Some other researchers have attempted to integrate the time series predictability of stock returns with the cross-sectional implications of asset pricing models in a framework known as conditional asset pricing models. These models allow for time variation in expected returns, covariances (betas), and risk premia and assume that the return distribution depends on a set of ex-ante observable variables. Although there are several advantages to justify the use of conditional asset pricing models, their theoretical

properties are not always attractive. In this Chapter, we review briefly the theory on asset pricing models and we discuss the empirical evidence bearing on them.

This Chapter is divided into three parts: In the first part, we discuss theory and empirical evidence on asset pricing models. We discuss unconditional tests as well as conditional versions of asset pricing models. We also present evidence for the ability of firm specific variables to explain the cross-section of stock returns. In the second part, we present evidence on stock return predictability using past returns and macroeconomic variables and we discuss briefly empirical evidence on seasonal patterns in returns. Finally, in the third part, we summarise the discussion and we provide the conclusions.

Part One: Asset Pricing Models and Stylised Facts

2.1 THE CAPITAL ASSET PRICING MODEL

2.1.1 Theoretical Background

One of the most important problems in the modern history of finance is the quantification of the tradeoff between risk and expected return. The first attempt to quantify risk and the reward for bearing it was the development of the Capital Asset Pricing Model (CAPM). The foundation of the CAPM is due to the work of Harry Markowitz. In 1959, Markowitz demonstrated how to create a frontier of optimal portfolios. He suggested that each of these portfolios can be characterised as mean-variance efficient because it has the highest expected return for a given level of risk. Focusing on the impact of optimal portfolio formation within a frictionless marketplace, financial economists began to investigate if Markowitz's model can influence the valuation of securities. Sharpe (1964), Lintner (1965), and Mossin (1966) showed that if investors have homogeneous expectations and optimally hold mean-variance efficient portfolios, then the market portfolio will also be mean-variance efficient if there are no market frictions. Furthermore, assuming that there is a risk-free asset such that investors can borrow or lend unlimited amounts at a risk-free rate of interest, Sharpe, Lintner, and Mossin developed the CAPM that can be written as follows,

$$E(R_i) = R_f + \beta_{im}[E(R_m) - R_f] \tag{2.1}$$

$$\beta_{im} = \frac{Cov(R_i, R_m)}{Var(R_m)} \tag{2.2}$$

The above model is derived under the additional assumptions that investors act as risk-averse individuals who maximise their expected utility in a single-period investment horizon. Furthermore, it is assumed that the quantities of assets are marketable and perfectly divisible and there are no market imperfections such as taxes, regulations, and short selling. Eq. (2.1) states that the expected rate of return on any asset i, $E(R_i)$, equals the risk-free rate of return, R_f, plus a risk premium, $\beta_{im}[E(R_m) - R_f]$. The risk premium can be thought of as the extra compensation above the risk-free rate of return that the investors require for investing in the market portfolio. It is the product of the quantity of risk, β_{im}, multiplied with the price of risk, $[E(R_m) - R_f]$. The quantity of risk, β_{im}, is defined in Eq. (2.2) as the covariance between the

return on the risky asset, R_i, and the return on the market portfolio, R_m, divided by the variance of the return on the market portfolio. The price of risk is the difference between the expected rate of return on the market portfolio, $E(R_m)$, and the risk-free rate of return, R_f.

In 1972, Black relaxed the assumption that there exists a risk-free asset such that investors can borrow or lend unlimited amounts at a risk-free rate of interest. In Black's model the risk-free rate is replaced by a zero-beta portfolio whose return is uncorrelated with the market portfolio return. The zero-beta CAPM postulates the same linear relationship as the original CAPM.

2.1.2 Empirical Evidence on the CAPM

An enormous amount of empirical evidence has been presented in the literature concerning the validity of the CAPM. The first major empirical studies of the CAPM were conducted in the early 1970s. Blume and Friend (1973) examined the relationship between return and risk for common stocks traded on the New York Stock Exchange (NYSE) for the 1955-1968 period. Consistent with the CAPM, they found a linear relationship between realised returns and systematic risk. Black et al. (1972) performed tests of the CAPM using all stocks traded on the NYSE during the period 1926 through 1965. For each year, they estimated the betas using monthly data for the previous five years. After estimating the betas, they ranked the stocks on the basis of betas and they formed ten portfolios. The ten percent of the stocks with the highest betas included into portfolio one, the next ten percent of the stocks with the highest betas included into portfolio two, and the same procedure was repeated through portfolio ten. After performing a time-series test of the CAPM, Black et al. found that high-risk portfolios have systematically lower returns than those implied by the CAPM, whereas low-risk portfolios have systematically greater returns than the returns implied by the model. They suggested that these findings are inconsistent with the predictions of the CAPM. On the other hand, after performing a cross-sectional examination of the model, the test results documented a linear relation between mean excess returns and beta. Furthermore, the results suggested that the excess return on the beta factor has nonzero mean and it is nonstationary over time. Black et al. suggested that this evidence supports a two-factor model to explain the cross-section of stock returns.

Fama and MacBeth (1973) examined two hypothesis: 1) if the relationship between risk and return is linear; and 2) if unsystematic risk has any impact on return. To examine these hypotheses, they run monthly cross-sectional regression equations for 20 portfolios composed of NYSE-listed stocks over the period 1935 to 1968. Their results suggested that in spite of non-linearities and a significant relationship between returns and unsystematic risk in some sub-periods, the data for the 33-year period as a whole are consistent with the basic implications of the CAPM.

In 1977, Roll criticised all the previously empirical tests of the CAPM that used proxies of the true market portfolio. He observed that the only legitimate test of the CAPM is whether or not the true market portfolio is mean-variance efficient. A proxy for the market portfolio might be mean-variance efficient even when the true market portfolio is not. Therefore, the tests of the model that are performed with any market portfolio other than the true market portfolio are not true tests of the CAPM unless the exact composition of the true market portfolio is known and used in the tests. However, a direct test of the proxy's mean-variance efficiency is difficult statistically because the sampling distribution of the efficient set is generally unknown. In response to Roll's criticism, Staumbaugh (1982) performed tests using broader market indexes that included bonds, government bonds, treasury bills, home furnishings, residential real estate and automobiles, in additions to common stocks. The results of this study showed that the zero-beta form of the CAPM seems to be more robust to various definitions of the market portfolio.

Reinganum (1981a) investigated empirically whether securities with different estimated betas systematically experience different average rates of return using daily returns of NYSE and AMEX companies during the years 1964 through 1979. The test results demonstrated that NYSE-AMEX stock portfolios with widely different estimated betas possess statistically indistinguishable average returns across securities during the years 1964 through 1979. The test results demonstrated further that the average returns of high beta stocks are not reliably different from the average returns of low beta stocks.

Corhay et al. (1987) reported evidence on the CAPM using U.K., France and Belgium data. Applying the Fama and MacBeth (1973) methodology and using monthly U.K. stock returns over the 1957-1983 period, they found that the market premium is insignificant and negative. After performing a month-to-month analysis, they found that systematic risk is positive only in April, whereas it is significantly negative in May and November. Contrary to the CAPM, Corhay et al. found that unsystematic risk is positively priced. On the other hand, using French and Belgium data, they reported that the market premium is positive in January but it is negative for the rest of the year. Other studies reporting negative risk premia in international markets are those by Poon and Taylor (1991) and Levis (1995) for the U.K., and Rubio (1988) for Spain.

With few exceptions, the empirical studies on the CAPM that were conducted in the early 1970s agree on the following conclusions: first, the evidence shows a significant relationship between realised returns and systematic risk as measured by beta; second, the relationship between risk and return appears to be linear; and third, versions of the model that include a squared beta term or unsystematic risk find that these explanatory factors are useful only in a small number of the time

periods sampled. On the other hand, reviewing the empirical studies on the CAPM after 1977, we can see that other studies support the CAPM, whereas other studies reject the CAPM. However, none of these studies provides a definite test on the CAPM.

Since the CAPM was not unanimously supported by the tests, many researchers formulated alternative equilibrium models by relaxing the CAPM assumptions. These include among others multiperiod CAPM models that relax the unrealistic assumption of the one-period investment horizon. According to these models, investors make their utility-maximising portfolio decisions by considering not only the returns on alternative assets over just one period but also in subsequent periods. In addition, trading in assets takes place continuously over time. Three of the extensions of the CAPM model that deserve particular attention are the multibeta CAPM derived by Merton (1973), the consumption CAPM derived by Rubinstein (1976), and the inflation CAPM derived by Friend et al. (1976).

2.2 FIRM SPECIFIC VARIABLES AND STOCK RETURN PREDICTABILITY

In the late 1970s less favourable evidence for the CAPM began to appear after a number of studies suggested that firm-specific variables such as the price to earnings ratio and the market capitalisation of common equity can be used to explain the cross section of stock returns beyond the beta of the CAPM. Other studies extended the list of the predictive variables to include the ratio of price to book, the debt to equity ratio, and other similar variables. These findings represent a set of stylised facts that can be used as a basis for more general multifactor asset pricing models. These stylised facts are discussed in the next sections.

2.2.1 The Size Effect

A number of studies have been presented in the literature to discuss the ability of firm size measured by the market value of common equity to explain the cross-sectional variation of stock returns. Klein and Bawa (1976) found that if insufficient information is available about a subset of securities, investors may not hold these securities because of uncertainty about the true parameters of the return distribution. They pointed out that if investors have access to different amount of information, they might limit their diversification to different subsets of all securities in the market. However, if the amount of information generated is related to the size of the firm, then many investors may not desire to hold the common stock of very small firms. In view of these considerations, Banz (1981) examined the empirical relationship between the return and the total market value of NYSE common stocks during the 1936-1975 period. He found that the common stock of smaller firms has higher risk-adjusted returns on average than larger firms. Banz also observed that securities held by only a small subset of the investors have higher risk-adjusted returns than those considered by all investors. He concluded that lack of

information about small firms might lead to limited diversification and therefore to higher returns for the undesirable stocks of small firms.

Reinganum (1981b) used daily data for NYSE and AMEX firms over the 1963-1977 period. He showed that portfolios of small firms have significantly higher average returns than portfolios of large firms. Roll (1981) suggested that the size effect might be a statistical artifact of improperly measured risk due to infrequent trading of small stocks. In response to this criticism, Reinganum (1982) used methods to account for nonsynchronous and infrequent trading in estimating betas. He reported that the magnitude of the size effect is not really sensitive to the particular method for estimating betas. Blume and Stambaugh (1983) observed that Reinganums' findings are actually overstated because transaction costs and bid-ask spread differentials are ignored. They suggested that if annual re-balancing is considered, then the magnitude of the size effect might be halved. In a latter study, Reinganum (1992) demonstrated that the small-firm premium for over-the-counter stocks is much lower than NYSE and AMEX stocks. He attributed these findings to differences in market microstructure.

Keim (1983) investigated the month to month stability of the size anomaly over the 1963-1979 period. He found that nearly fifty percent of the average magnitude of the risk-adjusted premium of small firms relative to large firms over this period is due to anomalous January abnormal returns. Brown et al. (1983) stated that at least part of the size effect can be explained by an omitted risk factor in the pricing model. Keim (1983) suggested that even if part of the average size effect is due to an unspecified risk variable, the behaviour observed in January cannot be attributed solely to this cause because risk alone cannot explain a return premium observed in the same month each year. Stoll and Whaley (1983) observed that transaction costs can explain the size effect because such costs prevent arbitrageurs from eliminating the average return differential. However, Keim (1983) suggested that only if transaction costs are seasonal in nature, implying some degree of market power for market makers, can such costs explain the January effect.

Chan and Chen (1988) attributed the size effect to measurement errors in betas that allow market capitalisation to serve as a proxy for the market beta. They showed that the explanatory power of size might disappear if data from a long period of time are used to estimate beta. On the other hand, Jegadeesh (1992) documented that the size effect cannot be explained by beta, and beta is indeed insignificant after sorting portfolios on the basis of size first and then beta. Fama and French (1992) also formed portfolios to allow for variation in beta that is unrelated to size by ranking by size first and then beta. They found no relation between beta and returns for non-financial firms in the NYSE, AMEX and NASDAQ stocks over the 1963-1990 period. In

addition, they reported that size appears to be always significant.

Other studies that examined the size effect were those of Corhay et al. (1987) for the U.K., Hawawini and Viallet (1987) for France, Rubio (1988) for Spain, Hawawini et al. (1989) for Belgium, Calvet and Lefoll (1989) for Canada, and Chan et al. (1991) for Japan. The findings of these studies suggest that there is evidence for a negative relationship between returns and size in all countries except Canada and France. The results of these and other international studies are discussed in more detail in Hawawini and Keim (1995).

2.2.2 The Price to Earnings Effect

Several studies documented that price to earnings (P/E) ratios can be used as indicators of the future investment performance of a security. Nicholson (1960) investigated the relationship between P/E multiples and subsequent total returns. He found that low P/E stocks consistently provide greater returns than the average stock. Basu (1977) investigated whether the investment performance of common stocks is related to their P/E ratios. A total number of 753 industrial firms that traded on the NYSE between September 1956 and August 1971 were selected for this study. Beginning from the first year, the P/E ratio of every sample security was computed. These ratios were then ranked and five portfolios were formed. The results suggested that low P/E portfolios seem to have higher absolute and risk-adjusted rates of return than the high P/E portfolios during the 1957-1971 period. Basu observed that this is also true when bias on the performance measures resulting from the effect of risk is taken into account. These results are consistent with the view that P/E ratio information is not fully reflected in security prices in a rapid manner as postulated by the semi-strong form of the efficient market hypothesis. Reinganum (1981b) reported that the P/E effect disappears for both NYSE and AMEX stocks after controlling for size but there is a significant size effect after controlling for the P/E ratio. These results suggest that the P/E ratio effect is a proxy for the size effect and not vice-versa.

Levis (1989a) showed that there is a significant P/E effect on the London Stock Exchange during the 1961-1985 period, whereas Chou and Johnson (1990) observed a significant P/E on the Taiwan Stock Exchange during the 1979-1988 period. The results of these studies as well as the results of other international studies are discussed in more detail in Hawawini and Keim (1995). In summary, the bulk of the evidence suggests that there is a significant P/E effect for U.S. and many other major Stock Exchanges.

2.2.3 The Cash Flow to Price and Cash Flow to Sales Ratios

An alternative to the P/E ratio is the Cash Flow to Price (CF/P) ratio. The cash flow is defined as reported accounting earnings plus depreciation. Hawawini and Keim (1995) suggested that

accounting earnings might be a biased estimate of economic earnings with which shareholders are concerned. On the other hand, the CF/P ratio might be a less biased estimate of economic earnings because cash flow per share is subject to less manipulation than accounting earnings. Furthermore, Hawawini and Keim observed that the distinction between reported earnings and cash flow is very important if we consider that many countries follow different practices regarding the reporting of earnings. Chan et al. (1991) found a significant relationship between expected returns and CF/P ratio for Japanese stocks.

An alternative to both P/E and CF/P ratios is the Price to Sales (P/S) ratio. This ratio is less influenced by accounting practices compared to P/E and CF/P ratios. The empirical studies of Shechack and Martin (1987) and Jacobs and Levy (1988) uncovered a significant P/S effect in U.S. All these studies found that low P/S ratio portfolios have higher returns on average than portfolios with high P/S ratios (Hawawini and Keim, 1995).

2.2.4 Book to Market Equity

A number of researches uncovered a systematic positive relationship between average returns and the firm's book value of common equity (BE) to its market value (ME). Stattman (1980) and Rosenberg et al. (1985) found that average returns on U.S. stocks are positively related to the BE/ME ratio.

Fama and French (1992) used all non-financial firms in the intersection of the NYSE, AMEX, and NASDAQ return files and the merged COMPUSTAT annual industrial files of income-statement and balance sheet data. The results of this study suggested that size and the BE/ME ratio provide a simple and powerful characterisation of the cross-section of average stock returns for the 1963-1990 period. The BE/ME ratio is the most significant explanatory variable, whereas beta is insignificant. Fama and French stated that the positive relation between average return and the BE/ME ratio is unlikely to be attributed to differences in betas since market betas vary little across BE/ME portfolios. It is rather possible that the risk factor captured by the BE/ME ratio is the relative distress factor of Chan and Chen (1991). According to Fama and French, the earnings prospects of firms are associated with a risk factor in returns. The firms that the market expects to have poor prospects possess higher BE/ME ratios (low stock prices relative to book value) and have higher expected stock returns than firms that the market expects to have strong prospects. In a subsequent study, Fama and French (1993) showed that size and BE/ME are related to systematic patterns in relative profitability and growth that can well be the source of common risk factors in returns.

Chan et al. (1991) also reported that the BE/ME ratio is a powerful variable for explaining

average returns on Japanese stocks. On the other hand, Capaul et al. (1993) documented a BE/ME effect on other international Stock Exchanges such as U.K., France, Germany, and Switzerland. More international evidence about the BE/ME effect can be found in Hawawini and Keim (1995). In summary, the findings from international studies suggest that the magnitude of the BE/ME in other markets is smaller than that reported for the U.S. market.

2.2.5 The Debt to Equity Ratio

Bhandari (1988) argued that beta should not be used as a measure of risk because it is estimated in a pre-test period which does not necessarily overlap with the corresponding test period. Furthermore, he claimed that errors might be involved in the estimation of beta and that beta might be time-varying. Bhandari suggested instead that the Debt to Equity (D/E) ratio should be used as a proxy for equity risk in addition to the estimated beta. In order to prove this argument, Bhandari formed twenty seven portfolios on the basis of size, estimated beta, and D/E ratio using NYSE stocks over the 1948-1979 period. The test results indicated that stock returns are positively related to D/E ratio after controlling for beta and size.

Fama and French (1992) also examined the effect of leverage. They used two leverage variables: the ratio of book assets (BA) to market equity (ME), and the ratio of book assets (BA) to book equity (BE). They interpreted the BA/ME ratio as a measure of market leverage, whereas the BA/BE ratio as a measure of book leverage. Fama and French used logarithms of the leverage ratios because preliminary tests suggested that logs are a proper functional form for modelling leverage effects in average returns. In addition, they have a simple interpretation of the relation between the roles of leverage and BE/ME in average returns. The test results indicated that the two variables are related to average returns but with opposite signs. Higher market leverage is associated with higher average returns, whereas higher book leverage is associated with lower average returns. However, Fama and French observed that the difference of the logarithms between the two variables is the logarithm of BE/ME. They stated that the close link between leverage and BE/ME can be used to explain the book-market effect in average returns. A high BE/ME ratio might suggest that the firm's leverage is high relative to its book leverage. The firm might have a large amount of market-imposed leverage because the market might expect that the prospects of this firm are poor and therefore, it should discount its stock price relative to its book value.

2.2.6 Interaction Among Firm-Specific Variables

A number of studies examined interactions among firm-specific variables. Reinganum (1981a) and Banz and Breen (1986) reported that size subsums the P/E effect. On the other hand, Basu (1983) showed that earnings to price ratios help to explain the cross-section of average returns

on U.S. stocks in tests that also included size and market beta. Peavy and Goodman (1983) suggested that the P/E effect can be an industry effect because firms in the same industry tend to have similar P/E multiples. However, in a subsequent study Goodman et al. (1986) found that size serves as a proxy for the P/E effect.

Jaffe et al. (1989) showed that after controlling for size, the P/E effect is significant in January as well as the other months as opposed to size which is more significant in January. They suggested that the findings of previous studies contradict to each other because they use short periods of time.

Levis (1989b) used data from the London Share Price Database (LSPD) monthly returns file and source file to test for irregularities in stock price behaviour of firms on the London Stock Exchange. Levis reported that investment strategies based on dividend yields, P/E multiples and share prices seem to be as profitable as strategies based on market size during the 1961-1985 period. Levis stated that it is also hard to distinguish between size and share price effects. It seems that these two variables are either proxies to each other or they are proxies for more fundamental determinants of expected returns on common stocks.

Chan et al. (1991) found that size and P/E do not provide additional insight about the cross-sectional variation in returns after controlling for the BE/ME and the CF/P effects. Fama and French (1992) showed that the combination of size and BE/ME captures the cross-section variation in average stock returns associated with size, P/E, BE/ME and leverage during the 1963-1990 test period. In a subsequent study, Fama and French (1993) expanded the set of asset returns to be explained. They also included U.S. government and corporate bonds as well as stocks. They suggested that if markets are integrated, then a single model might be able to explain bond returns. Furthermore, they expanded the set of variables used to explain returns. Apart from size and BE/ME, Fama and French also included term-structure variables that are likely to be related with bond returns. In implementing their model, they regressed monthly returns on stocks and bonds on the returns to a market portfolio of stocks and mimicking portfolios for size, BE/ME, and term structure in returns. The findings of this study suggested that size and BE/ME indeed proxy for sensitivity to common risk factors in stock returns. On the other hand, when the excess market return and the mimicking returns for size and BE/ME equity factors are used alone in bond regressions, they seem to capture common variation in bond returns. But, when the two-term structure is also included in the bond regressions, the explanatory power of the stock market factors mostly disappears. Lakonishok et al. (1994) found that P/E, CF/P and a five-year weighted average of growth in sales are the most significant variables to explain the cross-section of NYSE, AMEX and NASDAQ stocks during

the 1968-1989 period. They reported that the BE/ME effect is subsumed by the other three variables and the size has no predictive power even if it is considered individually.

2.2.7 Stock Returns and Financial Statement Information

The studies that investigated if publicly available financial statement information can predict future abnormal stock returns suggest two approaches to predict abnormal stock returns: the indirect approach and the direct approach (Setiono and Strong, 1998). The indirect approach suggests to predict earnings changes in the first stage and then to form an investment strategy based on these forecasts to predict abnormal returns in the second stage. This approach assumes a well-documented relation between earnings changes and returns and it also requires a good model to predict the one-year-ahead earnings changes based on current financial statement information. On the other hand, the direct approach suggests to predict stock returns directly by using a reliable and stable model to model the relationship between accounting numbers and share price changes.

Ou and Penman (1989a) followed the indirect approach to examine if publicly available financial statement information can predict future U.S. abnormal stock returns. More specifically, they used current year's financial statement information and they constructed a summary measure, labelled Pr, to model the likelihood of positive one-year ahead earnings changes. Their strategy suggests to take long positions in companies with a high forecast probability of a future earnings increase and short positions in companies with a low forecast probability of a future earnings increase. Their findings suggested that over the period 1973-1983 a two-year holding period returns to their strategy can earn substantial profits. Examining these findings, Greig (1992) investigated if differences in firm size can explain the results. Using individual first year observations, he regressed subsequent buy-and-hold twelve-months mean-adjusted returns against current and lagged values of Pr as well as the current equity market value. He found that size enters the regression with a significant negative coefficient that makes insignificant the previously significant coefficients on the Pr summary measure. Stober (1992) and Ball (1992) also examined the Pr strategy and they found that this strategy continues to earn abnormal returns up to six years following the portfolio formation date. They attributed these results to an omitted risk factor (Setiono and Strong, 1998).

Holthausen and Larcker (1992) attempted to predict stock return directly rather than predicting them indirectly via earnings. Their strategy suggests to buy stocks that are predicted to have positive abnormal returns, to sell stocks that are predicted to have negative abnormal returns, and then to hold these positions over a twelve months period. Applying this strategy, they found positive market-adjusted and size-adjusted returns. However, after replicating the Pr strategy

over their test period, they found negative abnormal returns for the long-short hedge over the 1983-1988 period (Setiono and Strong, 1998).

Bernard and Wahlen (1997) examined both the indirect and the direct approach to predict stock returns. Although they confirmed that both strategies earn positive market-adjusted returns, they concluded that the apparent profits of both strategies are likely to reflect compensation for risk. A different finding was reported by Setiono and Strong (1998) who also examined the direct Pr strategy as well as the indirect approach to predict U.K. abnormal returns. They reported evidence that an investor could use publicly U.K. financial information to predict subsequent - year earnings changes and then use these predictions to earn abnormal returns. On the other hand, they found little evidence for the direct approach of using financial statement information to forecast one year ahead stock returns. They claim that the results from the indirect approach are consistent with results of previous studies from Foster et al. (1984) and Bernard and Thomas (1989) that examined the post-earnings announcement drift. Setiono and Strong suggested that the little evidence from the direct approach to predict stock returns is attributed to the fact that market prices may react to several sources of value-relevant information simultaneously. Different sources of information may impact companies' accounts at different times and similar types of information may affect share prices and accounting numbers with various leads or lags. They claimed that this contrasts with companies' reporting of earnings and annual accounts where the market receives these at regular and known dates. Therefore, focusing on the earnings-return link isolates one clear transmission mechanism of how information gets into returns and using the direct approach to predict abnormal stock returns it is likely to add noise to the relation between financial statement information and stock returns.

Other studies that used financial statement information are those by Ou and Penman (1989b), Fairfield and Harris (1993), and Lev and Thiagarahan (1993) who reported abnormal returns using financial statement information, as well as the studies of Jones and Litzenberger (1970) who observed a tendency in stock prices to drift upwards following extreme earnings increases and drift downwards following extreme earnings declines.

2.2.8 Explanations About the Effects

Several explanations have been suggested in the literature for the ability of firm-specific variables to predict stock returns. Smidt (1968) stated that one potential reason for market inefficiency is inappropriate market responses to information. Inappropriate responses to information implicit in P/E ratios are believed to be caused by exaggerated investor expectations regarding growth in earnings and dividends. Exaggerated optimism might lead to

stocks with high P/E ratios, whereas exaggerated pessimism might lead to stocks with low P/E ratios. Ball (1978) suggested that the P/E is a catch-all proxy for unnamed factors in expected returns. Brown et al. (1983) stated that at least part of the size effect might be explained by an omitted risk factor in the pricing model. On the other hand, Keim (1983) suggested that even if part of the average size effect is due to an unspecified risk variable, the behaviour observed in January cannot be attributed solely to this cause because risk alone cannot explain a return premium observed in the same month each year. In a subsequent study, Keim (1988) suggested that some variables such as size (ME), leverage and P/E ratio all represent different ways to scale stock prices in order to extract relevant information in prices about risk and expected returns. Fama and French (1992) confirmed this view and examined if size, leverage, BE/ME, and P/E are also redundant in explaining stock returns. They found that combinations of size and BE/ME capture the cross-section variation in average stock returns associated with size, P/E, BE/ME and leverage during the 1963-1990 test period. Fama and French observed that if stocks are priced rationally, then their results suggest that stock risks are multidimensional. One dimension of risk is proxied by size (ME), whereas another dimension is proxied by the BE/ME ratio. In a subsequent study, Fama and French (1993) stated that size and BE/ME represent arbitrary indicator variables that for unexplained economic reasons are related to risk factors in returns. In order to fill this economic void, Fama and French (1995) investigated whether the behaviour of stock prices in relation to size and BE/ME reflects the behaviour of earnings using NYSE, AMEX, and NASDAQ firms over the 1963-1992 period. After grouping stocks on size and BE/ME, they found that high BE/ME stocks are less profitable than low BE/ME stocks four years before and at least five years after ranking dates. After controlling for BE/ME, Fama and French observed that small stocks tend to have lower earnings on BE than the earnings of big stocks. They suggested that the size effect in earnings is mainly due to the low profits of small stocks after 1980. Before 1980, ratios of earnings to BE are similar for small and big stocks. On the other hand, the earnings of small stocks are highly depressed during the recession between 1981 and 1982, whereas small stocks do not participate in the boom of the middle and late 1980s. Fama and French concluded that the market and size factors help to explain those in returns, whereas there is no link between BE/ME factors in earnings and returns.

Kothari et al. (1995) suggested that the Fama and French's (1992) findings are likely to be influenced by the survivorship bias in the COMPUSTAT database affecting the performance of high BE/ME stocks and the period-specific performance of both low BE/ME past "winner" stocks and high BE/ME past "loser" stocks. Re-examining the beta-return relation using annual rather than monthly data, Kothari et al. reported that there is a substantial ex post compensation for beta risk over the 1941 to 1990 period and even more over the 1927 to 1990 period. On the other hand, using the Standard & Poor's database during the 1947-1987 period, they found that

BE/ME is weakly related to average stock returns. They observed that the relation is statistically significant after using the 500 largest COMPUSTAT firms each year for the post-1962 period, but the effect is about forty percent lower than that obtained using all COMPUSTAT firms. Kothari et al. suggested that the returns of high BE/ME portfolios constructed from COMPUSTAT data might be spuriously inflated because several years of the surviving firms' historical data are included when COMPUSTAT added firms to the database. On the other hand, there are several firms with stock returns on the CRSP tapes but their financial data missing on COMPUSTAT. Furthermore, previous evidence indicated that the frequency of such firms' experiencing financial distress is quite high. Kothari et al. also observed that size and BE/ME are selected in a sequential process of examining and eliminating many other variables that Fama and French (1992) selected from past studies or other variables that are never mentioned in previous studies. Lo and MacKinlay (1990a) analysed similar issues and concluded that classical measures of statistical significance might overstate the true economic significance of the variables in the selection procedure. Considering this evidence about "data snooping", Kothari et al. suggested that the relation between BE/ME and stock returns might not be robust for longer periods as for the more recent decades. However, they observed that it is not possible to study the behaviour of low BE/ME stocks before 1963 because the data from the COMPUSTAT tapes are available in readable form only after 1963. They suggested that the alternative would be to examine the "winner" stocks discussed in the previous studies of DeBondt and Thaler (1985, 1987), Ball and Kothari (1989), and Ball et al. (1995). After performing this investigation, Kothari et al. observed that winners outperform the market before 1963 but they underperform it after 1962. Therefore, they concluded that the negative performance of low BE/ME stocks might be valid only for the specific period after 1963 and the same considerations might apply to high BE/ME "loser" stocks.

Lakonishok et al. (1994) suggested that value strategies based on size, P/E, BE/ME, and CF/P ratios might produce higher returns because they are contrarian to "naive" strategies that extrapolate past earnings information too far in the future. Contrarian investors tend to invest in stocks that are underpriced and they tend to underinvest in stocks that perform well in the past but they have become overpriced at present. Lakonishok et al. suggested that value strategies might not be fundamentally riskier if conventional approaches are used to evaluate risk.

If we summarise the above discussion, we can identify three explanations for the observed ability of firm-specific variables to explain stock returns: 1) they proxy for sensitivity to risk factors in returns and the correlation between the variables and returns reflects compensation for bearing risk; 2) they help to identify stocks that are mispriced because of systematic misjudgements of investors; and 3) their predictive ability is an artifact of "data snooping" and a

survivorship bias in the COMPUSTAT database. The first explanation of the ability of firm-specific variables to predict stock returns seems to be a more reasonable explanation compared to the other two, or at least, compared to the second one. If these stylised facts are a result of market mispricing, then they would be expected to disappear after a reasonable period of time. However, the consistent evidence that they persist provides ground for a rational explanation.

2.3 THE ARBITRAGE PRICING THEORY

2.3.1 Theoretical Background

The Arbitrage Pricing Theory (APT) was developed by Ross (1976) and it was proposed as one of the primary alternatives to the CAPM model. The CAPM predicts that security rates of return will be linearly related to the rate of return on the market portfolio. The APT model is based on similar intuition as the original CAPM but is much more general. It assumes that the rate of return on any security is a linear function of k factors. We can express this relationship as follows (Copeland and Weston, 1992),

$$R_i = E(R_i) + \sum \beta_{ik} f_{ik} + \varepsilon_i \qquad (2.3)$$

where, R_i is the random rate of return on the i^{th} asset, $E(R_i)$ is the expected rate of return on the i^{th} asset, β_{ik} is the sensitivity of the i^{th} asset's returns to the k^{th} factor, f_k is the mean zero k^{th} factor common to the return of all assets under consideration, and ε_i is a random zero mean noise term from the i^{th} asset.

The APT model is derived under the usual assumptions of perfectly competitive and frictionless capital markets. Individuals are assumed to have homogeneous beliefs that the random returns for the set of assets being considered is governed by the linear factor model given in Eq. (2.3). In addition, the theory requires that the number of assets under consideration, n, is much larger than the number of factors, k. The noise term, ε_i is the unsystematic risk component for the i^{th} asset and must be independent of all factors and all error terms for the other assets under consideration (Copeland and Weston, 1992).

2.3.2 Empirical Evidence on the APT

An enormous amount of evidence has been presented in the literature concerning the validity of the APT. Roll and Ross (1980) used daily return data for NYSE and AMEX firms over the 1962-1972 period. For this study, they selected 1260 securities and they divided them

alphabetically into groups of thirty securities. For each group, they applied a maximum-likelihood factor analysis procedure to identify the matrix of assets' sensitivities to the factors, B. Given the estimate of the matrix of the assets' sensitivities to the factors, $\widetilde{B}$, Roll and Ross performed cross-sectional regressions of assets returns on $\widetilde{B}$ as well as cross-sectional regressions of assets excess returns on $\widetilde{B}$ using generalised least squares. After applying this methodology, Roll and Ross reported that there are four factors that have significant risk premia. Brown and Weinstein (1983) tested the equality of the risk premia across subgroups of different assets using the same data set and the same time period as those used by Roll and Ross (1980). However, rather than dividing the total number of assets into groups of thirty, they divided them into groups of sixty. Furthermore, they divided each group of sixty assets into two subgroups of thirty assets. To obtain an estimate of the matrix of the assets' sensitivities to the factors, $\widetilde{B}$, for the groups of sixty assets as well as the subgroups of thirty assets, Brown and Weistein applied a maximum-likelihood factor analysis procedure. They tested the hypothesis of equal price of risk across subgroups of three, five, and seven factor models. They reported that the tests reject the hypothesis of equal prices of risk approximately fifty percent of the time even after using a posterior odds ratio approach to alter the size of the test to reflect the large sample.

Dhrymes at al. (1984) showed that one problem in interpreting the results from factor analysis is that the number of statistically significant factors increases as the number of securities increases. This argument is also supported in the empirical studies by Diacoyiannis (1986), and Beenstock and Chan (1988). Conway and Reinganum (1988) found that the number of factors tend to increase in small samples. The sensitivity of the number of factors to the number of assets may be attributed to some sources of within-industry cross-firm correlations that are insignificant sources of risk for a small number of assets, but they become significant as the number of assets increases.

A major drawback of factor analysis is that it does not provide insight about the nature of the risk factors. An alternative approach was suggested by Chan et al. (1985) who specified ex-ante a set of observable variables as proxies for the systematic state variables in the economy. These variables can be described as follows: 1) an equally-weighted NYSE index; 2) the growth rate of industrial production; 3) a measure of unanticipated inflation; 4) the change in expected inflation; 5) the difference in returns on long-term government bond portfolio and short-term Treasury bills; 6) the growth rate of the Net Business Formation Series from month t to t+1; and 7) the difference in returns on low-graded corporate bonds and long-term government bonds. Chan et al. investigated the firm size effect for the period 1958-1977 in the framework of a

multifactor pricing model. They divided the twenty-five years period into twenty overlapping intervals, each consisting of six years. During each six-year interval, firms on the NYSE that existed at the beginning of the interval and had price information on December of the fifth year were chosen and ranked according to market value at the end of the fifth year. These firms were then put into one of the twenty portfolios arranged in order of increasing size, each containing the same number of securities. The testing was done in the sixth year. The results of this study suggested that the firm size is essentially captured by a multifactor pricing model and it is doubtful whether any risk-adjusted profits can be realised in practice. Among the economic variables included, the measure of the changing risk premium explains a large portion of the size effect. After including a measure of the size of the firm (natural logarithm market value of equity) as an independent variable, Chan et al. found that the size variable seems to explain the returns in the multifactor model and the January residuals do not reveal any particular pattern.

Chen et al. (1986) also specified ex-ante a set of observable variables as proxies for the systematic state variables in the economy. Using variables more or less similar to those used by Chan et al. (1985), they utilised sixty months of time-series observations to estimate the matrix of asset's betas relative to the prespecified factors. After estimating the matrix of the factor sensitivities, $\widetilde{B}$, Chen et al. performed cross-sectional regressions of returns on $\widetilde{B}$ to estimate the returns of factor mimicking portfolios. At the beginning of each period, they formed twenty portfolios on the basis of market capitalisation of equity. Furthermore, they estimated the average risk premia for the full sample period from 1958 up to 1984 as well as three subperiods. The results suggested that the five prespecified factors provide a reasonable specification of the sources of systematic risk and priced risk in the economy. After controlling for factor risk, they reported that alternative measures of risk such as market betas and consumption betas do not seem to be priced. Other studies that tested the APT in ways similar to Chan et al. (1985) and Chen et al. (1986) include among others those by Burmeiser & Wall (1986), Connor and Uhlaner (1988), Ferson and Harvey (1991a, 1991b), and Cragg and MacDonald (1992).

Lehmann and Modest (1988) compared the CAPM and the APT using assets that were randomly selected from NYSE and AMEX firms over five-year subperiods between 1963 and 1982. They estimated the covariance matrix of daily returns using factor analysis. They obtained the returns on a number of factor mimicking portfolios by minimising the idiosyncratic risk under the assumption that the factor portfolio has sensitivity to only one factor. Forming various portfolios on the basis of market capitalisation, dividend yield and asset's variance, Lehmann and Modest regressed weekly returns on the portfolios against the factor mimicking returns. They also repeated these regressions with single-index portfolio proxies to examine the

CAPM. The results of this study rejected the CAPM during the 1963-1982 period, whereas they rejected the APT only when size portfolios were used. In a similar study, Connor and Korajczyk (1988) also compared the CAPM and the APT using monthly data on NYSE and AMEX firms over four five-year subperiods between 1964 and 1983. During this study, they performed two experiments. In the first experiment they used the whole sample of assets, whereas in the second experiment they grouped the individual assets into ten size-portfolios. Connor and Korajczyk found that the APT explains better than the CAPM the non-seasonal size anomaly using the size portfolios, whereas no major differences exist between the models using individual assets.

Bossaerts and Green (1989) tested dynamic versions of the APT. They found that dynamic models perform better than constant parameter models. Hollifield (1993) supports these findings. Bansal and Viswanathan (1993) applied a non-linear version of the APT using a value-weighted NYSE portfolio as an aggregate wealth proxy. They also used the yield on one month Treasury bills and the yield spread between six and nine month Treasury bills as the idiosyncratic risk free yields. The results of this study documented that non-linear versions of the APT perform better than linear versions of the model. Bansal and Viswanathan reported that a non-linear model based on the NYSE portfolio and the one-month Treasury bill performs better than a single factor model based on the NYSE portfolio, whereas adding the yield spread does not provide any additional benefits.

Clare and Thomas (1994) presented empirical evidence on the pricing of macroeconomic factors in the U.K. stock market between 1983 and 1990. Their sample size consisted of monthly returns on 840 randomly selected stocks. Starting with 18 macroeconomic variables, they selected the most significant variables and they excluded the rest of the variables. After introducing the market index into the model, they found that the eight most significant variables do not seem to be priced by the market and some of the other factors exhibit slightly lower coefficients. Clare and Thomas found that only three of the variables are significant, namely oil price, default risk and debentures, and loan redemption yield, whereas the rest of the variables are marginally significant at the 10% level. In addition, the constant seems to be significant indicating an inadequacy of the factor model.

More international evidence about the APT can be found in Connor and Korajczyk (1995). Overall, the empirical work on the APT suggests that more than one factor is important in determining asset returns. Studies comparing the APT and the CAPM suggest that the APT provides a better description of the expected returns on risky assets than the CAPM. The APT is more robust than the CAPM for several reasons: first, it makes no strong assumptions about

individuals' utility functions as well as assumptions about the empirical distribution of asset returns; second, investors are not assumed to select portfolios on the basis of expected return and variance; third, the APT allows the equilibrium returns of assets to be dependent on many factors; and fourth, the efficiency of market portfolio is not of a primary concern in the APT as the original CAPM. Therefore, the difficulties associated with the market portfolio are avoided.

However, the APT is not a perfect model and it has several disadvantages. Its major disadvantage is that the usual statistical methods are not adequate to test an approximate pricing relation. Therefore, the tests of the APT are joint tests of the model and additional assumptions are necessary to obtain exact pricing. On the other hand, the empirical tests to identify the factor structure in security returns have not produced consistent results. Connor and Korajczyk (1995) suggested that the APT would be a better model if the factors could be related to more identifiable sources of economic risk. Certainly, more developments in asset pricing theory and econometrics are required in order to understand the relationship between return factors and economic risks.

2.4 TIME-VARYING BETAS AND RISK PREMIA

2.4.1 Theory and Empirical Evidence

Most of the empirical studies mentioned so far can be viewed as unconditional tests of asset pricing models because they implicitly assume that the rates of return on stocks are serially independent and have a constant distribution which is assumed to be independent of any ex-ante known information. As a result of these restrictions, unconditional tests require short testing periods and implicitly assume that expected returns, variances, and covariances, betas, and risk premia are stationary over time. However, the empirical evidence casts some doubt on these restrictions and suggests that expected returns are indeed time-varying. This evidence together with the belief that investors use the latest available information in the market to form their expectations has given ground to the development of conditional asset pricing models that assume that the return distribution depends on a set of ex-ante observable variables.

Many empirical studies have examined the stability of betas and risk premia over time. Sharpe and Cooper (1972) used monthly returns on NYSE stocks over the 1937-1967 period to examine the stability of betas. Initially, they divided the whole sample period into six five-year subperiods. For each subperiod, they calculated the beta for each security and they grouped them to different risk classes based on their beta estimates. After assigning the securities to groups, Sharpe and Cooper studied if there are any changes in the groups over the next subperiod. After performing this study, they concluded that the successive betas for each

individual security are not constant over time.

Chan (1988) examined the stability of betas for "winner" and "loser" portfolios. He defined the "loser" portfolio as the decile with the lowest return over the previous three years period, and the "winner" portfolio as the decile with the highest return over the previous three years period. The test results indicated that the beta of the "loser" portfolio increased after a period of abnormal loss, whereas the beta of the "winner" portfolio decreased after a period of abnormal gain. These findings were supported by Ball and Kothari (1989), Kothari and Shanken (1992), Jones (1993), and Ball et al. (1995). Ball and Kothari (1989) observed that the performance of the portfolios that earned extreme returns is negatively correlated with changes in systematic risk measured by the market beta. They suggested that the time variation in betas could be due to shifts in leverage as well changes in the risk of the firm's underlying cash flows. This hypothesis was supported by Ferson and Harvey (1993a) who showed that when market betas are regressed on a set of accounting variables, they seem to be strongly related with variables such as leverage and earnings variability. This evidence has economic sense if we consider that companies with variable earnings are more likely to suffer during recessions and therefore are more risky than companies with more stable earnings. On the other hand, leverage increases the volatility of earnings and this is reflected in risk and beta. Therefore, it is reasonable to expect that if these variables change over time, then the betas will change as well. In another study, Ferson and Harvey (1993b) investigated the time variation in betas with respect to movements in the economy. For this purpose, they regressed the betas from unconditional asset pricing models to several state variables reflecting the state of the economy such as default spread, dividend yields etc. The results of this study uncovered that the movements in betas are due to changing economic conditions. Ferson and Korajcyzk (1994) supported these findings.

A number of empirical studies examined the variation of risk premia. In one of these studies, Black et al. (1972) showed that the market premium is nonstationary and it is negative for the 1957-1965 period. Fama and MacBeth (1973) found that the market premium fluctuates considerably from one month to the next and from one five-year subperiod to the next for the 1935-1968 period. Ferson and Harvey (1991a) examined the stability of fitted risk premia using a framework of a multifactor asset pricing model. They related the changes in the premia to movements on various ex-ante observable variables. Their results indicated that the risk premia associated to various macroeconomics factors are time-varying and most of them vary with the business conditions. Kothari and Shanken (1995) investigated the relationship between the excess return on the market and the book-to-market ratio. They found that during periods in which the book-to-market ratio is very low, the expected market premium is negative. Similar findings were reported by Pettengill et al. (1995).

Overall, the findings of the above empirical studies suggest that unconditional betas and risk premia are nonstationary over time. Furthermore, the empirical evidence suggests that the variation of betas and risk premia may be related to firm specific characteristics and changes in economic conditions. Therefore, it seems reasonable to expect the development of more flexible models that are able to incorporate time variation in betas and risk premia. These models are discussed in the next sections.

2.5 CONDITIONAL ASSET PRICING MODELS

Three approaches have been suggested in the literature to examine the validity of the conditional models: i) latent variable models where expected returns and risk premia are time-varying while betas are considered constant; ii) ARCH - M models that assume time-varying variances and covariances but assume constant price of risk; and iii) models based on instrumental variables that assume time-varying expected returns, variances, covariances, and hence betas and risk premia. These models will be discussed in the next sections.

2.5.1 Latent Variable Models

Latent variable models suggest that predictable asset returns are driven by a small number of expected risk premia that represent the unobservable or latent variables. Let us consider a set of asset returns R_{it} (i = 1,2...n) that are regressed on a vector of predetermined variables denoted as P_{t-1}. The conditional expected excess return can be expressed as follows (Ferson, 1995),

$$E(r_{it} / P_{t-1}) = \sum \gamma_{ij} P_{t-1,j}, \quad P_{t-1,j} \in I_{t-1} \tag{2.4}$$

where r_{it} is the excess return, and I_{t-1} represents all the information that investors use to set prices. Latent variable models study the cross-equation restrictions on the coefficients of the projections, γ_{ij}. Tests are performed to detect the reduced dimensionality across assets in the cross variation of expected returns. These tests are motivated from conditional beta models assuming that their betas are fixed parameters over time. The expected returns of the unobserved factor mimicking portfolios represent the latent variables (Ferson, 1995).

The latent variable models were introduced by Hansen and Hodrick (1983) and they further extended by Gibbons and Ferson (1985), Campbell (1987), Ferson (1990), and Ferson et al. (1993). Hansen and Hodrick (1983) examined whether a model with a single premium holds using a sample of currency premiums. They reported that the tests are unable to reject the

model. Gibbons and Ferson (1985) also examined whether a model with a single premium holds using the Dow Jones 30 common stocks. They reported that a single factor model with constant betas cannot be rejected over four equal sub-periods from 1962 to 1980. Using monthly U.S. data over the 1959-1979 and 1979-1983 periods, Campbell (1987) rejected a conditional asset pricing model with a single time-varying risk premium. He also rejected models in which betas are constant and risk premia are driven by time variation in one or two latent variables. Wheatley (1989) suggested that latent variable tests might have low power as tests of asset pricing models because their statistical assumptions are difficult to verify. Ferson (1990) used quarterly returns on size deciles of NYSE stocks and three fixed income excess returns. The results of this study uncovered two or three latent state variables in the time-varying expected returns. After relaxing the assumption that the conditional expected returns are functions of the predetermined instruments, Ferson et al. (1993) reported that a single risk premium is adequate to model expected stock and bond returns using two or three common factors. On the other hand, Mei and Saunders (1994) restricted their sample to insurer stocks and they found that time variation in expected returns can be explained by a latent variable model with one or two factors. Ferson and Foerster (1994) provided evidence on the power of latent variable tests that are implemented using the Generalised Methods of Moments approach. More evidence in tests of latent variable models can be found in Ferson (1995).

Typically, the tests of latent variable models indicate that the number of latent variables is usually small and no more than two or three. This evidence suggests that the expected returns can be modelled using a small number of common factors. However, latent variable models have several drawbacks:

- Latent variable models rely on the predictable variation of their power. If the assets' excess returns are independent and identically distributed over time, then the only predetermined variable that might enter into the predictive regression would be a constant term.
- The nature of the risk premia is undefined.
- The assumption of constant betas is rather restrictive.
- If a single premium model is accepted, latent variable models seem unable to answer the question whether or not the variance of a single unobservable factor is excessive compared to the one that would result from a portfolio choice model based on risk aversion.

2.5.2 ARCH-type Models

Various approaches have been suggested in the literature for modelling empirically time-varying second moments (Ferson, 1995). One approach assumes that the conditional second

moments have a specific functional form to a set of lagged instruments. Another approach uses lagged values of the squares and products of the innovations in returns to model the conditional second moments. A large number of studies have applied Autoregressive Conditional Heteroscedasticity (ARCH) models, Generalised ARCH (GARCH) models, and other related models to study security returns. The ARCH-type models assume a simple functional form between the conditional second moments and the lagged innovations of the variables. More specifically, ARCH-type models make the conditional variance of the time t prediction error a function of time, exogenous and lagged endogenous variables, system parameters, and past prediction errors. A variety of ARCH-type models have been developed over the last decades (Ferson, 1995). These include among others E-GARCH models (Nelson 1991), MARCH and TARCH models (Gourieroux and Monfort 1992), ARCH-M models (Engle et al. 1987), and GARCH-M models. The ARCH-M models are of particular interest in asset pricing because they imply restrictions between the conditional first and second moments of returns. These models were extended further to GARCH-M models that include both lagged conditional variances and squared innovations.

According to Nelson (1991), let ζ_t be a model's (scalar) prediction error, β a vector of parameters, x_t a vector of predetermined variables, and σ_t^2 the variance of ζ_t given information at time t. A univariate ARCH model sets $\zeta_t = \sigma_t \delta_t$ where $\delta_t \sim$ i.i.d and $E(\delta_t) = 0$, $Var(\delta_t) = 1$ and $\sigma_t^2 = \sigma^2(\zeta_{t-1}, \zeta_{t-2}, ..., t, x_t, \beta) = \sigma^2(\sigma_{t-1}\delta_{t-1}\sigma_{t-2}\delta_{t-2}, ..., t, x_t, \beta)$. This system can be easily extended to a multivariate interpretation if δ_t is an *n* by one vector and σ_t^2 is an *n* by *n* matrix. One common interpretation for $\sigma^2(..)$ are the linear ARCH and GARCH models (Engle 1982 and Bollerslev 1986, respectively), which make σ_t^2 linear in lagged values of $\zeta_t^2 = \sigma_t^2 \delta_t^2$ by defining the following equations (Nelson, 1991),

$$\sigma_t^2 = \theta + \sum_{j=1}^{p} \alpha_j \delta_{t-j}^2 \sigma_{t-j}^2 \text{, and} \tag{2.5}$$

$$\sigma_t^2 = \theta + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 + \sum_{j=1}^{p} \alpha_j \delta_{t-j}^2 \sigma_{t-j}^2 , \tag{2.6}$$

respectively, where θ, α_j, and β_i are nonnegative. It is obvious that Eq. (2.5) is a special case of Eq. (2.6). Therefore, we can refer to both Eqs (2.5) and (2.6) as GARCH models. The GARCH-M model of Engle and Bollerslev (1986), adds an additional equation which can be

written as follows (Nelson, 1991),

$$R_t = \alpha + \beta\sigma_t^2 + \zeta_t \tag{2.7}$$

where σ_t^2 is the conditional variance of R_t. The GARCH-M model states that σ_t^2 enters the conditional mean of R_t. If R_t is the return on a portfolio at time t, then its required rate of return may be linear in its risk as measured by σ_t^2.

French et al (1987) found that the risk premium on monthly returns on the market is positively related with their conditional variances which are calculated as a GARCH (1,2) model over the period 1928-1984. These findings were supported by Attanasio and Wadhwani (1990) who observed an important mean effect when the variances are estimated as GARCH (1,0) processes using monthly and annual U.S. data over the periods 1953-1988 and 1872-1986.

Harvey (1989) and Baillie and DeGennaro (1990) tested the sensitivity of the estimate of a risk aversion parameter under different model specifications. Harvey (1989) found that the estimate of this parameter is time-varying and its sign is dependent on the stage of the business cycle. Baillie and DeGennaro (1990) showed that the significance of the parameter is dependent on the form of the underlying distribution.

Some researches have formulated ARCH processes in a multivariate framework to explain excess returns on several assets. However, as the number of assets used in the multivariate model increases, the number of parameters to estimate increases tremendously. In order to simplify the computations, Bollerslev et al. (1988) assumed that the covariance of each asset to be dependent only on its lagged covariance and past forecast errors. After estimating a multivariate GARCH-M (1,1) model for quarterly returns on bonds, stocks, and Treasury bills during the 1959-1984 period, they found that the expected excess returns are significantly influenced by the conditional covariances. In a later study, Bollerslev (1990) simplified the calculations further by assuming that the conditional correlation of excess returns is constant over time, whereas the covariance matrix is allowed to change after changes in asset variances. Ng (1991) used this approach and conducted multivariate tests using monthly NYSE stocks over the 1926-1987 period. Assuming that the conditional covariance matrix follows a multivariate GARCH (1,1) process, she ranked the securities on the basis of betas as well as on the basis of size. The results of this study suggested that the conditional CAPM is rejected for the size-based portfolios, but it is not rejected for the beta-based portfolios. However, these findings were not supported by Bodurtha and Mark (1991).

The empirical evidence suggests that ARCH-type models are very useful tools because they provide a standard framework for testing conditional asset pricing models and are flexible to incorporate the time-variation in variance and covariance. However, ARCH type models have some major drawbacks that can be summarised as follows,

- The assumption about the functional form of the second moments is not based on any economic theory. Therefore, it is difficult to link temporal changes in stock returns and variances to economic conditions.
- The parameter estimates in ARCH-type models seem to be sensitive to different model specifications and estimation procedures.
- Any misspecification in the variance equations will lead to biased and inconsistent parameters in the mean equation.
- The evolution of the instrumental variables needs to be modelled, which can be a complex task if a multifactor model is being tested and a large set of assets of a portfolio is considered.

2.5.3 The Instrumental Variables Approach

The instrumental variables approach was proposed by Campbell (1987) and Harvey (1989) as an alternative to the GARCH-M model. Campbell and Harvey presented a model of market return that assumes that the expected return is linear in its own variance conditional on some vector I_t containing M instruments. This model can be expressed in the case of N assets as follows (Campbell et al. 1997),

$$E(R_{i,t+1}|I_t) = \delta_0 + \delta_1 Cov(R_{i,t+1}, R_{m,t+1}|I_t) \tag{2.8}$$

$$\varepsilon_{t+1} = R_{t+1} - I_t B \tag{2.9}$$

$$u_{t+1} = R_{t+1} - \delta_0 - \delta_1 (R_{t+1} - I_t B)(R_{m,t+1} - I_t \beta_m) \tag{2.10}$$

where $E(R_{i,t+1}|I_t)$ represents the expected returns of N assets conditional on some vector I_t containing M instruments, $R_{t+1} = (R_{1,t+1}, R_{2,t+1} \ldots R_{N,t+1})'$, $R_{m,t+1}$ is the market return, β_m is a vector of regression coefficients of the market return on the instruments, and B is a matrix of NM coefficients. The Eqs (2.8)-(2.10) imply that the errors $\varepsilon_{m,t+1}$ and $u_{m,t+1}$ are both

orthogonal to the instruments I_t. The total number of orthogonality conditions is 2(N+1)M, whereas the total number of parameters is M(L+1)+(L+2). Campbell and Harvey applied the Generalised Method of Moments to estimate the parameters.

Harvey (1989) generalised further the model to allow for time varying price of risk. This model can be expressed as follows (Campbell et al. 1997),

$$E(R_{i,t+1}|I_t) = \delta_0 + \delta_{1t} Cov(R_{i,t+1}, R_{m,t+1}|I_t) \tag{2.11}$$

$$h_{t+1} = (R_{m,t+1} - I_t B_m)^2 (I_t B - \delta_0 \nu) - (R_{t+1} - I_t B)(R_{m,t+1} - I_t \beta_m)(I_t \beta_m - \delta_0) \tag{2.12}$$

The parameter δ_{1t} in Eq. (2.11) varies through time but it is common to all assets. Assuming that δ_{1t} holds for the market portfolio, we can write it as follows (Campbell et al. 1997),

$$\delta_{1t} = \frac{E(R_{m,t+1}|I_t) - \delta_0}{Var(R_{m,t+1}|I_t)} \tag{2.13}$$

The error vector h_{t+1} is derived by substituting Eq. (2.13) into Eq. (2.11), multiplying by $Var(R_{m,t+1}|I_t)$, and taking into account that $E(R_{m,t+1}|I_t) = I_t \beta_m$ and $E(R_{t+1}|I_t) = I_t B$. However, Eqs (2.11) and (2.13) can be expressed as follows (Campbell et al. 1997),

$$E(R_{i,t+1}|I_t) = \delta_0 + \beta_{it} r_t \tag{2.14}$$

$$\beta_{it} = \frac{Cov(R_{i,t+1}, R_{m,t+1}|I_t)}{Var(R_{m,t+1}|I_t)} \tag{2.15}$$

where β_{it} is the conditional beta of the asset i with the market return, and $r_t = E(R_{m,t+1}|I_t) - \delta_0$ is the expected excess return on the market over a riskless return. Eqs (2.14) and (2.15) represent the conditional version of the CAPM. This model has been extended to conditional multibeta models such as the conditional version of Sharpe-Lintner-Mossin CAPM, the APT model of Ross (1976), the intertemporal asset pricing models of Merton (1973), and conditional approaches of relative risk aversion of Merton (1980).

Campbell (1987) imposed a linear restriction across the conditional first and second moments to

test a conditional version of the CAPM with time-varying variances using U.S. stock and bond returns over the 1959-1979 and 1979-1983 periods. He found that the conditional model cannot account for all predictable variation. Harvey (1989) used monthly returns on the NYSE stocks grouped in size deciles over the 1941-1987 period. He modelled expected returns, variances, and covariances as linear functions of several instruments reflecting the state of the economy. The results of this study indicated that the returns and the conditional covariances are predictable. Ferson and Harvey (1994) restricted the risk premia to depend on global information variables representing the state of the macroeconomy. They also decomposed the explained variance to estimate the contribution of time-varying risk premia and the contribution of time-varying betas. After applying single and multibeta models in eighteen different markets, they reported that conditional models capture a large part of return predictability for many countries. Furthermore, they reported that the multibeta models outperform single beta models.

The major drawbacks of the instrumental variables approach can be summarised as follows:

- The time series of the betas and risk premia have to be estimated using unconditional models first in order to test the conditional model later. This might give rise to the errors-in-variables problem. This problem may affect the results from the conditional models using instrumental variables.
- The validity of the model depends largely on the specification of the model used to describe the stochastic process of the returns. This means that the final inferences will be conditioned to the instruments used to forecast the returns, variances, covariances, and risk premia.

2.5.4 Unconditional Tests versus Conditional Asset Pricing Models

There are several advantages associated to conditional asset pricing models over unconditional ones:

- they allow for time variation in expected returns, betas, and risk premia and hence they are more consistent with the evidence reporting predictable component in the time series of returns;
- they provide a useful framework to integrate the time series predictability of stock returns with the cross-sectional implications of asset pricing models;
- they avoid the problem that the omission of conditioning information can mislead the conclusions regarding the validity of various asset pricing models.

However, conditional asset pricing models have two major drawbacks:

- a particular model for the conditional expectations has to be specified. The form of the function depends on the joint probability distribution of the returns and the instrumental variables. Although most studies have adopted a linear relation for the first moments, other relations may also be valid.
- it may be difficult to conclude anything about the validity of an asset pricing model given the fact that all information investors use to set prices is unobservable and that the econometrician only uses a reduced set of information.

Part Two: Stock Return Predictability Using Past Returns and Ex-Ante Observable Variables

2.6 PAST RETURNS AND STOCK RETURN PREDICTABILITY

One of the dominant themes in the academic literature since 1960s has been the efficient market hypothesis. This is one of the most important and pervasive as well as controversial concepts in the modern theory of finance. As a general definition, the capital market is said to be efficient if security prices fully reflect all available information. Fama (1970) has done a great deal to operationalise the notion of capital market efficiency by defining three types of efficiency based on a different notion of exactly what type of information is understood to be relevant in the phrase: all prices fully reflect all available information (Copeland and Weston, 1992). The three types of market efficiency suggested by Fama can be stated as follows: i) weak-form efficiency, ii) semi-strong form efficiency, and iii) strong-form efficiency.

The market is said to be *weak-form efficient* if all information contained in historical prices is fully reflected in current prices. Alternatively stated, weak-form efficiency implies that no investor can earn excess returns by developing trading rules based on historical price or return information. On the other hand, the market is said to be *semistrong-form efficient* if all publicly available information is fully reflected in current prices. This form of efficiency implies that no investor can earn excess returns from trading rules based on any publicly available information. Finally, the market is *strong-form efficient* if all information, whether public or private, is fully reflected in current prices. This type of efficiency is very strong indeed. If markets were efficient in their strong form, prices would fully reflect all information even though it might be held exclusively by a corporate insider (Copeland and Weston, 1992).

Most of the studies in market efficiency assumed a common equilibrium model of stock prices in which expected returns are constant through time. The model of constant expected returns

implies that prices should follow a random walk. In other words, if market is efficient then information in past price changes should not be relevant in predicting future changes. Previous studies have examined this notion of market efficiency by testing whether return autocorrelations are equal to zero. The empirical evidence from these studies is discussed in the next sections.

2.6.1 Evidence on Stock Return Autocorrelations

The first evidence on stock return autocorrelations is dated back in 1960s. Fama (1965) examined the autocorrelations of daily returns over the period 1957 to 1962 using the Dow Jones 30 industrial stocks. He found that approximately 75% of the stocks have significant positive autocorrelations. On the other hand, Fisher (1966) showed that significant portfolio autocorrelations might be due to infrequent trading. Reinganum (1981b) tested for autocorrelations using the daily returns of portfolios composed of infrequently traded small stocks and he found significant positive autocorrelations. Examining these results, Lo and MacKinlay (1990b) applied a model for nontrading and they reported that the model is unable to explain autocorrelations in the data. French and Roll (1986) examined the autocorrelations of daily returns of NYSE and AMEX stocks. They found that the estimated autocorrelations are inversely related to the market capitalisation of the stock. More specifically, smallest stock autocorrelations appear to be the most negative, and the stocks in the largest decile of market capitalisation appear to have positive autocorrelations on average. Foerster and Keim (1992) also examined autocorrelations of daily returns during the 1963-1990 period using the Dow Jones 30 industrial stocks. They found that around 80% of the stocks have significant positive autocorrelations and that the return autocorrelations of both small and large stocks are insignificantly different than zero in the last half of the 1980s. Poon and Taylor (1992) examined the U.K. Financial Times All Share Index and they found significant positive lag-one autocorrelations in daily returns during the 1965-1989 period. Baily et al. (1990) found lag-one autocorrelations during the 1977-1985 period for a number of different stock markets around the world including Hong Kong, Australia, Singapore, Taiwan, and Korea.

Some studies examined return autocorrelations using weekly data. Lo and MaKinlay (1988) tested for autocorrelations in weekly returns. They found strong positive autocorrelations for portfolios of small stocks and week positive autocorrelations for portfolios of large stocks. Conrad and Kaul (1988) also used weekly returns but they computed them using prices that were the result of actual transactions in order to test for nontrading. They found significant positive autocorrelations and they concluded that these results cannot be attributed to nontrading. Lehmann (1990) tested the economic significance of negative weekly autocorrelations based on the profitability of a trading rule using the price patterns. They found

that the trading rule is profitable for certain market participants but the profitability gains might deteriorate for most market participants after taking into account transaction costs.

Some other studies examined return autocorrelations using longer time intervals. Keim and Stambaugh (1986) used monthly returns over the 1928-1978 period to test the relation between monthly returns and autocorrelations. They found stronger positive autocorrelations for portfolios of small stocks and weaker positive autocorrelations for portfolios of large stocks. However, they reported that the results concerning the relation between market capitalisation and return autocorrelations are weaker than those reported by Lo and MacKinlay (1988) who used weekly returns.

Summers (1986) criticised previous tests on market efficiency because the techniques they used fail to detect large swings in stock prices away from their fundamental values. He modelled excess stock returns as an ARMA (1,1) process and he observed that part of the observed positive or negative returns are on average spurious due to shocks that deviate stock prices from their fundamental values. If prices are reverting to their mean values, then negative or positive returns might result implying negative serial correlation in long horizon returns. On the other hand, tests using short horizons might fail to capture the mean-reverting stock price components because returns might display little autocorrelations over short horizons.

Fama and French (1988) examined the autocorrelations for increasing holding periods. Using monthly returns on NYSE stocks during the 1926-1985 period, they modelled the stock prices as a sum of a random walk and a stationary AR(1) process. They found that the random walk price component produced white noise in returns, whereas the mean reversion stationary component produced negative autocorrelations in returns. The results of this study provided evidence for return autocorrelations that are close to zero for one-year horizons, become negative for two-year horizons, reach minimum values for three- to five-years horizons, and move towards zero for return horizons greater than five years. The sub-periods results suggested that the autocorrelations are largely due to the 1926-1940 period. Fama and French reported that past returns predict 25-40% of the variation of 3-5 years future returns. Poterba and Summers (1988) also suggested that long-horizon stock returns have large predictable components. Using variance-ratio tests, they found positive serial correlations for less than a year horizons, whereas they found negative serial correlations for more than a year horizons.

Some empirical studies reported seasonalities in stock returns related to calendar turning points through the year. These effects have become widely known as the Weekend effect (Gibbons and Hess 1981; Keim and Stambaugh 1984; Harris 1986; Lakonishok and Smidt 1988), the

January effect (Gultekin and Gultekin 1983; Blume and Stambaugh 1983) the Monthly effect (Ariel 1987; Lakonishok and Smidt 1988), and the Holiday effect (Ariel, 1990). A detailed discussion about these studies can be found in Keim and Hawawini (1995).

In summary, the empirical studies on stock return autocorrelations provide some evidence for return autocorrelations that are more apparent in long horizon returns. However, although the empirical studies suggest that the estimated autocorrelations might be statistically significant, they are not able to support the view that the return autocorrelations are economically significant as well. Some possible explanations for the return autocorrelations are reviewed in the next sections.

2.6.2 Possible Explanations for Return Autocorrelations

One explanation that has been suggested in the literature for the negative serial correlation in long horizon returns is stock market mispricing with prices taking long irrational temporary departures from their fundamental values. If there is market efficiency, then fundamental values will be fully reflected by market prices. On the other hand, if new information concerning the cash flows arises, then current prices will change to reflect this information quickly and efficiently. Under these assumptions, the variance of stock returns should be related to the amount of information. Examining this proposition, Shiller (1981) concluded that the variance of stock prices is too large for efficient markets.

In the last decade, empirical researchers have noted many apparent anomalies as far as the efficient market hypothesis is concerned. Probably, the most potentially devastating to the EMH anomaly was the observation that speculative assets appeared to be much too volatile to be explained by the Efficient Market Model. If this observation is correct, it implies that market may be efficient in some ways, but most speculative price movements cannot be explained as due to information about fundamentals (Shiller 1981; LeRoy and Porter 1981).

A number of researchers including Mankiew et al. (1985), and Campbell and Shiller (1988) tested the implications of the variability of stock prices relative to dividends. All found that simple models such as the constant expected return model are inconsistent with the data. It is difficult to imagine any reasonable model of equilibrium consistent with the efficient markets hypothesis that could also be consistent with these results. In addition, the stock market crash of October 1987 was extremely puzzling. The market dropped between 20% and 25% on a Monday following a weekend during which little surprising news announced. Because the crash of 1929 is still an enigma, it is doubtful that the more recent crash will be explained any time soon. Perhaps the two market crashes are evidence consistent with the bubbles theory of

speculative markets. According to the definition of the standard efficient markets model, a stock price is determined by the following arbitrage relationship (West, 1987),

$$P_t = \beta E(P_{t+1} + D_{t+1}) \setminus I_t,$$

(2.16)

where P_t is the real stock price in period t, β is the constant ex ante real discount rate, $0 < \beta = 1/(1+r) < 1$, r is the constant expected return, E denotes mathematical expectations, D_{t+1} is the real dividend paid to the owner of the stock in period t+1, and I_t is information common to traders in period t. I_t may contain current and past dividends as well as other variables that can be useful in forecasting dividends. According to West (1987), if we solve Eq. (2.16) recursively, we get

$$P_t = \sum_{1}^{n} \beta^i E D_{t+i} \setminus I_t + \beta^n E P_{t+n} \setminus I_t$$

(2.17)

If the condition $\lim_{n\to\infty} \beta^n E P_{t+n} \setminus I_t = 0$

(2.18)

holds, then we can have

$$P_t^* = \sum_{1}^{n} \beta^i E D_{t+i} \setminus I_t$$

(2.19)

where $P_t = P_t^*$. If condition (2.18) holds, then P_t^* is a unique forward solution to Eq. (2.16). On the other hand, if condition (2.18) fails, then there is a family of solutions to Eq. (2.16). Any P_t that satisfies the conditions (West, 1987),

$$P_t = P_t^* + G_t, \qquad E G_t \setminus I_{t-1} = \beta^{-1} G_{t-1}$$

(2.20)

is also a solution to Eq. (2.16). According to West (1987) G_t is a speculative bubble which can be interpreted as an extraneous event that affect stock prices because everyone expects it to do so.

West (1987) developed an ingenuous test for bubbles using the standard and Poor's Composite Price Index and the Dow-Jones data that were first used by Shiller (1981). The test involved a

comparison of estimates that were constructed using two different information tests. One information set was taken to be current and past dividends. The other information set was taken to be the market price under the hypothesis that constant expected returns are correct. Forecasting with a smaller information set than the market's should result in a larger innovation variance. However, West found the opposite and attributed a large part of the volatility of prices to bubbles. He argued that time-varying expected returns are unlikely to overturn the results.

De Bondt and Thaler (1985, 1987) mounted an aggressive empirical attack on market efficiency directed at unmasking irrational bubbles. De Bondt and Thaler examined an "overreaction" hypothesis, which states that people "overreact" to unexpected and dramatic new events. Applied to stock prices the hypothesis is that, as a result of overreaction, "loser" portfolios outperform the market after their formation. DeBondt and Thaler found that portfolios identified as the most extreme "losers" ("winners") over a three- to five-year period tend to have highest (lowest) market-adjusted returns during the following period. They attributed these results to market "overreaction" in which stock prices are driven temporarily away from their fundamental values due to irrational waves of optimistic or pessimistic news about firms.

After De Bondt and Thaler's (1985, 1987) findings, many contrarian investment strategies have suggested that profitable opportunities may arise as a result of market "overreaction" and negatively autocorrelated price changes. However, Chan (1988) and Ball and Kothari (1989) observed that the "winner-loser" results and the abnormal risk-adjusted returns reported for contrarian investment strategies are due to incorrect adjustment for risk. If the stock of a loser firm has declined and there is no decline in the value of the debt, then this firm might become more leveraged and therefore more risky. In a different study, Zarowin (1989) showed that there is a relationship between "winner-loser" results and the size of the firm. He observed that the "loser" firms are mostly small firms, whereas the "winner" firms are mostly large firms. After controlling for size, Zarowin found that contrarian investment strategies generate insignificant profits. On the other hand, Chopra et al. (1992) found that even after adjusting for size and beta there is still an economically important overreaction effect which is more evident for small stocks.

The tentative conclusion that neither rational bubbles nor traditional models of return determination could explain stock price volatility suggested that a nontraditional model for return determination might be required. Fads model was suggested as an alternative model. According to the traditional models of efficient financial markets, stock prices are non-stationary and financial returns are not predictable. Summers (1986) suggested that the existence of fads in the stock market might lead to long temporary price swings that can be

modelled as a slowly decaying stationary component in prices. It is then likely that the decay over time in this component may lead to mean reversion in stock prices. Following these observations, Fama and French (1988) and Cutler et al. (1991) proposed that both the traditional model and the idea of fads can be captured by the following model (Schaller and Van Norden, 1997),

$$p_t = p_t^* + u_t, \qquad p_t^* = p_{t-1}^* + e_t \tag{2.21}$$

where p_t is the log of the stock market price in period t, p_t^* is the non-stationary component of the log price, and e_t is white noise. According to the traditional model log prices are a random walk, $u_t = 0$, and returns are white noise. We can view p_t^* as the fundamental price because it does not include a fads element. According to Schaller and Van Norden (1997), the fads model implies that stock prices have a stationary component as follows,

$$u_t = p_\varepsilon u_{t-1} + \varsigma_t \tag{2.22}$$

According to the fads model $\sigma_u^2 > 0$ and $\rho_u > 0$, where σ_u^2 is the variance of u_t.

Cutler et al. (1991) suggested to use a proxy, p_t^f, for the fundamental price p_t^*. One way to measure the proxy p_t^f is to use measurement error w_t. Using the errors-in-variables approach, this can be modelled as follows (Schaller and Van Norden, 1997),

$$p_t^f = p_t^* + w_t \tag{2.23}$$

According to Schaller and Van Norden (1997), we can express returns in terms of differences between the proxy for fundamentals and log price. Using Eqs (2.21-2.23), this relationship can be modelled using regressions of the following form,

$$p_{t-1} - p_t = \beta_0 + \beta_b\left(p_t - p_t^f\right) + v_{t+1} \tag{2.24}$$

One example of proxy for the fundamental price is the log of the real dividend. Using this proxy, we can then view Eq. (2.24) as a regression of returns on the lagged log dividend-price ratio.

According to the fads explanation of the volatility tests, noise trading by naive investors plays a significant role in stock price determination. One simple way to think through the possible effects of fads is to add a factor due to noise trading to the level or log of what would be the fundamental price if expected returns were constant (Poterba and Summers, 1988). In one interpretation, fads mean that even after risk adjustments there are still profitable opportunities for smart investors with long enough horizons. In another interpretation, it means that while some fraction of trading is done by naive traders, another fraction of trading is done by sophisticated traders who ensure that there are no extraordinary expected return opportunities once risk is taken into account (West, 1987). This does not mean that stock prices are driven to whatever level they would be in the absence of fads. Risk is created by naive investors and sophisticated investors take into account this risk. Such risk, however, might not be captured by traditional models of return determination (West, 1987).

Some researchers suggested that the predictability of stock returns is a consequence of rational time variation in expected returns as business conditions, investment opportunities, and risk aversion change through time (Fama and French 1989; Fama 1990; Chen 1991). This seems to be a more rational explanation if we consider that the variation in expected returns is common across assets and related to business conditions.

Expected returns on stocks have been found to be higher around business cycle troughs when economic conditions are weak but anticipated to improve, and lower around business cycle peaks when economic conditions are strong but anticipated to deteriorate. A number of authors have claimed that the countercyclical behaviour of expected returns is consistent with consumption smoothing stories and intertemporal equilibrium models (Merton 1973; Breeden 1979; Balvers et al. 1990). These studies suggested that if income is high relative to wealth around business cycle peaks, investors might attempt to smooth their consumption by saving into future periods when output and income might be lower. If there is no increase in capital-investment opportunities, a desire to save would result in lower expected returns. On the other hand, expected returns might be high near business cycle troughs if economic conditions are poor and income is low.

Fama and French (1989) observed that the variation in capital investment opportunities might also affect stock returns. They suggested that poor prospects for future activity and investments near business peaks might result in low expected returns, whereas good prospects for future activity and investment near business cycle troughs might result in high expected returns.

Concerning the evidence that supports the view that markets are efficient, a number of

researchers have argued that the particular statistical tests that have been used are so weak that even an efficient market would pass them. Stambaugh (1986) suggested that a uniformly powerful test might exist only in the dreams of statisticians. Kim et al (1991) and McQueen (1992) suggested that the mean reversion documented in prior tests may be overstated due to two main statistical artifacts: first, the OLS estimates are inefficient because they place unnecessary weight to the economic depression during the World War II period that is characterised by larger error variances and strong mean-reverting tendencies; and second, the small sample sizes that result from the use of long horizon returns impair the power of the tests. Fama (1991) emphasised that past returns do indeed contain information about expected returns but they represent a very noisy signal. He suggested that a more powerful test should take into account explanatory variables that contain more precise information about expected returns. Studies that examined such tests are discussed in the next section.

2.7 EX-ANTE OBSERVABLE VARIABLES AND STOCK RETURN PREDICTABILITY

Univariate tests on long-horizon returns provide imprecise evidence of time-series return predictability. One of the main statistical shortcomings of these tests is that small sample sizes impede reliable inferences about the time-series properties of long-horizon returns. On the other hand, multivariate tests have been suggested in the literature as an alternative approach to stock return predictability. According to this approach, the returns are regressed on a set of ex-ante observable variables. Several ex-ante observable variables have been suggested in the literature as an alternative approach to stock return predictability. The most important of these variables are presented below.

2.7.1 Dividend Yields

There is a substantial amount of contradicting evidence in the recent literature whether dividend yields can be used to predict stock returns. A number of studies such as those by Rozeff (1984), Shiller (1984), Campbell and Shiller (1988), Fama and French (1988), Hodrick (1992), and Nelson and Kim (1993) among others, support the view that the so-called dividend yield can be used as a measure of expected stock returns. For example, Rozeff (1984) found that dividend yields explain 14% of the variation in the S&P composite index over the 1926-1981 period. Shiller (1984) also examined the predictability of annual S&P composite returns using the dividend yield as a predictor variable. After experimentation, he found that dividend yields explain nearly 16% of the variation in the S&P composite index during the 1946-1983 period. On the other hand, Shiller reported that both dividend yields and earning yields have little explanatory power during the 1898-1945 period.

Fama and French (1988) used dividend/price ratios, so called dividend yields, to forecast returns on the value- and equally-weighted portfolios of NYSE stocks for return horizons of one-month to four-years. Their tests confirmed the evidence that the predictable component of returns is a small fraction of short-horizon return variances. Dividend yields typically explain less than 5% of the variances of monthly or quarterly returns. An important finding, however, is that the predictable component of returns is a larger fraction of the variation of long-horizon returns. Regressions of returns on dividend yields often explain more than 25% of the variances of two- to four-year returns. The results of this study provide evidence of forecast power for sub-periods as well as for the 1927-1986 sample period.

Harvey (1991) found that U.S. dividend yields and term structure variables can be used to predict monthly returns on various common stock portfolios. Campbell and Hamao (1992) provide similar evidence for Japanese and U.K. stocks. Goetzmann and Jorion (1993) suggested that the stock return regressions that have been applied in previous studies suffer from the statistical problems of strong dependency structures and biases in the estimation of regression coefficients. Therefore, they claimed that the previous findings on the ability of dividend yields to predict stock returns might be overstated due to the small sample behaviour of commonly used inference methods. After applying a different sampling methodology, Goetzmann and Jorion concluded that there is no significant relationship between dividend yields and stock returns. However, Wolf (1997) argued that the sampling approach used by Goetzmann and Jorion (1993) is not backed up by strong theoretical properties while it requires a custom-tailoring to the specific situation at hand. To resolve this problem, Wolf proposed a new subsampling technique to make inference in the context of dependent and possibly nonstationary observations. This technique is based on the idea of re-computing the statistic of interest on smaller subsets of the entire data in order to approximate the sampling distribution of the estimator based on the complete data set. Wolf suggested that the advantage of this approach is the possibility to construct asymptotically correct confidence regions for unknown parameters under weak conditions. Using a variety of different indices and considering complete data starting in 1926 as well as post-war data starting in 1947, Wolf found strong evidence on stock return predictability at four-year horizons, whereas he found very weak evidence on stock predictability at short and medium horizons. However, a reorganisation of long-horizon returns avoiding increasing correlation in the residuals by means of summing dividend yields rather than returns was unable to support evidence on stock return predictability for all horizons.

2.7.2 Interest Rates

The relation between interest rate movements and common stock returns has been the subject of

a considerable amount of research in recent years. A number of studies tested the hypothesis that expected excess returns on common stock over Treasury bills are constant through time. For example, Fama and Schwert (1977) showed that in postwar U.S. data this hypothesis is strongly rejected. After regressing the monthly stock returns on the one-month bill rate, they reported that the estimated coefficient is significantly negative rather than unity as required by the hypothesis. Furthermore, the inclusion of an interest rate factor adds substantial explanatory power to a simple single-factor market model, where the return on an index of common stocks is used as a proxy for the market portfolio.

Oldfield and Rogalski (1980) analysed the response of common stock returns to statistical factors estimated on weekly returns on a set of U.S. Treasury bills during the 1964-1979 period. The empirical results of this study demonstrated that share returns are influenced by the same statistical factors that influence Treasury bill returns. Fama (1984) tested the hypothesis that expected excess returns on long-term Treasury bills or bonds over the short-term bill rate are constant. This hypothesis is more widely known as "the expectations theory of the term structure". Fama showed that this hypothesis is rejected for both bills and bonds in postwar U.S. data.

Flannery and James (1984) examined whether the interest rate sensitivity of common stock returns is related to the maturity composition of the firm's holdings of nominal contracts. He analysed the empirical relation between the interest rate sensitivity of common stock returns and the maturity composition of nominal contracts for a set of actively traded commercial banks and stock savings as well as loan associations. After experimentation, Flannery and James reported that the common stock returns of the firms included in their study are highly correlated to interest rate changes. Furthermore, they observed that the co-movement of bank stock returns and interest rate changes is positively related to the size and maturity difference between the bank's nominal assets and liabilities.

Campbell (1987) suggested that variables, which have been used to predict excess returns in the term structure, could be also used to predict excess stock returns in U.S. monthly data during the 1959-1979 and 1979-1983 periods. Therefore, there may be a payoff to simultaneous analysis of returns on bills, stocks, and bonds. For this purpose, Campbell utilised monthly time series on five asset returns and four instrumental variables. The five returns were one-month, two-month and six-month Treasury bills, a portfolio of five-to-ten-year Treasury bonds, and the value-weighted index of NYSE stocks. These returns were used to form four excess returns by subtracting the one-month bill rate from each individual return. The instruments that were used to forecast returns were the one-month bill rate, the spread between the two-month and one-

month rate, the spread between the six-month and one-month rate, and one lag of the excess return on two-month over one-month bills. The test results suggested that all asset returns, except the excess return on the bond portfolio, are predictable at conventional significant levels over the 1959-1979 sample period. Instruments that were used to forecast returns were also found to be indicators of the term structure of interest rates. In the second sample period from 1979 to 1983, the one-month bill rate and the lagged excess bill return were found significant for two-month excess returns, whereas the one-month bill rate was found significant for six-month returns. In general, Campbell's findings suggested that there are forecastable movements through time in excess returns on bills, bonds, and stocks. These movements are partially captured by a variety of term structure variables that add to the predictive power of the short interest rate alone. Although the evidence for predictability of bill and stock returns is very strong, the results for bonds are relatively weak.

Titman and Warga (1989) examined the relation between stock returns, interest rates and inflation. After applying a number of regressions, they documented that there is a positive relation between stock returns, interest rate changes, and future inflation during the 1979-1982 period. Breen et al. (1989) examined the ability of Treasury bill returns to forecast the return on the equally-weighted index as well as the value-weighted index of stocks traded on the NYSE during the 1954-1986 period. After performing a number of experiments, they found that Treasury bill interest rates can be used to forecast changes in the distribution of stock index excess returns when the index is the value-weighted portfolio. On the other hand, despite a significant negative correlation between Treasury bills and the equally-weighted index, the forecasting model did not show a significant forecasting ability in either statistical or economic terms. Breen et al. proposed that this result might be due to the leptokurtosis and January seasonal effect in the distribution of equally-weighted index excess returns. Fama and French (1989) analysed the forecast ability of long-term interest rate spreads. For this purpose, they used the difference between the yield on the Aaa bond portfolio and the one-month bill rate to examine its predictive power together with the dividend yield and the default spread over the 1927-1987 period. Fama and French concluded that the term spread is more closely related to short-term measured business cycles and captures a component of returns that is low around business cycle peaks and high around business cycle troughs. Additional evidence on the behaviour of the term spread over the business cycle is provided by Fama (1990) and Chen (1991) who reported a positive relation between the term spread and expected returns.

Lee (1992) used postwar U.S. data to investigate causal relations and dynamic interactions among stock returns, interest rates, real activity and inflation using a multivariate vector autoregression approach. The empirical results of this study demonstrated that stock returns

appear Granger-causally prior and help to explain real activity. However, if interest rates are included into the system, stock returns explain little variation in inflation while interest rates explain a substantial fraction of the inflation variation.

2.7.3 Aggregate Output

A number of studies provided empirical evidence for a linkage between expected returns and output. For example, Balvers et al. (1990) showed that changes in the equilibrium return on stocks might be predictable if there is predictability in aggregate output. To support this view, they proposed a simple equilibrium model relating output to consumption opportunities. According to Balvers et al., if investors expect lower output in the next period, they will attempt to transfer wealth to the future period of scarcity. To maximise utility, investors will attempt to smooth consumption by adjusting their required rate of return for financial assets. However, if there is a linkage between output and returns, then due to the predictability of output, returns should have a predictable component. Using annual returns on the value-weighted NYSE index over the 1947-1987 period, Balvers et al. found a negative relation between current output and future returns. Additionally, they found that the relation between output and returns appears to dominate the relation between dividend yield and returns found in Fama and French (1988).

Chen (1991) investigated the relation between state variables and changes in macroeconomy during the 1954-86 period. Among other variables, he used the lagged production growth rate over the previous 12 months as an indicator of the current state of the economy. After experimentation, Chen found that several state variables including the lagged annual production growth, the default spread, the term spread, the one-month Treasury-bill rate, and the dividend-price ratio are important determinants of future stock market returns. Another important finding was that the lagged production output is positively correlated with the growth rate of GNP from quarter $t-4$ to quarter $t-1$ as well as the next quarter t, but it is negatively correlated with the growth rate of GNP from quarter $t+1$ to quarter $t+4$. Chen reported that the lagged annual production growth appears to predict real and excess returns on the value-weighted NYSE index over the next four quarters. He interpreted the ability of state variables to forecast future market returns in terms of their correlation with changes in the macroeconomic environment. Overall, the empirical results of this study showed that the expected excess market return is negatively related to the recent growth of GNP, while it is positively related to its future growth. State variables that are positively (negatively) related to the recent growth of the economy are negatively (positively) related to the expected excess market return. Similarly, state variables that are positively (negatively) related to the future growth rates of the economy are positively (negatively) related to the expected excess market return.

2.7.4 Inflation

Early empirical findings documented a negative correlation between ex post nominal stock returns and inflation. Furthermore, empirical evidence documented a negative relation between ex ante nominal stock returns and ex ante inflation. Using U.S. data, Bodie (1976), Jaffe and Mandelker (1976), Nelson (1976), and Fama and Schwert (1977) reported that past rates of inflation are significantly and negatively associated with rates of return on common stocks for the post 1953 period. While expected stock returns and expected inflation in the U.S. have been found to be negatively related, Firth (1979) showed that in the U.K. there is a positive relation between nominal stock returns and inflation.

Considering the previous empirical evidence, Fama (1981) also attempted to explain the relation between stock returns and inflation. In a series of tests, he related real common stock returns first to other real variables, then to inflation measures, and finally to combinations of real variables and inflation measures using monthly, quarterly, and annual data during the 1953-1977 period. The test results indicated that real stock returns are positively related to measures of real activity like capital expenditures, the average real rate of return on capital, and output. The tests also indicated that stock returns exhibit strong negative simple correlation with measures of expected and unexpected inflation. However, in multiple regressions of stock returns on real variables and inflation measures, the results suggested that the most anomalous relation between the ex post stock returns and the ex ante expected inflation rate always disappears. Fama suggested that the negative stock return-inflation relations are induced by negative relations between inflation and real activity that in turn are explained by a combination of money demand theory and the quantity theory of money. According to Fama, an increase in anticipated real activity might lead to an increase in the demand for real money balances. Considering the level of nominal money, it is reasonable to assume that the increased demand for real money balances will be accommodated by a fall in the price level. If stock returns are assumed to be positively related to expected future real activity, then a negative relation between stock returns and inflation should not be considered surprising. However, this negative relation might be a proxy for a more fundamental relation between anticipated real activity and stock returns. Therefore, Fama suggested that the relation between stock returns and inflation is spurious.

Geske and Roll (1983) observed that there is a causal link in the observed negative relation between stock returns and changes in expected inflation. They suggested that according to the model proposed by Fama (1981) stock returns anticipate changes in real activity. Considering this view, changes in the government revenue are expected to vary inversely with changes in real activity. Given that the level of government expenditures is fixed, we should expect that

changes in revenue might lead to opposite changes in the government's deficit. According to Geske and Roll (1983), if the deficit is monetised, the change in money supply will result to a change in inflation. On the other hand, if the deficit is not monetised nominal interest rates may increase as a consequence of the increase in the real interest rate. However, if this process is anticipated, it is reasonable to assume that stock returns will signal changes in expected inflation. James et al. (1985) investigated the causal linkage among stock returns, real activity, money growth, and expected inflation. Using a vector autoregressive moving average (VARMA) technique, they examined the validity of a model that explains the negative relation between stock returns and inflation. The empirical results of this study confirmed the causality model proposed by Geske and Roll (1983). James et al. suggested that stock returns signal changes in expected inflation and nominal interest rates. They reported that expected changes in real activity and money supply growth can be used to predict changes in expected inflation while there is a strong link between stock returns, a proxy for expected real activity, and the growth in the monetary base.

Boudoukh and Richardson (1993) observed that previous studies on the stock return-inflation relations have focused mainly upon short-term asset returns with time horizons of one year or less. They argued that because many investors tend to hold stocks over longer periods, it is practically important to examine the manner in which stock returns move with inflation over longer horizons. Another motivation for examining the relation between stock returns and inflation over longer horizons is the fact that the results at short horizons appear to be very puzzling. Under these considerations, Boudoukh and Richardson investigated the relation between stock returns and inflation in both the U.S. and the U.K. markets using data on stocks, short-term and long-term bonds, and inflation over the 1820-1988 period. In contrast to the existing evidence at short horizons, they found that long-horizon nominal stock returns are positively related to both ex ante and ex post long-term inflation. Furthermore, Boudoukh and Richardson reported that these results are robust both in the U.S. and the U.K. markets as well as the particular subperiods chosen.

Two empirical studies that investigated the relation between stock prices and inflation in other countries using cross-sectional analysis are those by Branch (1974) and Cagan (1974). Branch reported that stocks are a partial inflation hedge while Cagan found that stocks are an inflation hedge for long-term holdings. Gultekin (1983) investigated the relation between stock returns and inflation in 26 countries including among others Canada, France, Germany, Japan, Sweden, U.K. and U.S. Using time series regressions, he found no reliable positive relation between nominal stock returns and inflation rates during the 1947-1979 period. The empirical findings of this study demonstrated that the relation between stock returns and inflation is not stable over

time while there are some differences among the countries. In general, countries with higher inflation rates appear to have higher nominal stock returns.

Part Three: Summary and Conclusions

2.8 DISCUSSION AND REMARKS

Several asset pricing theories that depict the intuitive relation between risk and return have been developed and tested empirically. One well-known model is the CAPM that evolved from the concepts of optimal portfolio selection using the Markowitz mean-variance framework. Empirical research has provided evidence that is inconsistent with the predictions of the CAPM. On the other hand, empirical research has documented evidence for the ability of firm specific variables to explain the cross-section of stock returns as alternatives to the CAPM. Indeed, firm specific variables such as Market Capitalisation, Book to Market Equity, Price-Earnings, Debt to Equity Ratios, and several other similar variables have been found to explain the cross-section of expected returns better than the original CAPM. Three explanations have been suggested in the literature for the observed ability of firm-specific variables to explain the cross-section of stock returns: first, they proxy for sensitivity to risk factors in returns and the correlation between the variables and returns reflects compensation for bearing risk; second, they help to identify stocks that are mispriced because of systematic misjudgements of investors; and third, their predictive ability is an artifact of "data snooping" and a survivorship bias in the COMPUSTAT database.

Many researchers observed that the findings for the ability of firm specific variables to explain stock returns provide evidence for multifactor alternatives to the CAPM. One of them is the APT model. Studies comparing the APT and the CAPM suggested that the APT is based on less restrictive assumptions and provides a better description of the expected returns on risky assets than the CAPM. The APT is more robust than the CAPM because it makes no strong assumptions about individuals' utility functions as well as assumptions about the empirical distribution of asset returns. On the other hand, the APT allows the equilibrium returns of assets to be dependent on many factors and investors are not assumed to select portfolios on the basis of expected return and variance. Finally, the efficiency of market portfolio is not of a primary concern in the APT as the original CAPM. Therefore, the difficulties associated with the market portfolio are avoided. However, the APT is not a perfect model and it has several disadvantages. Its major disadvantage is that the usual statistical methods are not adequate to test an approximate pricing relation. Therefore, the tests of the APT are joint tests of the model and additional assumptions are necessary to obtain exact pricing. On the other hand, the empirical tests to identify the factor structure in security returns have not produced consistent

results. It has been suggested that the APT would be a better model if the factors could be related to more identifiable sources of economic risk. Certainly, more developments in asset pricing theory and econometrics are required in order to understand the relationship between return factors and economic risks.

Many of the tests of asset pricing models assume that expected returns, betas, and risk premia are constant through time and independent of any ex-ante known information. These studies focus on average values and examine whether differences in average risk can explain differences in average returns across assets. The research on time series predictability documented evidence that expected returns are not constant through time. More recent research has attempted to integrate the time-series properties of conditional moments with the cross-sectional implications of asset pricing models in a framework known as conditional asset pricing models. Conditional asset pricing models allow for time-varying expected returns and assume that the return distribution depends on a set of ex-ante observable variables. There are several advantages associated to conditional asset pricing models over unconditional ones: first, conditional models allow for time variation in expected returns, betas, and risk premia and hence they are more consistent with the evidence reporting predictable component in the time series of returns; second, they provide a useful framework to integrate the time series predictability of stock returns with the cross-sectional implications of asset pricing models; and third, conditional asset pricing models avoid the problem that the omission of conditioning information can mislead the conclusions regarding the validity of various asset pricing models. However, conditional asset pricing models have two major drawbacks. The first drawback is that a particular model for the conditional expectations has to be specified. The form of the function depends on the joint probability distribution of the returns and the instrumental variables. Although most studies have adopted a linear relation for the first moments, other relations may also be valid. The second drawback is that it may be difficult to conclude anything about the validity of an asset pricing model given the fact that all information that investors use to set prices is unobservable and that the econometrician only uses a reduced set of information.

Most of the studies in market efficiency assumed a common equilibrium model of stock prices in which expected returns are constant through time. The model of constant expected returns implies that prices should follow a random walk. In other words, if market is efficient then information in past price changes should not be relevant in predicting future changes. Previous studies have examined this notion of market efficiency by testing whether return autocorrelations are equal to zero. Most of these studies documented return autocorrelations that were more evident for long horizon returns. However, although the empirical evidence suggests

that the estimated autocorrelations might be statistically significant, they are not able to support the view that the return autocorrelations will always be economically significant as well. Three explanations have been given in the literature to explain the negative serial correlation in long horizon returns. One explanation suggests that the predictability of stock returns is due to some form of irrationality such as fads, speculative bubbles or noise trading that deviates stock prices from their fundamental values and therefore generates abnormal returns. The second explanation suggests that the predictability of stock returns is a consequence of rational time variation in expected returns as business conditions, investment opportunities, and risk aversion change through time. This seems to be a more rational explanation if we consider that the variation in expected returns is common across assets and related to business conditions. Finally, the third explanation suggests that the particular statistical tests that have been used are so week that even an efficient market would pass them.

Univariate tests on long-horizon returns provide imprecise evidence of time-series return predictability. One of the main statistical shortcomings of these tests is that small sample sizes impede reliable inferences about the time-series properties of long-horizon returns. On the other hand, multivariate tests have been suggested in the literature as an alternative approach to stock return predictability. According to this approach, the returns are regressed on a set of ex-ante observable variables. Several ex-ante observable variables have been suggested in the literature as an alternative approach to stock return predictability. These variables include among others dividend yields, measures of inflation, interest rates, and aggregate output. Empirical studies documented evidence for the ability of these variables to predict stock returns. Finally, some temporal seasonal patterns in returns that are related to calendar turning points may have practical value only for those investors who are planning to trade in any event.

In summary, the empirical evidence presented in this Chapter suggests stock price behaviour that is inconsistent with the CAPM, which implies that security rates of return must be linearly related to the rate of return on the market portfolio. Other alternatives such as firm specific variables and measures of the state of the macroeconomy provide support for more general multifactor models.

We also observe that the models that have been applied so far to explain the cross-section of stock returns are not always suitable to deal with non-linearities and other complex patterns that have been documented in the financial processes. A more general model is required that will be more flexible to cope with non-linearities and other complex processes that are present in the financial data. These more general models are discussed in the next Chapter.

CHAPTER 3: CLASSIFICATION RULES - SUPERVISED LEARNING

The econometric methods that we discussed in the previous Chapter have been designed to detect either linear structures or strictly defined forms of non-linearity in the financial data. For example, the CAPM and the APT are based on linear models of expected returns, whereas ARCH- and GARCH-type models are based on strictly defined non-linear models. However, these models might not be suitable if the data is fuzzy, chaotic, or exhibits unpredictable non-linearities. Empirical evidence suggests that financial data exhibit complex non-linearities and other patterns that are time-varying, inconsistent and possibly chaotic in the time scale. For example, investors' attitudes toward risk and expected return are non-linear. The strategic interactions among market participants, the process by which information is incorporated into security prices, and the dynamic fluctuations of the economic environment are all inherently non-linear (Campbell et al., 1997). Furthermore, we have to consider that several factors such as human judgements, human emotions, human feelings, human expectations, psychology, politics, and other qualitative factors affect in a high degree the process that drives stock prices. Under these considerations, most of the models that have been used so far to predict stock returns may not be able to deal with the actual process in the financial data, if that data lacks a well-defined physical content. More powerful models should be applied that will be able to extract the hidden knowledge in the financial data that cannot be detected by either linear or well-defined non-linear models.

One consequence of the rapid development of computer power in the 1980s was the development of computer-intensive classification algorithms. These include among others neural networks, decision trees and rule induction techniques. These computer-intensive methods can be used to address the problem of stock predictability in the form of statistical classification.

Statistical classification is also known as supervised learning. Supervised learning is a form of learning from a sample of previously known objects. Each object is described by a set of observations and a class label. Given that the objects are known to come from one of C_{jt} distinct classes (j = 1, 2...k) of observations, we wish to find functions of these observations that will distinguish the classes, and that will enable us to assign a new object to one of these classes on the basis of m measured characteristics, x, associated with this object.

In our application, we are particularly interested in whether a particular share will be classified as a H (high) or L (low) excess return share based on some relevant information. Let us assume

that y_{it} is the 1-year-ahead excess return on some share i bought at time t, and x_{it} is the i vector of information attributes known at time t. The idea is to apply a classification method to assign y_{it} to one of the two classes $C_{jt}(j = H, L)$ depending on whether or not this return is above or below, say, the 25% threshold percentile that has been decided after ranking the returns in excess of an equally-weighted index. The models input is the vector x_{it} of variables that represent useful information. There are two distinct objectives here: first, class separation; and second, classification of future shares as H or L. The class separation is handled through the use of a discriminant function. The classification of future shares is handled through a classification rule.

Discriminant functions and classification rules are very closely related because the best function for class separation often provides the best allocation rule for future shares. This relationship can be described as follows. Initially, we attempt to construct the discriminant functions using a sample of objects that are described by measurement vectors x_{it} of relevant information attributes and the associated excess return class labels, y_{it}. This sample of previously known objects is known as the training set. Each object is associated with a particular vector of measurements that corresponds to a single point in the measurement space. Using the training set, we partition the measurement space into regions such that points falling in one region correspond to one class, whereas points falling to another region correspond to the other class. The discriminant functions that determine the partition of the measurement space into regions represent either decision hypersurfaces or have other forms depending on the classification method that is applied for each particular application. For example, in parametric and semi-parametric classification methods particular distributional forms are assumed for the discriminant functions. In non-parametric classification methods the discriminant functions have other forms that do not require distributional assumptions (Hand, 1997). In both cases, however, we use the information from the discriminant functions and we attempt to derive a general classification rule that allow us to predict the class label of future shares based on the measurement vectors of relevant information attributes associated with these shares. The main approaches of the supervised learning paradigm are discussed in this Chapter.

This Chapter is divided into four parts. In the first part, we discuss parametric, semi-parametric, and non-parametric smoothing classification rules. In the second part, we review neural networks, and data mining techniques focusing on decision tree and rule induction algorithms. In the third part, we discuss empirical evidence from comparative studies on supervised learning algorithms. Finally, in the fourth part, we present the summary and conclusions.

Part One: Parametric, Semi-Parametric, and Non-Parametric Smoothing Methods

3.1 THE THEORETICAL APPROACH TO DISCRIMINATION

3.1.1 Theoretical Background

The theoretical approach to discrimination was developed by Welch (1939) who considered classification as a decision problem. Let us assume that we have to classify a random vector $x_{it} = (x_{i1t}, x_{i2t}, \ldots, x_{imt})$ in one of the population classes H and L. Furthermore, let us assume that $f^H(x_{it})$ and $f^L(x_{it})$ are the probability density functions of x_{it} in population classes H and L, respectively. Welch suggested to partition the m-dimensional sample space into disjoint regions and classify x_{it} to a particular population class according to the region into which it falls. If we consider the partition of the m-dimensional space into regions R_H, R_L we classify x_{it} to the population class H if it falls in region R_H and to the population class L if it falls in region R_L. If the regions R_H and R_L are mutually exclusive, then their union includes the entire space R. In this case, we can write the total probability of misclassification as follows (Lachenbruch, 1975),

$$Pr_{R,f} = p_H \int_{R_L} f^H(x_{it})dx + p_L \int_{R_H} f^L(x_{it})dx \tag{3.1}$$

where p_H is the proportion of population class H in the entire population, and $p_L = 1 - p_H$ the proportion of population class L in the entire population. Eq. (3.1) can be minimised if R_H is chosen so that $p_L f^L(x_{it}) - p_H f^H(x_{it}) < 0$ for all points in R_H. Therefore, we assign x_{it} to population class H if $\frac{f^H(x_{it})}{f^L(x_{it})} > \frac{p_L}{p_H}$, and to the population class L otherwise. An alternative procedure would be to minimise the total cost of misclassification. It may be more costly to misclassify a member of population class L to class H than it is to misclassify a member of population class H to class L. Let as assume that c_H is the cost of misclassifying a member of population class H, and c_L is the cost of misclassifying a member of population class L. We wish to find R_H and R_L to minimise the following quantity (Lachenbruch, 1975),

$$J = c_H p_H \int_{R_L} f^H(x_{it})dx + c_L p_L \int_{R_H} f^L(x_{it})dx \tag{3.2}$$

The above expression can be minimised if R_H is chosen so that $c_L p_L f^L(x_{it}) < c_H p_H f^H(x_{it})$. This is equivalent to choosing population class H if $\frac{f^H(x_{it})}{f^L(x_{it})} > \frac{p_L c_H}{p_H c_L}$ is true. Welch (1939) observed that it is difficult to quantify the costs due to misclassification. Common practice is to set $c_L = c_H$. However, if prior probabilities are unknown then we can set additionally that $p_H = p_L$. In this case, we assign x_{it} to the population class $C_{jt}(j = H, L)$ that has the greater probability $f^j(x_{it})$. Therefore, the classification rule is to allocate x_{it} to population class H if $\frac{f^H(x_{it})}{f^L(x_{it})} > \alpha$ for some suitable constant, α, and otherwise to population class L. A detailed analysis of this approach in the case of more than two populations can be found in Lachenbruch (1975).

3.1.2 Bayes Theorem Approach

Another approach to derive the above classification rules would be to assign x_{it} to the class with the largest posterior probability. Let us assume that $f^j(x_{it})[j = H, L]$ denotes the conditional probability of x_{it} given the class j. The posterior probability of class j given x_{it} can be expressed by Bayes theorem as follows (Lachenbruch, 1975),

$$f(j \setminus x_{it}) = \frac{p_j f^j(x_{it})}{p_H f^H(x_{it}) + p_L f^L(x_{it})} \qquad j = H, L \tag{3.3}$$

The classification rule suggests to assign x_{it} to H if $f(H \setminus x_{it}) > f(L \setminus x_{it})$ which is equivalent to the rule that minimises the total probability of misclassification.

3.2 PARAMETRIC CLASSIFICATION RULES

The parametric approach includes four steps to construct a discriminant function or a classification rule: first, we construct a probability model for each of the population classes under consideration; second, we specify an objective function to be optimised; third, we derive the best population discriminant function or classification rule; and fourth, we estimate this function using the sample data. The mathematical formulation of this approach is discussed in detail below.

3.2.1 Theoretical Framework

A parametric discriminant function can be derived if we assume an appropriate form for the densities $f^j(x_{it})[j = H, L]$ in the two population classes H and L. One form that has been used extensively in the literature is the case of multivariate normal density functions because the resulting discriminant functions have been found more robust to departures from normality.

Let us assume that we have to classify the individual random vector x_{it} into one of the population classes H and L. If we assume that all elements of x_{it} are continuous, then we can assume that their joint distribution will be multivariate normal with means μ_H, μ_L and dispersion matrices V_H and V_L in population classes H and L, respectively. Furthermore, let us assume that $f^H(x_{it})$ and $f^L(x_{it})$ represent the population density functions of x_{it} into population classes H and L, respectively. If we substitute these functions in $\frac{f^H(x_{it})}{f^L(x_{it})} > \alpha$ and simplify, we can express the classification rule as follows (Krzanowski and Marriott, 1995),

$$\frac{1}{2}\log\left(\frac{|V_L|}{|V_H|}\right) - \frac{1}{2}\left[x'_{it}\left(V_H^{-1} - V_L^{-1}\right)x_{it} - 2x'_{it}\left(V_H^{-1}\mu_H - V_L^{-1}\mu_L\right) + \mu'_H V_H^{-1}\mu_H - \mu'_L V_L^{-1}\mu_L\right] > \log\alpha \tag{3.4}$$

If the above inequality holds we classify x_{it} to population class H, otherwise we classify x_{it} to L. The above rule is known as the quadratic discriminant function due to the presence of quadratics terms in elements of x_{it}. However, if we assume that $V_H = V_L = V$, then we allocate x_{it} to population class H if the following inequality holds (Krzanowski and Marriott, 1995),

$$(\mu_H - \mu_L)'V^{-1}\left[x - \frac{1}{2}(\mu_H + \mu_L)\right] > \log\alpha \tag{3.5}$$

and otherwise we classify x_{it} to population class L. The assumption of common dispersion matrices produced a linear discriminant function. The above analysis can be extended if we assume more than two populations. A detailed analysis in case of more two populations can be found in Krzanowski and Marriott (1995).

The rules presented in inequalities (3.4) and (3.5) are population classification rules that involve unknown population parameters. One possible way to estimate these rules is to obtain a sample of individuals from each of the population classes and base the estimation on these samples. Three approaches have been proposed in the literature for the estimation process: the estimative

approach, the predictive approach, and the testimative approach. A brief description of these approaches is given below (Krzanowski and Marriott, 1995).

- ***The Estimative Approach:*** According to this approach, we estimate all parameters from the training samples and then we replace each parameter by its estimate. If the prior probabilities, p_j, are unknown there are two possibilities: if sampling has been done from a mixture of population classes, then we can consider the sample sizes, n_i, as random variables with a multinomial distribution and we can estimate p_j as $\tilde{p}_j = \frac{n_j}{n}(j = H, L)$. On other hand, if sampling has been done for each population class, separately, then no sample estimates of the p_j are available from the data. In this case, we can make the assumption of equal prior probabilities if no other information is available.

- ***The Predictive Approach:*** This approach suggests to replace each parameter by just a single value which is supposed correct but which in practice may be wildly inaccurate. An alternative can be to incorporate variability of parameter estimates through their joint posterior distribution, and then to average the densities of x_{it} over this posterior distribution to obtain their predictive distributions. Use of these predictive distributions leads to a predictive allocation rule.

- ***The Testimative Approach.*** This approach suggests to estimate the ratio densities $\frac{f^H(x_{it})}{f^L(x_{it})} > \frac{p_L c_H}{p_H c_L}$ more directly through a likelihood ratio. In the case of two population classes, we calculate the likelihood ratio test statistic for the null hypothesis that x_{it} belongs to the population class H against the alternative that it belongs to population class L. The test statistic is taken as the maximum of the likelihood of two training samples one from each population class H, L with x_{it} added to the first sample, divided by the maximum of the likelihood of the same two samples but with x_{it} now added to the second sample.

3.3 SEMI-PARAMETRIC APPROACH

The semi-parametric approach suggests to specify a parametric form directly from the discriminant function or classification rule and then estimate the parameters of the function from the sample data. According to this approach, the parameters occur only in the function and not in the probability models. Functional forms that have been used in this semi-parametric fashion include the Fisher's linear discriminant function, the quadratic discriminant function, and the logistic function. These approaches are described in more detail below.

3.3.1 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) has a very important contribution in the allocation rule methodology. Although the basic algorithm was developed several decades ago, it is still widely used in a variety of extensions and implementations. The main idea of this algorithm is fairly simple. Let $f(H \setminus x_{it})$ be the probability that the share with measurement vector x_{it} will belong to class H. The aim of LDA is to construct an estimate of $f(H \setminus x_{it})$ and then to compare this estimate with a threshold. If the estimate is greater than the threshold, then the share with measurement vector x_{it} will be classified as H. If the estimate is lower than the threshold, then the share will be classified as L (Hand, 1997).

One way to estimate $f(H \setminus x_{it})$ is to construct a simple linear combination of the independent variables and then to use this combination as a basis for classifying future shares into one of the classes. Let us denote the linear function as $x\alpha'$. In the case of two classes H and L, we can estimate α by finding the direction in which the two classes are well separated. One criterion would be to use a linear combination of the observations, and choose the coefficients so that the ratio of the difference of the means of the linear combination in the two classes to the variance is maximised. If we assume that the covariance matrices in the two classes are equal so that $C_H = C_L = C$ and $\bar{x}_H, \bar{x}_L, V$ denote the sample means and the sample standard deviation in H, L, respectively, then we have to choose, α, to maximise (Lachenbruch, 1975),

$$\phi = \frac{(\alpha'\bar{x}_H - \alpha'\bar{x}_L)^2}{\alpha' V \alpha} \tag{3.6}$$

Differentiating ϕ with respect to α we get $\alpha = kV^{-1}(\bar{x}_H - \bar{x}_L)$ for arbitrary constant multiplying factor k which is usually taken as one.

According to this algorithm, we classify a share as H if $F = (\bar{x}_H - \bar{x}_L)' V^{-1}x$ is closer to $\bar{F}_H = (\bar{x}_H - \bar{x}_L)' V^{-1}\bar{x}_H$ and as L if F is closer to $\bar{F}_L = (\bar{x}_H - \bar{x}_L)' S^{-1}\bar{x}_L$. The midpoint of the interval between $\bar{F}_H$ and $\bar{F}_L$ is given as follows (Lachenbruch, 1975),

$$\frac{(\bar{F}_H + \bar{F}_L)}{2} = \frac{1}{2}(\bar{x}_H - \bar{x}_L)' V^{-1}(\bar{x}_H + \bar{x}_L) \tag{3.7}$$

F is closer to $\bar{F}_H$ if $|F - \bar{F}_H| < |F - \bar{F}_L|$ which occurs if $F > \frac{1}{2}(\bar{F}_H + \bar{F}_L)$ since $\bar{F}_H > \bar{F}_L$.

Figure 3.1 illustrates the LDA algorithm for the case of two predictor variables: Price-Earnings (P/E) ratio, and Book Equity to Market Equity (BE/ME) ratio. As we can see, the LDA imposes two elliptical distributions over the means of the two classes. The line connecting points of equal distance from the two neighbouring means defines a linear partition of the space into H and L regions.

The LDA model has been applied successfully in a wide variety of applications including financial distress prediction (Altman, 1968), bond ratings prediction (Kaplan and Urwitz, 1979) etc. Generally, this algorithm is favoured if there are linearities in the data. However, the LDA model has been extensively criticised in the literature because the validity of its results hinges on restrictive assumptions (Eisenbeis 1977; Altman and Eisenbeis 1978; Tollefson and Joy 1978; Ohlson 1980; Pinches 1980; Odom and Sharda 1990). For the linear discriminant function to provide a classification rule that minimises the probability of misclassification, the variables in each group must be from multivariate normal distributions and the covariance matrices for all groups must be equal. These requirements, however, have frequently been violated. Empirical evidence seems to indicate that most ratio distributions are either highly skewed, flat, and/or dominated by outliers (Deakin 1976; Karels and Prakash 1987). Empirical evidence also suggests that nonmultivariate normality influences the test for the equality of the dispersion matrices (Gilbert 1969; Mardia 1971). Remedial measures taken to improve the multivariate normality are often inadequate (Lachenbruch et al., 1973). Despite these criticisms, however, this algorithm is still used in a variety of applications because it offers statistical simplicity and interpretation.

3.3.2 Quadratic Discriminant Analysis

The LDA is based on the assumption of equal population covariance matrices for classes H and L, respectively. However, it should be noted that the equal covariance assumption is rarely satisfied. Suppose we assign x_{it} to class H if x_{it} is in some region R_H and to class L if x_{it} is in some region R_L. When the covariance matrices are quite different and normality holds then we assign x_{it} to class H if the following inequality holds (Lachenbruch, 1975),

$$Q(x_{it}) = \ln\left(\frac{f^H(x_{it})}{f^L(x_{it})}\right) > \left(\frac{1-p_H}{p_H}\right) =$$

$$\frac{1}{2}\ln\left|\frac{V_2}{V_1}\right| - \frac{1}{2}(x_{it}-\bar{x}_H)' V_H^{-1}(x_{it}-\bar{x}_H) + \frac{1}{2}(x_{it}-\bar{x}_L)' V_L^{-1}(x_{it}-\bar{x}_L) \qquad (3.8)$$

where, $f^H(x_{it})$ is the density function of x_{it} if it comes from class H, $f^L(x_{it})$ is the density function of x_{it} if it comes from class L, p_H is the proportion of class H in the sample, and $p_L = 1 - p_H$ is the proportion of class L in the sample. If inequality (3.8) does not hold then we assign x_{it} to class L. In this case, we have a quadratic discriminant function since $V_H^{-1} - V_L^{-1}$ does not vanish. In practice, deviations from normality tend to affect this function rather seriously.

3.3.3 Logistic Discriminant Function

Logistic functions represent another form of semi-parametric discrimination. This method is partially parametric because the ratios between the actual probability density functions of the classes are modelled rather than the actual probability density functions. More specifically, the algorithm requires that the logarithms of the prior odds p_H / p_L times the ratios of the probability density functions for the classes are modelled as linear functions of the attributes. For two classes, we can write the following expression (Michie et al. 1994),

$$\log \frac{p_H f^H(x_{it})}{p_L f^L(x_{it})} = \alpha + \beta' x_{it} \tag{3.9}$$

where α and β are the parameters to be estimated. In practice, the parameters can be estimated by maximum conditional likelihood. According to this method, the conditional class probabilities for classes H and L take the forms (Michie et al. 1994),

$$f(H \setminus x_{it}) = \frac{\exp(\alpha + \beta' x_{it})}{1 + \exp(\alpha + \beta' x_{it})} \tag{3.10}$$

$$f(L \setminus x_{it}) = \frac{1}{1 + \exp(\alpha + \beta' x_{it})} \tag{3.11}$$

If we assume independent samples from the two classes, the conditional likelihood for the parameters α and β can be written as follows (Michie et al. 1994),

$$CL(\alpha,\beta) = \prod f(H \setminus x_{it}) \prod f(L \setminus x_{it}) \tag{3.12}$$

The parameter estimates are the values that maximise this likelihood and they can be found by iterative methods (Day and Kerridge 1967). A new example is classified to the class that has the

highest posterior probability. A more detailed analysis of logistic discriminant analysis for the case of more than two populations can be found in Michie et al. (1994) and Krzanowski and Marriott (1995).

3.3.4 Advantages and Disadvantages of the Parametric and Semi-Parametric Approach

Parametric and semi-parametric classification rules have been developed by the explicit assumption of some model. Therefore, they offer statistical simplicity and interpretation. In addition, the formulation of the particular models make easy to assess the relative contribution of the input variables in the decision process. Furthermore, the learning process is fast and does not require excessive computational resources. One disadvantage of the parametric approach is that decision boundaries often depend on the tails of the assumed distributions and these are chosen so they can fit well in the centre of the classes. This makes necessary the replacement of multivariate normal by multivariate normal t-distributions (Ripley, 1992).

One disadvantage of the parametric and semi-parametric classification rules is that most of these rules will be fairly robust to certain assumptions for the assumed model. For example, two assumptions should be satisfied for the linear discriminant function to provide a classification rule that minimises the probability of misclassification: first, the variables in each group must be from multivariate normal distributions; and second, the covariance matrices for all groups must be equal. Empirical evidence suggests that these requirements have frequently been violated. Remedial measures taken to improve the multivariate normality are often inadequate.

3.4 NON-PARAMETRIC CLASSIFICATION RULES

Parametric and semi-parametric classification rules have been developed by the explicit assumption of some model. On the other hand, non-parametric discriminant procedures do not postulate models for the population-conditional distributions. They suggest instead to estimate first the probability density functions from the training data, and then apply the estimated functions to the individuals for classification. Some of the main approaches to non-parametric discrimination are discussed below.

3.4.1 Kernel Method

A common approach to non-parametric discriminant analysis is kernel discriminant analysis. According to this classification method, the class-conditional densities are replaced by their kernel estimates in the defining expressions for the posterior probabilities of class membership and consequent classification rule. Using the kernel function, the estimate of the probability

density at x_{it} using data sample X_{it} for class $C_{jt}(j = H, L)$ can be expressed as follows (Hand, 1997),

$$\tilde{f}^{j}(x_{it}) = \frac{1}{ns}\sum_{i} K\left(\frac{x_{it} - X_{it}}{s}\right) \tag{3.13}$$

where, K(.) is the kernel function, and s is a smoothing parameter. If K is a unimodal density function then $\tilde{f}^{j}(x_{it})$ will itself be a density so that $\int \tilde{f}^{j}(x_{it})dx = 1$ and $\tilde{f}^{j}(x_{it}) \geq 0$ for all x_{it}. The smoothing parameter s is a function of the i^{th} class-sample size n_i. Parzen (1962) investigated the large-sample properties of the kernel estimator in the univariate case, and Cacoulos (1966) extended the analysis in the multivariate case.

Although much work has been done on kernel methods, the main questions concern appropriate choice and standardisation of kernel functions, methods of estimating the smoothing parameter s, and development of fast algorithms to do the computations. A major practical drawback of the kernel method of density estimation is its inability to deal effectively with the tails of the distribution without oversmoothing the main part of the density.

3.4.2 Generalised Additive Models

Additive models represent an intermediate approach between kernel methods and linear methods. Kernel models use local models to model the function $f(j \setminus x)$ over a multivariate x_{it}, whereas linear models form a single global model. Additive models have the following form (Hand 1997),

$$f(j \setminus x_{it}) = \alpha_j + \sum_{k} g_{jk}(x_{kt}) + \varepsilon_t \tag{3.14}$$

The above expression represents a more general transformation of the variables than the one permitted by the linear model. If we allow g to represent a set of basis functions, this is equivalent to replacing the summation of the variables in model (3.14) by another summation that represents a larger set of transformed versions of the variables where each of the raw variables is used to generate a set of new derived variables. However, this approach would result in too many functions that would require a very large training set. A good alternative would be to use spline functions as basis functions. These functions consist of segments that are polynomial of degree k between specified values of x_{kt} which are called knots. In addition, the polynomials are differentiable of order $k-1$ at the knots. The overall function is piecewise

polynomial and its flexibility is determined by the order of the polynomial segments and the number of knots. The estimation of parameters in methods using basis functions can be made using ordinary regression on the extended variable space (Hand, 1997). One variant of the spline approach is multivariate adaptive regression splines (MARS) that were proposed by Friedman (1991).

3.4.3 Projection Pursuit Regression

Projection pursuit is a technique for finding projections of multivariate data into spaces of lower dimension. The original purpose of projection pursuit was to machine-pick interesting low dimensional projections of a high dimensional point cloud by numerically maximising a certain objective function or projection index.

Methods based on local averaging can perform poorly in high dimensions owing to the sparseness of the data. Projection pursuit regression tackles this problem by fitting a smooth function of a single linear function of x_{it}. This fitting is optimised by choosing the linear function to minimise some measure of residual variability. In the next step, the residuals from the first fit are fitted by a smooth function of another linear function and the process is continued until the fit is considered satisfactory. In the case of a single univariate output, the PPR can be written as follows (Hand, 1997),

$$f(j \mid x_{it}) = \alpha_j + \sum_k g_{jk}(\beta' x_{it}) + \varepsilon_t \qquad (3.15)$$

where the coefficients α_j, β and the functions $g_{jk}(.)$ are estimated from the data. Eq. (3.15) is a sum of transformations of the linear combinations $\beta' x$ of the raw variables.

In the case of two classes, the linear combination will define a direction in the x space. Hand (1997) observed that if only one g function is used, the estimated probabilities will provide constant contours in directions orthogonal to β and the gaps between these contours will depend on the g function. However, if more than one g function is used then it will be difficult to interpret the model.

Many of the methods of classical multivariate analysis turn out to be special cases of projection pursuit. Examples are principal components, discriminant analysis, and the quartimax and oblimax methods in factor analysis. The more interesting projection pursuit algorithms are able to ignore noise and information-poor variables. This is a distinct advantage over methods based

on interpoint distances like minimal spanning trees, multidimensional scaling, and most clustering techniques. PPR is a suitable procedure for multivariate non-parametric smoothing and prediction. However, these is some difficulty in interpreting the vectors selected by the procedure.

3.4.4 Nearest Neighbour

Nearest neighbour methods estimate the conditional probability that a share with measurement vector x_{it} will belong to class $C_{jt}(j = H, L)$. This probability is estimated by the proportion of the training set points in the neighbourhood area of x_{it} that belong to class j. The neighbourhood area can be defined by the distance from x_{it} to the k^{th} nearest point from the training set. Using this definition, we can estimate the probability, $\tilde{f}(j \setminus x_{it})[j = H, L]$, that the share with measurement vector x_{it} will belong to class j as the proportion of points in the training set that are in class j amongst the k nearest to measurement vector x_{it}. We then classify a future share with measurement vector x_{it} to the class with the largest estimated probability. The greater the k, the lower will be the variance in the probability estimates but it is likely that more bias will be introduced in the estimate $\tilde{f}(j \setminus x_{it})$. The smaller the k, the higher will be the variance in the probability estimates but it is likely that less bias will be introduced in the estimate $\tilde{f}(j \setminus x_{it})$ (Hand, 1997).

One problem with nearest neighbour methods is to determine the shape of the neighbourhood area of x_{it} in the multivariate case. The parameter k determines how large the neighbourhood area will be, but it does not determine its shape. The only way to determine the shape of the neighbourhood of x_{it} is to choose a distance metric in order to measure nearness. The most commonly used distance measure is the squared Euclidean distance. Sometimes its square root, the Euclidean distance, is also used. One disadvantage of the square Euclidean distance is that it depends on the units of measurement for the variables. Another disadvantage is that it is sensitive if variables are measured on different scales. For example, variables that are measured in larger numbers will contribute more to the computed distance than variables that are recorded in smaller numbers. One way to solve this problem would be to express all variables in standardised form. However, the form of standardisation is also crucial because it should not affect the discriminatory power of the variables.

There are several parameters of a k nearest neighbour classifier that can be varied to achieve a design goal. Such parameters can be the distance metric or the neighbourhood size. Some

success in building accurate and efficient independent nearest neighbour classifiers has been shown by selecting appropriate prototypes. The prototypes are the examples of a category that are best in some sense, such as the most typical. According to the prototype model of classification, a new example is classified as in or out of a category by computing a weighted similarity of the features possessed by the prototype and the example. Several prototype selection algorithms have been proposed in the literature. A very good review of prototype selection algorithms can be found in Skalak (1997). Prototype selection algorithms have two goals: first, to reduce the computational expense of applying the nearest neighbour algorithm; and second, to increase the accuracy of the nearest neighbour algorithm predictions. However, Skalak (1997) observed that most of the prototype selection algorithms either add an instance or remove it from the prototype set each time some triggering condition occurs. Although these algorithms change the cardinality of the set of prototypes, they do not investigate other prototype sets of that cardinality. This suggests that existing editing algorithms have an unfortunate search bias.

A variety of nearest neighbour methods have been proposed in the literature over the past decades. A very well known algorithm is the LVQ algorithm (Kohonen et al. 1988; Kohonen 1989). This algorithm is discussed in the following Section.

3.4.5 Learning Vector Quantization

The LVQ algorithm assumes that a mixture of distributions rather than a single elliptical distribution can approximate the distribution of each class. Therefore, we can see this algorithm as a generalisation of the LDA algorithm that imposes two elliptical distributions around the means of the two groups. This idea is illustrated in Figure 3.2 for the case of two predictor variables: P/E and BE/ME. As we can see in Figure 3.2, the data points are centred on distinct clusters of observations corresponding to H and L performing shares. The points around the centre of each cluster have equal likelihood in that cluster. If we connect the points with equal distance from the neighbouring means, we have a non-linear partition of the space into H and L regions.

According to the LVQ algorithm, a finite number of prototypes are chosen in the input space. During the learning process each sample is compared to all prototypes and the nearest one is selected. If the class of the selected prototype is the same with the class of the input sample, then the selected prototype is moved in the direction of the input sample. If the class of the selected prototype is not the same with the class of the input sample, then the selected prototype is moved in the opposite direction. During the test phase, there is no modification of the location of the prototype selection. The class label of the selected prototype gives the class of the input

sample. This algorithm is known as LVQ1. Kohonen et al. (1995) proposed two other versions of this algorithm namely LVQ2 and LVQ3. All these algorithms are discussed in detail below.

i) LVQ1

Let p_i be a number of free parameter vectors that are placed in the input space to approximate various domains of the input vector x_{it} by their quantized values. According to the LVQ algorithm, the x_{it} is decided to belong to the same class to which the nearest p_i belongs. Let us denote the nearest p_i to x_{it} by p_n. This can be written as follows (Kohonen et al., 1995),

$$n = \arg\min_{it}\{\|x_{it} - p_i\|\} \tag{3.16}$$

The values for the p_i that minimise the misclassification errors in the above classification rule can be found as asymptotic values in the following learning process (Kohonen et al., 1995),

$$\begin{aligned}
p_{n(t+1)} &= p_{nt} + \alpha_t\left[x_{it} - p_{nt}\right] \quad \text{if } x_{it} \text{ and } p_n \text{ belong to the same class} \\
p_{n(t+1)} &= p_{nt} - \alpha_t\left[x_{it} - p_{nt}\right] \quad \text{if } x_{it} \text{ and } p_n \text{ belong to different classes} \\
p_{i(t+1)} &= p_{it} \quad \text{for} \quad i \neq n
\end{aligned} \tag{3.17}$$

where x_{it} is the input sample, p_{it} are sequences of the p_i in the discrete-time domain, and α_t is a parameter that may be a constant or decrease monotonically with time.

The algorithm described in Eq. (3.17) is known as LVQ1 algorithm. Kohonen proposed the optimised LVQ1 (oLVQ1) algorithm in a way that an individual learning rate α_{it} is assigned to each p_i Taking into account this modification, we can write Eq. (3.17) as follows (Kohonen et al., 1995),

$$p_{n(t+1)} = \left[1 - h_t\alpha_{nt}\right]p_{nt} + h_t\alpha_{nt}x_{it} \tag{3.18}$$

where $h_t = +1$ if the classification is correct, and $h_t = -1$ if the classification is incorrect.

According to Kohonen et al. (1995) the optimal value of α_{nt} is determined by the following recursion,

$$\alpha_{nt} = \frac{\alpha_{n(t-1)}}{1 + h_t \alpha_{n(t-1)}} \tag{3.19}$$

Although the LVQ1 algorithm performs both a quantization and a classification task, no one of them is optimal. Specifically, the vector quantization is better if the classification task is not achieved and the boundaries between classes do not approximate the optimal Bayes boundary. In order to overcome this problem, Kohonen et al. (1995) suggested two improved versions of the LVQ1 algorithm named LVQ2 and LVQ3. These versions are described below.

ii) LVQ2

This algorithm is an improved version of the LVQ1 algorithm. According to this algorithm two free parameter vectors p_i and p_j that are the nearest neighbours to x_{it} are updated simultaneously. One of them is expected to belong to the correct class and the other to the wrong class. The LVQ2 algorithm can be written as follows (Kohonen et al., 1995),

$$p_{i(t+1)} = p_{it} - \alpha_t \left[x_{it} - p_{it} \right]$$
$$p_{j(t+1)} = p_{jt} + \alpha_t \left[x_{it} - p_{jt} \right] \tag{3.20}$$

where p_i and p_j are the two closest free parameter vectors to x_{it} so that x_{it} and p_j belong to the same class, while x_{it} and p_i belong to different classes. Furthermore, x_{it} must fall into a zone of values called window that is defined around the midplane of p_i and p_j. Thus, if we assume that d_i and d_j are the Euclidean distances of x_{it} from p_i and p_j, then x_{it} is defined to fall in a window if $\min\left(\frac{d_i}{d_j}, \frac{d_j}{d_i}\right) > h$ where $h = \frac{1-w}{1+w}$. The w represents the window width. A window width of 0.2 to 0.3 is recommended (Kohonen et al., 1995),

Although the LVQ2 algorithm is based on the idea of differentially shifting the decision borders towards the Bayes limits, no attention is paid to what is going to happen to the location of the p_i in the long run if this process is continued. This problem is addressed in the LVQ3 algorithm that is presented below.

iii) LVQ3

The LVQ3 algorithm can be represented mathematically as follows (Kohonen et al., 1995),

$$p_{i(t+1)} = p_{it} - \alpha_t \left[x_{it} - p_{it} \right]$$
$$p_{j(t+1)} = p_{jt} + \alpha_t \left[x_{it} - p_{jt} \right] \tag{3.21}$$

where p_i and p_j are the two closest free parameter vectors to x_{it} so that x_{it} and p_j belong to the same class, while x_{it} and p_i belong to different classes. Furthermore, x_{it} must fall into the following window (Kohonen et al., 1995),

$$p_{k(t+1)} = p_{kt} + \phi\alpha_t \left[x_{it} - p_{kt} \right] \tag{3.22}$$

for $k \in (i, j)$ if x_{it}, p_i, and p_j belong to the same class.

According to Kohonen et al. (1995) the optimal value of ϕ seems to depend on the size of the window. After a series of experiments, he found that a range of values of ϕ between 0.1 and 0.5 are more applicable.

iv) General Considerations

Kohonen et al. (1995) suggested that the accuracy of the LVQ algorithms in various classification tasks depends on two factors: first, an approximately optimal number of free parameter vectors assigned to each class and their initial values; and second, a proper learning rate and a proper criterion for stopping the learning. They observed that an upper limit to the total number of free parameter vectors is set by the restricted recognition time and the computer power available. In order to determine the initial values of the free parameter vectors, Kohonen et al. (1995) suggested to use samples of the real training data picked up from the respective classes and accept the samples that are not misclassified. According to this procedure, we classify a sample against all the other samples in the training set using a nearest neighbour algorithm and we accept it as a possible initial value only if this tentative classification is the same as the class of the sample. In the next step, we compute the medians of the shortest distances between the initial free parameter vectors of each class. If these distances are very different for the different classes, then we can add new free parameter vectors or delete old ones from the deviating classes. This training cycle that is based on the (oLVQ1) algorithm is run once and the whole procedure can be repeated a certain number of times. Kohonen et al. (1995) recommended that the medians of the shortest distances between the free parameter vectors should be smaller than the standard deviations of the input samples in all the respective classes.

As far as concerns learning, Kohonen et al. (1995) suggested to start with the oLVQ1 algorithm that has fast convergence. They observed that its asymptotic recognition accuracy will be achieved after a number of learning steps that is about 30 to 50 times the total number of free parameter vectors. However, if the initial learning period is included in the initialisation of the free parameter vectors, Kohonen et al. recommended that the oLVQ1 algorithm can be continued from those free parameter vector values that have been obtained in the initialisation phase.

3.4.6 Advantages and Disadvantages of Non-parametric Classification Rules

Non-parametric classification rules do not necessarily postulate models for the population-conditional distributions. They suggest instead to estimate first the probability density functions from the training data, and then apply the estimated functions to the individuals for classification. Therefore, non-parametric classification rules are not expected to be fairly robust to certain assumptions about the assumed model. However, most of these algorithms lack statistical simplicity and interpretation and require heavy computational resources to deal with the excessive training times. Furthermore, some of these algorithms suffer from the curse of dimensionality. For example, kernel and nearest neighbour methods estimate probabilities at a point x_{it} using the classes of neighbouring training set elements. Hand (1997) observed that bias is likely to be introduced by these methods if the probabilities at neighbouring training set elements are not the same with the estimated probabilities at x_{it}. High dimensional spaces can have a devastating effect where probability distributions are concerned. Hand (1997) observed that with one-dimensional normal distribution almost ninety percent of the probability will lie approximately within ± 1.6 standard deviations from the mean, whereas with a ten-dimensional spherical multivariate normal distribution only one percent of this probability will lie closer than 1.6 standard deviations from the mean. The mean of such distribution will be sparsely populated with sample points and most observations will lie far from the origin. To deal this problem a very large sample is required.

Part Two: Computer-Intensive Classification Techniques

3.5 NEURAL NETWORKS

3.5.1 Artificial Neural Networks

Neural networks have been used extensively in classification tasks. The first neural network architecture that was applied in classification tasks was the original perceptron that proposed by

Rosenblatt (1962). He showed that if two data sets are separated by a hyperplane, then a perceptron model would find a plane to separate them. Minsky and Papert (1969) demonstrated, however, that this algorithm would not converge if there was no separating hyperplane.

Multilayer perceptrons (MLP) have been applied successfully to a variety of problems by training them in supervised manner using a highly popular algorithm known as backpropagation. This algorithm consists of two phases: a forward pass and a backward pass. In the forward pass, an input vector propagates through the network layer by layer, and a set of outputs is applied as the actual response of the network. During this pass, all the weights of the network are fixed. During the backward pass the weights of the network are adjusted according with an error correction rule. The actual response of the network is subtracted from a target response to produce an error signal. This signal is then propagated backward through the network against the direction of synaptic connections. The synaptic weights are then adjusted so as to make the actual response of the network move closer to the desired response (Haykin, 1994).

A multilayer feedforward network consists of a set of neurons that are logically arranged into two or more layers. There is an input layer, an output layer, and one or more hidden layers between the input and output layers (Masters, 1995). The computational nodes of these layers are correspondingly called input neurons, hidden neurons, and output neurons. The source neurons in the input layer supply signals that are used as inputs to the second layer. The output signals of the second layer are used as inputs to the third layer and so on for the rest of the network. Information flows in one direction only (Haykin, 1994).

Figure 3.3 illustrates the layout of a three-layer feedforward neural network. The first layer on the left is the input layer, the second layer in the middle is the hidden layer, and the third layer on the right is the output layer . The neural network of Figure 3.3 is fully connected in the sense that every neuron in each layer is connected to every neuron in the adjacent forward layer. The function of the hidden neurons is to intervene between the external input and the network output. By adding one or more hidden layers, the network is able to extract higher-order statistics. This is particularly valuable when the size of the input layer is large (Haykin, 1994).

Classification of an individual with feature vector $x'_{it} = (x_{i1t}, x_{i2t} \ldots x_{imt})$ into one of the two classes H, L can be viewed as a mathematical process of transforming the m input units into two output units y_H, y_L that define the class allocation of the unknown vector. For example, $y_H = 1$ and $y_L = 0$ if the individual is to be allocated to class H or class L, respectively. The

MLP caries-out the transformation by treating the x_{it} values as m units in the input layer, the $y_j(j = H, L)$ as values of two units in the output layer, and using a number of hidden layers between these two layers (Krzanowski and Marriott 1995). A few excellent reviews on neural networks are provided by Ripley (1992), Cheng and Titterington (1994), Zhang et al. (1998), and Dunis and Jalilov (2001).

3.5.2 Probabilistic Neural Network

The probabilistic neural network (PNN) algorithm is a well-known representative of the family of non-linear models. This algorithm constructs estimates of the Probability Density Functions (PDF) for each class, $f^j(x_{it})(j = H, L)$, and then allocates a share with measurement vector x_{it} to the class with the largest $f^j(x_{it})$. Considering the prior probabilities and misclassification costs, we classify a particular share with measurement vector x_{it} as H if $p_H c_H f^H(x_{it}) > p_L c_L f^L(x_{it})$ where $p_H(p_L)$ is the prior probability of membership in the class H (L), $c_H(c_L)$ is the cost of misclassification into class H (L), and $f^H(x_{it})[f^L(x_{it})]$ is the Probability Density Function (PDF) of class H (L), respectively. The key factor in implementing the PNN algorithm is to construct the PDF for each class. Although we do not know the true PDF for each sample, we can estimate it from the sample data. Parzen (1962) developed a non-parametric technique for estimating a univariate PDF from a random sample, and Cacoulos (1966) extended Parzen's method to the multivariate case.

The PNN algorithm starts by positioning a separate distribution over each individual data point. Therefore, we can see this algorithm as generalisation of the LDA algorithm that imposes two elliptical distributions around the means of the two groups.

Let us assume that we have to assign a new observation x_{it} to one of the classes H or L. Our aim is to construct an estimate $\tilde{f}^j(x_{it})[j = H, L]$ for each class. The distance of x_{it} from classes H, L is the average of distances from all individual members x_{jt-i} that belong to classes H, L respectively. For example, if we assume that m_H is the number of members in class H, then the average of distances of x_{it} from all x_{jt-i} can be represented as a superposition of potential functions $\phi(x_{it}, x_{jt-i})$ over samples of $\tilde{f}^H(x_{it})$ as follows,

$$\tilde{f}^H(x_{it}) = \frac{1}{m_H}\sum_{j=1}^{m_H}\phi(x_{it}, x_{jt-i}) \qquad (3.23)$$

Although there is considerable freedom in choosing the potential function, we should notice that some functions might be more preferable than others. For example, a function that computes an equally weighted average sum of distances of x_{it} from all individual members x_{jt-i} that belong to class H, L respectively, might not be appropriate because it does not consider the decreasing influence of x_{it} upon a point x_{jt-i} as the distance $d(x_{it}, x_{jt-i})$ between the points increases. Using a Gaussian curve and let the size of the weight to decay as the distance $d(x_{it}, x_{jt-i})$ between the points increases will have more desirable properties. Using the Gaussian function, we can represent the density function for class H as follows,

$$\tilde{f}^{H}(x_{it}) = \frac{1}{\sqrt{2\pi}\sigma m_H} \sum_{j=1}^{m_H} \exp\left(-\frac{\left\| x_{it} - x_{jt-i} \right\|^2}{2\sigma^2} \right) \tag{3.24}$$

As we can see in Eq (3.24), the potential function shows the decreasing influence of a sample point x_{it} upon a point x_{jt-i} as the distance $d(x_{it}, x_{jt-i})$ between the points increases. The average of these potentials from samples of class H at a point x_{it} constitutes a measure of the degree of membership of point x_{it} in class H. The scaling parameter, σ, defines the width of the area of influence and should decrease as the sample size increases. The value of σ can be crucial for the positioning of a mixture of identical Gaussian distributions at each of the training sample points. If the value of σ is too small, then individual training cases may exert too much influence thus eliminating the gain of aggregate information. If the value of σ is too large, then the excess blurring may distort the density estimate because important details of the density are lost.

The original PNN architecture was proposed by Specht (1990). Specht showed that the algorithm could be split up into a large number of simple processes each of which has its own delicated task and most of which could run in parallel. A graphical illustration of the PNN architecture is given in Figure 3.4. As we can see in Figure 3.4, the above network consists of four layers: an input layer, a pattern layer, a summation layer, and an output layer. The input layer has as many neurons as the inputs. The pattern layer has one neuron for each training case. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects that measure to the neuron's activation function. The summation layer has one neuron for each class. Each summation neuron sums the pattern layer neurons corresponding to numbers of that summation neuron's class and estimates the density functions for classes H and L, respectively. The output neuron is a threshold

discriminator that decides which of its inputs from the summation units is the maximum (Masters, 1995).

The PNN has been proposed in the literature as an alternative technique to LDA for a variety of real world applications including financial distress (Tyree and Long 1996; Yang 1999), bond ratings (Albanis 1998a, 1998b), and stock selection (Albanis and Batchelor, 2000a). As opposed to LDA, the PNN does not suffer from departures from multivariate normality and equality of the group covariance matrices. However, the PNN has some other drawbacks: first, the performance of the network depends on having a thorough and representative training set due to the way it uses the potential functions around clusters of the training data; second, the entire training set must be stored each time an unknown case is classified (Masters, 1995). This results in large memory requirements and considerably slows the execution speed; and third, the success of the PNN paradigm is at a cost: an inherent inability to explain in a comprehensible form the process by which a given decision or output generated by the model has been reached.

3.5.3 Radial Basis Function Network

Basis function networks have the form depicted in Figure 3.5. As we can see in Figure 3.5, the input vector x_{it}, which consists of $x_{1t}, x_{2t}, \ldots, x_{mt}$ attributes, is passed to a set of basis functions. Each basis function returns a scalar value, $g_l, l = 1,2...k$. Typically, the basis functions are taken to be Gaussian - hence the name Radial Basis Function (RBF) networks. The motivation for using RBF networks is to transform the attribute space in a manner that increases the separability of different classes.

Let us assume that the input vectors have been normalised and that the standard deviation along each attribute is σ. The l-dimensional Gaussian surface which is centred at $\mu_{lt} = [\mu_{1lt} \cdots \mu_{lmt}]$ can be written as follows (Chen et al. 1991; Fritzke 1994; Kubat 1998),

$$g_l(x_{it}) = exp\left(-\frac{\|x_{it} - \mu_{lt}\|^2}{2\sigma^2}\right) \tag{3.25}$$

The activation function of the hidden neurons is radially symmetric in the input space. Eq. (3.25) suggests that the larger the distance between x_{it} and μ_{lt}, the smaller the value of $g_l(x_{it})$. More specifically, the magnitude of the activation given a particular datapoint is a decreasing function of the distance between the input vector of that datapoint and the centre of the basis function. The activation of the j^{th} output neuron is then calculated as

$y_{jt} = \sum_{l=0}^{k} w_{jl} g_l(x_{it})$ where w_{jl} is the weight of the link from the l^{th} hidden neuron to the j^{th} output neuron. The input vector is assigned to the H class at time t if $y_{Ht} = max_{\pi}(y_{\pi})$ (Kubat 1998).

The role of the hidden units is to perform a non-linear transformation of the input space into the space of activations of the hidden units. This transformation gives the RBF a much greater representational power than the linear perceptron. The output layer computes a linear combination of the activations of the basis functions, parameterised by the weights of the links between hidden and output layers.

However, the RBF networks have several drawbacks (Fritzke 1994; Kubat 1998):

- Most of the techniques that have been applied to identify the Gaussian centres are computationally expensive. For example, one technique to identify each Gaussian centre might be to use the coordinates of one input vector. However, if there is a large number of input vectors, the network might become extremely large and it is likely to overfit the data. On the other hand, using a random subset of the training data to identify the Gaussian centres might not be an ideal solution if the data is noisy.
- The ideal setting of the standard deviation, σ, might be different for each dimension if the different attributes have different scales. One solution might be to use a value propositional to the maximum distance between centres. Alternatively, a different σ for each hidden neuron calculated as a function of the distance between its centre and the centre of its nearest neighbour might be used. Although several heuristic procedures can be applied to find the proper setting of the parameter, σ, this might be a very expensive task in terms of computational resources.
- RBF networks might be particularly sensitive to irrelevant attributes that can distort similarity metrics as the one presented in Eq. (3.25). In this sense, RBF networks share some similar characteristics with nearest neighbour classifiers.
- Finally, if the number of neurons is large in the input and hidden layers, then hardware constraints might be important for the RBF due to the large number of links leading from the input to the hidden layer.

Considering the above limitations in the design and implementation of RBF networks, we decided not to include this particular model in our trading system. Furthermore, the lack of

source code to implement this algorithm was another constraint for not attempting this algorithm for our particular application.

3.5.4 DIPOL92

DIPOL92 is a hybrid algorithm that can be viewed as combination of statistical regression with a learning procedure quite similar to the one that is used for neural networks. DIPOL92 constructs an optimised piecewise linear classifier by using pairwise linear regression to position the discriminating hyperplanes. An error criterion function is used to optimise the initial positions of the hyperplanes in relation to the misclassified patterns. To minimise this function, a gradient descent procedure is applied for each individual hyperplane. If the classes have multimodal probability distributions, a clustering procedure is applied to decompose the classes into relevant subclasses. Let us assume that $x_{it} = (x_{i1t}, x_{i2t}, \ldots, x_{imt})$ represents a set of data such that $X \subset \Re^m$. DIPOL92 uses linear regression to discriminate two classes H and L by defining the dependent variable y as follows (Michie et al. 1994),

$$\begin{cases} \text{if} \quad x_{it} \in H \quad \text{then} \quad y = +1 \\ \text{if} \quad x_{it} \in L \quad \text{then} \quad y = -1 \end{cases} \tag{3.26}$$

Now, let as assume that $f(x_{it})$ is a linear regression function such that $f(x_{it}) = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_{im} x_{im}$ and $f: X \rightarrow \Re$. A pattern x_{it} is correctly classified to class H if $f(x_{it}) > 0$ and to class L if $f(x_{it}) < 0$. A discriminating regression function can be calculated for each pair of classes. A detailed analysis of the classification procedure can be found in Michie et al. (1994).

The advantage of DIPOL92 over typical neural network algorithms is its ability to determine the initial positions of the hyperplanes before learning starts. Another advantage of DIPOL92 over neural networks is that the problem of additional hidden layers is avoided by using Boolean variables for the description of class regions on a symbolic level and then using them in the classification procedure (Michie et al. 1994).

3.5.5 Advantages and Disadvantages of Neural Networks

Neural networks are likely to be most superior to other methods if the data exhibits unpredictable non-linearity that cannot be detected by linear models or other models that are based on strictly defined formulations. One of the main advantages of neural networks is their ability to discover patterns in the data that cannot be detected by standard statistical methods

because they detect patterns in a manner analogous to human thinking. If the data incorporate human judgement and other qualitative factors, then the robust performance of neural networks will be very important. Furthermore, neural networks are likely to have better generalisation performance than other statistical methods if the data is fuzzy, chaotic, or it is subject to possibly large error. This excellent performance of artificial neural networks should not be considered surprising if we consider that these algorithms have been built on strong theoretical foundations. The three-layer feedforward neural network has powerful approximation properties. Hornik et al. (1989) have proven theoretically that any continuous function that is defined over a constant subset of $\Re^m$ can be approximated for arbitrary accuracy given a sufficient number of neurons in the hidden layer. Specht (1990) has proven that under very general conditions the ability of PNN is asymptotically Bayes optimal (Masters, 1993). However, neural networks have several drawbacks. One main drawback is that the learning process is very slow. Another drawback is that the knowledge generated by neural networks is not implicitly represented in the form of rules or conceptual patterns but implicitly in the network itself as a vast number of weights. Therefore, the success of the neural network paradigm is at a cost: an inherent inability to explain in a comprehensible form the process by which a given decision or output generated by a neural network has been reached. In addition, neural networks might suffer from the problem of overfitting if the data set is small compared to the free parameters of the network.

3.6 DATA MINING SYSTEMS

3.6.1 Theoretical Background

Cognitive systems attempt to understand their environment by creating a simplified representation of the environment called a model. The creation of such a model is known as inductive learning. During the learning process, the cognitive system observes the environment in order to find similarities among objects that are part of the environment. After this process is complete, similar objects are grouped into classes and rules are constructed to predict future members of each separate class. The computer modelling of the inductive learning process has been the subject of a research area known as machine learning. A machine learning system uses a finite set of observations which is known as training set and represents information about the environment. However, if the training set is a database then the learning process is called data mining. The database is a large volume of data that have been stored for purposes other than learning processes. More often, these data are noisy and values of attributes are missing. The size of the database makes verification of certain hypothesis an extremely costly process. To

deal with these problems, statistical techniques are used to assist the user in the generation of certain hypotheses (Holsheimer and Siebes, 1991).

Let us assume that E represents the environment. In addition, let $X_t = x_{i1t}, x_{i2t}, \ldots, x_{imt}$ be a set of attributes with domains $D_t = d_{i1t}, d_{i2t}, \ldots, d_{imt}$. A training set T_t is a table over X_t, whereas an example is a tuple in the training set. The tuples in the database represent properties of the objects and not relationships of the objects. Therefore, the data mine system has to infer the rules that govern the classification of database objects. In the supervised learning, this requires that the user has to define one or more classes. A class $C_{jt}(j = H, L)$ is a subset of the training set T_t consisting of all objects that satisfy the class condition $Cond_{jt}$ as follows (Holsheimer and Siebes, 1991),

$$C_{jt} = \left\{(obj_{it}) \in T_t \setminus Cond_{jt}(obj_{it})\right\} \tag{3.27}$$

Objects that satisfy the class condition $Cond_{jt}$ are positive examples of the class j, whereas objects that do not satisfy the class condition $Cond_{jt}$ are negative examples of class j.

The database contains one predicted attribute denoting the class of existing examples, whereas the remaining attributes are the predicting attributes associated with each class. The data mine system has to infer classification rules that predict the class of new examples from the predicting attributes. In other words, the system has to find the description of each class. These descriptions consist of conditions on the attributes. A description is non-empty disjunction of elementary conditions such that $X_{1t} = k_1 \wedge \ldots \wedge X_{mt} = k_n$ under the assumptions that $X_{it} \in X_t$ $\left(i \neq j \rightarrow X_{it} \neq X_{jt}\right)$ and $k_i \in d(X_{it})$. The set of possible descriptions forms the description space, Des (Holsheimer and Siebes, 1991).

All the examples that satisfy the description Des, are said to be covered by Des. This can be denoted as $\sigma_{Des}(T)$. A classification rule consists of a description and a class symbol C_{jt} and can be denoted as follows (Holsheimer and Siebes, 1991),

$$\forall(obj_{it}) \in \Omega_t : (obj_{it}) \in \sigma_{Des}(T_t) \rightarrow (obj_{it}) \in C_{jt} \tag{3.28}$$

where Ω_t represents the entire universe so that $\Omega_t = d_{i1t} \times d_{i2t} \times,\ldots,\times d_{imt}$ and it is a full relation over X_t. The expression (3.28) represents a classification rule and states that any object that satisfies Des belong to class $C_{jt}(j = H, L)$. A classification rule is said to be correct with respect to the training set T_t if the description of this rule covers each positive example and none of the negative examples. In other words, $\sigma_{Des}(T_t) = C_{jt}$.

Several data mine systems have been proposed in the literature. These include among others the ID3 system, the C4.5 system, the AQ15 system (Michalski et al. 1986), the CN2 system (Clark and Niblett 1989), the DBLearn system (Han et al. 1992) etc. Some of these systems such as the ID3 and C4.5 generate decision trees, whereas others such as the AQ15 and CN2 induce classification rules. Decision trees and rule induction algorithms are discussed in more detail in the next Sections.

3.7 RECURSIVE PARTITIONING CLASSIFICATION METHODS

3.7.1 Decision Trees

Decision trees are hierarchical classification structures that recursively partition a set of observations. They are a particularly useful tool because they perform classification by a sequence of simple tests whose semantics are easy to understand by domain experts. Decision trees are often described as non-parametric since they do not assume any underlying family of probability distributions. They represent a way to describe rules underlying the data.

Most of the existing decision tree algorithms use a greedy top-down approach to build a decision tree. This approach can be described as follows: if all training examples at the current node h belong to class $C_{jt}(j = H, L)$, then we create a leaf node with the class C_{jt} and halt. Otherwise, we score each one of the set of possible splits, S, using a goodness measure. We then choose the best split s' as the test at the current node and create as many leaf nodes as there distinct outcomes of s'. We then label edges between the current and leaf nodes with outcomes of s' and we partition the training data using s' into the leaf nodes.

A new share with measurement vector x_{it} is classified by passing it through the tree starting at the root node. The test at each internal node along the path is applied to the attributes of x_{it} to determine the next edge along with this example should be down. The label at the leaf node at which x_{it} ends up is outputted as its classification. The tree misclassifies the new share if the predicted classification is not the same as the share's class label.

Figure 3.6 gives a graphical illustration of a decision tree that classifies shares into H and L classes for the case of two predictor variables: P/E and BE/ME. A hypothetical sequential representation of this tree is given in Figure 3.7. To classify a new share at the top or root of the tree, we first test whether the P/E ratio is greater than 0.2. If the P/E is not greater than 0.2, the share is H. If the P/E ratio is greater than 0.2, then we test if the BE/ME is less than 0.7. If the BE/ME is greater than 0.7 the share is H. If it is less than 0.7, then we test if the linear combination BE/ME-3(P/E) is less than 0.2 and so on. Following this path, we work down the tree, recursively testing conditions of the predictor variables at the nodes of the tree, and decide which path to follow depending on whether or not the conditions of the predictor variables are satisfied. As we can see in the tree structure presented in Figure 3.7, each non-terminal node is associated with a single variable and a partition of that variable into two classes determines which path a new example should follow.

When building a tree, we have to decide the partitioning criteria that determine which variable should we use at each internal node, which internal node should be split, and what would be the nature of the split. The usual approach to solve these problems is to use an "impurity" index at each node of the tree. The "impurity" index is a measure of the differences between the probabilities of belonging to each class. For example, let us assume that there are 20 shares with measurement vectors x_{it} (i=1,2...20) known at time t at an internal node and that 10 of these vectors are from class H, whereas the other 10 vectors are from class L. Furthermore, let us assume that all vectors from class H have P/E values less than 0.2, whereas all vectors from class L have P/E values greater than 0.2. This node is relatively impure because it has equal numbers from each of the two classes. However, if we decide to split this node and build the tree further by splitting this node using P/E at the value 0.2, then we have two offspring nodes that are perfectly pure since one would have only class H vectors, whereas the other would have only class L vectors (Hand, 1997).

Several impurity measures have been suggested in the literature. These include among others the Information Index, the Gini Index, Max Minority, Sum Minority etc. A very effective impurity index is also the Twoing Rule that was proposed by Breiman et al. (1984). This impurity index is given as follows (Murthy, 1997),

$$TV = \left(\frac{|N_L|}{n}\right) * \left(\frac{|N_R|}{n}\right) * \left(\sum_{i=1}^{k} \left|\frac{CL_L}{|N_L|} - \frac{CR_H}{|N_R|}\right|\right)^2 \qquad (3.29)$$

where TV is the Twoing value, $|N_L|$ is the number of vectors on the left of a split at node v, $|N_R|$ is the number of vectors on the right of a split at node v, n is the total number of vectors at node v, CL_L is the number of vectors in category L on the left of the split, and CR_H is the number of vectors in category H on the right of the split.

We could continue to partition the nodes of a decision tree until all the leaf nodes contained only a single class H or class L vector x_{it}. However, this procedure would lead to a very large tree with many leaf nodes. It is likely that such a tree would overfit the training data and would have poor generalisation performance to new data. It would simply model random variation in the training set rather than modelling the true underlying structure of the data. One way to solve this problem would be to stop the growth of the decision tree by adopting a stopping rule. A possible stopping rule would be to stop the growth of the tree when the maximum reduction in impurity is less than some threshold. Obviously, a small threshold would lead to many small leaf nodes, whereas a large threshold would lead to few leaf nodes. However, this approach has a disadvantage if the growth of the tree is sequential. In that case, a split is possible at a node of the tree only if the reduction in impurity exceeds some value, independently on what is going to happen at a lower node of the tree as a result of the split. An alternative stopping rule to avoid this problem would be to build a large tree that overfits the data, and then to prune this tree using some pruning criterion (Hand, 1997).

A commonly used criterion in order to decide which leaves of the tree to prune is $\max_j \tilde{p}(j/h)\phi(h)$. This criterion measures the value of each leaf node (Hand, 1997). The first factor, $\tilde{p}(j/h)$, serves as an estimate of the probability that points belonging to class $C_{jt}(j = H, L)$ and falling at node h are correctly classified at class j, whereas the second factor, $\phi(h)$, serves as weight indicating the proportion of objects from the training set that fall in terminal node h. Objects arriving in this node are assigned to the class maximising $\tilde{p}(j/h)$. If the value of the above criterion is small, it means that either the node is small or that the probability estimates at that node are not particularly in favour of the largest class. If the value of the criterion is large, it means that it correctly classifies most of a large number of objects that arrive at it (Hand, 1997).

Rather than trying to build a tree in order to optimise classification performance, we should also consider that the size of the tree may be of critical importance. Given that all other things being equal, simpler trees are to be preferred to large trees since they minimise the possibility of

overfitting. If we take into account the size of the tree, we can then represent the overall quality of the tree as $\sum_{h\in t}\left[1-\max_j \tilde{p}(j/h)\right]\phi(h)+\alpha|N|$, where $|N|$ is the number of terminal nodes. The parameter α determines the payoff between predictive accuracy on the training set and the complexity of the tree. Varying α and choosing the tree that minimises this criterion, a number of best trees is produced. The selection of the final tree can be based on an error rate estimate after using a test set or after applying cross-validation procedures. These cross-validation procedures are discussed in Section 3.9 (Hand, 1997).

3.7.2 Variants of Decision Trees

A very popular decision tree algorithm is the CART algorithm that was developed by Breiman et al. (1984). This algorithm is based on the idea of Kendall et al. (1983). This idea suggests to examine the data on each variable in turn and try to separate the training data into their corresponding classes by successive partitioning of the variables. This is achieved by identifying the variable, x_{i_1t}, and two associated constants d_1, d_2 such that all the training data with $x_{i_1t} < d_1$ fall in one class, all the training data with $x_{i_1t} > d_2$ fall in the other class, and the smallest possible number of training elements fall in the region $d_1 \le x_{i_1t} \le d_2$. The elements that fall into the latter region are then subjected to a similar search process using the remaining variables and this procedure is continued until no further classification is possible (Krzanowski and Marriott, 1995). CART automates this procedure by using binary splits of the variables without to incorporate the region $d_1 \le x_{i_1t} \le d_2$. A binary decision tree is constructed from the training data in a forward/backward stepwise manner and optimal splits are determined by minimising an index of impurity of classification. Brieman et al. (1984) used the Gini index to minimise the impurity index. The CART uses the minimal cost complexity cost pruning technique to prune the decision tree. This technique treats pruning as a tradeoff between getting the right size of the tree and getting accurate estimates of the probabilities of misclassification.

One extension of CART is the IndCART algorithm. The IndCART algorithm differs from CART in using a different way to handle missing values as well as in using different pruning techniques. Furthermore, several researches modified the CART algorithm by adopting a variety of splitting rules and averaging techniques. A detailed review of related work can be found in Krzanowski and Marriott (1995).

A system that had a very important impact on machine learning research in recent years is the C4.5 algorithm. The C4.5 algorithm is actually an extension of the ID3 algorithm that was developed by Quinlan (1983, 1986). ID3 is a supervised learning system that constructs decision trees from a set of pre-defined examples. The search space for this particular algorithm

consists of all possible trees that can be constructed with attributes and values in the test set. Among all trees in the search space, the ID3 system attempts to find the best quality tree. The two criteria that determine the quality of the tree are classification performance in the test set and simplicity. For example, between two trees that classify examples in the test set correctly, the algorithm will prefer the less complex tree against the more complex tree.

The ID3 system constructs a decision tree from a set of examples where the domain of each attribute in these examples represents a small number of either symbolic or attribute values. The process of constructing a tree begins by selecting an attribute as the root of the tree and forming branches that correspond to different values of the selected attribute. This tree is then used to classify the training set. If the examples at a particular leaf node of the tree belong to the same class, the leaf node is labelled with this class. However, if a leaf node of the tree is not labelled with a class, then the node is labelled with an attribute that does occur on the path of the tree and new branches are created for possible values of the new attribute. The new tree is then used to classify the training set and the same process is repeated until all leaves are labelled with a class. The condition on an attribute at an internal node of the tree is a test on the value of the attribute with branches for all possible values of this attribute. This process is more appropriate with symbolic attributes but it is not convenient if the test is based on a range of numerical attributes. This problem was addressed on the C4.5 algorithm. The C4.5 algorithm allows to test an inequality of numerical attributes such that $x_{gt} \leq q$ with two corresponding possible branches. Furthermore, the C4.5 algorithm allows to test whether the value of an attribute belongs to a particular set of values so that the node is labelled with the attribute and the branches are labelled with sets of values (Holsheimer and Siebes, 1991).

The information gain of such a test can be computed by sorting the examples on the values of the attribute being considered. If there is only a finite number of attributes, for example $x_{1t}, x_{2t}, \ldots, x_{mt}$, then the algorithm will search for $m-1$ possible splits on this particular attribute. It is obvious that this search might be a particularly expensive task in terms of computational resources if the examples are not sorted. However, if the examples are sorted, the search process can be performed in one pass updating the class distributions to the left and right of the threshold on the fly. For each possible threshold, an information gain can be computed and used in the process of selecting the next test (Holsheimer and Siebes, 1991).

Another important feature of the C4.5 algorithm is to test whether the value of a particular attribute belongs to a particular set of values. To perform this test, the node is labelled with the attribute, whereas the branches are labelled with sets of values. Assuming that there are m

different values, there are $2^{m-1}-1$ binary partitions. This reduces substantially the possibility of a very exhaustive search for the best partitioning of the attribute space. On the other hand, this process may actually reduce the complexity of the tree if the sets of values are related to each other. After dividing the individual values of the attribute under consideration into groups, the C4.5 algorithm uses an irrevocable bottom up search for merging the initial groups. At each separate cycle, C4.5 evaluates the consequences of merging every pair of groups. This process terminates when just two value groups remain or when the gain cannot be improved any further after merging the groups (Holsheimer and Siebes, 1991).

According to Quinlan (1996a), the process of generating a decision tree by applying the C4.5 algorithm to a set T of cases can be described as follows: If T satisfies a stopping criterion, the tree for T will be used as a leaf associated with the most frequent class in T. In the next step, a test K with mutually exclusive outcomes $K_1, K_2 ... K_L$ can be used to partition T into subsets $T_1, T_2, ..., T_L$ where T_i will contain only those cases that have outcome K_i. As a result of this process, the tree for T has test K at its root with one subtree for each outcome K_i that is constructed by applying the same procedure recursively to the cases in T_i. The default tests used by C4.5 are $x_{gt} = ?$ for a discrete attribute x_{gt} with one outcome for each value of x_{gt}, and $x_{gt} \leq q$ for a continuous attribute x_{gt} with two outcomes, true and false. To find the threshold q that maximises the splitting criterion, Quinlan suggests to sort the cases in T on the values of attribute x_{gt} to give ordered distinct values $v_1, v_2, ..., v_N$. For each pair of values, there is a potential threshold $q = (v_i + v_{i+1})/2$ and a corresponding partition of T. The values of the thresholds are then compared, and the threshold that yields the best value of the splitting criterion is finally selected.

The default splitting criterion used by C4.5 is an information-based measure known as the gain ratio. This measure takes into account different probabilities of test outcomes. According to Quinlan, the residual uncertainty about the class to which the case in T belongs can be expressed as follows (Quinlan, 1996a),

$$I(T) = \sum_{j=1}^{C} p(T,j) \times log_2[p(T,j)] \tag{3.30}$$

where C denotes the number of classes and $p(T,j)$ represents the proportion of cases in T that belong to the j^{th} class. Similarly, the information gained by a test K with q outcomes can be expressed as follows (Quinlan, 1996a),

$$G(T,K) = I(T) - \sum_{i=1}^{q} \frac{|T_i|}{|T|} \times I(T_i)$$

(3.31)

The information gained by a test is strongly affected by the number of outcomes and is maximal when there is one case in each subset T_i. The potential information obtained by partitioning a set of cases is based on knowing the subset T_i. This split information can be written as shown below (Quinlan, 1996a),

$$S(T,K) = -\sum_{i=1}^{q} \frac{|T_i|}{|T|} \times log_2\left(\frac{|T_i|}{|T|}\right)$$

(3.32)

The split information $S(T,K)$ tends to increase with the number of outcomes of a test. The gain ratio then assesses the desirability of a test as the ratio of its information gain to its split information. For every possible test, the gain ratio is determined and the split with maximum gain ratio is selected. However, there are situations where every possible test might split T into subsets that have the same class distribution. In those situations, all tests have zero gain, and C4.5 uses this condition as an additional stopping criterion.

The above strategy might result in trees that are consistent with the training data. However, if noise is present, then the system may lead to complex trees that are likely to overfit the data. Quinlan suggests that several pruning techniques can be applied to prune the initial tree and increase the predictive accuracy of the proposed algorithm.

Another decision tree algorithm is the Cal5 algorithm that was primarily designed to handle continuous and ordered discrete valued attributes. However, a later version of Cal5 was able to handle unordered discrete value attributes as well. Let us assume that X_t is a sample of data elements and each of these elements is expressed with m attributes. Cal5 separates the data elements into subsets X_t of samples such that $X_{it} \in X \ (i = 1,2...,m)$ and the class $C_{jt}(j = H,L)$ exists with probability $p(C_{jt}) > \alpha$ where $\alpha \leq 1$ is a decision threshold. The class areas are bounded by hyperplanes that are parallel to the axes of the feature space. A detailed derivation of this algorithm can be found in Michie et al. (1994).

3.7.3 Axis-parallel versus Oblique Splits

Many variants of decision trees have been proposed in the literature over the last decades. One well-known variant is axis parallel trees. Axis-parallel trees use tests at each internal node of the form $x_{it} > k$ where x_{it} is one of the attributes and k is a constant. These tests produce partitioning of the attribute space in the form of hyper-rectangles that are parallel to the feature axis. It is obvious, however, that if the underlying concept is defined by a polygonal partitioning of the attribute space, then axis-parallel trees would not be an ideal tool. In that case, we need more general trees to model the underlying concept. A good alternative to axis-parallel trees that has been proposed in the literature is oblique decision trees. Oblique decision trees use tests at each internal node of the form (Murthy, 1997),

$$\sum_{i=1}^{m} \alpha_{it} x_{it} + \alpha_{m+1} = 0 \qquad (3.33)$$

where m is the number of the attributes and α_{it} are real-valued coefficients. These tests are equivalent to hyperplanes at an oblique orientation of the axis.

Some of the existing oblique decision tree induction methods are CART with linear combinations (Breiman et al. 1984), Linear Machine Decision Trees (Utgoff, 1989), Simulated Annealing of Decision Trees (Heath et al., 1996), and Linear Programming based methods. The core idea of the CART algorithm with linear combinations is to find a value that maximises the goodness of the split. However, Murthy (1997) observed that this algorithm is fully deterministic since there is no built-in mechanism for escaping local minima. This algorithm makes adjustments that sometimes increase the impurity of the split. Because of this feature, there is no upper bound on the time spent at any node in the decision tree.

The Linear Machine Decision (LMDS) system uses a different approach from the CART algorithm. Each internal node in the tree is a Linear Machine. The training algorithm presents examples repeatedly at each node until the linear machine converges. However, since convergence cannot be guaranteed, the LMDS uses heuristics to determine when the node has stabilised. A thermal training method is used to make training stable even when the set of training instances is not linearly separable (Murthy, 1997).

Simulated Annealing Decision Trees (SADT) use simulated annealing to find good values for the coefficients of the hyperplane at each node of the tree. They first place a hyperplane in a canonical location, and then iteratively perturb all the coefficients by small random amounts.

Murthy (1997) observed that the use of randomisation allows SADT to avoid local minima, but it compromises on efficiency. In addition, it runs slower than either CART or LMDS.

An alternative way of finding splits is linear programming (LP). LP methods find a split by minimising the distance of misclassified points to the decision boundary. Murhty (1997) observed that even though this method is very competitive in terms of efficiency and effectiveness, it is unlikely to be robust to non-uniformly distributed noise in the data. In addition, most LP methods produce null/useless solutions when the two sets to be separated have the same centres. Finally, it has been claimed that LP-methods attempt to find a split that is good for the whole data set which may not exist. This may result in large sized trees.

Taking into account the above considerations, Murthy (1997) suggested the OC1 oblique classifier in an attempt to find locally good splits but not to spend excessive computational effort on improving the quality of these splits. This algorithm is described in more detail below.

3.7.4 Oblique Classifier (OC1)

This OC1 algorithm finds first the best axis-parallel split at a node before looking for an oblique split. It uses the oblique split only if it improves over the best axis-parallel split. The equation of the current hyperplane Y at a node of a decision tree is equivalent to Eq. (3.33). If we substitute a point T_{jt} from the training set into the equation for Y, we get (Murthy, 1997),

$$\sum_{i=1}^{m} \alpha_{it} x_{it} + \alpha_{m+1} = G_{jt} \tag{3.34}$$

The sign of G_{jt} tells us whether the point T_{jt} is above or below the hyperplane. If $G_{jt} > 0$ then T_{jt} is above Y. If Y splits the training set perfectly, then all points belonging to the same category will have the same sign for T_{jt}. OC1 adjusts the coefficients of Y by treating the coefficient α_{mt} as a variable and all the other coefficients as constants. Then, T_{jt} can be viewed as a function of α_{mt}. The condition that T_{jt} is above Y is equivalent to (Murthy, 1997),

$$G_{jt} > 0, \qquad \alpha_{mt} > \frac{\alpha_{mt} - G_{jt}}{x_{jm}} = F_{jt} \tag{3.35}$$

According to this definition for F_{jt}, the point T_{jt} is above Y if $\alpha_{mt} > F_{jt}$ and below otherwise. The problem with this approach, however, is to find a value for α_{mt} that satisfies as many of

the constraints as possible. Murthy suggested that this problem can be solved simply by sorting all the values F_{jt} and consider setting α_{mt} to the midpoint between each pair of different values. For each distinct placement of the coefficient α_{mt}, we compute the impurity of the resulting split. For example, if some of the values F_{jt} belonging to the same category are misclassified, then we have to find the best one-dimensional split of these values. This requires considering just $n-1$ values for α_{mt}. The value α'_{mt} obtained by solving this one-dimensional problem is then considered as a replacement for α_{mt}. Let Y_1 be the hyperplane obtained by perturbing α_{mt} to α'_{mt}. If Y has better impurity than Y_1, then Y_1 is discarded. If Y_1 has lower impurity, then Y_1 becomes the new location of the hyperplane. If Y and Y_1 have identical impurities, then Y_1 replaces Y with probability P_{stag}. This is the probability that a hyperplane is perturbed to a location that does not change the impurity measure. To prevent the impurity from remaining stagnant for a long time, Murthy suggested to decrease P_{stag} by a constant amount each time OC1 makes a stagnant perturbation. P_{stag} is reset to 1 every time the global impurity measure is improved.

The perturbation procedure halts when the split reaches a local minimum of the impurity measure. A local minimum occurs when no perturbation of any single coefficient of the current hyperplane will decrease the impurity measure. In order to escape local minima, Murthy suggested either to perturb the hyperplane with a random vector or to restart the perturbation algorithm with a different random initial hyperplane.

If the underlying problem requires an oblique split, then the oblique trees would be more accurate than the axis-parallel trees. In addition, building a decision tree that is based on both oblique and axis-parallel splits broadens the range of problems for which this tree would be useful. We should consider, however, that axis-parallel splits are simpler than oblique decision trees because the description of the split uses only one attribute at each node. Oblique decision tree algorithms use oblique splits only when their impurity is less than the impurity of the best axis-parallel split. Obviously, this increases the computational complexity compared to axis-parallel trees. As a result of these complexity considerations, the OC1 classifier attempts to generate small trees, but it is not looking for the smallest tree. Using the greedy approach to generate a decision tree, the OC1 classifier tries to generate small trees by finding locally good splits but not spending excessive computational effort to improve the quality of these splits (Murthy, 1997).

The OC1 classifier uses the Cost Complexity pruning method to guide pruning (Breiman et al., 1984). The idea behind this method is to create a set of trees of decreasing size compared with the original tree. All these trees are applied to classify a pruning set that is chosen randomly from the training data. The smallest tree that is chosen is the one whose accuracy is within k standard errors squared of the best accuracy obtained. If k=0, then the tree with the highest accuracy in the training set is selected. Otherwise, if k>0, then the smallest tree size is selected rather than the highest accurate tree (Murthy, 1997).

3.7.5 Advantages and Disadvantages of Decision Trees

Decision tree algorithms have a number of desirable properties: first, they can model a wide range of data distributions since only a few assumptions are made about the model and the data distribution; second, they are based on the hierarchical decomposition which implies better use of available features and computational efficiency in classification; third, they are very able to handle complex interactions between variables; fourth, they perform classification by a sequence of simple easy to understand tests whose semantics are intuitively clear to domain experts; fifth, they reduce the volume of data by transforming them in a more compact form which preserves the essential characteristics and provides an accurate summary. However, the main disadvantage of Decision Trees is that they tend to grow very large for realistic applications and are thus difficult to interpret by humans. Hence, there has been some research in transforming decision trees into other representations.

3.8 RULE INDUCTION ALGORITHMS

Inductive logic programming (ILP) is a research area that has emerged at the intersection of machine learning and computational logic. Using techniques from both fields, ILP aims to develop tools for inducing hypotheses from observations or using substantial knowledge from experience. ILP uses an initial background theory T and some evidence E consisting of positive (P), and negative (N) examples so that $E = P \cup N$. It then attempts to induce a hypothesis I that together with T explains E. In the usual case I, T, and E have to satisfy syntactic restrictions that are well known as bias B. The bias B includes prior expectations and assumptions that formulate the space that we have to build the hypothesis. In other words, the induced hypothesis I must satisfy the bias B and explain E with respect to theory T.

Several tools and techniques of ILP have been developed over the past decades. A very popular technique is rule induction. Rule induction algorithms have been established as basic component of many data mining systems and they represent an important area of research in the Artificial Intelligence (AI) science. A rule induction system uses conditions of the attributes in the training set and attempts to infer rules that govern the classification of previously defined

classes. It then uses these rules to find the class of unlabelled vectors in the test set. The condition of the attributes that built the rule set is a selection of the conditions from the relational algebra. Let $x_{it} = x_{i1t}, x_{i2t}, \ldots, x_{imt}$ be a set of attributes with domains $D_t = d_{1t}, d_{2t}, \ldots, d_{mt}$. A class $C_{jt}(H,L)$ is a subset of the training set T_t consisting of all vectors that satisfy the class condition $Cond_{jt}$ so that $C_{Ht} = (\theta_H \in T_t \setminus Cond_{Ht})$ and $C_{Lt} = (\theta_L \in T_t \setminus Cond_{Lt})$. Vectors that satisfy the condition R_H are positive examples of the class H, whereas vectors that do not satisfy this condition are negative examples of the class H. The conditions for each class that partition the training set T_t into subsets H, L are defined by the user. The task of the rule induction algorithm is to construct classification rules that consist of descriptions Des_H, Des_L for H and L, respectively.

There are three requirements that rule induction algorithms should meet in order to be useful tools for real world applications: first, they should induce rules that are able to classify future unlabelled vectors x_{it} correctly even if the data is noisy; second, they should induce rules that are as short as possible so that they minimise the possibility of overfitting in the presence of noise; and third, they should scale efficiently with the sample size and the time required for rule generation should be linear in the size of the data set. Several rule induction algorithms have been proposed in the literature over the last decade. These include among others the AQ15 system, the CN2, the DB-Learn, the RADIX/RX, and the Ripper-Rule Induction algorithm. These algorithms are discussed below. A more detailed presentation of these algorithms can be found in Holsheimer and Siebes (1991) and Cohen (1993, 1995).

3.8.1 AQ15

The AQ15 system is an inductive learning system that generates decision rules through constructive induction. Constructive induction is the process of using domain knowledge to generate new attributes that are not present in the data. To build a decision rule, the AQ15 system performs a heuristic search to find those logical expressions that account for all positive examples and no negative examples. The most preferred rule is selected based on a preference criterion that is defined by the user. The algorithm starts by focusing on one selected positive example and then generates a set of complexes that cover the positive example and no negative examples. The complex is a conjunction of attribute value conditions that relate an attribute to a value or to a disjunction of values. Using user-defined criteria, the best complex is selected and is added to a disjunction of complexes that consist the cover of the rule. The initial cover of the rule is either empty or supplied by the user. The AQ15 system has been tested for several medical domains including prognosis of breast cancer and diagnosis of lymphography using

small training sets. The system discovered rules that were competitive in accuracy with human experts. A disadvantage of the AQ15 algorithm is that the algorithm handles noise using pre- and post-processing techniques known as rule truncating.

3.8.2 CN2

The CN2 system is an adaptation of the AQ15 system and was designed to remove the dependency on pre- and post-processing techniques by incorporating a noise handling technique. The output of the algorithm is an ordered set of if-then rules. The main difference of this algorithm from the AQ15 algorithm is that the conditional part is a complex and not a disjunction of complexes that consist the cover of the rule. The complexes are specialised during the search process by adding a conjunctive attribute value condition or by removing a disjunctive value in one of the existing attribute value conditions. Two tests are performed to determine if the complex is both accurate and significant in the training set. The primary advantage of CN2 over the AQ15 algorithm is that the CN2 algorithm searches a larger area of the search space and does not restrict its search to only those rules that are consistent with the training examples. It constructs instead probabilistic rules where the conditional part covers examples of a single class as well as a few examples of other classes as well.

3.8.3 DB-LEARN

The DB-Learn system uses domain knowledge in the form of hierarchies of attribute values to generate descriptions for predefined subsets in a relational database. For each separate class, the DB-Learn algorithm constructs a relational table that is a disjunction of conjunctions of attribute-value pairs. The search process attempts to generalise this table to a much smaller table that covers all the examples belonging to a specific class. In this way, the class description is constructed. The maximum number of disjunction of conjunctions that are associated with this table are specified by a threshold which is defined by the user. The choice of this threshold is very crucial. If the threshold is small, it may result in overgeneralization and loss of valuable information. If the threshold is large it may result in a very complex rule with poor generalisation performance.

3.8.4 RADIX/RX

Another more specialised algorithm is the RADIX/RX system which is used for the discovery of relationships in a clinical database. A very important feature that differentiates this system from the previous systems is the incorporation of time in the set of examples. The training set consists of objects and each object stores information about a single patient at different times. Therefore, the generated knowledge can be represented by causal relationships. The set of examples can be represented as a three dimensional matrix. The first dimension represents the

number of patients, the second dimension represents the attributes for each patient, and the third dimension represents the different moments at which the value of each attribute is recorded. However, the main drawback of the RX system is that it does not use domain knowledge to guide search.

3.8.5 Ripper Rule Induction

Most of the rule induction algorithms use an overfit-and-simplify learning strategy to handle noisy data. According to this strategy a hypothesis is formed by first growing a rule set which overfits the data and then simplifying or pruning the rule set. A variety of methods have been proposed in the literature to prune rule sets. One well-studied method is known as Reduced Error Pruning (REP) proposed by Pagallo and Haussler (1990) and Brunk and Pazzani (1991). According to this method the training data is split into a growing set and a pruning set and a rule set is formed that overfits the growing set using some heuristic method. The rule set is then simplified at successive stages by applying pruning operators that reduce the error on the pruning set. This process terminates when applying any pruning operator would increase the error on the pruning set.

Empirical evidence suggests that REP improves generalisation performance on noisy data. Cohen (1993) showed that REP is computationally expensive for large data sets. He proposed instead an alternative method called Grow. Applying this method to a set of benchmark problems, he showed that Grow is competitive with REP with respect to error rates and is an order of magnitude faster.

Furnkranz and Widmer (1994) proposed a novel learning algorithm called Incremental Reduced Error Pruning (IREP). According to this algorithm a rule set is built in a greedy fashion, one rule at a time. After a rule is found, all examples covered by the rule are deleted. This process terminates when no positive examples exist or when the rule found by IREP has a large error rate.

Cohen (1995) implemented the IREP algorithm by randomly partitioning the uncovered examples into a growing set and a pruning set. The process of building a rule begins by greedily adding conditions to an empty conjunction of conditions and considers adding to this any condition of the form $x_{it} \leq \theta$ or $x_{it} \geq \theta$ where x_{it} is a continuous variable and θ is some value for x_{it} that exists in the training data. The building of the rule is continued until all positive examples are covered in the growing data set. After the rule is grown, it is then simplified by deleting any final sequence of conditions from the rule so as to improve its

performance on the pruning data set. During this phase of pruning, ad hoc heuristic measures are used to guide the greedy search for new conditions and simplifications. The conditions that are used to form the rules will have the following form:

if ($x_{it1} < 0.4$ *and* $x_{it2} > 0.6$ *and* $x_{it4} > 0.8$) *then H*
if ($x_{it1} < 0.2$ *and* $x_{it3} < 0.7$) *then H*
if ($x_{it1} < 0.8$ *and* $x_{it5} > 0.8$) *then H*
Else L

Once the rule set is complete, it is then optimised through multiple passes so as to reduce its size and improve its fit to the training data. Each optimisation pass involves constructing for each rule $Rule_i$, two alternative rules: a replacement rule and a revision rule. The replacement rule can be formed by growing and then pruning $Rule_i$ so as to minimise error of the entire rule set on the pruning data. The revision rule can be formed in a similar fashion as the replacement rule but the only difference is to add conditions to $Rule_i$ greedily rather than the empty rule. After both the replacement and the revised rule are formed, a decision can be made as to whether the final theory should include the revised rule, the replacement rule, or the original rule. This decision can be taken by inserting a variant of $Rule_i$ into the rule set and then start deleting rules that decrease the Total Description Length (TDL) of the rules and examples. The TDL of the examples and the simplified rule set can then be used to compare variants of $Rule_i$. Finally, rules can be added to cover any remaining positive examples. Cohen (1995) called the new algorithm RIPPER. Figure 3.8 gives an example of rules developed by the RRI algorithm.

3.8.6 Advantages and disadvantages of Rule Induction Techniques

Although a variety of other representations have also been used in machine learning, a great deal of research has focused on rule induction for the following reasons: first, rules are often easier for people to understand; second, certain types of prior knowledge can be easily incorporated in the learning process; and third, rule induction techniques overcome the use of the limited-knowledge propositional logic formalism and they can be easily extended to the first order logic (Cohen, 1995). This property should be considered particularly important because one of the well-established findings of artificial intelligence and machine learning is that the use of domain knowledge is essential for achieving intelligent behaviour.

One disadvantage of rule induction methods is that they scale poorly for large data sets. Although this is not of critical importance for our application, the RRI classifier has been shown to perform efficiently on large noisy data sets and scales nearly linearly with the number of examples in a data set. In addition, it should be mentioned that despite the fact that rule induction techniques offer interpretable rules, they are not expert systems. The knowledge engineer has still a substantial amount of work to perform in order to generate rules that perform well and are also sensible so that they can enhance the knowledge of domain experts. Despite this weakness, however, rule induction systems result in simple rules that are more preferable than other machine learning representations.

3.9 TESTING SUPERVISED LEARNING ALGORITHMS

3.9.1 Error Rates and their Estimation

The most widely used criterion for assessing the performance of a particular classification method is the misclassification rate or error rate. The misclassification rate or error rate indicates the proportion of objects that classified incorrectly after applying a particular classification rule. If the conditional probability distribution for each class is known, then the minimum error rate that can be achieved given a set of measurements is the Bayes error rate that can be written as follows (Hand, 1997),

$$\varepsilon_{Bj} = \int\left[1 - \max_j f(j / x_{it})\right] f(x_{it})dx \qquad j = H, L \tag{3.36}$$

where $f(x_{it})$ represents the overall distribution of measurement vectors x_{it} and $f(j / x_{it})[j = H, L]$ denotes the probability of belonging to class j at x_{it}. The Bayes error rate represents an optimal error rate because it provides a lower bound on any error rate that can be achieved by a classification rule.

If we assume that the data have multivariate normal distributions with unequal covariance matrices, then the decision surface will be quadratic. If we apply a linear classification rule to this problem, then the minimum possible error that can be achieved given a set of measurements will provide an indication of how far the applied classification rule is from the best rule that can be applied to the given problem. Therefore, this error rate will be greater than the Bayes error rate (Hand, 1997).

Another type of error rate is the one that can be achieved if a particular classification method estimates $f(j \setminus x_{it})[j = H, L]$ and on the basis of these estimates forms regions R_j in which it classifies examples to each class. This error rate is known as the actual rate of the classifier and can be expressed as follows (Hand, 1997),

$$\varepsilon_{Cj} = \int [1 - f(j \setminus x_{it})] f(x_{it}) dx \tag{3.37}$$

The actual error rate ε_{Cj} of a classifier is also known as the conditional error rate because it is conditional on the training set used in building a particular classifier. The expected value of ε_{Cj} over training sets of a given size is known as the unconditional or expected error rate ε_{Uj}. The conditional error rate ε_{Cj} can be used if we attempt to compare estimates, whereas the expected error rate ε_{Uj} can be used if we attempt to compare estimators.

One approach to estimate the actual error rate of a classifier is to re-apply the classification rule that has been derived from the training set to classify any individual in the training set and consider the proportion of misclassified individuals from each class as the class's estimated conditional error rate. However, this method is highly overoptimistic as an estimator of the error rate because the same data are being used both to construct the classification rule and assess its performance. The classification rule will be optimised to maximise separation between classes in the training set but it will not be an ideal rule for future objects that do not possess the same distributional properties as the objects in the training set. If the test set was available at the time we construct the classifier, then we could overcome this problem by using the training set to build the classification rule, and then use the test set to calculate the error rate. However, this would not be the ideal solution from a practical point of view because if a test set was available at the time we construct the classification rule, then we could use both the training and test sets to build a better classification rule without having any concern to assess the actual error rate in the test set (Hand, 1997). An alternative method that we can use if a test set is not available would be to partition randomly the training data into two parts. The classification rule can be formed using one part, whereas the error rate can be estimated by applying the classification rule in the second part and count the misclassifications. Krzanowski and Marriott (1995) and Hand (1997) observed that one problem with this method is deciding how to split the training set in order to balance the trade-off between obtaining a good classification rule and a good estimate of its performance. If the part that is used as a test set does not have the same

distribution as the part that is used as the training set, then we should not expect a correct assessment of the actual error rate. This method requires really large samples.

Computer-intensive methods of error-rate estimation suggest that an alternative approach to assess the actual error rate of a particular classifier would be to partition the training data into two separate sets several times. For each separate spilt, we can use one part to build the classification method and the other part to assess its performance. After this procedure is complete, we can simply average the results over the different splits. This approach is known as cross-validation (Hand, 1997). Variations of this approach include among others the bootstrap method, the rotation method, the jackknife method, and the leave-out method. These techniques are discussed in more detail below.

3.9.2 The Leave-One-Out Method

The most common solution to test the performance of the classifier if the training set is small is the leave-one-out method. According to this method, a single element is held out from the training set and the remaining $n-1$ elements are used to build the classification rule which is then used to classify the element that was left-out. In the next step, this single element is returned into the training set and a different element is removed. The classifier is then retrained using the remaining $n-1$ elements and it is tested on the new element that was left-out. By repeating this process for any individual element in the training set, every known element is used to both training and testing. The proportion of elements that are misclassified from each class gives a direct estimate of the actual error rate in that class.

The advantage of the leave-one-out method is that the test set is almost as large as the training set. This means that the estimate of the error rate is approximately unbiased. Any bias will occur only from the extra variation that will result from using $n-1$ data elements instead of n elements. Although this is a very desirable theoretical property, empirical evidence suggests that this method might have a relatively large variance (Hand, 1997). Hand (1987) discussed several alternatives that even though might be biased, they result in smaller mean square error than the leave-one-out method. Another problem with the leave-one-out method is that it is computationally costly since n separate classifiers must be built. In addition, if the classification rule requires computation of the inverse or a determinant of a matrix and the data set is large, then the recalculation of these quantities each time an element is omitted will result in extensive computational demands. Several researches proposed algebraic solutions that simplify the computational demands. A detailed review of these studies can be found in Krzanowski and Marriott (1995).

A close relative of the leave-one-out method is the rotation method. According to this method, a number m of mutually exclusive subsets are defined. The m-1 subsets are used to build the classification rule which is then used to classify the subset that was left-out. Repeating this procedure m times, all observations are tested out-of-sample.

3.9.3 Jackknife

The jackknife method is closely related to the leave-one-out method but it is based on a different idea. Let us assume that ε_{Mj} (j=H,L) are the class-conditional apparent error rates that result if we re-apply a classification rule that has been derived from the training set to classify any individual in the training set. Furthermore, let us assume that ε'_{Mj} are the class-conditional apparent error rates that result if we remove the i^{th} element from the training set and use the remaining $n-1$ elements to construct a new classification rule. The average of these apparent error rates that will result if we repeat the same process for each individual element in the training set will be $\varepsilon_{Mj(A)} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon'_{Mj}\left(j = H, L\right)$. Therefore, the jackknife estimate of bias will be $(n-1)\left(\varepsilon_{Mj} - \varepsilon_{Mj(A)}\right)$. Using this estimate, we can write the jackknife estimate of the expected-unconditional error rate as follows (Krzanowski and Marriott, 1995),

$$\varepsilon_{Oj(U)} = \varepsilon_{Mj} + (n-1)\left(\varepsilon_{Mj} - \varepsilon_{Mj(A)}\right) \tag{3.38}$$

To find the jackknife estimate of the actual-conditional error rate, we need to find the class-conditional probabilities that result if all n elements in the training set are classified by the rule that was formed after removing the i^{th} element. If we assume that $\varepsilon'_{Mj(A)}$ is the average of the error rates in class j over possible omissions, we can express the jackknife estimate of the actual error rate as follows (Krzanowski and Marriott, 1995),

$$\varepsilon_{Oj(C)} = \varepsilon_{Mj} + (n-1)\left(\varepsilon'_{Mj(A)} - \varepsilon_{Mj(A)}\right) \tag{3.39}$$

3.9.4 Bootstrap

The basic idea of the bootstrap method is to use the original dataset to obtain new datasets of equal size to the complete dataset by sampling with replacement. To obtain an estimate of the error rate, a large number of bootstrap samples are chosen randomly. Each sample is a replicate of the original dataset and is taken by sampling with replacement. Sampling with replacement means that some data elements will be omitted from the bootstrap sample while others will

appear more than once. Each bootstrap sample is used to construct a classification rule which is then used to predict the classes of the elements that were not used in the training set. The average error rate over all bootstrap samples is then used to give an estimate of the error rate of the original classification rule (Efron and Tibshirani 1993; Michie et al. 1994).

Let us assume that $\varepsilon_{Mj(\beta)}$ is the proportion of the bootstrap sample from class j that is misclassified by the classification rule computed from the bootstrap sample. In addition, let us assume that $\varepsilon'_{M_j(\beta)}$ is the proportion of the original training set from class j that is misclassified by the classification rule computed from the bootstrap sample. The difference between $\varepsilon_{Mj(\beta)}$ and $\varepsilon'_{M_j(\beta)}$ can be written simply as $f_{j(\beta)} = \varepsilon_{Mj(\beta)} - \varepsilon'_{Mj(\beta)}$. If we repeat this procedure N times, then the average $\bar{f}_{j(\beta)} = \sum_{\beta=1}^{N} \frac{f_{j(\beta)}}{N}$ is the bootstrap estimate of bias in the apparent error rate from class j, whereas the difference $\varepsilon_{Mj} - \bar{f}_j$ is the bootstrap bias estimate of the actual error rate. (Krzanowski and Marriott, 1995).

Part Three: Comparative Studies on Supervised Learning Algorithms

3.10 EMPIRICAL EVIDENCE

There are various studies in the literature that compared supervised classification rules. Although there are various possible measures of performance, more often studies compared classification rules in terms of their error rates. Some of these empirical studies are discussed below.

3.10.1 General Comparative Studies

LDA has been sharply criticised in the literature because its proper functioning hinges on restrictive assumptions (Eisenbeis 1977; Altman and Eisenbeis 1978; Tollefson and Joy 1978; Ohlson 1980; Pinches 1980; Karels and Prakash 1987; and Odom and Sharda 1990). For the linear discriminant function to provide a classification rule that minimises the probability of misclassification, the variables in each group must be from multivariate normal distributions and the covariance matrices for all groups must be equal. While many of the ratio applications employ methodologies that rely on either univariate or multivariate normality assumptions and parametric test procedures, little is known about the distributional properties of financial ratios. Empirical evidence seems to indicate that most ratio distributions are either highly skewed, flat, and/or dominated by outliers. After examining the cross-sectional distributions of 11 ratios over

the 1953-1972 period for large populations of manufacturing firms, Deakin (1976) concluded that the normality assumption is not tenable except for the debt to total assets ratio. Departures from normality may occur when the population contains some extreme observations that can dominate parameter estimates when they are present. Cochran (1963) noted that such outliers have the effect of increasing the sample variance and decreasing precision. Frecka and Hopwood (1983) extended Deakin's (1976) study by examining the effects of outliers on the cross-sectional distributional properties of selected financial ratios. Their results indicated that by deleting outliers, normality or approximate normality can usually be achieved for their population of manufacturing firms and for specific industry groupings. Mardia (1971) developed tests for multivariate skewness and kurtosis that have been programmed. Applications of this approach indicated that the business data analysed did not meet the assumption of multivariate normality.

Nonmultivariate normality influences the test for the equality of the dispersion matrices. Mardia (1971) found that tests for the equality of dispersion matrices are sensitive to nonmultivariate kurtosis, but perhaps not to nonmultivariate skewness. Gilbert (1969) concluded that linear and quadratic rules can produce significantly different classification results that are directly related to the differences in the dispersion matrices, the number of variables, and the separation between groups. Lachenbruch et al. (1973) found that classification results for the linear discriminant function are affected greatly by nonmultivariate normality, while the quadratic classification rule does not produce better results.

Studies indicate that quadratic rules produce more accurate estimates if the sample is large relative to the number of variables, if the difference between the dispersion matrices is large, and if the data have multivariate normal distributions. In other situations, linear rules may produce more accurate estimates of the probabilities of misclassification (Pinches and Mingo, 1973).

Titterington et al. (1981) compared independence models, models with categorical interactions, latent variable models, kernel methods, LDA, QDA, and logistic discrimination using data from 1000 patients with severe head injuries. They found that LDA and independence models perform better than kernel models or models with categorical interaction. Empirical studies that report the superiority of logistic regression over LDA can be found in McLachlan (1992).

Hand (1983) compared LDA with kernel methods using multivariate binary data sets and he found little difference in the estimated true error rates. Huang and Lippmann (1987) compared various neural networks with decision trees and statistical algorithms and they reported that

neural networks outperform the other algorithms. These findings were supported by Sethi and Otten (1990) and Bonelli and Parodi (1991). Gorman and Sejnowski (1988) compared the backpropagation algorithm with a nearest neighbour method in terms of their ability to classify sonar targets. They reported that the backpropagation algorithm outperforms the nearest neighbour method.

Kirkwood et al. (1989) compared the ID3 algorithm with LDA in classifying the gait cycle of artificial limbs. They found that the ID3 algorithm outperforms LDA. Shavlik et al. (1991) compared the same algorithms including a perceptron. They found that the backpropagation algorithm performs as well as or better than the ID3 algorithm. However, they reported that the backpropagation algorithm is significantly slower than ID3. Weiss and Kulikowski (1991) compared the backpropagation algorithm with decision tree classifiers. They found that decision tree classifiers outperform the backpropagation algorithm.

Tsaptsinos et al. (1990) compared the ID3 algorithm with two other neural network algorithms on an engineering control problem. They found that the ID3 outperforms the neural network algorithms. Atlas et al. (1990) also compared neural networks with tree classifiers. They reported that both methods produce comparable error rates. Brown et al. (1993) support these findings.

Ripley (1993) compared various statistical methods with neural networks as well as a decision tree classifier using the Tsetse fly data. He reported that the nearest neighbour algorithm, the backpropagation, the projection pursuit, and the decision tree algorithm produce very favourable error rates. In addition, he reported that the decision tree algorithm produces more interpretable results, whereas the neural networks are significantly slower than the other methods.

Michie et al. (1994) performed a large comparative study using 20 different classification methods. These methods included among others LDA, QDA, PPR, kernel methods, nearest neighbour methods, causal network methods, tree methods, radial basis function methods, and other neural network structures. These methods were applied on different datasets that included among others datasets involving costs, credit risk datasets, image related datasets, and other datasets such as shuttle control, diabetes, DNA, technical data, Belgian Power I and II, machine faults, and tse-tse fly distribution. Michie et al. reported that LDA and QDA perform as might be expected on the basis of relevant theory. When the data is sufficiently large and the class covariance matrices are very dissimilar, the results suggest that the classification performance of QDA is better than the classification performance of LDA but it requires more computational

resources. Minor differences were found in the classification performance of ordinary and logistic classification rules. Concerning the nearest neighbour methods, Michie et al. reported that although these methods have a very satisfactory performance, they are extremely slow for very large data sets. However, they found that the LVQ algorithm has about the same error as a k-nearest neighbour algorithm, but it is six times faster and it uses 25% less storage. The experimentation results on neural networks suggested that the neural networks have the best predictive performance in most of the datasets with only exception the datasets involving cost matrices. However, in terms of computational complexity, neural networks are found more complex than other algorithms. The experimentation results on decision tree algorithms suggested that different variants of decision trees have about the same classification performance. Michie et. al found no evidence that a particular splitting criterion is better than the other, but they reported that substantial benefits may result after using pruning techniques. However, they were unable to identify what percentage of the data set should be used for pruning. Concerning rule induction algorithms, Michie et al. reported that the CN2 algorithm produces very favourable results and its performance is also good on datasets involving costs. However, Michie et al. observed that since a decision tree may be expressed in the form of rules there is no practical reason of choosing rule-based methods unless the complexity of the data requires more simplifying representations.

In terms of error rates, Michie et al. reported that the algorithms have different performance on different datasets. In the credit datasets, the best algorithms are SMART, Cal5, DIPOL92, C4.5, and IndCART. Michie et al. suggested that a logical explanation why decision trees algorithms such as Cal5, C4.5 and IndCART perform well on credit datasets might be that credit datasets are partitioning datasets since they were constructed by humans who classified the data on the basis of attributes. In the image datasets, the results were quite different. Non-linear algorithms such as the k-nearest neighbour algorithm, the LVQ algorithm, the QDA, and the backpropagation algorithm have very satisfactory performance, whereas DIPOL92 also produces favourable results. Michie et al. observed that non-linear algorithms are expected to perform well on image datasets given the non-linear structures that are present in these datasets. In the datasets with costs, the LDA and logistic discriminant analysis share the best performance, whereas the performance of the QDA is also competitive. In the remaining datasets, it is more difficult to find algorithms that have the same performance on different datasets. For example, the DIPOL92 algorithm performs very well on the diabetes, DNA, and machine faults datasets, the SMART algorithm performs very well on the Belgian datasets, whereas the CN2 algorithm performs very well on shuttle control and tse-tse fly distribution.

In terms of other evaluation criteria such as whether the particular programs can deal with a cost matrix, missing values, and different types of data as well as whether they are understandable and user-friendly, Michie et al. reported that decision tree and rule induction algorithms have the highest ratings.

Jutten (1995) compared seven classification methods including a k-nearest neighbour (KNN) classifier, a Gaussian quadratic classifier (GQC), a LVQ algorithm, a multi-layer perceptron (MLP), a sub-optimal Bayesian classifier (IRVQ) which is based on radial Gaussian kernels and uses an unsupervised learning method based on vector quantization to obtain a low-memory kernel density estimator, a piecewise linear separation (PLS) incremental classifier which tries to find in an incremental way the best linear approximation for the discriminant function, and the restricted Coulomb energy (RCE) classifier which is one of the first incremental models of neural networks. According to this classifier, decision units are characterised by their influence region which is defined by the hypersphere around the unit and whose radius is equal to the threshold of the unit. The input space is then divided into zones and each zone is represented by the different decision units. Jutten used two different types of datasets to compare the classifiers: artificial datasets and real datasets. The artificial datasets were generated according to three main requirements: first, heavy interaction of the class distributions; second, high degree of non-linearity of the class boundaries; and third, various dimensions of the vectors. The real datasets were selected according to four main requirements: first, datasets that were used extensively in classification; second, already published results on these datasets; third, various dimensions of the vectors; and fourth, sufficient number of vectors. After implementing the classifiers, Jutten reported that the KNN, the GQC, the IRVQ, the LVQ, and the MLP outperform the RCE and the PLS classifiers on both artificial and real data sets.

3.10.2 Comparative Studies on Financial Applications

Several studies compared neural networks and LDA in terms of their ability to predict financial distress (Williams 1985; Odom and Sarda 1990; Webb and Lowe 1990; Coats and Fant 1993; to name a few). All these studies indicated that neural network models outperform LDA in terms of overall accuracy to correctly classify companies and predict financial distress using financial ratio data. Tyree and Long (1996) compared the PNN model with the LDA methodology in terms of their ability to predict financial distress. Their test results indicated that the PNN performs better than LDA analysis. Albanis et al. (1997) compared and contrasted the PNN with the LDA methodology in terms of their ability to correctly classify companies into homogeneous industrial sectors. Although they found evidence of the existence of linearities in the data set, the test results indicated that the classification performance of the PNN model is as

good as or better than the LDA model if the assumptions of multivariate normality and equality of the group dispersion matrices are violated.

Srinivisan and Kim (1987) compared parametric, non-parametric, and judgmental classification procedures in terms of their ability to analyse corporate credit granting. They found that a decision tree algorithm performs better than the other methods. Carter and Catlett (1987) used probability trees to assess credit card applications, whereas Michie (1989) used decision trees to analyse the customer credit granting.

A number of studies applied statistical models that performed remarkably well in explaining and predicting the ratings of a large cross section of corporate bonds (Horrigan 1966; Pogue and Soldofky 1969; Pinches and Mingo 1975; Ang and Patel 1975; Kaplan and Urwitz 1979). A number of studies have compared various regression models and neural networks in terms of their ability to predict bond ratings. Dutta and Shekhar (1988) applied both multiple regression models and neural networks to detect AA and non-AA rated bonds. They selected 47 bonds at random and they used 30 patterns for training and 17 for testing. The selected bonds had approximately the same maturity and they selected from different industrial sectors. The test results indicated that several neural networks architectures correctly classify between 76.5% and 88.3% of the bonds in the test set, whereas regression is less successful and classifies correctly only 64.5% of the bonds in the test set. Singleton and Surkan (1994) compared LDA with Neural Networks in terms of their ability to classify the ratings of the Bell Telephone companies divested by American Telephone and Telegraph (AT&T) in 1982. They found that neural networks correctly classify 89.5% of the Standard and Poor's ratings and 61.1% of the Moody's ratings, whereas LDA correctly classifies 84.2% of the Standard and Poor's ratings and 55.6% of the Moody's ratings.

A variety of studies have applied artificial neural networks to time-series prediction. These studies suggest that several of these prediction and decision-taking tasks present sufficient non-linearities to justify the use of ANNs (Moody et al. 1993; Refenes 1994). The non-linear models that have been applied to predict time-series incorporate three types of explanatory variables: i) technical variables that depend on the past-price sequence; ii) micro-economic stock specific data; and iii) macro-economic variables which give information about the business cycle. Levin (1995) implemented multilayer feedforward neural networks for predicting a stock's excess return based on its exposure to various technical and fundamental factors. In this work, a portfolio optimizer was used to limit exposure to undesired factors and to contain turnover. Levin also used zero-investment strategy portfolios that were designed to have effectively zero exposure to certain kinds of risk by holding long and short positions. In spite of the very low

signal to noise ratio of the raw data, the model is able to extract meaningful relationships between factor exposures and expected returns. When utilised to construct hedged portfolios, the predictions achieve persistent returns with very favourable risk characteristics.

Bengio (1996) explored the question of whether the same neural network should be used for all stocks at the same time or a different network for each individual stock. To investigate this question, he performed a series of experiments in which different subsets of parameters were shared across different models. The experiments were performed on 9 years of data concerning 35 large capitalisation companies of the Toronto Stock Exchange. During these experiments, the networks were not trained to predict the future return of stocks, but instead to directly optimise a financial criterion. Bengio reported that a partial sharing of parameters is more preferable and it achieves more consistent results. Another interesting finding was that very large returns might be obtained at risks comparable to the market if a combination of partial parameter sharing and training with respect to a financial criterion is used together with a small number of explanatory input features that include technical, micro-economic, and macro-economic information.

John et al. (1996) approached the problem of stock selection from the perspective of knowledge discovery in databases. Using a large database of historical information on many stocks, they attempted to select some portfolio of stocks that were likely to exhibit exceptional returns over a future period of time. Using measures of trends in the stocks' prices as well as fundamental data on the companies, they applied a system that is able to induce a set of classification rules to model the data it is given. After implementing this system, John et al. reported that their portfolio produces a total return of 258% over a four-year period, significantly outperforming the benchmark, which returns 93.5% over the same period. They claimed that this performance is not attributable to growth/value or size effects alone.

Albanis (1998a) examined the applicability of rule induction techniques to predict long-term bond ratings using financial ratios. In addition, he compared and contrasted the LDA, the PNN and the RRI algorithm in terms of their ability to classify bond issues of 132 U.K. companies into three boundary rating categories namely A, AA, and AAA. He reported that both the PNN and the RRI algorithms have better classification performance than the LDA model. In a subsequent study, Albanis (1998b) compared and contrasted the LDA, the PNN, and the RRI algorithm in terms of their ability to classify long-term bond ratings if the data set is small. After implementing the models to classify a number of bond issues into boundary rating groups, he found that the PNN has better classification performance than the RRI algorithm if the data set is relatively small, whereas the RRI has better classification performance than the PNN

model if the data set is relatively large. However, both the PNN model and the RRI algorithm classify significantly better than the LDA model. Albanis concluded that one of the main reasons for the poor performance of the LDA model is the violation of the assumption of multivariate normality. This study is discussed in detail in Chapter 4.

Leung et al. (2000) compared the performance of various classification models to a group of level estimation methods using the S&P, FTSE 100, and Nikkei 225 market indices over the 1967-1995 period. The classification models were LDA, PNN, logit and probit. On the other hand, the level estimation counterparts were exponential smoothing, multivariate transfer function, vector autoregression with Kalman filter, and multilayered feedforward neural networks. The classification models were aimed at forecasting the sign (direction) of index return, whereas the level estimation models were aimed at estimating the value of the return. The comparisons between the two types of models were conducted on the basis of forecasting performance and investment return. After experimentation, Leung et al. reported that the classification models outperform the level estimation models in terms of predicting the direction of the stock market movement and maximising returns from investment trading. They also reported that the classification models are able to generate higher trading profits than the level estimation methods.

Part Four: Summary and Conclusions

3.11 DISCUSSION AND REMARKS

In this chapter we discussed the main approaches to discriminant analysis and we presented the main supervised classification algorithms including parametric classifications rules, semi-parametric classification rules, non-parametric smoothing methods, neural networks, recursive partitioning methods, and rule-induction methodologies.

Parametric and semi-parametric classification rules have been developed by the explicit assumption of some model. Therefore, most of these rules will be fairly robust to certain assumptions for the assumed model. For example, two assumptions should be satisfied for the linear discriminant function to provide a classification rule that minimises the probability of misclassification: first, the variables in each group must be from multivariate normal distributions; and second, the covariance matrices for all groups must be equal. Empirical studies suggest that these requirements have frequently been violated. Evidence suggests that linear and quadratic classification rules suffer from departures from multivariate normality and equality of the group dispersion matrices. On the other hand, remedial measures taken to

improve the multivariate normality are often inadequate. These results were more than enough to support the development of non-parametric classification rules. Non-parametric classification rules do not postulate models for the population-conditional distributions. They suggest instead to estimate first the probability density functions from the training data, and then apply the estimated functions to the individuals for classification.

One consequence of the rapid development of computer power in the 1980s was the development of computer-intensive classification algorithms. These include among others neural networks, decision trees and rule induction techniques. Neural networks are likely to be most superior to other methods if the data is fuzzy, chaotic, or exhibits unpredictable non-linearity that cannot be detected by linear models or other models that are based on strictly defined models. Furthermore, neural networks are likely to have better generalisation performance than other statistical methods if the data incorporates human judgement or other qualitative factors because they detect patterns in data in a manner analogous to human thinking. This excellent performance of artificial neural networks should not be considered surprising if we consider that these algorithms have been built on strong theoretical foundations. For example, the three-layer feedforward neural network and the PNN have powerful approximation properties. However, neural networks have several drawbacks as well. One major drawback of neural networks is that the learning process is very slow. Another major drawback of neural networks is their inherent inability to explain in a comprehensible form the process by which a given decision or output generated by the model has been reached. In addition, neural networks might suffer from the problem of overfitting if the data set is small compared to the free parameters of the network.

Decision tree algorithms are powerful for classification and prediction. These algorithms are able to model a wide range of data distributions since only a few assumptions are made about the model and the data distribution. In addition, they are based on the hierarchical decomposition that implies better use of available features and computational efficiency in classification. Therefore, they are very able to handle complex interactions between variables. Furthermore, a major advantage of decision trees is that they perform classification by a sequence of simple easy to understand tests whose semantics are intuitively clear to domain experts. The main disadvantage of decision trees is that they tend to grow very large for realistic applications and are thus difficult to interpret by humans. In response to this limitation, there has been some research in transforming decision trees into other representations using rule induction techniques. Although a variety of other representations have also been used in machine learning, a great deal of research has focused on rule induction for the following reasons: first, rules are often easier for people to understand; second, certain types of prior

knowledge can be easily incorporated in the learning process; and third, rule induction techniques overcome the use of the limited-knowledge propositional logic formalism and they can be easily extended to the first order logic. However, one disadvantage of rule induction methods is that they scale poorly for large data sets. In addition, despite the fact that rule induction techniques offer interpretable rules, they are not expert systems. The knowledge engineer has still a substantial amount of work to perform in order to generate rules that perform well and are also sensible so that they can enhance the knowledge of domain experts. Despite this weakness, however, rule induction systems result in simple rules that are more preferable than other machine learning representations.

There are various studies in the literature that compared supervised classification rules. Although there are various possible measures of performance, more often studies compared classification rules in terms of their error rates. Only a few studies compared other important aspects of performance such as interpretation of the results, speed of classification, cost of classification, particular types of data, important constraints on the problem that the classification method should satisfy, the amount of prior knowledge on the problem under investigation, data preprocessing etc. Hand (1997) observed that the results of most studies should be interpreted with caution because there were often undertaken by researchers with special interest in a particular method over the other methods. Therefore, it is likely that some of the comparisons of different methods might be unfair because researchers have implemented properly only the methods where they were expert and either they have ignored or they have not optimised properly the methods on which they were less expert. Most of these studies used data sets that were likely to favour one method over the others. For example, if the data were generated from two mutlivariate normal classes with equal covariance matrices, Hand (1997) observes that LDA would be expected to perform better than the other methods.

Considering the empirical evidence but also the "strong" and "wild" capabilities of supervised classification rules, we explored the possibility of using these algorithms to address the problems of stock return predictability and stock selection. Real datasets such as financial data do not artificially favour one method over the other because they exhibit patterns that are inconsistent and chaotic in the time scale. Therefore, we selected a small number of heterogeneous algorithms for our application because we considered that one individual algorithm would be insufficient to deal with the complex financial processes. The results of this study are summarised in the following Chapters.

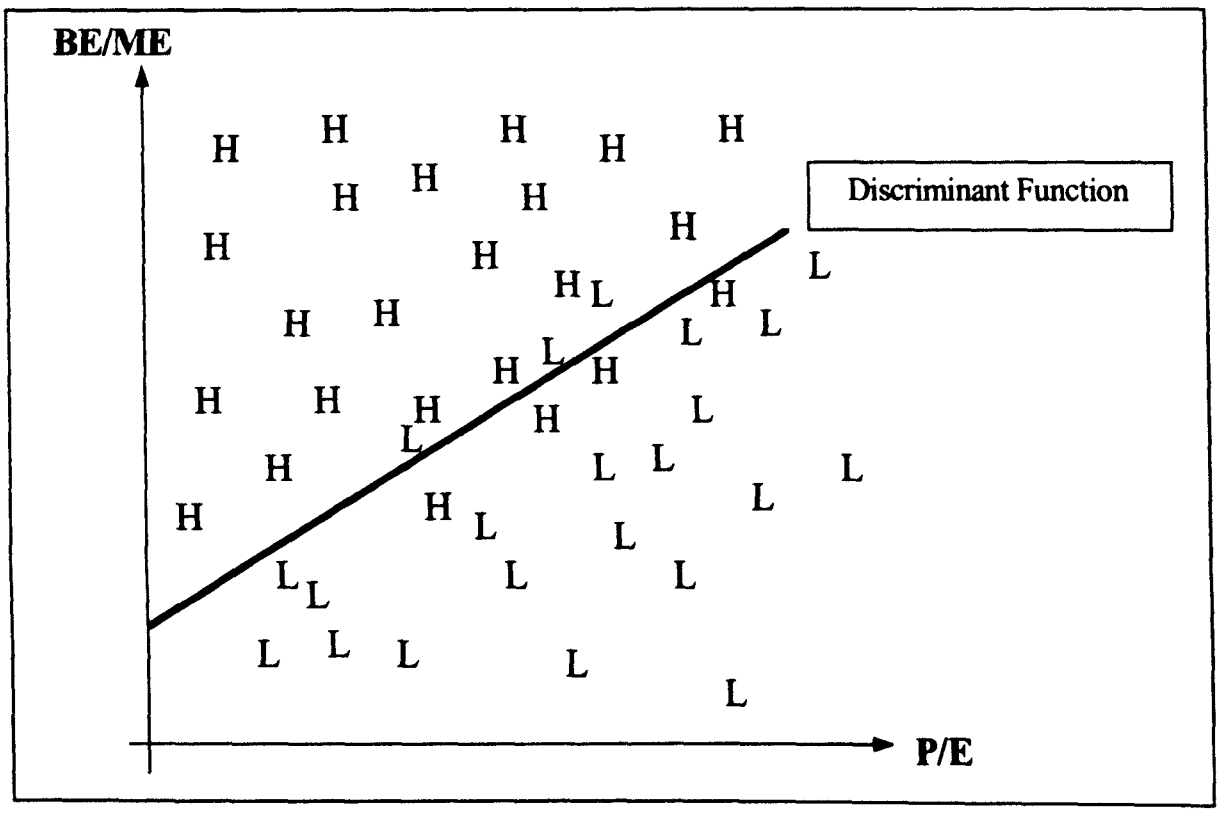


Figure 3.1: Linear Discriminant Analysis (LDA)

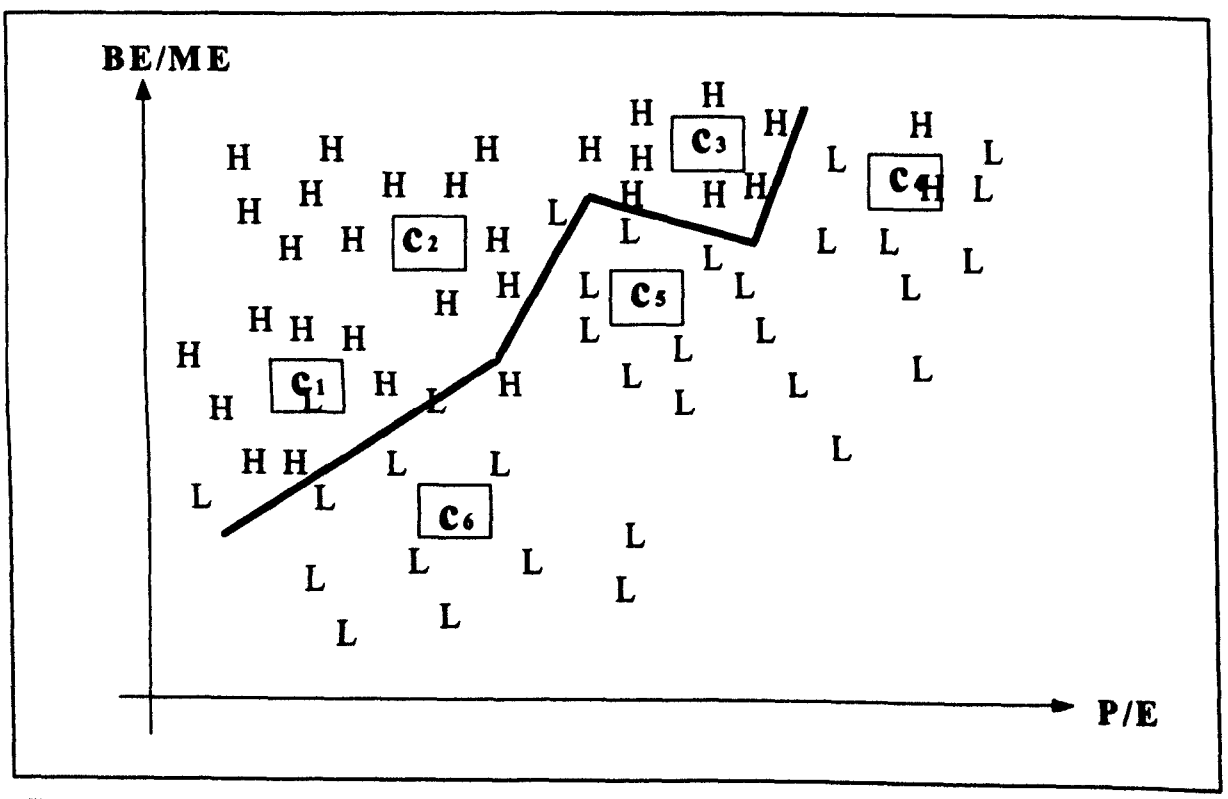


Figure 3.2: The Learning Vector Quantization (LVQ)

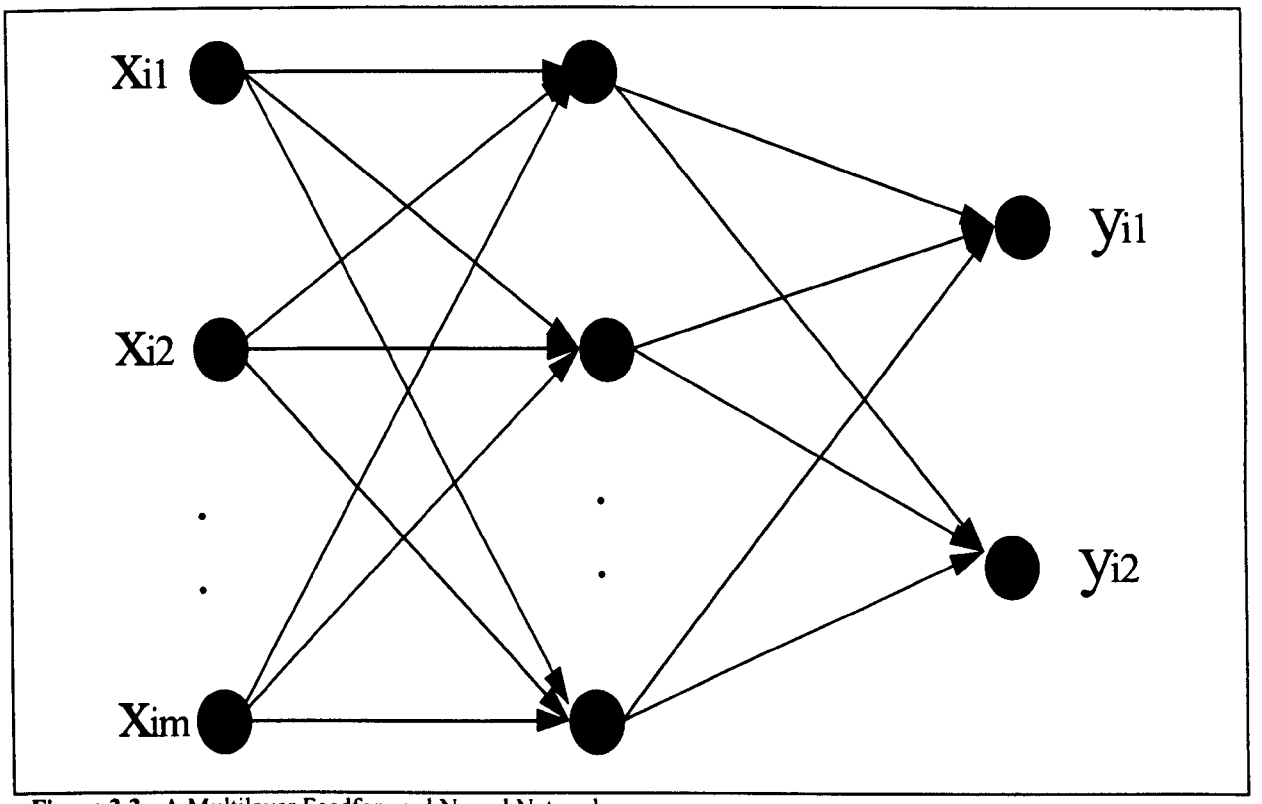


Figure 3.3: A Multilayer Feedforward Neural Network.

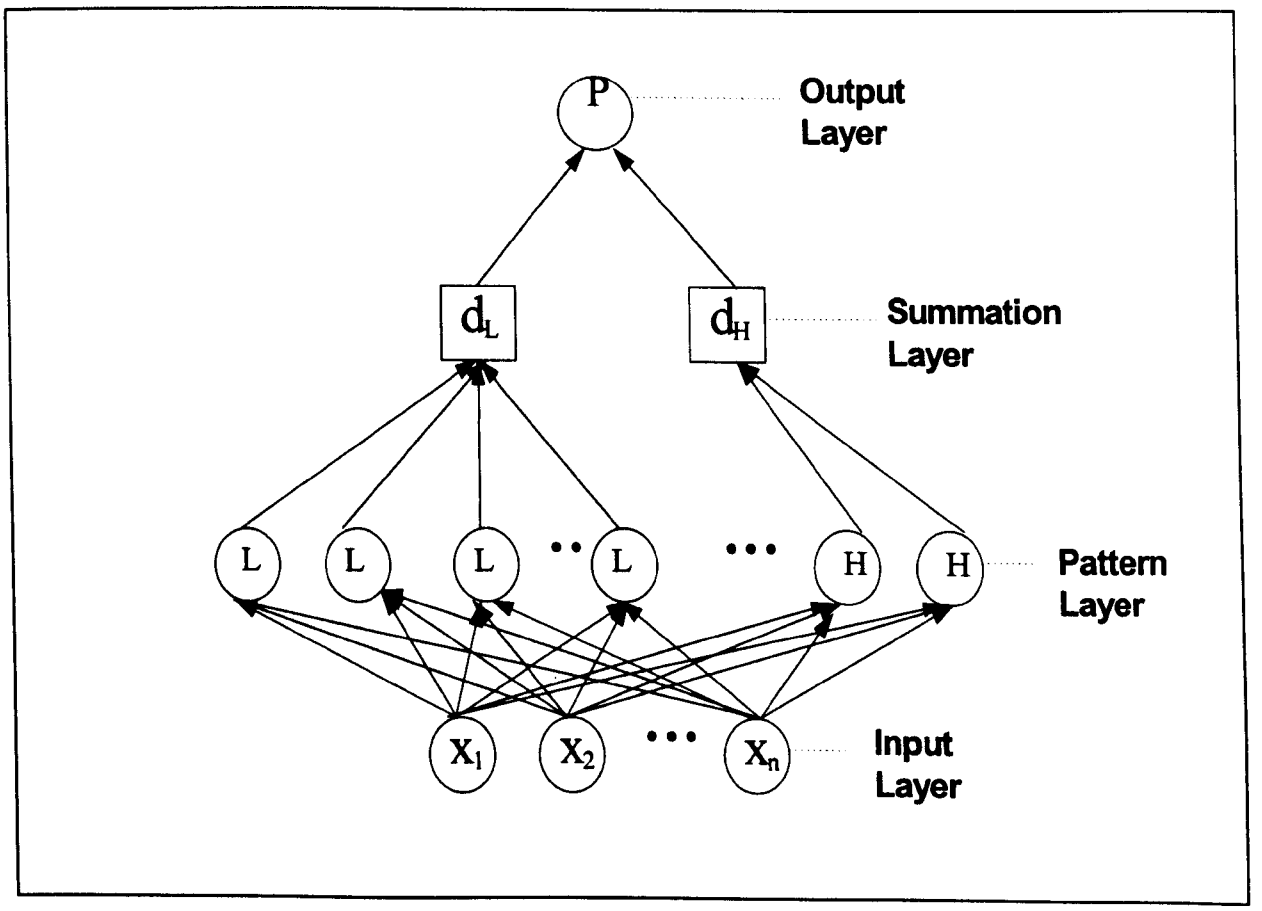


Figure 3.4: The Probabilistic Neural Network (PNN)

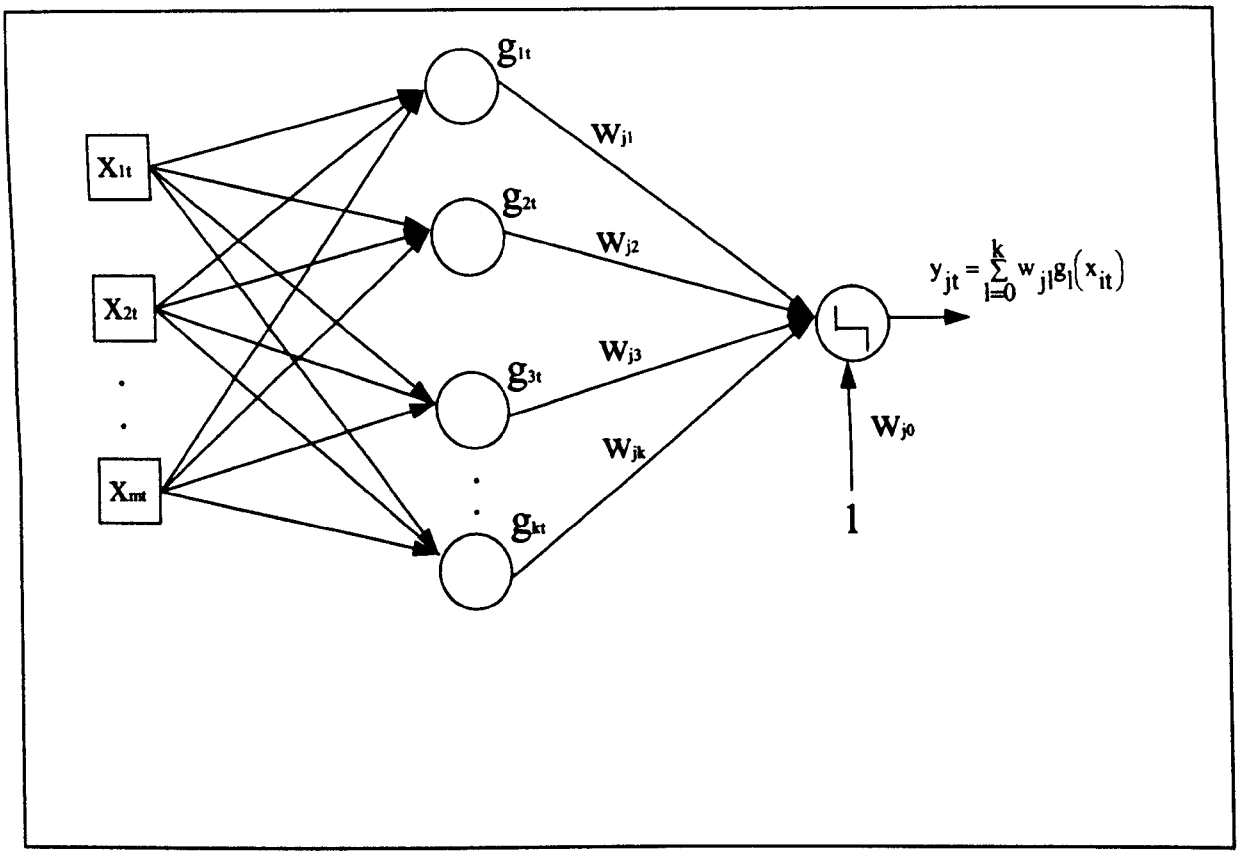


Figure 3.5: The Radial Basis Function (RBF) Network

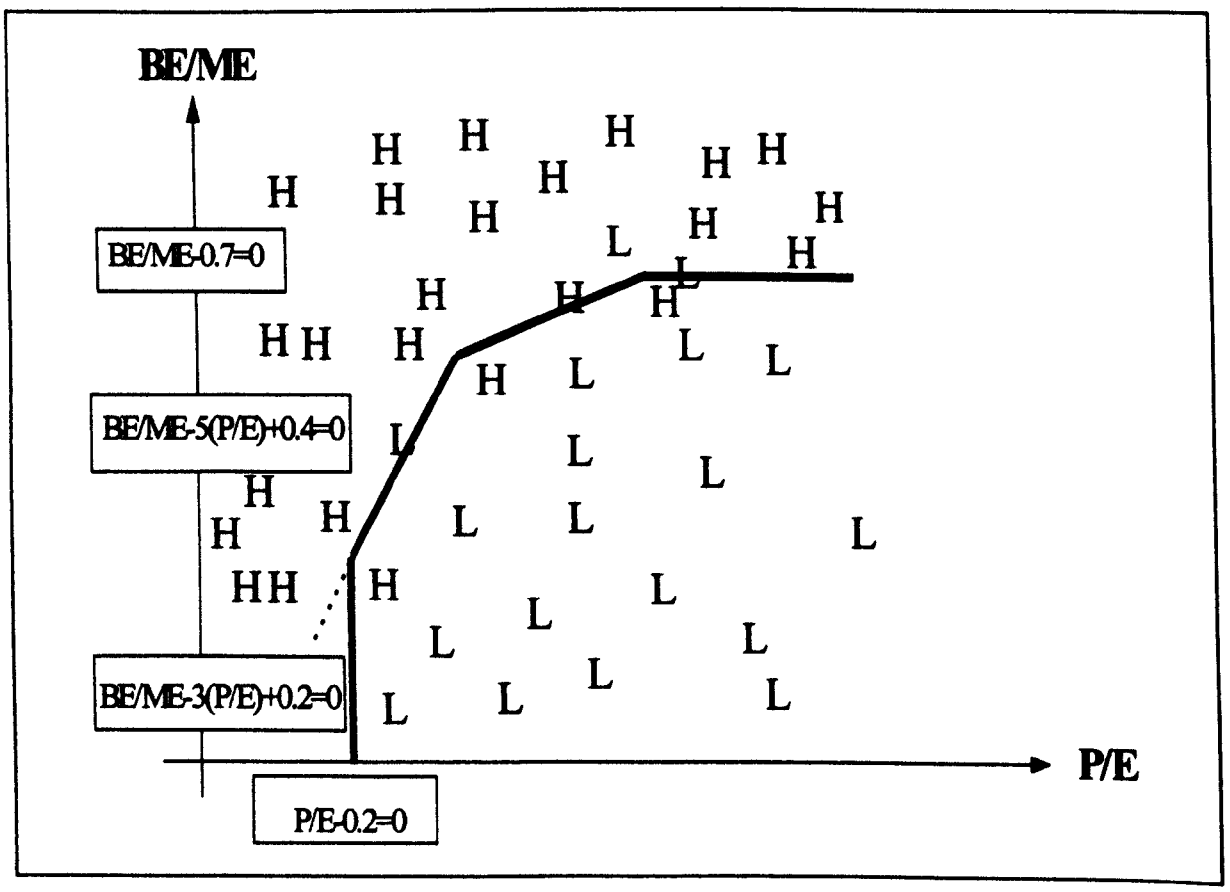


Figure 3.6: An Imaginary Decision Tree with Oblique Splits

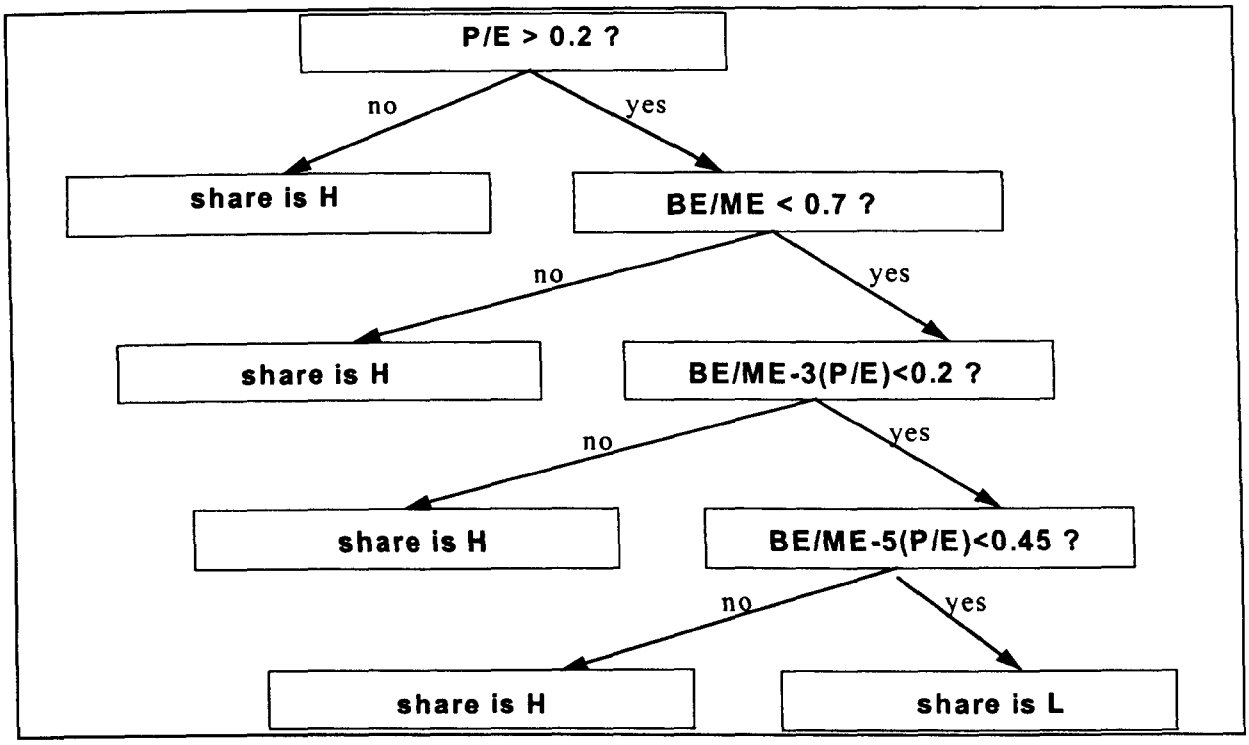


Figure 3.7: A Sequential Representation of a Decision Tree with Oblique Splits

Initial hypothesis:

1. **High:** EPS(%)>=11.19, DEBT/TA<=13.91, PAT(%)<=34.44, DEBT/TA>=3.16, PBT/TA>=8.51, NI/TCE<=19.55.
2. **High:** P/E<=888.65, D/Y>=6.16, CF/TA<=1.36, CF/TA>=- 2.48.
3. **High:** CF/TA>=8.88, EPS (%)>=8.66, TA (%)<=12.56.
4. **High:** PBT/CL<=-14.55.
5. **High:** CF/MKBD>=7.50, P/E>=931.05, MKBD (%)>=94.63, CA/CL<=208.08.
6. **High:** EY>=7.56, EPS (%)<=-33.96.
7. **High:** EY>=7.51, MKBD (%)>=31.50, TA (%)<=10.48, P/E>=1260.3.
8. **High:** EPS (%)>=26.83, SR/TA>=152.56, CF/SR<=1.55.
9. **High:** PBT/CL>=23.66, PAT/TA<=6.82, CL/EQ>=47.67, SR/TA<=105.26.
10. **High:** CF/TCE>=3.20, NI/TCE<=6.89, PAT/TA>=4.68.
11. **High:** PBT/CL<=6.60, PAT/TA>=1.18, TA (%)<=-0.69.
12. **High:** MKBD (%)<=-30.54, PBT/TA<=0.47, EPS (%)>=-106.9, DEBT/TA>=13.05.
13. **High:** PBT/CL>=33.22, CL/EQ>=48.37, PBT/CL>=63.04, BE/ME>=26.32.
14. **High:** DEBT/EQ<=16.26, EY>=8.33, PAT/TA<=7.53, SR (%)>=2.94.
15. **High:** SR (%)<=-4.76, NI/TCE>=10.28, NI/TCE<=11.30.
16. **High:** PAT/SR<=3.01, PBT/CL>=13.12, EPS (%)<=-32.56.
17. **High:** CF/TA>=10.45, TA/EQ>=272.77, TA (%)<=5.78.
18. **Low :** true.

Figure 3.8: Rules developed by RRI

CHAPTER 4: AN INVESTIGATION OF THE DISTRIBUTIONAL PROPERTIES OF FINANCIAL RATIOS - APPLICATION TO BOND RATINGS

In the previous Chapter, we reviewed a variety of classification methods that may be more suitable than either linear models or strictly defined non-linear models to deal with unpredictable non-linearities and other complex processes in the financial data. In this Chapter, we investigate the distributional properties of financial ratios and we compare three heterogeneous classifiers namely, Linear Discriminant Analysis (LDA), Probabilistic Neural Network (PNN), and Ripper Rule Induction (RRI) in terms of their ability to predict long-term bond ratings.

We performed this experiment for two main reasons – first, to examine the distributional properties of financial ratios; and second, to test the robustness of linear and non-linear models under different distributional assumptions. To test the robustness of linear and non-linear models under different distributional assumptions, we have chosen the LDA as a representative of the family of linear models and the PNN and RRI classifiers as representatives of the family of non-linear models.

The results of this experiment demonstrated that non-linear models do not depend on distributional assumptions in the same degree as the linear model and that they are more flexible to deal with unpredictable non-linearities and other complex processes in the financial data. We suggest that the PNN should be implemented if the data set is relatively small because it does not suffer from the assumption of multivariate normality in the distributions of the financial ratios in the same degree as the LDA. On the other hand, the RRI should be implemented if the data set is relatively large because not only it classifies better than the LDA and PNN models, but it also offers interpretable rules that are easy to visualise and understand, as opposed to the inherent inability of the PNN to explain in a comprehensible form the process by which a given decision generated by the model has been reached. The outcome of this implementation is a new methodology to predict long-term bond ratings using probabilistic neural networks and rule induction techniques.

The Chapter is organised as follows. In Section 4.1, we discuss the data and methodology that we used in this study. In Section 4.2, we present the results of our experimentation. Finally, in Section 4.3, we discuss the results and we provide the conclusions.

4.1 DATA AND METHODOLOGY

In this application, we are particularly interested in whether a particular bond will be classified as A, AA or AAA based on accounting information. Let us assume that r_{it} is the rating on some bond i at time t, and x_{it} is the vector of accounting information attributes for company i known at time t. The idea is to apply a classification method to assign r_{it} to one of the three classes $C_{jt}(j = A, AA, AAA)$ using as inputs the vector x_{it} of variables that represent accounting information at time t.

Industrial corporate bonds are assigned quality ratings by rating agencies such as Standard & Poor's (S&P's), Moody's, Fitch etc. The importance of these ratings is demonstrated in many cases where such ratings have a significant effect on the offering yield of the bond (West, 1973). Furthermore, regulators commonly accept bond ratings as indicative of suitability for institutional holdings. Consequently, institutional investors, regulatory commissions in the fields of banking and insurance, corporate financial managers, and bond issuers have come to rely upon corporate bond ratings. More extensive discussion on the importance of bond ratings is provided by Hickman (1958), Pogue and Soldofksy (1969), West (1973), Ross (1976), and Kaplan and Urwitz (1979).

As part of the rating process, the rating agencies use financial variables and particularly emphasise the importance of the subjective judgement of the analyst in determining the rating of the bond. Indicators measuring company's performance such as leverage, coverage and profitability are considered by analysts as prime determinants of the quality of the bond. On the other hand, a number of researches have developed statistical models in predicting bond ratings. Statistical bond rating methods use only the quantifiable financial data of the firm and the financial variables chosen are those which the researchers consider the most appropriate proxies for liquidity, debt capacity, debt coverage, size, variability of earnings, and indenture provisions such as subordination. Rating agencies have claimed that statistical analysis lacks the sophistication necessary to model the expert judgement required to rate bonds. Despite these claims, however, a number of studies have been able to develop a statistical model that performed remarkably well in explaining and predicting the ratings of a large cross section of corporate bonds (Horrigan 1966; West 1970; Pogue and Soldofky 1969; Pinches and Mingo 1973, 1975; Altman and Katz 1976; Ang and Patel 1975; Kaplan and Urwitz 1979; Dutta and

Shekhar 1988; Singleton and Surkan 1994). Earlier studies such as those by Horrigan (1966), West (1970), Pogue and Soldofky (1969), Pinches and Mingo (1973, 1975), and Altman and Katz (1976) reported that variables that have been found significant to predict bond ratings include among others total assets, subordination, debt to equity ratio, debt coverage, debt capacity, turnover, measures of earnings, and liquidity. On the other hand, more recent studies such as those by Dutta and Shekhar (1988) and Singleton and Surkan (1994) reported that variables that have been found important in determining bond ratings include among others debt proportion, turnover, financial strength, earnings, past five-year revenue growth rate, projected next five-year revenue growth rate, working capital, logarithm of total assets, and the subjective prospect of the company.

Considering the more recent studies in bond ratings prediction, we can easily conclude that most of the variables that have been found important in predicting bond ratings can be viewed either as indicators of growth or as indicators of company performance. In view of these considerations, in our study we selected many of the variables that were found important in previous studies to predict bond ratings. Apart from these variables, however, we also examined the predictive ability of a few other variables that have not received considerable attention in previous studies. In a series of preliminary experiments, several combinations of the variables were attempted to predict bond ratings. After these experiments, a combination of six performance ratios and three growth indicators were finally selected to be included in our study. The performance ratios that we selected for our study can be described as follows: Return on Capital Employed (ROCE), Profit Margin (PM), Total Debt/Equity (TDE), Earnings Yield (EY), Dividend Yield (DY), and Price-Earnings (P/E) ratio. We also selected three growth indicators, namely: Total Assets (TA), Turnover (S), and Profit Before Tax (PBT). These growth indicators were calculated as a percentage change between the beginning and end of the year. We selected the above ratios in order to assess the impact of both the performance and the growth of the company on the bond rating. The company data were collected from the EXTEL service.

We have to mention that some variables that have been found important in previous studies to predict bond ratings such as subordination, interest coverage and liquidity were also attempted in our experiments but they were not found important. A statistical explanation for this result might be that the multivariate combination of our variables could make redundant some of the variables that were found important in previous studies. We should also consider that previous studies have used different combinations of variables as well as different data sets. On the other hand, it is highly possible that some researchers might have used a different methodology in calculating a specific variable, while they are not giving a lot of details of this calculation given

the practical importance of their studies.

Three boundary bond ratings were chosen according to the rating definitions given by the S&P's bonds guide: 1) AAA: the highest rating assigned – capacity to pay interest and principal very strong; 2) AA: very strong capacity to pay interest and principal – differ from highest rated issues only in small degree; and 3) A: strong capacity to repay interest and principal but may be susceptible to adverse changes in economic conditions. The bond ratings data were collected from the S&P's bonds guides for the years 1991-1997 (Standard and Poor's, 1991-97).

One problem in practical applications of classification methods concerns the normalisation of the inputs. For some classification methods such as PNN, the input variables must be normalised so as to be commensurate with the loss functions practical limits. Although normalisation is not always a necessary prerequisite for LDA and RRI, we decided that the comparison of the models might be more fair after applying the same data preprocessing technique for each individual classification method. Our preliminary results suggested that normalisation of the data does not affect the discriminating power of methods such as LDA and RRI whose loss functions do not depend necessarily upon normalisation. On the opposite side, we found some evidence of improvements in accuracy after normalising the data for RRI. Therefore, the accounting data were normalised in the range $(0,1)$ by applying the following transformation,

$$x_{git}^{n} = \left(\frac{x_{git} - \min_{gkt}}{\max_{gkt} - \min_{gkt}} \right) \tag{4.1}$$

where x_{git}^{n} is the i^{th} normalised observation of the g^{th} variable at time t, x_{git} is the i^{th} raw observation of the g^{th} variable at time t, $\min_{gkt}$ is the lowest value of the g^{th} variable for all k (k=1,2..N) observations of this variable at time t, and $\max_{gkt}$ is the highest value of the g^{th} variable for all k observations of this variable at time t.

The most common solution in order to test the performance of the classifier is to save out a sample of the known patterns from the training set and use the remaining patterns to build the classification rule which we can use in turn to classify the left-out patterns. The classifier's performance on this test set is a better measure of what its performance will be in the general population. However, if the training set is small, then it will be wasteful to save out part of it. An important subset of the population may be excluded from the training. In this case, the

classifier will fail to take these patterns into account and will misclassify them during the test phase (Masters, 1995). A better way to solve this problem is to use the leave-one-out method that we described in Section 3.9. According to this method, a single pattern is held out from the training set and the remaining $n-1$ patterns are used to build the classification rule which is then used to classify the pattern that was left-out. In the next step, this single pattern is returned into the training set and a different pattern is removed. The classifier is then retrained using the remaining $n-1$ patterns and it is tested on the new pattern that was left-out. By repeating this process for any individual pattern in the training set, every known pattern is used to both training and testing. The proportion of patterns from each class that are misclassified after applying the particular classification rule gives a direct estimate of the actual error rate in that class.

The LDA, PNN, and RRI classification methods were first implemented to classify bond issues of 132 U.K. industrial companies. Our data set consists of 18 triple AAA, 45 double AA, and 69 single A rated bonds. All selected bonds have different maturity dates. In order to test the ability of the models to classify bonds into boundary rating classes if the data is small, we excluded randomly 52 patterns from the initial data set and we implemented the models using the remaining 80 patterns. In this second implementation, the data set consists of 16 triple AAA, 22 double AA, and 42 single A rated bonds.

We applied the LDA, the PNN, and the RRI methods to classify bonds into the three classes of bond ratings at the same time as well as into one class against the other. A variety of tests for multivariate normality have been proposed in the literature (Malkovich and Afifi, 1973). A simple tactic is to examine the distributions of each of the variables individually. If any of the variables have markedly non-normal distributions, then there is reason to suspect that the assumption of multivariate normality is violated.

A number of transformations of the variables were performed in order to meet the assumption of normality in the distributions of the financial ratios that we selected for the implementation of LDA, PNN, and RRI. Unfortunately, none of these transformations was always found appropriate to achieve normality and symmetry in the distribution of each individual variable that we selected for this study. Finally, we decided to choose a trigonometric transformation of the variables because it was more effective than the other transformations in achieving a high degree of improvement in the distributions of the financial ratios that did not meet the assumption of normality. This transformation is shown below,

$$x^{t}_{git} = \arcsin\left(\sqrt{\frac{x^{n}_{git}}{\sum_{i=1}^{N} x^{n}_{git}}}\right) \tag{4.2}$$

where x^{t}_{gti} is the i^{th} normalised transformed observation of the g^{th} variable at time t, x^{n}_{git} is the i^{th} normalised observation of the g^{th} variable, and N is the total number of observations of the g^{th} variable at time t. The transformation presented in Eq. (4.2) is a variation of the transformation that was proposed for count variables from small populations. The original transformation can be found in Acton (1959).

The common density estimator that we used for the implementation of the PNN model is given below,

$$\tilde{f}^{j}(x) = \sum_{i=1}^{n} e^{-\frac{(x-x_{ji})^2}{2\sigma}} \tag{4.3}$$

In order to optimise the value of σ, we simply selected many values and chose the one that performed better on the training set.

As far as concerns the RRI algorithm, we computed the TDL of the rule set and the examples after adding a rule[1]. We then simplified the rule set by examining each rule in turn and kept deleting rules so as to reduce the TDL. For each rule in turn, we constructed a replacement rule and a revised rule and we used the MDL heuristic to decide whether the final theory should include the revised rule, the replacement rule, or the original rule. In the final step, we added rules to cover any remaining positive examples as suggested by Cohen (1993, 1995).

4.2 RESULTS

Table 4.1 compares the classification results of LDA, PNN, and RRI after applying the leave-one-out method. Bonds are classified into one of the classes: AAA, AA and A. The data set consists of 69 single A, 45 double AA, and 18 triple AAA rated bonds. The diagonal elements are the number of bonds classified correctly into the classes. The off-diagonal elements are the numbers of bonds classified incorrectly into classes. The overall percentage refers to the total

[1] We would like to thank W. Cohen for very kindly providing the source code for the implementation of the RRI classifier. This code was slightly modified for the purpose of our studies.

percentage of correct classifications regardless of the type of rating. Applying the LDA, we found that 21 out of 69 bonds rated as A classified correctly as A, 31 bonds out of 45 rated as AA classified correctly as AA, and 14 bonds out of 18 rated as AAA classified correctly as AAA. Applying the PNN, we found that 41 bonds out of 69 rated as A classified correctly as A, 37 bonds out of 45 rated as AA classified correctly as AA, and 17 bonds out of 18 rated as AAA classified correctly as AAA. Finally, applying the RRI, we found that 52 bonds out of 69 rated as A classified correctly as A, 33 bonds out of 45 rated as AA classified correctly as AA, and 13 bonds out of 18 rated as AAA classified correctly as AAA. Overall, the LDA classified correctly 50% of the test patterns, the PNN classified correctly 71.96% of the test patterns, and the RRI classified correctly 74.24% of the test patterns.

Applying the $X^2 - Test$ for normality we found that the distributions of the financial ratios selected for this study were non-normal. We therefore applied the trigonometric transformation presented in Eq. (4.2) to ensure normality and symmetry in the distributions of the financial ratios and we then applied the $X^2 - Test$ again. The results of this test are presented in Table 4.2. As we can see in Table 4.2, even after variable transformation we were not able to achieve normality for the financial ratios selected for the implementation of the models. Therefore, we have reason to suspect that the assumption of multivariate normality is violated. However, the histograms presented in Figures 4.1(a-r) show a high degree of improvement in the distributions of the financial ratios that we selected for the implementation of LDA, PNN, and RRI.

Table 4.3 compares the classification results of LDA, PNN, and RRI after variable transformation and after applying the leave-one-out method. Overall, the LDA classified correctly 56.1% of the test patterns, the PNN classified correctly 72.72% of the test patterns, and the RRI algorithm classified correctly 74.24% of the test patterns. Comparing the results in Tables 4.1 and 4.3, we can see that the transformation of the variables improved the classification performance of the LDA to a higher degree than the PNN but it did not affect the classification performance of the RRI.

Extending the above analysis, we classified bonds into the one of two classes each time instead of classifying bonds into one of the three classes directly. This involves three pairs of classes A against AA, A against AAA, and AA against AAA so that each bond is classified in a class one class against the other. Table 4.4 compares the classification results of LDA, PNN and RRI when each bond is classified in a class one class against the other after applying the leave-one-out method. A Type I error (T1) is one in which bonds on the left of the first column were classified incorrectly. A Type II error (T2) is one in which bonds on the right of the first column

were classified incorrectly. For example, the first row of Table 4.4 shows that 21 out of 69 bonds rated as A classified incorrectly as AA, and 11 out of 45 bonds rated as AA classified incorrectly as A, if the LDA is applied. Applying the PNN, we found that 14 out of 69 bonds rated as A classified incorrectly as AA, and 6 out of 45 bonds rated as AA classified incorrectly as A. Finally, applying the RRI, we found that 13 out of 69 bonds rated as A classified incorrectly as AA, and 5 out of 45 bonds rated as AA classified incorrectly as A. Total error (T) refers to the total incorrect classifications regardless of type.

Table 4.5 compares classification results of LDA, PNN and RRI expressed as percentage of success. As we can see, the PNN model and the RRI algorithm outperformed the LDA model for three classes at the same time as well as for one class against the other. The RRI, however, performed better than the PNN.

To test the hypothesis that the percentage of accuracy of LDA for a given class is greater than or equal to the percentage of accuracy of PNN and RRI, we performed a $Z-Test$. Let us denote $\tilde{P}_{LDA_{it}}$ (i=56.1%, 71.92%, 70.11%, and 84.12%) the percentages of bonds classified correctly into the three classes at the same time as well as A against AA, A against AAA, and AA against AAA, respectively, if the LDA is applied, $\tilde{P}_{PNN_{it}}$ (i=72.72%, 82.45%, 82.75%, and 93.65%) the percentages of bonds classified correctly into the same classes if the PNN classifier is applied, and $\tilde{P}_{RRI_{it}}$ (i=74.24%, 84.21%, 86.20%, and 93.65%) the percentages of bonds classified correctly into the same classes if the RRI classifier is applied. The test statistic used to test the hypothesis that the percentage of accuracy of the LDA for a given class is greater than or equal to the percentage of accuracy of the PNN classifier is given below,

$$Z_{1jt} = \frac{\left(\tilde{P}_{PNN_{it}} - \tilde{P}_{LDA_{it}}\right) - 0}{\left(\frac{\tilde{P}_{PNN_{it}}\tilde{Q}_{PNN_{it}} + \tilde{P}_{LDA_{it}}\tilde{Q}_{LDA_{it}}}{n}\right)^{0.5}} \qquad (4.4)$$

where $\tilde{P}+\tilde{Q}=1$ and n is the number of bonds.

The test statistic used to test the hypothesis that the percentage of accuracy of the LDA for a given class is greater than or equal to the percentage of accuracy of the RRI classifier is given below,

$$Z_{2jt} = \frac{\left(\tilde{P}_{RRI_{it}} - \tilde{P}_{LDA_{it}}\right) - 0}{\left(\frac{\tilde{P}_{RRI_{it}}\tilde{Q}_{RRI_{it}} + \tilde{P}_{LDA_{it}}\tilde{Q}_{LDA_{it}}}{n}\right)^{0.5}} \qquad (4.5)$$

where $\tilde{P} + \tilde{Q} = 1$ and n is the number of bonds.

Since the calculated Z_{1jt} values (2.86, 1.90, 1.98, and 1.72 for three classes at the same time as well as for A against AA, A against AAA, and AA against AAA, respectively) are greater than the critical value (1.64) at the 0.05 level of significance, we reject the null hypothesis that the LDA performed better than the PNN. On the other hand, since the calculated Z_{2jt} values (3.15, 2.26, 2.61, and 1.72 for three classes at the same time as well as for A against AA, A against AAA, and AA against AAA, respectively) are greater than the critical value (1.64) at the 0.05 level of significance, we reject the null hypothesis that the LDA performed better than the RRI.

Tables 4.6 and 4.7 show examples of the rules learned and tested by the RRI classifier using the variables after transformation. We have to mention, however, that the RRI is not an expert system, even though it does learn rules. Expert systems are collections of rules written by a human knowledge engineer, while RRI's rules are learned automatically from a database. Since rules are overlapping the rating that the RRI finally predicts for a bond is a combination of the predictions made by each matching rule.

To test the ability of the models to classify bonds into boundary rating classes if the data set is small, we excluded randomly 52 patterns from the data set and we implemented the models using the remaining 80 patterns. Table 4.8 compares the classification results of LDA, PNN, and RRI if bonds are classified into one of the three classes and after applying the leave-one-out method. Table 4.9 shows the Type I and Type II errors of the models when each bond is classified in a class one class against the other. As we can see in Tables 4.8 and 4.9, the PNN outperformed both the LDA and the RRI if the sample size (SS) reduces from 132 to 80 patterns. Figures 4.2-4.4 compare the results in Tables 4.3, 4.4 and 4.8, 4.9 for the LDA, the PNN, and the RRI, respectively. As we can see in these figures, the small SS affects the classification performance of LDA and RRI in a greater degree than the PNN. However, the PNN outperforms the LDA and the RRI if the sample size reduces from 132 to 80 patterns.

4.3 SUMMARY AND CONCLUSIONS

In this Chapter, we compared and contrasted the LDA, the PNN, and the RRI in terms of their ability to classify long-term bond ratings using financial ratios. We found that the PNN, which is a non-linear classifier, and the RRI, which is a non-linear rule induction classifier, not only significantly outperform the LDA, but they are also more robust to different distributional assumptions compared to LDA which is affected from the assumption of multivariate normality.

The final outcome of this experiment is a new quantitative system to predict long-term bond ratings using probabilistic neural networks and rule induction techniques as an alternative to LDA. Our results suggest that the PNN should be implemented to classify bond ratings if the data set is relatively small because it does not suffer from the assumption of multivariate normality in the distributions of the financial ratios in the same degree as the LDA.On the other hand, the RRI should be implemented if the data set is relatively large because not only it classifies better than the LDA and the PNN classifiers, but it also offers interpretable rules that are easy to visualise and understand, as opposed to the inherent inability of the PNN to explain in a comprehensible form the process by which a given decision generated by the model has been reached.

Overall, the findings of this experiment support the evidence for the existence of non-linearities and other complex processes in the financial data. No model is perfectly robust and small variations in the sample size may affect the overall performance of any given model. Therefore, rather than applying a single model for any specific financial application, it is more preferable and wise to use a variety of different models to deal with different sample sizes, unpredictable non-linearities, and inconsistent patterns in the financial data.

In the next Chapter, we apply the classifiers we applied in this experiment as well as other non-linear classifiers to identify high performing shares which is the central idea in this thesis.

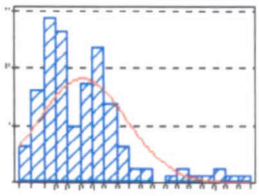
		LDA			PNN			RRI		
Actual Class	Patterns	Predicted Class Membership								
	n=132	A	A A	A A A	A	A A	A A A	A	A A	A A A
A	69	21	21	27	41	10	18	52	9	8
AA	45	10	31	4	8	37	0	7	33	5
AAA	18	0	4	14	1	0	17	2	3	13
overall(%)		50 %			71. 96 %			74.24 %		

Table 4.1: Classification results of LDA, PNN, and RRI after applying the leave-one-out method to predict long-term bond ratings - all classes

Variables	Before Transformation		DF	p-value	Decision
	calculated value	critical value			
ROCE	56.62	11.34	3	0.000	reject
PM	153.65	11.34	3	0.000	reject
TDE	105,727,25.0	11.34	3	0.000	reject
EY	11.77	11.34	3	0.008	reject
DY	33548.15	11.34	3	0.000	reject
P/E	20.27	11.34	3	0.000	reject
TA	40545	11.34	3	0.000	reject
S	33.06	11.34	3	0.000	reject
PBT	76.84	11.34	3	0.000	reject
Variables	After Transformation		DF	p-value	Decision
	calculated value	critical value			
ROCE	6.09	11.34	3	0.100	**accept**
PM	55.47	11.34	3	0.000	reject
TDE	2735.96	11.34	3	0.000	reject
EY	2.61	11.34	3	0.450	**accept**
DY	38.10	11.34	3	0.000	reject
P/E	1.28	11.34	3	0.730	**accept**
TA	-0.92	11.34	3	0.000	reject
S	88.83	11.34	3	0.000	reject
PBT	10.13	11.34	3	0.014	**accept**

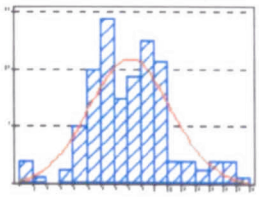
Table 4.2: X^2 - Test for normality

4.1(a): Histogram before transformation



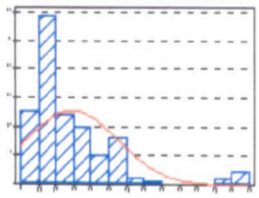
ROCE

4.1(b): Histogram after transformation



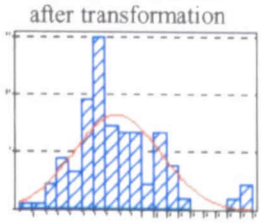
ROCE

4.1(c): Histogram before transformation



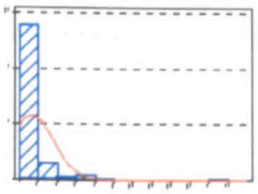
PM

4.1(d): Histogram after transformation



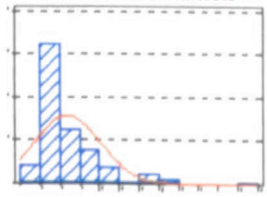
PM

4.1(e): Histogram before transformation



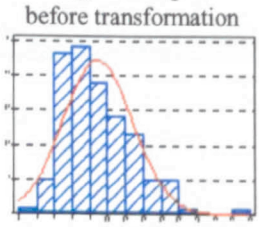
TDE

4.1(f): Histogram after transformation



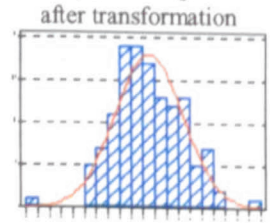
TDE

4.1(g): Histogram before transformation



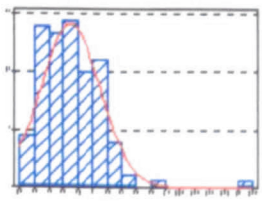
EY

4.1(h): Histogram after transformation



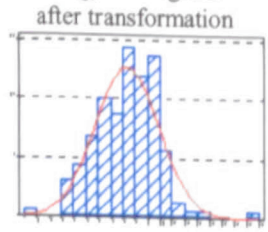
EY

4.1(i): Histogram before transformation



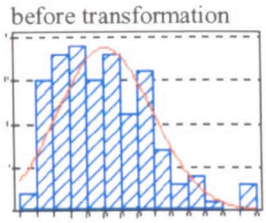
DY

4.1(j): Histogram after transformation



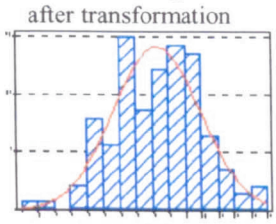
DY

4.1(k): Histogram before transformation



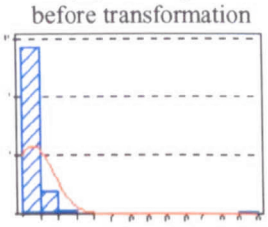
P/E

4.1(l): Histogram after transformation



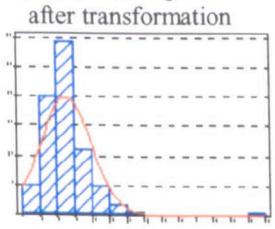
P/E

4.1(m): Histogram before transformation



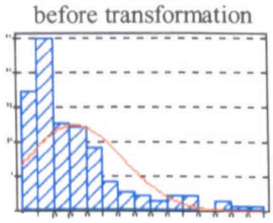
TA

4.1(n): Histogram after transformation



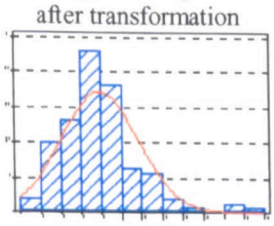
TA

4.1(o): Histogram before transformation



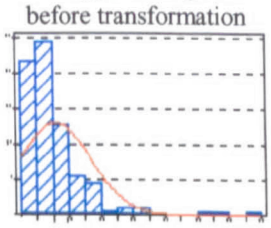
PBT

4.1(p): Histogram after transformation



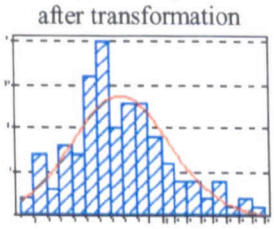
PBT

4.1(q): Histogram before transformation



S

4.1(r): Histogram after transformation



S

Figures 4.1(a-r): Histograms of the financial ratios that we used to predict long-term bond ratings before and after variable transformation

		LDA			PNN			RRI		
Actual Class	**Patterns**	**Predicted Class Membership**								
	n=132	A	AA	AAA	A	AA	AAA	A	AA	AAA
A	69	31	18	20	41	9	19	52	9	8
AA	45	10	29	6	6	38	1	7	33	5
AAA	18	1	3	14	1	0	17	2	3	13
overall(%)		**56.1 %**			**72.72 %**			**74.24 %**		

Table 4.3: Classification results of LDA, PNN, and RRI after variable transformation and after applying the leave-one-out method to predict long-term bond ratings - all classes

		LDA			PNN			RRI		
Classes	**Patterns**	**T1**	**T2**	**T**	**T1**	**T2**	**T**	**T1**	**T2**	**T**
A ↔ AA	69 ↔ 45= 114	21	11	**32**	14	6	**20**	13	5	**18**
A ↔ AAA	69 ↔18 = 87	24	2	**26**	1	14	**15**	2	10	**12**
AA ↔ AAA	45 ↔18 = 63	7	3	**10**	4	0	**4**	3	1	**4**

Table 4.4: Classification results of LDA, PNN, and RRI after variable transformation and after applying the leave-one-out method to predict long-term bond ratings - one class against the other

Classes	LDA	PNN	RRI
AAA ↔ AA ↔ A	56.1 %	72.72 %	74.24 %
A ↔ AA	71.92 %	82.45%	84.21%
A ↔ AAA	70.11 %	82.75 %	86.20 %
AA ↔ AAA	84.12 %	93.65%	93.65 %

Table 4.5: (%) Classification results of LDA, PNN, and RRI after variable transformation and after applying the leave-one-out method to predict long-term bond ratings

Extract Rules...

Rule 1: P/E ≥ 0.2532, TA ≥ 0.0195, DY ≥ 0.1066, TDE ≤ 0.0137
Rule 2: DY ≤ 0.0954, PBT ≤ 0.1084, TDE ≤ 0.0273
Rule 3: PBT ≥ 0.2945, PBT ≤ 0.3355, TA ≤ 0.0335, TDE ≤ 0.0285
Rule 4: DY ≤ 0.0912, DY ≥ 0.0905, TDE ≤ 0.0294
Rule 5: P/E ≤ 0.2278, TDE ≤ 0.3373
Rule 6: DE ≥ 0.2893, TDE ≤ 1
Rule 7: DY ≤ 0.1365, TDE ≤ 0.0225
Rule 8: P/E ≤ 0.3263, TDE ≤ 0.0190
Rule 9: true (50/1).

Table 4.6: Examples of the rules learned by the RRI and used to predict bond ratigs

Test the Rules...

						A	AA	AAA
Rule 1:	**9**	**Bs**	**rated as**	**AAA**	**classified as**	**0**	**0**	**9**
Rule 2:	**4**	**Bs**	**rated as**	**AAA**	**classified as**	**1**	**0**	**3**
Rule 3:	**1**	**Bs**	**rated as**	**AAA**	**classified as**	**0**	**0**	**1**
Rule 4:	**2**	**Bs**	**rated as**	**AAA**	**classified as**	**0**	**0**	**2**
Rule 5:	**46**	**Bs**	**rated as**	**AA**	**classified as**	**14**	**31**	**1**
Rule 6:	**5**	**Bs**	**rated as**	**AA**	**classified as**	**0**	**5**	**0**
Rule 7:	**8**	**Bs**	**rated as**	**AA**	**classified as**	**2**	**6**	**0**
Rule 8:	**3**	**Bs**	**rated as**	**AA**	**classified as**	**2**	**1**	**0**
Rule 9:	**54**	**Bs**	**rated as**	**A**	**classified as**	**50**	**2**	**2**

Table 4.7: Classification results of the rules used by the RRI to predict long-term bond ratings after applying the leave-one-out method

		LDA			PNN			RRI		
Actual Class	**Patterns**	**Predicted Class Membership**								
	n = 80	A	AA	AAA	A	AA	AAA	A	AA	AAA
A	42	7	13	22	28	5	9	27	6	9
AA	22	1	17	4	4	18	0	2	17	3
AAA	16	3	3	10	2	0	14	1	2	13
overall(%)		**42.5 %**			**75 %**			**71.25 %**		

Table 4.8: Classification results of LDA, PNN, and RRI after applying the leave-one-out method to predict long-term bond ratings if the dataset is small- all classes

		LDA			PNN			RRI		
Classes	**Patterns**	**T1**	**T2**	**T**	**T1**	**T2**	**T**	**T1**	**T2**	**T**
A ↔AA	42 ↔ 22 = 64	13	5	18	6	4	10	8	4	12
A ↔AAA	42 ↔16 = 58	14	2	16	3	8	11	6	6	12
AA ↔AAA	22 ↔16 = 38	3	1	4	2	0	2	3	1	4

Table 4.9: Classification results of LDA, PNN, and RRI after applying the leave-one-out method to predict long-term bond ratings if the dataset is small - one class against the other

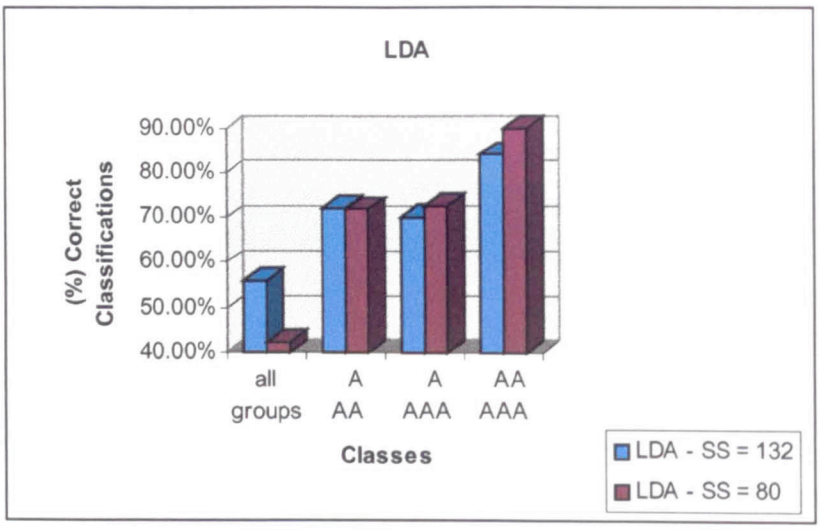


Figure 4.2: The sample size (SS) effect on the classification performance of the LDA to predict long-term bond ratings after applying the leave-one-out method

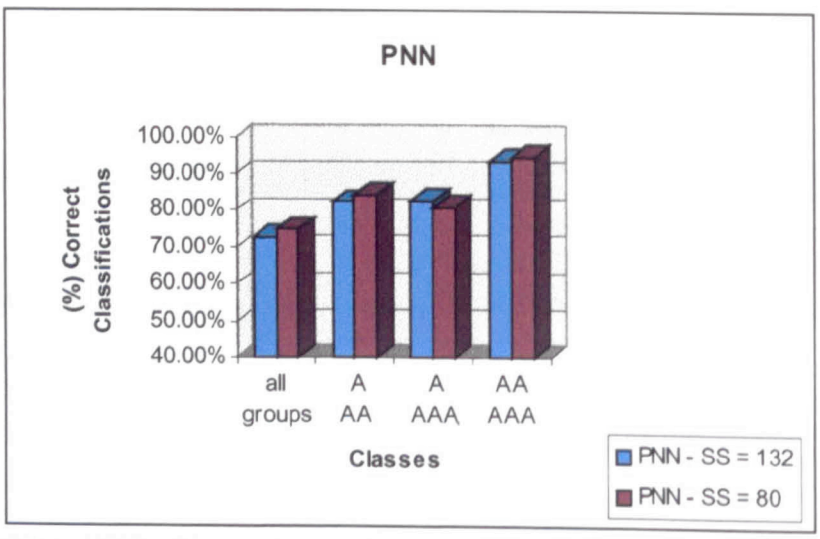


Figure 4.3: The sample size (SS) effect on the classification performance of the PNN to predict long-term bond ratings after applying the leave-one-out method

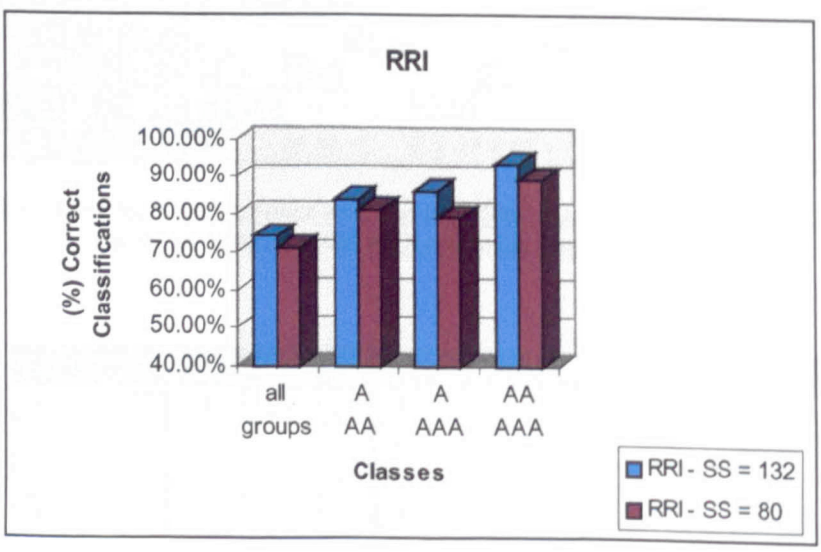


Figure 4.4: The sample size (SS) effect on the classification performance of the RRI to predict long-term bond ratings after applying the leave-one-out method

CHAPTER 5: PREDICTING HIGH PERFORMANCE STOCKS USING FIVE STATISTICAL CLASSIFICATION ALGORITHMS

In this Chapter, we present the first implementation of the central idea in this thesis which is a new methodology to identify outperforming shares using five classification methods namely, the Linear Discriminant Analysis (LDA), the Probabilistic Neural Network (PNN), the Learning Vector Quantization (LVQ), the Oblique classifier (OC1), and the Ripper-Rule Induction (RRI) classifier. These classifiers have been chosen as representatives of five different model families: the LDA is a well-known linear classifier, the PNN is a kernel density non-linear classifier, the LVQ is a vector quantization classifier that also gives a non-linear partitioning of the data, the oblique OC1 classifier is a recursive partitioning classifier, and the RRI is a rule induction algorithm that depends on the induction of logical rules. We use these classifiers to explore the potential for identifying high performing shares on the London Stock Exchange. The model inputs are 38 accounting ratios for around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. We compared and contrasted the classifiers in terms of classification accuracy and profitability in the target years 1993-97. We found that all classification methods produce consistent excess returns in ex ante forecasting, whereas there are some inconsistencies in the classification accuracy and profitability of the classifiers from one year to the next.

We organise this Chapter as follows: In Section 5.1, we discuss the data and trading rules that we used in our study. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This data consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our predictor variables are 38 accounting ratios drawn from published accounting statements. In Section 5.2, we discuss the methodology that we used to implement our classification methods. To make the classification methods more robust, we applied data preprocessing techniques such as data normalisation and stepwise variable elimination procedures to identify the best subset of variables for each individual classification method. We compared and contrasted the classifiers in terms of classification accuracy and profitability and we also considered the trade-off between predicted returns and risk. In Section 5.3, we report the results of our experimentation. We found that all classification methods produce consistent excess returns in ex ante forecasting, whereas there are some inconsistencies in the classification accuracy and profitability of the classifiers from one year to the next. In Section 5.4, we summarise our empirical results and we provide the conclusions.

5.1 DATA AND TRADING RULES

Our target data are total returns on all shares that are traded on the London Stock Exchange in the years 1993-97. The total share returns data cover the period 1991-97. A share was included in our sample only if there was a complete set of annual accounts available on the EXTEL service as well as a date of publication of the company's annual report. We calculated the share returns on an annual basis using end-month price data adjusted for dividends received during the year. The price data were collected in the month of publication of the company's annual report with different companies having different reporting cycles. The predictions were made for the performance of these companies' shares over 1-year holding periods during the years 1993-97. Using the price data, we calculated excess returns for each share by subtracting the corresponding total return on an equally-weighted index of all sampled shares from the individual share return. The share prices data were collected from the DATASTREAM service.

The reason of using an equally-weighted index of all sampled shares to benchmark the profitability of our classification methods should be made particularly clear. As we mentioned above, a share was included in our sample only if there was a complete set of annual company accounts available on the EXTEL service as well as the date of publication of the company's annual report. After a very extensive investigation, we were able to find the date of publication of the companies' annual reports for a sample of around 700 companies per year starting with 626 companies in 1993 and rising up to 718 companies in 1997. In deciding our benchmark, we had to consider that the random sample of shares we included in our study was substantially smaller than the number of all shares that were traded on the London Stock Exchange during the 1993-97 period (over 1400 shares as reported in the DATASTREAM service for the 1993-97 period). Therefore, it might be unfair to compare the profitability of our trading system against a much more general benchmark – the All Shares index, for example – consisting of all shares traded on the London Stock Exchange during the 1993-97 period. On the other hand, an equally-weighted index consisting of 700 sampled shares on average is a very fair benchmark to evaluate the profitability of our trading system. From a trading perceptive the attempt to identify a small subset of shares from an initial sample of 700 shares and investigate if the smaller subset can outperform the whole sample in terms of profitability seems entirely sensible. On the other hand, we also have to consider that our sample contains a number of small capitalisation stocks. Therefore, once again it might be unfair to compare the performance of our trading

system with some other benchmarks, such as the FTSE 100 index, for example, which refers to large capitalisation stocks. Although it would be very desirable, if our trading system outperforms every single benchmark, the benchmark that we use in our study is both fair and reasonable from a trading perspective.

We also have to mention that there are other ways we can benchmark our classification methods. For example, one way is to benchmark our classification models against level estimation methods as, for example, in Leung et al. (2000). Although it would be very interesting to benchmark our classification methods against level estimation methods, we have to mention that this not the focus of our study. As we mentioned in Chapter 1 (Section 1.2), the focus of this thesis is to predict events in financial markets by using individually and combining a "portfolio" of five heterogeneous classifiers namely LDA, PNN, LVQ, OC1 and RRI using majority voting (MV) and unanimous voting (UV) techniques that are described in more detail in the experiments presented in the following Chapters. The general hypothesis that we investigate in this thesis is that the ability of any classifier to predict high performing shares can be improved or exceeded through composite classifier architectures that combine a small number of heterogeneous classifiers using voting procedures. Therefore, in this thesis we investigate if classification algorithms from different model families can be either combined or applied individually to correctly classify and predict which shares are likely to have exceptional returns in the future. Furthermore, we investigate what types of information we have to use in order to increase the accuracy of either an individual classifier or a composite classifier architecture to correctly classify and predict which shares are likely to have exceptional returns in the future. Finally, we investigate what data preprocessing techniques we have to undertake in order to improve the classification accuracy and reduce the computational expense of the proposed algorithms. Therefore, a comparison of our classification methods relative to level estimation methods is not relevant with the focus of this thesis even though this comparison may prove that our classification methods outperform level estimation techniques. Certainly, this investigation could be another subject of future research. We also have to emphasise that our non-linear classification methods are always benchmarked against a more traditional method such as LDA.

The central task of our classification methods is to predict whether or not a share will be high- or low-performing share for the out-of-sample years 1993-97. The distinction between high- and low performing shares is based on the excess returns of the shares over our equally-weighted benchmark index. Shares with excess returns in the top 25% in a given year are marked as high-performing shares (H) and shares in the bottom 75% in the same year are marked as low-performing shares (L). The 25% cut-off point between H and L performing

shares is decided so as to achieve a reasonable difference in the mean returns on the H and L classes and at the same time ensure a reasonably large sample size for these classes. After performing a few preliminary experiments on different cut-off points between H and L classes (i.e. 15%, 20%, and 30%), we found that there was no improvement over the 25% cut-off point as far as the classification accuracy and profitability of our classification methods are concerned. However, a few other scenarios should also be mentioned. From a trading perspective, for example, it might have been better to classify as L shares those with excess returns in the bottom 25% which would have enabled long/short strategies, where not only the high performing shares are bought, but also the low performing shares are sold. Although this strategy might be more appropriate from a trading perspective, it might not be particularly appropriate from an implementation perspective for a number of reasons: If we had to classify as L shares those with excess returns in the bottom 25% quartile and as H performing shares those in the upper 25% quartile, then for each particular out-of-sample year, we would have to predict an additional class of shares – the shares between the upper 25% and the bottom 25% quartiles. It is reasonable to assume that the classification performance as well as the profitability of our classification methods would be affected negatively if we were applying them to predict three classes rather than two. In fact, a few preliminary experiments verified this assumption. Another possible scenario would be to attempt to classify as H shares those with excess returns in the upper 75% quartile and as L shares those with excess returns in the bottom 25% quartile. Although this strategy might be appropriate from an implementation perspective, it might not be particularly feasible from a trading perspective because it would improve the profitability of the short side but it would decrease the profitability of the long size (lower return stocks would be included in the portfolio of high performing stocks). It is obvious that these as well as other similar strategies could be investigated to enable the performance of our trading system by incorporating long/short strategies. However, the time frame that we had available to complete our empirical experiments did not allow us to investigate all possible scenarios. A few preliminary experiments showed the general methodology that we had to follow in order to achieve the best results given the time limitations. Certainly, this investigation could be another subject of future research.

To predict whether a particular share will be H or L in a given year, we collected previous years of accounting information published in the annual accounts of the companies and we calculated 38 accounting indicators. A detailed list of the variables that were selected to implement the classification methods in order to classify a new observation as H performing or L performing share over 1-year holding periods is given in Table 5.1. The idea of collecting these variables in the month of publication of the company's annual report was to develop a trading system that will be able to incorporate the impact of the public announcement of the accounting information

to the share price and examine the interaction of this information with other types of information such as economic information. The interaction between accounting and non-accounting information is examined in Chapter 7.

Using the accounting variables and the share returns data, we aimed to find rules that classify a particular share as H or L performing share. We used five classification methods namely, LDA, PNN, LVQ, OC1 and RRI. To implement these classifiers, we used two years of data to predict the next year. For example, to predict relative excess returns for 1993, we first trained the classification methods on the two preceding years 1991 and 1992 using cross-validation procedures, such as the leave-one-out method and other rotation procedures, to find the optimal values of parameters for each individual classifier. After selecting the optimal values of parameters, we applied the classification methods to predict the out-of-sample year 1993. We then moved the implementation one-year ahead and we used information available from 1992 and 1993 to predict 1994 and so on. We use only two previous years of data to predict relative excess returns for the next year because we believe that only recent accounting information may be relevant to predict relative excess returns in the next year.

5.2 METHODOLOGY

In our application, we are particularly interested in whether a particular share will be classified as H or L excess return share based on accounting information. Let us assume that y_{it} is the 1-year-ahead excess return on some share i bought at time t, and x_{it} is the vector of accounting information attributes for company i known at time t. The idea is to apply a classification method to assign y_{it} to one of the two classes C_{it} = H or L depending on whether or not this return is above or below the 25% threshold percentile that has been decided after ranking the returns in excess of an equally-weighted index. The models input is the vector x_{it} of variables that represent current month accounting information.

One of the desirable end products of discriminant analysis is identification of good predictor variables. A standard methodology that has been applied in the literature to identify an optimal subset of ratios for the LDA model is based on a stepwise variable selection algorithm using criteria which eliminate overfitting of variables, thereby improving the model's classification of out-of-sample patterns and its robustness over time. According to this method, the first variable included in the analysis has the largest acceptable value for the selection criterion. After the first variable is entered, the value of the criterion is reevaluated for all variables not in the model, and the variable with the largest acceptable criterion value is entered next. At this point, the

variable entered first is reevaluated to determine whether it meets the removal criterion. If it does, it is removed from the model. Variable selection terminates when no more variables meet entry or removal criteria. The criterion used for stepwise variable selection is minimisation of Wilks' lambda. The significance of the change in Wilks' lambda when a variable is entered or removed from the model is based on an F- transformation.

In contrast with LDA, there are few techniques in the production of PNN, LVQ, RRI, and OC1 algorithms which minimise overfitting of variables in an attempt to improve out-of-sample classification and robustness over time. In order to determine optimal ratios for these algorithms, we followed the methodology suggested by Tyree and Long (1996). According to this methodology, the algorithms are implemented using all thirty-eight ratios and the misclassification rate is recorded. In the next step, a single variable is removed and the algorithms are implemented again. If the misclassification rate is lower, the variable is removed permanently; otherwise, it is returned back to the pool of variables to be included in the final model. This procedure is repeated for each individual variable.

The variable elimination procedure was implemented using the years 1991, 1992 to build the models and then using the year 1993 to test them. A detailed list of the variables that were finally selected for the five classification methods after applying the stepwise variable elimination procedure is given in Table 5.2. The subsets of variables presented in Table 5.2 were also used for the next out-of-sample years 1994-97.

The accounting data were normalised in the range $(0,1)$ by applying Eq. (4.1). In this first implementation, we did not transform the accounting data because we thought that it would be a good idea to assess initially the performance of the classification methods using the minimum of data preprocessing. Variable transformations and other data preprocessing techniques were applied in further experiments that are described in the next Chapters.

The common density estimator that we used for the implementation of the PNN model is similar to the one presented in Eq. (4.3). In order to optimise the value of σ, we simply selected many values and chose the one that performed better on the training set.

We implemented the LVQ algorithm using the optimised (oLVQ1) version[2]. In order to determine the initial values of the free parameter vectors, we used samples of the training data

[2] We would like to thank T. Kohonen for very kindly providing the source code for the implementations of the LVQ classifier. This code was slightly modified for the purpose of our studies.

picked up from the respective classes and we accepted the samples that were not misclassified. According to this procedure, we tentatively classified a sample against all the other samples in the training set using a nearest neighbour algorithm and we accepted it as a possible initial value only if this tentative classification was the same as the class of the sample. In the next step, we computed the medians of the shortest distances between the initial free parameter vectors of each class. If these distances were very different for the different classes, then we added new free parameter vectors and we deleted old ones from the deviating classes. After initialising the codebook, we started the training using the oLVQ1 algorithm. However, since we used the oLVQ1 algorithm already once in the initialisation phase, we shorted the training phase by that amount. In total, the number of learning steps was chosen to be 30 to 40 times the total number of free parameter vectors as suggested by Kohonen et al. (1995).

The impurity measure we used to implement the OC1 classifier is the Twoing value[3] (see Eq. 3.29). We applied the OC1 classifier by minimising the reciprocal of this value. In order to avoid local minima, we added a random vector to the coefficients of the current hyperplane to perturb it to a new location and we repeated this procedure a number of times until to find a hyperplane that decreased the overall impurity. However, if no improvement of the impurity was possible, then we halted the algorithm and we used the current hyperplane as the split for the current tree node as suggested by Murthy (1997).

As far as concerns the RRI algorithm, we computed the TDL of the rule set and the examples after adding a rule. We then simplified the rule set by examining each rule in turn and keep deleting rules so as to reduce the TDL. For each rule in turn, we constructed a replacement rule and a revised rule and we used the MDL heuristic to decide whether the final theory should include the revised rule, the replacement rule, or the original rule. In the final step, we added rules to cover any remaining positive examples as suggested by Cohen (1993, 1995).

After implementing the models, we found that all classifiers were more accurate to classify L performing shares than H performing shares. In order to increase the classification accuracy for H performing shares, we changed the loss functions to incorporate prior probabilities and misclassification costs. For example, in the case of the PNN we classified an unknown share with measurement vector x_{it} as H performing share if the following inequality was true:

[3] We would like to thank K.V.S. Murthy for very kindly providing source code for the implementation of the OC1 classifier. This code was slightly modified for the purpose of our studies.

$$\left(\frac{p_H}{p_L}\right)\cdot\left(\frac{c_H}{c_L}\right)\cdot \tilde{f}^H(x) > \tilde{f}^L(x) \qquad (5.1)$$

where p_H is the prior probability for H performing shares, p_L is the prior probability for L performing shares, c_H is the cost of misclassification for H performing shares, and c_L is the cost of misclassification for L performing shares. A similar more or less to the above tactic was followed for the other classifiers as well.

We compared and contrasted the classifiers in terms of classification accuracy and profitability and we also considered the trade-off between predicted returns and risk. According to our trading system, if a share is classified as H, we buy equal amounts of this share at the end of the reporting month and we hold it for one year. The profitability of each classification method is therefore calculated by the cumulative profits generated by the resulting portfolio of H performing shares. The benefit of this approach is that it minimises transaction costs while it is not affected by price fluctuations around the reporting date. Each share is traded at most once per year and trades can be done at the end of the month in a basket of no more than 13-16 shares bought and sold in the ideal trading strategy. On average, the H performing portfolio will turn over 1/12 of its constituents each month.

5.3 RESULTS

In this Section, we present the results of our experimentation. We compare the five classification methods in terms of classification accuracy and profitability in the out-of-sample years 1993-97. We mentioned that we used the target year 1993 to select the best subset of variables. Therefore, it might be more correct to consider the target year 1993 as the test out-of-sample year and the target years 1994-97 as the genuine out-of-sample years. Table 5.3 shows the classification performance of LDA, PNN, LVQ, OC1 and RRI for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97. Applying the LDA to predict H and L performing shares for the target year 1993, we found that 98 out of 163 actual H shares classified correctly as H performing shares, and 360 out of 488 actual L shares classified correctly as L performing shares. Applying the PNN, we found that 96 shares out of 163 actual H shares classified correctly as H performing shares, and 365 shares out of 488 actual L shares classified correctly as L performing shares. Applying the LVQ, we found that 90 out of 163 actual H shares classified correctly as H performing shares, and 334 shares out of 488 actual L shares classified correctly as L performing shares. Applying the OC1, we found that 88 out of 163 actual H shares classified correctly as H performing shares, and 325 out of 488 actual L

shares classified correctly as L performing shares. Finally, applying the RRI, we found that 91 shares out of 163 actual H shares correctly classified as H performing shares, and 339 out of 488 actual L shares classified correctly as L performing shares. Overall, the LDA classified correctly 70.35% of the test patterns, the PNN classified correctly 70.81% of the test patterns, the LVQ classified correctly 65.13% of the test patterns, the OC1 classifier classified correctly 63.44% of the test patterns, and the RRI classified correctly 66.05% of the test patterns for the target year 1993. The classification results for the target years 1993-97 are also presented in Figure 5.1.

As it is shown in Figure 5.1, the PNN and the LDA produced very good results for the target year 1993 but is seems that their classification performance deteriorates for the next years compared to the other classifiers. The OC1 is the best classifier for the target years 1994 and 1995, whereas the LVQ is the best classifier for the target year 1996. This pattern is slightly different for the target year 1997 where the LDA and the PNN have the best classification performance compared to the other classifiers. We have to notice from the results presented in Figure 5.1 that the LDA with 12 inputs produced excellent results for the out-of-sample test year 1993 on which the input dimensionality reduction was conducted, but the performance of the model was not also exceptional for the genuine out-of-sample years 1994-97. This may suggest that the stepwise variable elimination procedure was too parsimonious for the LDA. One the other hand, the non-linear models performed better than LDA for the genuine out-of-sample years 1994-97.

Table 5.3 also shows that a substantial number of L performing shares are incorrectly classified as H in spite the adjustment in the loss function to prefer H performing shares against L performing shares. For example, 128 out of 226 of the shares that are predicted as H are actually low for the LDA for the target year 1993, whereas 163 out of 251 of the shares that are predicted as H are actually low for the OC1 classifier for the same year. Therefore, it may be of practical importance to assess the relative performance of the classifiers in predicting H performing shares alone. These results are presented in Figure 5.2 for the target years 1993-97. As we can see, the LDA and PNN still outperform the other classifiers for the target year 1993, whereas OC1 is still the best classifier for the target year 1994. However, the OC1 is not the best classifier for the target year 1995 since LDA and PNN produce more favourable results for H performing shares this year compared to the other classifiers. On the other hand, the RRI predicts more accurately H performing shares for the target year 1996 compared to LVQ which is the best classifier in terms of overall accuracy for both H and L performing shares. However, LDA and PNN are still the best classifiers for the target year 1997 and outperform the other

classifiers in terms of classification accuracy for H performing shares as well as in terms of overall classification accuracy for H and L performing shares.

Overall, the classification results suggest that the non-linear classifiers except RRI outperformed on average the LDA. The results also show some consistency in the classification performance of LVQ, OC1, and PNN. On the other hand, we have to notice that there is no decay on the overall classification performance of the classifiers for the genuine out-of-sample years 1994-97. On the contrary, the classification rates for all classifiers are better for the genuine out-of-sample year 1997 compared to year 1994.

Although the classification performance is a very important factor to evaluate a particular classifier, it is not the primary concern for this particular application. The ultimate purpose of our trading system is profitability. We therefore compared the average returns and excess returns over the index of the portfolios of actual H and L shares in our data in all the 12-month holding periods starting each year, with the respective average returns and excess returns of the portfolios of H and L shares predicted by the classifiers.

Table 5.4 compares the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the classifiers. These results are also presented in Figures 5.3-5.6 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, all the classifiers produce positive returns and excess returns. PNN and LDA produce very good results for the target year 1993 and outperform the other classification methods. However, the financial returns deteriorate for the target year 1994 even though all classifiers produce positive results. The target year 1995 is a year of high profitability for all classifiers. LDA and PNN produce the highest financial returns for the target year 1995, whereas the other classifiers also produce favourable results. The financial results deteriorate for the target years 1996 and 1997 even though all the classifiers produce positive financial results. However, the RRI classifier seems to be weaker in terms of financial results compared to the other classifiers.

Overall, the financial results suggest that all classification methods outperform the equally-weighted benchmark index and produce consistent excess returns. However, we have to notice that the improvements in classification that are achieved by the non-linear models for the genuine out-of-sample years 1994-97 are not reflected to the financial returns in the same degree and there are only minor inconsistencies in the profitability of the classifiers from one

year to the next. The differences in the profitability of the classifiers are more obvious for the calibration year 1993, whereas these are less obvious for the genuine out-of-sample years 1994-97. However, we have to notice some degree of correlation between classification accuracy and profitability. For example, the LDA and PNN which have the best classification performance for the test out-of-sample year 1993 also produce the highest returns for that year. On the other hand, the LVQ and OC1 which are the best classifiers for the genuine out-of-sample years 1994-97 also produce very favourable returns for these years, whereas the RRI which is the more weak classifier compared to the other classifiers also produces the lowest financial returns.

There are several aspects that we have to consider in evaluating the importance of our trading system. The first aspect is the trade-off between predicted returns and risk. After a careful examination of the results presented in Table 5.4, we have to notice that greater excess returns are achieved in years where the index rises more sharply. For example, greater excess returns are achieved for the target year 1995 where the index rises to 24.9% from 5.8% in 1994. After regressing the return of the PNN H class on the index gives a beta estimate of 1.5. This result indicates that high returns tend to be achieved only at the expense of high risk. On the other hand, it is also true that the actual H portfolios and to a lesser degree the predicted H portfolios contain a disproportionate number of small capitalisation stocks. This may impose additional risk if we consider that the market in these shares is relatively illiquid. Portfolios from other classifiers and years are similar. Despite these considerations, however, we have to emphasise that the strong advantage of our trading system is that all classification methods produce consistent returns and excess returns for the genuine out-of-sample years 1994-97. The portfolio of H performing shares produces an average return over 15% for RRI and over 18% for the other four classifiers over the genuine out-of-sample years 1994-97. We have to recognise that these returns exceed the 6.50% average GBP rate over the same period. On the other hand, the portfolio of H performing shares produces an excess return of around 4% for the RRI classifier and over 6% for the other four classifiers over the genuine out-of-sample period 1994-97. Obviously, the average excess returns of our classification methods do not exceed the 6.50% average GBP rate in contrast with the average return of the portfolio over the period 1994-97. Definitely, a number of improvements on our trading system should be considered for outperforming the average GBP rate in terms of excess returns. These improvements are presented in the following Chapters. Overall, our empirical results on this Chapter suggest that our trading system is profitable on a consistent basis despite the fact that the theoretical measures of risk make the results vulnerable to the accusation that high returns are achieved only at the expense of high risk.

5.4 SUMMARY AND CONCLUSIONS

In this Chapter, we presented the first implementation of the main idea in this thesis which is a new methodology to predict high performing shares that are likely to have exceptional returns in the future by applying five heterogeneous classifiers namely, LDA, PNN, LVQ, OC1 and RRI. Using a database of several years of accounting information for around 700 companies with shares traded on the London Stock Exchange in the years 1991-97, we compared and contrasted the classifiers in terms of their ability to classify shares as H and L performing shares as well as in terms of their ability to predict H and L returns and excess returns over an equally-weighted benchmark index for the target years 1993-97.

Overall, the classification results suggest that the non-linear classifiers except RRI outperform on average the LDA. On the other hand, the financial results suggest that the improvements in classification that are achieved by the non-linear models are not reflected by the financial returns in the same degree and there are only minor inconsistencies in the profitability of the classifiers from one year to the next. The differences in the profitability of the classifiers are more obvious for the calibration year 1993, whereas these are less obvious for the genuine out-of-sample years 1994-97. Despite these differences, however, we found that all methods produce consistent excess returns and outperform the benchmark.

Differences in risk between the H and L portfolios must be carefully controlled. We observed that greater excess returns are achieved in years where the index rises more sharply and we indicated that high returns tend to be achieved only the at the expense of high risk. We also noted that the actual H portfolios and to a lesser degree the predicted H portfolios contain a disproportionate number of small capitalisation stocks. However, the scale and consistency of the excess returns from all our classifiers and for all target years seems to us to outweigh any possibility that the high returns are achieved at the expense of high risk.

Furthermore, we have to emphasise that the results we obtained from this experiment are the minimum results we can get from our trading system since a number of improvements could have been considered. For example, in view of the success of forecast combination in more conventional forecasting exercises such the one reported from Makridakis et al. (1982), it would be worth investigating if we could improve the performance of our trading system by combining the predictions of the five classifiers that we applied in this Chapter. The theoretical issues in forecast combination and the application of relevant techniques in our study are discussed in the next Chapter.

Return on Capital	PBT/TA, PBT/TCE, NI/TCE, CF/TA, CF/TCE
Profitability	PBT/SR, PAT/SR, NI/SR, CF/SR, PAT/EQ, CF/MKBD
Financial Leverage	DEBT/EQ, DEBT/TCE, DEBT/TA, TL/EQ, TA/EQ, BA/MKBD
Investment	P/E, DY, EY, BE/ME
Growth (%)	TA, PAT, PBT, EPS, MKBD, SR
Short-Term Liquidity	CA/CL, CL/TA, CL/EQ
Return on Investment	NI/TA, PAT/TA
Efficiency	SR/TA, DRS/SR
Risk	PBT/CL, PAT/CL, NI/CL, CF/CL

PBT: Profit Before Taxes; TA: Total Assets; TCE: Total Capital Employed; CF: Cash Flow; PAT: Profit after Taxes; SR: Sales Revenue; NI: Net Income; EQ: Shareholders' Equity; MKBD: Market Capitalisation at Balance Sheet Date; DEBT: Debt; TL: Total Liabilities; BA: Book Assets; P/E: Price/Earnings Ratio; EY: Earnings Yield; DY: Dividend Yield; BE: Book Equity; ME: Market Equity; EPS: Earnings Per Share; CA: Current Assets; CL: Current Liabilities; DRS: Debtors.

Table 5.1: Initial list of the accounting variables that we collected to predict high and low performing shares

LDA	PBT/TA, PBT/SR, SR/TA, DBS/SR, PAT/TA, P/E, CF/TCE, DY, CF/CL, EQ/MKBD, TA/MKBD, CF/MKBD
PNN	SR/TA, NI/TA, EPS (%), PAT/TA, P/E, NI/TCE, DEBT/TCE, CF/TA, CF/SR, EY, DY, PBT/CL, PAT/CL, NI/CL, EQ/MKBD, TA/MKBD, CF/MKBD
RRI	DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, EQ/MKBD, TA/MKBD, CF/MKBD
LVQ	SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, EPS (%), SR (%), MKBD (%), PBT (%), NI/TCE, DEBT/TCE, CF/TA, CF/TCE, CA/CL, CL/EQ, CF/CL, EQ/MKBD, TA/MKBD, CF/MKBD
OC1	DEBT/TA, PBT/SR, TA/EQ, NI/SR, DEBT/EQ, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), PAT (%), P/E, NI/TCE, CF/TA, CF/TCE, CF/SR, CL/TA, CL/EQ, NI/CL, TA/MKBD, CF/MKBD

Table 5.2: List of the accounting variables that we finally selected to predict high and low performing shares after applying stepwise variable elimination procedures

		LDA		PNN		LVQ		OC1		RRI	
Actual Class	Patterns	Predicted Class Membership									
1993		H	L	H	L	H	L	H	L	H	L
H	163	98	65	96	67	90	73	88	75	91	72
L	488	128	360	123	365	154	334	163	325	149	339
Overall (%)		70.35 %		70.81 %		65.13 %		63.44 %		66.05 %	
1994		H	L	H	L	H	L	H	L	H	L
H	163	92	71	84	79	84	79	99	64	94	69
L	488	212	276	199	289	206	282	193	295	233	255
Overall (%)		56.53 %		57.30 %		56.22 %		60.52 %		53.61 %	
1995		H	L	H	L	H	L	H	L	H	L
H	173	106	67	103	70	90	83	98	75	99	74
L	519	208	311	193	326	188	331	183	336	233	286
Overall (%)		60.26 %		61.99 %		60.84 %		62.72 %		55.64 %	
1996		H	L	H	L	H	L	H	L	H	L
H	188	106	82	100	88	105	83	96	92	113	75
L	561	262	299	254	307	216	345	216	345	253	308
		54.07 %		54.34 %		60.08 %		58.88 %		56.21 %	
1997		H	L	H	L	H	L	H	L	H	L
H	188	105	83	103	85	100	88	100	88	96	92
L	564	214	350	199	365	203	361	212	352	220	344
Overall (%)		60.51 %		62.23 %		61.30 %		60.11 %		58.51 %	

Table 5.3: Out-of-sample classification results of LDA, PNN, LVQ, OC1, and RRI for 1993-97 using accounting information to predict high and low performing shares

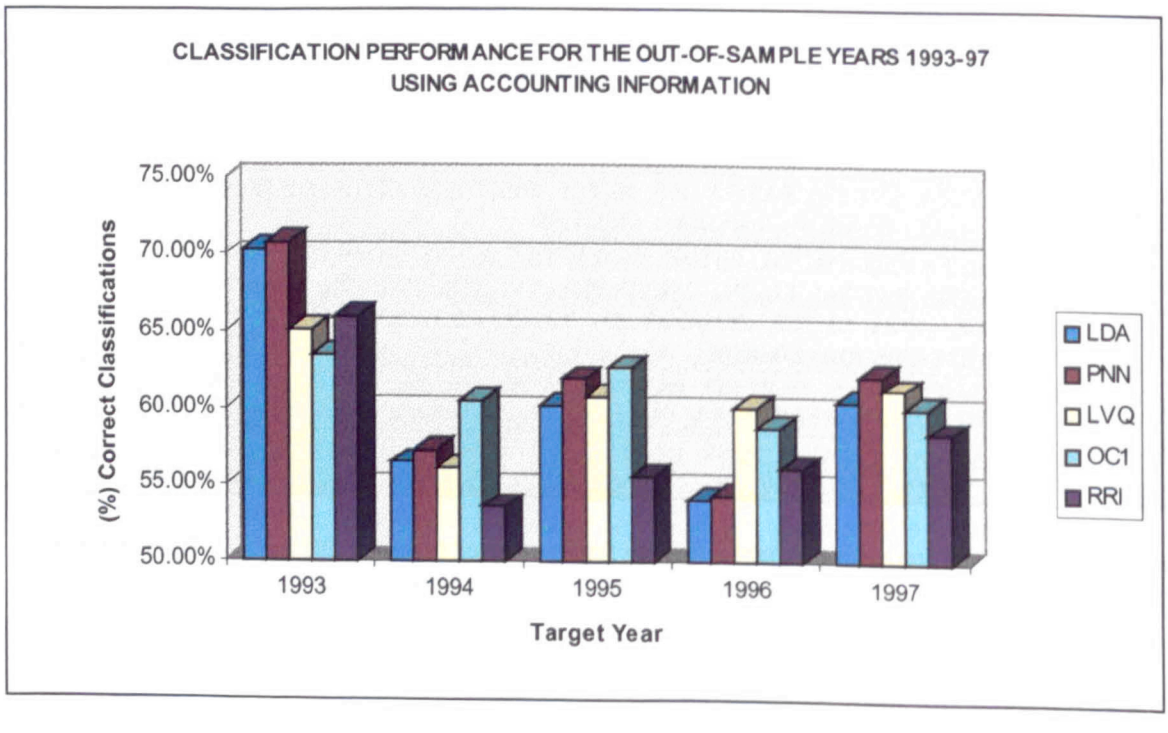


Figure 5.1: Out-of-sample classification results of LDA, PNN, LVQ, OC1, and RRI for 1993-97 using accounting information to predict high and low performing shares

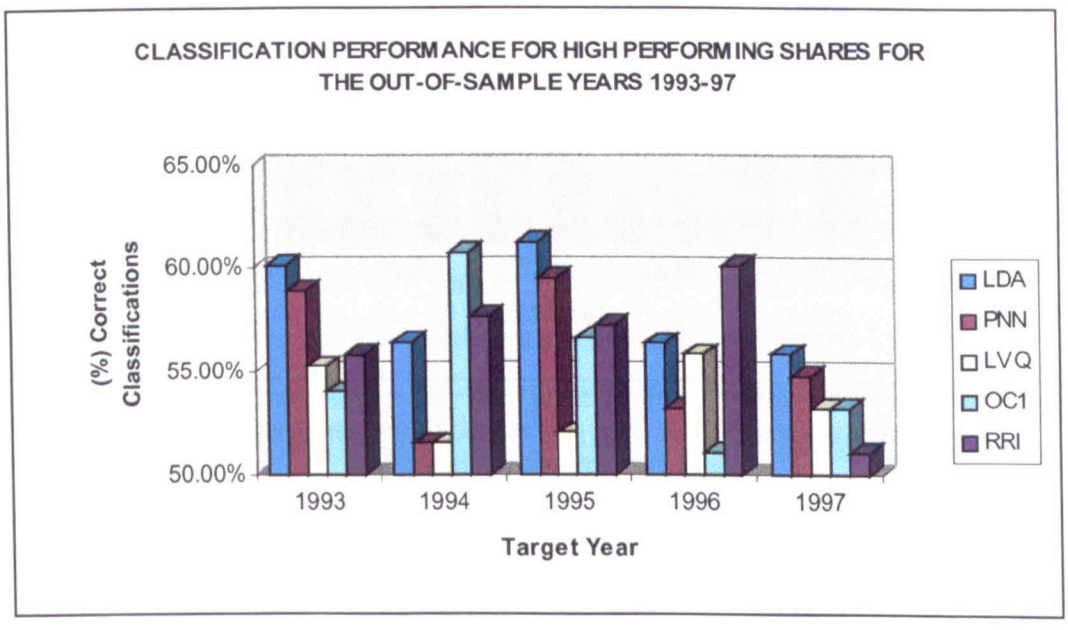
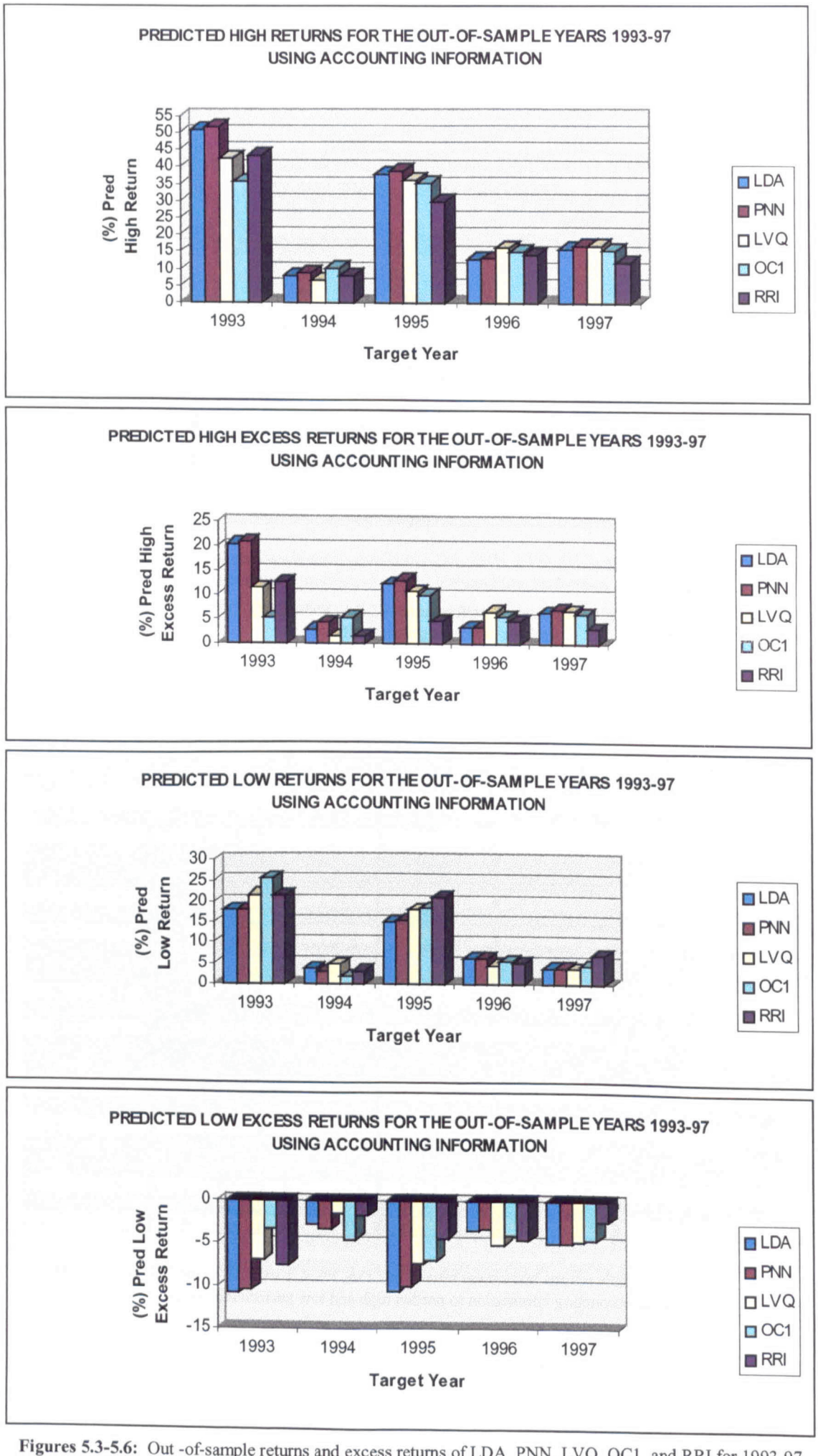


Figure 5.2: Out-of-sample classification performance of LDA, PNN, LVQ, OC1, and RRI for 1993-97 for high performing shares only

		LDA		PNN		LVQ		OC1		RRI	
1993		Predicted Returns & Excess Returns									
Actual Return	Index	H	L	H	L	H	L	H	L	H	L
H= 90.0 L= 9.2	H= 31.3	50.9	18.1	51.8	18.1	42.5	21.6	35.9	25.4	43.3	21.3
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L
H= 58.7 L= -19.5	L= 28.7	20.4	-10.8	20.8	-10.5	11.6	-6.9	5.5	-3.4	12.9	-7.5
1994											
Actual Return	Index	H	L	H	L	H	L	H	L	H	L
H= 45.5 L= -7.6	H= 5.8	8.1	3.6	9.1	3.1	6.6	4.9	10.5	1.8	8.2	3.2
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L
H= 39.7 L= -13.2	L= 5.6	3.1	-2.7	4.4	-3.4	1.8	-1.4	5.5	-4.5	1.7	-1.8
1995											
Actual Return	Index	H	L	H	L	H	L	H	L	H	L
H= 79.8 L= 7.1	H= 24.9	37.8	14.9	38.7	15.3	36.0	18.1	35.4	18.4	29.9	21.0
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L
H= 54.9 L= -18.3	L= 25.4	12.6	-10.5	13.3	-9.9	10.9	-7.3	10.2	-7.0	4.7	-4.3
1996											
Actual Return	Index	H	L	H	L	H	L	H	L	H	L
H= 54.5 L= -5.3	H= 9.7	13.2	6.3	13.2	6.5	16.4	4.6	15.3	5.6	14.2	5.4
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L
H= 44.8 L= -15.1	L= 9.8	3.4	-3.4	3.5	-3.3	6.8	-5.2	5.6	-4.1	4.7	-4.7
1997											
Actual Return	Index	H	L	H	L	H	L	H	L	H	L
H= 58.4 L= -7.2	H= 9.0	16.3	3.9	17.0	3.9	17.1	3.8	15.8	4.5	12.0	7.1
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L
H= 49.4 L= -16.5	L= 9.3	6.6	-5.0	7.3	-5.0	7.0	-4.8	6.3	-4.6	3.2	-2.4

Table 5.4: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, and RRI for 1993-97 using accounting information to predict high and low performing shares



Figures 5.3-5.6: Out -of-sample returns and excess returns of LDA, PNN, LVQ, OC1, and RRI for 1993-97 using accounting information to predict high and low performing shares

CHAPTER 6: COMBINING HETEROGENEOUS CLASSIFIERS TO PREDICT HIGH PERFORMANCE STOCKS

In recent decades, various researchers have considered combining (averaging) forecasts in the hope that a combination of forecasts can be found which yield final classifications superior to those of each individual forecast. There is little doubt that combining individual forecasts improves forecast accuracy for quantitative prediction. The conclusion holds true for statistical forecasting and judgmental forecasting (Clemen 1989, Makridakis 1989). Empirical results suggest three regularities. First, combining individual forecasts improves forecasting accuracy in the sense that it reduces the variance of forecasting errors. Second, simple averaging models work as well as more complex combination methods. Third, combining can generally be done with little or no increase in overall forecasting costs.

Apart from lower forecast errors, combinations of forecasts allow for greater flexibility in the model switching sense to utilise a broader information set. In this content, a combination of diverse models may be able to utilise a broader information set than a combination of similar models. However, a better utilisation of the information set may lead to an increase in forecasting accuracy. Therefore, forecast diversity may be an essential technique in improving forecasting accuracy.

In our sphere of interest, many researchers have applied composite classifier architectures and mixture models to various financial applications (Faria and Souza, 1995; Waterhouse, 1997). However, although there have been a number of successful efforts at combining homogeneous classifiers including multiple decision trees (Breiman et al. 1984), multiple neural networks (Maclin and Shavlik, 1995), and nearest neighbour algorithms (Skalak, 1997), searching the more complex space of sets of heterogeneous component classifiers has not been shown necessarily to achieve high accuracy (Battiti and Colla, 1994; Breiman 1994). The value of combining classification methods is therefore not so well established as the value of combining point forecasting methods. Our study provides a good testbed for further investigation.

In this Chapter, we therefore explore the potential to apply composite classifier architectures to predict high performing shares by intelligently combining five heterogeneous classifiers, namely LDA, PNN, LVQ, OC1 and RRI. A composite classifier architecture that combines successfully a small number of heterogeneous component classifiers may be of particular

importance for our application in predicting high performing shares. The empirical results that we presented in Chapter 5 suggested that there are some inconsistencies in the classification accuracy and profitability of the classifiers from one year to the next. Although the inconsistencies are more obvious in the classification performance of the classifiers, this evidence supports the view that the actual process in the financial data is complex and inconsistent in the time scale. A linear classifier such as the LDA may be able to predict high performing shares with the same accuracy as non-linear classifiers if the actual process in the data is linear, whereas the PNN will be more accurate than LDA if the actual process in the data is non-linear. However, since we do not know what the actual process in the financial data is going to be over the coming year, we are not in a position to select the more suitable classifier for each particular target year. Therefore, a combination of five heterogeneous classifiers may be more accurate on average than the individual classifiers to predict high performing shares because it may be more flexible to deal with the complex and unpredictable structures that are evident in the financial data.

Applying this system, we attempt to answer the following questions: first, how can classifiers from different model classes be combined to create a composite classifier with higher accuracy in predicting high performing shares than the individual component classifiers ? and second, what design criteria should we take into account for the selection of the individual component classifiers that will be used to build the composite classifier architecture ?

This Chapter is divided into three parts: in the first part, we discuss design specifications to construct composite models and we discuss the literature review in combining forecasts. In the second part, we investigate empirically the possibility of combining our five classification methods, namely LDA, PNN, LVQ, OC1 and RRI and apply the resulting predictions to predict high performing shares. We discuss our proposed composite classifier architecture and we explain in detail the criteria we considered for the selection of the individual component classifiers that we used to build this architecture. Finally, in the third part, we summarise the discussion and we provide the conclusions.

Part One: Design Considerations to Construct Composite Architectures for Classification or Regression

6.1 TECHNIQUES FOR CONSTRUCTING COMPONENT MODELS FOR COMPOSITE ARCHITECTURES

Many different approaches have been presented in the literature to create diverse component models whose predictions can be combined effectively. A good taxonomy of the different methods presented in the literature is provided by Skalak (1997) as follows: 1. Reasoning strategy combination; 2. Divide and conquer; 3. Model class combination; 4. Architecture and parameter modification; 5. Randomised search; 6. Feature selection; and 7. Training set resampling. These methods are discussed in brief in the next Sections. A more detailed presentation of these methods can be found in Skalak (1997).

6.1.1 Reasoning Strategy Combination

An active area for research is the combination of case-based systems and other reasoning systems. Hybrid case-based systems give particular emphasis on the control strategy or the implementation framework for integrating the component modules. Hybrid case-based systems apply larger amounts of domain knowledge and therefore they require an effective control strategy. A case-based reasoning system also incorporates a measure of case similarity which is selected according to the particular application.

A number of researches have combined case-based systems with reasoning strategies. Case-based systems that have combined reasoning strategies for classification include the ANAPRON system (Golding and Rosenbloom, 1991), the CASEY system (Koton, 1988), and the PROLEXS system (Walker, 1992). Other hybrid case-based systems have been proposed by Rissland and Skalak (1991), and Risland et al. (1993).

6.1.2 Divide and Conquer

The divide-and-conquer or recursive partitioning strategy can be applied to achieve model diversity by applying the component models to separate regions of the space of instances.

Jacobs et al. (1991b) applied the divide-and-conquer strategy to the hierarchical mixtures of experts (HME) algorithm. This algorithm is an extension of the mixture of experts (ME) model

that was first introduced by Jacobs et al. (1991a). The ME model is motivated by the principle that if a problem may be broken down into smaller problems, then we might be able to solve more easier the overall problem. The model assumes that the data are generated from a series of processes. The experts model the distinct processes in the data and a gating network models the decision to use one of the distinct processes. In this sense, a prediction is made up of a series of predictions from the individual experts. The HME can be represented as a tree structure where the terminal nodes of the tree contain the expert networks and the non-terminal nodes contain the gating networks. This model implies that each process that generated the data is itself a decomposition of processes that are selected stochastically. The experts model the lowest level of the process, whereas the gating network models the decomposition of processes in a successive fashion.

Cooper et al. (1982) applied the divide-and-conquer strategy to a classification algorithm known as Nestor Learning System (NLS). This consists of an array of Restricted Coulomb Energy System (RCE) classifiers whose classification decisions are combined using a proprietary class selection device. An RCE classifier is created by sequentially fitting basins of attraction around a subset of training patterns. A test pattern is given the class of a training pattern only if it falls within the basin of attraction of that training pattern. The highest priority RCE classifier that gives an unambiguous response during classification is the prediction of the composite algorithm (Skalak, 1997).

Other studies that applied the divide-and-conquer strategy are those of Chan and Stolfo (1995) who combined a set of classifiers in a binary tree structure, and Edelman (1995) who proposed a two level stacked generalisation framework using radial basis functions. A detailed description of these studies can be found in Skalak (1997).

6.1.3 Model Class Combination

This strategy implies that models from similar or dissimilar model classes are combined to create a model with higher accuracy than the component models. Empirical work in this area has been presented by Batiti and Colla (1994) and Breiman (1994). Battiti and Colla (1994) combined MLP perceptrons and LVQ. Breiman (1994) investigated stacked generalisation for function approximation in an approach known as stacked regression. He combined several types of CART and two types of linear regression: subset regression and ridge regression. The results of this experiment were very promising (Skalak, 1997).

6.1.4 Architecture and Parameter Modification

This strategy implies that a number of different models can be created by varying one or more

parameters. Battiti and Colla (1994) investigated the combination of MLP networks with different numbers of units. Varying both the number of input and hidden nodes of the networks as well as the random weight initialisations, they attempted to create networks with uncorrelated errors. Battiti and Colla observed that the predictions of the resulting networks are not independent even though their theoretical analysis is based on the common assumption of independence. Maclin and Shavlik (1995) demonstrated that competitive learning can be used to initialise neural network weights to increase the number of local minima that may be reached by the networks.

6.1.5 Randomised Search

Diverse models can also be created through randomised search. Opitz and Shavlik (1995) used a genetic algorithm called ADDEMUP to create accurate and diverse neural networks to be used as individual component models. This algorithm applies mutation and crossover operators that perturb and combine the topologies of the member networks. English (1996) developed an evolutionary algorithm to evolve a collection of recurrent networks to be applied in a stacked generaliser for time-series prediction.

6.1.6 Training Set Resampling

A very effective way to achieve diversity is to train component models on different samples of the training set. Breiman (1994) proposed a form of stacked generalisation in which bootstrap sampling is used to create different training sets for a set of predictors. After training, the predictions of the component models are aggregated by averaging numeric predictions or voting symbolic ones.

6.1.7 Feature Selection

Feature selection can also be used to enhance model diversity. After studying voting and related combination algorithms, Battiti and Colla (1994) observed that different input features might lead to increases in accuracy. Wolpert (1992) developed a stacked generaliser which incorporates three nearest neighbour components that are each given a different input feature from the representation of training instances. He also reported increases in classification accuracy (Skalak, 1997).

6.2 THREE GENERAL FRAMEWORKS FOR COMBINING COMPONENT MODELS

Many architectures have been suggested in the literature to combine component models. The most important architectures for combining component models can be summarised as follows: i) stacked generalisation; ii) boosting; and iii) recursive partitioning. The architectures are

discussed in the next Sections. A more detailed presentation of these architectures can be found in Skalak (1997).

6.2.1 Stacked Generalisation

The idea of stacked generalisation has its origin back to the work of Selfridge's (1959) pandemonium for pattern recognition and Nilsson's (1990) committee machines. Stacked generalisation is a recursive layered framework in which each layer of component models is used to combine the predictions of the models at the previous layer. A single classifier or predictor at the top ultimate layer makes the final prediction. Stacked generalisation can be a two-layer architecture or it may contain as many layers as possible. The model at each layer receives as input a vector of predictions of the components in the previous layer. Therefore, the layering can be extended to include additional layers. Using higher-lever layers to learn the types of errors made by the immediately preceding layers may be an effective technique to minimise the generalisation error.

Some of the most recent studies on the stacked generalisation framework are those of Wolpert (1992, 1993), Perrone (1993), Schaffer (1994), Murphy and Aha (1994), Kong and Dietterich (1995), Dietterich and Bakiri (1995), and Skalak (1997). Wolpert (1992, 1993) concentrated on the two-level framework in which the classifiers to be combined represent the level-0 classifiers and the combining classifier is the level-1 classifier. He applied this architecture to the NetTalk problem for prediction of letter sequences to test whether stacked generalisation can be used to combine different pieces of incomplete information. He reported that stacked generalisation achieves a generalisation accuracy 20% over the level-0 classifiers. Perrone (1993) attempted implementations of stacked generalisation in which the models to be stacked were neural networks. Schaffer (1994) investigated an extension of stacked generalisation named β-level stacking. According to this architecture, the combining classifier has also access to the inputs of the level-0 classifiers as well as the predictions of the component classifiers. Comparing β-level stacking with stacked generalisation on five U.C.I. Machine Learning Repository data sets of Murphy and Aha (1994), Schaffer reported that β-level stacking outperforms stacking in only three out of fifteen experiments using three combining classifiers in each of five different data sets. The three classifiers combined were a rule induction, a decision tree, and a neural network trained with backpropagation.

Kong and Dietterich (1995) and Dietterich and Bakiri (1995) investigated a variant of stacked generalisation called error correction output coding (ECOC) which transforms a multi-class problem into a large number of two-class problems. A large number of component classifiers

are trained on a set of distinct problems and each component classifier is applied to learn one bit-position of a binary codeword that is assigned to each class. After collecting the predictions of the individual components in a vector, a nearest neighbour algorithm is applied to find the codeword that is more close to the vector of predictions. The class of that codeword represents the prediction of the ECOC classifier (Skalak, 1997).

Skalak (1997) investigated whether the accuracy of a standard nearest neighbour classifier can be improved or exceeded through composite classifier architectures that incorporate a small number of diverse component nearest neighbour classifiers each of which stores only a small number of prototypes. Using a variety of data sets from the U.C.I. Machine Learning Repository of Murphy and Aha (1994), Skalak investigated the benefits of stacking minimal nearest neighbour component classifiers. He reported that combining the predictions of a set of minimal nearest neighbour components is significantly more accurate than a full nearest neighbour classifier on approximately half of the data sets. However, he observed that after increasing the number of components, the accuracy of stacked generalisation improves monotonically on three data sets, whereas in two other data sets the accuracy seems to decrease.

6.2.2 Boosting

Boosting has been proposed by Schapire (1990) and has been investigated further by Drucker et al. (1994), Freund and Schapire (1995), Drucker and Cortes (1996), and Quinlan (1996b). Boosting is a framework that attempts to increase the generalisation accuracy of a given model by creating complementary component models while filtering the training set. The ultimate prediction is a weighted vote over the predictions of the initial model as well as the predictions of the newly created components.

The main difference of boosting from stacked generalisation is that boosting assumes a single model whose accuracy can be increased by creating complementary component models while filtering the training set, whereas stacked generalisation assumes that each layer of component models is used to combine the predictions of the components at the immediately preceding layer. Another difference is that boosting uses a voting scheme to combine the individual predictions, whereas stacked generalisation may use other combining algorithms as well. Finally, the results from training each component model in boosting are used to form the training set for the next component model, whereas only some forms of stacked generalisation allow the combining algorithm to have access to the inputs of the component models in the lower level (Skalak, 1997).

Schapire's (1990) boosting algorithm is applied in four steps as follows (Skalak, 1997),

Step 1: Train the base component model C_1 on a set of training patterns.

Step 2: Toss a fair coin. If the coin comes up heads, then apply C_1 to new instances until C_1 makes a mistake and add the mistaken instance to the training set for C_2. If the coin comes up tails, then apply C_1 to instances until C_1 is correct and then apply the correct instance to a training set for C_2. Repeat this process until a sufficient number of patterns have been selected to train C_2.

Step 3: Drawn new instances from a distribution of training instances and classify or predict them by using the base component classifier C_1 and the newly created component model C_2. If the two models disagree, then add the instance to the training set for model C_3 and stop the training when C_3 has been trained on sufficiently many patterns.

Step 4: To give the ultimate prediction, take a majority vote of the base model C_1 and the newly created component models C_2 and C_3.

Quinlan (1996b) demonstrated in a number of empirical tests that although boosting generally increases accuracy, it leads to a deterioration on some data sets. An extension to boosting is the Adaboost algorithm proposed by Freud and Shapire (1995). This algorithm differs from boosting in that it combines a large number of weak learners by taking a weighted average over their predictions and changing the distribution of training examples as each weak learner is trained serially. An extension of the Adaboost family of algorithms is the Rankboost algorithm that was proposed by Freud et al. (1998). This algorithm works by combining many "weak" rankings of the given instances. These rankings may be only weekly correlated with the target ranking that they attempt to approximate. Freud et al. showed how to combine such rankings into a single highly accurate ranking.

6.2.3 Recursive Partitioning

Recursive partitioning is a framework for model combination which recursively sub-divides the domain into different regions and then applies individual models that are specialised in each region to make a prediction. This architecture uses a divide-and-conquer strategy to partition an instance space into regions that contain instances of only one class. A tree of component models is created by recursively calling a construction function on the instances that fall to each node. This function determines if the instances are all from one class. If this is the case, the class is returned. Otherwise, a single component model is trained on those instances and then applied to them to partition the instances for the next recursive step.

Utgoff (1989) proposed a recursive partitioning algorithm that combines a univariate decision

tree with linear threshold units. This algorithm starts by determining whether a subspace is linearly separable by applying some heuristic measure. If the subspace is linearly separable, a linear threshold is applied. If it is not linearly separable, an information-theoretic measure is applied to split the subspace into separate regions. Brodley (1992) suggested a recursive tree-structured hybrid classifier which combines decision trees, linear discriminant functions, and instance-based classifiers using heuristic knowledge (Skalak, 1997).

The recursive partitioning algorithms have two advantages: first, a classifier or predictor can be selected whose bias is appropriate for a specific region of the instance space; and second, the application of the framework is efficient since only component models along a path through the tree need to be applied to each instance. Recursive partitioning algorithms have two main disadvantages: first, the process of sub-dividing the domain space into regions may result in overfitting since there are fewer training instances to fit; and second, only one component model is ultimately used to make a prediction for each subspace excluding the possibility that other component models may provide useful predictions for instances in that subspace (Skalak, 1997).

6.3 EMPIRICAL MODELS IN COMBINING FORECASTS

In this section, we review methods that have been implemented in the literature to combine forecasts. These methods can be structured according to whether the components output a real value or a class label. In the former case, we attempt to identify a model that can be applied to combine the forecasts. Common approaches that have been implemented in the literature include the minimum-variance approach, the regression approach, the Bayesian approach, econometric models, and other extensions. In the later case, we attempt to identify the appropriate method to perform classification. Two common approaches that have been implemented in the literature are the voting and non-voting methods (Skalak, 1997). A more detailed discussion on these methods is provided in the next Sections.

6.3.1 Minimum-Variance versus Regression

The first studies in the combination of forecasts are dated back in the 1960s (Crane and Cotty 1967; Zarnowitz 1967). However, the starting point for the building of relevant theory in this field was given by Reid (1968, 1969) and Bates and Granger (1969) who pioneered the approach for combining forecasts. Bates and Granger (1969) proposed that if a number of unbiased forecasts of the future variable are available, then we can combine them in such a way that the composite forecast will have a variance less than or equal to the variance of the individual forecasts. Using a classical statistical framework, they showed that a linear weighted

combination of forecasts is the optimal forecast if we restrict the weights to sum to one and we assume that the forecasts are unbiased. Let us assume that we want to combine n individual forecasts $g_{1t}, g_{2t}, \ldots, g_{nt}$ to obtain an estimate of the unknown quantity Y. According to Bates and Granger, the solution to the problem is to consider the linear combination $\sum \alpha_i g_i$ and choose α_i to minimise $E\left(\sum \alpha_i g_i - Y\right)^2$ under the assumption that $\sum \alpha_i = 1$. If the weights are restricted to sum to one some of them might be negative. This will be a desirable property, however, because a weighted forecast with negative weights may lie outside the range of the individual forecasts. Following the formulation of Bates and Granger (1969), Dickinson (1973, 1975) discussed the sampling distribution of combining weights and emphasised that the poor reliability of negative weights may reduce the practical usefulness of the combining procedure.

In their original paper, Bates and Granger (1969) observed that if a set of forecasts are combined with a set of unbiased forecasts and consistently overestimate the true values, they may lead to forecasts that are biased. Therefore, they suggested that the common practice should be to check the individual sets of forecasts and if we find them biased then to correct for absolute percentage bias. To deal with these issues, Granger and Ramanathan (1984) suggested the use of Ordinary Least Squares (OLS) with an intercept and no restriction on the weights to add up to unity. In other words, contrary to Bates and Granger's formulation, they proposed a combining scheme using a regression framework. More specifically, they proposed to estimate Y using a model of the form $\alpha_0 + \sum \alpha_i g_i$ where the weights are chosen to minimise $E\left(\alpha_0 + \sum \alpha_i g_i - Y\right)^2$ but without constraining them to sum to one. In implementing this model, Granger and Ramanathan suggested to perform a simple OLS regression with the actual value as the dependent variable and the forecast values as the independent variables. Clemen (1986) observed that this general model may lead to the smallest sum of squared errors for the data used to fit the regression but it may not always be better than the Bates and Granger's model in terms of predictive accuracy outside the fitting data. He showed that the inclusion of the constraints might improve the predictive accuracy, whereas in some cases the increased efficiency might offset any incurred bias. After empirical tests, Clemen reported that the combined forecasts do not perform better than the individual forecasts when the constraints are omitted, whereas the imposition of the constraints seems to improve the predictive accuracy of the combining forecasts over the individual forecasts. In a subsequent paper, Trenkler and Liski (1986) proposed an F-test as a good compromise when the truth of the linear constraints is questionable.

Diebold (1988) suggested that the regression-based approach proposed by Granger and Ramanathan (1984) makes all the standard results of the linear model immediately applicable to forecast combination but emphasis have to be paid to properties of residuals from combined forecasts. If the combined forecast errors are serially correlated, then the OLS estimates of the combining weights might be inefficient and their associated standard deviation estimates might be inconsistent. The combined forecast might not be the best unbiased linear combination if serial correlation is present. Diebold demonstrated how explicit modelling of the serial correlation can improve the combined predictions.

Holden and Peel (1989) considered the problem of determining whether forecasts are unbiased and examined the implications that this might have on combining different forecasts. They showed that if a constant is included in an unconstrained regression, then it might correct for any bias in the combination of the individual forecasts, but it might combine these forecasts with the unconditional mean of the series. If a constant is included in the regression and the weights are constrained to add to unity, then the presence of the constant might correct for any bias, whereas the unconditional mean of the series might not be introduced as an additional forecast. However, if a constant is not included in the regression and any of the forecasts are biased, then the resulting least squares parameters might be biased. Holden and Peel suggested that if the forecasts are thought to be unbiased, then there is no justification for including a constant since most economic forecasts rely on models which attempt to explain changes in the mean. If forecasts are thought to be biased, then the most effective way to combine them might be to constrain the weights they are given and to include a constant term to remove the bias. However, this approach might lead to an improved forecast only if the patterns between the different forecasts are stable.

Coulson and Robins (1993) generalised Diebold's (1988) treatment of the dynamics in the combining regression using the theory that was proposed by Hendry and Mison (1978). This theory suggests that the process of whitening residuals through Cochrane-Orcutt or similar methods is equivalent to OLS estimation of dynamic regressions with non-linear restriction of the parameters. Extending Hendry and Mison's model, Coulson and Robins suggested a parsimonious method of accounting for the dynamics through the use of a lagged dependent variable but without lagged forecasts. They found that this method improves forecast combination procedures and they reported that improvements are also obtained when the data are nonstationary.

Chandrasekharan et al. (1994) emphasised that the reliability and precision of the weights used in combining individual forecasts may be also important in evaluating a combined forecast.

They suggested that a combination of forecasts will be reliable if the weights are not changing significantly for small variations in the individual forecast errors, whereas an individual forecast will be insignificant to a combination of forecasts if we cannot statistically reject the hypothesis that the weight assigned to the combined forecast is equal to zero.

6.3.2 The Bayesian Approach

A number of studies suggested a Bayesian framework in combining forecasts. Roberts (1965) and Geisel (1973) proposed Bayesian methodologies for weighting or combining forecasts. They suggested that the Bayesian process can be applied in those situations where neither the structure of the model nor the values of its parameters are known with any degree of certainty. Bunn (1975, 1977, 1978) used Bayesian theory to allow the incorporation of subjective opinion into the synthesis of forecasts. Let us consider a common linear combining model as follows (Faria and Souza, 1995),

$$\phi(y_t \setminus V) = \sum_{i=1}^{n} w_i \phi_i(y_t) = L'\phi \qquad (6.1)$$

where,

$y_1, y_2 \ldots y_t$ = a sequence of observations from time *t=1* to *t=T* of the variable of interest Y,

$g_1, g_2, \ldots, g_n$ = the n forecasters that generate one period ahead forecast of variable Y,

$\phi_i(.)$ = the predictive density that specifies the forecast generated by g_i,

w_i = the weight of $\phi_i(.)$ in the combination.

Bunn suggested a formulation that allows sampling in respect of the performance of individual forecasting models in the form of a multinomial distribution whose parameters are the probabilities of the best performance of a model over the remaining ones. Considering a synthesis of n individual forecasts at time t = T, we can represent the multinomial distribution as follows (Faria and Souza, 1995),

$$P_{D_1, D_2 \ldots D_{n-1}}(d_1, \ldots, d_{n-1}) = \frac{T!}{\prod_{i=1}^{n} d_i!} \prod_{i=1}^{n} w_i^{d_i}, \qquad \sum_{i=1}^{n} w_i = 1 \qquad (6.2)$$

where,

d_i is the number of times the individual forecast i performed better than the other

forecasts, and w_i is the probability that forecast i will outperform all the other forecasts in the future.

To estimate parameters, w_i, Bunn used the Dirichlet distribution to incorporate the prior information, and then he used the Bayes Theorem to update this distribution in a continuous fashion based on the performances of the forecasting models. To consider the revision of the relative probabilities associated with the outperformed individual forecasts, Bunn used a formulation based on the matrix beta distribution. After a series of experiments, he demonstrated that this model is preferable to the minimum-variance method of Bates and Granger (1969) if there is little prior information. In a similar study, Bunn (1979) showed that the standard Bayesian combination forecast is inferior to a minimum-variance linear composite forecast that is obtained from the class of linear unbiased estimators. Furthermore, in a later study, Bunn (1985) reported that the accuracy of a Bayesian combination of simulated forecasts is inferior to that of regression combinations in terms of Mean Square Error (MSE). Examining these results, Walz and Walz (1989) observed that the Bayesian procedure produces the least accurate combined forecasts because the simulated data created by Bunn are generated from a stable process with specified variance. Therefore, these data contain no shifts in the relative forecasting abilities of the set of predictors. Using macroeconomic data, Walz and Walz compared the accuracy of combining forecasts derived by a Bayesian methodology with the accuracy of composite forecasts derived by multiple regression. This experiment is based on the forecasts of four macroeconomic variables from five econometric models. The results of this experiment suggested that the Bayesian combination procedure produces more accurate composite forecasts than a regression combination procedure using a version of Theil's (1966) U^2 statistic. Walz and Walz suggested that the multiple regression technique can be used to combine econometric forecasts only if three conditions are satisfied: 1) there is some belief that the processes involved are stable; 2) the data represent time frames that are small enough to ensure the assumption of stability; and 3) there is a priori information concerning the nature and timing in the relative forecasting abilities within the set of predictors.

Smith and Makov (1978) proposed a quasi-Bayes method as a simplification of pure Bayesian estimation which presents a serious problem in the calculation of the posterior probability distribution. In the Bayesian approach, the posterior distribution of W given y_τ can be written as follows (Faria and Souza, 1995),

$$p(W \setminus y_\tau) \propto \phi(y_\tau \setminus W) p(W \setminus y_{\tau-1}) \tag{6.3}$$

where $y_\tau = (y_1,\ldots,y_\tau)$.

A successive computation of the Eq. (6.3) until time $t = \tau$ becomes prohibitive with n^τ terms of products of population densities. The form of the linear combination induces a combinational explosion in terms of the likelihood function and this prevents the derivation of the exact posterior distribution in most practical problems. To overcome such computational problems, Smith and Makov assumed a quasi-Bayes simplification as follows (Faria and Souza, 1995),

$$p(W \setminus y_\tau) = D(W; c_{1,\tau},\ldots,c_{n,\tau}) \tag{6.4}$$

The parameters of Eq. (6.4) can be updated as shown below (Faria and Souza, 1995),

$$c_{i,\tau} = c_{i,\tau-1} + \frac{\phi_i(y_\tau)c_{i,\tau-1}}{\sum_{i=1}^{n}\phi_i(y_\tau)c_{i,\tau-1}} \quad i=1,2\ldots N \tag{6.5}$$

The mean of the posterior distribution for w_i at time $t = T$ is given as follows (Faria and Souza, 1995),

$$\overline{w}_{i,\tau} = \frac{c_{i,\tau}}{\left(\sum_{i=1}^{n} c_{i,\tau}\right) + T} \quad i=1,2\ldots n \tag{6.6}$$

According to the quasi-Bayes approach, the sequential classification of the variable Y as belonging to one of the populations is treated as a multinomially distributed random variable. The weight of each predictive density in the mixture is interpreted as the probability of the observation belonging to the population its density specifies. It is attributed a prior Dirichlet distribution to the weights that are sequentially estimated by a posterior update on the parameters of this Dirichlet. Hence, at each time t, the weights of each individual predictor in the quasi-Bayes combination are obtained from Eq. (6.6) through Eq. (6.5) (Faria and Souza, 1995).

Bordley (1982) showed that the variance-covariance approach that was suggested by Bates and Granger (1969) and enhanced by Dickinson (1975) and Winkler (1981) can be deducted from a Bayesian approach under the following assumptions: a) the vector of the expert forecast errors is normally distributed; b) the expert forecast errors are uncorrelated with the unknown

quantity; and c) the decision maker assigns a uniform prior distribution to the unknown quantity. In a subsequent study, Bordley (1986) showed that the Granger and Ramanathan (1984) formulation may deduce from Bayesian principles if we modify assumptions b) and c). More specifically, assuming that the decision maker has a normally distributed prior over the unknown quantity and that the expert forecast errors are correlated with the unknown quantity, then the resulting Bayesian model treats the decision maker's prior about the unknown quantity as an extra forecast that is handled in the same way as the forecasts of the individual experts. Bordley observed that this formulation will lead to Granger and Ramanathan (1984) model for the combination of forecasts. He observed that the Granger and Ramanathan (1984) model for combining n unbiased forecasts is equivalent to Bates and Granger's (1969) formula for combining $n+1$ forecasts where the $n+1^{th}$ forecast is the decision maker's prior estimate of the unknown quantity. Tibilleti (1994) extended further the original model by relaxing the assumption which states that the vector of the expert forecast errors is normally distributed. Tibilleti observed that there are situations in which the decision maker evaluates that the expert is more likely to overestimate than to underestimate the unknown quantity in his forecast. In these situations, Bordley (1982) suggested to use a procedure based on the lognormality assumptions and the final solution is a geometric average of the experts' forecasts. Tibilleti (1994) generalised this result by showing that under moderate assumptions this solution is a consistent, monotonic, quasi-linear average of the experts' forecasts $g_1, g_2, \ldots, g_n$. In other words, there is an increasing function z such that the optimal combined forecast admits the expression $z^{-1}\left(\sum_{i=1}^{n}\alpha_i z(g_i)\right)$ where α_i i=1,2,...n are the appropriate weights. Tibilleti showed that in the normal and lognormal case, this formula reduces to Bordley's results.

Morris (1974, 1977) used a Bayesian approach to address the general problem of aggregating diverse information from a variety of sources. This approach treats the expert opinions as data and assesses the likelihood function for these data. The posterior distribution of the decision maker about the quantity of interest can be derived by applying the Bayes' rule. This approach is based on the assumption that there is independence among the opinions of experts. However, if we consider that the experts' opinions may be influenced by similar factors such as similar experience, training procedures, and shared information then it is likely that these opinions will be dependent. Bayesian models for combining dependent information can be found in Winkler (1981), French (1980, 1981), Lindley (1983, 1985), Agnew (1985), Clemen (1984, 1987), and Chang (1985).

Wiper and French (1995) suggested to combine Experts' opinions using a Normal-Wishart

model. This model utilises the relationship between the decision maker and the experts and examines how a Bayesian decision maker might update his distribution for continuous variables upon hearing experts' forecasts as quantiles. Therefore, this model avoids the problems associated with different scales and ranges of the variables by assuming that the decision maker transforms the experts' quantiles in terms of his own prior distribution for each continuous variable. Wipper and French tested the Normal-Wishart model using the DSM data set used in Cooke et al. (1989) and analysed in more detail in Wiper (1990). This data set consists of forecasts of twelve experts for eight continuous variables related to flanges on engineering equipment. Each expert states his 0.05, 0.5, and 0.95 quantiles for every variable. To compare the performance of different experts with the model, Wiper and French evaluated the experts' forecasts using a bilinear loss function as follows,

$$L(\delta, x) = \begin{cases} p_1|\delta - y| & \text{for} \quad \delta \le y \\ p_2|\delta - y| & \text{for} \quad \delta > y \end{cases} \tag{6.7}$$

Under the bilinear loss, an expert minimises his expected loss by stating his $p_2 / (p_1 + p_2)$ quantile. Three bilinear loss functions are chosen so that the 0.05, 0.5, and 0.95 quantiles can be used to form the Bayes estimators of Y. The results of this experiment suggested that the Normal-Wishart model has very satisfactory performance and performs better than either the opinion pool or the best expert under all three loss functions.

6.3.3 Econometric Models and Time Series

The use of econometric models to produce macroeconomic forecasts has been extensively criticised in the literature. It has been argued that forecasts from these models are no better than those produced from simple extrapolative or regression procedures. One the other hand, the economic theory predicts that the parameters incorporated in the behavioural functions of these models will not be constant over time because they will reflect the changes in government policy. Investigating these problems, Nelson (1972), Cooper and Nelson (1975), and Granger and Newbold (1975) recognised that econometric models may be unable to incorporate all relevant information and they proposed a different approach to make the use of these models more efficient. They suggested that the forecasts derived from the time-series approach to forecasting can be combined with econometric forecasts and the combined forecasts might outperform those derived from applying each approach individually. Longbottom and Holly (1985) compared forecasts from structural econometric and time-series models. They suggested that such comparisons might provide insight for the respecification of the econometric model. Bischoff (1989) examined combining the forecasts produced from Chase Econometrics with

those produced using the Box-Jenkins ARIMA technique. Using six series of quarterly ex-ante and simulated ex-ante forecasts over 37 periods and 10 horizons, he combined the forecasts using seven different methods. The results suggested that the best combined forecasts, in terms of average relative root-mean-square error, are better than the Chase forecasts for three variables and they are inferior for two variables. However, averaging across all six variables the Chase forecasts are slightly better. The forecasts produced for the last half of the sample after applying a two-step procedure outperform slightly the Chase forecasts.

A variety of studies have investigated the problem of forecasting contemporaneously aggregated time series. Some of the earliest studies are those by Rose (1977), Tiao and Guttman (1980), Wei and Abraham (1981), Kohn (1982), and Lutkepohl (1984), to name a few. Most of these studies suggest that if the disaggregated data are generated by a known ARMA process, then forecasting them directly might not be the most suitable approach. A more preferable approach might be to forecast the disaggregated series separately using some univariate model and then aggregate these forecasts. Lutkepohl (1984) suggested that if the data generation process is unknown then it might better to forecast the aggregate series directly rather than applying a multi-model time series approach. Riise and Tjostheim (1984) supported this view and suggested that the multi-model time-series methodologies may be very sensitive to structural change and they may not forecast as well as univariate models. Mills and Stephenson (1985) attempted to integrate the approach of forecasting contemporaneously aggregated time-series and the approach of combining alternative forecasts of time-series for two of the U.K. monetary aggregates namely £M3 and M0. They found that forecasts from a time-series model for aggregate £M3 outperform forecasts from individual models used for the components or counterparts of £M3, whereas greater gains in forecasting accuracy are achieved after forming a linear combination of these alternatives. However, the findings are different for M0 where aggregated forecasts of its components outperform the forecast from the aggregate as well as a linear combination of the two.

Some researchers were generally concerned about the properties of model-based predictions in the presence of structural change. To account for structural changes over time, Engle et al. (1985) used the model of autoregressive conditional heteroskedasticity (ARCH). Using this formulation, they attempted to model the evolution of prediction error variances and covariances over time. This approach suggests using the full sample to produce a sequence of time-varying weights on a systematic fashion rather than using a recent subset of observations as a basis for the calculations. Diebold and Pauly (1987) observed that ARCH-combined forecast has several problems: first, it produces an extremely noisy weight sequence; second, it does not improve upon the individual forecasts; third, it does not compare favourably with a

fixed-weight combination; and fourth, it requires the modelling of the entire conditional covariance matrix over time which is an extremely costly task in terms of computational resources. They proposed instead that a regression-based weighted least squares (WLS) combining method may be more flexible than the ARCH combined-forecast. Furthermore, they emphasised that the modelling of time-varying regression parameters is more tractable than the modelling of the evolution of the variances and covariances under the ARCH framework in terms of the available theory.

Lawrence et al. (1985) investigated the benefits from combining judgmental forecasts as well as the benefits from combining judgmental and statistical forecasts. They found that judgmental forecasts contribute to improved forecasting performance, whereas mechanical combination is superior to judgmental combination or adjustment.

6.3.4 Extensions and Generalisations

Lawrence and Reeves (1981) used goal programming to determine optimal weights for combining forecasts. They suggested that the goals can be flexibly defined to satisfy a variety of management objectives. In a subsequent study, Reeves and Lawrence (1982) used multiple objective linear programming to combine forecasts. This model allows for the systematic combination of multiple forecasts and considers multiple objectives in the forecasting selection process. To test the decision support capabilities of the model, Reeves and Lawrence used 30 consecutive monthly observations of actual chemical product sales data, and three forecasting methods namely, exponential smoothing, harmonic smoothing, and multiple regression. After implementing the model, Reeves and Lawrence reported 16 efficient solutions. The results suggested that 15 out of 16 efficient solutions favour a combination of all three forecasting methods, whereas only one solution favours a combination of two instead of three forecasting methods. Overall, the findings suggested that it might be inefficient to use the three forecasts individually.

Russel and Everett (1987) suggested a methodology that selectively chooses models from a set of ten individual models and then combines the selected models to forecast. The selection of the models to be included in the combination was based upon their individual forecasting accuracy which is different from either using all models in the set or equally weighting the most economical and most easily understood models. Nine different model combinations were investigated in this study using thirty-one time series. These time series were evaluated over 30 periods and each series was divided into two sets: the first 18 periods were used for fitting the models, whereas the last 12 periods were used for experimentation. The results of this study supported a model combination that selects the best 3 to 5 models from the set of ten individual

models and weights the selected models according to the inverse proportion of their individual accuracy as measured by the Mean Squared Error (MSE).

Gupta and Wilton (1987, 1988) proposed a method for combining forecasts that allows inclusion of subjective and empirical information about the forecasts while at the same time provides weights that are intuitively meaningful and they are not dependent upon large numbers of prior forecast accuracy. This method is known as the Odds-Matrix approach. This approach uses a matrix of pairwise odds on outperformance to derive the required set of weights. Let us assume that the true weights are represented from the vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$. Each element, o_{ij}, in the matrix, O, is positive and it has the property that $o_{ij} = 1/o_{ji}$. The element, o_{ij}, can be interpreted as the odds that model i will outperform model j. It is easy to show that $O\alpha = k\alpha$. If the matrix, O, is known then the underlying weight vector α can be provided by the solution $(O - kI)\alpha = 0$. If the matrix, O, has unit rank and its trace is equal to k, then we can accept that only one of its eigenvalues will be nonzero and its value will be equal to k. However, usually the weights, α_i, are unknown and the odds matrix has to be estimated. Gupta and Wilton suggested that this problem can be solved as follows: if we represent as π_{ij} the probability that the i^{th} model will outperform the j^{th} model in the next forecast realisation, then we can write that $o_{ij} = \pi_{ij} / \pi_{ji}$. To specify the Odds-Matrix, we have to specify the $\binom{k}{2}$ pairwise probabilities of outperformance. These binomial probabilities can be estimated individually from available data, managerial judgement, or both. After performing a simulation experiment, Gupta and Wilton showed that their method has the following desirable properties: first, it does not rely upon large quantities of data for estimating the weights; second, it allows updating of the weights as new information becomes available; third, it is computationally simpler if the data set is large; fourth, it uses the decision maker's subjective judgement to modify the odds in order to reflect any changes if the process becomes nonstationary; and fifth, the decision maker does not have to estimate the matrix of error variance-covariances among the models directly because this information is contained in the odds.

Anandalingam and Chen (1989) generalised the Bayesian methodology for combining forecasts by assuming that the forecasters may be biased and correlated with each other. Assuming that the likelihood function of the forecasts is a multivariate normal distribution, they suggested a combination procedure that combines information from past and current sources recursively at each time epoch in a two stages fashion. In the first stage, the biasedness is taken from the individual forecasts, whereas in the second stage the decision maker's prior forecast is updated

from the information provided by the individual forecasts. According to this model, the combination weights are a function of the expected precision of the experts and the decision maker. Anandalingam and Chen showed that this model provides a combined forecast that is unbiased and has minimum variance if the individual forecasts are biased. They showed that models that assume unbiased forecasts are special cases of their model.

LeSage and Magura (1992) proposed a multimodel approach to combine forecasts from alternative sources. They criticised the Granger and Ramanathan (1984) model in three different aspects: 1) it does not allow the weights to change over time; 2) it is subject to the influence of outlying data points that might be critical in determining the least-squares weights; and 3) the individual forecasts may be highly collinear and this might decrease the precision of the least squares estimated weights. To overcome these problems, LeSage and Magura estimated the weights using a multiprocess mixture model that was initially introduced by Gordon and Smith (1988, 1990) for modelling biomedical time series. The multiprocess mixture model uses recursive Bayesian updating procedures that extend the standard Kalman filter and enable the running of multiple models in parallel. The multiple models are processed in the time frame and the posterior probabilities of each individual model can be used to draw inferences regarding the presence of outliers or shifts in the parameters of the model. Structural shifts in the weight parameters associated with the individual forecasts will reflect changes in the accuracy of the competing forecasts that occur due to changes in business-cycle conditions, changes in the model providing the forecast or even changes in judgmental adjustments made by individuals. The actual basis of the multimodel approach is a single dynamic linear model that was introduced initially by Harrison and Stevens (1976) and extended by West and Harrison (1989). LeSage and Magura observed that the multiprocess mixture model approach provides separate models that are able to account for outlying observations, structural shifts in the weight parameters, and a steady state relationship.

The majority of studies that deal with the combination of forecasts combined forecasts that have been derived from alternative sources with identical timing. A typical example is the combination of two or more forecasts of the monthly values of the variable of interest. Some particular studies examined the combination of forecasts when the timing is not the same. Cholette (1982) examined the use of benchmarks by modifying forecasts from autoregressive integrated moving average (ARIMA) models. This framework suggests that the monthly observations of some variable of interest (i.e. the U.S. unemployed rate) may be modelled by an ARIMA model, whereas a benchmark of the variable for the last quarter of the upcoming year may be provided by an outside source such as a group of economic experts or an econometric model. Cholette provided a numerical solution to address this problem by combining the outside

information with the monthly prediction made from the ARIMA model. Greene et al. (1986) investigated the problem of modifying quarterly forecasts from an econometric model using additional monthly information from an independent source. This approach starts with econometric forecasts and then considers the use of outside information to improve the forecast accuracy of the econometric model. Adopting this approach, Greene et al. showed that in a non-linear system optimal combining weights will vary over time. Trabelsi and Hillmer (1989) adopted a similar approach. However, rather than starting with econometric forecasts and considering the use of outside information to improve forecast accuracy of the econometric model as Greene et al., Trabelsi and Hillmer started with extrapolative forecasts and then considered the impact of an outside forecast like the econometric forecast. More specifically, they proposed a statistical-based approach to optimally combine forecast derived from an extrapolative model such as an ARIMA time-series model, with forecasts of a particular characteristic of the same series obtained from some independent sources. They investigated the implications of the approach in the context of seasonal ARIMA models and they provided an empirical example to illustrate their methodology.

Some studies have used non-linear functions to combine forecasts (Jacobs et al. 1991a; Zhang et al. 1992; Jordan and Jacobs 1993; Edelman, 1995). The motivations for using non-linear combining algorithms is that there is no a priori reason to believe that linear models are adequate to model the combination of arbitrary function approximators. It is expected that the component models will have different relative contributions in different parts of the space. Therefore, a linear function of the component predictions cannot track the selective contributions of the components (Skalak, 1997). Despite the efforts, however, non-linear combining algorithms have received much less attention.

Gunter and Aksu (1997) investigated Non-Negativity Restricted Least Squares (NRLS) combinations of forecasts and they examined various optimising and heuristic computational algorithms for estimating NRLS combination models. They empirically compared the combination weights identified by the alternative algorithms and their computational demands based on a total of more than 66,000 models that were estimated to combine the forecasts of 37 firm specific accounting earnings series. The ex-ante prediction accuracies of combined forecasts from the optimising algorithms were compared versus the heuristic algorithms. The results of this comparison suggested that computationally simple heuristic algorithms have very competitive performance compared to optimising algorithms. However, Gunter and Aksu reported that they are unable to draw any conclusions regarding on which algorithm should be applied based on series and forecast characteristics.

6.3.5 Voting Methods

The default method for classifier combination is majority voting (MV). Let us assume that we have $I_{it} = n$ classifiers and $C_{jt}(j = 1,2...k)$ classes. Further, let us denote by $g_{it}^{j}(j = H, L)$ the estimate of classifier i for class j at time t. In a voting scheme, we combine votes for classes through a weighted sum and then assign the input to the class C_{jt} receiving the maximum vote. Using the appropriate notation, we can express this relationship as shown below (Alpaydin 1993, 1998),

$$d_{it} = \sum_{i=1}^{n} g_{it}^{j} w_{it} \qquad w_{it} \geq 0 \qquad \sum_{i=1}^{n} w_{it} = 1$$
$$C_{jt} = \arg\max_{i=1}^{n}(d_{it}) \tag{6.8}$$

Unless there is an a priori reason to favour one voter over another weights are taken as equal $w_{it} = 1/n$. This is simple voting. If we take a simple MV rule and apply similar experts, whose predictions generally coincide, there is no way allowing a more expert to express a more valuable opinion over the other experts. In that case, we would prefer a consensus among heterogeneous experts.

In business forecasting, experience suggests that most benefits are captured by equally-weighted schemes where $w_{it} = 1/n$ (Diebold and Lopez, 1995). In artificial intelligence applications the evidence is unclear. Hansen and Salamon (1990) showed that applying a simple MV rule and using independent classifiers with success probability higher than 1/2, it will increase classification accuracy as the number of voting classifiers increases. Mani (1991) showed that a simple voting rule will decrease variance as the number of independent voters increases. Benediktsson and Swain (1992) proposed a voting scheme where data from different sources are integrated based on consensus theory.

Battiti and Colla (1994) compared several combination mechanisms for combining the classifications of multilayer perceptrons trained to recognize handwritten digits. The combination schemes they compared were unanimous voting (UV), MV, averaging raw output activation values, and several mechanisms based on thresholding the confidence of the component networks in their classification predictions. They found that MV outperforms a combination rule requiring unanimity of prediction among the components.

Ali and Pazzani (1996) combined stochastically generated rules and decision trees using voting and three probabilistic algorithms as combiners, namely Bayesian Combination, Distribution

Summation, and Likelihood Combination. Using twenty-nine data sets, they found that voting is the most accurate combiner.

Although voting methods seem a reasonable way to combine individual decisions, the empirical findings are not always impressive. Heath et al. (1996) found that increasing the set of components and voting their predictions does not reliably lead to higher generalisation accuracy. Skalak (1997) emphasised that the difference between a good classifier and a superior one is determined by their behaviours on difficult instances. Both will probably get the easy ones right. If an example is difficult to classify, then the average classifier will get it wrong, and taking a majority vote of classifiers that are more likely to get it wrong will not increase but may decrease classification accuracy.

Two other approaches to the voting framework are the Halving algorithm and Weighted Majority algorithm proposed by Littlestone (1987) and Littlestone and Warmuth (1989), respectively. A more detailed description of these algorithms can be found in Skalak (1997).

6.3.6 Non-voting Methods

The most important non-voting methods that are applied to classifier combination include nearest neighbour proposed by Wolpert (1992) and Dietterich and Bakiri (1995), ranking algorithms proposed by Ho et al. (1994), and an algorithm based on the Dempster-Schaffer theory of evidence proposed by Xu et al. (1992). A more detailed description of these studies can be found in Skalak (1997). Here, we will describe them in brief.

Wolpert (1992) used a nearest neighbour combiner in a single NetTalk example. Dietterich and Bakiri (1995) used an l_1 metric in error correcting output coding to compute the distance between each static codeword and a vector of the outputs of the classifiers used to predict each bit in the codeword. Ho et al (1994) attempted to amalgamate the predictions of classifiers by applying a consensus mechanism known as borda count. According to this mechanism each of the classifiers outputs a ranked list of predictions in order of the confidence of the classifier in each predicted class. Each class has a borda count given by the sum of the number of classes ranked below it by each classifier. The class with the highest borda count is the ultimate prediction.

The Dempster-Schaffer theory assigns a probability interval to sets of propositions instead of making a scalar probability estimate to a single proposition. The interval not only reflects the strength of the evidence in favor of the set of propositions, but it also reflects the amount of information that is available. A disadvantage of this framework, however, is that it requires a

large amount of computation (Skalak, 1997).

6.4 APPLICATIONS OF COMBINING FORECASTS

In this Section, we review the applications of combining forecasts. A very extensive review of a variety of applications of combining forecasts can be found in Clemen (1989). Here, we focus on economic applications, whereas we discuss other applications only in brief.

6.4.1 Combining Macroeconomic Forecasts

A variety of studies investigated the accuracy of combining macroeconomic forecasts. Most of these studies have used data from the Blue Chip Economic Indicators (BCEI) service run by Bob Eggert who has published consensus macroeconomic forecasts since 1976. Some of the most recent studies that have used the BCEI service are those by McNees (1987, 1992), Batchelor (1990), and Batchelor and Dua (1990a, 1990b, 1992, 1995). McNees (1987) used five quarter ahead forecasts of 23 participants in the BCEI service. He reported that averages of large numbers of published forecasts are outperformed by a sizeable minority of their individual components. In a subsequent study, McNees (1992) used 20 Blue Chip participants plus two prominent commercial forecasting services that did not participate in Blue Chip Economic Indicators service. The 22 individual forecasts plus the consensus were used to provide forecasts of seven variables from 1978 to 1988. The results of this study suggested that approximately one-third of individual forecasters are more accurate than the consensus for each specific variable. However, the results also suggested that the individual forecasters' performance varies widely across variables and even though everyone can beat the consensus for some variable no one is superior for all variables.

Batchelor and Dua (1992) used the track record of a panel of 14 U.S. Blue Chip economic forecasters to test for conservatism and consensus-seeking behaviour in the published forecasts of four variables namely, the real GNP growth, the inflation in the GNP deflator, the unemployment rate, and a short-term interest rate. Conservatism in forecasting occurs when an individual revises his forecast on the arrival of new information by less than would be suggested by Bayes' Theorem (Phillips and Edwards 1966; Edwards 1968). Consensus-seeking behaviour bias occurs when an individual revises his forecast towards that made by a group of forecasts, by more than would suggested by Bayes' Theorem (Asch 1951). Batchelor and Dua (1992) used data for 125 forecasting months, and 11 target years. The results of this study suggested that economic forecasters are conservative rather than consensus-seeking. In other words, most forecasts made by members of the Blue Chip panel might have been improved if the forecasters are prepared to move their forecasts towards the consensus and less attached to their own

forecasts when making monthly revisions. However, Batchelor and Dua (1992) suggested that the demand for economic forecasts may be related to factors other than accuracy. For example, end-users may mistrust forecasters who make frequent revisions of their forecasts or produce figures very different from the consensus. Therefore, suppliers of economic forecasts may be willing to adjust their forecasts in order to support their credibility. In this sense, conservatism and consensus-seeking behaviour may be regarded as symptoms of irrationality and these biases may be inconsistent with a rational application of Bayes' Theorem. This proposition is explained in more detail by Batchelor and Dua (1990b). In a subsequent study, Batchelor and Dua (1995) introduced a measure of the benefit from combining to take into account the variation in the performance of combined forecasts. This measure is the probability of a reduction in error variance. Starting from the observation that combining a set of unbiased forecasts will on average increase forecast accuracy, Batchelor and Dua noted that this approach might be risky because the precise set chosen by the user might not be as satisfactory as a randomly chosen individual forecast. Using data of a panel of 22 U.S. economic forecasters and a survey of their econometric methods provided from the BCEI service, they examined if such risk can de reduced by combining forecasts in a nonrandom way. The results of this study suggested that the average benefits from combining published economic forecasts are lower than those achieved by pure time-series. Batchelor and Dua suggested that one explanation for these results might be that some of the most inaccurate forecasts are not always published. Another explanation might be that individual forecasters typically use a variety of forecasting techniques and, therefore, the forecasts that they publish can be considered as averages of the forecasts produced by "textbook" techniques. Another important finding of this study was that there are substantial benefits from combining even a small number of forecasts that are produced by different theories, whereas there are also benefits by combining the same number of forecasts using different forecasting techniques.

Holden and Thomson (1997) investigated similarities and differences between the optimal combination of forecasts, forecast encompassing, and forecast efficiency. For this study, they used forecasts from four different major U.K. macroeconomic forecasting organisations namely, Cambridge Econometrics (CAM), Liverpool University Research Group in Macroeconomics (LPL), the Centre for Economic Forecasting at London Business School (LBS), and the National Institute of Economic and Social Research (NI). The CAM model is as a large annual Keynesian model, the LPL model is a small annual neo-classical expectations model, the LBS model is a medium sized international monetarist quarterly model, and the NI model is a quarterly Keynesian income expenditure model with some rational expectations and monetarist influences. Holden and Thompson analysed the forecasts that are produced in October/November for the current year t and years t+1, t+2, t+3 and t+4 for 1981-94. This time

period was decided as the most appropriate for comparisons because all forecasts were made available at around the same time. The variables considered for this study were the growth of real gross domestic product (percentage terms), the rate of inflation (percentage terms), and the rate of unemployment (millions). To evaluate the forecasts, Holden and Thompson used the root mean square forecast errors (RMSEs) and the mean absolute errors (MAEs). After comparing the forecasts, they found that the forecasts are mainly unbiased, whereas the correlations of the forecasts are high for the current year and decline with the forecast horizon. In terms of forecasting accuracy, they reported differences in the accuracy of the forecasts with the accuracy declining while moving from year t to year t+4. On the other hand, they found that the arithmetic mean of the four forecasts is superior to the individual forecasts, whereas no one forecasting organisation dominates for every variable.

6.4.2 Other Economic Applications

Cacciatore and Nowlan (1994) used a recurrent network to gate a set of experts which are linear or non-linear controllers. If x represents the input vector and y represents the output vector, then the model states that the indicator variable v_j^n for each expert obeys a first-order Markov assumption stated as $P\left(v_j^n = 1 \setminus v^{n-1}, x^n\right)$. The likelihood of this model is maximised through an on-line gradient descent procedure. Cacciatore and Nowlan suggested that this model is more capable to control time series that contain switching or jump effects than the standard ME model (Waterhouse, 1997).

Waterhouse and Robinson (1995) used linear regression experts gated by a multinomial logit model to predict future values of a sunspot time series. They reported results comparable to multilayer perceptron models. Weigend et al. (1995) designed an architecture in which they used gates and experts. They applied this architecture to various problems including time-series prediction and they reported satisfactory results.

Pawelzik et al. (1996) used an ensemble of radial basis functions that described it as a mixture of experts (ME) without a gating network. They applied this algorithm to the segmentation of time-series and they derived a weighting factor for each expert based on its likelihood of explaining the current sample in the time-series. After this implementation, they reported good segmentation of switching time-series and better prediction accuracy on a real time-series.

Zeevi et al. (1997) developed a universal approximation proof for ME which suggests that the ME are capable to model the same function as multilayer perceptron models. They applied this model to time-series prediction and reported comparable results with multilayer perceptrons.

Kehagias and Petridis (1997) applied predictive modular neural networks for segmentation of time series into distinct regions. They also reported satisfactory results.

Combinations of forecasts have been applied in a variety of other real world forecasting economic applications such as predicting the gross national product (Reid 1968); inflation (Engle et al. 1985; Hafer and Hein 1985); money supply (Mills and Stephenson 1987); exchange rates (Bilson 1983; Blake et al. 1986; Guerard 1989); forecasting and trading currency volatility (Dunis et al. 2000; Dunis and Huang 2001); stock prices (Stael Von Holstein 1972; Virtanen and Yli-Olli 1987); corporate earnings (Cragg and Malkiel 1968; Elton et al. 1981; Conroy and Harris 1987; Guerard 1987; Guerard and Beidleman 1987; Newbold et al. 1987); sales forecasting (Sewall 1981; Moriarty and Adams 1984; and Schnaars 1986a-b).

6.4.3 Other Applications

Other applications of combining forecasts are described by Clemen (1989) and include among others applications in meteorology (Stael Van Holstein 1971; Winkler et al. 1977; Clemen and Murphy 1986a-b; Murphy et al. 1988); prediction of social and technological events (Kaplan et al. 1950); forecasting of city populations (Schmitt 1954); psychiatric diagnosis (Goldberg 1965, 1970); prediction of football game outcomes (Winkler 1971); prediction of livestock prices (Bessler and Brandt 1981; Brandt and Bessler 1981, 1983; Bessler and Chamberlain 1987); electrical demand (Bunn 1987, Smith 1989); tourism (Reinmuth and Geurts 1979; Fritz et al. 1984); insurance (Taylor 1985); political risk (Bunn and Mustafaoglu 1978); and Sunspot cycles using Wolf data (Morris 1977; Poskitt and Tremayne 1986).

6.5 GENERAL COMPARATIVE STUDIES IN COMBINING FORECASTS

6.5.1 Empirical Review

Newbold and Granger (1974) examined the relative performance of a number of forecasting techniques. After a number of experiments, they demonstrated that in-sample weighted average forecast outperforms either the individual forecasts or a simple average. Their findings suggested that it is usually better to ignore the effects of correlations in estimating combining weights. A more extensive comparison of forecasting methods was performed by Makridakis et al. (1982, 1983) as part of a forecasting competition where a variety of time-series forecasting methods were applied to 1001 different economic time-series. This competition was concerned mainly with the post-sample forecasting accuracy of extrapolative time-series methods but combining schemes were also applied. The results from combining forecasts demonstrated that a simple average of six individual methods performs better than the individual methods

included in the combination. A weighted average of the same methods based on the sample covariance matrix of fitting errors performs well but not as well as the simple average. Armstrong et al. (1983) prepared an extensive commentary on this competition. Several experts were invited to participate in this commentary and they were asked to write their commentaries on the Makridakis competition. The objective of this effort was to provide a forum for discussion by experts who were likely to have different perspectives on the results presented in the competition. One of the suggested topics for discussion was the effectiveness of the methods that were applied to combine forecasts. The individual experts expressed different opinions in the combination schemes that were tried in the competition. Gardner (1983) disagreed with Makridakis' view that a simple average of six individual forecasts should be preferred over the individual forecasts. He emphasised that this method is superior to individual components in terms of Mean Average Percentage Error (MAPE) but not in median Average Percentage Error (APE) or Mean Square Error (MSE). He observed that Holt or Holt-Winters methods do about the same as a simple combination of six forecasts in terms of median APE at all horizons. Furthermore, he identified maintenance problems associated with running six different methods at the same time. These maintenance problems may arise because four of the six methods use fixed parameters. Therefore, if repetitive forecasts are made over time, then the fixed-parameter methods would have to be refitted periodically. To adjust for outliers and bias, these methods should be monitored with tracking signals between the refittings. Winkler (1983) disagreed with Gardner's view and supported Makridakis' recommendation that the combining schemes should be preferred over the individual forecasts. He observed that pairwise comparisons between the combining schemes and the individual methods suggested that the combining schemes outperform the individual methods in a large number of cases, whereas the individual methods outperform the combining schemes in a smaller number of cases. Overall, the pairwise comparisons suggested that the performance of the combining schemes are more robust than the individual forecasts. Geurts (1983) also emphasised in his report that the results from combining individual forecasts are very promising but he raised the question about the choice of the method to combine forecasts. He observed that the models that were combined using a simple average scheme were five exponential smoothing models and the Carbone-Longini filtered method (1977). This raised the problem that the weights from past data in the forecasting equation are not necessarily confined to an exponential weighting scheme. The inclusion of a single smoothing method that performs poorly when trend is present in the time-series might affect the effectiveness of the combining method. In his own view, Geurts suggested that combinations of forecasts from Box-Jenkins, Bayesian, FORSYS and ARARMA models should also be examined in a competition. Furthermore, he seemed to be surprised about the superiority of the simple average scheme to combine forecasts over the weighted average scheme. The argument is that if equal weighting is the optimum weighting scheme, then the

weights based on the sample covariance should have generated the equal weighting scheme. Carbone (1983) disagreed with this view and he emphasised that reality sometimes differs from theory and the often noted inclination to question data is rather unfortunate. In a later study, Winkler and Makridakis (1983) used the weighting techniques of Newbold and Granger (1974) to generate combined forecasts for the series used in the competition. After a series of experiments, they found that the techniques relating the weights to reciprocals of sums of squared errors outperform those relating the weights to an estimated covariance matrix of forecast errors. Furthermore, the improved weighted schemes perform better than a simple average. In a different study, Makridakis and Winkler (1983) found that a simple average of forecasts derived from several different forecasting techniques generally perform better than forecast of any single model in terms of mean absolute percentage deviations.

Zarnowitz (1984) examined the accuracy of a large number of individual forecast series and of the corresponding average forecast series from a quarterly survey that he conducted for the National Bureau of Economic Research in collaboration with the American Statistical Association (ASA). The survey questionnaire was mailed by the ASA in the middle month of each quarter to a list of professionals involved in forecasting the course of the economy. The regular reports on the results were released in the third month. A variety of professionals were polled in this survey such as economists, independent consulting firms, government agencies, academic and research organisations. The study covered 79 individuals who participated in at least 12 of the 42 surveys in the period from the fourth quarter of 1968 to the first quarter of 1979. The forecasts related to rates of change in four variables, namely gross national product in current dollars, gross national product in constant dollars, the GNP implicit price deflator, and the consumer expenditures for durable goods. The survey averages included many econometric and judgmental adjustments. The results of this study suggested that the group mean forecasts from a series of surveys are on average more accurate than most of the corresponding sets of individual predictions. This conclusion applies to all variables and predictive horizons. The results also suggested that forecasts from the major econometric model services are influenced positively from judgmental adjustments. In an extended literature review on combining forecasts, Armstrong (1986) found forecast error reductions which varies from zero to 23%, whereas summarising across eight different studies he reported that the unweighted average error reduction is 6.6%.

Faria and Souza (1995) applied the combining expert opinions to the forecasts of petroleum prices generated by a group of experts from Petrobras, the Brazilian state-owned oil company. They studied the following international markets: USA, North Sea, Russia, Nigeria, and the Persian Gulf. The time horizons they considered in this study were 4, 8 and 12 weeks. The three

methods they used in the combination of forecasts was optimal combination (Bates and Granger, 1969), outperformance, and quasi-Bayes. These methods were applied to 15 data sets, each data set consisting of six forecasted series referred to one of the five markets and one of the three forecasting horizons. The relative performances of both individual forecasters and combination methods were calculated from the forecast errors and real prices by means of the Square Root of the Percentage Relative Medium Square Error (PRMSE) given as follows,

$$\sqrt{\text{PRMSE}} = \left[\frac{1}{\tau}\cdot\sum_{i=1}^{n}\left(\frac{\varepsilon_t}{Y_t}\right)^2\cdot 100\right]^{\frac{1}{2}} \tag{6.9}$$

where Y_t is the real average price at week t, $\varepsilon_t = Y_t - \tilde{Y}_t$ is the forecast error, and τ is the total number of observations.

For the classical optimal combination method, Faria and Souza used a method of optimal combination that assumes independence between individual forecasts and calculates the weight w_i in the actual time as follows,

$$w_{i,\tau} = cw_{i,\tau-1} + (1-c)\frac{(s_i)^{-1}}{\sum_{i=1}^{n}(s_i)^{-1}} \tag{6.10}$$

$$s_i = \sum_{d=\tau-v}^{\tau-1}(\varepsilon_{i,d})^2 \; ; \; \varepsilon_{i,d} = y_d - \tilde{y}_{i,d} \; ; \; \tilde{y}_{i,d} = E\left[\phi_i(y_d)\right] \; , \; i = 1,2\ldots n. \tag{6.11}$$

where c is the weight given to the past w_i, and v is the number of past errors to be introduced in the calculation of s_i.

Faria and Souza reported that the optimal combination is the most robust and stable method even in the worst case of inadequate choice of parameters. The Bayesian methods perform well, whereas the quasi-Bayes method presents the smallest error for the horizon of 12 weeks. Faria and Souza observed that these results confirm the expectation that the quasi-Bayes method performs well over large variances for the individual predictive distributions. Faria and Souza concluded that it seems reasonable in practice to use the optimal combination with optimal parameters for horizons of 4 and 8 weeks ahead and the quasi-Bayes intervention for 12 weeks ahead.

6.6 SUMMARY OF THE EMPIRICAL EVIDENCE IN COMBINING FORECASTS

6.6.1 Summary and Discussion

In this part, we reviewed the literature on combining forecasts. We discussed design considerations for constructing composite architectures and we reviewed a variety of existing models and methodologies on combining forecasts. We also presented a variety of applications on combining forecasts and we discussed comparative studies on combining forecasts. The empirical evidence that we presented in this part suggests that combining individual forecasts improves forecast accuracy. The conclusion holds true in statistical forecasting, judgmental estimates as well as averaging statistical and subjective predictions. Empirical results suggest three factors that encourage combinations of forecasts: first, combining individual forecasts improves forecasting accuracy and reduces the variance of forecasting errors; second, simple combination models work as well as more complex combinations; and third, combining can be done with little or no increase in cost. However, we have to emphasise that the results that we have presented from various studies on combining forecasts are conditional on the information sets that various researchers have used to implement the particular combining models as well as the way they have implemented the models. Therefore, these results should be interpreted with caution.

Furthermore, we have to observe that although the majority of studies applied composite architectures for regression, only a few studies investigated composite architectures for classification. In the next part, we investigate the applicability of combining classifiers in our study and we present a new methodology to predict high performing shares by combining five heterogeneous classifiers namely LDA, PNN, LVQ, OC1, and RRI using voting techniques.

Part Two: Combining Heterogeneous Classifiers to Predict High Performance Stocks

In this part, we propose a new methodology to predict high performance stocks by combining five heterogeneous classifiers namely, LDA, PNN, LVQ, OC1, and RRI using Majority Voting (MV) and Unanimous Voting (UV) schemes and applying the composite architectures to classify stocks that are likely to have exceptional returns in the future. The motivation for combining heterogeneous classifiers through MV and UV schemes is the possibility that by combining a set of heterogeneous classifiers, we may be able to perform classification better than using the individual classifiers. This motivation is supported by the idea that a group

decision based on "different" experts may be more reliable on average than the decision of the individual expert or the decision of "similar" experts. It is expected that a composite classifier architecture which combines heterogeneous component classifiers will be able to hedge the classification bets better than the individual classifiers or a composite architecture which combines homogeneous component classifiers. In this sense, a "portfolio" of classifiers is analogous to a "portfolio" of financial instruments where diversification is used to reduce risk (Skalak, 1997). Furthermore, combinations of heterogeneous classifiers may allow for greater flexibility in the model switching sense to utilise a broader information set. In this content, a combination of diverse models may be able to utilise a broader information set than a combination of similar models. However, a better utilisation of the information set may lead to an increase in forecasting accuracy.

This part is organised as follows: In Section 6.7, we discuss the data and methodology that we used to implement our classification methods. For this implementation, we used the same data and the same data preprocessing techniques that we used to implement the five classifiers to predict high performing shares in Chapter 5. We implemented the classifiers separately and we also combined their forecasts using MV and UV schemes whereby a share is not assigned to the high-performing portfolio unless either the majority of classifiers or all classifiers, respectively, agree on their predictions. In Section 6.8, we report the results of experimentation and we discuss the economic implications of our findings. Our results suggest that the UV scheme produces significant improvements in both classification and profitability over the individual classifiers and reduces substantially the trading volume. In Section 6.9, we discuss the practical implications of our study and we also discuss the possibilities for further improvements in our methodology.

6.7 DATA AND METHODOLOGY

In this application, we are particularly interested in whether a particular share will be classified as H or L excess return share based on accounting information. Let us assume that y_{it} is the 1-year-ahead excess return on some share i bought at time t, and x_{it} is the vector of accounting information attributes for company i known at time t. The idea is to apply the five classification methods namely LDA, PNN, LVQ, OC1, and RRI, to assign y_{it} to one of the two classes C_{it} = H or L and then combine their predictions of these classifiers using MV and UV schemes. We recall that the ranking of the shares as H and L performing shares in a given year was decided on whether or not their excess returns is above or below the 25% threshold percentile that has been decided after ranking the returns in excess of an equally-weighted index. The models input is the vector x_{it} of variables that represent current month accounting information.

To perform this experiment, we used the same data and the same data preprocessing techniques that we used in Chapter 5. Therefore, our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This data consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our predictor variables are 38 accounting ratios drawn from published accounting statements. The input variables for each individual classifier were reduced after applying the stepwise variable elimination procedures that we described in Chapter 5. Therefore, the list of variables that were finally selected for the implementation of each individual classifier is similar to the one presented in Table 5.2. These data were normalized using Eq. (4.1).

The general hypothesis examined in this part can be stated as follows:

The ability of any classifier to predict high performing shares can be improved or exceeded through composite classifier architectures that combine a small number of heterogeneous classifiers using voting procedures.

We investigate two voting schemes to combine component classifiers: MV and UV schemes. In both cases, the voters have equal weight in their predictions. Assuming that we have $g_{it}(i=5)$ classifiers and $C_{jt}(j=H,L)$ classes, we assign a share to the H class if the following conditions hold,

$$V_t = \sum_{i=1}^{5} w_{it} g_{it}^{H} \geq V_t^{*} \tag{6.12}$$

$$w_{it} \geq 0, \quad \sum_{i=1}^{n} w_{it} = 1 \tag{6.13}$$

where w_{it} are weights on each prediction, g_{it}^{H} is the estimate of classifier i for class H, and V_{it}^{*} a threshold value which is equal to $V_{it}^{*} = 3$ if the MV rule is applied, whereas $V_{it}^{*} = 5$ if the UV rule is applied. If inequality (6.12) does not hold, then we assign the share to class L. Unless there is an a priori reason to favour one voter over another weights are taken as equal $w_j = 1/5$. This is simple voting. In this experiment, we have not applied weighted voting schemes because the in-sample performance of the classifiers was broadly similar.

Skalak (1997) proposed three design criteria for building composite classifiers: accuracy, diversity, and efficiency. A brief explanation of these criteria is given below,

1) The accuracy criterion ensures that the component classifiers that are used to build the composite classifier are also accurate when applied individually to make predictions. If the predictions of the component classifiers that are being combined are not accurate, then the ultimate prediction of the composite classifier might not be highly accurate.

2) The diversity criterion ensures that the component classifiers make diverse errors when applied individually to make predictions. If the predictions of the component classifiers are exactly the same, then a combination of these predictions will not improve the predictive accuracy of the composite classifier. If the individual components make some different predictions, then there is a hope that combining their predictions using the appropriate mechanisms might improve the predictive accuracy of the composite architecture.

3) As far as concerns the efficiency of the composite classifier, two principles should be taken into account: first, we have to prefer fewer component classifiers over more; and second, we have to prefer computationally inexpensive ones over expensive ones. These principles ensure that the composite classifier is not only accurate but it is also fast in reaching the ultimate prediction.

Taking into account these criteria, we combined the five classifiers namely LDA, PNN, LVQ, OC1, and RRI to predict high performing shares for three main reasons: first, all classifiers were found accurate when they were applied individually; second, these classifiers are representatives of different model families and this ensures diverse component models in our composite architecture; and third, we prefer fewer components over more to keep the computational complexity as minimum as possible. We have chosen MV and UV schemes to combine the five classifiers because they are simple to implement and they can be applied with no additional computational cost.

To implement the five classifiers namely LDA, PNN, LVQ, OC1 and RRI, we followed implementation strategies exactly similar to those that we described in Chapter 5. Following these strategies, we used two years of data to predict the next year. For example, to predict relative excess returns for 1993, we first trained the classification methods on the two preceding years 1991 and 1992 using cross-validation procedures, such as the leave-one-out method and other rotation procedures, to find the optimal values of parameters for each individual classifier. After selecting the optimal values of parameters, we applied the classification methods to

predict the out-of-sample year 1993 and we combined their predictions using the MV and UV techniques. We then moved the implementation one-year ahead and we used information available from 1992 and 1993 to predict 1994 and so on. We use only two previous years of data to predict relative excess returns for the next year because we believe that only recent accounting information may be relevant to predict relative excess returns in the next year.

We compared the five classifiers and the two voting methodologies in terms of classification accuracy, profitability, and trading volume but we also considered the trade-off between predicted returns and risk. To evaluate the profitability of the five classifiers and the two voting methodologies, we calculated average returns and excess returns over the index for the portfolios of actual H and L shares in our data in all the 12-month holding periods starting each year, and then we compared them with the respective averages for the portfolios of H and L shares predicted by the classification methods. To examine if transaction costs can have an important impact in our trading system, we also compared the classification methods and the two voting methodologies for the predicted number of shares included in the portfolios of H performing shares that are traded in the target years 1993-97.

6.8 RESULTS

In this Section, we summarise the results of experimentation. We compared the five algorithms and the two voting methodologies in terms of classification accuracy, profitability, and trading volume for the test out-of-sample year 1993 - on which the variable reduction was conducted - as well as for the genuine out-of-sample years 1994-97.

Table 6.1 shows the classification results after implementing the classifiers and the voting methodologies for the test out-of-sample target year 1993 as well as for the genuine out-of-sample years 1994-97. These results are also presented in Figure 6.1. As we can see, the UV methodology outperforms significantly the MV and the individual classifiers for the target years 1993-97. PNN, LDA and MV produce very good results for the target year 1993 but is seems that their classification performance deteriorates for the next years compared to the other classifiers. As we can see, the OC1 is the second best classifier for the target year 1994 with a very balanced predictive accuracy for both H and L performing shares as we can see in Table 6.1. The OC1 classifier is also the second best classifier for the target year 1995 even though its classification accuracy seems to favour more L performing shares against H performing shares compared to LDA, PNN, and MV that favour more H performing shares against L performing shares. The LVQ is the second best classifier for the target year 1996 but it seems that the RRI predicts more accurately than the LVQ H performing shares. The pattern is slightly different for the target year 1997 where the PNN and the MV have the second best classification

performance and achieve a very good predictive accuracy for both H and L performing shares.

Table 6.2 shows the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the classifiers and voting methodologies using accounting information only. These results are also presented in Figures 6.2-6.5 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, all the classifiers and voting methodologies produce positive returns and excess returns. However, the UV methodology outperforms significantly the other classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 and produces the highest financial results. PNN, LDA and MV produce very good results for the target year 1993, whereas it seems that the financial returns deteriorate for the next target year 1994 even though all classifiers and voting methodologies produce positive results. The target year 1995 is a year of high profitability for all classifiers and voting methodologies. The UV rule produces the highest financial results in this year, whereas there are only minor differences in the financial results produced by the other classifiers. The financial returns deteriorate slightly for the next target years 1996 and 1997 that are years of positive financial returns as well.

Although the financial returns are a primary factor to evaluate a particular trading system, we should also examine the transaction costs involved in trading the number of shares predicted by the classification methods. Empirical evidence suggests that the financial profits that emerge from quantitative models might be eaten up if trading occurs more than once or twice per year (Pesaran and Timmermann 1994, 1995). Taking into account these findings, we calculated the predicted number of shares included in the portfolios of H performing shares that are traded for the target years 1993-97. This is the number of actual H performing shares that were correctly predicted as H by the classification methods as well as the number of actual L performing shares that were incorrectly predicted as H.

Table 6.3 compares the classification methods for the number of shares predicted to be H in all the 12-month holding periods for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97. Figure 6.6 gives a graphical illustration of the results. As we can see in Figure 6.6, the UV rule produces the smallest trading volume for the target years 1993-97 and outperforms significantly the other classification methods. However, we should emphasise that according to our trading system, if a share is classified as H, we buy equal amounts of this share at the end of the reporting month and we hold it for one year. The profitability of each classification method is therefore calculated by the cumulative profits generated by the resulting

portfolio of H performing shares only. The benefit of this approach is that it minimises transaction costs while it is not affected by price fluctuations around the reporting date. Given that there are 157-180 H shares each year, each share is traded at most once per year and trades can be done at the end of the month in a basket of no more than 13-16 shares bought and sold in the ideal trading strategy if one of the five classifiers or the MV methodology is applied. As we can see, the UV methodology results in additional reductions in the trading volume. Therefore, the transaction costs are not expected to affect the UV methodology as well as the other classification methods by more than 2% per year on average.

Once more, we have to mention that the above findings are vulnerable to the conclusion that high returns are achieved at the expense of high risk. Table 6.4 shows averages of selected attributes affecting risk for the actual and predicted H portfolios from the PNN for the out-of-sample year 1995. As we can see, the actual H portfolios contain a number of small stocks that have smaller-than-average market capitalisation. Therefore, these stocks are expected to be riskier and less liquid. Furthermore, the results suggest that the actual and predicted H companies have a high debt to equity ratio and low current after tax profit. They tend to pay low dividends and to be growth companies rather than value companies. Growth companies are characterised with a high market-to-book value.

To adjust for risk, we used the CAPM to obtain estimates of the betas (gearing to the equally-weighted index) and alphas estimates (excess returns not due to gearing) for actual H and L portfolios for the five classifiers and the two voting methodologies for the out-of-sample year 1995. These results are presented in Table 6.5. As we can see, the actual H portfolio does have a high beta of around 1.4. The predicted H portfolios from LDA, PNN and UV have also a high beta of around 1.4, whereas the other classification methods produce a beta very close to 1. In terms of alphas estimates, OC1, LVQ and UV produce very similar results that are much better than the other classification methods which do not produce the same good results. The PNN H portfolio has a return of 19.0% but a beta of 1.38 giving an excess return of 3.5%, whereas the OC1 H portfolio has a return of 18.9% with a beta of 0.9 giving an excess return of 8.1%. On the other hand, the UV H portfolio has a higher return of 24.8% and a much higher beta of 1.39 giving a risk-adjusted return of 9.3%.

6.9 SUMMARY OF THE RESULTS

In this part, we investigated the potential to apply composite classifier architectures to predict high performing shares by combining five heterogeneous classifiers, namely LDA, PNN, LVQ, OC1 and RRI using MV and UV techniques. Our target data were total returns on all shares

traded on the London Stock Exchange in the years 1993-97, whereas our predictor variables were accounting ratios drawn from published accounting statements. After experimentation, we found that using MV to combine the classifiers does not improve classification accuracy and profitability, whereas using UV to combine the classifiers improves substantially overall accuracy and profitability and this improvement is achieved with lower transaction costs. However, we emphasised that findings like this may be vulnerable to the conclusion that high returns are achieved at the expense of high risk. After adjusting for market risk using the CAPM, we found that LDA and PNN are less attractive in terms of risk-adjusted returns compared to OC1 which achieves a better deal of risk adjusted returns.

Part Three: Summary and Conclusions

6.10 DISCUSSION AND REMARKS

During the past years, researchers in machine learning, computational learning theory, pattern recognition, and statistics have applied composite architectures in a variety of application domains that require classification and prediction. The primary conclusion of this research is that combining individual forecasts improves forecasting accuracy and reduces the variance of forecasting errors. Furthermore, simple combination models often work reasonably well relative to more complex combinations. Moreover, combining can be done with little or no increase in cost. These conclusions hold true in statistical forecasting, judgmental estimates as well as averaging statistical and subjective predictions (Clemen 1989, Makridakis 1989).

Apart from lower forecast errors, combinations of forecasts allow for greater flexibility in the model switching sense to utilise a broader information set. In this content, a combination of diverse models may be able to utilise a broader information set than a combination of similar models. However, a better utilisation of the information set may lead to an increase in forecasting accuracy. Therefore, forecast diversity may be an essential technique in improving forecasting accuracy.

In the recent years, the interest in combining forecasts has been extended to classification problems. Many researchers have applied composite classifier architectures and mixture models to various financial applications. However, although there have been a number of successful efforts at combining homogeneous classifiers including multiple decision trees, multiple neural networks, and nearest neighbour algorithms, searching the more complex space of sets of heterogeneous component classifiers has not been shown necessarily to achieve high accuracy.

The attraction of combining heterogeneous models is very obvious. If the predictions of the component classifiers are exactly the same, then a combination of these predictions will not improve the predictive accuracy of the composite classifier. On the other hand, if the individual components make some different predictions, then there is a hope that combining their predictions using the appropriate mechanisms might improve the predictive accuracy of the composite architecture. Furthermore, a composite architecture based on heterogeneous component models might be more flexible to capture the structure of a data set which may consists of sub-regions with different underlying processes. For example, if the data set is non-linear, then a composite architecture that combines linear components may not perform well. On the other hand, a composite architecture that combines both linear and non-linear components will be more flexible to capture the underlying structure of the data.

In this Chapter, we investigated the potential to apply composite classifier architectures to predict high performing shares by combining five heterogeneous classifiers, namely LDA, PNN, LVQ, OC1 and RRI using MV and UV techniques. We found that using MV does not improve classification accuracy and profitability, whereas using UV to combine the classifiers improves substantially overall accuracy and profitability and this improvement is achieved with lower transaction costs. However, we emphasised that findings like this may be vulnerable to the conclusion that high returns are achieved at the expense of high risk. After adjusting for market risk using the CAPM, we found that LDA and PNN are less attractive in terms of risk-adjusted returns compared to OC1 which achieves a better deal of risk adjusted returns.

The results we presented so far are of course conditional on the information sets that we have used, the way we have implemented the various models, and the trading rules we have assumed. Therefore, these results may be improved if a number of improvements is considered. One possible improvement may be to use more general economic indicators and patterns in the time series of share prices in addition to the set of company-specific indicators. Another possible improvement may be to apply more sophisticated data preprocessing techniques such as variable transformations, and data winsorisation techniques. These improvements are discussed in detail in the next Chapter.

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
Actual Class	Patterns	Predicted Class Membership													
1993		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	163	98	65	96	67	90	73	88	75	91	72	96	67	32	131
L	488	128	360	123	365	154	334	163	325	149	339	123	365	33	455
Overall (%)		70.35 %		70.81 %		65.13 %		63.44 %		66.05 %		70.81 %		74.81 %	
1994		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	163	92	71	84	79	84	79	99	64	94	69	93	70	34	129
L	488	212	276	199	289	206	282	193	295	233	255	193	295	63	425
Overall (%)		56.53 %		57.30 %		56.22 %		60.52 %		53.61 %		59.60 %		70.51 %	
1995		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	173	106	67	103	70	90	83	98	75	99	74	101	72	42	131
L	519	208	311	193	326	188	331	183	336	233	286	194	325	64	455
Overall (%)		60.26 %		61.99 %		60.84 %		62.72 %		55.64 %		61.56 %		71.82 %	
1996		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	188	106	82	100	88	105	83	96	92	113	75	104	84	45	143
L	561	262	299	254	307	216	345	216	345	253	308	234	327	80	481
Overall (%)		54.07 %		54.34 %		60.08 %		58.88 %		56.21 %		57.54 %		70.23 %	
1997		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	188	105	83	103	85	100	88	100	88	96	92	101	87	41	147
L	564	214	350	199	365	203	361	212	352	220	344	196	368	58	506
Overall (%)		60.51 %		62.23 %		61.30 %		60.11 %		58.51 %		62.36 %		72.74 %	

Table 6.1: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

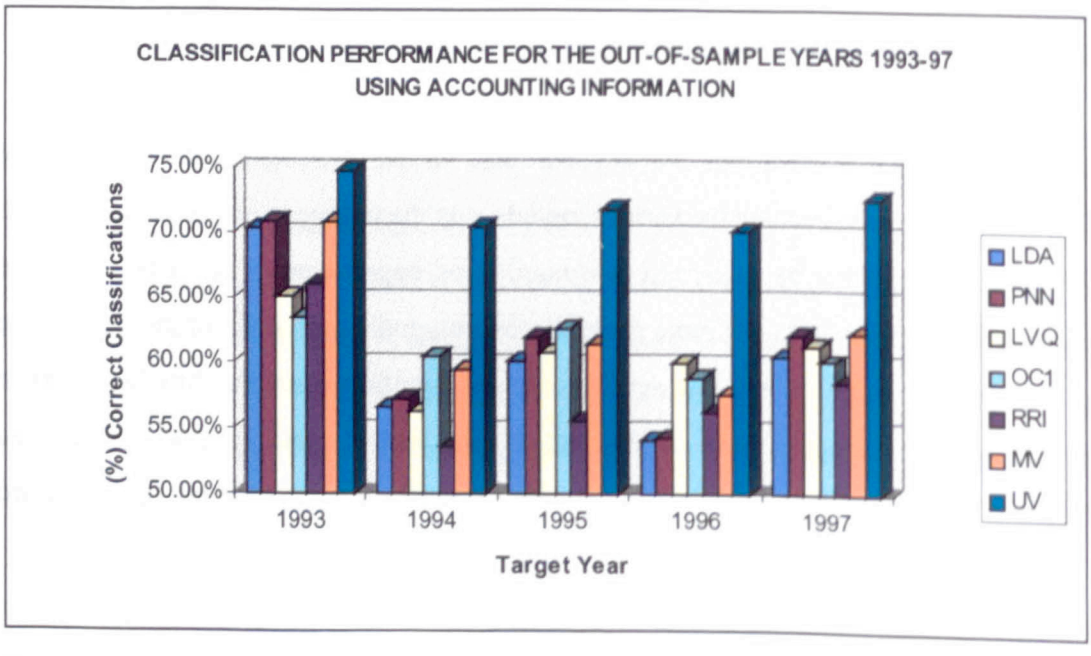
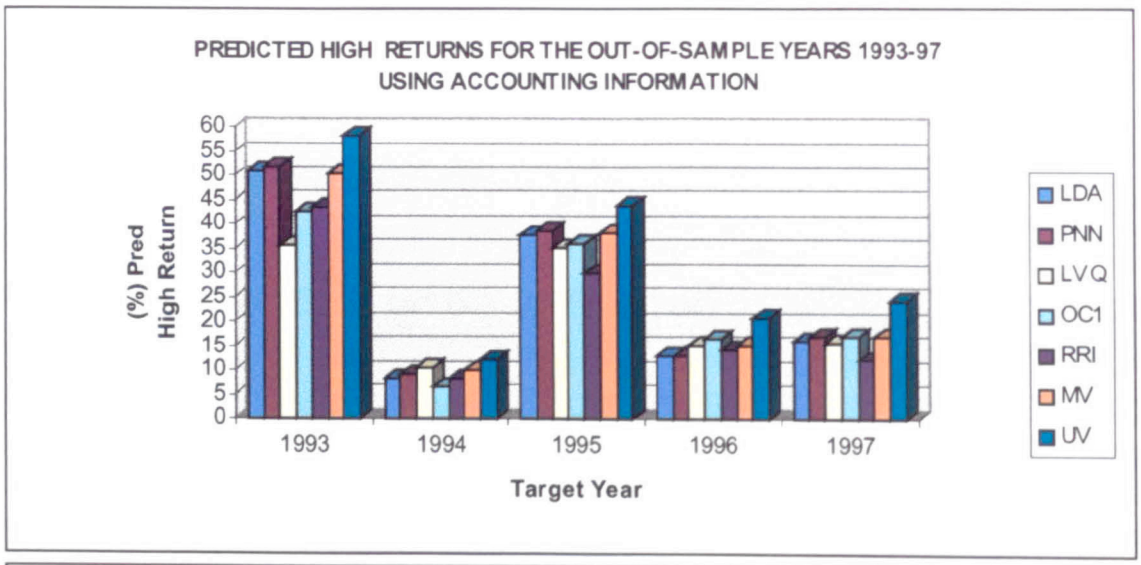
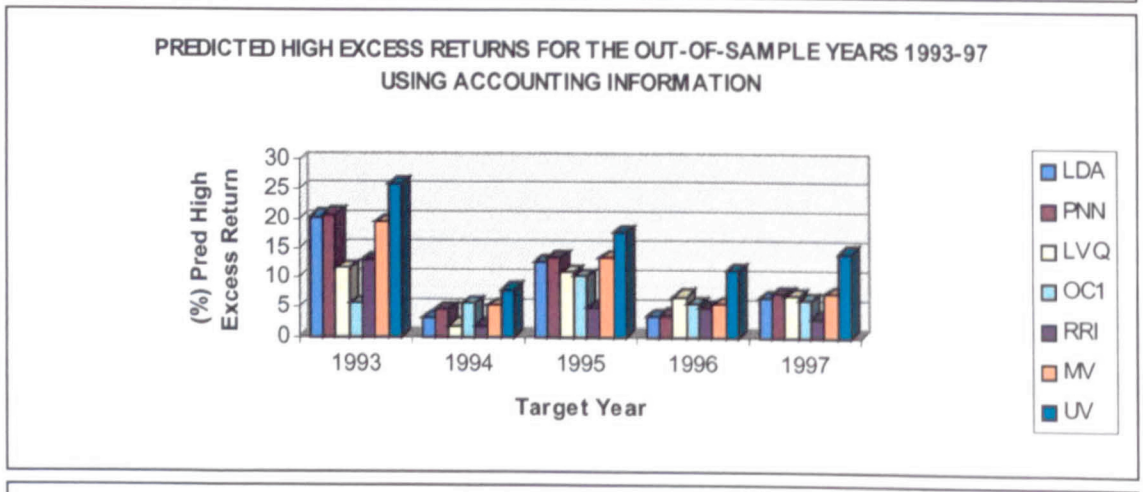


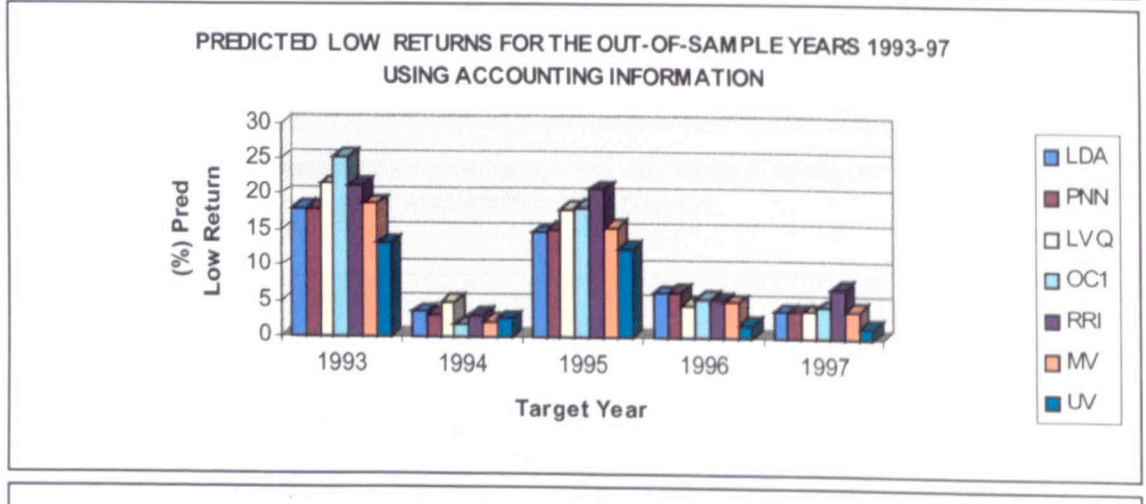
Figure 6.1: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

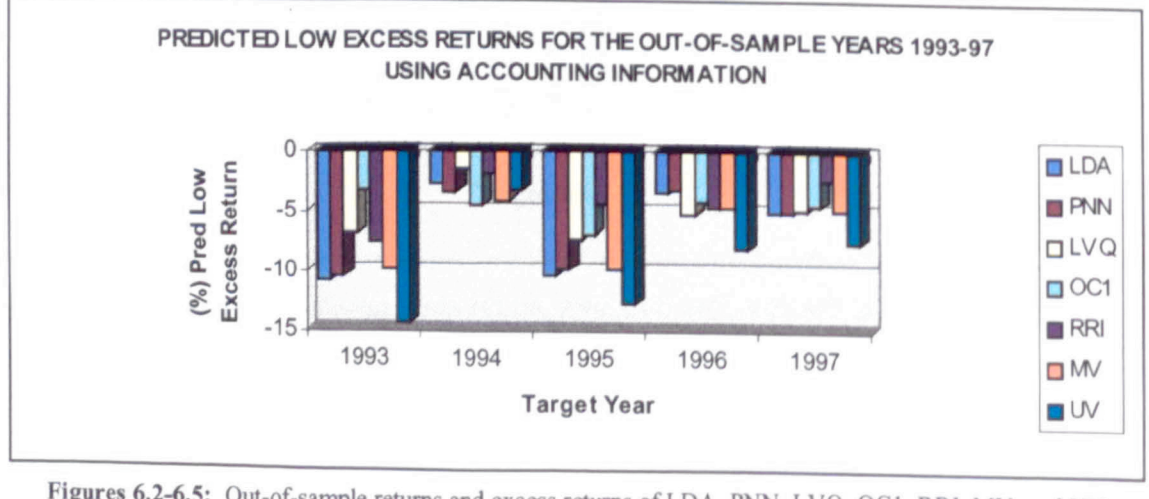
		LDA		PNN		LVQ		OC1		RRI		MV		UV	
1993		Predicted Returns & Excess Returns													
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 90.0 L= 9.2	H= 31.3	50.9	18.1	51.8	18.1	42.5	21.6	35.9	25.4	43.3	21.3	50.5	18.8	58.2	13.3
Actual Excess Ret		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 58.7 L= -19.5	L= 28.7	20.4	-10.8	20.8	-10.5	11.6	-6.9	5.5	-3.4	12.9	-7.5	19.5	-9.8	26.0	-14.5
1994															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 45.5 L= -7.6	H= 5.8	8.1	3.6	9.1	3.1	6.6	4.9	10.5	1.8	8.2	3.2	10.2	2.1	12.3	2.6
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 39.7 L= -13.2	L= 5.6	3.1	-2.7	4.4	-3.4	1.8	-1.4	5.5	-4.5	1.7	-1.8	5.2	-4.1	7.8	-3.2
1995															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 79.8 L= 7.1	H= 24.9	37.8	14.9	38.7	15.3	36.0	18.1	35.4	18.4	29.9	21.0	38.4	15.6	43.7	12.6
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 54.9 L= -18.3	L= 25.4	12.6	-10.5	13.3	-9.9	10.9	-7.3	10.2	-7.0	4.7	-4.3	13.3	-9.9	17.7	-12.8
1996															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 54.5 L= -5.3	H= 9.7	13.2	6.3	13.2	6.5	16.4	4.6	15.3	5.6	14.2	5.4	15.3	5.0	21.0	1.8
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 44.8 L= -15.1	L= 9.8	3.4	-3.4	3.5	-3.3	6.8	-5.2	5.6	-4.1	4.7	-4.7	5.6	-4.7	11.3	-8.1
1997															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 58.4 L= -7.2	H= 9.0	16.3	3.9	17.0	3.9	17.1	3.8	15.8	4.5	12.0	7.1	17.0	4.1	24.3	1.6
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 49.4 L= -16.5	L= 9.3	6.6	-5.0	7.3	-5.0	7.0	-4.8	6.3	-4.6	3.2	-2.4	7.3	-4.9	14.3	-7.5

Table 6.2: Out-of-sample high and low returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares









Figures 6.2-6.5: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

	LDA	PNN	LVQ	OC1	RRI	MV	UV
1993	226	219	244	251	240	219	**65**
1994	304	283	290	292	327	286	**97**
1995	314	296	278	281	332	295	**106**
1996	368	354	321	312	366	338	**125**
1997	319	302	303	312	316	297	**99**

Table 6.3: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

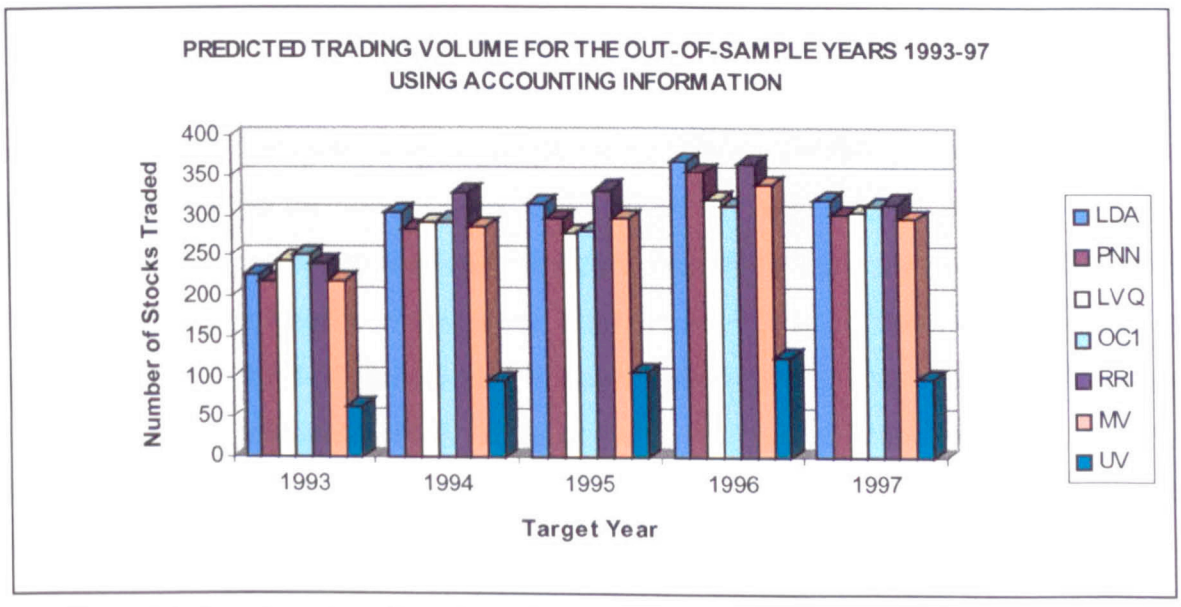


Figure 6.6: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

Portfolio	Return %	Size £m	DEBT/EQ	PAT/SR	P/E	DY	ME/BE
Actual H	79.8	266.5	69.8	-10.9	14.1	3.1	167.7
Pred H	38.7	205.9	78.4	-14.4	9.8	3.0	197.7
Actual L	7.1	586.4	53.5	3.0	16.2	3.9	154.1
Pred L	15.3	731.0	42.0	10.0	20.1	4.1	127.5

Table 6.4: Attributes of actual and predicted high and low portfolios from the PNN for the out-of-sample year 1995 using accounting information to predict high and low performing shares

Method	Return	Alpha	Beta
Actual	12.2	0	1
Actual H	59.1	43.2	1.39
LDA	18.3	2.9	1.38
PNN	19.0	3.5	1.38
LVQ	18.6	5.7	1.11
OC1	18.9	8.1	0.90
RRI	15.8	3.6	1.10
MV	19.7	5.1	1.30
UV	24.8	9.3	1.39

Table 6.5: Returns, alphas, and betas of predicted high portfolios of LDA, PNN, LVQ, OC1, RRI, MV, and UV for the out-of-sample year 1995 using accounting information to predict high and low performing shares

CHAPTER 7: PREDICTING HIGH PERFORMING SHARES USING ACCOUNTING AND NON-ACCOUNTING INFORMATION

In this Chapter, we extend our methodology to predict high performing shares by combining five statistical classifiers namely LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes and using accounting information, economic information, past share and index returns information as well as information about the industrial classification of around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. Our findings suggest that the UV principle, whereby a share is not assigned to the high performing portfolio unless all classifiers agree, produces significant improvements in classification and profitability if compared to the individual classification methods and reduces substantially the trading volume. Using accounting information alone results in a very accurate and profitable trading system. But there are additional benefits for individual classifiers and voting methodologies from adding non-accounting information.

This Chapter is organised as follows. In Section 7.1, we discuss the data and trading rules that we used in our study. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our input variables are 38 accounting ratios drawn from published accounting statements, 17 economic variables, 6-month, 1-year, 2-years, and 3-years past total share and index returns, and the industrial classification of the companies included in our sample.

In Section 7.2, we discuss the methodology that we used to implement the classification methods.

To make the models more robust and reduce the possibility of overfitting, we applied data preprocessing techniques such as data winsorisation, data normalisation, and data transformation. We performed two different experiments: In the first experiment, we compared and contrasted the five classifiers namely, LDA, PNN, LVQ, OC1 and RRI and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume using three different subsets of information: first, using accounting information only; second, using economic information, past share and index returns information, and industrial classification information only; and third, using all the available information by

mixing accounting and non-accounting information. In this experiment, we applied both the MV and UV methodologies for each individual implementation, respectively. In the second experiment, we applied the UV methodology over two parallel implementations of the classifiers using accounting and non-accounting information, respectively. According to this latter implementation, a share is not assigned to the high performing portfolio unless the five classifiers from the first implementation based on accounting information as well as the same classifiers from the second implementation based on non-accounting information agree unanimously on their decisions.

In Section 7.3, we report the results of experimentation. The results from the first experiment suggest that the UV scheme produces significant improvements in both classification and profitability over the individual classifiers and reduces substantially the trading volume. Using accounting information only results in a very accurate and profitable trading system. But additional benefits for individual classifiers and voting methodologies arise after adding non-accounting information. The results from the second experiment suggest that there are additional benefits after

(1) implementing the classifiers in parallel using accounting and non-accounting subsets of information, respectively, and

(2) then assigning a share to the high performing portfolio only if the five classifiers from the first implementation based on accounting information, as well as the same classifiers from the second implementation based on non-accounting information, agree unanimously on their decisions.

This implementation not only results in greater gains in profitability but also results in substantial reductions in the trading volume.

In Section 7.4, we discuss the practical implications of our study and we also discuss the possibilities for further improvements in our methodology.

7.1 DATA AND TRADING RULES

In this application, we are particularly interested in whether a particular share will be classified as H or L excess return share based on accounting information, economic information, past share and index returns information, as well as information about the industrial classification of the companies. To perform our experiments, we used the same target data that we used in the previous experiments. We recall that this data are total returns on all shares traded on the London Stock Exchange in the years 1993-97 starting with 626 shares in 1993 and rising up to 718 shares in 1997. To predict whether a particular share will be H or L in a given year, we used as input variables the 38 accounting indicators that we used in the previous experiments and we also collected 17-economic indicators, 6-month, 1-year, 2-years, and 3-years past total share and index returns information, as well as information about the industrial classification of the companies. The incorporation of previous years total share and index returns resulted in a small reduction in the number of companies we used in our previous experiments because a few shares were traded more recently on the London Stock Exchange and therefore we were not able to calculate the previous years total share and index returns for these shares. Therefore, the whole pattern was excluded from the data if either a share or index return indicator was missing. The company data were collected from the EXTEL service, whereas the share prices, the economic data, and the industrial classification information were collected from the DATASTREAM service.

A detailed list of the variables that we selected to implement the classification methods is given in Table 7.1. The idea of collecting the variables listed in Table 7.1 in the month of publication of the company's annual report was to develop a trading system that will be able to incorporate the impact of the public announcement of the accounting information to the share return and examine the interaction of this information with economic information, past share and index returns information, as well as information about the industrial classification of the companies.

We included information about the industrial classification of the companies in our sample because we expect some relation between share returns and the nature of the activities of the company. For example, one of the most important trends in the U.K. economy is the continuing shift of activity to services. The service sector accounted for around 60% of the U.K. output in 1998 compared to 50% in 1950. On the other hand, the manufacturing sector accounted for less than 25% of the U.K. output in 1998 compared to 33% in 1950. Another important trend

concerns the profitability of U.K. industrial sectors. The net rate of return on capital (NROC) for service companies was 14.0% in 1998 compared to 14.5% in 1990. The NROC for manufacturing companies was 11.0% in 1998 compared to 7.0% in 1990 (Warton, 1999). Including an indication of the industrial classification of the companies, we might be able to identify if these and other relevant industrial trends affect the share returns that we included in our sample. The companies that have included in our sample come from six separate industrial sectors namely services (S), manufacturing (M), property (P), utility (U), extractive (E), and financial (F).

Based on the variables presented in Table 7.1, we aimed to find rules that classify a particular share as H or L performing share using the two previous years of data to predict the next year. For example, to predict relative excess returns for 1993, we first trained the classification methods on the two preceding years 1991, 1992 and we tested them on the data available on 1993. We then moved the implementation one-year ahead and we used information available from 1992 and 1993 to predict 1994 and so on. We use only two previous years of data to predict relative excess returns for the next year because we believe that only recent accounting and non-accounting historical information may be relevant to predict relative excess returns in the next year.

7.2 METHODOLOGY

Let us assume that y_{it} is the 1-year-ahead excess return on some share i bought at time t, and x_{it} is a vector of variables that represent information for company i known at time t. The idea is to apply the five classification methods namely LDA, PNN, LVQ, OC1, and RRI, to assign y_{it} to one of the two classes C_{it} = H or L and then combine their predictions of these classifiers using MV and UV schemes. We recall that the ranking of the shares as H and L performing shares in a given year was decided on whether or not their excess returns is above or below the 25% threshold percentile that has been decided after ranking the returns in excess of an equally-weighted index. The models input is the vector x_{it} of variables that represent current month accounting information, economic information, past share and index returns information, as well as information about the industrial classification of the companies that we included in our sample.

In Chapter 4, we provided evidence that supports the view that financial ratio distributions are highly non-normal. To make our classification methods more robust and smooth the effect of outliers, we applied various data preprocessing techniques such as data winsorisation, data

normalisation, and triangular transformations of the variables. We first winsorised the input variables by ranking all observations on each individual variable in a given year, and setting values in the lower and upper 5% tails equal to the 5^{th} and 95^{th} percentile values. Then, we normalised the variables into the range $(0,1)$ by applying Eq. (4.1) and we transformed the normalised variables using Eq. (4.2). These data preprocessing techniques were applied to the predictor variables only but not in the target data. It is obvious that the same transformations were applied in the training and test sets.

We performed two experiments. In the first experiment, we compared and contrasted the five classifiers namely, LDA, PNN, LVQ, OC1 and RRI and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume using three different subsets of information: first, using accounting information only (AI); second, using economic information, past share and index returns information, and industrial classification information only (ERIIC); and third, using all the available information by mixing accounting and non-accounting information (ALL). In this experiment, we applied both the MV and UV methodologies for each individual implementation, respectively. In the second experiment, we applied the UV methodology over two parallel implementations of the classifiers using accounting and non-accounting information, respectively. A graphical illustration of this idea is presented in Figure 7.1. As we can see in Figure 7.1, we first implement the classifiers using accounting information only and then we implement them using non-accounting information only. For each separate implementation, a unanimous vote is taken over the five classifiers (UV-AI and UV-ERIIC, respectively) and then a vote is taken over the two votes that correspond to two separate implementations (UV-2V). According to this architecture, a share is not assigned to the high performing portfolio unless the five classifiers from the first implementation based on accounting information as well as the same five classifiers from the second implementation based on non-accounting information agree unanimously on their decisions.

One of the desirable end products of discriminant analysis is identification of good predictor variables. To identify an optimal number of predictor variables for each individual classification method we applied the same stepwise variable elimination procedures that we described in Section 5.2. However, we adapted this methodology given the different types of information we used to implement our classification methods. For example, if the implementation was based on either accounting or non-accounting information only, we selected the best subset of variables as follows: initially, we implemented the classification methods to predict the out-of-sample year 1993 using either all accounting variables or all non-accounting variables at the same time

and we recorded the misclassification rate. In the next step, we removed a single variable and we implemented the classification methods again. If the misclassification rate was lower, we removed the variable permanently; otherwise, we returned this variable back to the pool of variables to be included in the final model. Then, we repeated this procedure for each individual variable. We used the selected best subset of variables to predict the next target years 1994-97 while rolling the model one-year ahead and using two previous years of data to predict the next out-of-sample year.

However, if the implementation of the classification methods was based on mixing accounting and non-accounting variables at the same time, we thought that the above selection procedure might not be efficient because the large number of variables from both subsets of information might affect the selection of variables due to the possibility of overfitting. Therefore, we adapted this methodology as follows: after selecting the best accounting variables using a complete variable elimination procedure, we added to them all the non-accounting variables and we implemented the classification methods to predict 1993. We then removed one variable from the non-accounting variables and we implemented the classification methods again. If the misclassification rate was lower the non-accounting variable was removed permanently as before. Otherwise, we returned this variable back to the pool of the variables to be included in the final model. The same procedure was repeated for each individual non-accounting variable. As before, we used the selected best subset of variables to predict the next out-of-sample years 1994-97 while rolling the model one year ahead and using two previous years of data to predict the next out-of-sample year. We implemented the above variable selection procedures for all five classification methods to make the comparisons of the models equal. We have to mention, however, that the opposite experiment of selecting the best subset of non-accounting variables first, and then adding to them the accounting variables and repeating the variable selection procedure was also attempted but the models were not proven robust. This result should not be considered surprising, however, if we take into account that the variable elimination procedure applied in the tests in order to select the best subset of variables for our non-linear classifiers is not necessarily optimal. As we mentioned in Chapter 5, there are only a few techniques in the production of PNN, LVQ, RRI and OC1 classifiers which minimise overfitting of variables in an attempt to improve out-of-sample classification and robustness of the classification rule over time. Although these procedures might be efficient in minimising the possibilityof overfitting, they are not based on statistical criteria which eliminate overfitting of variables. On the other hand, for non-linear classification models, a certain combination of variables may be found significant for a given setting of model parameters, whereas the same combination of variables may be found insignificant for a different setting of parameters. Testing all possible combinations of parameters for each individual variable might be an expensive task in terms of

computational resources if the data set is large. We also have to consider that for recursive partitioning algorithms such as decision trees, the ordering of the variables might be particularly important in building the most accurate tree. A minor change in the ordering of the variables may lead to a substantial change in the tree topology. To deal with these limitations of ad hoc variable elimination techniques for non-linear classification methods, we investigate alternative techniques to dimensionality reduction. These techniques are discussed in more detail in Chapter 9.

Table 7.2 summarises the variables that we finally selected for the implementation of the algorithms under three different implementations: first, using accounting information only; second, using economic, past share and index returns, and industrial classification information only; and third, using all the available information at the same time.

To implement the five classifiers namely, LDA, PNN, LVQ, RRI, and OC1 and the two voting methodologies namely MV and UV, we followed implementation strategies exactly similar to those that we described in the previous Chapter. We compared and contrasted the five classifiers and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume. To evaluate the profitability of the five classifiers and the two voting methodologies, we calculated average returns and excess returns over the index for the portfolios of actual H and L shares in our data in all the 12-month holding periods starting each year, and then we compared them with the respective averages for the portfolios of H and L shares predicted by the classification methods. To examine if transaction costs can have an important impact in our trading system, we also compared the classification methods and the two voting methodologies for the predicted number of shares included in the portfolios of H performing shares that are traded in the target years 1993-97.

7.3 RESULTS

In this section, we summarise the results of experimentation. We performed two experiments: in the first experiment, we compared the five algorithms namely, LDA, PNN, LVQ, OC1 and RRI, and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume for the test out-of-sample year 1993 - on which the dimensionality reduction was conducted - and for the genuine out-of-sample years 1994-97. We implemented the five classifiers and the two voting methodologies using three different sets of information: first, using accounting information only (AI); second, using economic information, past share and index returns information, and information about the industrial classification of

the companies only (ERIIC); and third, using all the available information (ALL). In the second experiment, we first implemented the classifiers using accounting information only and then we implemented them using non-accounting information only. For each separate implementation, a unanimous vote was taken over the five classifiers (UV-AI and UV-ERIIC, respectively) and then a vote was taken over the two separate votes (UV-2V). Therefore, a share was not assigned to the to the high performing portfolio unless the five classifiers from the first implementation based on accounting information as well as the same five classifiers from the second implementation based on non-accounting information agreed unanimously on their decisions.

Experiment 1

Table 7.3 shows the classification results after implementing the classifiers and the voting methodologies using accounting information only. These results are also presented in Figure 7.2. As we can see, the UV methodology outperforms significantly the MV and the individual classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97. The PNN and the LDA also produce very good results for the target year 1993 but is seems that their classification performance deteriorates for the next years compared to the other classifiers. As we can see, the MV voting is the second best classifier for the target year 1994 with a very balanced predictive accuracy for both H and L performing shares. The OC1 is the second best classifier for the target year 1995 even though its classification accuracy seems to favour more L performing shares against H performing shares compared to LDA, PNN, and RRI that favour more H performing shares against L performing shares. The PNN is again the second best classifier for the target year 1996 but it seems that the MV predicts more accurately than the PNN H performing shares. The pattern is slightly different for the target year 1997 where the LVQ and the MV have the second best classification performance and achieve a very good predictive accuracy for both H and L performing shares.

Table 7.4 shows the classification results after implementing the classifiers and the voting methodologies using economic, past share and index returns information, as well as information about the industrial classification of the companies. These results are also presented in Figure 7.3. As we can see, the UV outperforms clearly the other classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97. The RRI is the second best classifier for the target year 1993, whereas the PNN is the second best classifier for the target year 1994. The classification performance of the PNN is also good for the target year 1995 but the LVQ and the MV are very close to the PNN and produce very favourable results. Although the pattern remains the same for the target year 1996 with PNN, LVQ and MV sharing again the second best classification performance, we have to notice that the RRI classifier predicts more accurately H performing shares compared to PNN, LVQ and MV. The pattern will change

slightly for the target year 1997 where the PNN is the second best classifier in terms of overall classification accuracy, whereas LDA, OC1, RRI, and MV predict more accurately than the PNN H performing shares.

Table 7.5 shows the classification results after implementing the classifiers and voting methodologies using all the available information. Figure 7.4 gives the graphical presentation of the results. As we can see, UV once more outperforms significantly the other classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97. PNN is the second best classifier for the target year 1993, whereas MV is the second best classification rule for the target years 1994 and 1995. On the other hand, OC1 has the second best classification accuracy for the target year 1996 but MV predicts more accurately than OC1 H performing shares for this year. The same will not happen the next target year 1997 where MV is the second best classification rule in terms of overall classification accuracy but not in terms of predictive accuracy for H performing shares since PNN and LDA produce more favourable results for H performing shares this year.

Figures 7.5-7.11 demonstrate the classification performance of each individual classifier and voting methodology separately under three different types of information: first, using accounting information only (AI); second, using economic, past share and index returns information, and information about the industrial classification of the companies only (ERIIC); and third, using all available information (All). As we can see, the classification performance of LDA is better after using accounting information only for the target years 1993-96, whereas the classification performance of the model is better after using all the available information for the target year 1997.

The classification performance of the PNN follows a more inconsistent pattern than the LDA under the three different types of information. As we can see, the model classifies as well as or even better after using accounting information rather than any other type of information for the target years 1993 and 1996, whereas the model classifies better after using non-accounting information only for the target years 1994 and 1995. However, the classification performance of the model is significantly better after using all the available information for the target year 1997.

The LVQ classifier favours more the use of all available information for the target years 1993, 1995 and 1996 but it seems that any additional information does not help for the target years 1994 and 1997 since the model classifies better these years using non-accounting and accounting subsets of information, respectively. The OC1 classifier favours the use of

accounting and non-accounting subsets of information for the target years 1993 and 1994, respectively, but the model prefers the use of all available information for the latest target years 1995-97. The RRI follows a different pattern than the OC1 classifier. It classifiers better using all the available information for the target years 1993, 1994 and 1997 but it prefers non-accounting and accounting subsets of information for the target years 1995 and 1996, respectively.

The classification performance of MV follows a more consistent pattern than the other classifiers. As we can see, MV classifies better after using all available information rather than subsets of information only for the target years 1993-97. However, the pattern is not the same for UV that seems to be more sensitive under the different types of information. As we can see, UV classifies better after using all available information for the target years 1994, 1995 and 1997, whereas it is more accurate after using non-accounting and accounting information subsets for 1993 and 1996, respectively.

Overall, the classification results suggest that the UV outperforms significantly the other classification methods for the target years 1993-97. PNN and LDA favour the use of either accounting or non-accounting subsets of information for the target years 1993-96, whereas they favour more the use of all available information in the latest target year 1997. LVQ, OC1, RRI and MV follow a different pattern than LDA and PNN. Their classification performance is as well as or even better after using all available information rather than using either accounting or non-accounting subsets of information in most out-of-sample years. However, the improvements in classification accuracy using all available information seem to be more significant for LVQ and OC1 compared to RRI and MV. On the other hand, UV seems to follow a more inconsistent pattern and it is more difficult to extract a general conclusion about the relationship between type of information and classification performance for this particular methodology.

Although the classification performance is a very important factor to evaluate a particular classifier, it is not the primary concern for this particular application. The ultimate purpose of our trading system is profitability. We therefore compared the average returns and excess returns over the index of the portfolios of actual H and L shares in our data in all the 12-month holding periods starting each year, with the respective average returns and excess returns of the portfolios of H and L shares predicted by the classifiers and voting methodologies.

Table 7.6 shows the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the

financial returns and excess returns of the portfolios of H and L shares predicted by the classifiers and voting methodologies using accounting information only. These results are also presented in Figures 7.12-7.13 and 7.14-7.15 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, all the classifiers and voting methodologies produce positive returns and excess returns. However, the UV methodology outperforms significantly the other classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 and produces the highest financial results. We recall that the financial results for the UV methodology represent the returns of actual H performing shares that correctly classified as H based on a unanimous decision from the five classification methods as well as the number of L performing shares that incorrectly classified as H based on a unanimous decision from the five classification methods as well. As far as concerns the other classifiers, PNN and LDA produce very good results for the target year 1993. It seems that the financial returns deteriorate for the next target year 1994 even though all classifiers and voting methodologies produce positive results. As we can see, there only minor differences between the MV, which is the second most profitable classification rule this year, and the other classifiers that produce slightly lower financial returns. The next target year 1995 is a year of high profitability for all classifiers and voting methodologies. The UV produces the highest financial results this year, whereas there are only minor differences in the financial results between the other classifiers. This pattern will not change significantly for the next target years 1996 and 1997 that are years of positive financial returns as well. The PNN is the second most profitable classifier for the target year 1996, whereas LVQ is the second most profitable classifier for the target year 1997.

Table 7.7 shows the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the classifiers and voting methodologies using economic, past share and index returns information as well as information about the industrial classification of the companies. These results are also presented in Figures 7.16-7.17 and 7.18-7.19 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, the UV methodology outperforms significantly the MV and the individual classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 and produces impressive financial results. The RRI is the second most profitable classifier for the target year 1993 but the profitability of MV and the other classifiers is also high as well. The next target year 1994 is a year of very high profitability for UV and its predicted high return (20.3%) and high excess return (14.8%) are almost twice the respective returns of the PNN which is the second most profitable classifier. The other classifiers produce lower but still positive results for the target year 1994. The

profitability increases significantly for the target year 1995 and deteriorates slightly for the latest target years 1996 and 1997. However, the UV still produces very impressive financial returns for these years whereas there are only minor differences in the financial returns predicted by the other classifiers.

Table 7.8 shows the financial returns and excess returns over the index for the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the classifiers and voting methodologies using all the available information. These results are also presented in Figures 7.20-7.21 and 7.22-7.23 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, the UV methodology once more outperforms significantly the other classifiers for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 and produces impressive financial results. MV and PNN are the second most profitable classification rules for the target year 1993, whereas the other classifiers also produce impressive financial results. The next target year 1994 is a year of very high profitability for UV, whereas the other classifiers produce lower but positive financial returns. As we can see, UV outperforms significantly the other classifiers and its financial return (30.9%) is almost two times the return of the LVQ (12.1%) which is the second most profitable classifier for the target year 1994. However, even more impressive is the excess return produced by UV (26.1%) that is four times higher than the excess return produced by MV and LVQ (6.2%) which are the second most profitable classifiers in excess returns. The next target year 1995 is a year of high profitability for all classifiers and voting methodologies. The UV is still the most profitable methodology for this year, whereas there are only minor differences in financial returns among the other classifiers. The profitability deteriorates for the target 1996 but the financial returns are still positive for all classifiers and voting methodologies. The UV rule clearly outperforms the other classifiers whereas there are only minor differences in the financial returns among the other classifiers. The UV rule also produces very impressive results for the last target year 1997 with impressive high return (41.4 %) and excess return (29.3%) that are almost twice the respective returns of the MV and the PNN that are the second most profitable classifiers.

Figures 7.24-7.51 demonstrate the financial returns of each individual classifier and voting methodology separately under three different types of information: first, using accounting information only (AI); second, using economic, past share and index returns information as well as information about the industrial classification of the companies (ERIIC); and third, using all available information (All). As we can see in Figures 7.24-7.27, the financial returns of LDA are better after using accounting only for the target years 1993-96, whereas the returns

produced by the model are better after using all the available information for the target year 1997.

The financial returns of the PNN follow a different pattern. As we can see in Figures 7.28-7.31, the model produces greater returns using accounting information for the target year 1993, whereas it seems that the model favours the use of non-accounting information for the target year 1994. On the other hand, there are only minor differences in the financial returns produced by the model under different types of information for the target years 1995 and 1996, whereas it seems that the model favours the use of all available information for the latest target year 1997.

Figures 7.32-7.35 illustrate the financial results of LVQ. The results suggest that the LVQ classifier favours the use of accounting information for the target years 1993, 1996 and 1997, whereas it favours the use of all available information for the target years 1994 and 1995. On other hand, Figures 7.36-7.39 suggest that the OC1 is not particularly sensitive under the different types of information even though it seems that this classifier favours slightly the use of all available information for the target years 1993, 1995 and 1997, and the use of non-accounting information for the target years 1994 and 1996. The RRI follows an even more consistent pattern than OC1 with only minor differences in the financial results produced by the model under the different types of information for the target years 1993-97 as we can see in Figures 7.40-7.43.

The MV seems to prefer the use of all available information. As we can see in Figures 7.44-7.47, the MV rule produces greater financial returns using all available information for the target years 1993, 1995 and 1997, whereas there are minor differences in the financial returns produced by this rule for the target years 1994 and 1996 under the three different types of information. On the other hand, the UV seems to follow a different pattern than MV. As we can see in Figures 7.48-7.51, the UV produces greater financial returns after using all available information only for the target years 1994 and 1997, whereas there are only minor differences in the financial returns produced by this rule for the other years under different types of information.

Overall, the predicted financial returns suggest that UV outperforms significantly the other classification methods for the target years 1993-97. PNN and LDA seem to prefer the use of either accounting or non-accounting subsets of information for the target years 1993-96, whereas they favour more the use of all available information for the target year 1997. If we compare these results with the classification performance of the algorithms we might notice some degree of correlation between classification accuracy and profitability. LVQ, OC1, RRI

follow a different pattern than LDA and PNN and their financial returns are as well as or even better after using all available information rather than using either accounting or non-accounting subsets of information for specific out-of-sample years. However, the improvements in financial returns using all available information are less obvious for LVQ and RRI compared to OC1. On the other hand, MV and UV seem to favour the use of all available information rather than the use of specific subsets of information for most out-of-sample years.

Although the financial returns are a primary factor to evaluate a particular trading system, we should also examine the transaction costs involved in trading the number of shares predicted by the classification methods. For this purpose, we calculated the predicted number of shares included in the portfolios of H performing shares that are traded for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97. It is obvious that this is the number of actual H performing shares that correctly predicted as H by the classification methods as well as the number of actual L performing shares that incorrectly predicted as H.

Table 7.9 compares the classification methods for the number of shares predicted to be H in all the 12-month holding periods for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 using accounting information only. Figure 7.52 gives a graphical illustration of the results. As we can see in Figure 7.52, the UV rule produces the smallest trading volume for the target years 1993-97 and outperforms significantly the other classification methods. PNN and LDA predict the second smallest trading volume for the target year 1993, whereas all classification methods seem to be expensive enough if compared to UV for the target years 1994-97.

Table 7.10 compares the classification methods for the number of shares predicted to be H in all the 12-month holding periods for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 using economic, past share and index returns information as well as information about the industrial classification of the companies. Figure 7.53 gives a graphical illustration of the results. As we can see in Figure 7.53, the UV produces again the smallest trading volume for the target years 1993-97 and outperforms significantly the other classification methods. The PNN predicts more H performing shares for the target year 1993, whereas all classification methods seem to be really expensive if compared to UV for the target years 1994-97.

Table 7.11 compares the classification methods for the number of shares predicted to be H in all the 12-month holding periods for the test out-of-sample year 1993 as well as for the genuine out-of-sample years 1994-97 using all the available information. Figure 7.54 gives a graphical

illustration of the results. As we can see in Figure 7.54, the UV produces once more the smallest trading volume for the target years 1993-97 and outperforms significantly the other classification methods. PNN and MV predict the second smallest trading volume for the target year 1993, whereas once again all classifiers seem to be really expensive if compared to UV for the target years 1994-97.

A very careful examination of the results presented above might reveal some degree of correlation in the predictions of the classifiers in terms of classification accuracy and profitability. For example, as we can see in Figures 7.2 and 7.12-7.13, some classifiers such as LDA, PNN, and UV which have the best classification performance for the test out-of-sample year 1993 are the most profitable classifiers during the same year. To investigate this observation further, we compared the predictions of the PNN in terms of classification accuracy (CA), predicted H return (HR), and predicted H excess return (HER) for the out-of-sample period 1993-97. The results of this comparison are illustrated in Figures 7.55-7.57 for three different types of information, respectively: first, using accounting information; second, using economic, past share and index returns information as well as information about the industrial classification of the companies; and third, using all available information. As we can in Figures 7.55-7.57, although it seems that there is a slight parallel movement between classification accuracy and predicted H excess return for the PNN during the 1993-97 period, this relation is not particularly clear. Furthermore, a close examination of the results presented in Figures 7.2-7.4 and 7.12-7.13, 7.16-7.17 and 7.20-7.21 seems to confirm the view that when classification accuracy of the PNN as well as the other classifiers rises more sharply, profitability also rises. However, the fairly robust classification performance of the classifiers during the 1993-97 period does not allow us to extract some definite conclusions on the relation between classification accuracy and profitability of the classifiers. We expect that as we improve the classification performance of the classifiers in the future, we will be able to observe more clearly if further improvements in the classification accuracy of the classifiers will also improve the profitability of the classifiers as well.

Experiment 2

The results from the previous experiment suggested that UV outperforms significantly the other classifiers in terms of classification accuracy, profitability, and trading volume. In the second experiment, we first implemented the classifiers using accounting information only and then we implemented them using non-accounting information only. For each separate implementation, a unanimous vote was taken over the five classifiers (UV-AI and UV-ERIIC, respectively) and then a vote was taken over the two separate votes (UV-2V). According to this implementation, a share was not assigned to the high performing portfolio unless the five classifiers from the first

implementation based on accounting information as well as the same five classifiers from the second implementation based on non-accounting information agreed unanimously on their decisions.

Table 7.12 shows the classification performance of the UV methodology under four different implementations: first, using accounting information to implement the five classifiers and then taking a UV over their predictions (UV-AI); second, using economic information, past share and index returns information, and information about the industrial classification of the companies to implement the five classifiers and then taking a UV over their predictions (UV-ERIIC); third, using all the available information to implement the classifiers and then taking a UV over their predictions (UV-ALL); and fourth, using accounting information and non-accounting information to implement the five classifiers, separately, taking a UV of the predictions of the classifiers under each separate implementation, and then taking a UV of the two votes that were taken after each separate implementation (UV-2V). The classification results are also illustrated in Figure 7.58. As we can see in Figure 7.58, the UV-ERIIC methodology outperforms the other methodologies for the test out-of-sample year 1993, whereas the UV-2V methodology outperforms the other methodologies for the genuine out-of-sample years 1994 and 1995. UV-AI, UV-ALL, and UV-2V have similar classification performance for the genuine out-of-sample year 1996, whereas the UV-ALL methodology outperforms the other methodologies for the genuine out-of-sample year 1997.

Although the classification performance is a very important factor to evaluate a particular classifier, it is not the primary concern for this particular application. As we said before, the ultimate purpose of our trading system is profitability. Table 7.13 shows the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the four different implementations of the UV methodology. These results are also presented in Figures 7.59-7.62 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, the UV-2V methodology outperforms significantly the other methodologies for the target years 1993 and 1995, whereas it also produces the highest returns and excess returns for the target year 1994. UV-ALL and UV-2V produce very similar results for the target year 1997.

Although the financial returns are a primary factor to evaluate a particular trading system, we should also examine the transaction costs involved in trading the number of shares predicted by the classification methods. Table 7.14 compares the four UV methodologies for the number of shares predicted as H in all the 12-month holding periods for the test out-of-sample year 1993

as well as for the genuine out-of-sample years 1994-97. Figure 7.63 gives a graphical illustration of the results. As we can see in Figure 7.63, the UV produces the smallest trading volume for the target years 1993-97 and outperforms significantly the other classification methods.

As we have mentioned in Section 5.1, a share was included in our sample only if there was a complete set of annual company accounts available on the EXTEL service as well as the date of publication of the company's annual report. After very extensive investigation, we were able to find the date of publication of the companies' annual reports for a sample of around 700 companies per year starting with 626 companies in 1993 and rising up to 718 companies in 1997. Therefore, the sample of shares that we used for our study was substantially smaller than the sample of all shares traded on the London Stock Exchange in the years 1993-97. The use of an equally-weighted index to benchmark the performance of our trading system gave us the flexibility to examine the robustness of our trading system under different market conditions. For example, if we examine the performance of the index in Table 7.6, we can see that for the years 1993 and 1995, the index rises more sharply while for other years, for example 1994, the index is particularly low. As it is shown in Tables 7.6-7.8, although the performance of our trading system deteriorates in 1994 when the index is particularly low, our classification methods are still profitable. This result gives a good indication about the performance of our trading system under different market conditions, for example, in a bear market.

Another issue concerns the practical applicability of our trading strategy that is based on the idea of taking positions and holding them for one-year horizons. Certainly, the implementation of our strategy within an asset management company might not be an ideal strategy in the real world. On the other hand, the implementation of our strategy might be more feasible within a hedge fund operation, particularly if a number of further improvements were considered. For example, in the real world environment, many hedge funds practice long/short strategies, where not only the high performing shares are bought, but also the low performing shares are sold. Furthermore, many traders put stop-loss and limit orders in place, and these might be conditional on the degree of confidence in the model predictions. Although the time frame we were given to complete this thesis does not allow such an extensive investigation, it would be really worthy to investigate all these improvements in future research. However, we have to make very clear that our focus on this thesis is to develop and test the initial prototype of a general trading strategy that will enable the integration of other models in the future to facilitate the idea of more frequent trading. If our trading strategy can guarantee a given level of return over one year trading horizons, then incorporation of other models into our trading platform for either hedging on a day-to-day basis or enabling more frequent trading, could further improve

our financial returns. Definitely, the description of a fully functional commercial trading system goes well beyond the ambitions of the thesis.

7.4 SUMMARY AND CONCLUSIONS

In this Chapter, we applied five statistical classification methods from different model families to identify H and L performing shares for the target years 1993-97 and we examined the possibility of combining their forecasts using MV and UV principles. The model inputs we used were accounting information, economic information, past share and index returns information as well as information about the industrial classification of around 700 companies. We performed two experiments: In the first experiment, we compared the five algorithms namely, LDA, PNN, LVQ, OC1 and RRI, and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume for the test out-of-sample year 1993 - on which the dimensionality reduction was conducted - and for the genuine out-of-sample years 1994-97. We implemented the five classifiers and the two voting methodologies using three different sets of information: first, using accounting information only (AI); second, using economic information, past share and index returns information, and information about the industrial classification of the companies only (ERIIC); and third, using all the available information (ALL). We found that all classification methods produce consistent excess returns. However, greater gains resulted from UV were a share is not classified as H performing share unless all classifiers agree. The UV principle not only produces significant greater returns than the other methods, but it also results in substantial reductions in the number of shares traded. The results also suggest that there are substantial gains of using all available information rather than using subsets of information only especially for classifiers such as OC1, LVQ, and RRI. These results should not be considered surprising, however, if we take in account the way these classifiers form their hypothesis for different clusters of the sample data.

In the second experiment, we applied the UV methodology over two parallel implementations of the classifiers using accounting and non-accounting information, respectively. According to this implementation, a share is not assigned to the high performing portfolio unless the five classifiers from the first implementation based on accounting information as well as the same classifiers from the second implementation based on non-accounting information agree unanimously on their decisions. After performing this experiment, we found greater gains in profitability and substantial reductions in the trading volume.

Our work in this Chapter extends previous research that examined the predictability of share returns mostly under restricted forms of linear and non-linear models. Our results provide

substantial evidence for the ability of non-linear classification methods over the linear model to identify high performing shares. Non-linear models are more flexible to deal with the complex relationships that are evident in the financial data compared to the linear model that is able to handle more simple linear patterns. The main advantage of our methodology is model flexibility that is essential for the complex financial processes that are chaotic and inconsistent in the time-scale. Furthermore, our results confirm previous research that reports predictable patterns in stock returns.

We have to notice, however, that our classification methods should be treated with caution due to their large number of parameters. To avoid the possibility of overfitting, we applied sophisticated data pre-processing techniques in order to eliminate the effect of outliers and increase the robustness of the models.

A very interesting investigation might be to apply our methodology to different industrial sectors and examine the benefits of applying our trading system to homogeneous industrial sectors. The details of this implementation are presented in the next Chapter.

ACCOUNTING VARIABLES:	ECONOMIC VARIABLES:	PAST SHARE & INDEX RETURNS VARIABLES:	INDUSTRIAL CLASSIFICATION :
Return on Capital	U.K. Industrial Production (UIP)	6-Month Aggregate Return (6MAR)	Manufacture (M)
PBT/TA	U.K. Effect. Exchange Rate Index (UEERI)	6-Month All Share Index Return (6MASIR)	Financial (F)
PBT/TCE	U.K. Retail Price Index (URPI)	1-Year Aggregate Return (1YAR)	Service (S)
NI/TCE	U.K. Import Price Index (UIPI)	1-Year All Share Index Return (1YASIR)	Extractive (E)
CF/TA	U.K. Export Price index (UEPI)	2-Year Aggregate Return (2YAR)	Utility (U)
CF/TCE	U.K. Volume of Retail Sales (UVRS)	2-Year All Share Index Return (2YASIR)	Property (P)
	U.K. Average Earn. Index (UAEI)	3-Year Aggregate Return (3YAR)	
Profitability	U.K. Unemployment Rate (UUR)	3-Year All Share Index Return (3YASIR)	
PBT/SR			
PAT/SR	U.K. 10 Year Gov. Bond Yield (U10YGBY)		
NI/SR	U.K. Corporate Bond Yield (UCBY)		
CF/SR	U.K. Gross Red.. Yld on 20 Year Gilts (UGRY20YG)		
PAT/EQ			
CF/MKBD	U.K. 3-M Disc. Treas. Bill Rate (U3MDTBR)		
	U.K. 3-M Disc. Bank Bill Rate (U3MDBBR)		
Financial Leverage			
DEBT/EQ	Fuel Oil Prices (Pounds/Gal) (FOP)		
DEBT/TCE			
DEBT/TA	U.S. $ TO U.K. £ Exchange Rate (US/UKER)		
TL/EQ	German Mark to U.K. £ Exchange Rate (GM/UKER)		
TA/EQ	Japanese Yen to U.K. £ Exchange Rate (JY/UKER)		
BA/MKBD			
Investment			
P/E			
DY			
EY			
BE/ME			
Growth (%)			
TA			
PAT			
PBT			
EPS			
MKBD			
SR			
Short-Term Liquidity			
CA/CL			
CL/TA			
CL/EQ			
Return on Investment			
NI/TA			
PAT/TA			
Efficiency			
SR/TA			
DRS/SR			
Risk			
PBT/CL			
PAT/CL			
NI/CL			
CF/CL			

PBT: Profit Before Taxes; TA: Total Assets; TCE: Total Capital Employed; CF: Cash Flow; PAT: Profit after Taxes; SR: Sales Revenue; NI: Net Income; EQ: Shareholders' Equity; MKBD: Market Capitalisation at Balance Sheet Date; DEBT: Debt; TL: Total Liabilities; BA: Book Assets; P/E: Price/Earnings Ratio; EY: Earnings Yield; DY: Dividend Yield; BE: Book Equity; ME: Market Equity; EPS: Earnings Per Share; CA: Current Assets; CL: Current Liabilities; DRS: Debtors.

Table 7.1: Initial list of the accounting and the non-accounting variables that we collected to predict high and low performing shares

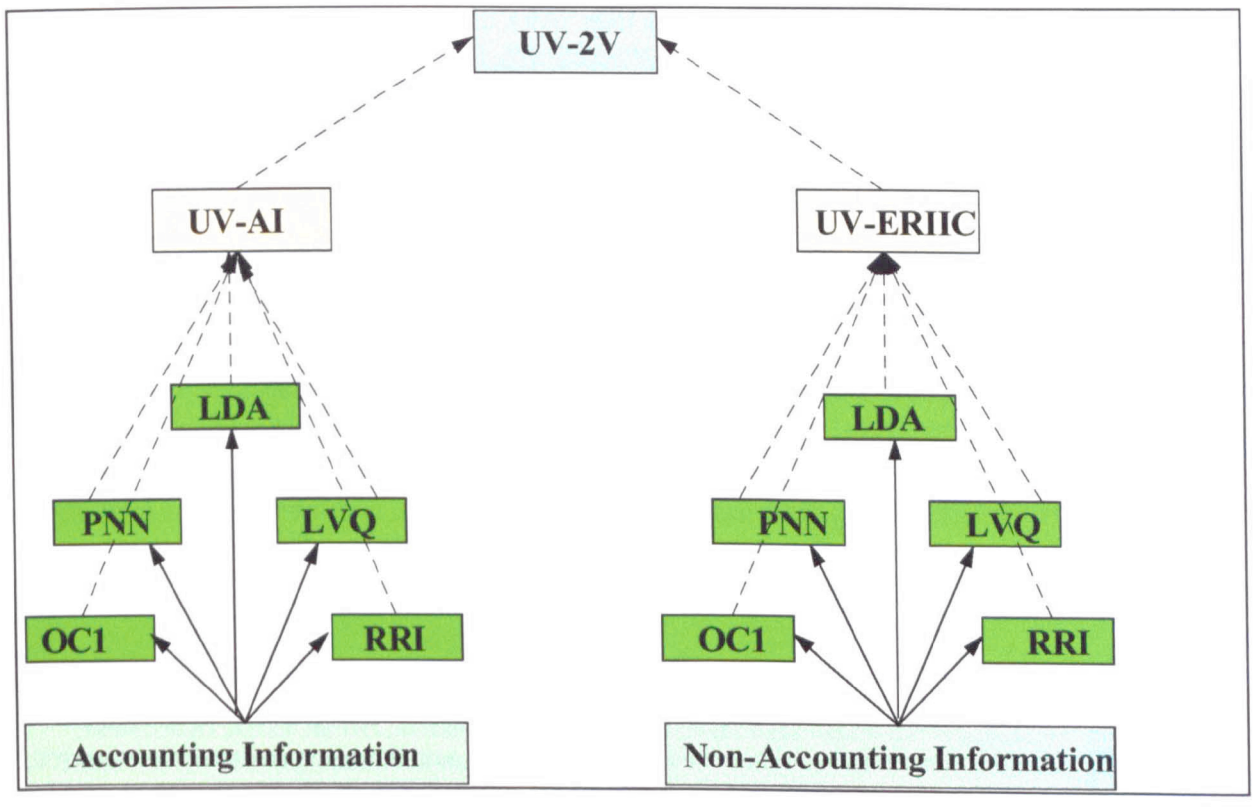


Figure 7.1: A two-level unanimous voting framework

LDA	**Accounting Variables Subset:** PBT/SR, SR/TA, DEBT/EQ, PAT/SR, EPS (%), SR (%), PAT/TA, PBT(%), NI/TCE, PBT/TCE, TL/EQ, CF/SR, DY, CA/CL, CL/TA, PBT/CL, PAT/CL, NI/CL, CF/CL, BE/ME, TA/MKBD **Non-Accounting Variables Subset:** UEERI, URPI, UVRS, U10YGBY, UCBY, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR **Accounting & Non-Accounting Subsets:** DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD, UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P
PNN	**Accounting Variables Subset:** DEBT/TA, PAT/EQ, SR/TA, DRS/SR, PAT/TA, TA (%), MKBD (%), CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PAT/CL, NI/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD **Non-Accounting Variables Subset:** URPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, US/UKER, GM/UKER, JY/UKER, 6MAR, 6MASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P **Accounting & Non-Accounting Subsets:** DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD, UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P
LVQ	**Accounting Variables Subset:** SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, EPS (%), SR (%), MKBD (%), PBT (%), NI/TCE, DEBT/TCE, CF/TA, CF/TCE, CA/CL, CL/EQ, CF/CL, BE/ME, TA/MKBD, CF/MKBD **Non-Accounting Variables Subset:** UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P **Accounting & Non-Accounting Subsets:** DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD, UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P
OC1	**Accounting Variables Subset:** DEBT/TA, PBT/SR, TA/EQ, NI/SR, DEBT/EQ, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), PAT (%), P/E, NI/TCE, CF/TA, CF/TCE, CF/SR, CL/TA, CL/EQ, NI/CL, TA/MKBD, CF/MKBD **Non-Accounting Variables Subset:** UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P **Accounting & Non-Accounting Subsets:** DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD, UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P
RRI	**Accounting Variables Subset:** DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD **Non-Accounting Variables Subset:** UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P **Accounting & Non-Accounting Subsets:** DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD, UIP, UEERI, URPI, UIPI, UEPI, UVRS, UAEI, U10YGBY, UCBY, U3MDTBR, U3MDBBR, UGRY20YG, FOP, US/UKER, GM/UKER, JY/UKER, UUR, 6MAR, 6MASIR, 1YAR, 1YASIR, 2YAR, 2YASIR, 3YAR, 3YASIR, M, F, S, E, P

Table 7.2: Subsets of accounting and non-accounting variables that we finally selected to predict high and low performing shares after applying stepwise variable elimination procedures

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
Actual Class	Patterns	Predicted Class Membership													
1993		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	157	95	62	95	62	95	62	95	62	98	59	99	58	36	121
L	469	116	353	121	348	180	289	180	289	173	296	139	330	25	444
Overall (%)		71.57 %		70.77 %		61.34 %		61.34 %		62.94 %		68.53 %		76.68 %	
1994		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	155	83	72	84	71	83	72	83	72	80	75	88	67	25	130
L	463	193	270	183	280	188	275	199	264	201	262	178	285	52	411
Overall (%)		57.12 %		58.90 %		57.93 %		56.15 %		55.34 %		60.36 %		70.55 %	
1995		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	160	90	70	90	70	82	78	88	72	92	68	91	69	31	129
L	479	183	296	185	294	177	302	168	311	225	254	177	302	65	414
Overall (%)		60.41 %		60.09%		60.09 %		62.44 %		54.15 %		61.50 %		69.64 %	
1996		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	171	94	77	93	78	98	73	93	78	88	83	99	72	34	137
L	510	195	315	189	321	212	298	219	291	188	322	197	313	37	473
Overall (%)		60.06 %		60.79 %		58.15 %		56.39 %		60.21 %		60.50 %		74.45 %	
1997		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	180	106	74	106	74	110	70	97	83	96	84	112	68	32	148
L	538	211	327	209	329	180	358	200	338	230	308	182	356	39	499
Overall (%)		60.31 %		60.58 %		65.18 %		60.58 %		56.27 %		65.18 %		73.96 %	

Table 7.3: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

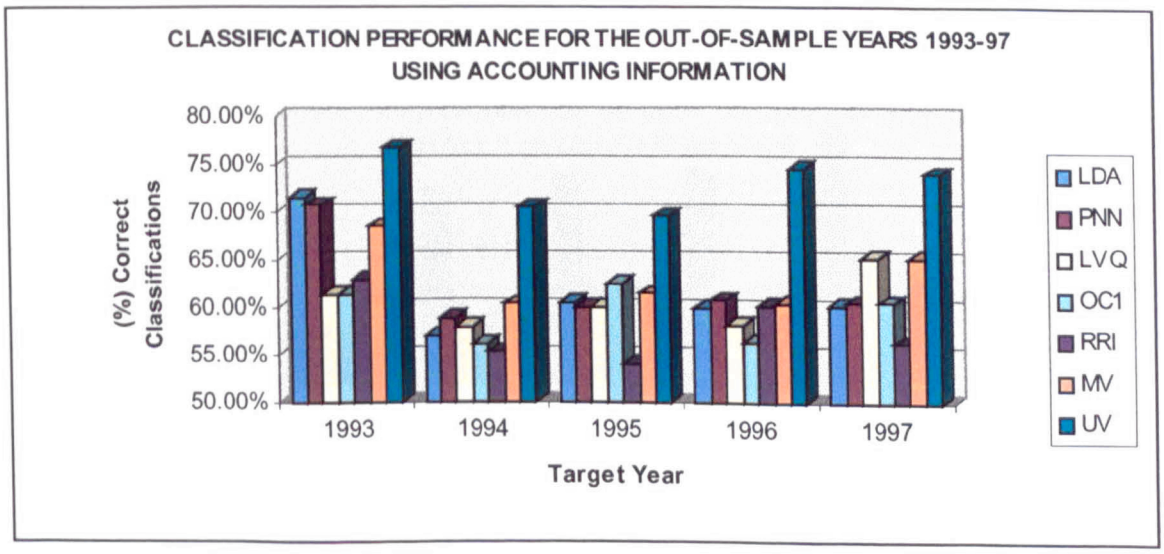


Figure 7.2: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
Actual Class	Patterns	Predicted Class Membership													
1993		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	157	94	63	90	67	81	76	81	76	93	64	85	72	37	120
L	469	199	270	162	307	170	299	173	296	153	316	148	321	15	454
Overall (%)		58.15 %		63.42 %		60.70 %		60.22 %		65.34 %		64.86 %		78.43 %	
1994		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	155	85	70	87	68	83	72	86	69	88	67	84	71	27	128
L	463	208	255	175	288	174	289	191	272	198	265	187	276	36	427
Overall (%)		55.02 %		60.68 %		60.19 %		57.93 %		57.12 %		58.25 %		73.46 %	
1995		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	160	89	71	100	60	98	62	84	76	93	67	102	58	30	130
L	479	188	291	188	291	191	288	179	300	213	266	191	288	44	435
Overall (%)		59.47 %		61.19 %		60.41 %		60.09 %		56.18 %		61.03 %		72.77 %	
1996		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	171	95	76	94	77	91	80	86	85	105	66	94	77	22	149
L	510	226	284	209	301	204	306	210	300	240	270	204	306	45	465
Overall (%)		55.65 %		58.00 %		58.30 %		56.68 %		55.07 %		58.74 %		71.51 %	
1997		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	180	105	75	99	81	96	84	104	76	106	74	108	72	39	141
L	538	247	291	201	337	212	326	248	290	251	287	239	299	74	464
Overall (%)		55.15 %		60.72 %		58.77 %		54.87 %		54.74 %		56.69 %		70.06 %	

Table 7.4: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares

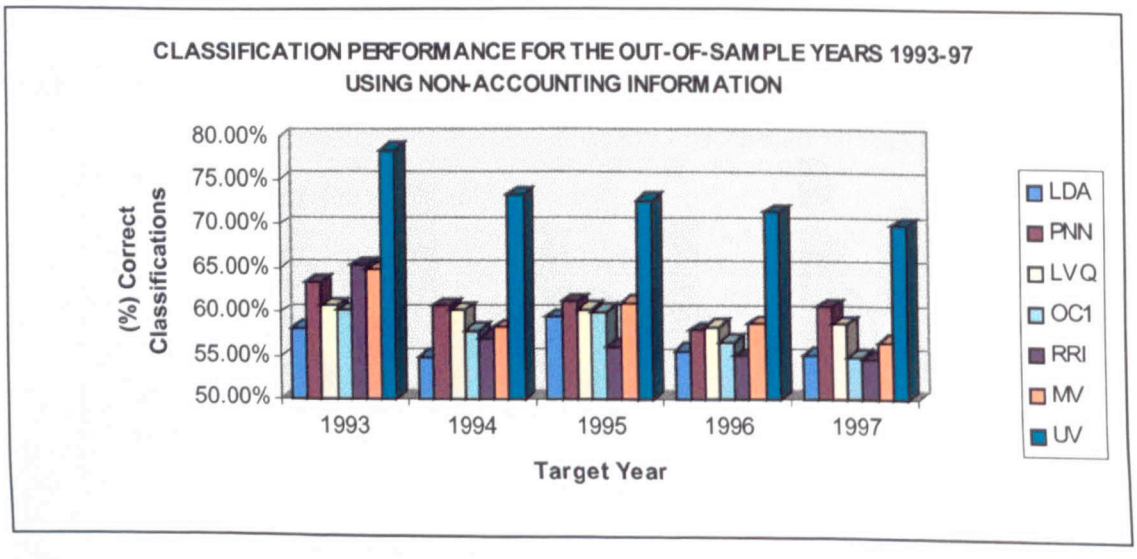


Figure 7.3: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
Actual Class	Patterns	Predicted Class Membership													
1993		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	157	95	62	93	64	83	74	91	66	94	63	89	68	44	113
L	469	162	307	119	350	157	312	179	290	152	317	125	344	28	441
Overall (%)		64.22 %		70.77 %		63.10 %		60.86 %		65.65 %		69.17 %		77.48 %	
1994		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	155	82	73	84	71	81	74	90	65	85	70	91	64	21	134
L	463	197	266	195	268	172	291	197	266	177	286	181	282	25	438
Overall (%)		56.31 %		56.96 %		60.19 %		57.61 %		60.03 %		60.36 %		74.27 %	
1995		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	160	90	70	91	69	96	64	89	71	86	74	100	60	16	144
L	479	201	278	180	299	166	313	152	327	206	273	156	323	26	453
Overall (%)		57.59 %		61.03 %		64.01 %		65.10 %		56.18 %		66.20 %		73.40 %	
1996		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	171	93	78	94	77	89	82	86	85	92	79	90	81	31	140
L	510	213	297	214	296	190	320	175	335	218	292	182	328	35	475
Overall (%)		57.27 %		57.27 %		60.06 %		61.82 %		56.39 %		61.38 %		74.30 %	
1997		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	180	105	75	107	73	97	83	100	80	102	78	99	81	44	136
L	538	190	348	183	355	195	343	197	341	229	309	166	372	32	506
Overall (%)		63.09 %		64.35 %		61.28 %		61.42 %		57.24 %		65.60 %		76.60 %	

Table 7.5: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares

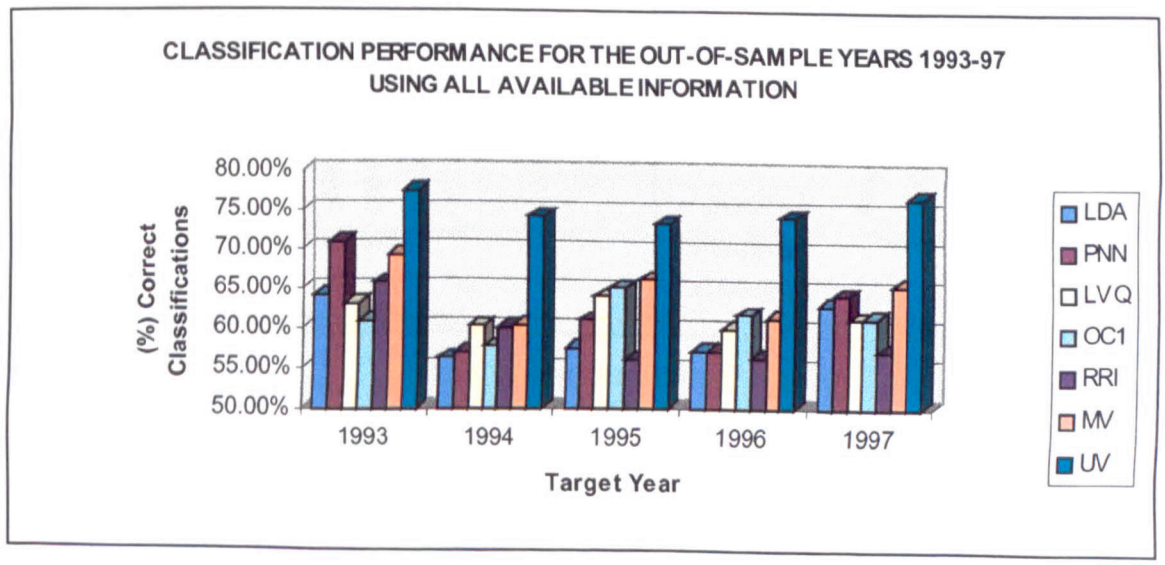
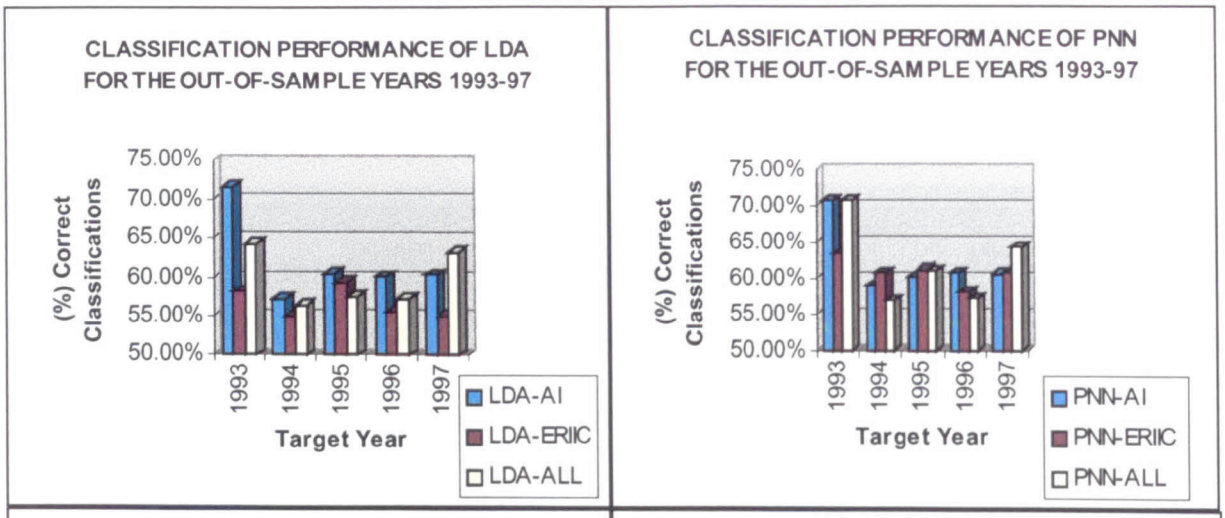
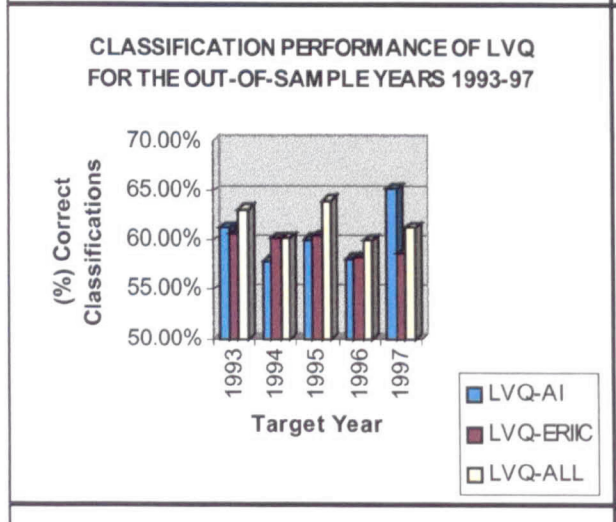
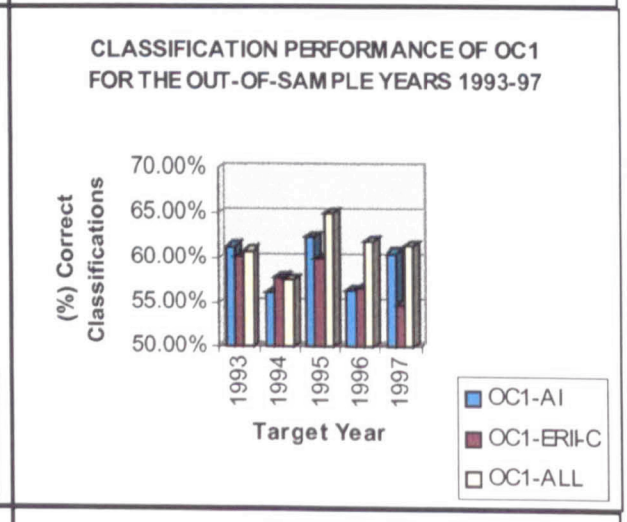
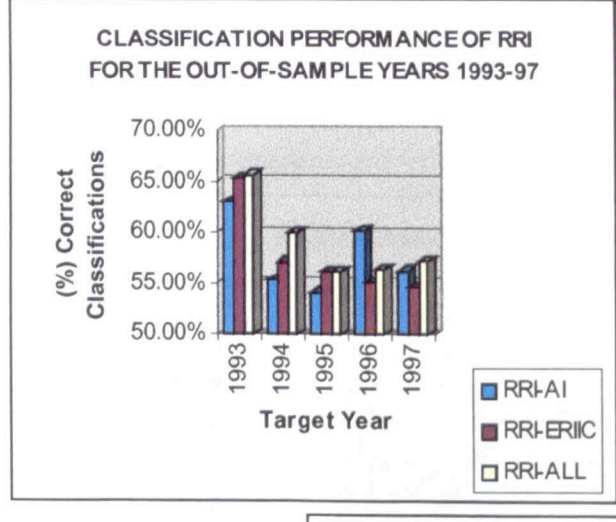


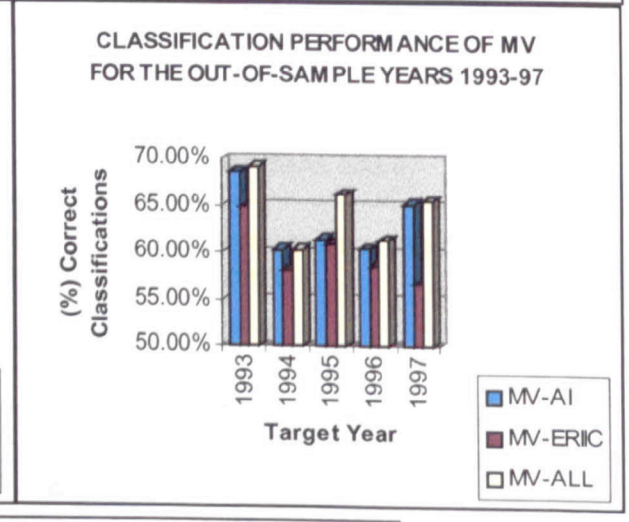
Figure 7.4: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares

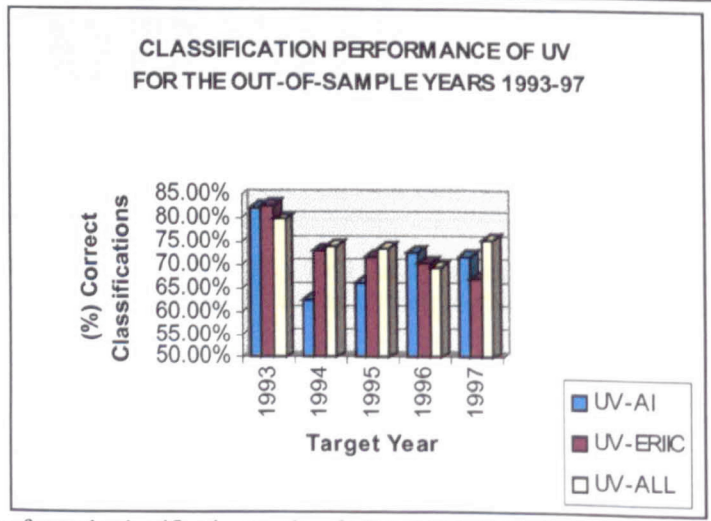








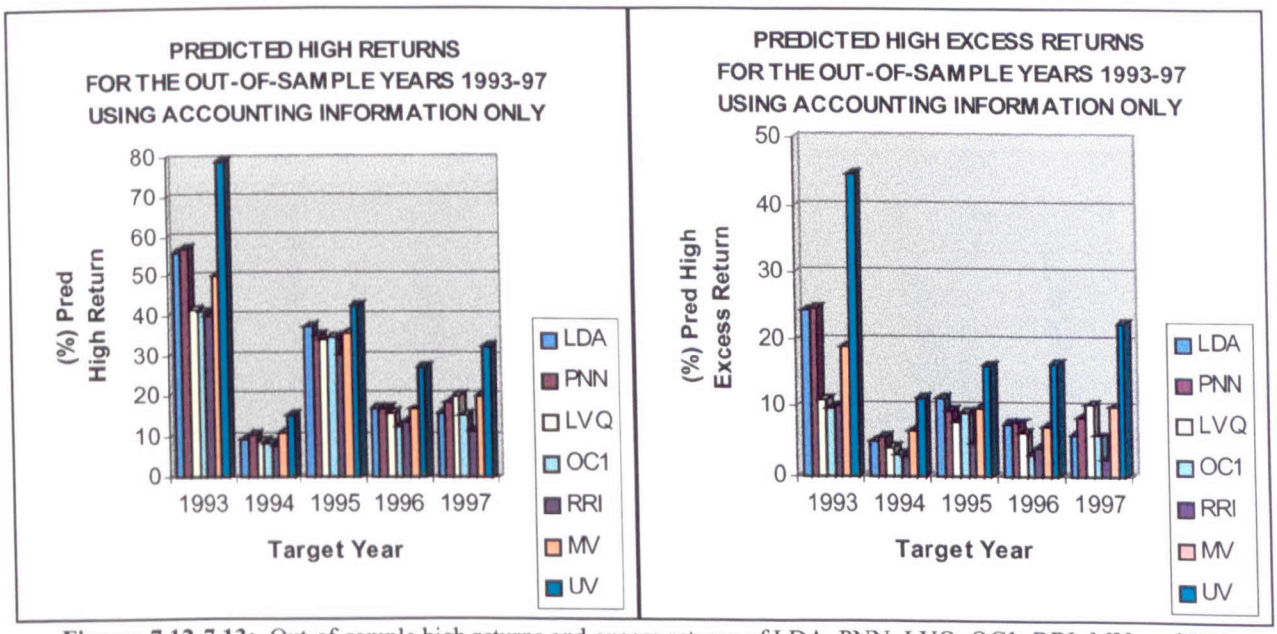




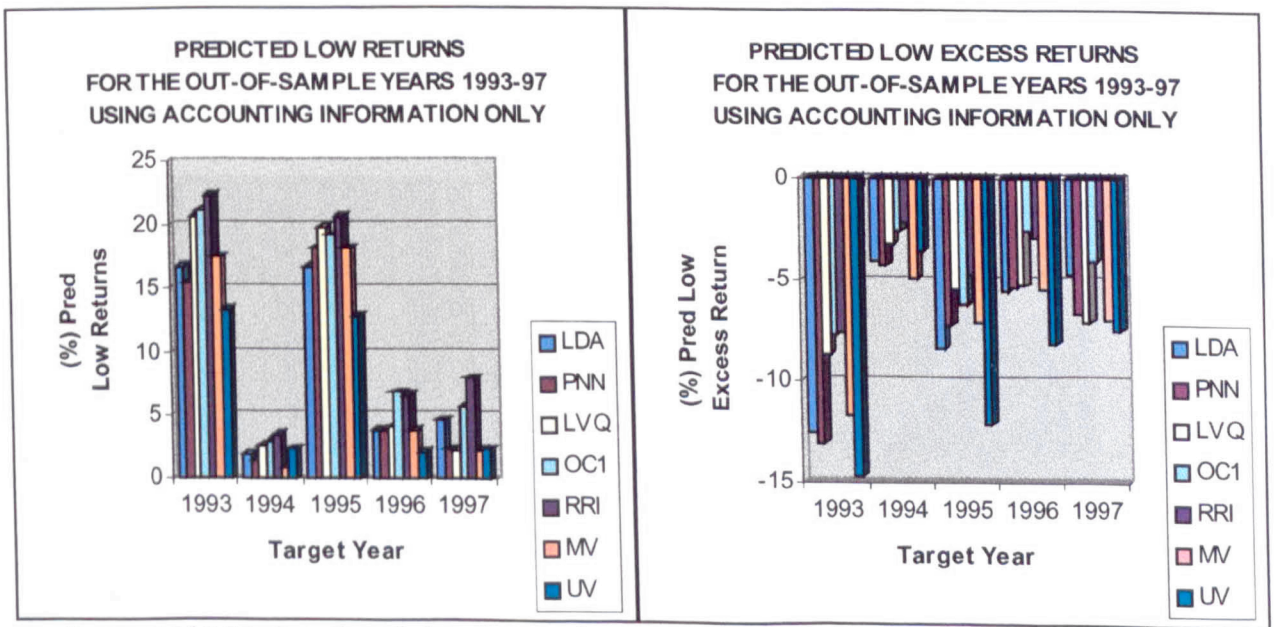
Figures 7.5-7.11: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using three different types of information to predict high and low performing shares

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
1993		Predicted Returns & Excess Returns													
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 90.1 L= 9.8	H= 31.4	55.9	16.7	57.1	15.6	41.7	20.7	41.2	21.1	40.0	22.3	50.2	17.5	79.1	13.3
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 59.0 L= -19.8	L= 29.6	24.5	-12.5	24.8	-13.1	11.1	-8.7	9.8	-7.7	10.0	-7.6	19.1	-11.7	44.7	-14.7
1994															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 44.2 L= -7.7	H= 5.5	9.6	1.9	10.5	1.4	8.7	2.7	8.1	3.0	7.6	3.5	11.2	0.9	15.1	2.4
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 38.7 L= -12.9	L= 5.3	5.1	-4.1	5.7	-4.3	4.1	-3.2	3.1	-2.6	2.7	-2.2	6.5	-4.9	11.4	-3.5
1995															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 78.8 L= 7.6	H= 25.0	37.2	16.7	35.0	18.2	33.8	19.8	34.7	19.2	30.4	20.6	35.5	18.2	42.5	12.9
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 53.7 L= -17.9	L= 25.6	11.3	-8.4	9.4	-7.2	8.0	-5.5	9.2	-6.2	4.8	-4.7	9.9	-7.1	16.0	-12.1
1996															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 50.5 L= -4.4	H= 9.0	16.8	3.9	17.1	3.9	15.6	4.1	12.5	6.8	13.6	6.5	16.6	3.8	27.0	2.1
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 41.44 L= -13.9	L= 9.5	7.6	-5.6	7.7	-5.4	6.3	-5.3	3.1	-2.6	4.2	-2.9	7.2	-5.5	16.5	-8.1
1997															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 58.8 L= -6.9	H= 9.7	15.9	4.6	18.7	2.4	20.3	2.3	15.1	5.7	11.5	8.0	20.0	2.3	32.5	2.4
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 49.1 L= -16.4	L= 9.5	5.9	-4.7	8.5	-6.7	10.5	-7.1	5.8	-4.1	2.4	-2.0	10.1	-7.0	22.5	-7.5

Table 7.6: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares



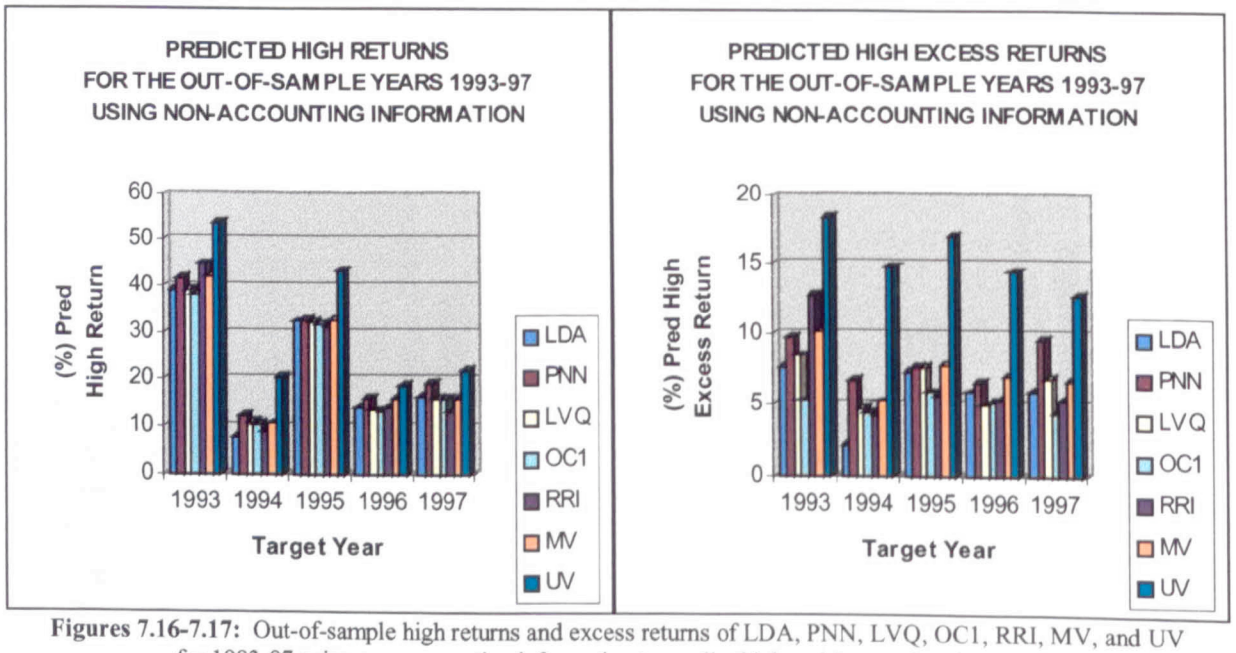
Figures 7.12-7.13: Out-of-sample high returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares



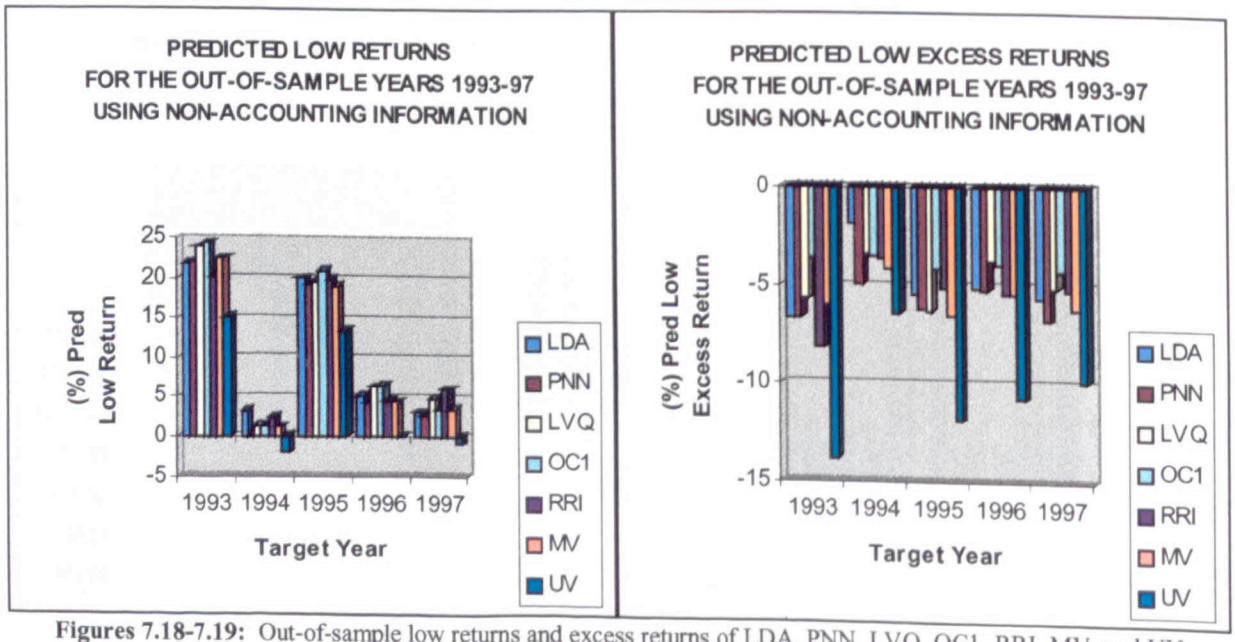
Figures 7.14-7.15: Out-of-sample low returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
1993		Predicted Returns & Excess Returns													
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 90.1 L= 9.8	H= 31.4	39.1	21.8	41.8	22.0	39.0	23.9	38.1	24.4	44.9	20.3	42.5	22.5	53.6	15.2
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 59.0 L= -19.8	L= 29.6	7.5	-6.6	9.7	-6.6	8.4	-5.7	5.2	-3.6	12.7	-8.2	10.2	-6.0	18.4	-13.8
1994															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 44.2 L= -7.7	H= 5.5	7.6	3.3	12.3	0.3	10.9	1.4	10.4	1.2	8.7	2.4	10.6	1.2	20.3	-2.0
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 38.7 L= -12.9	L= 5.3	2.1	-1.9	6.6	-4.9	4.7	-3.4	4.4	-3.5	4.2	-3.6	5.2	-4.1	14.8	-6.4
1995															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 78.8 L= 7.6	H= 25.0	32.5	20.1	32.8	19.4	32.4	19.7	31.8	21.0	31.5	19.9	32.9	19.1	43.5	13.5
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 53.7 L= -17.9	L= 25.6	7.2	-5.5	7.6	-6.2	7.6	-6.3	5.8	-4.1	5.5	-5.1	7.7	-6.5	17.0	-11.9
1996															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 50.5 L= -4.4	H= 9.0	14.0	5.2	15.8	4.2	13.3	6.4	13.1	6.5	13.9	4.7	15.7	4.5	18.6	0.1
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 41.44 L= -13.9	L= 9.5	5.8	-5.1	6.5	-5.2	4.9	-3.7	5.1	-3.9	5.3	-5.5	6.9	-5.4	14.5	-10.8
1997															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 58.8 L= -6.9	H= 9.7	16.3	3.1	19.1	2.7	16.0	4.8	15.8	3.6	13.1	6.1	16.0	3.6	21.9	-0.9
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 49.1 L= -16.4	L= 9.5	5.9	-5.7	9.5	-6.8	6.8	-5.1	4.4	-4.2	5.3	-5.2	6.7	-6.2	12.8	-9.9

Table 7.7: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares



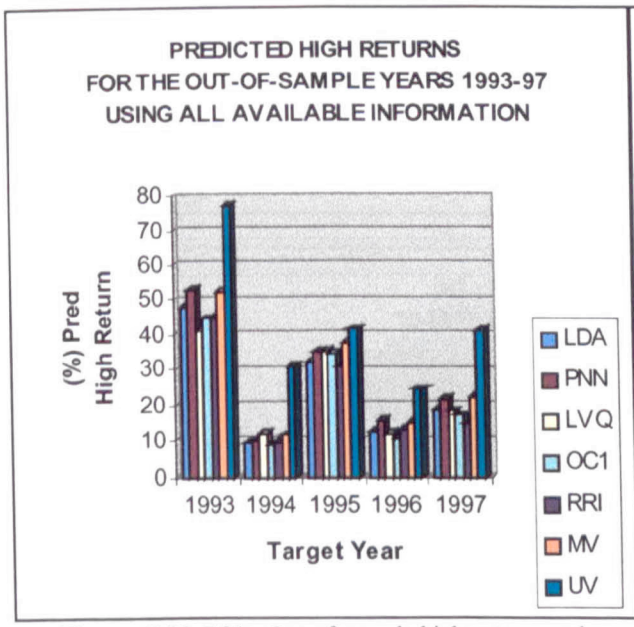
Figures 7.16-7.17: Out-of-sample high returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares

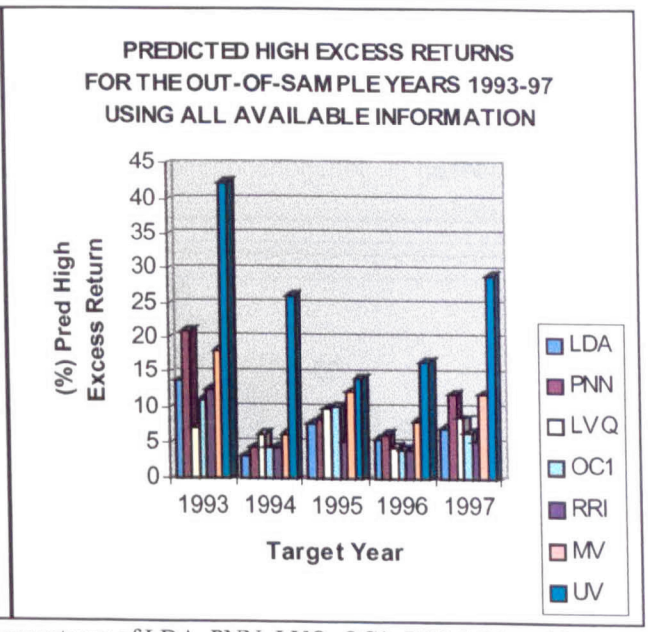


Figures 7.18-7.19: Out-of-sample low returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares

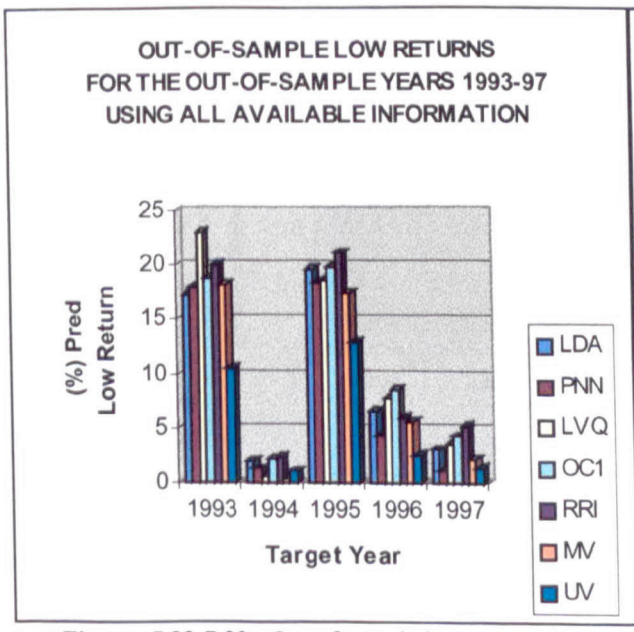
		LDA		PNN		LVQ		OC1		RRI		MV		UV	
1993		Predicted Returns & Excess Returns													
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 90.1 L= 9.8	H= 31.4	48.0	17.3	53.0	18.1	41.0	23.0	44.6	18.8	45.1	20.1	52.3	18.3	77.1	10.5
Actual Excess Ret		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 59.0 L= -19.8	L= 29.6	13.8	-9.6	21.0	-10.7	7.1	-4.4	11.2	-8.5	12.6	-8.2	18.3	-9.5	42.3	-13.5
1994															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 44.2 L= -7.7	H= 5.5	9.4	2.0	10.3	1.3	12.1	0.6	9.0	2.2	9.4	2.4	11.7	0.4	30.9	1.1
Actual Excess Ret		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 38.7 L= -12.9	L= 5.3	3.3	-2.7	4.5	-3.7	6.2	-4.3	4.3	-3.7	4.4	-3.2	6.2	-4.9	26.1	-3.2
1995															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 78.8 L= 7.6	H= 25.0	32.3	19.7	35.0	18.4	35.3	18.6	34.5	19.9	30.6	21.1	37.5	17.4	41.5	12.9
Actual Excess Ret		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 53.7 L= -17.9	L= 25.6	7.7	-6.4	8.5	-6.3	9.9	-6.9	10.2	-6.2	5.3	-4.4	12.4	-8.3	14.3	-13.6
1996															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 50.5 L= -4.4	H= 9.0	12.8	6.6	15.4	4.4	11.8	7.7	10.7	8.5	13.4	6.0	15.1	5.6	24.5	2.6
Actual Excess Ret		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 41.44 L= -13.9	L= 9.5	5.8	-4.7	6.2	-5.1	4.4	-3.0	4.1	-2.5	4.1	-3.4	8.0	-5.3	16.8	-9.5
1997															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 58.8 L= -6.9	H= 9.7	18.9	3.1	21.9	1.2	18.4	3.6	17.1	4.3	14.7	5.2	22.2	2.2	41.4	1.4
Actual Excess Ret		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 49.1 L= -16.4	L= 9.5	7.3	-5.1	12.1	-8.2	8.6	-5.9	6.5	-4.6	5.2	-4.5	12.0	-7.0	29.3	-6.9

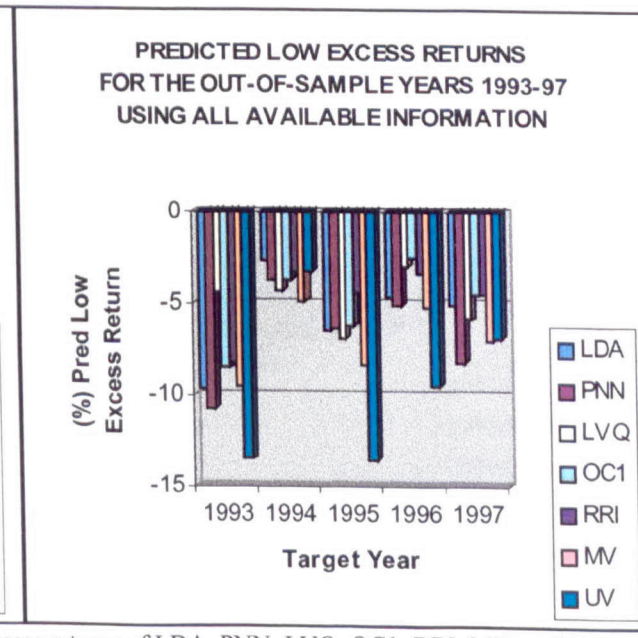
Table 7.8: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares



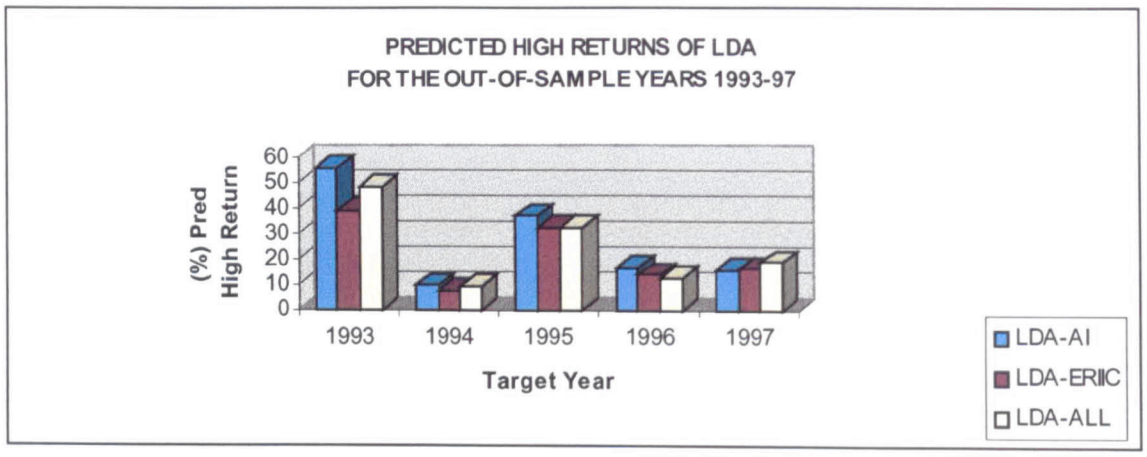


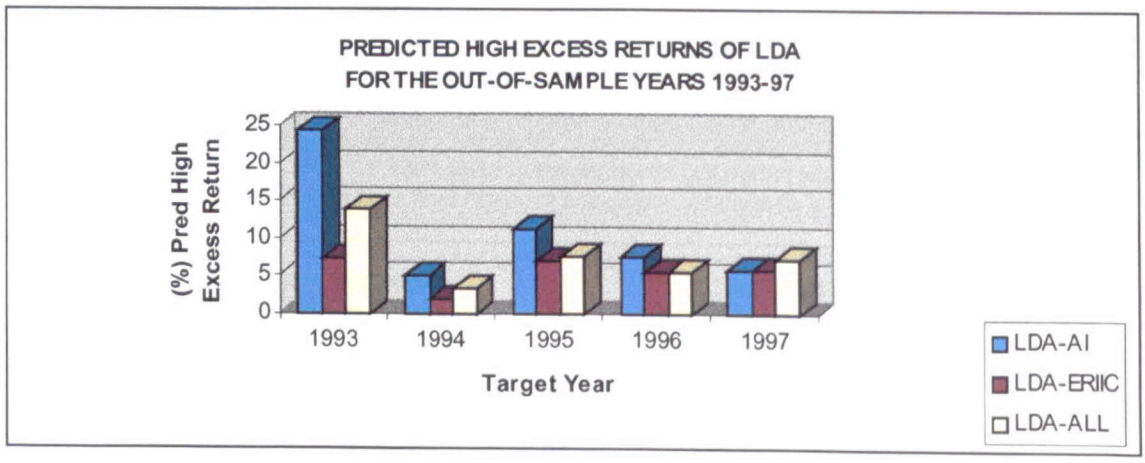
Figures 7.20-7.21: Out-of-sample high returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares

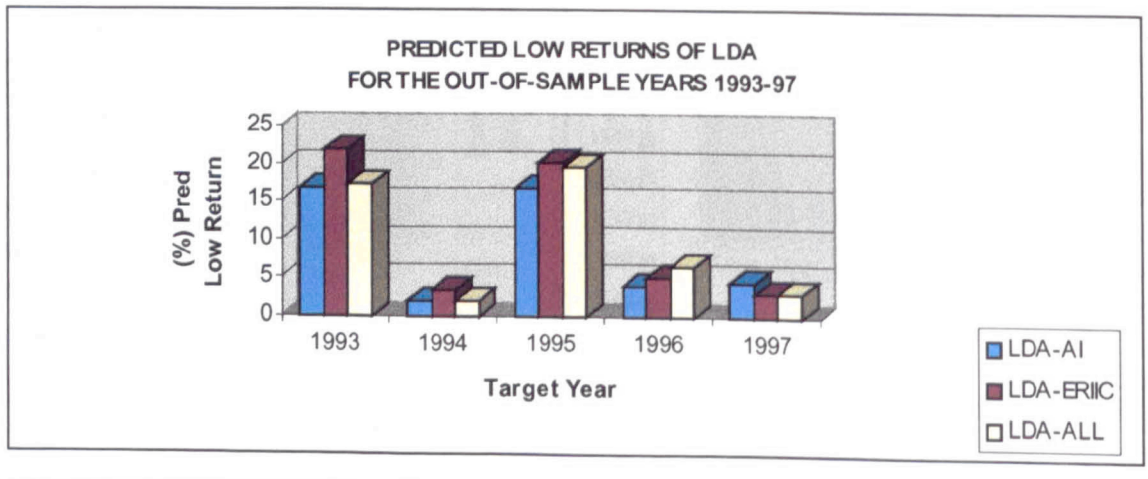


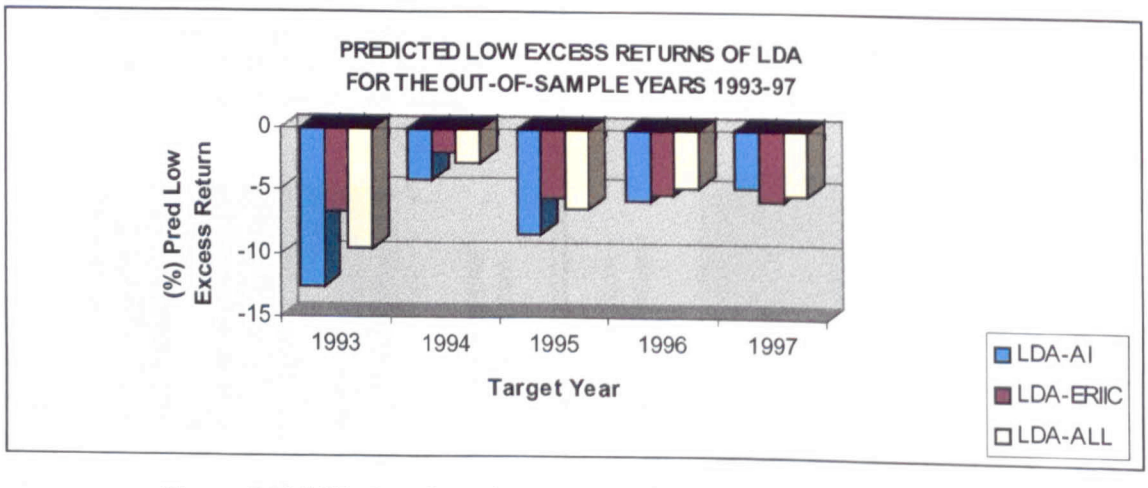


Figures 7.22-7.23: Out-of-sample low returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares

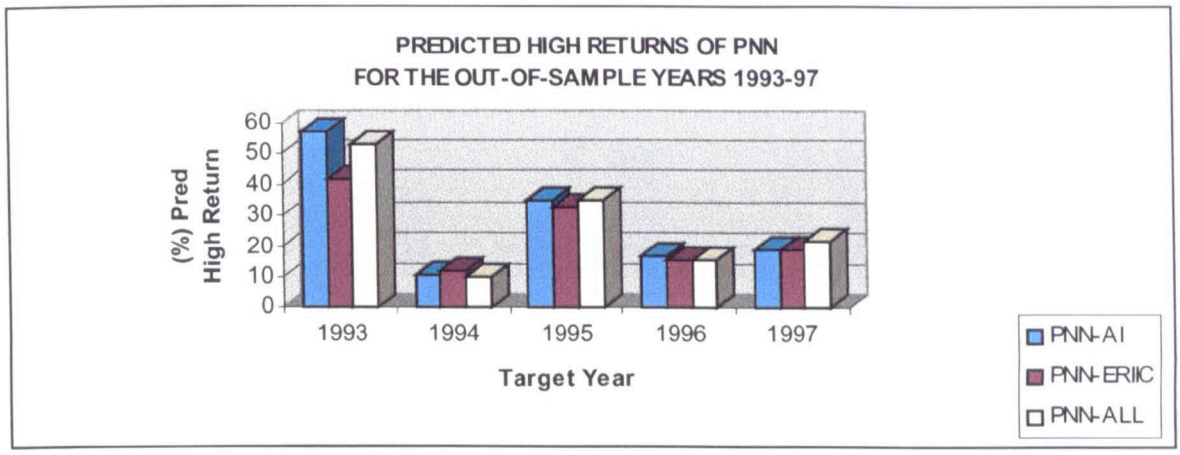


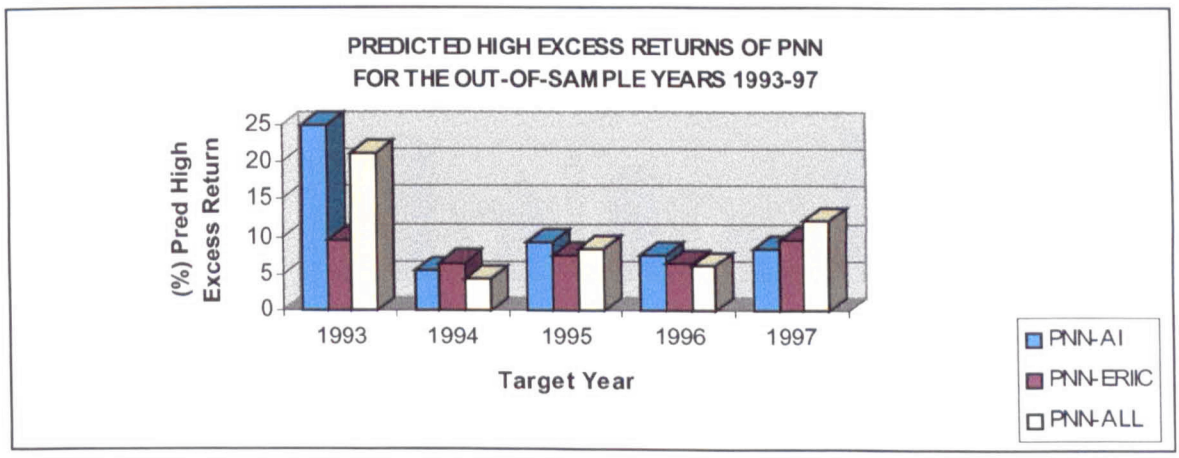


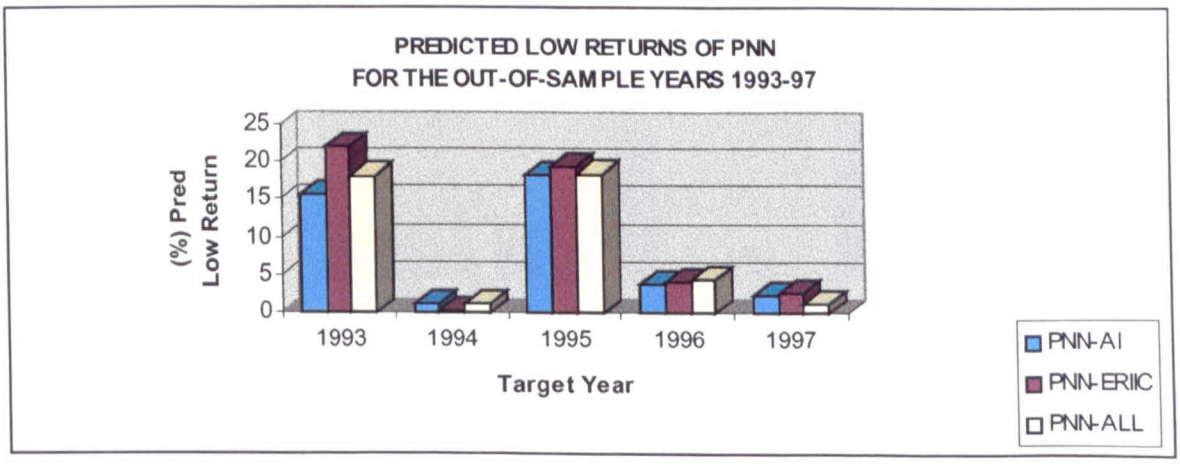


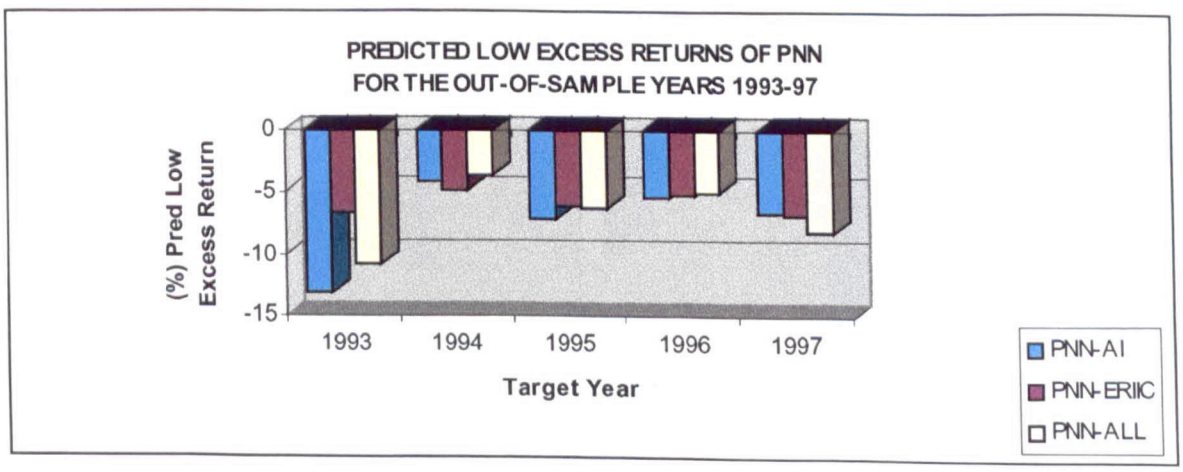


Figures 7.24-7.27: Out-of-sample returns and excess returns of LDA for 1993-97 using three different types of information to predict high and low performing shares

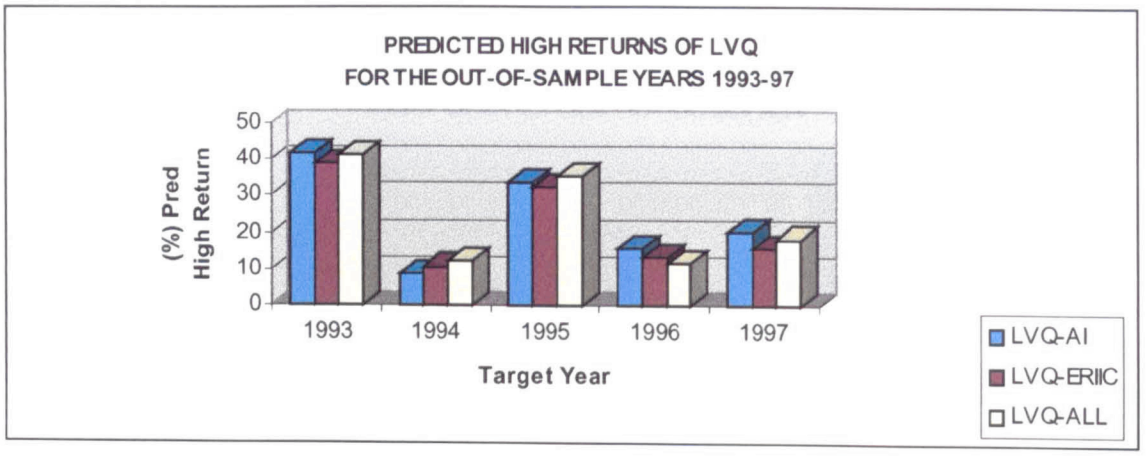


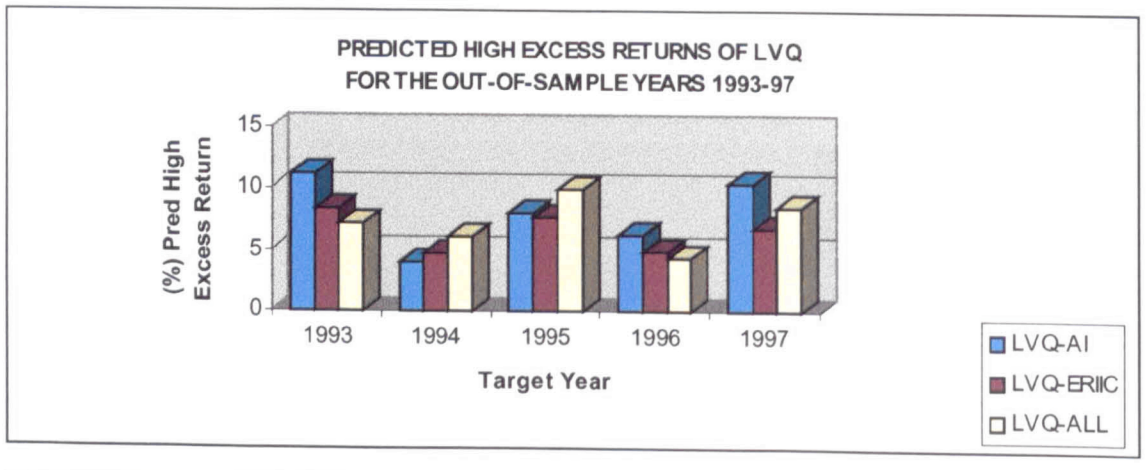


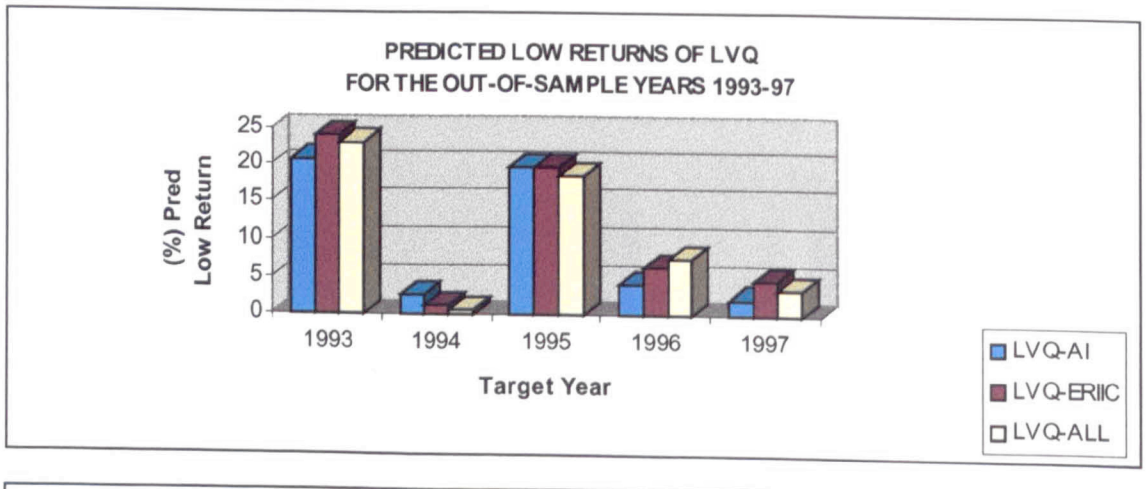


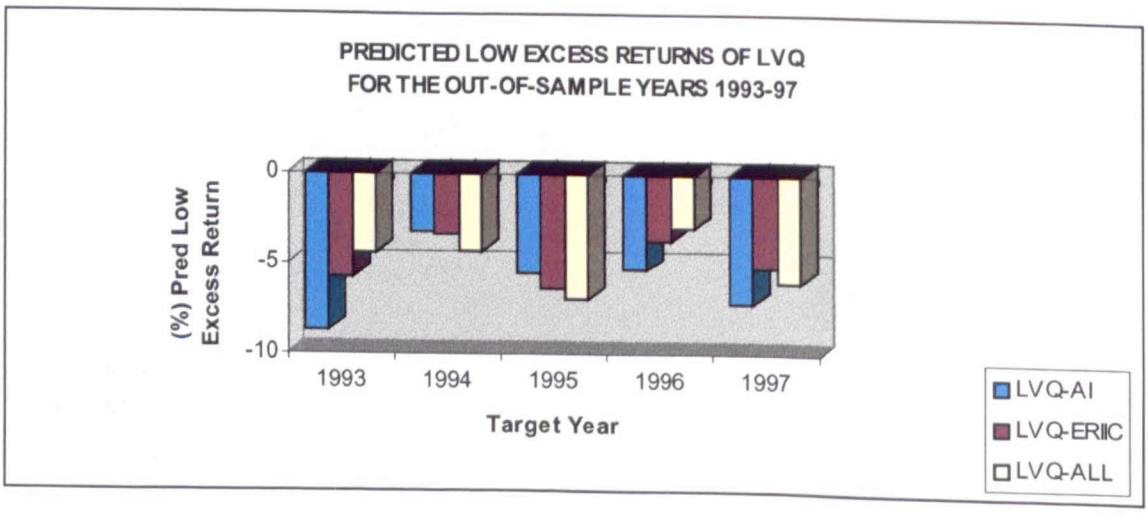


Figures 7.28-7.31: Out-of-sample returns and excess returns of PNN for 1993-97 using three different types of information to predict high and low performing shares

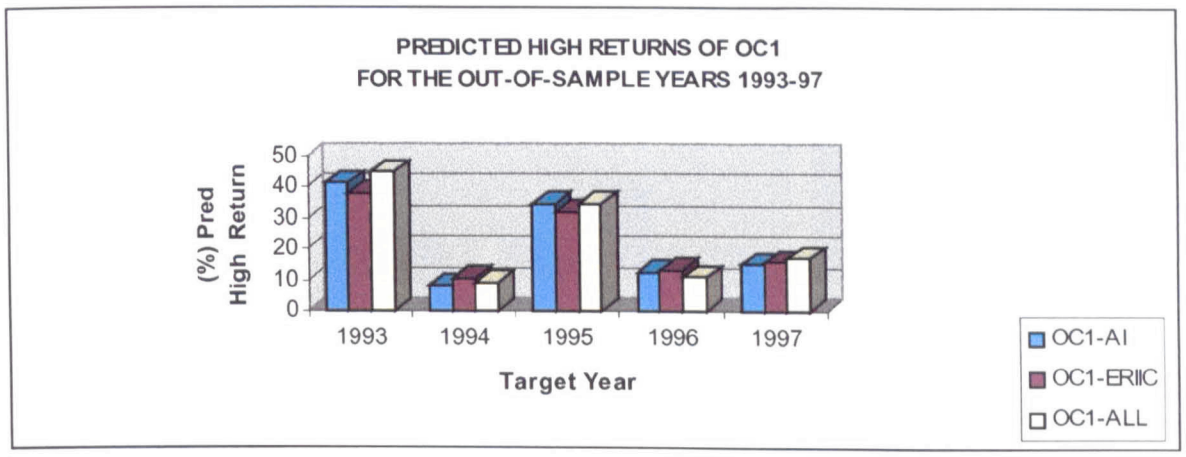


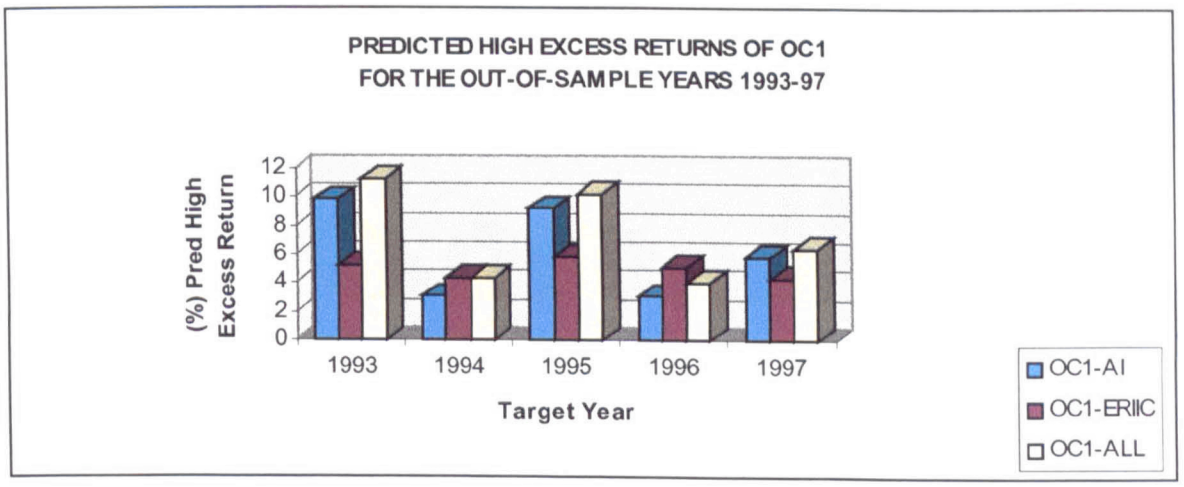


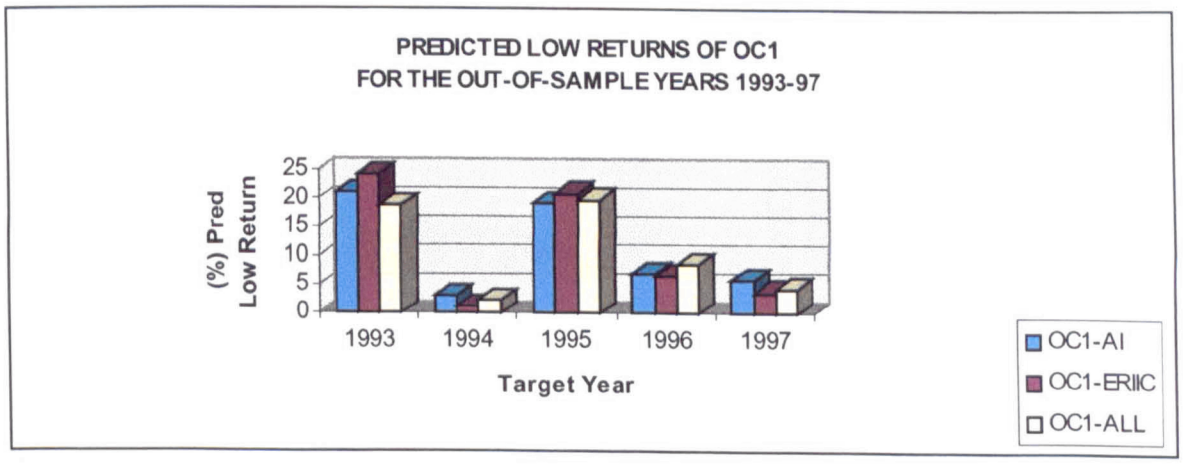


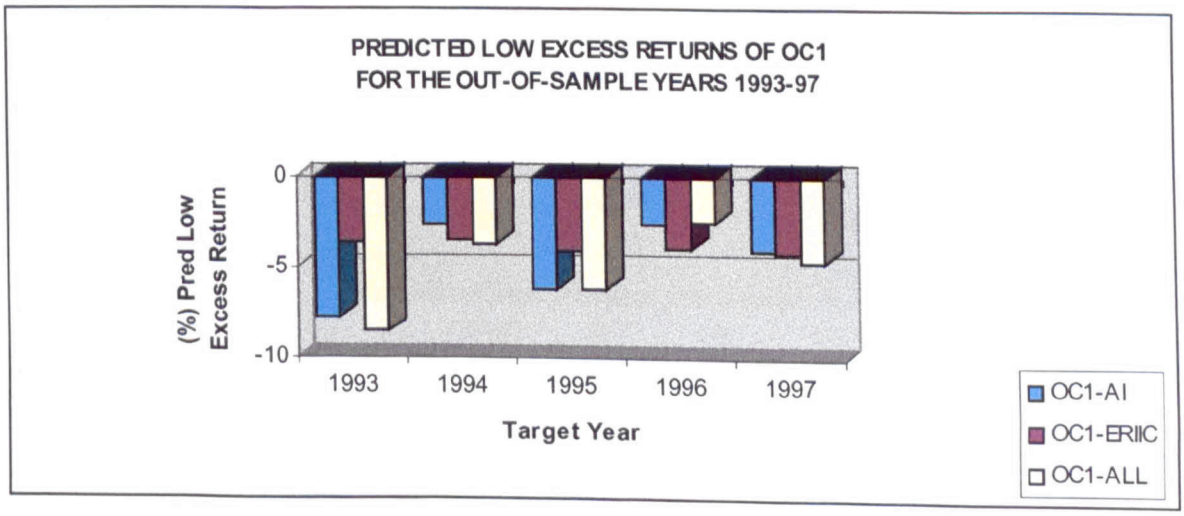


Figures 7.32-7.35: Out-of-sample returns and excess returns of LVQ for 1993-97 using three different types of information to predict high and low performing shares

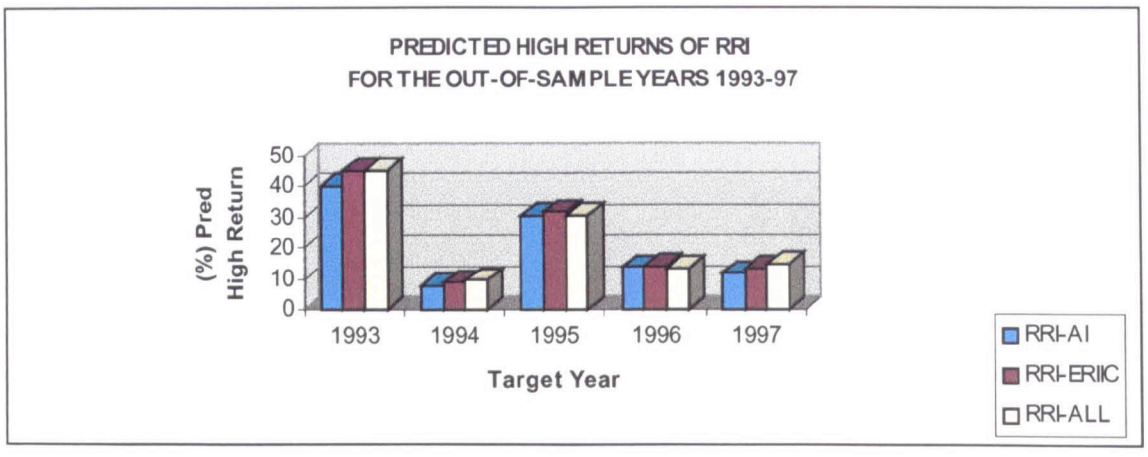


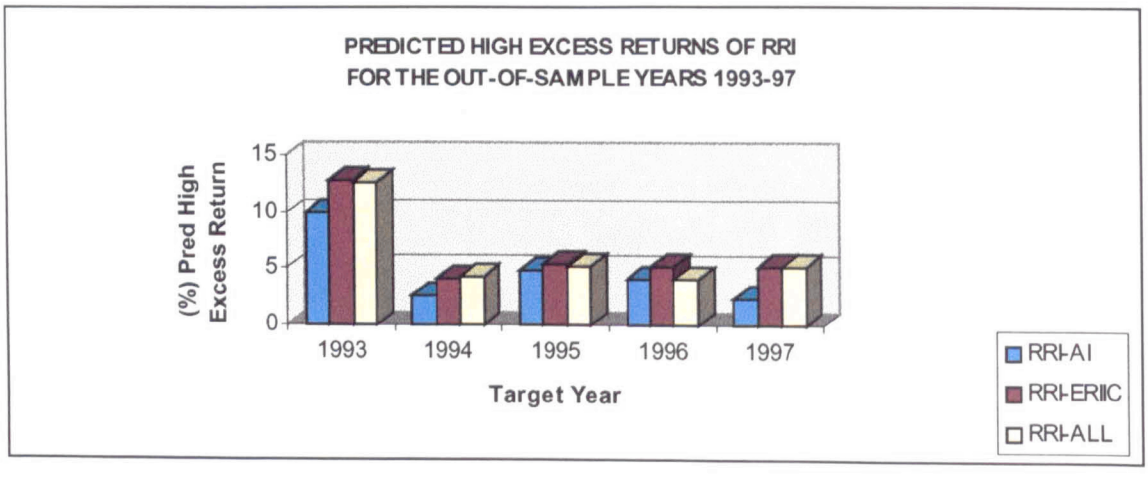


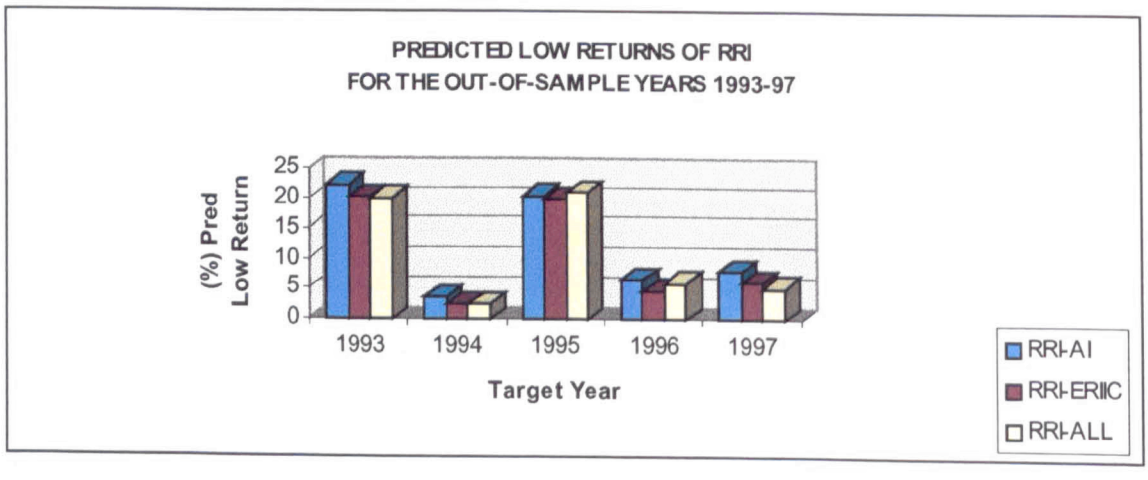


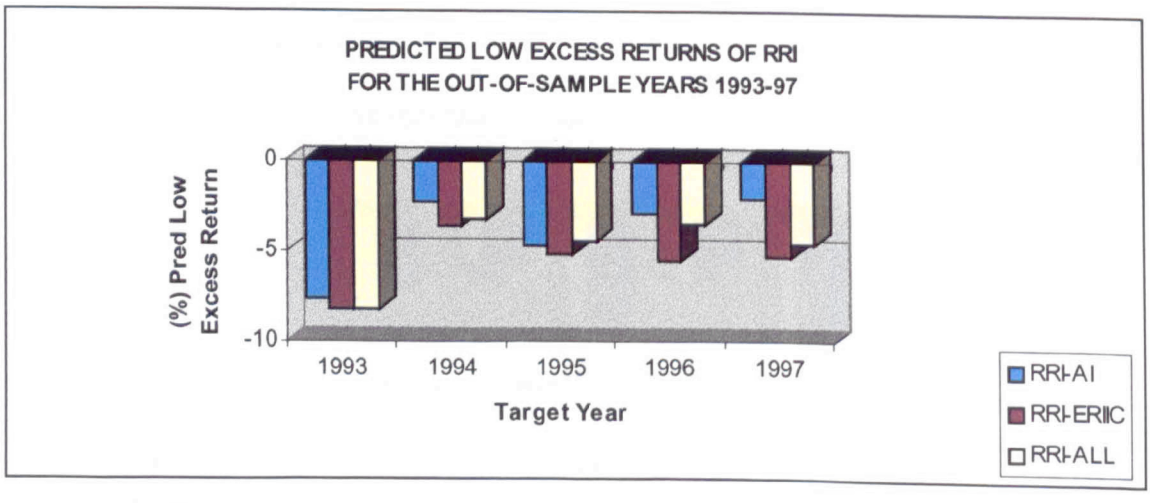


Figures 7.36-7.39: Out-of-sample returns and excess returns of OC1 for 1993-97 using three different types of information to predict high and low performing shares

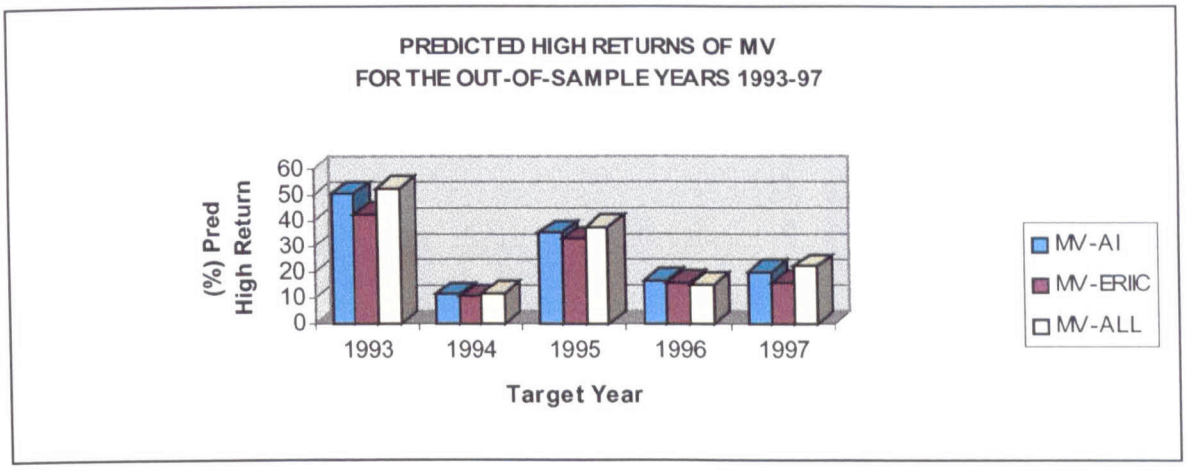


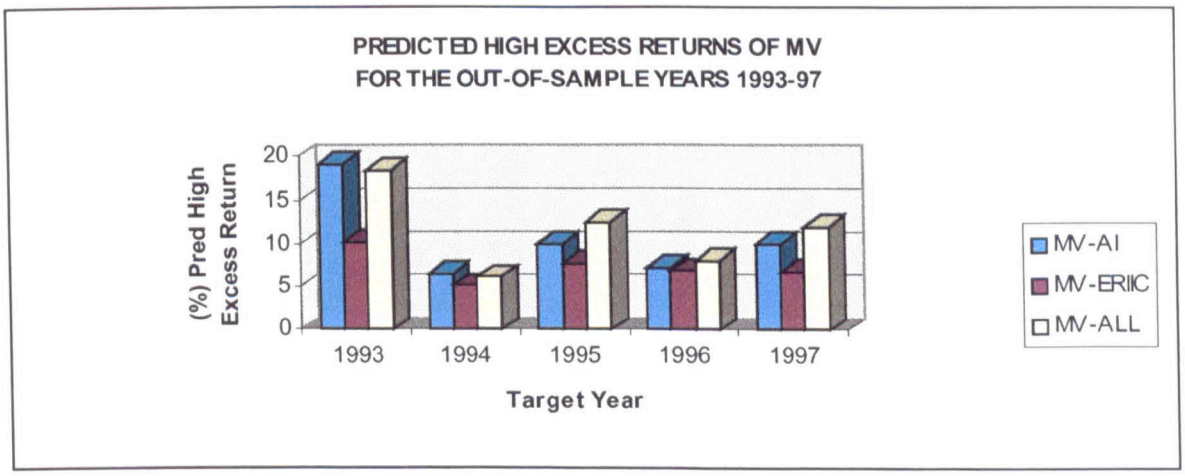


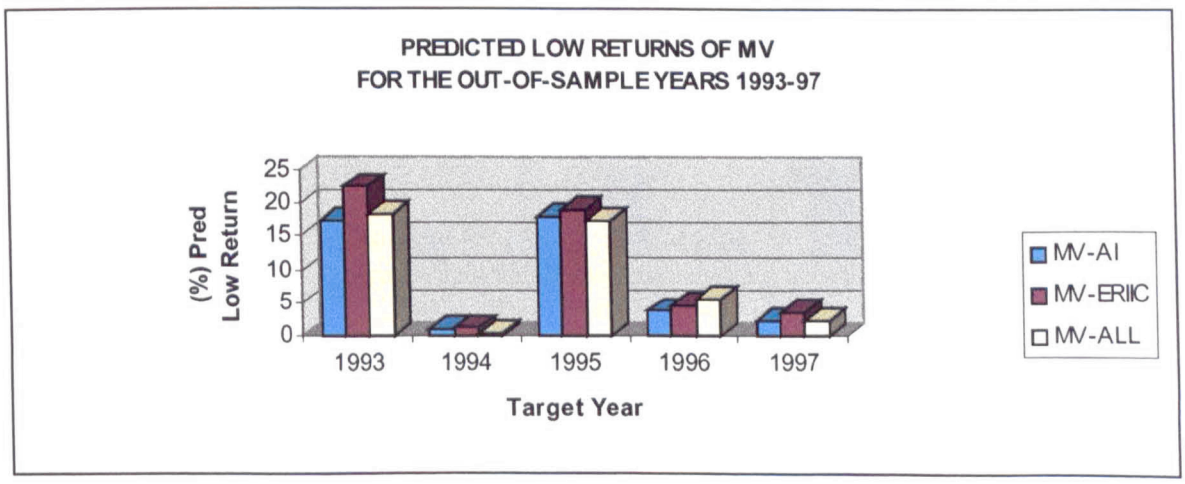


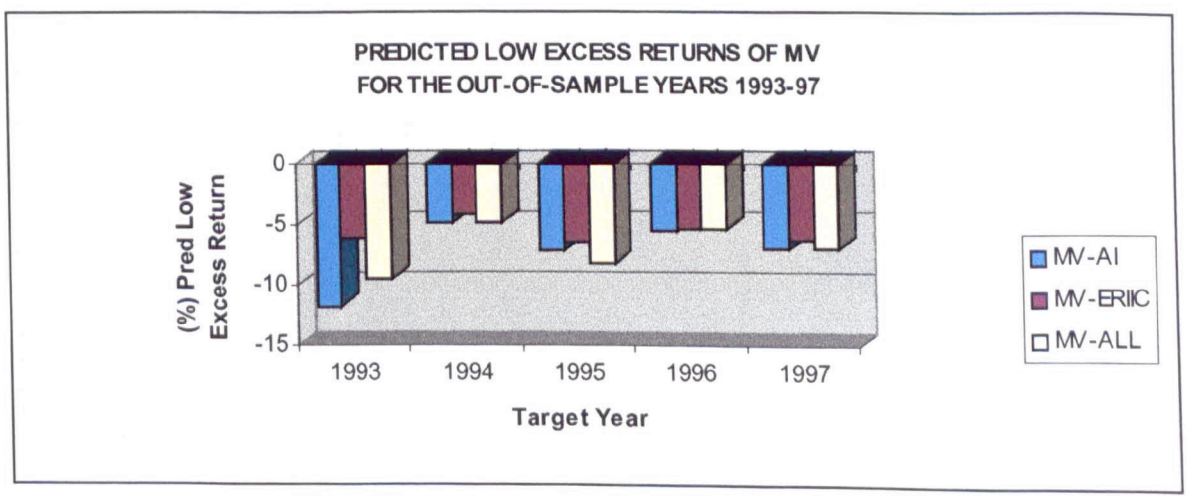


Figures 7.40-7.43: Out-of-sample returns and excess returns of RRI for 1993-97 using three different types of information to predict high and low performing shares

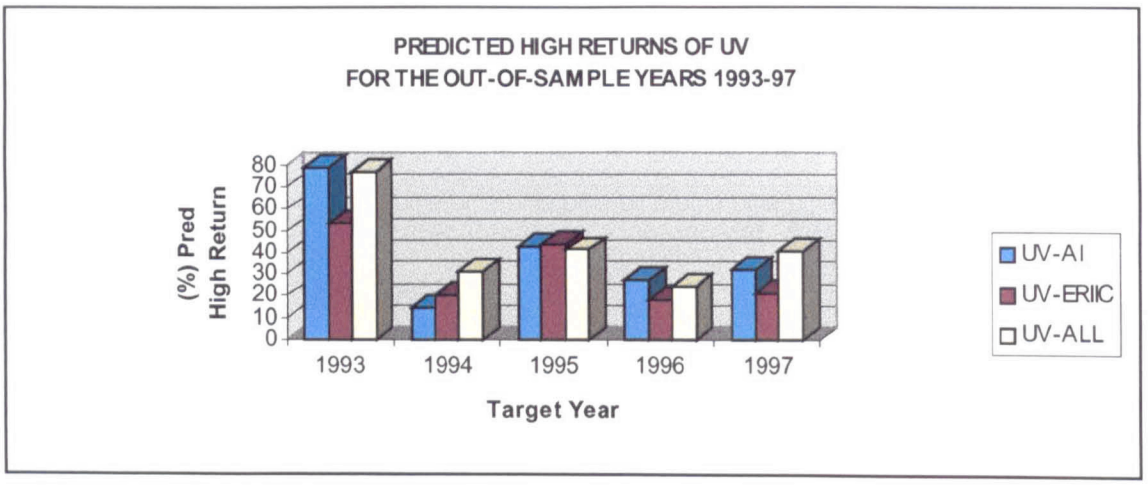


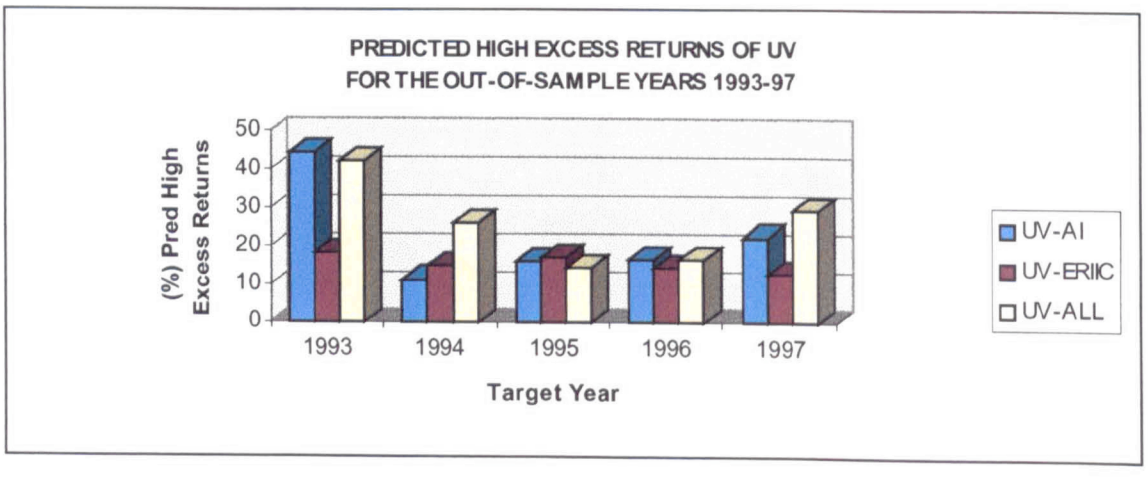


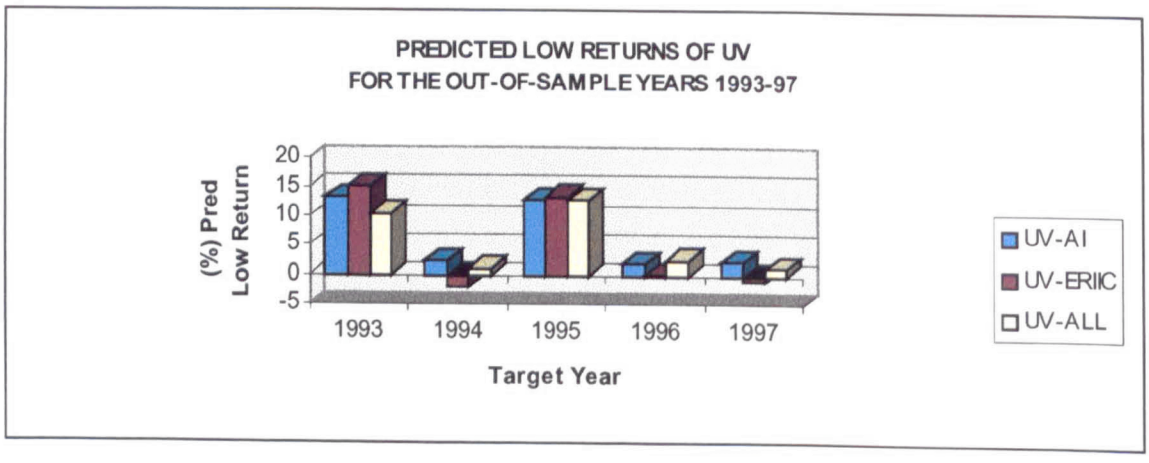


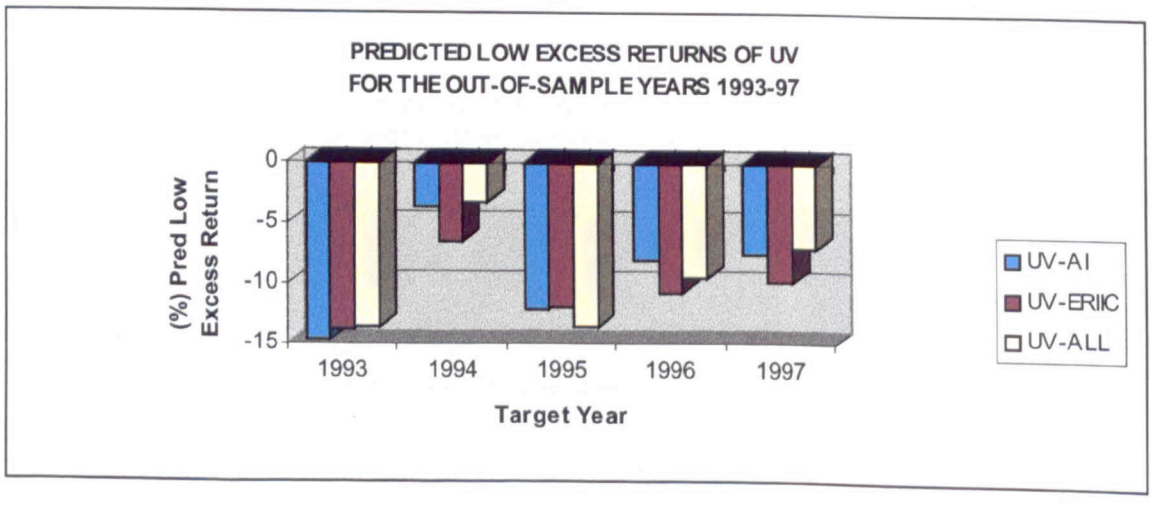


Figures 7.44-7-47: Out-of-sample returns and excess returns of MV for 1993-97 using three different types of information to predict high and low performing shares









Figures 7.48-7.51: Out-of-sample returns and excess returns of UV for 1993-97 using three different types of information to predict high and low performing shares

	PNN	LDA	LVQ	OC1	RRI	MV	UV
Target Year	*Predicted Trading Volume*						
1993	216	211	275	275	271	238	61
1994	267	276	271	282	281	266	77
1995	275	273	259	256	317	268	96
1996	282	289	310	312	276	296	71
1997	315	317	290	297	326	294	71

Table 7.9: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

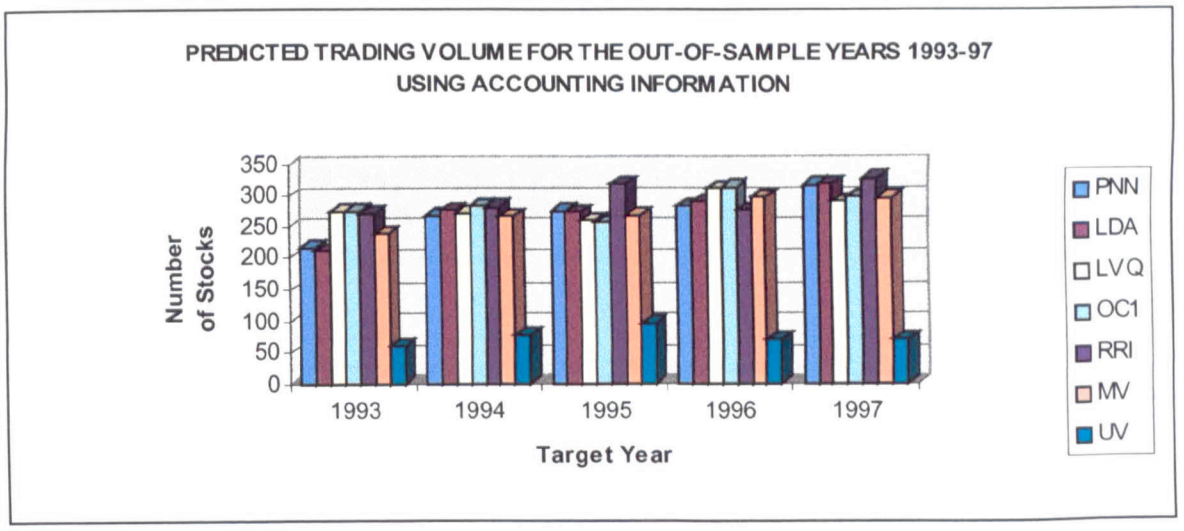


Figure 7.52: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using accounting information to predict high and low performing shares

	PNN	LDA	LVQ	OC1	RRI	MV	UV
Target Year	*Predicted Trading Volume*						
1993	252	293	251	254	246	233	90
1994	262	293	257	277	286	271	63
1995	288	277	289	263	306	293	74
1996	303	321	295	296	345	298	67
1997	300	352	308	352	357	347	113

Table 7.10: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares

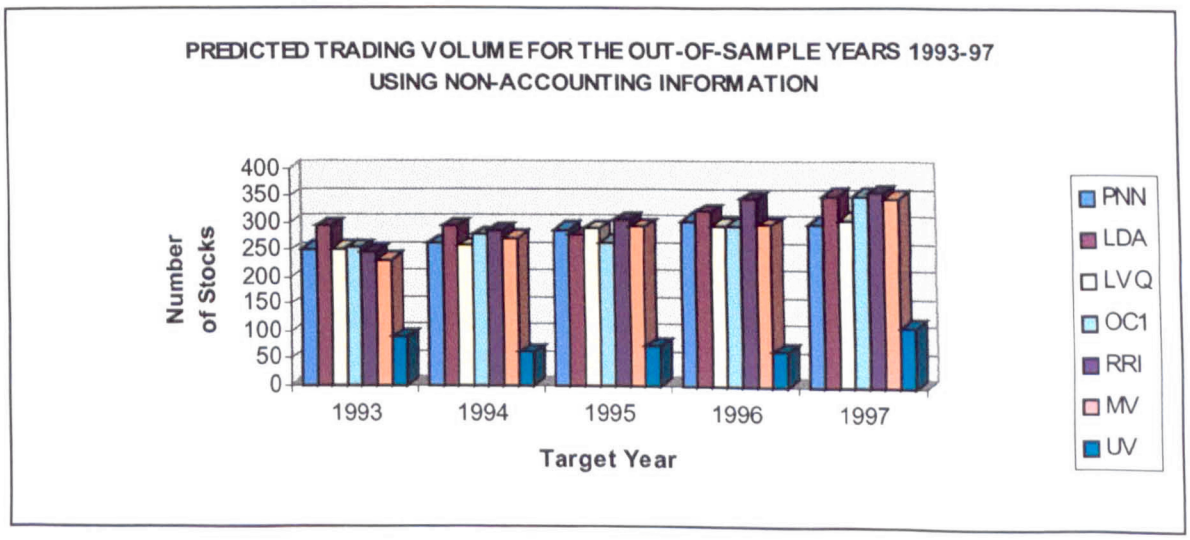


Figure 7.53: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using non-accounting information to predict high and low performing shares

	PNN	LDA	LVQ	OC1	RRI	MV	UV
Target Year	*Predicted Trading Volume*						
1993	212	257	240	270	246	214	72
1994	279	279	253	287	262	272	46
1995	271	291	262	241	292	256	42
1996	308	306	279	261	310	272	66
1997	290	295	292	297	331	265	76

Table 7.11: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares

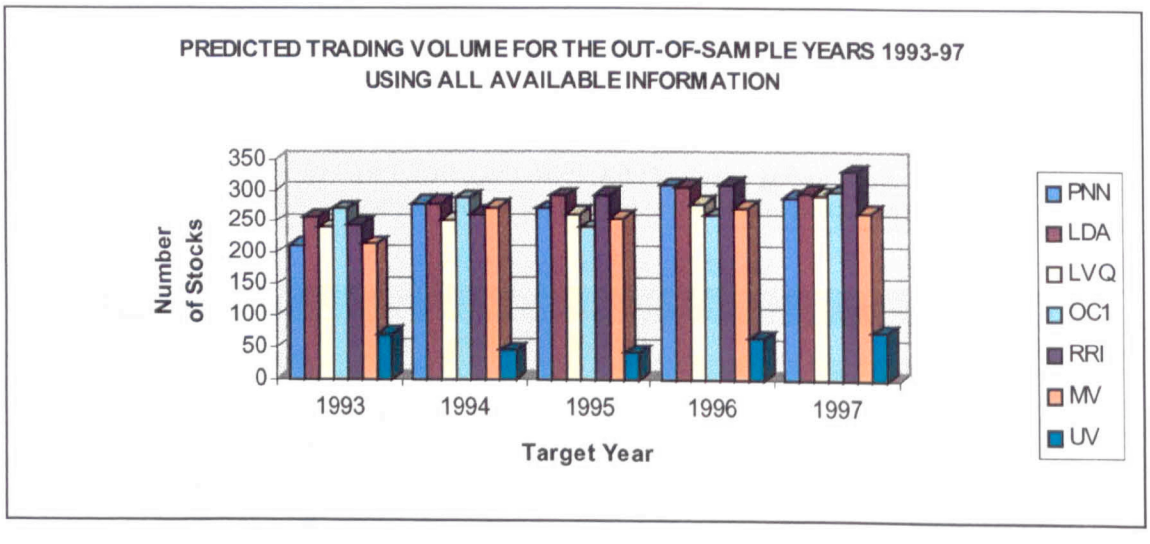
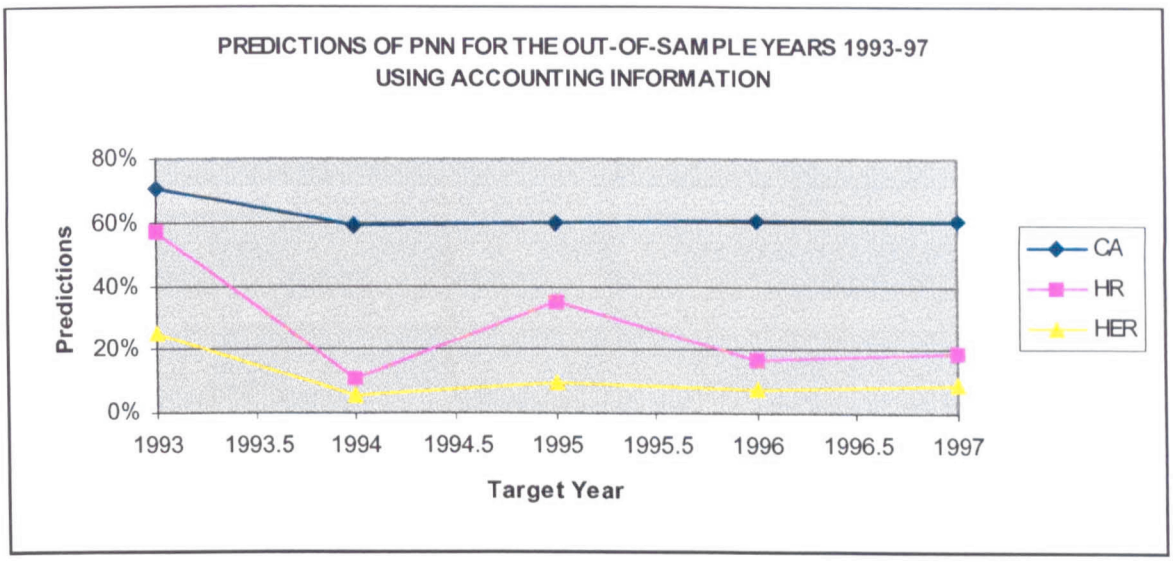
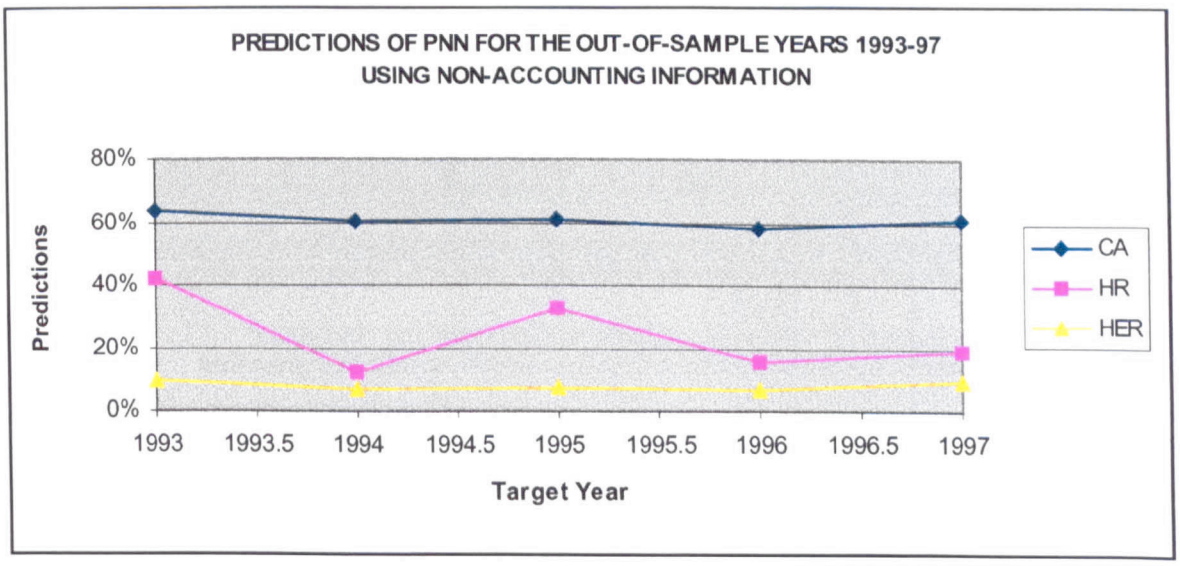
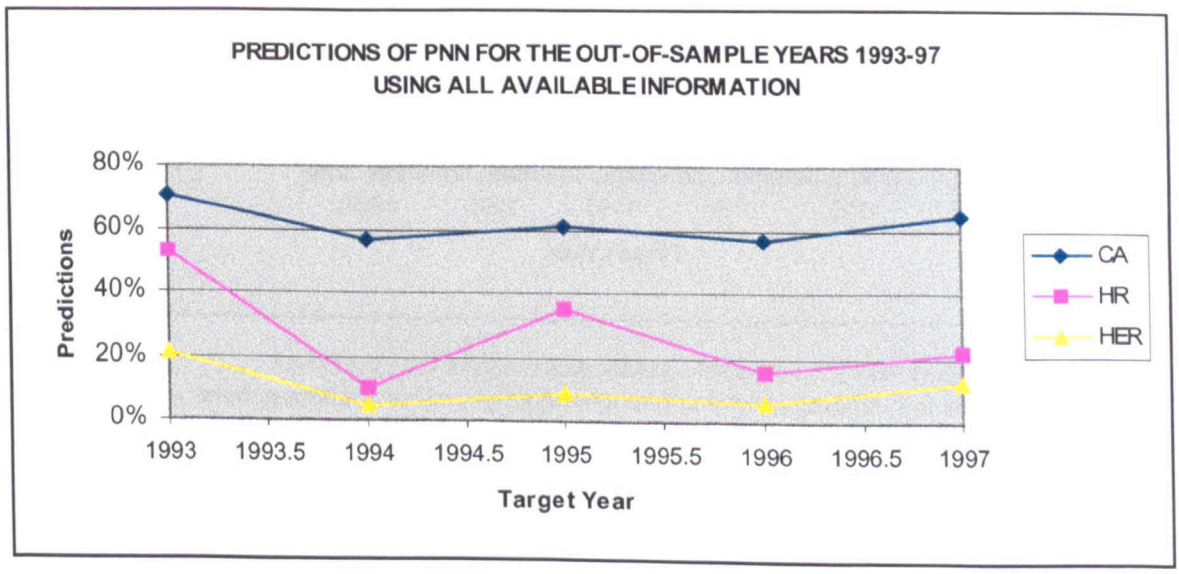


Figure 7.54: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1993-97 using all available information to predict high and low performing shares







Figures 7.55-7.57: Out-of-sample predictions of PNN for 1993-97 using three different types of information to predict high and low performing shares

		UV-AI		UV-ERIIC		UV-ALL		UV-2V	
Actual Class	Patterns	Predicted Group Membership							
1993		H	L	H	L	H	L	H	L
H	157	36	121	37	120	44	113	14	143
L	469	25	444	15	454	28	441	3	466
Overall (%)		76.68 %		78.43 %		77.48 %		76.68 %	
1994		H	L	H	L	H	L	H	L
H	155	25	130	27	128	21	134	8	147
L	463	52	411	36	427	25	438	6	457
Overall (%)		70.55 %		73.46 %		74.27 %		75.24 %	
1995		H	L	H	L	H	L	H	L
H	160	31	129	30	130	16	144	6	154
L	479	65	414	44	435	26	453	6	473
Overall (%)		69.64 %		72.77 %		73.40 %		74.96 %	
1996		H	L	H	L	H	L	H	L
H	171	34	137	22	149	31	140	31	140
L	510	37	473	45	465	35	475	35	475
Overall (%)		74.45 %		71.51 %		74.30 %		74.30 %	
1997		H	L	H	L	H	L	H	L
H	180	32	148	39	141	44	136	11	169
L	538	39	499	74	464	32	506	11	527
Overall (%)		73.96 %		70.06 %		76.60 %		74.93 %	

Table 7.12: Out-of-sample classification results of UV for 1993-97 performing four different implementations to predict high and low performing shares

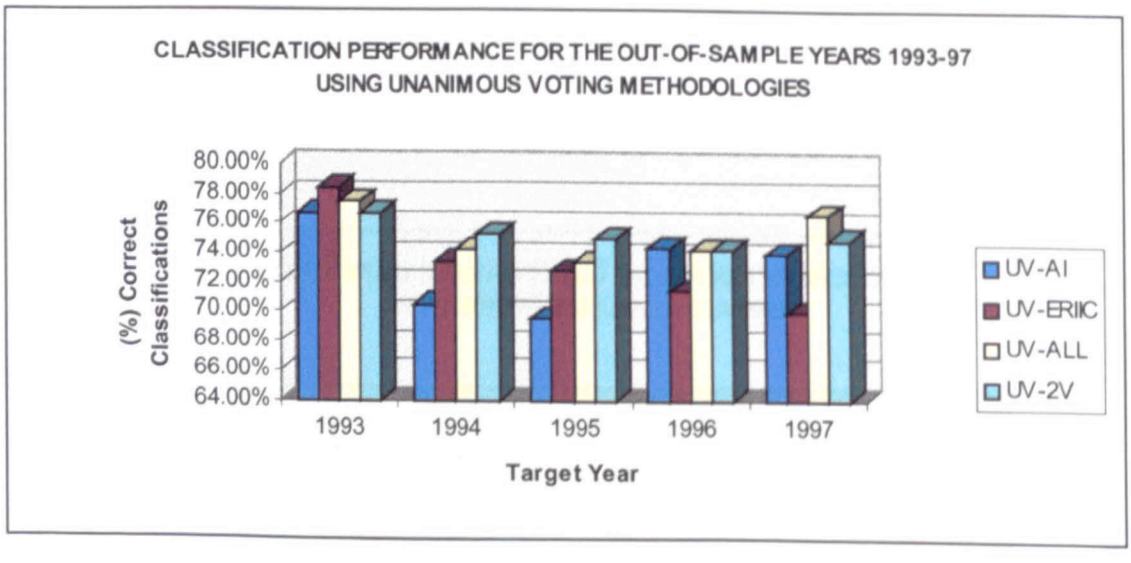
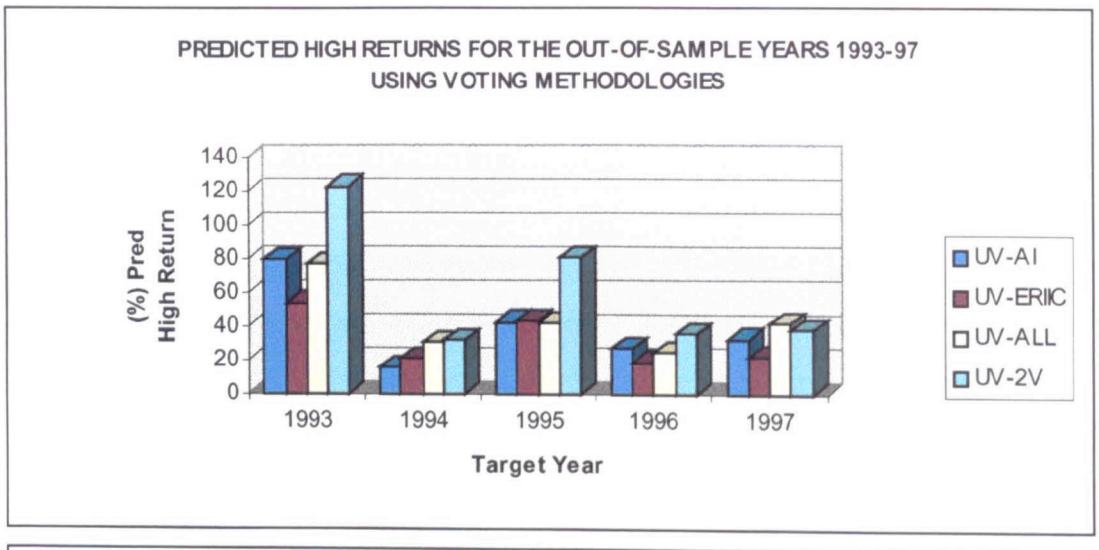
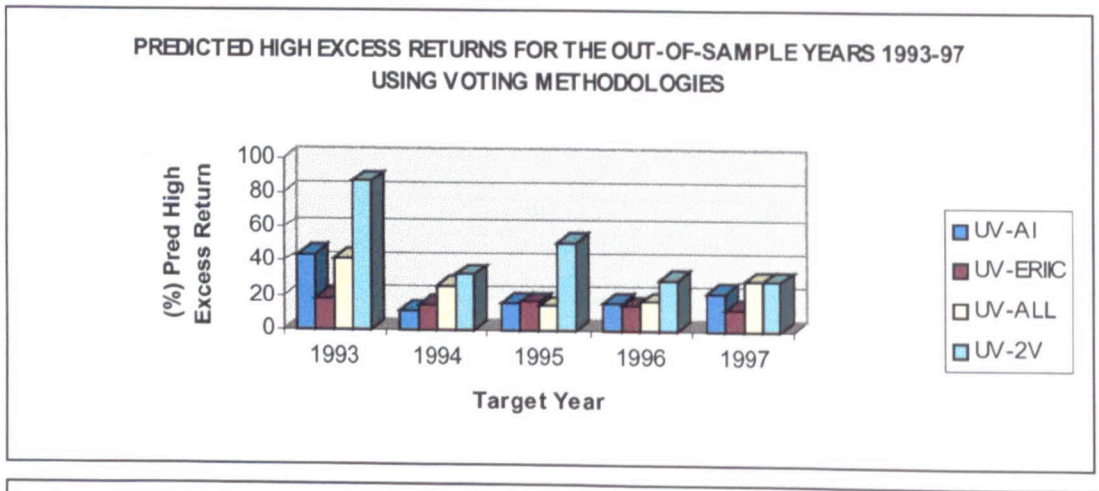


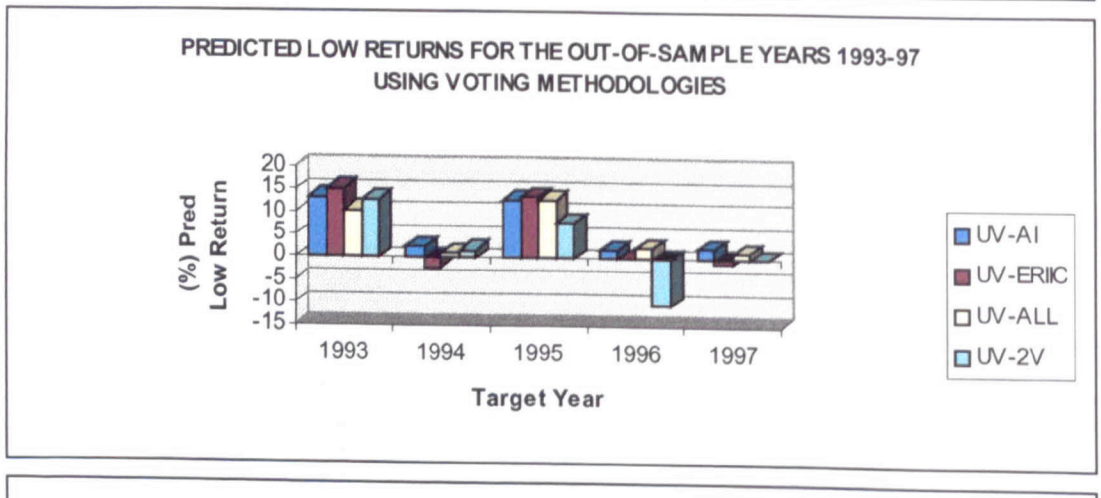
Figure 7.58: Out-of-sample classification results of UV for 1993-97 performing four different implementations to predict high and low performing shares

		UV-AI		UV-ERIC		UV-ALL		UV-2V	
1993		Predicted Returns & Excess Returns							
Actual Return	Index	H	L	H	L	H	L	H	L
H= 90.1 L= 9.8	H= 31.4	79.1	13.3	53.6	15.2	77.1	10.5	122.47	12.87
Actual Excess Ret	Index	H	L	H	L	H	L	H	L
H= 59.0 L= -19.8	L= 29.6	44.7	-14.7	18.4	-13.8	42.3	-13.5	86.78	-15.38
1994									
Actual Return	Index	H	L	H	L	H	L	H	L
H= 44.2 L= -7.7	H= 5.5	15.1	2.4	20.3	-2.0	30.9	1.1	32.22	1.83
Actual Excess Ret	Index	H	L	H	L	H	L	H	L
H= 38.7 L= -12.9	L= 5.3	11.4	-3.5	14.8	-6.4	26.1	-3.2	32.94	-2.37
1995									
Actual Return	Index	H	L	H	L	H	L	H	L
H= 78.8 L= 7.6	H= 25.0	42.5	12.9	43.5	13.5	41.5	12.9	80.35	7.91
Actual Excess Ret	Index	H	L	H	L	H	L	H	L
H= 53.7 L= -17.9	L= 25.6	16.0	-12.1	17.0	-11.9	14.3	-13.6	51.70	-16.41
1996									
Actual Return	Index	H	L	H	L	H	L	H	L
H= 50.5 L= -4.4	H= 9.0	27.0	2.1	18.6	0.1	24.5	2.6	35.84	-10.49
Actual Excess Ret	Index	H	L	H	L	H	L	H	L
H= 41.44 L= -13.9	L= 9.5	16.5	-8.1	14.5	-10.8	16.8	-9.5	29.56	-20.94
1997									
Actual Return	Index	H	L	H	L	H	L	H	L
H= 58.8 L= -6.9	H= 9.7	32.5	2.4	21.9	-0.9	41.4	1.4	37.85	0.08
Actual Excess Ret	Index	H	L	H	L	H	L	H	L
H= 49.1 L= -16.4	L= 9.5	22.5	-7.5	12.8	-9.9	29.3	-6.9	29.08	-9.17

Table 7.13: Out-of-sample returns and excess returns of UV for 1993-97 performing four different implementations to predict high and low performing shares







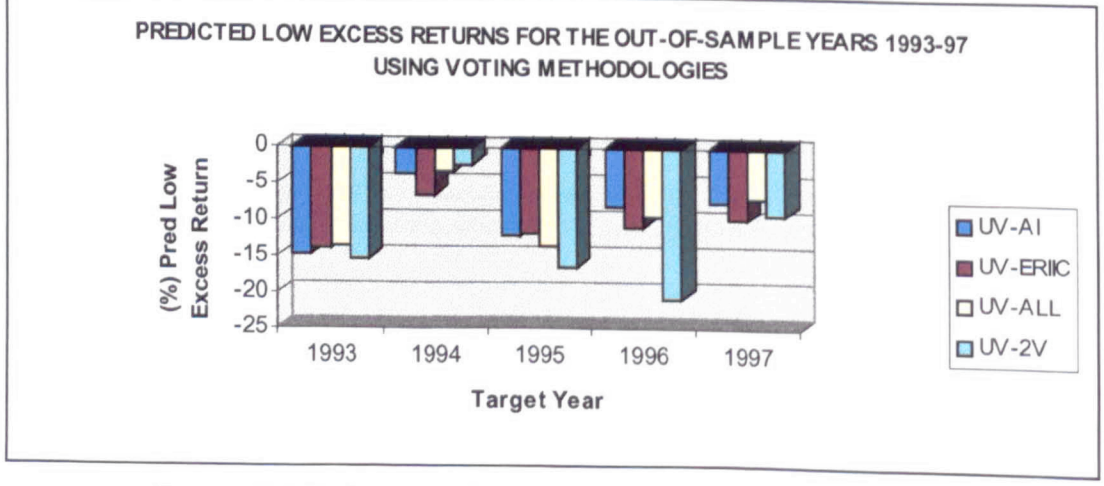


Figure 7.59-7.62: Out-of-sample returns and excess returns of UV for 1993-97 performing four different implementations to predict high and low performing shares

	UV-AI	UV-ERIIC	UV-ALL	UV-2V
Target Year	*Predicted Trading Volume*			
1993	61	90	72	17
1994	77	63	46	14
1995	96	74	42	12
1996	71	67	66	66
1997	71	113	76	22

Table 7.14: Out-of-sample trading volume of UV for 1993-97 performing four different implementations to predict high and low performing shares

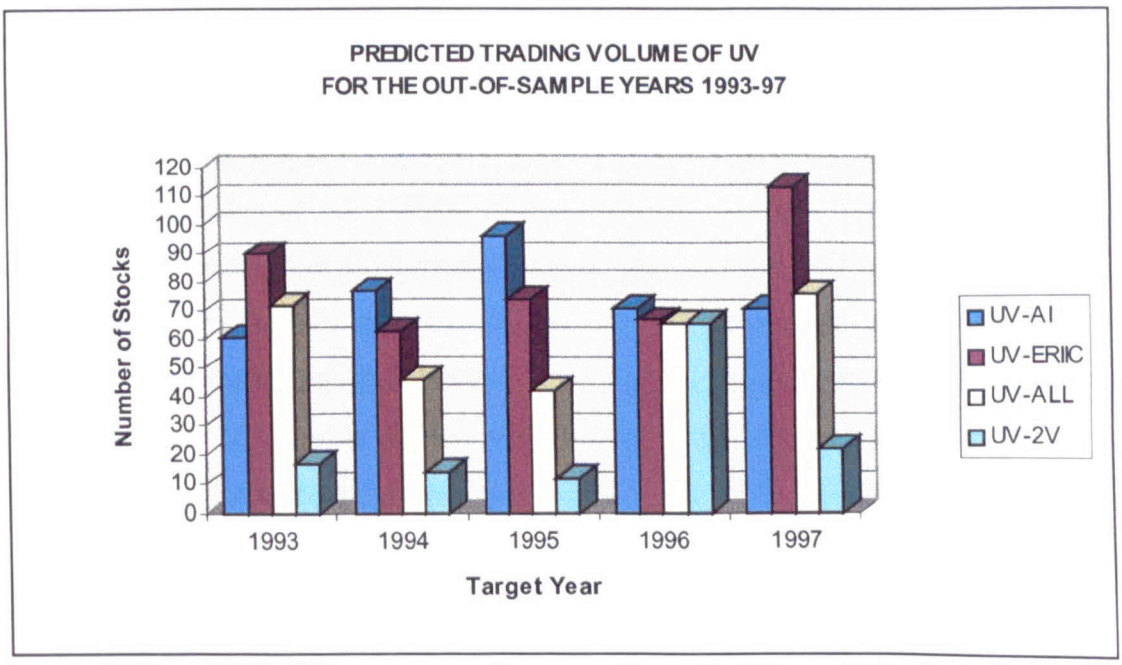


Figure 7.63: Out-of-sample trading volume of UV for 1993-97 performing four different implementations to predict high and low performing shares

CHAPTER 8: STOCK RETURN PREDICTABILITY IN UK INDUSTRIAL SECTORS

In this Chapter, we examine the potential of identifying outperforming shares in homogeneous U.K. industrial sectors by combining our five classification methods namely, LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes and using accounting information of around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. We restrict the sample to companies drawn from homogeneous industries for the following reasons: first, a more homogeneous sample might increase the classification performance of those classification methods that are sensitive to outliers and require a normally distributed input space; second, if shares of individual sectors are found to be more profitable than shares of other sectors in terms of abnormal returns, then we might be able to isolate those sectors that are more profitable and therefore to increase the profitability of our classification methods; third, national statistical figures concerning the performance of U.K. industrial sectors suggest that there are some important trends that took place in the U.K. economy in recent years. For example, one of the most important trends is the continuing shift of activity to services. The service sector accounted for around 60% of the U.K. output in 1998 compared to 50% in 1950. On the other hand, the manufacturing sector accounted for less than 25% of the U.K. output in 1998 compared to 33% in 1950. Another important trend concerns the profitability of U.K. industrial sectors. The net rate of return on capital (NROC) for service companies was 14.0% in 1998 compared to 14.5% in 1990. On the other hand, the NROC for manufacturing companies was 11.0% in 1998 compared to 7.0% in 1990 (Warton, 1999). Restricting the sample to homogeneous industries, we might be able to identify if these trends affect the share returns in our sample. The companies that have been included in our sample come from six separate industrial sectors namely services (S), manufacturing (M), property (P), utility (U), extractive (E), and financial (F).

This paper is organised as follows: In Section 8.1, we discuss the data and trading rules that we used in our study. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1994-97. Our predictor variables are 38 accounting ratios drawn from published accounting statements.

In Section 8.2, we discuss the methodology that we used to implement our classification methods. To make the models more robust and reduce the possibility of overfitting, we applied data preprocessing techniques. We performed two separate implementations: In the first implementation, we compared the LDA and the PNN in terms of classification accuracy and profitability for the out-of-sample years 1994-97 using information from either individual industrial sectors or combinations of different sectors. In the second implementation, we compared the five classification methods namely, LDA, PNN, LVQ, OC1 and RRI and the two voting methodologies namely, MV and UV in terms of classification accuracy and profitability for the out-of-sample years 1994-97 using information from S companies only.

In Section 8.3, we report the results of experimentation and we discuss the economic implications of our findings. After applying the LDA and the PNN, we found that both classifiers produce consistent excess returns if we restrict the sample to S companies, whereas additional benefits may arise after adding U, F, and P companies. On the other hand, the high performing portfolios that result if we restrict the sample to M and E companies are not particularly profitable and fail to produce consistent excess returns. After restricting the sample to S companies and combining five classification methods through MV and UV schemes, we found that the UV principle produces significant improvements in classification and profitability if compared to the individual classification methods and reduces substantially the trading volume.

In the final Section of the paper, we discuss the practical implications of our study and we also discuss the possibilities for further improvements in our methodology.

8.1 DATA AND TRADING RULES

In this experiment, we are interested in whether a particular share will be classified as H or L excess return share based on accounting information only. To perform this experiment, we used the same share returns data and the same accounting indicators that we used in the experiment described in Chapter 7. We recall that our share returns data are total returns on all shares traded on the London Stock Exchange in the years 1991-97. On the other hand, our accounting data are 38 accounting ratios drawn from published accounting statements. In Chapter 7, we mentioned that the incorporation of previous years total share and index returns resulted in a small reduction in the number of companies that we used to perform the experiments described in Chapters 5 and 6 because a few shares were traded more recently on the London Stock Exchange and therefore we were not able to calculate the previous years total share and index returns for these shares. To perform this experiment, we decided to use the same target data and the same accounting indicators that we used in Chapter 7 despite the fact that we have not incorporated any previous years total share and index returns. We left this possibility open for a future implementation and therefore it was considered more economical and practical to perform this implementation once.

A detailed list of the accounting variables that we used for this experiment is given in Table 8.1. As we can see, these variables are exactly the same with the variables presented in Table 5.1. Based on these variables, we aimed to find rules that classify a particular share as H or L performing share using three previous years of data to predict the next year. For example, to predict relative excess returns for 1994, we first trained the classification methods on the three preceding years 1991, 1992, 1993 and we tested them on the data available on 1994. We then moved the implementation one year ahead and we used information available from the years 1992, 1993, 1994 to predict the target year 1995 and so on. This implementation is slightly different from implementations in the previous experiments in which we used two years of data to predict the next year. We have not attempted the same implementation in this experiment because after restricting the sample to homogeneous industrial sectors resulted in a smaller

sample size. Therefore, we used three years of data to predict relative excess returns for the next year in order to avoid a potential problem of overfitting.

The companies that we have included in our sample come from six separate broad industrial sectors namely services (S), manufacturing (M), property (P), utility (U), extractive (E), and financial (F). The S sector comprises three main categories of companies: The first category covers wholesale and retail trade such as food and drug retailers, home entertainment, gaming, discount stores, hotels, pubs and restaurants, tourism, leisure facilities, and personal services such as hairdressers, laundries, and cleaning. These are mainly small firms who sell services to final consumers. The second category covers the activities of real estate, renting of vehicles and other machinery and equipment, business support, computer services, education and training, software houses, computer consultancy, supply of computer systems, hospital management, research and development, legal services, accountancy, advertising and cleaning, and media agencies. Companies in this category vary in size and provide their services to other companies rather than the final customers. The third category covers airlines and airports, road and rail transport, passenger and freight transport, postal activities and telecommunications including mobile phone networks, media equipment, and supplies. Most of the companies in this category are some large, capital intensive firms which used to be public corporations prior their privatisation.

The M sector comprises four main categories of manufactures. The first category covers food producers and processors, manufacturers of soft drinks, and tobacco. The second category covers manufacturers of chemical products including pharmaceuticals. The third category covers manufacturers of electrical and optical equipment, manufactures of electronic equipment and computer hardware, and manufacturers of telecommunications equipment such as computers, faxes, mobile telephones, and videos. The fourth category covers manufacturers of textiles and leather goods, paper and printing, plastic and rubber products, metals and cutlery, machinery equipment, distillers and vintners, transport equipment, commercial vehicles, defence systems, furnitures, toys and games, engineering, engineering contractors, and engineering fabricators.

The U sector covers electricity supply, gas distribution and water supply. The P sector covers builders merchants, construction and building materials, property agencies, house building, and real estate development. The E sector covers extracting industries, mining finance, oil and gas, exploration and production, oil services, and other mining. The F sector covers investment banks, asset managers, insurance brokers, consumer finance, and other financial companies.

Our data set consists of 1641 S, 1752 M, 668 P, 63 U, 87 E, and 94 F companies. Obviously, the M and S companies represent the largest proportion in our data, whereas the other categories represent a smaller proportion. National statistical figures concerning the performance of U.K. industrial sectors suggest that the S sector accounted for around 60% of the U.K. output in 1998 compared to 50% in 1950. On the other hand, the M sector accounted for less than 25% of the U.K. output in 1998 compared to 33% in 1950. Another important trend concerns the profitability of U.K. industrial sectors (Warton, 1999). The NROC for the S sector was 14.0% in 1998 compared to 14.5% in 1990. On the other hand, the NROC for the M sector was 11.0% in 1998 compared to 7.0% in 1990. The annual rates of return for the M and the S sectors for the U.K. economy are illustrated in Figure 8.1. As we can see, the profitability of the S sector has remained stable. The worst year for the S sector was in 1992 as the NROC reached 12.4% compared to 13.8% in 1991. The profitability of the S sector increased to 14.4% in 1994 and remained relatively stable over the next three years reaching 14.3% and 15.0% in 1996 and 1997, respectively. The S companies that were particularly profitable in 1996 and 1997 were those specialising in transport and communications, whereas companies specialising in real estate, renting, and consultancy also showed substantial growth in earnings during the same period. However, the same growth was not sustained in 1998 because only companies specialising in leisure industry showed substantial growth in profits that year. The main reason that affected the growth in profits for S companies in 1998 was new international competition motivated by the use of information technology as well as the use of new pricing strategies (Warton, 1999).

The profitability of the M sector experienced more wild fluctuations than the S sector. As we can see, the NROC of the M sector reached 4.3% in 1991 compared to 7.0% in 1990. Over the next years, the M sector improved the NROC reaching 11.0% in 1998. Manufacturers that were particularly profitable in 1997 and 1998 were those producing electrical and optical equipment as well as chemical products. The national statistical figures also suggest that the profitability of M companies have been accompanied by productivity gains. The output of the M sector fall in most of 1997 and 1998, whereas investments in information technology facilitated the productivity at the same time (Warton, 1999).

The strength of sterling in 1996-98 had more adverse effects in the M sector rather than the S sector because M companies export a larger proportion of their output than S companies. Increased import competition in low-value consumer goods as well as steel, plastics, and textiles reduced import prices and squeezed the profits on home sales. Exporters focused on low cost strategies in order to be competitive and retain overseas markets. On the other hand,

exchange rate movements put pressure on export prices that fall in the last two years (Warton, 1999).

National statistical figures also suggest that the profitability of both S and M sectors hasbeen affected by the average growth in earnings. In the middle of 1997, the earnings of the M sector exceeded those of the S sector by 0.2 percentage points, whereas the earnings of the S sector have been growing more strongly than the earnings of the M sector since the beginning of 1998. In addition, we should consider that profits of companies are closely related to investment through retained earnings. The S companies were more capital intensive since 1990 and this might be an explanation of the pattern in profitability for this sector. The building of cable television and mobile phone networks, investments in new technology such as computer equipment and computer software, and large scale investments in retail, telecommunications, catering, transport, and rental have built a new structure for the S industry. As a result, the net average capital employed increased by 64% (187 billion pounds at current prices) between 1990 and 1998. National statistical figures also suggest that the S companies accounted for 15% average annual growth in the private S sector from 1994 to 1998, 15% growth in investment in the wholesale and retail sectors of the economy in 1998, and 70% of total business investment in 1999. On the other hand, the growth of capital investment for the M sector was not the same as the S sector because of competitive trading conditions and high costs of capital. The net average capital employed for the M sector increased only 16% (36 billion pounds at current prices) between 1990 and 1998, whereas the investment of this sector increased only to 8% from 1995 to 1998 (Warton, 1999).

Another trend concerning the S and M sectors is the expansion of employment for the S sector with a respective decline of employment for the M sector. Obviously, this affected the S companies costs (Warton, 1999).

8.2 METHODOLOGY

To make the classification methods more robust and smooth the effect of outliers, we applied the same data preprocessing techniques that we used in the experiments described in the previous Chapters. We recall that these data preprocessing techniques include data winsorisation, data normalisation, and triangular transformations of the variables. In this experiment, however, we have not repeated the stepwise variable elimination procedures that we applied in all previous experiments, but we have used instead the optimal subsets of accounting variables that we found in the experiment described in Chapter 7. Although we consider that this might not be the best optimal procedure, we do not expect that it may affect

the performance of the classifiers in a substantial degree since we use the same data. Therefore, the list of the variables that were applied for the implementation of the classifiers is presented in Table 8.2.

We performed two separate implementations: in the first implementation, we compared the LDA and the PNN in terms of classification accuracy and profitability for the out-of-sample years 1994-97. We implemented the two classifiers using five different sets of information: first, using information from all industrial sectors at the same time; second, using information from M+E companies at the same time; third, using information from S+F+U companies at the same time; fourth, using information from S+F+U+P companies at the same time; and fifth, using information from S companies only. In the second implementation, we compared the five classification methods namely, LDA, PNN, LVQ, OC1 and RRI and the two voting methodologies namely, MV and UV in terms of classification accuracy and profitability for the out-of-sample years 1994-97 using information from S companies only.

To implement the five classifiers namely, LDA, PNN, LVQ, RRI, and OC1 and the two voting methodologies namely MV and UV, we followed implementation strategies similar to those that we described in Chapter 7. We compared and contrasted the five classifiers and the two voting methodologies in terms of classification accuracy, profitability, and trading volume.To evaluate the profitability of the five classifiers and the two voting methodologies, we calculated average returns and excess returns over the index for the portfolios of actual H and L shares in our data in all the 12-month holding periods starting each year, and then we compared them with the respective averages for the portfolios of H and L shares predicted by the classification methods. On the other hand, to examine if transaction costs can have an important impact in our trading system, we also compared the classification methods and the two voting methodologies for the predicted number of shares included in the portfolios of H performing shares that are traded in the target years 1994-97.

8.3 RESULTS

In this section, we summarise the results of experimentation. We performed two separate implementations: in the first implementation, we compared the LDA and the PNN in terms of classification accuracy and profitability for the out-of-sample years 1994-97. We implemented the two classifiers using five different sets of information: first, using information from all industrial sectors (all sectors) at the same time; second, using information from manufacturing and extractive companies (M+E) at the same time; third, using information from service,

financial, utility, and property companies (S+F+U+P) at the same time; fourth, using information from service, financial, and utility companies (S+F+U) at the same time; and fifth, using information from service companies (S) only. In the second implementation, we compared the five classification methods (LDA, PNN, LVQ, OC1 and RRI) and the two voting methodologies (MV and UV) in terms of classification accuracy and profitability for the out-of-sample years 1994-97 using information from S companies only. We recall that we have not performed any stepwise variable elimination procedure for the target year 1994 and therefore we will consider all the target years 1994-97 as genuine out-of-sample years.

Tables 8.3 and 8.4 show the classification performance of LDA and PNN, respectively, for the out-of-sample years 1994-97 using the five different sets of information. These results are also presented in Figures 8.2 and 8.3 for LDA and PNN, respectively. As we can see in Figure 8.2, the classification performance of LDA reaches the highest peaks using information from M+E companies at the same time for the target years 1994 and 1996, whereas it reaches the lowest levels using information from the same group of companies for the target years 1995 and 1997. On the other hand, the classification performance of the model reaches the highest peak using information from S companies only for the target year 1995, whereas it reaches the highest peak using information from S+U+F+P companies at the same time for the target year 1997. The classification results in Figure 8.2 suggest that the classification performance of LDA is on average more consistent after using information from S companies only, whereas it exhibits more wild fluctuations after using information from the other groups of companies.

The classification performance of the PNN seems to follow a pattern similar to LDA. As we can see in Figure 8.3, the PNN favours the use of information from M+E companies for the target years 1994 and 1996, whereas the classification performance of the model deteriorates substantially using information from the same group of companies for the target years 1995 and 1997. On the other hand, the classification performance of the model reaches the highest peak using information from S+F+U companies at the same time for the target year 1995, whereas it reaches the highest peak using information from S companies only for the target year 1997. On average, the classification performance of the PNN seems to be more consistent after using either information from S+F+U companies at the same time or information from S companies only for the target years 1994-97, whereas it exhibits more wild fluctuations after using information from M+E companies.

Although the classification performance is a very important factor to evaluate a particular classifier, it is not the primary concern for this particular application. The ultimate purpose of our trading system is profitability. We therefore compared the average returns and excess

returns over the index of the portfolios of actual H and L shares in our data in all the 12-month holding periods starting each year, with the respective average returns and excess returns of the portfolios of H and L shares predicted by LDA and PNN under the five different types of information.

Table 8.5 compares the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the LDA. These results are also presented in Figures 8.4-8.7 for H returns and excess returns and for L returns and excess returns, respectively. As we can see in Figures 8.4-8.7, the financial returns of LDA are better after using information from M+E companies for the target year 1994, whereas this pattern is not consistent for the following target years as the financial returns of LDA deteriorate substantially after using information from M+E companies only. The LDA produces greater financial returns after using information from S companies only for the target year 1995, whereas the model produces greater financial returns using either information from S companies only or information from S+U+F+P companies for the target years 1996 and 1997, respectively.

Table 8.6 compares the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the PNN. These results are also presented in Figures 8.8-8.11 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, the PNN follows a more consistent pattern than LDA. The model produces greater returns using information from S companies only for the target years 1994-97, whereas it also produces favourable results using information from S+U+F+P and S+U+F companies at the same time during the same period. On the other hand, the financial returns predicted by the model deteriorate after using information from M+E companies for the target years 1994-97.

Overall, both the classification results and the predicted financial returns suggest that both PNN and LDA seem to prefer the use of information from S companies only, whereas the models also produce favourable results after using information from S+U+F and S+F+U+P companies at the same time. On the other hand, the information extracted from M+E companies does not seem to benefit the models on a consistent basis.

We recall that according to our trading system, if a share is classified as H, we buy equal amounts of this share at the end of the reporting month and we hold it for one year. The

profitability of each classification method is therefore calculated by the cumulative profits generated by the resulting portfolio of H performing shares only. The benefit of this approach is that it minimises transaction costs while it is not affected by price fluctuations around the reporting date. Given that there are 157-180 H shares each year, each share is traded at most once per year and trades can be done at the end of the month in a basket of no more than 13-16 shares bought and sold in the ideal trading strategy. On average, the H performing portfolio will turn over 1/12 of its constituents each month. Therefore, the transaction costs are not expected to reduce the profitability of the UV methodology as well as the profitability of the other classification methods by more than 2% per year on average. Given that there are substantial benefits in the trading volume from using a sample of companies as small as possible we compared the five classification methods (LDA, PNN, LVQ, OC1 and RRI) and the two voting methodologies (MV and UV) in terms of classification accuracy and profitability for the out-of-sample years 1994-97 after using S companies only.

Table 8.7 compares the five classification methods and the two voting methodologies in terms of classification accuracy for the target years 1994-97 using information from S companies only. These results are also illustrated in Figure 8.12. As we can see, the UV methodology outperforms significantly the other classification methods for the target years 1994-97. LVQ and OC1 also produce very good results for the target year 1994, whereas OC1 and MV are the second best classifiers for the target year 1995. The classification accuracy of all classifiers deteriorates slightly for the next target years 1996 and 1997 but it is still satisfactory. The PNN is the second best classifier for the target year 1996, whereas the OC1 classifier is the second best classifier for the target year 1997.

Table 8.8 compares the financial returns and excess returns over the index of the portfolios of actual H and L shares in all the 12-month holding periods starting in each year, with the financial returns and excess returns of the portfolios of H and L shares predicted by the classification methods. These results are also presented in Figures 8.13-8.16 for H returns and excess returns and for L returns and excess returns, respectively. As we can see, all the classifiers and voting methodologies produce positive returns and excess returns. However, the UV methodology outperforms significantly the other classifiers for the target years 1994-97 and produces the highest financial results. The financial returns for all classification methods are relatively low for the target year 1994 but they are still positive and significantly greater than zero. On the other hand, the financial returns for all classification methods increase substantially for the target year 1995 and they remain impressive for the last target years 1996 and 1997. It is quite clear from the graphs that that the UV methodology outperforms significantly the other classification methods for the target years 1995 and 1997 and it clearly outperforms the other

classification methods for the target years 1994 and 1996. In addition, we have to notice that there are only minor differences among the other classification methods in terms of predicted financial returns.

Although the financial returns are a primary factor to evaluate a particular trading system, we should also examine the transaction costs involved in trading the number of shares predicted by the classification methods. Therefore, we calculated the predicted number of shares included in the portfolios of H performing shares that are traded for the out-of-sample years 1994-97. The results from this comparison are illustrated in Figure 8.17. The results suggest that UV outperforms significantly the other classifiers in terms of trading volume for the target years 1994-97. However, we should emphasise that according to our trading system, if a share is classified as H, we buy equal amounts of this share at the end of the reporting month and we hold it for one year. The profitability of each classification method is therefore calculated by the cumulative profits generated by the resulting portfolio of H performing shares only. The benefit of this approach is that it minimises transaction costs and it is not affected by price fluctuations around the reporting date. Given that there are 60-70 H shares each year, each share is traded at most once per year and trades can be done at the end of the month in a basket of no more than 7-13 shares bought and sold in the ideal trading strategy. On average, the H performing portfolio will turn over 1/12 of its constituents each month. Therefore, the transaction costs are not expected to affect the UV methodology as well as the other classification methods by more than 1% per year on average if we restrict the sample to S companies.

If we compare the results of the UV methodology presented in Chapter 7 (Tables 7.6 and 7.9) with the results of the UV methodology presented in Chapter 8 (Table 8.8 and Figure 8.17), we can see that restricting the sample to homogeneous industrial sectors is highly beneficial in terms of profitability and trading volume. Indeed, this comparison reveals that after restricting the sample to homogeneous industrial sectors the calculated average return, excess return, and trading volume (36.8%, 24.5%, and 45 shares, respectively - Table 8.8 and Figure 8.17), for the UV methodology over the 1994-97 period, have been improved substantially over the corresponding results of the UV methodology after including all industrial sectors in the sample (29.7%, 16.6%, and 79 shares for average returns, excess returns, and trading volume, respectively - Tables 7.6 and 7.9) for the same period.

8.4 SUMMARY AND CONCLUSIONS

In this Chapter, we examined the potential of identifying outperforming shares in homogeneous U.K. industrial sectors by combining five heterogeneous statistical classification algorithms

namely, LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes and using accounting information of around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. Our target data were total returns on all shares traded on the London Stock Exchange in the years 1994-97. Our input variables were 38 accounting ratios drawn from published accounting statements.

After applying the LDA and the PNN, we found that both classifiers produce consistent excess returns after restricting the sample to service companies, whereas additional benefits arise after adding utility, financial and property companies. On the other hand, the high performing portfolios that result after restricting the sample to manufacturing and extractive companies are not particularly profitable and fail to produce consistent excess returns. After implementing the five classifiers using service companies only, we found that all classification methods produce consistent excess returns. However, greater gains result from UV where a share is not classified as H performing share unless all classifiers agree. The UV principle not only produces significantly greater returns than the other methods, but it also results in substantial reductions in the number of shares traded.

Our work provides substantial evidence for the ability of statistical classification methods to identify high performing shares if the sample is homogeneous. There are three main benefits that result for our trading system after restricting the sample to homogeneous industrial sectors: first, less data are required for the implementation of the models and therefore less time and effort are required to optimise the classification methods; second, the trading volume is reduced substantially and this results in more efficient trading strategies; and third, the transaction costs are minimised because less shares are traded for each particular year.

One potential explanation of the apparent profitability of the five classification methods and voting methodologies after restricting our sample to companies drawn from homogeneous industries is that there are some important trends that took place in the U.K. economy in recent years. For example, one of the most important trends is the continuing shift of activity to services. The service sector accounted for around 60% of the U.K. output in 1998 compared to 50% in 1950. On the other hand, the manufacturing sector accounted for less than 25% of the U.K. output in 1998 compared to 33% in 1950. Another important trend concerns the profitability of U.K. industrial sectors. The net rate of return on capital (NROC) for service companies was 14.0% in 1998 compared to 14.5% in 1990. On the other hand, the NROC for manufacturing companies was 11.0% in 1998 compared to 7.0% in 1990 (Warton, 1999). It is highly likely that these trends were reflected in the stock prices and consequently in the returns of the shares that we included in our sample.

We have to repeat, however, that our classification methods should be treated with caution due to their large number of parameters. To avoid the possibility of overfitting, we applied sophisticated data pre-processing techniques in order to eliminate the effect of outliers and increase the robustness of the models. In the next Chapter, we investigate a new data preprocessing technique related to dimensionality reduction. More specifically, we investigate the applicability of dimensionality reduction techniques based on neural networks as an alternative to ad hoc techniques to select the best subset of variables for classifiers such as LVQ, PNN, OC1, and RRI. The details of this methodology are provided in the next Chapter.

Return on Capital	PBT/TA, PBT/TCE, NI/TCE, CF/TA, CF/TCE
Profitability	PBT/SR, PAT/SR, NI/SR, CF/SR, PAT/EQ, CF/MKBD
Financial Leverage	DEBT/EQ, DEBT/TCE, DEBT/TA, TL/EQ, TA/EQ, BA/MKBD
Investment	P/E, DY, EY, BE/ME
Growth (%)	TA, PAT, PBT, EPS, MKBD, SR
Short-Term Liquidity	CA/CL, CL/TA, CL/EQ
Return on Investment	NI/TA, PAT/TA
Efficiency	SR/TA, DRS/SR
Risk	PBT/CL, PAT/CL, NI/CL, CF/CL

PBT: Profit Before Taxes; TA: Total Assets; TCE: Total Capital Employed; CF: Cash Flow; PAT: Profit after Taxes; SR: Sales Revenue; NI: Net Income; EQ: Shareholders' Equity; MKBD: Market Capitalisation at Balance Sheet Date; DEBT: Debt; TL: Total Liabilities; BA: Book Assets; P/E: Price/Earnings Ratio; EY: Earnings Yield; DY: Dividend Yield; BE: Book Equity; ME: Market Equity; EPS: Earnings Per Share; CA: Current Assets; CL: Current Liabilities; DRS: Debtors.

Table 8.1: Initial list the accounting variables that we collected to predict high and low performing shares in different industrial sectors

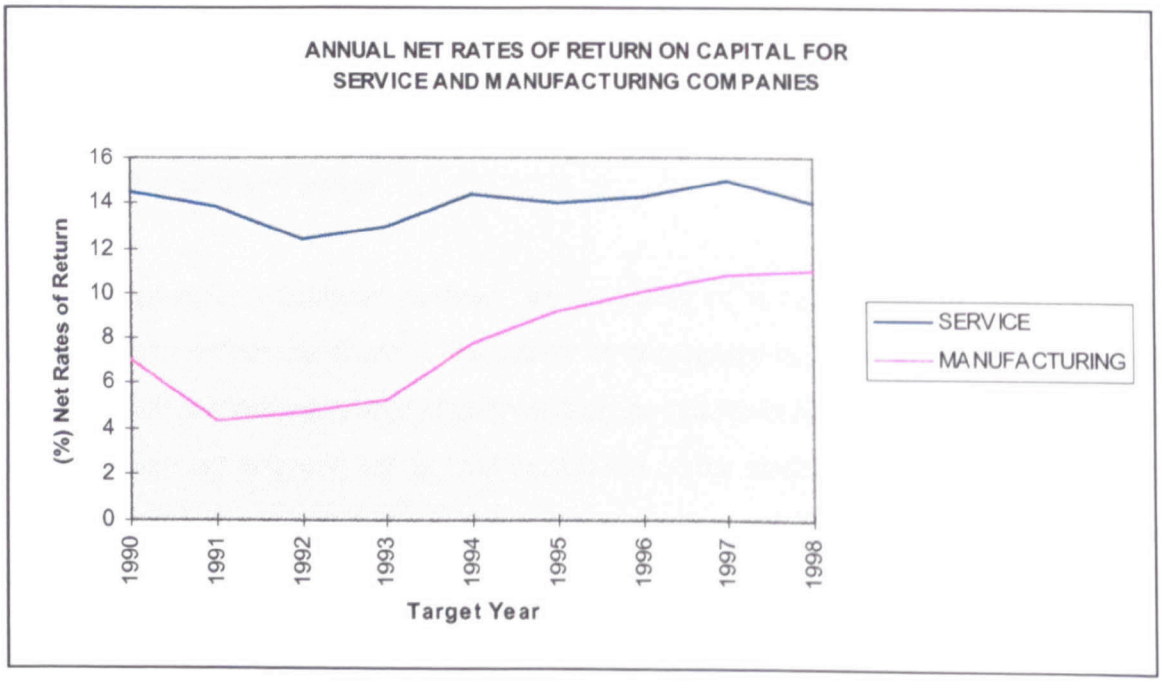


Figure 8.1: Annual net rates of return on capital for service and manufacturing companies

LDA	PBT/SR, SR/TA, DEBT/EQ, PAT/SR, EPS (%), SR (%), PAT/TA, PBT(%), NI/TCE, PBT/TCE, TL/EQ, CF/SR, DY, CA/CL, CL/TA, PBT/CL, PAT/CL, NI/CL, CF/CL, BE/ME, TA/MKBD
PNN	DEBT/TA, PAT/EQ, SR/TA, DRS/SR, PAT/TA, TA (%), MKBD (%), CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PAT/CL, NI/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD
LVQ	SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, EPS (%), SR (%), MKBD (%), PBT (%), NI/TCE, DEBT/TCE, CF/TA, CF/TCE, CA/CL, CL/EQ, CF/CL, BE/ME, TA/MKBD, CF/MKBD
OC1	DEBT/TA, PBT/SR, TA/EQ, NI/SR, DEBT/EQ, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), PAT (%), P/E, NI/TCE, CF/TA, CF/TCE, CF/SR, CL/TA, CL/EQ, NI/CL, TA/MKBD, CF/MKBD
RRI	DEBT/TA, PAT/EQ, PBT/TA, PBT/SR, SR/TA, TA/EQ, NI/SR, DEBT/EQ, PAT/SR, NI/TA, EPS (%), SR (%), PAT/TA, TA (%), MKBD (%), PAT (%), PBT (%), P/E, NI/TCE, PBT/TCE, DEBT/TCE, CF/TA, CF/TCE, CF/SR, EY, DY, CA/CL, CL/TA, CL/EQ, PBT/CL, PAT/CL, CF/CL, BE/ME, TA/MKBD, CF/MKBD

Table 8.2: Subsets of accounting variables that we finally selected to predict high and low performing shares in different industrial sectors after applying stepwise variable elimination procedures

LDA		All Sectors			M+E			S+F+U+P			S+F+U			S	
Actual Class		Predicted Class Membership													
	Patterns			Patterns			Patterns			Patterns			Patterns		
1994		H	L		H	L		H	L		H	L		H	L
H	155	83	72	66	35	31	89	48	41	66	34	32	59	31	28
L	463	193	270	196	71	125	267	133	134	196	93	103	179	83	96
Overall (%)		57.12 %			61.07 %			51.12 %			52.29 %			53.36 %	
1995		H	L		H	L		H	L		H	L		H	L
H	160	90	70	68	33	35	93	59	34	69	43	26	62	38	24
L	479	183	296	201	95	106	277	107	170	204	72	132	186	63	123
Overall (%)		60.41 %			51.67 %			61.89 %			64.10 %			64.92 %	
1996		H	L		H	L		H	L		H	L		H	L
H	171	94	77	72	41	31	99	58	41	73	43	30	65	36	29
L	510	195	315	215	79	136	295	124	171	218	97	121	196	76	120
Overall (%)		60.06 %			61.67 %			58.12 %			56.36 %			59.77 %	
1997		H	L		H	L		H	L		H	L		H	L
H	180	106	74	76	35	41	104	71	33	77	48	29	68	44	24
L	538	211	327	228	115	113	310	113	197	231	93	138	207	80	127
Overall (%)		60.31 %			48.68 %			64.73 %			60.39 %			62.18 %	

Table 8.3: Out-of-sample classification performance of LDA for 1994-97
using accounting information from different industrial sectors to predict high and low performing shares

PNN		All Sectors			M+E			S+F+U+P			S+F+U			S	
Actual Class		Predicted Class Membership													
	Patterns			Patterns			Patterns			Patterns			Patterns		
1994		H	L		H	L		H	L		H	L		H	L
H	155	84	71	66	39	27	89	53	36	66	37	29	59	33	26
L	463	183	280	196	70	126	267	122	145	196	75	121	179	74	105
Overall (%)		58.90 %			62.98 %			55.62 %			60.31 %			57.98 %	
1995		H	L		H	L		H	L		H	L		H	L
H	160	90	70	68	35	33	93	59	34	69	50	19	62	40	22
L	479	186	293	201	94	107	277	95	182	204	69	135	186	61	125
Overall (%)		59.94 %			52.79 %			65.14 %			67.77 %			66.53 %	
1996		H	L		H	L		H	L		H	L		H	L
H	171	93	78	72	44	28	99	58	41	73	44	29	65	39	26
L	510	189	321	215	74	141	295	121	174	218	81	137	196	71	125
Overall (%)		60.79 %			64.46 %			58.88 %			62.20 %			62.84 %	
1997		H	L		H	L		H	L		H	L		H	L
H	180	106	74	76	41	35	104	68	36	77	50	27	68	44	24
L	538	209	329	228	109	119	310	125	185	231	92	139	207	76	131
Overall (%)		60.58 %			52.63 %			61.11 %			61.36 %			63.64 %	

Table 8.4: Out-of-sample classification performance of PNN for 1994-97
using accounting information from different industrial sectors to predict high and low performing shares

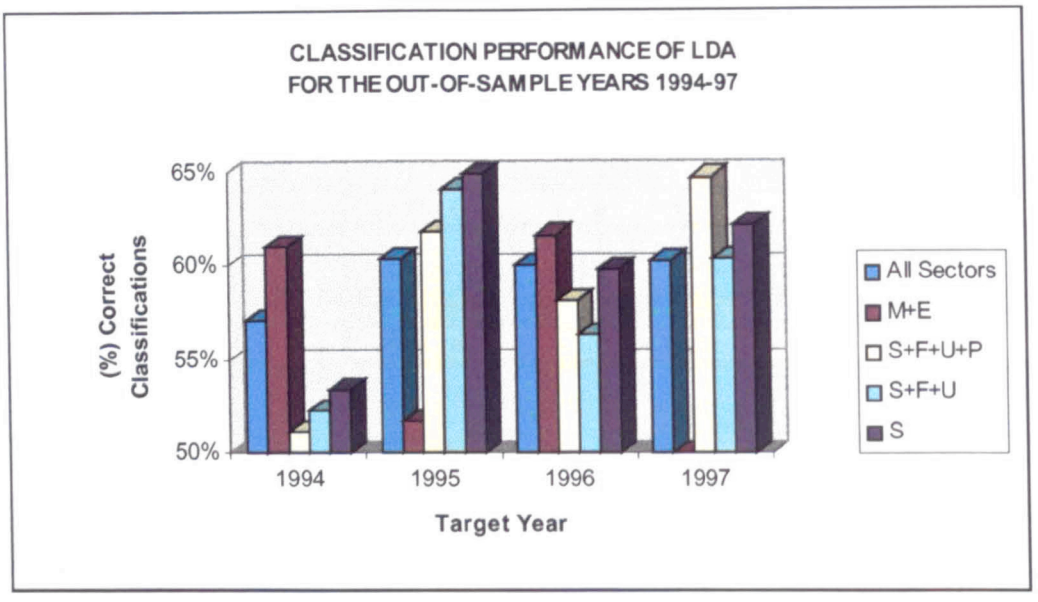


Figure 8.2: Out-of-sample classification performance of LDA for 1994-97 using accounting information from different industrial sectors to predict high and low performing shares

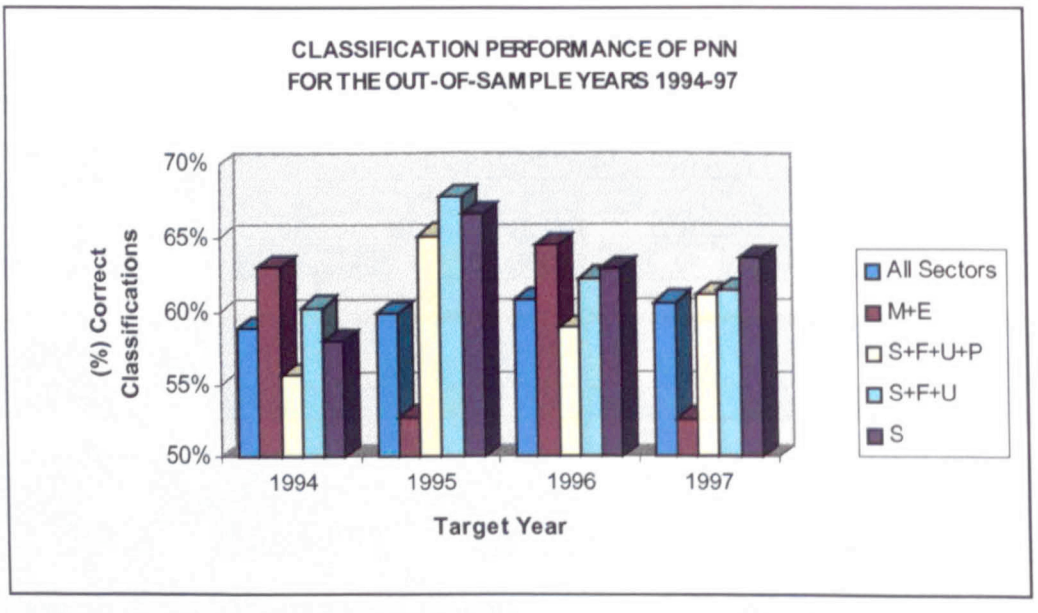
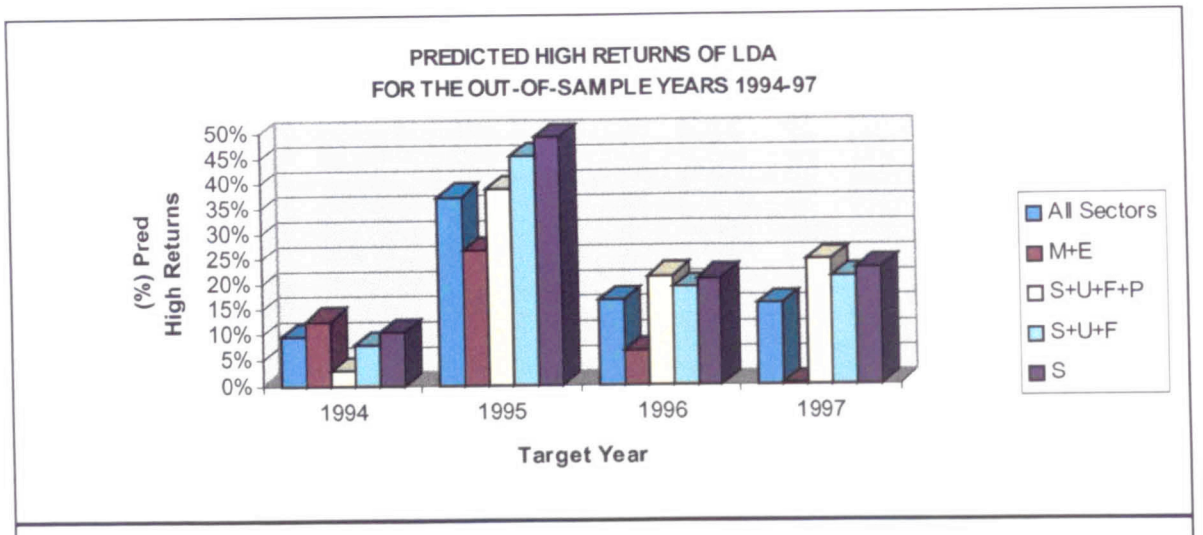
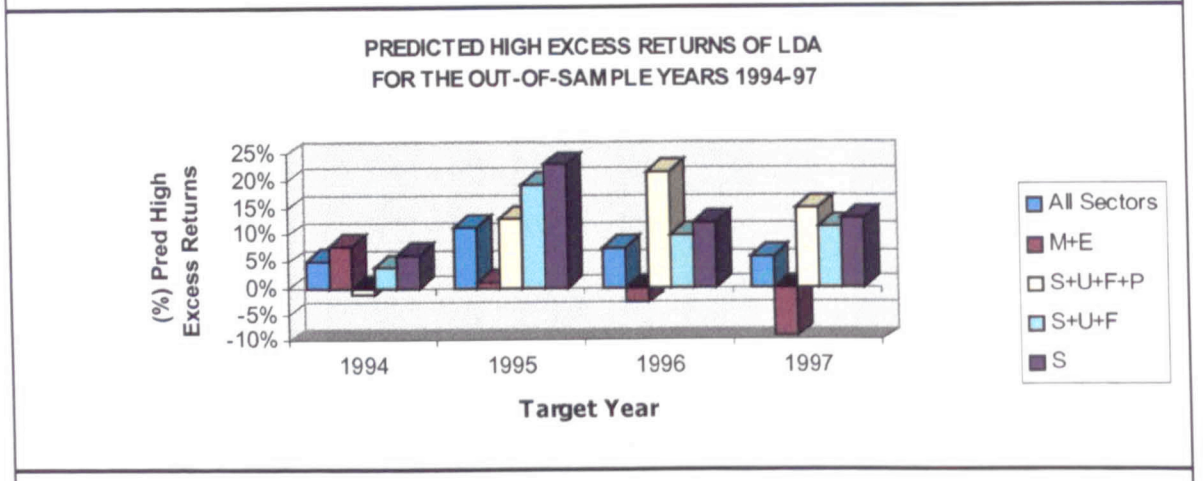


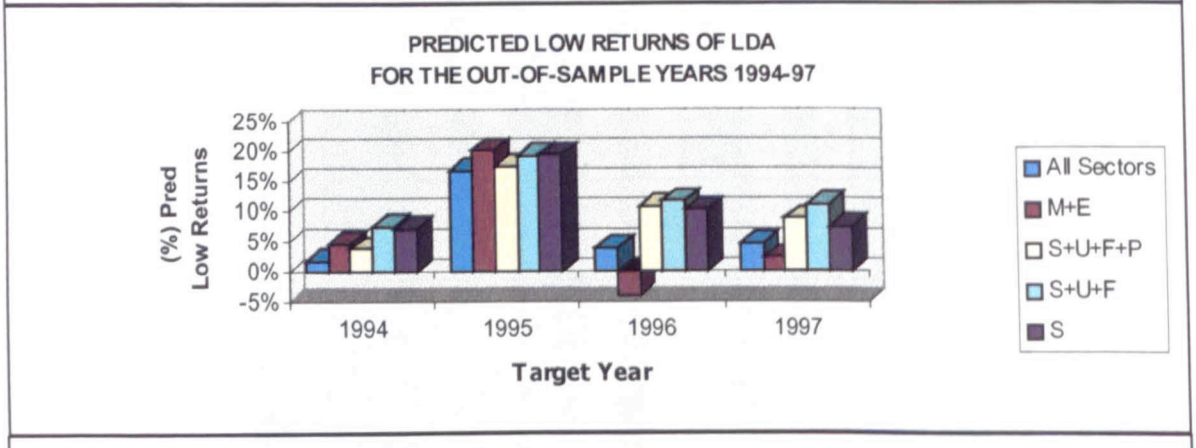
Figure 8.3: Out-of-sample classification performance of PNN for 1994-97 using accounting information from different industrial sectors to predict high and low performing shares

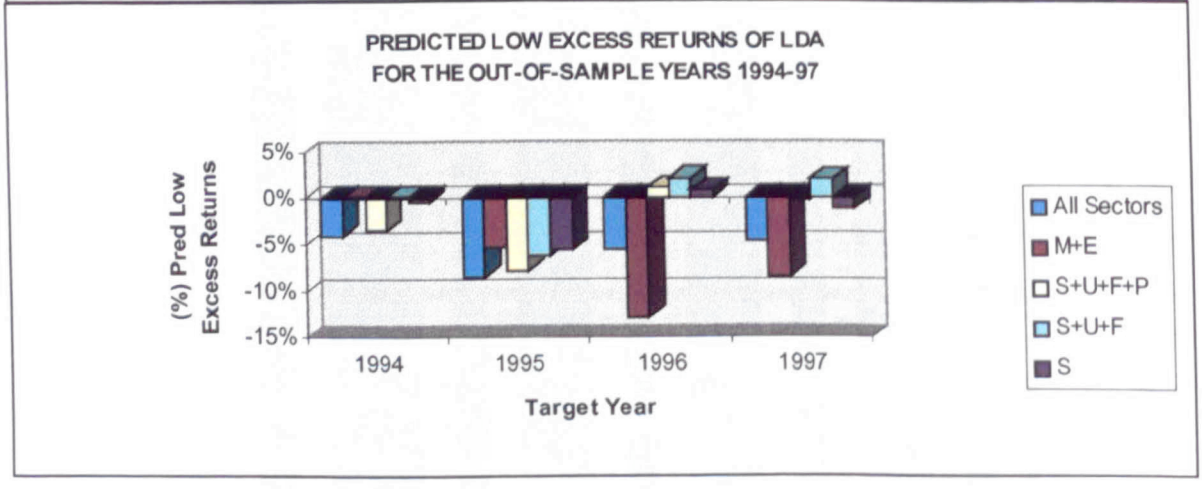
LDA	All Sectors			M+E				S+U+F+P				S+U+F				S			
1994		Pred Class				Pred Class				Pred Class				Pred Class				Pred Class	
Actual Return		H	L	Actual Return		H	L	Actual Return		H	L	Actual Return		H	L	Actual Return		H	L
H= 44.2 L= -7.7	5.5	9.6	1.9	H= 44.0 L= -4.2	4.5	12.8	4.6	H= 44.1 L= -10.1	6.0	3.1	3.8	H= 46.8 L= -5.4	5.0	7.9	7.5	H= 49.5 L= -4.7	5.0	10.5	7.1
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 38.7 L= -12.9	5.2	5.1	-4.1	H= 39.5 L= -9.0	4.8	7.9	0.1	H= 38.1 L= -15.9	5.8	-1.2	-3.5	H= 41.8 L= -11.4	6.0	4.1	0.1	H= 44.5 L= -11.0	6.3	6.2	-0.4
1995																			
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 78.8 L= 7.6	25.1	37.2	16.7	H= 80.9 L= 3.9	25.3	26.7	20.2	H= 77.0 L= 10.2	25.5	38.7	17.4	H= 80.7 L= 13.1	25.3	45.2	19.3	H= 83.4 L= 14.1	25.4	49.0	19.4
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 53.7 L= -17.9	25.5	11.3	-8.4	H= 55.6 L= -21.7	25.6	1.2	-5.2	H= 51.5 L= -15.2	25.4	12.9	-7.7	H= 55.4 L= -12.6	25.7	19.3	-6.1	H= 58.0 L= -11.3	25.4	23.0	-5.6
1996																			
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 50.5 L= -4.4	9.1	16.8	3.9	H= 37.7 L= -11.8	9.3	6.9	-3.9	H= 58.8 L= 1.3	9.4	21.5	10.8	H= 61.3 L= -0.2	9.7	19.1	11.6	H= 62.6 L= -0.9	9.7	21.1	10.2
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 41.4 L= -13.9	9.5	7.6	-5.6	H= 28.4 L= -21.0	9.2	-2.9	-12.8	H= 49.4 L= -8.2	9.5	21.5	1.0	H= 51.6 L= -9.5	9.3	9.9	2.1	H= 52.9 L= -9.9	9.0	12.2	0.8
1997																			
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 58.8 L= -6.9	9.7	15.9	4.6	H= 51.3 L= -15.5	11.1	0.1	2.3	H= 62.6 L= 0.0	9.0	24.5	8.7	H= 64.8 L= -0.5	9.2	21.3	11.1	H= 65.0 L= -2.0	9.4	23.0	7.6
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 49.1 L= -16.4	9.5	5.9	-4.7	H= 40.2 L= -25.2	9.7	-9.2	-8.4	H= 53.6 L= -9.3	9.3	14.9	-0.2	H= 55.6 L= -10.1	9.6	11.5	2.0	H= 55.6 L= -11.5	9.5	12.9	-1.2

Table 8.5: Out-of-sample returns and excess returns of LDA for 1994-97
using accounting information from different industrial sectors to predict high and low performing shares





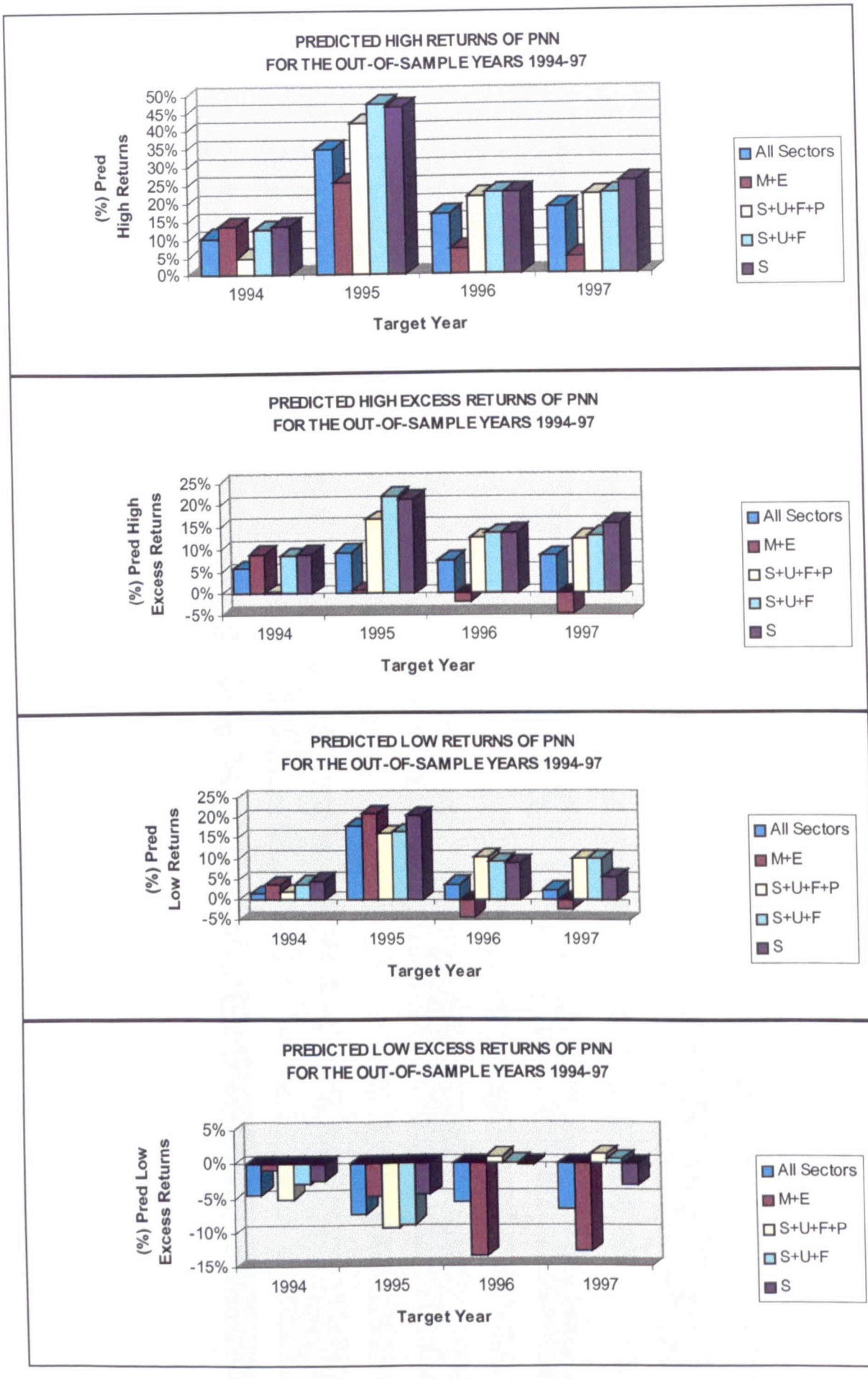




Figures 8.4-8.7: Out-of-sample returns and excess returns of LDA for 1994-97 using accounting information from different industrial sectors to predict high and low performing shares

PNN	All Sectors			M+E				S+U+F+P				S+U+F				S			
1994		Pred Class				Pred Class				Pred Class				Pred Class				Pred Class	
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 44.2 L= -7.7	5.5	10.5	1.4	H= 44.0 L= -4.2	4.5	13.7	3.8	H= 44.1 L= -10.1	6.0	5.0	1.9	H= 46.8 L= -5.4	5.0	12.8	3.9	H= 49.5 L= -4.7	5.0	13.7	4.7
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 38.7 L= -12.9	5.2	5.7	-4.3	H= 39.5 L= -9.0	4.8	8.9	-0.9	H= 38.1 L= -15.9	5.8	0.4	-5.0	H= 41.8 L= -11.4	6.0	8.5	-2.9	H= 44.5 L= -11.0	6.3	9.0	-2.3
1995				1995				1995				1995				1995			
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 78.8 L= 7.6	25.1	35.0	18.2	H= 80.9 L= 3.9	25.3	25.7	21.1	H= 77.0 L= 10.2	25.5	42.1	16.2	H= 80.7 L= 13.1	25.3	47.6	16.7	H= 83.4 L= 14.1	25.4	46.9	20.9
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 53.7 L= -17.9	25.5	9.4	-7.2	H= 55.6 L= -21.7	25.6	0.7	-4.7	H= 51.5 L= -15.2	25.4	16.6	-9.2	H= 55.4 L= -12.6	25.7	21.9	-8.8	H= 58.0 L= -11.3	25.4	21.2	-4.4
1996				1996				1996				1996				1996			
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 50.5 L= -4.4	9.1	17.1	3.9	H= 37.7 L= -11.8	9.3	7.3	-4.1	H= 58.8 L= 1.3	9.4	21.8	10.7	H= 61.3 L= -0.2	9.7	22.9	9.5	H= 62.6 L= -0.9	9.7	22.9	9.0
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 41.4 L= -13.9	9.5	7.7	-5.4	H= 28.4 L= -21.0	9.2	-1.8	-13.4	H= 49.4 L= -8.2	9.5	12.5	1.1	H= 51.6 L= -9.5	9.3	13.5	0.1	H= 52.9 L= -9.9	9.0	13.8	-0.2
1997				1997				1997				1997				1997			
Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L	Actual Return	Index	H	L
H= 58.8 L= -6.9	9.7	18.7	2.4	H= 51.3 L= -15.5	11.1	5.0	-2.4	H= 62.6 L= 0.02	9.0	21.9	10.3	H= 64.8 L= -0.5	9.2	22.5	10.1	H= 65.0 L= -2.0	9.4	26.0	5.7
Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L	Actual Excess Ret		H	L
H= 49.1 L= -16.4	9.5	8.5	-6.7	H= 40.2 L= -25.2	9.7	-4.8	-12.7	H= 53.6 L= -9.3	9.3	12.3	1.4	H= 55.6 L= -10.1	9.6	13.0	0.7	H= 55.6 L= -11.5	9.5	15.7	-3.1

Table 8.6: Out-of-sample returns and excess returns of PNN for 1994-97 using accounting information from different industrial sectors to predict high and low performing shares



Figures 8.8-8.11: Out-of-sample returns and excess returns of PNN for 1994-97 using accounting information from different industrial sectors to predict high and low performing shares

		LDA		PNN		LVQ		OC1		RRI		MV		UV	
Actual Class	Patterns	Predicted Class Membership													
1994		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	59	31	28	33	26	33	26	35	24	32	27	34	25	9	50
L	179	83	96	74	105	68	111	71	108	74	105	76	103	21	158
Overall (%)		53.36 %		57.98 %		60.50 %		60.08 %		57.56 %		57.56 %		70.17 %	
1995		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	62	38	24	40	22	41	21	44	18	37	25	43	19	20	42
L	186	63	123	61	125	67	119	56	130	74	112	55	131	20	166
Overall (%)		64.92 %		66.53 %		64.52 %		70.16 %		60.08 %		70.16 %		75.00 %	
1996		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	65	36	29	39	26	40	25	39	26	47	18	39	26	27	38
L	196	76	120	71	125	79	117	78	118	94	102	75	121	41	155
Overall (%)		59.77 %		62.84 %		60.15 %		60.15 %		57.09 %		61.30 %		69.73 %	
1997		H	L	H	L	H	L	H	L	H	L	H	L	H	L
H	68	44	24	44	24	43	25	42	26	36	32	44	24	20	48
L	207	80	127	76	131	74	133	67	140	76	131	73	134	20	187
Overall (%)		62.18 %		63.64 %		64.00 %		66.18 %		60.73 %		64.73 %		75.28 %	

Table 8.7: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1994-97 using accounting information from service companies to predict high and low performing shares

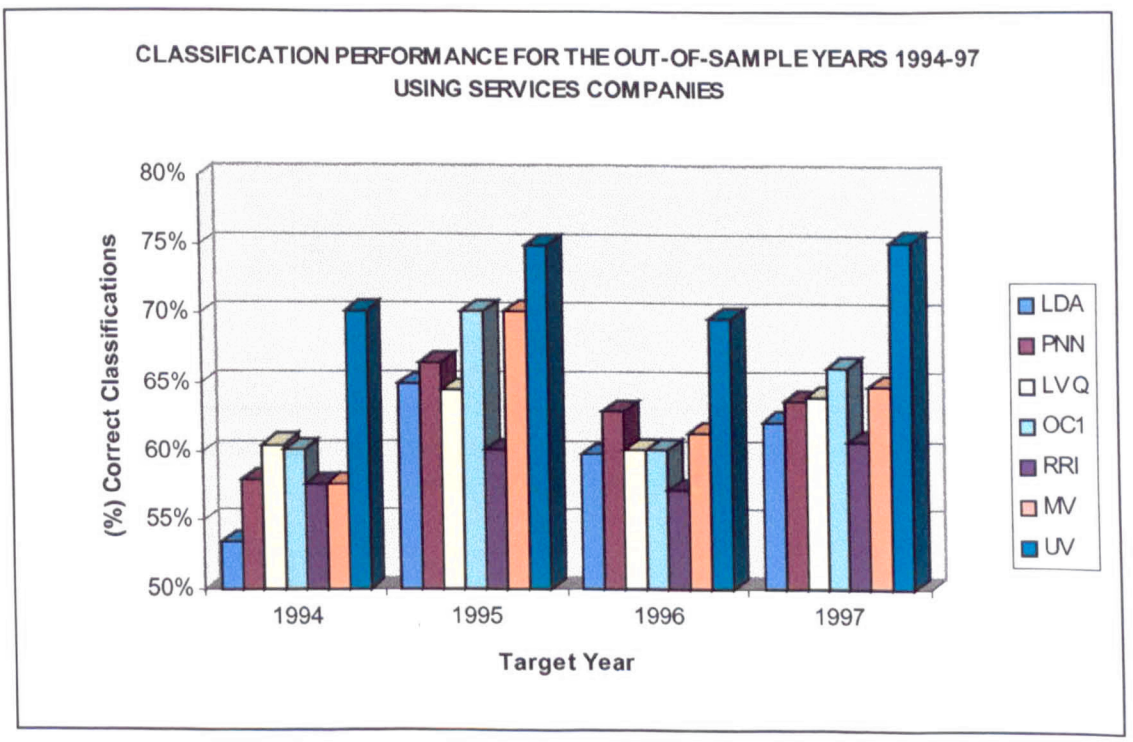
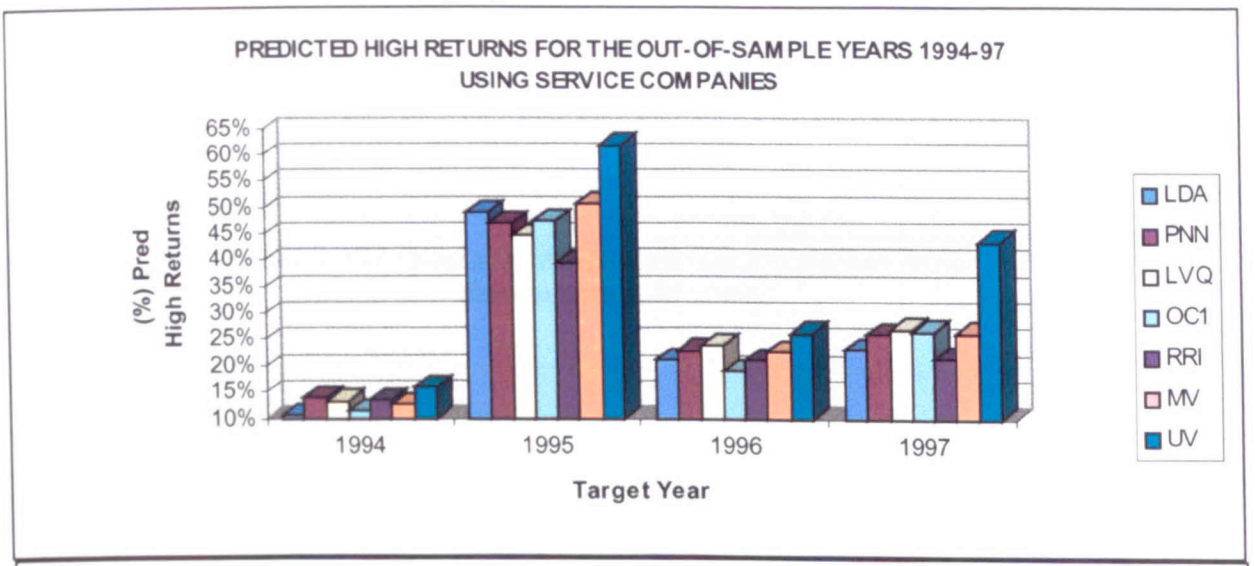
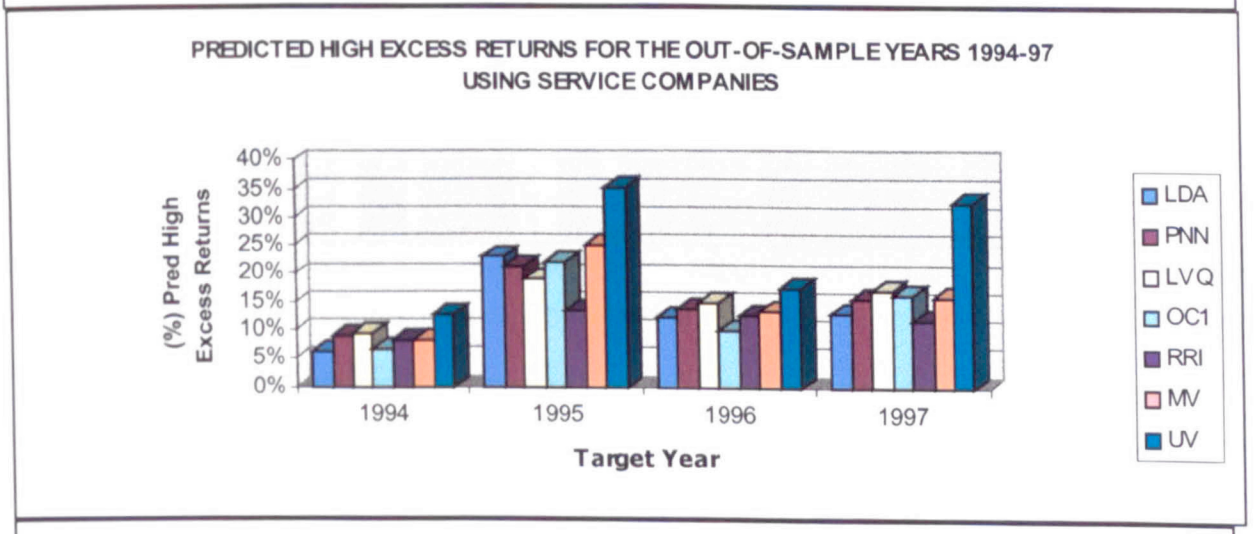


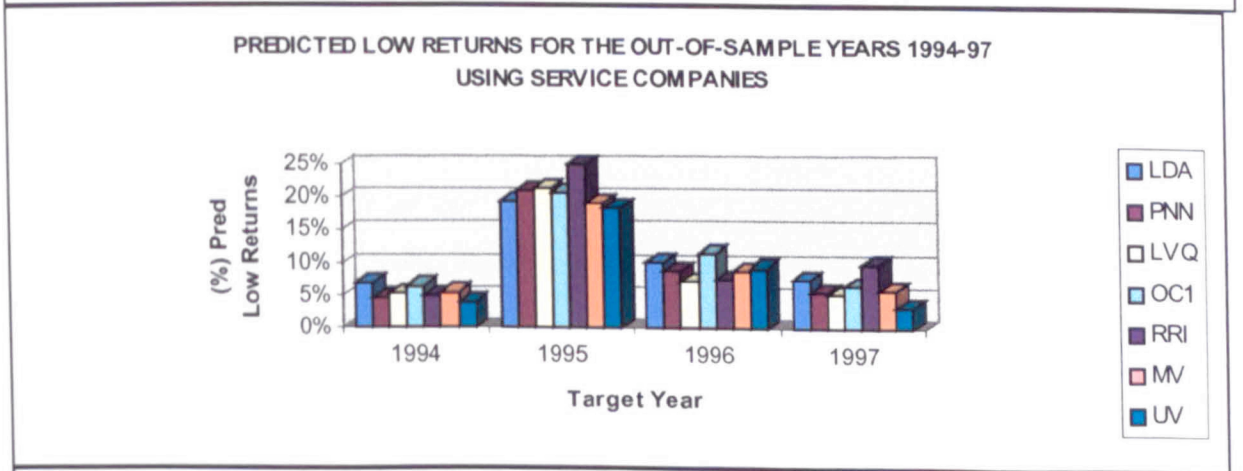
Figure 8.12: Out-of-sample classification results of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1994-97 using accounting information from service companies to predict high and low performing shares

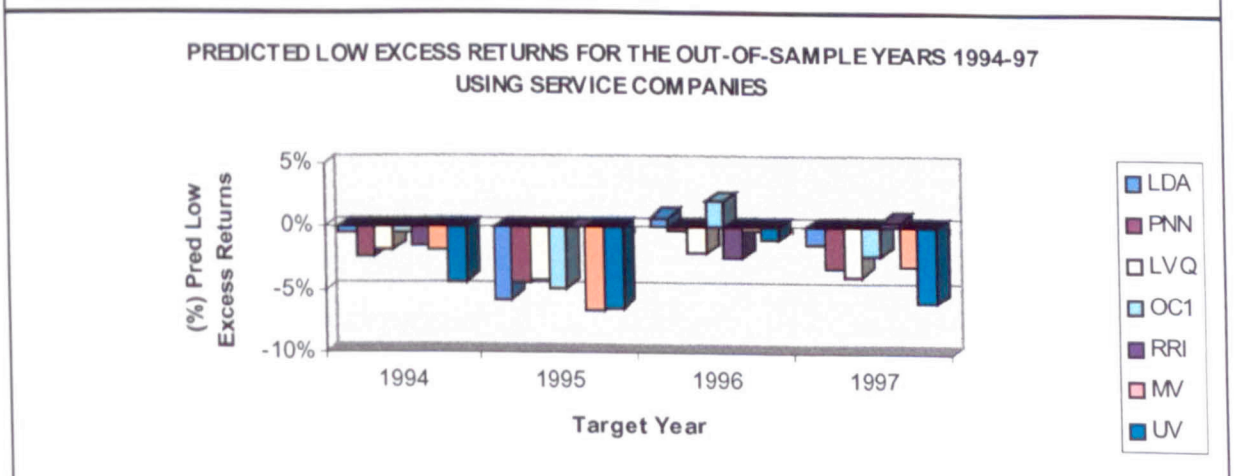
		LDA		PNN		LVQ		OC1		RRI		MV		UV	
1994		Predicted Returns & Excess Returns													
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 49.5 L= -4.7	H= 5.0	10.5	7.1	13.7	4.7	13.2	5.4	11.5	6.5	13.5	4.9	12.5	5.5	15.8	4.0
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 44.5 L= -11.0	L= 6.3	6.2	-0.4	9.0	-2.3	9.1	-1.8	6.7	-0.4	8.0	-1.4	8.2	-1.8	12.5	-4.3
1995															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 83.4 L= 14.1	H= 25.4	49.0	19.4	46.9	20.9	44.6	21.3	47.5	20.6	39.5	24.9	50.7	18.9	61.8	18.4
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 58.0 L= -11.3	L= 25.4	23.0	-5.6	21.2	-4.4	19.1	-4.1	22.1	-4.8	13.5	0.01	25.1	-6.5	35.1	-6.3
1996															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 62.6 L= -0.9	H= 9.7	21.1	10.2	22.9	9.0	24.0	7.3	19.2	11.4	21.0	7.7	22.7	8.9	26.2	9.4
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 52.9 L= -9.9	L= 9.0	12.2	0.8	13.8	-0.2	14.9	-2.0	10.1	2.1	12.5	-2.3	13.4	-0.3	17.6	-1.0
1997															
Actual Return	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 65.0 L= -2.0	H= 9.4	23.0	7.6	26.0	5.7	27.0	5.4	26.6	6.7	21.5	9.8	26.1	6.0	43.6	3.4
Actual Excess Ret	Index	H	L	H	L	H	L	H	L	H	L	H	L	H	L
H= 55.6 L= -11.5	L= 9.5	12.9	-1.2	15.7	-3.1	17.2	-3.8	16.3	-2.2	11.9	0.5	15.9	-2.9	32.7	-5.9

Table 8.8: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1994-97 using accounting information from service companies to predict high and low performing shares









Figures 8.13-8.16: Out-of-sample returns and excess returns of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1994-97 using accounting information from service companies to predict high and low performing shares

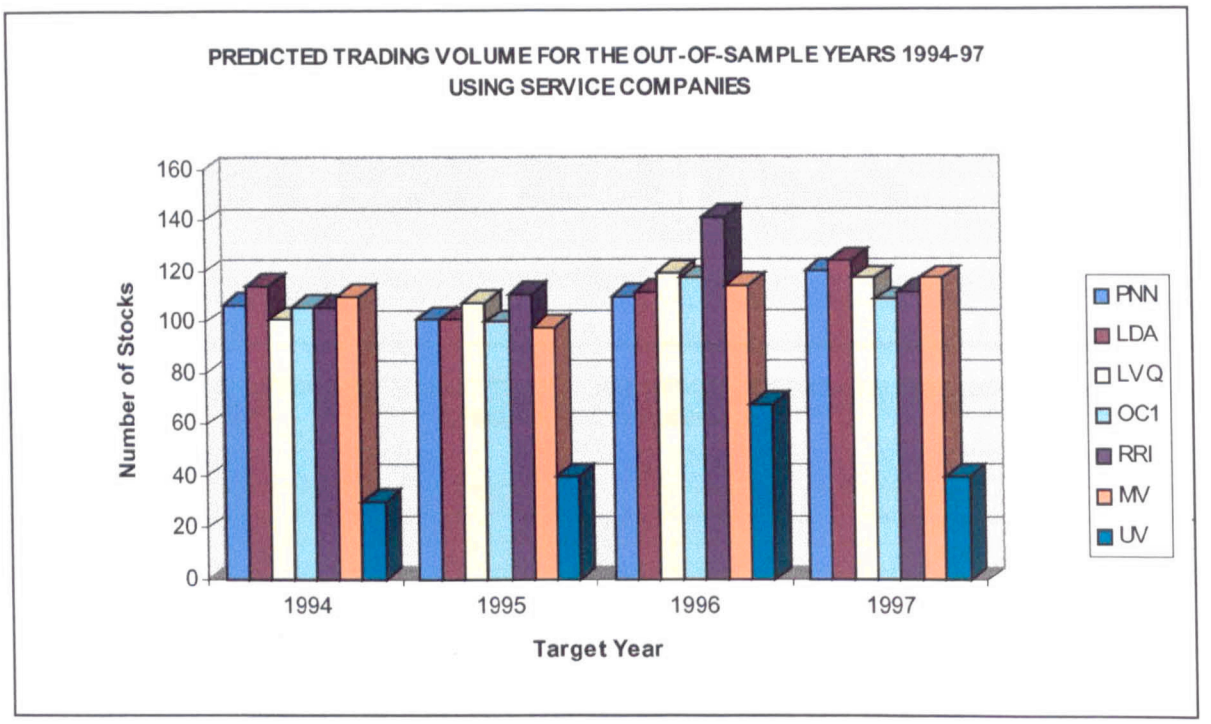


Figure 8.17: Out-of-sample trading volume of LDA, PNN, LVQ, OC1, RRI, MV, and UV for 1994-97 using accounting information from service companies to predict high and low performing shares

CHAPTER 9: DIMENSIONALITY REDUCTION TECHNIQUES BASED ON NEURAL NETWORKS - APPLICATIONS TO STOCK SELECTION AND CREDIT RATINGS

Dimensionality reduction is an important step in pre-processing data input to highly parameterised non-linear forecasting models.

One possibility is simply to drop unpromising variables. However, as mentioned in the previous Chapter there are only a few ad hoc techniques available to select the best subset of variables for classifiers such as PNN, LVQ, OC1, and RRI. For example, to select an optimal number of variables for these classifiers, we implement them using all thirty-eight ratios at the same time and we record the misclassification rate. In the next step, we remove one variable at a time and we decide whether to include each variable in the model or exclude it from the model based on the degree of improvement in the misclassification rate. However, this approach has several drawbacks. One is that selection of the variables may be dependent on the choice of the model parameters rather than the discriminating power of the variables. For example, a specific variable may be found significant for a given setting of model parameters, whereas the same variable may be found insignificant for a different setting of parameters. Searching all possible combinations of parameters for each individual variable might be an expensive task in terms of computational resources. Another drawback of this approach is that it might be time-consuming if the data set is large because we have to implement the same procedure for each individual variable.

In this Chapter, we investigate alternative approaches to dimensionality reduction including principal component analysis (PCA), neural network linear PCA (NN-PCA), and neural network non-linear PCA (NN-NLPCA). We use these techniques to reduce the dimensionality of our accounting data and we apply five classification methods namely, LDA, PNN, LVQ, OC1, and RRI to test the ability of the resulting Principal Components (PCs) to predict which shares are likely to have exceptional returns in the future. After experimentation, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology and the resulting PCs maintain important discriminating power to identify which shares are likely to have exceptional returns in the

future. We also found that the resulting PCs are easier to interpret if extracted from homogeneous groups of financial ratios.

In this Chapter, we also propose a new methodology to assess the long-term standing of debt issuers by applying NN-PCA and NN-NLPCA dimensionality reduction techniques. We use these techniques to reduce the dimensionality of a small data set of financial ratios and we apply three classification methods namely, LDA, PNN, and BPNN to test the ability of the resulting PCs to predict the long-term credit standing of debt issuers. The results of this experiment confirm the findings of the first experiment. More specifically, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology. Furthermore, we found that the PCs extracted from NN-PCA and NN-NLPCA are better discriminators than the PCs extracted from PCA and are easier to interpret if extracted from homogeneous groups of financial ratios. Overall, the results of this experiment suggest that linear and non-linear dimensionality reduction techniques based on neural networks can be an efficient tool to assess the long-term credit standing of debt issuers and at the same time provide an efficient solution to overfitting.

This Chapter is organised as follows: in the first part, we use dimensionality reduction techniques based on neural networks and we apply five classification methods namely, LDA, PNN, LVQ, OC1, and RRI to test the ability of the resulting PCs to predict which shares are likely to have exceptional returns in the future. We discuss the data and methodology that we used for this implementation and we present the results. In the second part, we apply the same dimensionality reduction techniques to assess the long-term credit standing of debt issuers by applying LDA, PNN and BPNN. We also discuss the data and methodology that we used for this implementation and we present the results. Finally, in the third part, we present the summary of this Chapter and we provide the overall conclusions.

Part One: Predicting High Performing Shares Using Dimensionality Reduction Techniques Based on Neural Networks

The central idea of dimensionality reduction is to identify a subset of variables that are functions of the original problem variables, and efficiently capture the information contained in the original data set. Let X_t be an n×m matrix that contains n observations on m variables. The superficial dimensionality of data is the number of individual observations constituting one measurement vector. The intrinsic dimensionality of data, on the other hand, is the number of independent variables underlying the significant nonrandom variations in the observations (Kramer, 1991). The superficial dimensionality of data is often greater than the intrinsic dimensionality. Dimensionality reduction techniques allow us to extract a new set of variables that contain the same information as X_t, but have smaller intrinsic dimension (Kramer 1991; Dong and McAvoy 1995). Various dimensionality reduction techniques have been proposed in the literature. One well-known technique is PCA. PCA is a technique for linearly mapping multidimensional data onto lower dimensions with minimal loss of information. PCA can be used to approximate n points in m dimensions by fitting a p-dimensional ($p<m$) plane through the middle of the points so that the sum of the distances between the points and their projections onto the plane is minimised (Malthouse, 1996). PCA has been applied in almost every discipline such as chemistry, engineering, biology, meteorology etc. (Dong and McAvoy, 1995). Reduction of dimensionality by PCA has been shown to facilitate many types of multivariate analysis including quality control (MacGregor 1989; Dong and McAvoy 1995), correlation and prediction (Stephanopoulos and Guterman 1989; Dong and McAvoy 1995), and fault detection (Wise and Ricker 1989; Dong and McAvoy 1995). Alternatively, many different more or less neural PCA subspace or principal component estimation algorithms have been proposed in the literature during the last years (Baldi and Hornik 1989; Oja 1992; Cichoski and Unbehauen 1993). However, PCA is a linear method and most real problems are non-linear. It has been shown that if PCA is applied in non-linear problems, minor components might contain important information (Chang 1983; Dong and McAvoy 1995). Therefore, if minor components are discarded important information is lost. It is therefore proposed that a non-linear Principal Component Analysis (NLPCA) should be applied to deal with these problems (Kramer 1991; Dong and McAvoy 1995).

The NLPCA is a general-purpose feature extraction algorithm producing features that retain the

maximum possible amount of information from the original data set. The main difference between PCA and NLPCA is that the former involves linear mappings between the original and reduced dimension spaces, whereas the latter involves non-linear mappings. If non-linear correlations between variables exist and sufficient data to support the formulation between more complex mapping functions are available, then NLPCA will describe the data with greater accuracy than PCA and by fewer PCs (Kramer 1991; Dong and McAvoy 1995).

Another drawback of PCA is that the relative sizes of the elements in a variable weight vector associated with a particular PC indicate the relative contribution of the variable to the variance of the PC. Therefore, the patterns of variable weights for a particular PC are used to interpret the PC. A problem is identified, however, if more than a few variables have a significant contribution to the variance of a particular PC. In this case, the interpretation of this PC becomes extremely difficult.

In this part, we discuss the theoretical properties of PCA as well as the theoretical properties of dimensionality reduction techniques based on neural networks. To test the effectiveness of these techniques, we apply them to reduce the dimensionality of our accounting data and we use the resulting PCs as inputs to five classification methods namely, LDA, PNN, LVQ, OC1, and RRI. Then, we apply these classification methods to predict which shares are likely to have exceptional returns in the future. We organise this part as follows: in Section 9.1, we discuss the PCA methodology. In Section 9.2, we present the NN-PCA model. In Section 9.3, we present the NN-NLPCA model. In Section 9.4, we discuss the data and methodology that used in this study. Our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our predictor variables are 38 accounting ratios drawn from published accounting statements. In Section 9.5, we present the results of our experimentation. After experimentation, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology and the resulting PCs maintain important discriminating power to identify which shares are likely to have exceptional returns in the future. On the other hand, we found that the resulting PCs are easier to interpret if extracted from homogeneous groups of financial ratios. In Section 9.6, we summarise the results.

9.1 PRINCIPAL COMPONENT ANALYSIS

The central idea of PCA is to reduce the dimensionality of a data set which consists of a large number of interrelated variables into a substantially smaller set of uncorrelated variables that

represent the maximum amount of information in the original set of variables. A small set of uncorrelated variables is much easier to understand and use in further analyses than a larger set of correlated variables.

PCA can be used to approximate n points in m dimensions by fitting a p-dimensional (p<m) plane through the middle of the points so that the sum of the distances between the points and their projections onto the plane is minimised (Malthouse, 1996).

Let X_t be an n×m data matrix that is centred about the mean so that the column totals are zero. According to Jolliffee (1986), the variance-covariance matrix of X_t can be estimated as $V_t = X_t'X_t/(n-1)$. If we assume that α_{it} and x_{it} are $m \times 1$ data vectors, then any linear combination $\alpha_{it}'x_{it}$ has an estimated variance $\alpha_{it}'V_t\alpha_{it}$. The PCs can be derived by solving the following equation (Jolliffee, 1986),

$$\frac{\partial}{\partial \alpha_{it}}(\alpha_{it}'V_t\alpha_{it} - l_{it}\alpha_{it}'\alpha_{it}) = 0 \tag{9.1}$$

where l_{it} is the Langrange multiplier so that $V_t\alpha_{it} = l_{it}\alpha_{it}$.

The values of l_{it} are the m eigenvalues of V_t and to each of these eigenvalues $l_{1t}, l_{2t}, \ldots, l_{mt}\left[l_{it} > l_{jt}, (i > j)\right]$ corresponds an eigenvector so that the variance of $\alpha_{it}'x_{it}$ is equal to l_{it}. The vectors $\alpha_{it}'x_{it}$ are the PCs. The vector α_{it} contains the coefficients of the i^{th} PC and $X_t\alpha_{it}$ gives the scores of the n elements on the i^{th} PC. It is obvious that $\alpha_{it}'\alpha_{it} = 1$ and $\alpha_{it}'\alpha_{jt} = 0, i \neq j$.

The magnitude of the variances of the PCs provides an indication of how well they account for the variability in the data. The relative sizes of the elements in a variable weight vector associated with a particular PC indicate the relative contribution of the variable to the variance of the PC. Therefore, the patterns of variable weights for a particular PC are used to interpret the PC. A problem is identified, however, if more than a few variables have a significant contribution to the variance of a particular PC. In this case, the interpretation of this PC becomes extremely difficult.

PCA can be viewed as an optimal transformation of X_t into two matrices: a scores matrix, and

a loadings matrix. Let S_t be the n×p scores matrix, F_t be the m×p loadings matrix, and E_t be the n×m matrix of residuals. Then, we can write the following expression (Kramer, 1991),

$$X_t = S_t F_t' + E_t \tag{9.2}$$

where p is the number of first PCs (p<m). According to Kramer (1991), the condition of optimality on the factorisation is that the Euclidean norm of the residual matrix, E_t, must be minimised for a given number of PCs. This criterion is satisfied if the columns of F_t are the eigenvectors corresponding to the p largest eigenvalues of the covariance matrix of X_t.

The Spectral or Jordan decomposition of the square symmetric matrix X_t can be written as $X_t'X_t = F_t L_t F_t'$ where the columns of F_t are the unit-length eigenvectors of $X_t'X_t$ and the diagonal elements of $L_t = \text{diag}(l_{1t}, l_{2t}, \ldots, l_{mt})$ are the eigenvalues of $X_t'X_t\left[l_{it} > l_{jt}, (i > j)\right]$. The eigenvectors form a basis for R^m and the coordinates of X_t relative to the eigenbasis are the PC scores. Therefore, we can write $S_t = X_t F_t$ (Malthouse, 1996).

PCA approximates X_t by projecting X_t onto the subspace spanned by a subset of the eigenvectors given by the columns of F_t. If F_{pt} is a matrix of the first p<m columns of F, then the approximation of matrix X_t is given as follows (Malthouse, 1996),

$$X_{it}' = Pr_{X_{it} \to F_{pt}} = S_{pt} F_{pt}' \tag{9.3}$$

where S_{pt} is the matrix of PC scores. The loadings of matrix F_t are the coefficients of the linear transformation. The information lost in this projection is given as follows (Malthouse, 1996),

$$E_{pt} = X_t - Pr_{X_{it} \to F_{pt}} \tag{9.4}$$

where E_{pt} is the matrix of errors.

PCA can be used in discriminant analysis by replacing the original set of variables by the first PCs that represent the maximum amount of variation in the original set of variables. However, Jolliffee (1986) argues that this procedure may be unsatisfactory for two reasons: first, the within-group covariance matrix may be different for different groups; and second, there is no

guarantee that the separation between groups will be in the direction of the high-variance PCs. The latter argument can be justified in the case of two completely specified normal populations that differ only in mean. In that case, the linear discriminant function is a monotonically decreasing function of the squared Mahalanobis distance, d^2, which can be written as follows,

$$d^2 = \left(\bar{x}_H - \bar{x}_L\right)' \sum^{-1} \left(\bar{x}_H - \bar{x}_L\right) \tag{9.5}$$

Chang (1983) showed that the Mahalanobis distance based on the k^{th} PC is a monotonic increasing function of $\left[\alpha_{ik}'\left(\bar{x}_H - \bar{x}_L\right)\right]^2 / l_{ik}$ where α_{ik} and l_{ik} are the vector coefficients and the variance in the k^{th} PC, respectively. If this is the case, however, then the PC with the largest discriminatory power is the one that maximises $\left[\alpha_{ik}'\left(\bar{x}_H - \bar{x}_L\right)\right]^2 / l_{ik}$ and this will not necessarily correspond to the first PC which maximises l_{ik}. On the other hand, if α_{it}' is orthogonal to $\left(\bar{x}_H - \bar{x}_L\right)$, then the first PC will have no discriminatory power. An alternative technique to PCA is to apply a neural network PCA (NN-PCA). This technique is discussed in more detail in the next Section.

9.2 NEURAL NETWORK LINEAR PCA

It is well known that many meaningful information processing operations can be done by simple neural networks whose input-output mappings become linear after learning. Oja (1992) suggests that several of the unsupervised learning algorithms of such networks are neural realisations of PCA. Let us consider a multi-layer perceptron of the form shown in Figure 9.1 having m neurons in the input layer, m neurons in the output layer, and p neurons in the hidden layer, with p<m as suggested by Malthouse (1996).

The network presented in Figure 9.1 can be used to map vectors x_{it}^n in a m-dimensional space $\left(x_{i1t}, x_{i2t}, \ldots, x_{imt}\right)$ onto vectors s_{it}^n in an p-dimensional space $\left(s_{i1t}, s_{i2t}, \ldots, s_{ipt}\right)$ where p<m. The targets used to train the network are simply the inputs themselves. Therefore, the network is trained to map each input vector onto itself. The activation functions for all nodes are linear and there are no direct connections between input and output nodes. The hidden layer is called a bottleneck layer because the m-dimensional inputs must pass through this p<m dimensional layer before producing the inputs. Data compression therefore occurs in the bottleneck layer (Bishop 1995; Malthouse 1996).

If we assume a whole data set of N vectors x^n , then the network can be trained to minimise the following objective function (Malthouse, 1996),

$$E = \min \sum_{i=1}^{n} \sum_{i=1}^{m} \left(x_{ijt} - \tilde{x}_{ijt} \right)^2 \tag{9.6}$$

Bourland and Kamp (1988) and Baldi and Hornik (1989) showed that if the hidden units have linear activation functions then it can be shown that the error function has a unique global minimum and at this minimum the network performs a projection onto the p-dimensional sub-space which is spanned by the first p principal components of the data. The weight vectors between the input and the bottleneck layers span the same subspace as the p eigenvectors in PCA.

PCA networks are useful in optimal feature extraction and data compression, and they have a number of possible applications in different areas. However, they have an important limitation that makes them less attractive from a neural network point of view: the input-output mapping becomes generally non-linear and linear PCA networks are able to perform only linear input-output mappings. We might think that this limitation of the linear PCA network could be overcome if we used non-linear activation functions for the hidden units in Figure 9.1. However, Bourland and Kamp (1988) showed that non-linear activation functions in the hidden layer make no difference and that the minimum error solution is again given by the projection onto the principal component sub-space. However, the results would be different if additional hidden layers with non-linear activation functions were permitted into the network. In that case, we could have a non-linear PCA network that is discussed in more detail below.

9.3 NEURAL NETWORK NON-LINEAR PCA

The NLPCA is a general purpose feature extraction algorithm producing features that retain the maximum possible amount of information from the original data set. The main difference between PCA and NLPCA is that the former involves linear mappings between the original and reduced dimension spaces, whereas the latter involves non-linear mappings. If non-linear correlations between variables exist and sufficient data to support the formulation between more complex mapping functions are available, then NLPCA will describe the data with greater accuracy than PCA and by fewer PCs (Kramer, 1991). In summary, the motivations of using non-linearities in PCA-type networks are illustrated below (Baldi and Hornik 1989; Kramer 1991; Karhunen and Joutsensalo 1994; Dong and McAvoy 1995),

- In optimising non-quadratic criteria, closed-form solutions are not usually available and iterative algorithms must be used. The gradient type neural learning algorithms are iterative by nature and a suitably chosen non-linearity such as the sigmoid can be easily implemented via analog hardware.
- Higher-order statistics are introduced into the computations in an implicit way. Higher-order statistics are needed for a good representation of non-Gaussian data.
- The outputs of standard PCA networks are not usually independent which would be more desirable in many cases. Adding non-linearities to a PCA network increases the independence of the outputs so that the original signals can sometimes be roughly separated from their mixture.

Let x_{it} represents a row of the X_t, a single data vector, and s_{it} represents a row of the scores matrix, S_t. By applying NLPCA, we seek a mapping in the form (Kramer, 1991),

$$s_{it} = \phi_i(x_{it}) \tag{9.7}$$

where ϕ is a non-linear vector function, composed of p individual non-linear functions $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$ analogous to the columns of the loadings matrix, F_t. The information lost in this mapping can be assessed by reconstruction of the measurement vector by reversing the projection back to R^m as follows (Kramer 1991; Dong and McAvoy 1995),

$$x'_{jt} = \theta_j(s_{it}) \tag{9.8}$$

where $\theta_j = (\theta_1, \theta_2, \ldots, \theta_m)$ is a second non-linear function. The loss of information is measured by $E_{it} = X_{it} - X'_{jt}$. It is obvious that the functions ϕ_i and θ_j are selected to minimise $|E_{it}|$ for individual measurement vectors, or $|E_t|$ for the whole data set. According to Cybenko (1989) ϕ and θ can be modelled by fitting functions of the following form,

$$v_k = \sum_{j=1}^{M_2} w_{jk2}\sigma\left(\sum_{i=1}^{M_1} w_{ij1}\varepsilon_i + \zeta_{j1}\right) \tag{9.9}$$

where $\sigma(x_{it})$ is a continuous monotonically increasing function so as $\sigma(x_{it}) \to 1$ as $x_{it} \to +\infty$, and $\sigma(x_{it}) \to 0$ as $x_{it} \to -\infty$. However, if $\sigma(x_{it})$ is the sigmoid function, then

Eq. (9.9) describes a feedforward neural network with M_1 inputs, a hidden layer with M_2 nodes and sigmoidal transfer functions, and a linear output node for each k. The w_{ijk} is the weight on the connection from node i in layer k to node j in layer $k+1$, whereas ζ are nodal biases that are treated as adjustable parameters like the weights.

According to Kramer (1991), the neural networks that represent ϕ and θ can be described as follows: the neural network that represents the function ϕ operates on the rows of the n×m data matrix X_t and has m inputs. The hidden layer of this network is a mapping layer that contains N_1 nodes with sigmoidal transfer functions so that $N_1 > p$. The output of the network is then a projection of the input vector into feature space and therefore has p nodes with linear or sigmoidal activation functions. The function ϕ_i is the i^{th} non-linear factor that is defined by the weights and biases of the connections from the input to the i^{th} output. On the other hand, the network that represents the inverse mapping function θ_i takes the rows of the scores matrix S_t as inputs and therefore it has p inputs. The hidden layer of this network is a demapping layer that contains N_2 nodes with sigmoidal transfer functions so that $N_2 > p$. The output layer of the network contains m nodes that can be linear or sigmoidal and give the reconstructed data matrix X_t'. The function θ_i is defined by the weights and biases that connect the inputs to the i^{th} output node. Kramer (1991) combined the two networks that represent functions ϕ and θ as in Figure 9.2. The particular network architecture presented in Figure 9.2 employs five layers: input layer (1), mapping layer (2), bottleneck layer (3), de-mapping layer (4), and output layer (5). This five-layer NLPCA architecture has m nodes in the input layer, p nodes in the bottleneck layer, and m nodes in the output layer. The mapping and de-mapping layers of the network (2 and 4, respectively) have sigmoidal activation functions. The bottleneck and output layers (3 and 5, respectively) have linear activation functions. Direct connections are allowed between layers 1 and 3 and between layers 3 and 5, but direct connections are not allowed to cross bottleneck layer 3 (Malthouse, 1996). The bottleneck layer has fewer nodes than the input or output layer. The network is trained to perform the identity mapping, where the input is approximated at the output layer. Because the dimension of the bottleneck layer is smaller than both the input and output layers, the network is forced to develop a compact representation of the input data (Malthouse, 1996). If network training finds an acceptable solution, a good representation of the input must exist in the bottleneck layer. This implies that data compression caused by the bottleneck layer may force hidden units to represent significant features in data. Kramer (1991) proposed that the outputs of the bottleneck layer represent the non-linear PCs.

NLPCA reduces the dimension of the inputs by fitting a curve or surface through the data. The

first three layers (1,2, and 3) of the network project the original data onto the curve or surface and the activation values of the bottleneck layer, called scores, give the location of the projection. The last three layers (3, 4 and 5) define the curve or surface. Let $s_f: \Re^m \rightarrow \Re^p$ denote the function modelled by layers 1,2 and 3, and let $f: \Re^p \rightarrow \Re^m$ denote the function modelled by layers 3, 4 and 5. The weights in the autoassociative NLPCA network are determined under the following objective function (Malthouse, 1996),

$$\min_{f,s_f} \sum_{i=1}^{n} \left\| x_{it} - f\left(s_f\left(x_{it}\right)\right) \right\|^2 \tag{9.10}$$

Many methods have been proposed in the literature to minimise any smooth non-linear function $f: \Re^m \rightarrow \Re$ if the number of variables m is large. These methods include among others the Newtons's method, variations of the Newtons' method, the partitioned quasi-Newton method, the conjugate gradient method, the limited memory quasi-Newton method (L-BFGS), and the methods that combine cycles of BFGS and conjugate gradient steps (Liu and Nocedal, 1989).

Numerical tests that performed on medium size problems have shown that L-BFGS methods require fewer function evaluations than conjugate gradient methods even when little additional storage is added. In addition, empirical evidence suggests that L-BFGS methods are faster than methods that combine cycles of BFGS and conjugate gradient steps and they are better able to use additional storage to accelerate convergence (Liu and Nocedal, 1989). However, evidence also suggests that partitioned quasi-Newton methods are better than L-BFGS methods for problems where the user is able to supply the information on the objective function that the method requires as well as for problems where the element functions depend on less than 3 or 4 variables (Liu and Nocedal, 1989). On the other hand, L-BFGS methods are better than quasi Newton-methods for problems where the element functions depend on more than 3 or 4 variables because they are very simple to implement, they are faster, they require only function and gradient values, and they require no other information about the problem (Liu and Nocedal, 1989).

According to Liu and Nocedal (1989), limited memory quasi-Newton methods can be viewed as extensions of the conjugate gradient methods and they are more suitable for large-scale problems because the amount of storage to be used can be controlled by the user. On the other hand, L-BFGS methods can be viewed as implementations of quasi-Newton methods in which storage is restricted. The main advantage of L-BFGS methods is that they do not require knowledge of the sparcity structure of the Hessian. Let us denote the iterates by x_k and let us

define $\sigma_k = x_{k+1} - x_k$ and $\gamma = g_{k+1} - g_k$. According to Liu and Nocedal, the L-BFGS methods use the inverse BFGS formula in the form $H_{k+1} = B_k^T H_k B_k + r_k \sigma_k \sigma_k^T$ where $r_k = 1/\gamma_k^T \sigma_k$ and $B_k = 1 - r_k \gamma_k \sigma_k^T$. The L-BFGS algorithm can be described as follows:

Step 1: choose x_0, μ, and θ, θ' so that $0 < \theta' < \frac{1}{2}$ and $\theta' < \theta < 1$. In addition, choose a symmetric and positive definite starting matrix H_0 and set k=0.

Step 2: compute $\delta_k = -H_k g_k$ and $x_{k+1} = x_k + \alpha_k \delta_k$ where α_k satisfies the wolfe conditions:

$f(x_k + \alpha_k \delta_k) \leq f(x_k) + \theta' \alpha_k g_k^T \delta_k$ and $g(x_k + \alpha_k \delta_k)^T \delta_k \geq \theta g_k^T \delta_k$.

Step 3: let $\tilde{\mu} = \min(k, \mu - 1)$. Update H_0 $\tilde{\mu} + 1$ times using the pairs $\left[\gamma_j, \sigma_j\right]_{j=k-\tilde{\mu}}^{k}$. In that case,

$$H_{k+1} = \left(B_k^T \ldots B_{k-\tilde{\mu}}^T\right) H_0 \left(B_{k-\tilde{\mu}}, \ldots, B_k\right) + r_{k-\tilde{\mu}} \left(B_k^T, \ldots, B_{k-\tilde{\mu}+1}^T\right) \sigma_{k-\tilde{\mu}} \sigma_{k-\tilde{\mu}}^T \left(B_{k-\tilde{\mu}+1}^T, \ldots, B_k\right) + \ldots r_k \sigma_k \sigma_k^T$$

Step 4: set k:= k + 1 and go to step 2.

Liu and Nocedal note that the matrices H_k are not formed explicitly, but the previous values of γ_j and σ_j are stored separately.

9.4 DATA AND METHODOLOGY

To apply the three dimensionality reduction techniques namely PCA, NN-PCA, and NN-NLPCA we used the same accounting data that we used to apply the five classification methods in Chapter 5[4]. Therefore, our target data are total returns on all shares traded on the London Stock Exchange in the years 1993-97. This consists of around 700 shares per year starting with 626 shares in 1993 and rising up to 718 shares in 1997. Our input variables are 38 accounting ratios drawn from published accounting statements. These data are the thirty-eight accounting ratios presented in Table 5.1.

We implemented the dimensionality reduction techniques for each group of ratios separately, and we extracted nine PCs - one for each group of ratios, respectively - instead of using all the thirty-eight ratios at the same time. Applying this methodology, we achieved two clear advantages: first, to reduce the possibility of overfitting; and second, to be able to interpret the PCs since they are extracted from homogeneous groups of ratios.

[4] We would like to thank C. Malthouse for very kindly providing the source code for the implementation of the NN-PCA and NN-NLPCA dimensionality reduction techniques. This code was slightly modified for the purpose of our studies.

To examine the discriminating power of the PCs extracted from PCA, NN-PCA, and NN-NLPCA, we used them as inputs to five classification methods namely, LDA, PNN, LVQ, OC1, and RRI to classify H and L performing shares. The five classifiers were implemented as follows: first, we trained the classifiers on one year and then we tested their performance on the following year. For example, to predict relative excess returns for 1992, we first trained the classifiers on the previous year 1991 and we tested them on the out-of-sample year 1992. We then moved the implementation one-year ahead and we used information available from 1992 to predict 1993 and so on. After this implementation, we compared the classification results with the classification results obtained after implementing the classifiers using the subsets of variables presented in Table 5.2 which is the best subsets that we found after applying stepwise variable elimination procedures. We recall that these classification results were obtained after using two years of data to train the classifiers in order to predict the following year. Therefore, if the predictions of these two implementations are found to be close, then this would prove that the PCs derived from the neural network dimensionality reduction techniques not only maintain discriminating power, but they also reduce the possibility of overfitting even if the data set is small.

The data used for the implementation of the NN-PCA and NN-NLPCA were normalised within the range (0,1) using Eq. (4.1) We minimised the objective functions (9.6) and (9.10) by applying the LBFGS optimisation routine as suggested by Liu and Nocedal (1989) and Malthouse (1996).

9.5 RESULTS

Table 9.1 shows the Percentage of Variance Explained (PVE) by PCA, NN-PCA, and NN-NLPCA if the thirty-eight ratios are grouped into homogeneous groups of ratios based on conceptual clustering, one architecture is applied for each group, and only one PC is extracted from each group for the years 1993-1997. The PVE by the individual PCs was found after implementing the architectures every two successive years. For example, the PVE by the individual PCs for 1993 was found after implementing the architectures using the data for 1992 and 1993. The bottom part of Table 9.1 shows the average PVE by the individual PCs for the period 1993-97. As we can see, the PCs extracted from NN-NLPCA explain a greater proportion of variance in the original data set than the other two dimensionality reduction techniques. These results indicate the existence of strong non-linearities in the original data set. The average PVE by the individual PCs is also presented in Figure 9.3.

To determine whether the average PVE by the architectures is equal, we performed statistical tests. First, we tested the NN-PCA against the PCA, and then we tested the NN-NLPCA against the NN-PCA. In the former case, the null hypothesis is that the average PVE by the PCA is greater than or equal to the average PVE by the NN-PCA. In the latter case, the null hypothesis is that the average PVE by the NN-PCA is greater than or equal to the average PVE by the NN-NLPCA. To perform these tests, we applied the following statistics,

$$Z_1 = \frac{\left(\tilde{P}_{NN-PCA} - \tilde{P}_{PCA}\right) - 0}{\left(\frac{\tilde{P}_{NN-PCA}\tilde{Q}_{NN-PCA} + \tilde{P}_{PCA}\tilde{Q}_{PCA}}{n}\right)^{\frac{1}{2}}} \qquad Z_2 = \frac{\left(\tilde{P}_{NN-NLPCA} - \tilde{P}_{NN-PCA}\right) - 0}{\left(\frac{\tilde{P}_{NN-NLPCA}\tilde{Q}_{NN-NLPCA} + \tilde{P}_{NN-PCA}\tilde{Q}_{NN-PCA}}{n}\right)^{\frac{1}{2}}}$$

where $\tilde{P}$ is the average percentage of variance explained (APVE) , $\tilde{Q}$ is the percentage of variance unexplained (APNU), and n is the average number of companies.

As we can see from the p-values presented in Table 9.2, the null hypothesis is rejected for 8 out of 9 Z_1 tests and is also rejected for all Z_2 tests at the 5% level of significance. These results suggest that the average PVE by the NN-PCA is significantly greater than the PVE by the PCA, whereas the PVE by the NN-NLPCA is significantly greater than the PVE by NN-PCA and therefore by PCA.

Table 9.3 compares the classification performance of LDA for the out-of-sample years 1993-97 after five implementations: 1) using all variables; 2) using the best subset of variables that we found after variable reduction; 3) using PCA for each homogeneous group of ratios and then using the resulting PCs as independent variables; 4) using NN-PCA for each homogeneous group of ratios and then using the resulting PCs as independent variables; and 5) using NN-NLPCA for each homogeneous group of ratios and then using the resulting PCs as independent variables. The classification performance of LDA for the out-of-sample target years 1993-97 is also illustrated in Figure 9.4. As we can see, the classification performance of LDA is better after using the best subset of variables for the out-of-sample year 1993. On the other hand, the classification performance of LDA is more unstable after using the PCs from the PCA methodology for the out-of-sample years 1994-97, whereas there are only minor inconsistencies in the performance of this classifier after applying the other four dimensionality reduction techniques over the same period.

Table 9.4 shows the classification performance of the PNN for the out-of-sample years 1993-97 after applying the five dimensionality reduction techniques. These results are also illustrated in Figure 9.5. As we can see, the classification performance of the PNN is better after using the

best subset of variables for the out-of-sample years 1993 and 1994. On the other hand, the classification performance of the PNN seems to be more unstable after using the PCs from the PCA methodology for the out-of-sample years 1995-97, whereas there are only minor inconsistencies in the performance of this classifier after using the other dimensionality reduction techniques over the same period.

Table 9.5 shows the classification performance of LVQ for the out-of-sample years 1993-97 after applying the five dimensionality reduction techniques. These results are also illustrated in Figure 9.6 As we can see, the classification performance of LVQ is significantly better after using the best subset of variables for the out-of-sample year 1993, whereas it is significantly better after using the PCs from NN-NLPCA for the out-of-sample years 1994, 1995 and 1997. On the other hand, there are only minor inconsistencies in the performance of this classifier for the out-of-sample year 1996 under all five dimensionality reduction techniques.

Table 9.6 shows the classification performance of OC1 for the out-of-sample years 1993-97 after applying the five dimensionality reduction techniques. These results are also presented in Figure 9.7. As we can see, the classification performance of OC1 is significantly better after using the PCs from NN-NLPCA as well as after using the best subset of variables for the out-of-sample years 1993-97. On the other hand, the classification performance of OC1 is affected significantly after using all variables as well as after using the PCs from PCA and NN-PCA over the same period.

Table 9.7 shows the classification performance of RRI for the out-of-sample years 1993-97 after applying the five dimensionality reduction techniques. These results are also presented in Figure 9.8. As we can see, there are large inconsistencies in the classification performance of RRI after applying the five methodologies. The results suggest that the classification performance of RRI is significantly better after using the best subset of variables for the out-of-sample year 1993, whereas it is significantly better after using all variables for the out-of-sample year 1994. On the other hand, the classification performance of RRI is significantly better after using the PCs from PCA for the out-of-sample years 1995 and 1997, whereas the NN-PCA outperforms the other techniques for the out-of-sample year 1996.

Table 9.8 shows the average total percentage of correct classifications of the classification methods for the whole out-of-sample period 1993-97. These results are also illustrated in Figure 9.9. As we can see, the classification performance of LDA, PNN, LVQ, and OC1 is affected seriously after applying the PCA methodology. The LDA classifies on average better after using the best subset of variables, the PCs from NN-PCA, and the PCs from NN-NLPCA, whereas the

PNN produces very similar results to LDA under the same dimensionality reduction techniques. On the other hand, the classification performance of LVQ and OC1 is on average better after using the best subset of variables as well as the PCs from NN-NLPCA, whereas the classification performance of RRI is on average better after using all variables as well as after using the best subset of variables.

Table 9.9 shows the average actual H and L returns and the respective average H and L returns predicted by the classifiers for the out-of-sample years 1993-97 after applying the five dimensionality reduction techniques. The bottom part of Table 9.9 shows the average actual H and L returns and the respective average H and L returns predicted by the classifiers during the whole out-of-sample period 1993-97. The average results are also presented in Figures 9.10 and 9.11 for H and L returns, respectively. As we can see in Figure 9.10, the LDA predicts more accurately H returns after using the best subset of variables, whereas it also produces very favourable results after using the PCs from the NN-PCA and the NN-NLPCA dimensionality reduction techniques. The PNN produces similar results to LDA but it also favours the use of all variables. The OC1 and RRI classifiers predict more accurately H performing shares after using the PCs from NN-NLPCA, whereas the RRI produces very similar results under the five dimensionality techniques. The conclusions for L returns are similar for the individual classifiers and the different dimensionality reduction techniques as we can see in Figure 9.11.

Table 9.10 shows the average actual H and L excess returns and the respective average H and L excess returns predicted by the classifiers for the out-of-sample target years 1993-97 after applying the five dimensionality reduction techniques. The bottom part of Table 9.10 shows the average actual H and L excess returns and the respective average H and L excess returns predicted by the classifiers during the whole out-of-sample period 1993-97. The average results are also presented in Figures 9.12 and 9.13 for H and L excess returns, respectively. As we can see in Figure 9.12, LDA and PNN predict more accurately H excess returns after using the best subset of variables, whereas LVQ and OC1 predict more accurately H excess returns after applying the NN-NLPCA dimensionality reduction technique. On the other hand, the RRI produces very similar results under the five dimensionality techniques. The conclusions for L returns are similar for the individual classifiers and the different dimensionality reduction techniques as we can see in Figure 9.13.

9.6 SUMMARY OF THE RESULTS

In this part, we applied neural network implementations of linear principal component analysis (NN-PCA) and non-linear PCA (NN-NLPCA) to reduce the dimensionality of a large number

of accounting variables. Then, we used the resulting PCs to predict high performing shares that are likely to have exceptional returns in the future by applying five classifiers namely, LDA, LVQ, PNN, OC1, and RRI. Our target data were total returns on all shares traded on the London Stock Exchange in the years 1993-97. Our input variables were 38 accounting ratios drawn from published accounting statements. After experimentation, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology (Table 9.1). On the other hand, we found that the resulting PCs from NN-PCA and NN-NLPCA are competitive to other dimensionality reduction techniques in maintaining important discriminating power to identify which shares are likely to have exceptional returns in the future. For example, the LDA and the PNN predict more accurately H returns after using either the best subset of variables found from stepwise variable deletion or all variables, respectively, while they both produce very competitive returns after using the PCs from NN-PCA and NN-NLPCA dimensionality reduction techniques (Table 9.9). On the other hand, the OC1 and RRI classifiers predict more accurately H returns using the PCs from NN-NLPCA. Finally, the RRI classifier produces very similar financial returns under the five dimensionality techniques. The conclusions for excess returns are more or less the same (Table 9.10).

Part Two: Assessing the Long-Term Credit Standing of Debt Issuers Using Dimensionality Reduction Techniques Based on Neural Networks - An Alternative to Overfitting

In this part, we apply neural network dimensionality reduction techniques to homogeneous groups of financial ratios and then we use the derived PCs to assess the long-term credit standing of U.K. debt issuers by applying three classification methods namely, LDA, PNN, and BPNN. This part is organised as follows: in Section 9.7, we discuss the data and methodology that we used in this study. Our target data are 15 triple AAA, 38 double AA, and 67 single A rated debt issuers. On the other hand, our predictor variables are thirty financial ratios that were drawn from published accounting statements. In Section 9.8, we present the results of our experimentation. The results from this experiment confirmed the findings of the experiment we described in the first part. More specifically, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology. Furthermore, we found that the PCs extracted from NN-PCA and NN-NLPCA are better discriminators than the PCs extracted from PCA and are easier to interpret if extracted from homogeneous groups of financial ratios. Overall, the results of this experiment suggest that linear and non-linear dimensionality reduction techniques based on neural networks can be an efficient tool to assess the long-term credit standing of debt issuers and at the same time

provide an efficient solution to overfitting. In Section 9.9, we summarise the results.

9.7 DATA AND METHODOLOGY

Standard & Poor's (S&P's) issuer credit ratings express a current opinion of an obligor's overall capacity to pay its financial obligations. Although this opinion focuses on the obligor's overall capacity and willingness to meet its financial commitments as they common due, it does not take into account the nature of the provisions of the obligation, its standing in bankruptcy or liquidation, statutory preferences, or the enforceability of the obligation. In addition, this opinion does not take into account the creditworthiness of the guarantors, insurers, or other forms of credit enhancement on the obligation (S&P's 1991-97).

The issuer credit rating is not a recommendation to purchase, sell, or hold a financial obligation issued by an obligor because it does not comment on market price or suitability for a particular investor. Issuer credit ratings are based on current information provided by obligors or obtained by S&P's from sources it considers reliable. However, S&P's does not perform an audit in connection with any issuer credit rating and may, on occasion, rely on unaudited financial information. Issuer credit ratings may be changed, suspended, or withdrawn as a result of changes in, or unavailability of, such information, or based on other circumstances (S&P's 1991-97).

Three boundary credit ratings were chosen for this study according to the rating definitions given by the S&P's bonds guide: 1) AAA: the highest rating assigned - the obligor's capacity to meet its financial commitments is extremely strong, 2) AA: the obligor's capacity to pay its financial commitments is very strong - differ from highest rated issues only in small degree, and 3) A: the obligor's capacity to meet its financial commitments is strong but may be more susceptible to adverse changes in economic conditions than higher-rated categories (S&P's 1991-97).

In this application, we are particularly interested in whether a particular debt issuer will be classified as A, AA or AAA based on accounting information. Let us assume that c_{it} is the rating on some issuer i at time t, and x_{it} is the vector of accounting information attributes for company i known at time t. The idea is to apply a classification method to assign c_{it} to one of the three classes $C_{jt}(j = A, AA, AAA)$ using as inputs the vector x_{it} of variables that represent accounting information at time t.

Our data set consists of 15 triple AAA, 38 double AA, and 67 single A rated issuers. Thirty financial ratios were used in this study to predict long-term credit ratings. These ratios are presented in Table 9.11. As we can see in Table 9.11, the ratios are divided in eight representative groups namely Return on Capital (RCAP), Profitability (PROF), Financial Leverage (FLEV), Investment (INV), Growth (GRT), Short-Term Liquidity (ST-LIQ), Interest Coverage (ICOV), and Risk (RISK). The growth ratios are expressed as a percentage change between the beginning and end of the year. The influence of the variable on the credit rating was the primary factor in the selection of variables.

In the first step, we applied the PCA, the NN-PCA, and the NN-NLPCA to extract PCs using all thirty ratios at the same time. After performing this implementation, we found that the architectures overfit the data. In order to reduce the possibility of overfitting, we then applied the architectures for each group of ratios separately, and we extracted eight PCs - one for each group of ratios, respectively - instead of using all the thirty ratios at the same time. Applying this methodology, we achieved two clear advantages: first, to avoid the problem of overfitting; and second, to be able to interpret the PCs since they were extracted from homogeneous groups of financial ratios.

The data used for the implementation of the PCA, NN-PCA, and NN-NLPCA were normalised within the range (0,1) by applying Eq. (4.1). We minimised the objective functions (9.6) and (9.10) by applying the LBFGS optimisation routine as suggested by Liu and Nocedal (1989).

We implemented the PNN to classify 120 U.K. debt issuers using the PCs extracted from NN-PCA and NN-NLPCA. On the other hand, we implemented the LDA using the PCs extracted from PCA and NN-PCA. We also implemented a BPNN architecture to classify issuer credit ratings. The purpose of doing this latter experiment was to test the sensitivity of the results under this particular algorithm that has powerful approximation properties.

We applied the LDA, the PNN, and the BPNN algorithms to classify issuers into the three classes of ratings at the same time as well as into one class against the other. In the latter case, we classified issuers using three separate discriminant functions between pairs of ratings, namely A-AAA, A-AA, AA-AAA. Out of sample performance for the LDA and PNN models was determined by applying the leave-one out method. Leaving out each of the cases in turn, calculating the function based on the remaining n-1 cases, and then classifying the left-out case, all observations were tested out of sample. A different procedure was applied to determine out of sample performance for the BPNN. According to this procedure, we omitted 20% of the cases in turn from the training data, we trained the network using the remaining 80% of the

cases, and then we tested it using the omitted cases. Repeating this procedure five times, all observations were tested out of sample.

9.8 RESULTS

Table 9.12 shows the Percentage of Variance Explained (PVE) by PCA, NN-PCA, and NN-NLPCA if all thirty ratios are used to extract the PCs and the number of PCs increases. As we can see, the PCs extracted from NN-NLPCA explain a greater proportion of variance in the original data set than the other two dimensionality reduction techniques. The PVE increases as the number of PCs increases. Table 9.13 shows the classification results of LDA and PNN after applying the leave-one-out method. For this implementation, we have used the PCs extracted from PCA and NN-PCA to implement the LDA model, and the PCs extracted from NN-PCA and NN-NLPCA to implement the PNN model. The LDA and PNN models were implemented to classify debt issuers into three boundary ratings at the same time namely A, AA, and AAA. The data set size consists of 15 triple AAA, 38 double AA, and 67 single A rated issuers. As we can see, the PNN model performs better than the LDA when the PCs extracted from NN-PCA are used to implement the model.

Table 9.14 shows the PVE by the NN-NLPCA architecture if 80% of the data set is used for training, 20% is used for testing, all thirty ratios are used at the same time to extract eight PCs, and the number of hidden layers increases from three to eight. As we can see, the PVE by the NN-NLPCA is lower in the test set than in the training set. As the number of hidden layers increases the performance in the training set increases while the performance in the test set decreases. These results are indicative that the NN-NLPCA architecture overfits the data if all thirty ratios are used at the same time as the input variables.

Table 9.15 shows the performance of NN-NLPCA architectures in the training and test sets if the thirty ratios are grouped into homogeneous groups of ratios based on conceptual clustering, one NN-NLPCA architecture is applied for each group, and only one PC is extracted from each group. As we can see in Table 9.15, there is no much difference in the PVE in the training and test sets after applying the NN-NLPCA architectures to extract one individual PC from each homogeneous group. These results reduce the possibility that the NN-NLPCA architecture overfits the data.

Table 9.16 shows the PVE by the PCA, the NN-PCA, and the NN-NLPCA architectures if all observations are used for training, one architecture is applied to each homogeneous group of ratios, and one PC is extracted from each individual architecture. As we can see in the last

column, the NN-NLPCA explains a greater proportion of variance than the PCA and NN-PCA.

Tables 9.17 and 9.18 show the classification performance of LDA and PNN, respectively, if the individual PCs that are extracted from the eight homogeneous groups of ratios are combined and used as independent variables to implement the models. The diagonal elements are the number of issuers classified correctly after applying the LDA using the PCs extracted from PCA and NN-PCA, and the PNN using the PCs extracted from NN-PCA and NN-NLPCA. The elements off the diagonals are the number of issuers classified incorrectly into classes. As we can see, the PNN performs better than LDA if the PCs extracted from NN-PCA are used as independent variables. The first part of Table 9.18 shows that 57 out of 67 A rated debt issuers classified correctly as A, 32 out of 38 AA rated debt issuers classified correctly as AA, and 14 out of 15 AAA rated debt issuers classified correctly as AAA if the PNN model is applied using the PCs extracted from NN-PCA. Overall, the LDA model classified correctly 65% of the debt issuers using the PCs extracted from PCA, and 67.5% using the PCs extracted from NN-PCA. On the other hand, the PNN classified correctly 85.83% of the debt issuers using the PCs extracted from NN-PCA, and 75.83% of the debt issuers using the PCs extracted from NN-NLPCA.

Tables 9.19 and 9.20 show the classification performance of LDA and PNN, respectively, if each debt issuer is classified in a class one class against the other after applying the leave-one-out method. A Type I error is one in which debt issuers on the left of the first column are classified by the predictor as debt issuers on the right of the first column. On the other hand, a Type II error is one in which debt issuers on the right of the first column are classified by the predictor as debt issuers on the left. A total error refers to the total incorrect classifications for the set, regardless of type.

Tables 9.21 and 9.22 show the classification performance of LDA and PNN, respectively, if the classification results are expressed as percentage of success. Classification results are presented for the three classes at the same time as well as for one class against the other. As we can see in Tables 9.21 and 9.22, the PNN performs as well as or even better if the PCs extracted from NN-PCA are used as independent variables rather than the PCs extracted from NN-NLPCA. Certainly, this is not what we would expect if we consider that the PVE by the NN-NLPCA architecture is greater than the percentage of variance explained by the NN-PCA architecture. On the other hand, we have to consider that the NN-PCA architecture is a linear dimensionality reduction technique as opposed to the non-linear NN-NLPCA architecture. These results can be justified, however, if we consider that we have used a particular non-linear classifier that has certain approximation properties.

In order to verify the above results, we used the BPNN architecture that has more universal approximation properties. Tables 9.23, 9.24, 9.25 and 9.26 show the classification performance of PNN and BPNN using the PCs extracted from NN-PCA and NN-NLPCA to implement the models. These results were based on a five-fold rotation method. According to this method, we omitted 20% of the patterns from the training data, we trained the networks using the remaining 80% of the patterns, and then we tested them using the omitted patterns. Repeating this procedure five times, all observations are tested out-of-sample. As we can see in Tables 9.24 and 9.26, the results obtained from the PNN model are similar to the results presented in Tables 9.18 and 9.20. The PNN performs better if the PCs extracted from NN-PCA are used as independent variables rather than the PCs extracted from NN-NLPCA. The BPNN, on the other hand, performs better if the PCs extracted from NN-NLPCA are used as independent variables rather than the PCs extracted from NN-PCA as we can see in Tables 9.23 and 9.25.

A comparison of the results presented in Tables 9.23 and 9.24 indicates that the PNN has better classification performance on average (78.33% and 74.16% for NN-PCA and NL-NLPCA, respectively) than the BPNN (65.83% and 67.5% for NN-PCA and NN-NLPCA, respectively) if we classify issuers in three classes of ratings at the same time and we use either the PCs extracted from NN-PCA or the PCs extracted from NN-NLPCA as input variables to implement the models. On the other hand, if we compare the results presented in Tables 9.25 and 9.26, we can observe that that the PNN has a better classification performance (39 errors on average) than the BPNN (46 errors on average) if we classify issuers in a class one class against the other and we use the PCs extracted from NN-PCA to implement the models. On the other hand, the BPNN has better classification performance (37 errors on average) than the PNN (48 errors on average) if we classify issuers in a class one class against the other and we use the PCs extracted from NN-NLPCA to implement the models.

9.9 SUMMARY OF THE RESULTS

In this part, we applied neural network architectures of both PCA and NLPCA to homogeneous subsets of financial ratios and used the derived PCs to assess the long-term credit standing of U.K. debt issuers. After experimentation, we found that both NN-PCA and NN-NLPCA explain a higher amount of variation in the original set of variables than the common PCA methodology and the derived PCs are easier to interpret if extracted from homogeneous groups of financial ratios. Although the NN-NLPCA explains a higher proportion of variation in the original set of variables than the NN-PCA, the PCs extracted from NN-PCA discriminate better than the PCs extracted from NN-NLPCA when the PNN model is applied to classify debt issuers into

boundary rating classes. On the other hand, the PCs extracted from NN-NLPCA discriminate better than the PCs extracted from NN-PCA when the BPNN model is applied. In terms of classification accuracy, however, the PNN model performs better than the BPNN. Overall, the results of this experiment show that NN-PCA and NN-NLPCA architectures can be successfully implemented as a preliminary step to assess the credibility of U.K. debt issuers and at the same time provide an alternative solution to overfitting.

Part Three: Summary and Conclusions

9.10 DISCUSSION AND REMARKS

In this Chapter, we investigated alternative methodologies to reduce the dimensionality of our data in ways other than ad hoc stepwise variable elimination procedures. The alternative methods we investigated were PCA as well as linear and non-linear dimensionality reduction techniques based on neural networks.

The attractiveness of PCA is that the model parameters can be computed directly by diagonalizing the sample covariance matrix of the data set. On the other hand, PCA is the optimal linear technique in terms of mean square error when compressing a set of high dimensional vectors into lower dimensions and then decompressing. Compressing and decompressing are easy to perform because they require only matrix multiplications given the model parameters. Despite its attractiveness, however, PCA has several disadvantages. First, the method might not be appropriate for high dimensional data. For example, if we try to diagonalize a sample covariance matrix of n data vectors in a space of p dimensions when n and m are several hundreds or thousands then difficulties can arise in the form of data scarcity because we will not have enough data in high dimensions for the sample covariance matrix to be a full rank. On the other hand, direct diagonalization of a symmetric matrix thousands of rows in size can be an extremely costly operation.

PCA can be used as a dimensionality reduction technique within some other type of analysis such as discriminant analysis, cluster analysis, canonical correlation analysis etc. However, this procedure may be unsatisfactory for two reasons: first, the within-group covariance matrix may be different for different groups; and second, there is no guarantee that the separation between groups will be in the direction of the high-variance PCs. Another problem with PCA is that the relative sizes of the elements in a variable weight vector associated with a particular PC indicate the relative contribution of the variable to the variance of the PC. Therefore, the patterns of

variable weights for a particular PC are used to interpret the PC. A problem is identified, however, if more than a few variables have a significant contribution to the variance of a particular PC. In this case, the interpretation of this PC is extremely difficult. Finally, we should consider that PCA is a linear method and most real problems are non-linear. It has been shown that if PCA is applied in non-linear problems, minor components might contain important information. Therefore, if minor components are discarded important information is lost. It is therefore proposed that a NLPCA should be applied to deal with these problems. NLPCA uncovers both linear and non-linear correlations among variables without restriction on the character of the non-linearities presented in the data.

As an alternative to PCA, we applied dimensionality reduction techniques based on neural networks and we examined the ability of the resulting PCs to predict which shares are likely to have exceptional returns in the future by applying five heterogeneous classifiers namely, LDA, PNN, LVQ, OC1, and RRI. We examined the effectiveness of the neural network dimensionality reduction techniques by applying them using the minimum number of available observations and we compared them with ad hoc procedures to identify the best subset for variables that require more observations in order to avoid the problem of overfitting and are more time-consuming than the state-of-the-art methodology. After experimentation, we found that NN-PCA and NN-NLPCA explain a higher proportion of variation in the original set of variables than the common PCA methodology. On the other hand, we found that the resulting PCs from NN-PCA and NN-NLPCA are competitive to other dimensionality reduction techniques in maintaining important discriminating power to identify which shares are likely to have exceptional returns in the future. Furthermore, we found that NN-PCA and NN-NLPCA dimensionality reduction techniques can be used as an alternative to ad hoc methodologies for variable selection because they require less effort and their resulting PCs classify as well as or even better than the optimal subsets of variables that we found after applying stepwise variable elimination procedures. Taken together, the results from this experiment suggest that there are discoverable patterns relating fundamental data to the published financial performance of companies and their subsequent share price performance. But these patterns are non-linear, and are more likely to be detected using non-linear classifiers with non-linear data pre-processing.

In this Chapter, we also applied neural network architectures of both PCA and NLPCA to homogeneous subsets of financial ratios and used the derived PCs to assess the long-term credit standing of U.K. debt issuers. The results of this experiment confirmed the findings from the first experiment. More specifically, we found that both NN-PCA and NN-NLPCA explain a higher amount of variation in the original set of variables than the common PCA methodology and the derived PCs are easier to interpret if extracted from homogeneous groups of financial ratios. Furthermore, the PCs derived from either NN-PCA or NN-NLPCA maintain important

discriminating power and they can be used as input to non-linear classification techniques to predict the long-term credit standing of debt issuers. Taken together, the results of this experiment suggest that NN-PCA and NN-NLPCA architectures can be successfully implemented as a preliminary step to assess the credibility of U.K. debt issuers and at the same time provide an alternative solution to overfitting.

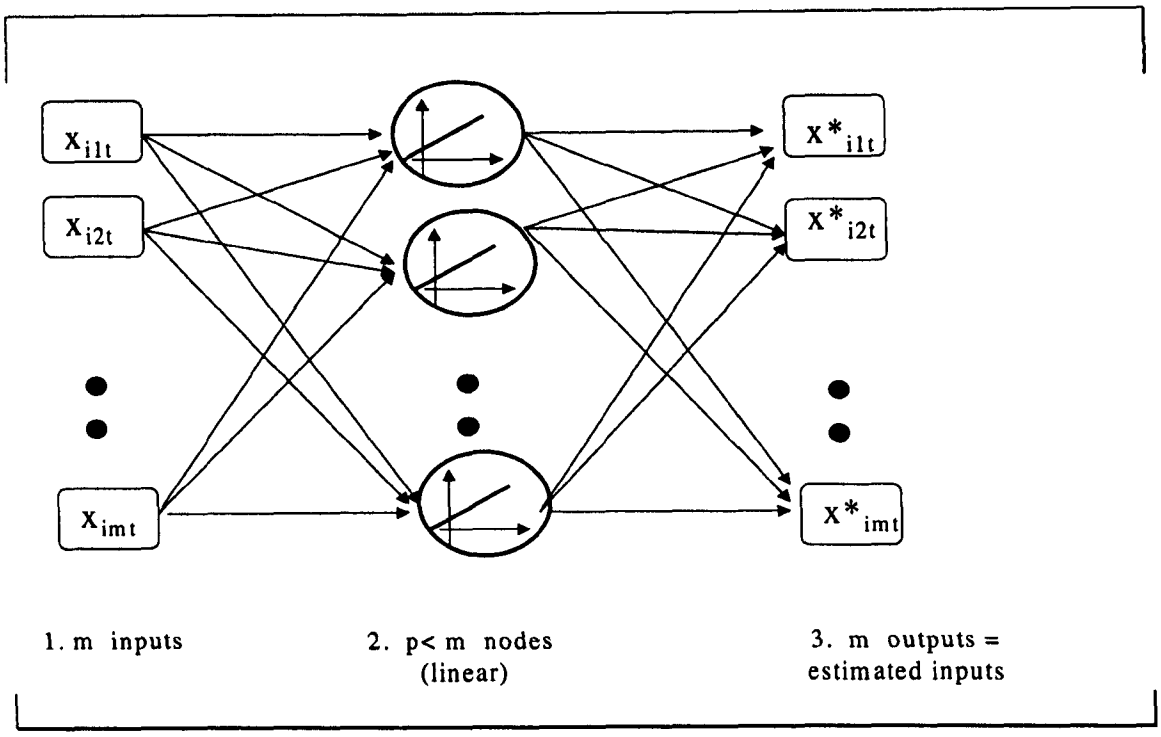


Figure 9.1: Neural network PCA (NN-PCA)

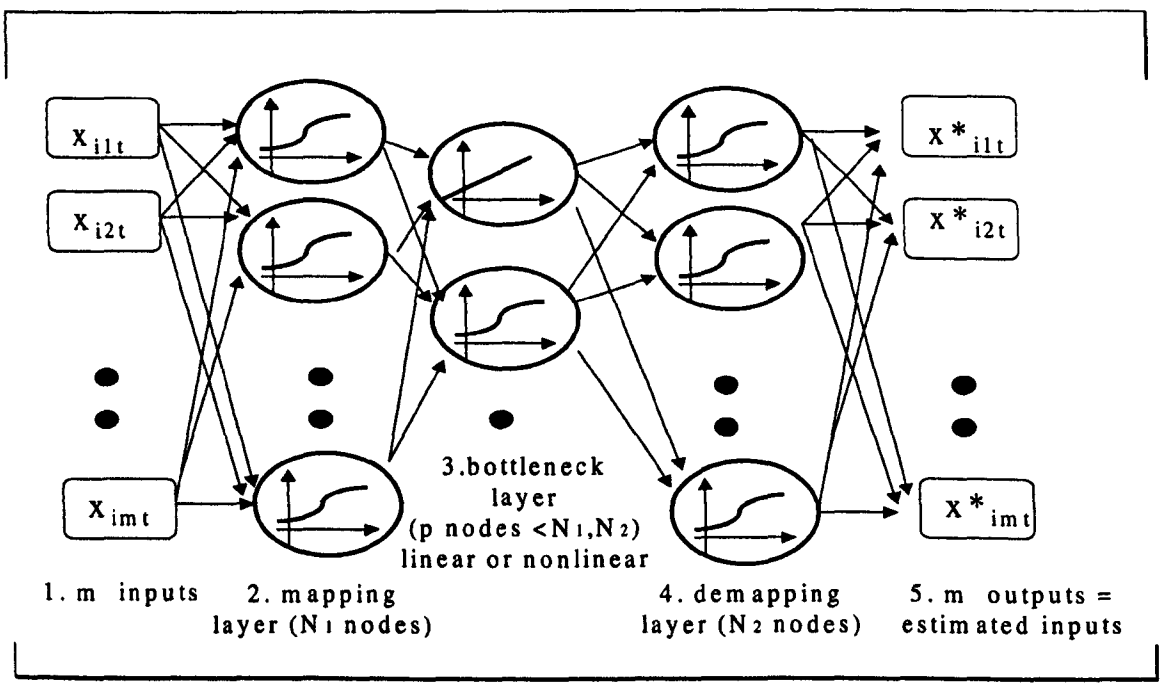


Figure 9.2: Neural network non-linear PCA (NN-NLPCA)

Groups	PERCENTAGE OF VARIANCE EXPLAINED (PVE)				
1993	Initial Dimension	PCs Extracted	PCA	NN-PCA	NN-NLPCA
RCAP	5	1	46.76 %	64.99 %	76.05 %
PROF	6	1	46.37 %	53.56 %	79.75 %
FLEV	6	1	39.96 %	37.90 %	85.24 %
INV	4	1	29.81 %	69.72 %	87.50 %
GRT	6	1	29.34 %	57.58 %	80.78 %
ST-LIQ	3	1	44.75 %	75.20 %	95.90 %
ROI	2	1	88.42 %	94.53 %	99.61 %
EFF	2	1	59.34 %	70.08 %	98.80 %
RISK	4	1	72.66 %	79.51 %	80.82 %
1994	Initial Dimension	PCs Extracted	PCA	NN-PCA	NN-NLPCA
RCAP	5	1	43.12 %	39.43 %	78.22 %
PROF	6	1	51.14 %	52.67 %	78.27 %
FLEV	6	1	42.62 %	41.38 %	81.47 %
INV	4	1	30.69 %	62.96 %	85.85 %
GRT	6	1	28.79 %	45.67 %	80.82 %
ST-LIQ	3	1	44.32 %	76.69 %	95.08 %
ROI	2	1	99.92 %	99.93 %	99.93 %
EFF	2	1	59.43 %	69.94 %	98.74 %
RISK	4	1	74.55 %	90.43 %	94.95 %
1995	Initial Dimension	PCs Extracted	PCA	NN-PCA	NN-NLPCA
RCAP	5	1	40.79 %	47.18 %	87.30 %
PROF	6	1	52.03 %	56.36 %	95.27 %
FLEV	6	1	36.78 %	62.11 %	71.63 %
INV	4	1	29.00 %	83.04 %	90.09 %
GRT	6	1	27.24 %	45.82 %	78.27 %
ST-LIQ	3	1	41.83 %	84.69 %	95.80 %
ROI	2	1	99.85 %	99.85 %	99.85 %
EFF	2	1	59.14 %	76.71 %	99.02 %
RISK	4	1	74.10 %	85.68 %	91.18 %
1996	Initial Dimension	PCs Extracted	PCA	NN-PCA	NN-NLPCA
RCAP	5	1	41.78 %	44.91 %	88.04 %
PROF	6	1	50.94 %	48.83 %	94.79 %
FLEV	6	1	42.78 %	43.89 %	92.14 %
INV	4	1	27.56 %	31.84 %	72.46 %
GRT	6	1	34.15 %	39.95 %	75.71 %
ST-LIQ	3	1	41.97 %	83.18 %	97.12 %
ROI	2	1	99.86 %	99.86 %	99.86 %
EFF	2	1	58.10 %	80.05 %	99.27 %
RISK	4	1	74.13 %	73.47 %	84.18 %
1997	Initial Dimension	PCs Extracted	PCA	NN-PCA	NN-NLPCA
RCAP	5	1	58.15 %	54.89 %	77.71 %
PROF	6	1	50.23 %	51.61 %	80.07 %
FLEV	6	1	41.98 %	79.00 %	95.74 %
INV	4	1	28.11 %	37.42 %	76.24 %
GRT	6	1	37.34 %	56.41 %	76.05 %
ST-LIQ	3	1	45.18 %	87.29 %	97.37 %
ROI	2	1	99.97 %	99.97 %	99.97 %
EFF	2	1	57.90 %	82.81 %	99.45 %
RISK	4	1	75.07 %	92.75 %	97.80 %
AVERAGE PVE FOR THE WHOLE PERIOD 1993-97					
RCAP	5	1	46.12%	50.28%	81.46%
PROF	6	1	50.14%	52.61%	85.63%
FLEV	6	1	40.82%	52.86%	85.24%
INV	4	1	29.03%	57.00%	82.43%
GRT	6	1	31.37%	49.09%	78.33%
ST-LIQ	3	1	43.61%	81.41%	96.25%
ROI	2	1	97.60%	98.83%	99.84%
EFF	2	1	58.78%	75.92%	99.06%
RISK	4	1	74.10%	84.37%	89.79%

Table 9.1: Percentage of variance explained (PVE) by PCA, NN-PCA, and NN-NLPCA after conceptual clustering of the accounting variables that we selected to predict high and low performing shares

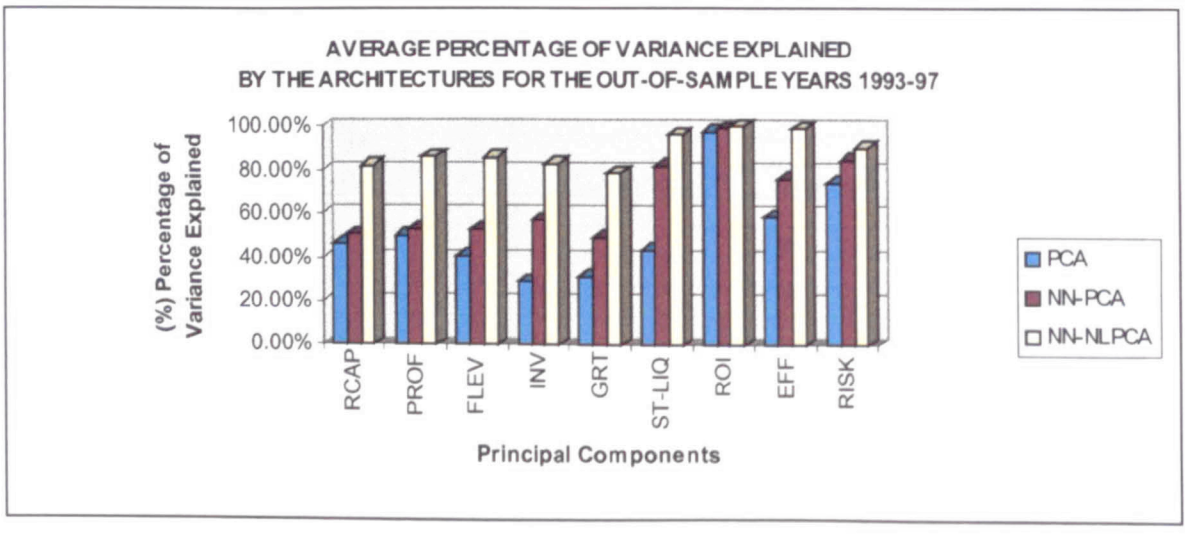


Figure 9.3: Average PVE by PCA, NN-PCA, and NN-NLPCA after conceptual clustering of the accounting variables that we selected to predict high and low performing shares

	n	PCA	NN-PCA	NN-NLPCA	Z_1	p-value	Z_2	p-value
RCAP	1378	46.12%	50.28%	81.46%	2.19 *	.0143	18.28 *	.0000
PROF	1378	50.14%	52.61%	85.63%	1.30	.0968	20.09 *	.0000
FLEV	1378	40.82%	52.86%	85.24%	6.38 *	.0000	19.63 *	.0000
INV	1378	29.03%	57.00%	82.43%	15.46 *	.0000	15.12 *	.0000
GRT	1378	31.37%	49.09%	78.33%	9.64 *	.0000	16.76 *	.0000
ST-LIQ	1378	43.61%	81.41%	96.25%	22.26 *	.0000	12.72 *	.0000
ROI	1378	97.60%	98.83%	99.84%	2.44 *	.0073	3.27 *	.0007
EFF	1378	58.78%	75.92%	99.06%	9.76 *	.0000	19.60 *	.0000
RISK	1378	74.10%	84.37%	89.79%	6.70 *	.0000	4.26 *	.0000

Note: * denotes significance at the 5% level

Table 9.2: The one-tail $Z-\text{Test}$ for differences between proportions

LDA		All Variables		12 Variables		PCA		NN-PCA		NN-NLPCA	
Actual Class	Patterns	Predicted Class Membership									
1993		H	L	H	L	H	L	H	L	H	L
H	163	107	56	98	65	97	66	102	61	105	58
L	488	176	312	128	360	145	343	152	336	153	335
Total (%)		64.36 %		70.35 %		67.59 %		67.28 %		67.59 %	
1994		H	L	H	L	H	L	H	L	H	L
H	163	89	74	92	71	88	75	87	76	88	75
L	488	212	276	212	276	242	246	212	276	214	274
Total (%)		56.07 %		56.53 %		51.31 %		55.76 %		55.61 %	
1995		H	L	H	L	H	L	H	L	H	L
H	173	109	64	106	67	85	88	107	66	108	65
L	519	211	308	208	311	295	224	200	319	209	310
Total (%)		60.26 %		60.26 %		44.65 %		61.56 %		60.40 %	
1996		H	L	H	L	H	L	H	L	H	L
H	188	107	81	106	82	101	87	103	85	100	88
L	561	278	283	262	299	241	320	255	306	257	304
		52.07 %		54.07 %		56.21 %		54.61 %		53.94 %	
1997		H	L	H	L	H	L	H	L	H	L
H	188	109	79	105	83	103	85	108	80	107	81
L	564	255	309	214	350	205	359	224	340	229	335
Total (%)		55.59 %		60.51 %		61.44 %		59.57 %		58.78 %	

Table 9.3: Out-of-sample classification performance of LDA for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

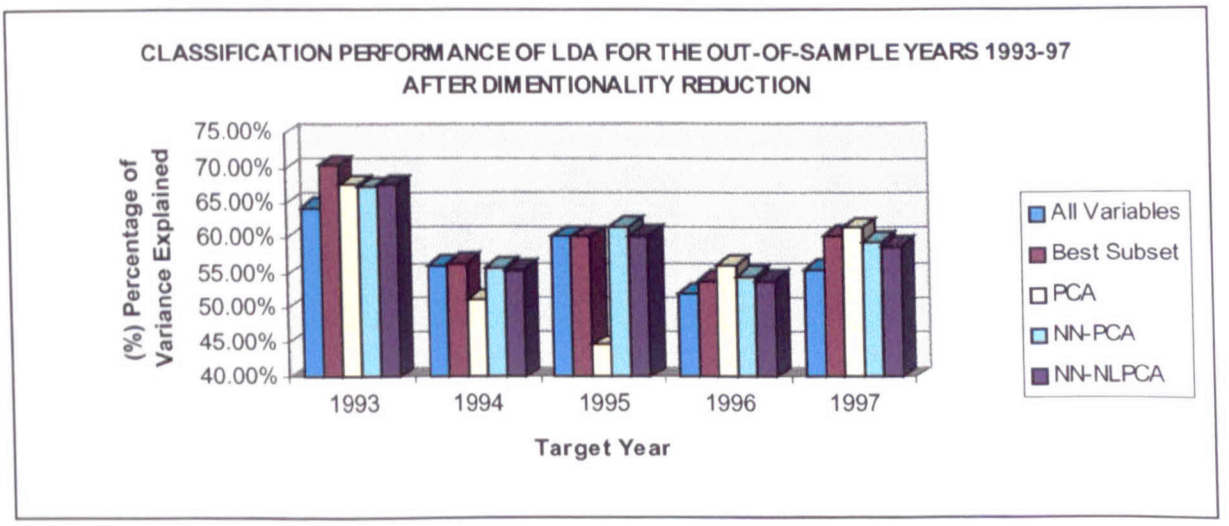


Figure 9.4: Out-of-sample classification performance of LDA for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

PNN		All Variables		17 Variables		PCA		NN-PCA		NN-NLCA	
Actual Class	Patterns	Predicted Class Membership									
1993		H	L	H	L	H	L	H	L	H	L
H	163	94	69	96	67	90	73	93	70	93	70
L	488	138	350	123	365	136	352	130	358	138	350
Total (%)		68.20 %		70.81 %		67.90 %		69.28 %		68.05 %	
1994		H	L	H	L	H	L	H	L	H	L
H	163	83	80	84	79	89	74	83	80	83	80
L	488	221	267	199	289	221	267	218	270	229	259
Total (%)		53.76 %		57.30 %		54.69 %		54.22 %		52.53 %	
1995		H	L	H	L	H	L	H	L	H	L
H	173	104	69	103	70	84	89	106	67	96	77
L	519	192	327	193	326	264	255	187	332	182	337
Total (%)		62.28 %		61.99 %		48.99 %		63.29 %		62.57 %	
1996		H	L	H	L	H	L	H	L	H	L
H	188	99	89	100	88	98	90	96	92	96	92
L	561	223	338	254	307	238	323	236	325	236	325
		58.34 %		54.34 %		56.21 %		56.21 %		56.21 %	
1997		H	L	H	L	H	L	H	L	H	L
H	188	102	86	103	85	102	86	102	86	102	86
L	564	192	372	199	365	201	363	203	361	212	352
Total (%)		63.03 %		62.23 %		61.84 %		61.57 %		60.37 %	

Table 9.4: Out-of-sample classification performance of PNN for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

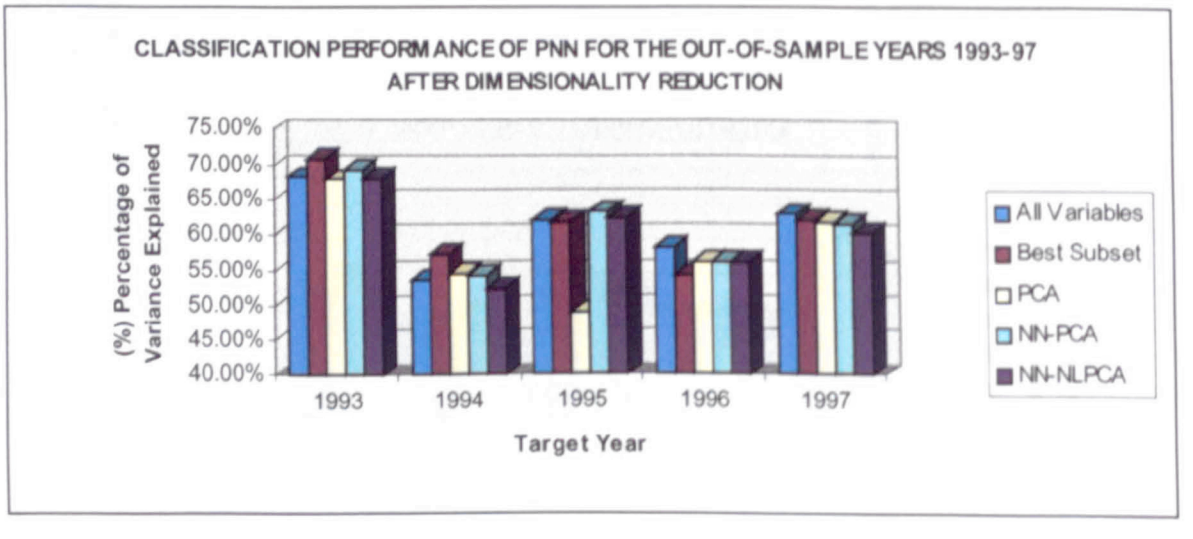


Figure 9.5: Out-of-sample classification performance of PNN for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

LVQ		All Variables		19 Variables		PCA		NN-PCA		NN-NLCA	
Actual Class	Patterns	Predicted Class Membership									
1993		H	L	H	L	H	L	H	L	H	L
H	163	84	79	90	73	91	72	88	75	90	73
L	488	210	278	154	334	233	255	205	283	187	301
Total (%)		55.61 %		65.13 %		53.15 %		56.99 %		60.06 %	
1994		H	L	H	L	H	L	H	L	H	L
H	163	87	76	84	79	82	81	83	80	84	79
L	488	227	261	206	282	218	270	219	269	181	307
Total (%)		53.46 %		56.22 %		54.07 %		54.07 %		60.06 %	
1995		H	L	H	L	H	L	H	L	H	L
H	173	98	75	90	83	98	75	89	84	85	88
L	519	239	280	188	331	231	288	227	292	155	364
Total (%)		54.62 %		60.84 %		55.78 %		55.06 %		64.88 %	
1996		H	L	H	L	H	L	H	L	H	L
H	188	99	89	105	83	101	87	99	89	96	92
L	561	221	340	216	345	229	332	207	354	207	354
		58.61 %		60.08 %		57.81 %		60.48 %		60.08 %	
1997		H	L	H	L	H	L	H	L	H	L
H	188	106	82	100	88	118	70	95	93	103	85
L	564	256	308	203	361	281	283	257	307	201	363
Total (%)		55.05 %		61.30 %		53.32 %		53.46 %		61.97 %	

Table 9.5: Out-of-sample classification performance of LVQ for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

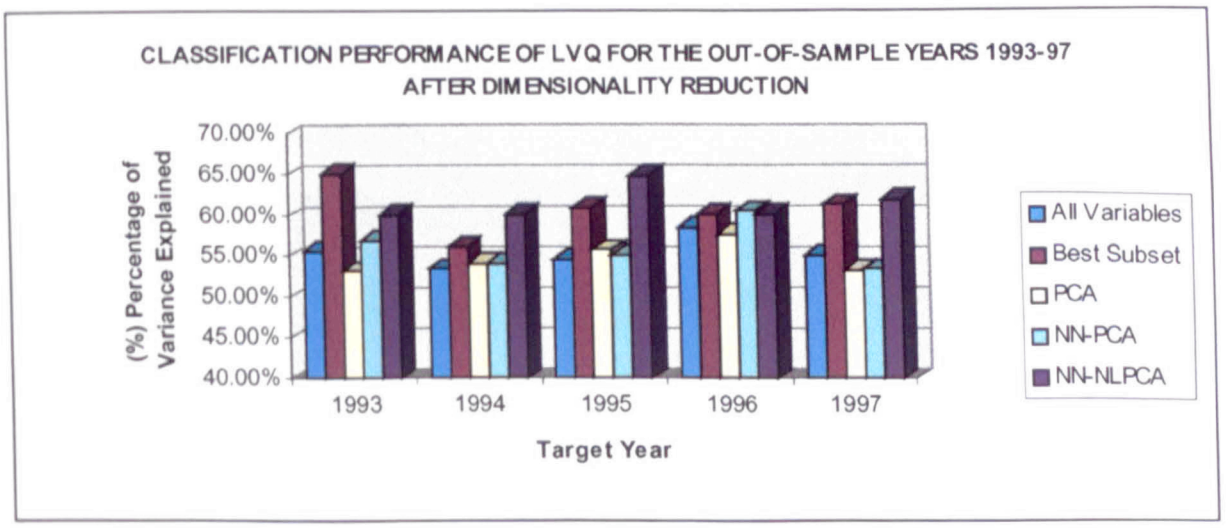


Figure 9.6: Out-of-sample classification performance of LVQ for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

OC1		All Variables		21 Variables		PCA		NN-PCA		NN-NLCA	
Actual Class	Patterns	Predicted Class Membership									
1993		H	L	H	L	H	L	H	L	H	L
H	163	92	71	88	75	98	65	87	76	95	68
L	488	227	261	163	325	236	252	241	247	174	314
Total (%)		54.22 %		63.44 %		53.76 %		51.31 %		62.83 %	
1994		H	L	H	L	H	L	H	L	H	L
H	163	83	80	99	64	78	85	77	86	83	80
L	488	200	288	193	295	229	259	223	265	196	292
Total (%)		56.99 %		60.52 %		51.77 %		52.53 %		57.60 %	
1995		H	L	H	L	H	L	H	L	H	L
H	173	87	86	98	75	95	78	87	86	97	56
L	519	228	291	183	336	220	299	254	265	214	305
Total (%)		54.62 %		62.72 %		56.99 %		50.87 %		58.09 %	
1996		H	L	H	L	H	L	H	L	H	L
H	188	106	82	96	92	99	89	106	82	100	88
L	561	279	282	216	345	233	328	264	297	222	339
		51.80 %		58.88 %		57.01 %		53.81 %		58.61 %	
1997		H	L	H	L	H	L	H	L	H	L
H	188	111	77	100	88	114	74	97	91	109	79
L	564	275	289	212	352	246	318	251	313	221	343
Total (%)		53.19 %		60.11 %		57.45 %		54.52 %		60.11 %	

Table 9.6: Out-of-sample classification performance of OC1 for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

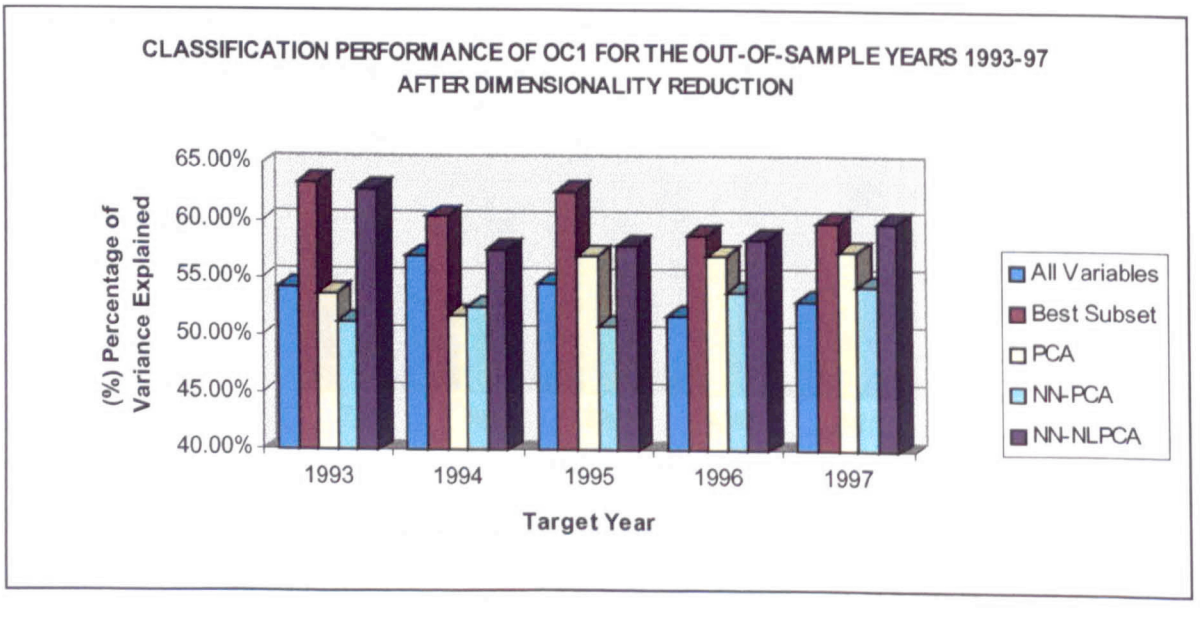


Figure 9.7: Out-of-sample classification performance of OC1 for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

RRI		All Variables		35 Variables		PCA		NN-PCA		NN-NLCA	
Actual Class	Patterns	Predicted Class Membership									
1993		H	L	H	L	H	L	H	L	H	L
H	163	93	70	91	72	90	73	95	68	88	75
L	488	189	299	149	339	216	272	188	300	176	312
Total (%)		60.22 %		66.05 %		55.61 %		60.68 %		61.44 %	
1994		H	L	H	L	H	L	H	L	H	L
H	163	88	75	94	69	90	73	88	75	89	74
L	488	197	291	233	255	226	262	217	271	228	260
Total (%)		58.22 %		53.61 %		54.07 %		55.15 %		53.61 %	
1995		H	L	H	L	H	L	H	L	H	L
H	173	97	76	99	74	104	69	105	68	95	78
L	519	225	294	233	286	212	307	251	268	246	273
Total (%)		56.50%		55.64 %		59.39 %		53.90 %		53.18 %	
1996		H	L	H	L	H	L	H	L	H	L
H	188	99	89	113	75	97	91	101	87	97	91
L	561	261	300	253	308	246	315	229	332	226	335
		53.27 %		56.21 %		55.01 %		57.81 %		57.68 %	
1997		H	L	H	L	H	L	H	L	H	L
H	188	102	86	96	92	101	87	104	84	99	89
L	564	226	338	220	344	208	356	246	318	248	316
Total (%)		58.51 %		58.51 %		60.77 %		56.12 %		55.19 %	

Table 9.7: Out-of-sample classification performance of RRI for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

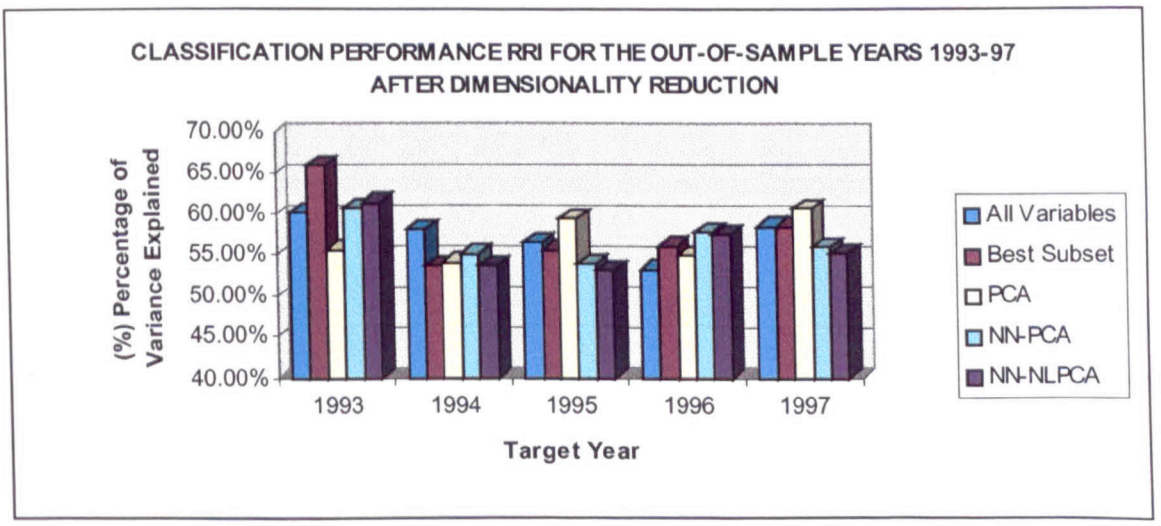


Figure 9.8: Out-of-sample classification performance of RRI for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

	All Variables	Best Subset	PCA	NN-PCA	NN-NLPCA
LDA	57.67 %	60.34%	56.24%	59.76%	59.26%
PNN	61.12%	61.33%	57.93%	60.91%	59.95%
LVQ	54.16%	61.13%	55.40%	52.61%	59.45%
OC1	55.47%	60.71%	54.83%	56.01%	61.41%
RRI	57.34%	58.00%	56.97%	56.73%	56.22%

Table 9.8: Out-of-sample average classification results of LDA, PNN, LVQ, OC1 and RRI for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

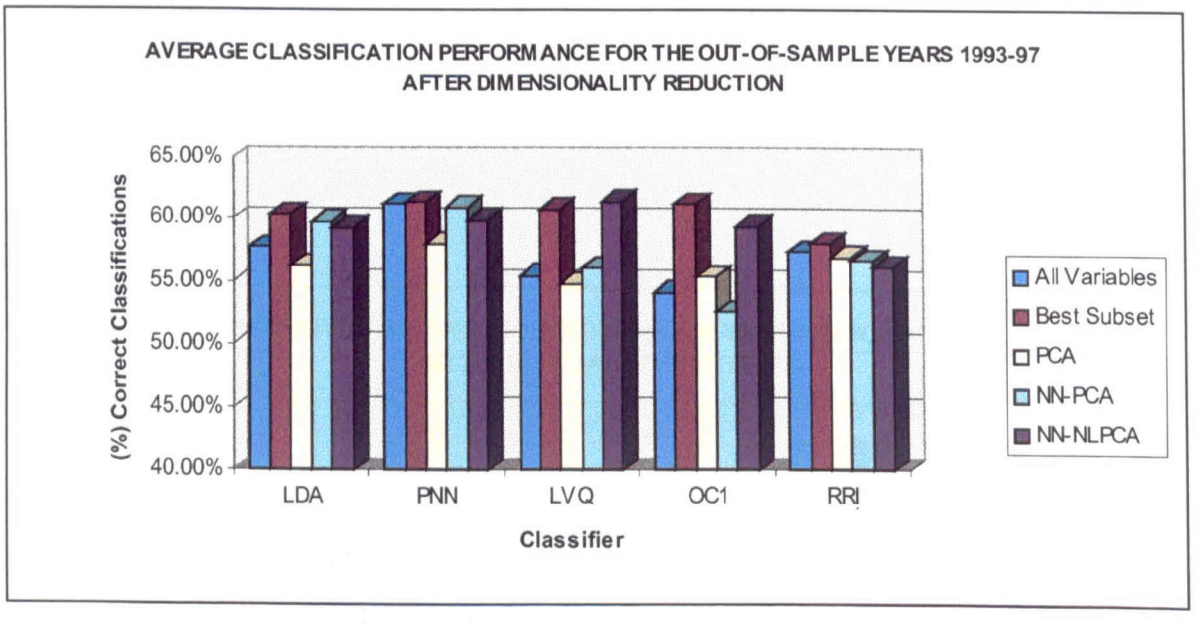


Figure 9.9: Out-of-sample average classification results of LDA, PNN, LVQ, OC1 and RRI for 1993-97 after applying different dimensionality reduction techniques to predict high and low performing shares

		Actual Returns		PREDICTED RETURNS FOR THE TARGET YEARS 1993 –97					
1993					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	89.98	High	44.36	47.62	34	33.15	37.36
		Low	9.22	Low	17.97	19.38	25.68	25.88	23.39
	Best Subset	High	89.98	High	50.86	51.77	42.49	35.87	43.31
		Low	9.22	Low	18.06	18.12	21.62	25.41	21.34
	PCA	High	89.98	High	48.17	45.82	35.14	34.87	36.42
		Low	9.22	Low	18.36	20.73	23.8	23.73	23.25
	NN-PCA	High	89.98	High	47.36	49.2	31.32	30.3	38.94
		Low	9.22	Low	17.98	19.15	27.9	28.57	22.14
	NN-NLPCA	High	89.98	High	47.68	48.06	39.37	41.03	42.06
		Low	9.22	Low	17.47	19.2	22.09	21.28	20.84
1994					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	45.48	High	6.73	7.5	6.28	9.13	9.67
		Low	-7.61	Low	4.76	4.07	5.11	3.01	2.56
	Best Subset	High	45.48	High	8.05	9.06	6.62	10.5	8.17
		Low	-7.61	Low	3.59	3.07	4.91	1.75	3.15
	PCA	High	45.48	High	5.46	8.76	9.59	7.37	8.52
		Low	-7.61	Low	5.89	2.87	2.33	4.16	2.99
	NN-PCA	High	45.48	High	6.67	8.56	9.62	6.04	8.43
		Low	-7.61	Low	4.83	3.19	2.26	5.36	3.24
	NN-NLPCA	High	45.48	High	7.11	6.96	10.23	7.97	6.83
		Low	-7.61	Low	4.43	4.49	2.54	3.95	4.57
1995					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	79.78	High	36.97	39.77	29.48	28.64	31.03
		Low	7.12	Low	15.24	14.46	21.31	22.49	20.29
	Best Subset	High	79.78	High	37.81	38.7	36.04	35.4	29.93
		Low	7.12	Low	14.89	15.26	18.06	18.37	21.01
	PCA	High	79.78	High	23.79	26.39	33.48	31.69	34.36
		Low	7.12	Low	27.11	24.17	17.86	19.94	17.66
	NN-PCA	High	79.78	High	37.61	36.89	28.76	22.41	31.6
		Low	7.12	Low	15.46	16.77	22.37	28.09	18.6
	NN-NLPCA	High	79.78	High	34.5	35.49	41.78	32.93	28.64
		Low	7.12	Low	17.5	18.44	16.53	19.05	22.03
1996					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	54.45	High	12.05	15.93	14	11.05	13.09
		Low	-5.34	Low	7.12	4.92	6.41	8.18	6.48
	Best Subset	High	54.45	High	13.15	13.21	16.35	15.32	14.16
		Low	-5.34	Low	6.28	6.47	4.63	5.61	5.36
	PCA	High	54.45	High	14.11	13.94	15.71	14.07	11.16
		Low	-5.34	Low	5.91	6.17	4.89	6.14	8.38
	NN-PCA	High	54.45	High	12.7	13.32	15.71	13.34	12.09
		Low	-5.34	Low	6.87	6.64	5.47	6.06	7.74
	NN-NLPCA	High	54.45	High	13.23	13.83	16.24	16.47	12.8
		Low	-5.34	Low	6.4	6.34	5.18	4.51	7.27
1997					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	58.44	High	13.07	17.81	11.13	12.8	15.74
		Low	-7.23	Low	5.53	3.64	7.36	5.36	4.1
	Best Subset	High	58.44	High	16.29	17	17.13	15.76	12
		Low	-7.23	Low	3.94	3.93	3.81	4.51	7.14
	PCA	High	58.44	High	16.84	16.66	13.67	15.98	17.27
		Low	-7.23	Low	3.87	4.134	4.1	2.94	3.53
	NN-PCA	High	58.44	High	15.94	16.54	10.14	9.22	14.41
		Low	-7.23	Low	3.83	4.15	8.34	9.14	4.63
	NN-NLPCA	High	58.44	High	15.62	16.88	18.17	18.12	13.55
		Low	-7.23	Low	3.98	3.66	3.08	2.18	5.43
AVERAGE PREDICTED RETURNS FOR THE WHOLE PERIOD 1993-97									
		Actual Returns		Predicted Returns					
1993-97					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	65.62	High	22.63	25.72	18.97	18.95	21.37
		Low	-0.76	Low	10.12	9.29	13.17	12.98	11.36
	Best Subset	High	65.62	High	25.23	25.94	23.72	22.57	21.51
		Low	-0.76	Low	9.35	9.37	10.60	11.13	11.6
	PCA	High	65.62	High	21.67	22.31	21.51	20.79	21.54
		Low	-0.76	Low	12.22	11.61	10.59	11.38	11.16
	NN-PCA	High	65.62	High	24.05	24.90	19.11	16.26	21.09
		Low	-0.76	Low	9.79	9.98	13.26	15.44	11.27
	NN-NLPCA	High	65.62	High	23.62	24.24	25.15	23.30	20.77
		Low	-0.76	Low	9.95	10.42	9.88	10.19	12.02

Table 9.9: Out-of-sample high and low returns predicted by LDA, PNN, LVQ, OC1, and RRI for 1993-97 after applying different dimensionality reduction techniques

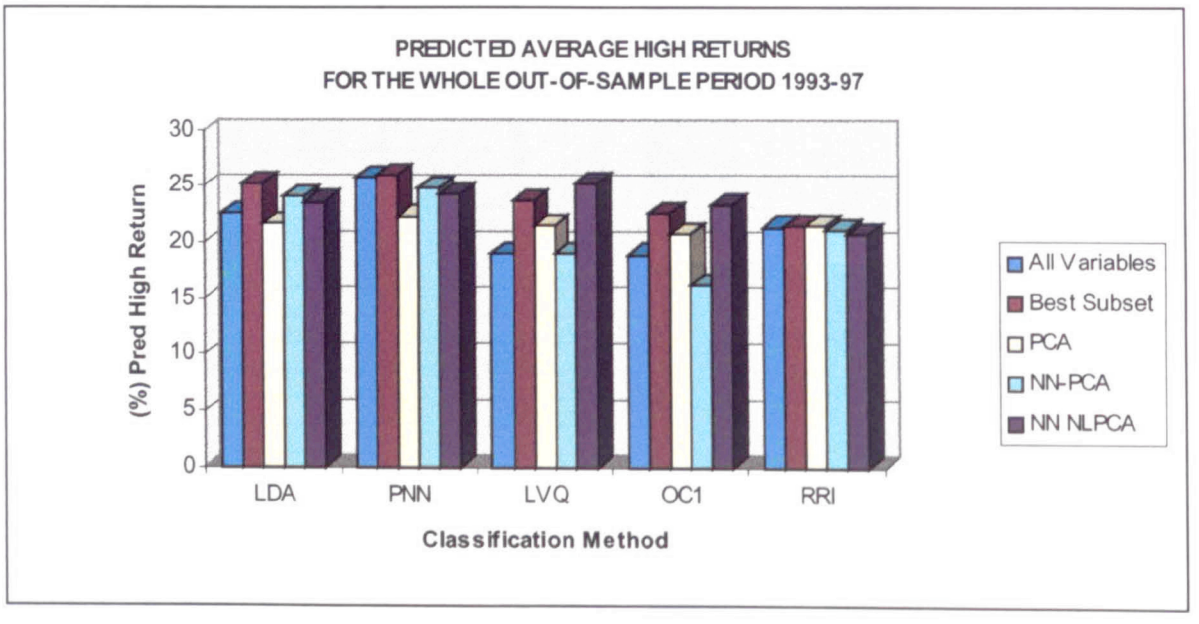


Figure 9.10: Out-of-sample average high returns predicted by LDA, PNN, LVQ, OC1, and RRI for 1993-97 after applying different dimensionality reduction techniques

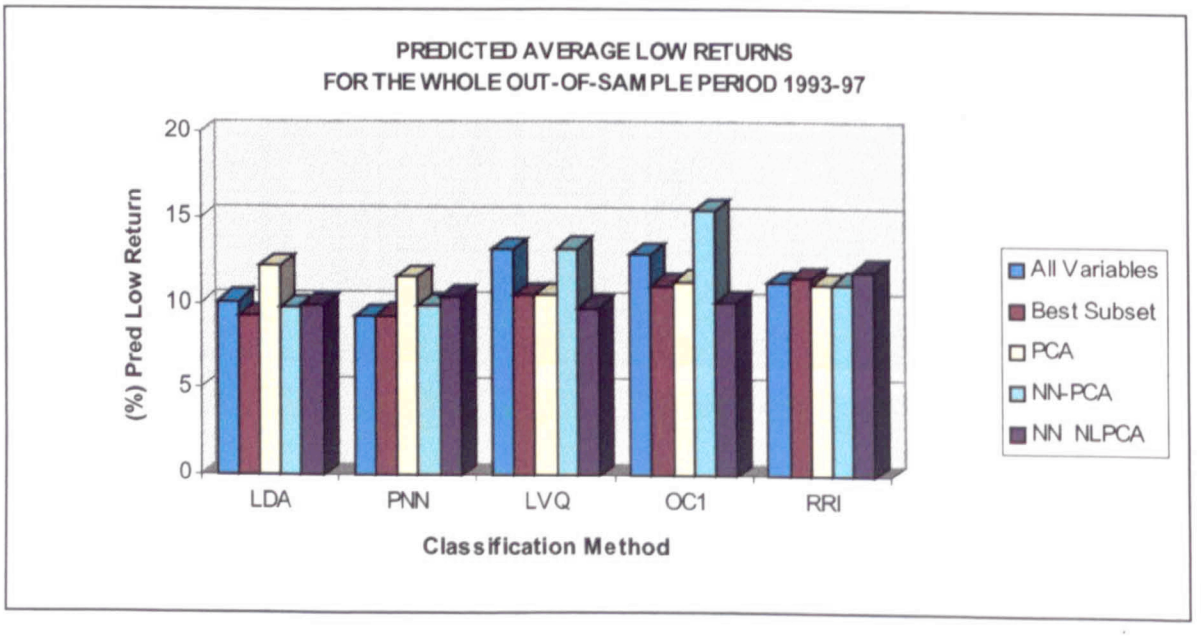


Figure 9.11: Out-of-sample average low returns predicted by LDA, PNN, LVQ, OC1, and RRI for 1993-97 after applying different dimensionality reduction techniques

		Actual Excess Returns		PREDICTED EXCESS RETURNS FOR THE YEARS 1993 -97					
1993					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	58.65	High	13.73	16.93	4.97	2.98	8.17
		Low	-19.53	Low	-10.48	-9.3	-4.01	-2.77	-6.17
	Best Subset	High	58.65	High	20.4	20.77	11.61	5.46	12.94
		Low	-19.53	Low	-10.78	-10.46	-6.89	-3.35	-7.49
	PCA	High	58.65	High	16.66	15	5.13	4.67	5.94
		Low	-19.53	Low	-9.79	-7.9	-4.99	-4.83	-5.19
	NN-PCA	High	58.65	High	15.53	17.6	2.09	0.58	7.94
		Low	-19.53	Low	-9.86	-9.1	-1.63	-0.5	-6.03
	NN-NLPCA	High	58.65	High	16.03	16.79	8.64	10.53	11.45
		Low	-19.53	Low	-10.45	-9.17	-6.32	-7.34	-7.74
1994					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	39.67	High	1.98	2.85	0.29	3.42	3.12
		Low	-13.24	Low	-1.7	-2.5	-0.27	-2.62	-2.42
	Best Subset	High	39.67	High	3.08	4.37	1.77	5.54	1.73
		Low	-13.24	Low	-2.7	-3.36	-1.42	-4.5	-1.75
	PCA	High	39.67	High	0.8	4.58	3.42	1.17	3.4
		Low	-13.24	Low	-0.82	-4.16	-2.92	-1.04	-3.21
	NN-PCA	High	39.67	High	1.68	3.86	3.66	0.37	3.66
		Low	-13.24	Low	-1.42	-3.31	-3.16	-0.31	-3.22
	NN-NLPCA	High	39.67	High	2.24	2.3	5.65	2.95	2.2
		Low	-13.24	Low	-1.94	-2.12	-3.87	-2.21	-2.09
1995					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	54.89	High	11.72	14.14	4.18	2.99	5.73
		Low	-18.3	Low	-10.09	-10.58	-3.98	-2.51	-5
	Best Subset	High	54.89	High	12.61	13.27	10.86	10.17	4.67
		Low	-18.3	Low	-10.49	-9.93	-7.3	-6.96	-4.32
	PCA	High	54.89	High	-2.15	0.21	7.97	6.39	9.17
		Low	-18.3	Low	2.6	-0.23	-7.23	-5.35	-7.72
	NN-PCA	High	54.89	High	12.29	11.51	3.61	-2.55	6.3
		Low	-18.3	Low	-9.81	-8.46	-3.04	2.47	-6.69
	NN-NLPCA	High	54.89	High	9.09	10.08	16.49	7.25	3.29
		Low	-18.3	Low	-7.7	-6.78	-8.76	-5.93	-3.21
1996					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	44.84	High	2.42	6.14	4.08	1.83	3.22
		Low	-15.1	Low	-2.69	-4.74	-3.15	-2.06	-3.1
	Best Subset	High	44.84	High	3.41	3.53	6.8	5.58	4.73
		Low	-15.1	Low	-3.41	-3.28	-5.21	-4.09	-4.64
	PCA	High	44.84	High	4.21	4.04	5.96	4.53	1.14
		Low	-15.1	Low	-3.65	-3.4	-4.8	-3.72	-1.07
	NN-PCA	High	44.84	High	3.03	3.48	6	3.88	2.27
		Low	-15.1	Low	-2.89	-2.97	-4.24	-3.91	-1.9
	NN-NLPCA	High	44.84	High	3.44	4.22	6.47	6.79	2.99
		Low	-15.1	Low	-3.25	-3.47	-4.5	-5.22	-2.37
1997					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	49.37	High	3.23	7.93	1.58	3.45	5.81
		Low	-16.54	Low	-3.15	-5.19	-1.59	-3.78	-4.61
	Best Subset	High	49.37	High	6.63	7.25	7	6.29	3.18
		Low	-16.54	Low	-4.99	-4.97	-4.83	-4.57	-2.42
	PCA	High	49.37	High	7.25	7.12	4.25	6.39	8.07
		Low	-16.54	Low	-5.13	-4.91	-4.94	-5.99	-5.74
	NN-PCA	High	49.37	High	6.5	7.17	0.9	0.01	5.24
		Low	-16.54	Low	-5.25	-5	-0.91	-0.13	-4.68
	NN-NLPCA	High	49.37	High	5.88	7.06	8.49	8.1	3.69
		Low	-16.54	Low	-4.87	-5.17	-5.87	-6.45	-3.28
PREDICTED AVERAGE EXCESS RETURNS FOR THE WHOLE PERIOD 1993-1997									
		Actual Excess Returns		Predicted Excess Returns					
1993-97					LDA	PNN	LVQ	OC1	RRI
	All Variables	High	49.48	High	6.616	9.598	3.02	2.934	5.21
		Low	-16.54	Low	-5.622	-6.462	-2.6	-2.75	-4.26
	Best Subset	High	49.48	High	9.226	9.838	7.608	6.608	5.45
		Low	-16.54	Low	-6.474	-6.4	-5.13	-4.69	-4.12
	PCA	High	49.48	High	5.354	6.19	5.346	4.63	5.544
		Low	-16.54	Low	-3.358	-4.12	-4.98	-4.19	-4.59
	NN-PCA	High	49.48	High	7.806	8.724	3.252	0.458	5.082
		Low	-16.54	Low	-5.846	-5.768	-2.6	-0.48	-4.5
	NN-NLPCA	High	49.48	High	7.336	8.09	9.148	7.124	4.724
		Low	-16.54	Low	-5.642	-5.342	-5.86	-5.43	-3.74

Table 9.10: Out-of-sample average high and low excess returns predicted by LDA, PNN, LVQ, OC1, and RRI for 1993-97 after applying different dimensionality reduction techniques

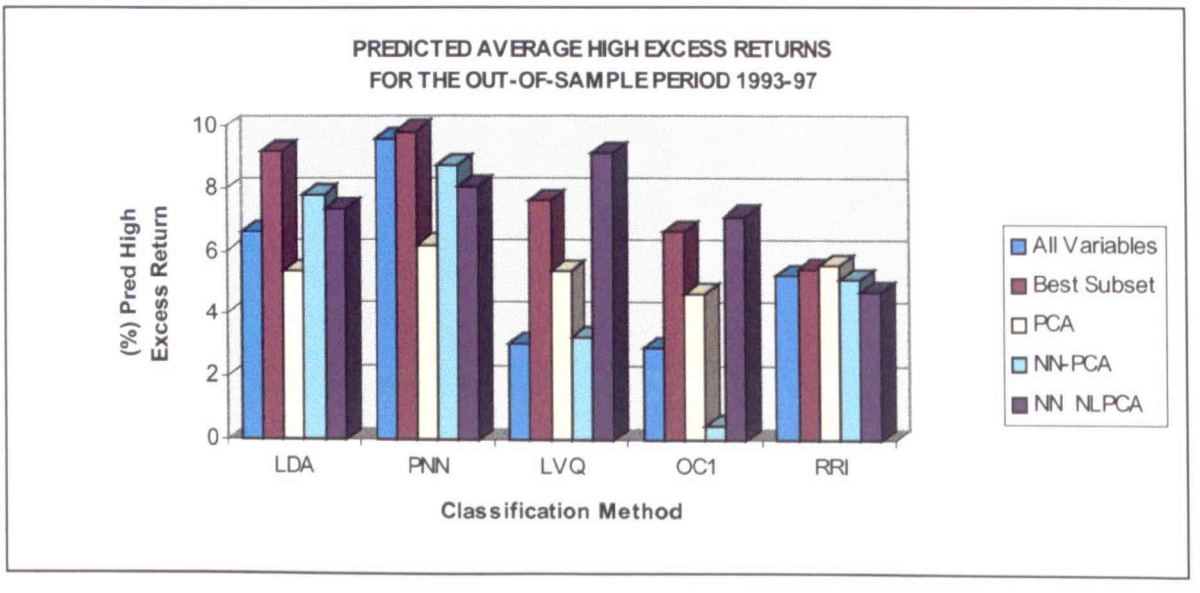


Figure 9.12: Out-of-sample average high excess returns predicted by LDA, PNN, LVQ, OC1, and RRI for 1993-97 after applying different dimensionality reduction techniques

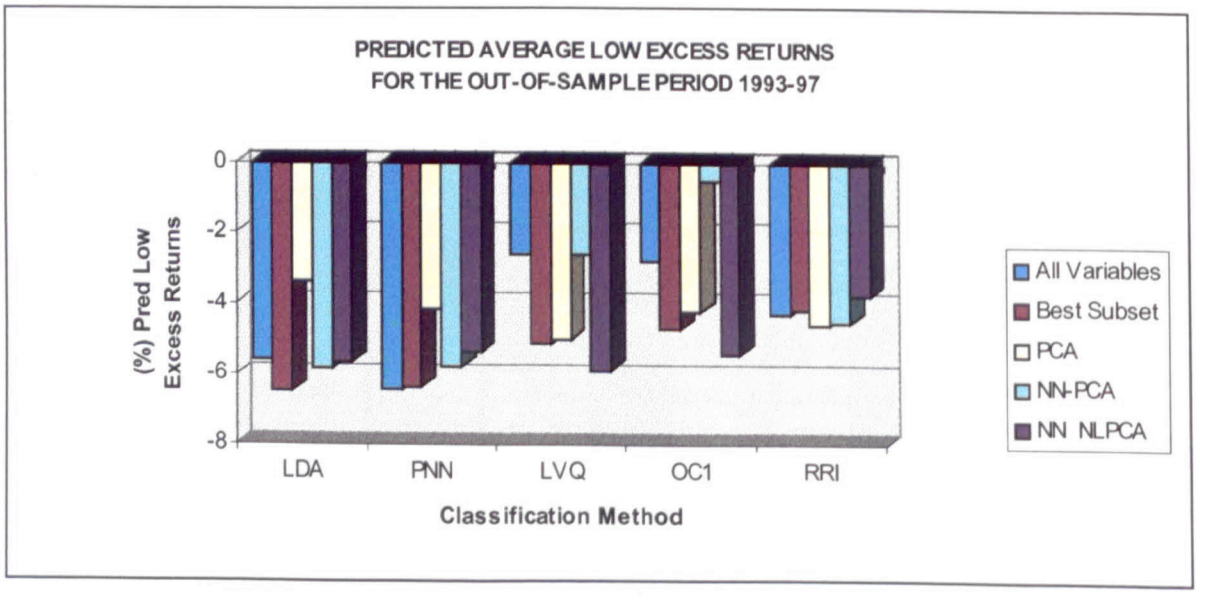


Figure 9.13: Out-of-sample average low excess returns predicted by LDA, PNN, LVQ, OC1, and RRI for 1993-97 after applying different dimensionality reduction techniques

RCAP	ROCE, EBIT/TA, EBIT/TCE, NI/TCE, CF/TA, CF/TCE
PROF	PM, PAT/S, NI/S, CF/S
FLEV	DBT/EQ, DBT/TCAP, DBT/TA
INV	DY, P/E, GEPS, DCOV
GRT	TA, S, PBT
ST-LIQ	CA/CL, CL/TA, CL/SFUNDS, (CA+IS)/IS
ICOV	(EBIT+IS)/IS, (CF+IS)/IS
RISK	PBT/CL, PAT/CL, NI/CL, CF/CL

ROCE: Return on Capital Employed, EBIT: Earnings Before Interest and Taxes, TA: Total Assets, TCE: Total Capital Employed, CF: Cash Flow, PM: Profit Margin, PAT: Profit after Taxes, S: Turnover, NI: Net Income, DBT: Debt, EQ: Equity, TCAP: Total Capital, DY: Dividend Yield, P/E: Price-Earnings Ratio, GEPS: Gross Earnings Per Share, DCOV: Dividend Cover, CA: Current Assets, CL: Current Liabilities, SFUNDS: Shareholders Funds, IS: Interest Charges.

Table 9.11: List of accounting variables that we selected to predict long-term credit ratings

PCs	Percentage of Variance Explained (PVE)		
	PCA	NN-PCA	NN-NLPCA
1	18.9 %	28.82 %	63.98 %
2	32.6 %	42.24 %	89.52 %
3	44.8 %	51.42 %	95.57 %
4	53.1 %	59.72 %	97.77 %
5	60.3 %	66.33 %	98.31 %
6	66.7 %	71.67 %	98.89 %
7	72.8 %	76.72 %	99.38 %
8	76.9 %	81.30 %	99.60 %
9	80.7 %	83.98 %	99.70 %
10	84.2 %	86.59 %	99.76 %
11	87.4 %	88.99 %	99.84 %
12	90.3 %	88.99 %	99.92 %

Table 9.12: PVE by PCA, NN-PCA, and NN-NLPCA before conceptual clustering of the accounting variables that we selected to predict long-term credit ratings

PCs	Percentage of Correct Classifications (PCC) - Jackknife			
	LDA (PCA)	LDA (NN-PCA)	PNN (NN-PCA)	PNN (NN-NLPCA)
1	58.3 %	60 %	59.16 %	52.5 %
2	62.5 %	61.7 %	61.66 %	63.33 %
3	62.5%	61.7 %	64.16 %	68.33 %
4	62.5 %	69.2 %	77.5 %	67.5 %
5	62.5 %	69.2 %	84.16 %	70 %
6	62.5 %	62.5 %	78.33 %	73.33 %
7	71.7 %	69.2 %	80.83 %	74.16 %
8	71.7 %	68.3 %	85.83 %	67.5 %
9	71.7 %	67.5 %	82.5 %	78.33 %
10	71.7 %	69.2 %	81.66 %	80.83 %
11	70.8 %	60 %	83.33 %	79.16 %
12	70.8 %	61.7 %	82.5 %	80 %

Table 9.13: Total percentage of correct classifications after applying LDA and PNN to predict credit ratings and using all the accounting variables at the same time to apply PCA, NN-PCA, and NN-NLPCA

Inputs	PCs	Hidden Layers	NN-NLPCA	
			Training Set	Test Set
30	8	3	90.22 %	76.06 %
30	8	4	92.51 %	69.05 %
30	8	5	92.51 %	69.05 %
30	8	6	93.67 %	69.44 %
30	8	7	94.93 %	44.80 %
30	8	8	95.84 %	42.67 %

Table 9.14: PVE by NN-NLPCA in the training and test sets before conceptual clustering of the accounting variables that we selected to predict long-term credit ratings

Groups	Inputs	PCs	NN-NLPCA	
			Training Set	Test Set
RCAP	6	1	94.30 %	93. 65 %
PROF	4	1	83.47 %	77.71 %
FLEV	3	1	87.51%	80.69 %
INV	4	1	80.01 %	72.88 %
GRT	3	1	91.68 %	83.58 %
ST-LIQ	4	1	82.03 %	80.28 %
ICOV	2	1	94.69 %	94.03 %
RISK	4	1	80.66 %	75.62 %

Table 9.15: PVE by NN-NLPCA in the training and test sets after conceptual clustering of the accounting variables that we selected to predict long-term credit ratings

	Percentage of Variance Explained by Individual PCs		
	PCA	NN-PCA	NN-NLPCA
RCAP	50.1 %	58.71 %	94.21 %
PROF	75.9 %	79.55 %	91.55 %
FLEV	43.5 %	65.94 %	98.93 %
INV	41.2 %	49.42 %	84.14 %
GRT	54.4 %	63.46 %	93.02 %
ST-LIQ	48.4 %	54.64 %	90.71 %
ICOV	98 %	97.99 %	98.64 %
RISK	75 %	76.47 %	86.94 %

Table 9.16: PVE by PCA, NN-PCA and NN-NLPCA after conceptual clustering of the accounting variables that we selected to predict long-term credit ratings

		LDA (PCA)			LDA (NN-PCA)		
Actual Class	Patterns	Predicted Class Membership					
		A	AA	AAA	A	AA	AAA
A	67	48	9	10	49	10	8
AA	38	4	26	8	2	27	9
AAA	15	4	7	4	4	6	5
overall(%)		65 %			67.5 %		

Table 9.17: Leave-one-out classification results after applying the LDA to predict long-term credit ratings for three classes at the same time and using conceptual clusters of the accounting variables to apply PCA and NN-PCA

		PNN (NN-PCA)			PNN (NN-NLPCA)		
Actual Class	Patterns	Predicted Class Membership					
		A	AA	AAA	A	AA	AAA
A	67	57	6	4	48	5	14
AA	38	4	32	2	4	31	3
AAA	15	0	1	14	1	2	12
Overall (%)		85.83 %			75.83 %		

Table 9.18: Leave-one-out classification results after applying the PNN to predict long-term credit ratings for three classes at the same time and using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

		LDA (PCA)			LDA (NN-PCA)		
Classes	Patterns	T1	T2	T	T1	T2	T
A ↔ AA	67 ↔ 38= 105	13	7	20	14	8	22
A ↔ AAA	67 ↔ 15 = 82	18	5	23	16	3	19
AA ↔ AAA	38 ↔ 15 = 53	11	2	13	10	3	13

Table 9.19: Leave-one-out classification results after applying the LDA to predict long-term credit ratings for one class against the other and using conceptual clusters of the accounting variables to apply PCA and NN-PCA

		PNN (NN-PCA)			PNN (NN-NLPCA)		
Classes	Patterns	T1	T2	T	T1	T2	T
A ↔ AA	67 ↔ 38= 105	7	5	12	6	7	13
A ↔ AAA	67 ↔ 15 = 82	2	4	6	7	6	13
AA ↔ AAA	38 ↔ 15 = 53	3	1	4	2	1	3

Table 9.20: Leave-one-out classification results after applying the PNN to predict long-term credit ratings for one class against the other and using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

Classes	LDA (PCA)	LDA (NN-PCA)
AAA ↔ AA ↔ A	65 %	67.5 %
A ↔ AA	80.95 %	79.04 %
A ↔ AAA	71.95 %	76.82 %
AA ↔ AAA	75.47 %	75.47 %

Table 9.21: (%) Leave-one-out classification results after applying the LDA to predict long-term credit ratings using conceptual clusters of the accounting variables to apply PCA and NN-PCA

Classes	PNN (NN-PCA)	PNN (NN-NLPCA)
AAA ↔ AA ↔ A	85.83 %	75.83 %
A ↔ AA	88.57 %	87.61 %
A ↔ AAA	92.68 %	84.14 %
AA ↔ AAA	92.45 %	94.33 %

Table 9.22: (%) Leave-one-out classification results after applying the PNN to predict long-term credit ratings using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

		BPNN (NN-PCA)			BPNN (NN-NLPCA)		
Actual Class	**Patterns**	**Predicted Class Membership**					
		A	**AA**	**AAA**	**A**	**AA**	**AAA**
A	67	40	23	4	49	16	2
AA	38	5	30	3	8	30	0
AAA	15	2	4	9	6	7	2
Overall (%)		**65.83 %**			**67.5 %**		

Table 9.23: Leave-one-out classification results after applying the BPNN to predict long-term credit ratings for three classes at the same time and using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

		PNN (NN-PCA)			PNN (NN-NLPCA)		
Actual Class	**Patterns**	**Predicted Class Membership**					
		A	**AA**	**AAA**	**A**	**AA**	**AAA**
A	67	60	3	4	46	9	12
AA	38	10	27	1	6	29	3
AAA	15	7	1	7	0	1	14
Overall (%)		**78.33 %**			**74.16 %**		

Table 9.24: Leave-one-out classification results after applying the PNN to predict long-term credit ratings for three classes at the same time and using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

		BPNN (NN-PCA)			BPNN (NN-NLPCA)		
Classes	**Patterns**	**T1**	**T2**	**T**	**T1**	**T2**	**T**
A ↔ AA	67 ↔ 38= 105	4	20	24	5	16	21
A ↔ AAA	67 ↔ 15 = 82	4	7	11	3	7	10
AA ↔ AAA	38 ↔ 15 = 53	3	8	11	3	3	6

Table 9.25: Leave-one-out classification results after applying the BPNN to predict long-term credit ratings for one class against the other and using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

		PNN (NN-PCA)			PNN (NN-NLPCA)		
Classes	**Patterns**	**T1**	**T2**	**T**	**T1**	**T2**	**T**
A ↔ AA	67 ↔ 38= 105	11	6	17	12	20	32
A ↔ AAA	67 ↔ 15 = 82	6	5	11	5	7	12
AA ↔ AAA	38 ↔ 15 = 53	4	7	11	2	2	4

Table 9.26: Leave-one-out classification results after applying the PNN to predict long-term credit ratings for one class against the other and using conceptual clusters of the accounting variables to apply NN-PCA and NN-NLPCA

CHAPTER 10: SUMMARY OF THE THESIS AND FUTURE RESEARCH

Our study has implications for two lines of research in financial markets. At one level, the finding of predictability in stock and bond markets can be regarded as further evidence against the efficient markets hypothesis. At a more practical level, the techniques developed here have immediate application to support trading systems in financial institutions.

In Section 10.1 we summarise the findings of our work, examine these claims, and in Section 10.2 look ahead to ways in which the research reported here might be improved and extended.

10.1 SUMMARY AND CONCLUSIONS

The efficient markets hypothesis – the proposition that stock prices impound publicly available information promptly and without bias – was until recently regarded as one of the strongest pillars of modern finance theory.

However, empirical research has documented evidence for the ability of firm specific variables to explain systematic variations in the cross-section of stock returns beyond what might be expected on the basis of the Capital Asset Pricing Model. Indeed, variables such as market capitalisation, book to market equity, price-earnings ratios, debt to equity ratios, and several other similar variables have been found to explain the cross-section of expected returns better than the original CAPM. Three explanations have been suggested in the literature for this. First, these variables help to identify stocks that are mispriced because of systematic misjudgements of investors, and so genuinely indicate market inefficiency. Second, these variables proxy for sensitivity to risk factors beyond the market risk assumed by the CAPM, so that the correlation between the variables and returns reflects compensation for bearing risk. Third, their predictive ability is an artifact of "data snooping" and a survivorship bias in the U.S. COMPUSTAT database.

Other empirical studies suggest further that stock returns can be predicted by publicly available information such as time-series data on various economic variables, especially those with an important business cycle component. This conclusion holds true over different investment horizons and across different markets. Economic variables that have been found to be highly

correlated with stock returns include among others dividend yields, measures of interest rates, measures of inflation, and growth rates of industrial production. Most researchers interpret the ability of macroeconomic variables to explain stock returns as evidence for multi-factor alternatives to the CAPM such as the Arbitrage Pricing Theory.

Empirical studies have also documented significant return autocorrelations, especially for long horizon returns. Again there are interpretations of this evidence which are consistent with both rational and irrational investor behaviour. One explanation is that the predictability of stock returns is due to some form of irrationality such as fads, speculative bubbles or noise trading that deviates stock prices from their fundamental values and therefore generates abnormal returns. Another explanation is that the predictability of stock returns is a consequence of rational time variation in expected returns as business conditions, investment opportunities, and risk aversion change through time. This seems to be a more convincing explanation if we consider that the variation in expected returns is common across assets and related to business conditions. Yet other researchers argue that the statistical tests that have been used to test patterns in autocorrelations have low power, and leave the issue of market efficiency undecideable.

Ultimately, all tests of market efficiency are tests of a joint hypothesis – that some underlying asset pricing model is valid, and that investors behave rationally in the context of this model. Investor preferences, investor expectations, availability of information, business and economic conditions should all be controlled for when testing efficiency. However, there are some obvious problems in handling these factors – for example, it is difficult to quantify investor preferences and expectations. The form of the asset pricing model specifies a joint probability distribution of returns and some instrumental variables proxying these factors. Again there are difficulties. For example, most studies have adopted a linear relation for the first moments, but other relations may also be valid. Indeed, it is often difficult to conclude anything about the validity of the asset pricing model even in principle. The full information set of investors is unobservable, and the econometrician necessarily relies on a reduced set of easily measured information.

The objective of our study has not been the impossible goal of establishing whether markets are inefficient in the strict sense. Rather, our study is an examination of relative efficiency, in the sense of Lo and MacKinlay (1999). That is, we examine whether we can develop a competitive trading strategy that gives us a trading advantage compared to other strategies. This is based on the observation that advances in computing technology over the last decade suggest that highly

computational models and strategies can exploit profitable opportunities that could not have been exploited several years ago.

To develop a competitive trading strategy, we considered that most of the econometric methods that have been used so far to predict stock returns have been designed to detect either linear structure or strictly defined forms of non-linearities in the financial data. For example, the CAPM and the APT are based on linear models of expected returns, whereas ARCH-type and GARCH-type volatility models are based on strictly defined non-linear models. However, all these models might not be the most suitable if the data is fuzzy, chaotic, or exhibits unpredictable non-linearities. Empirical evidence suggests that financial data exhibit unpredictable non-linearities and other patterns that are inconsistent and chaotic in the time scale. Furthermore, we have to consider that several factors such as human judgements, human emotions, human feelings, human expectations, psychology, politics, and other qualitative factors affect in a high degree the process that drives stock prices. Under these considerations, it is obvious that most of the models that have been used so far to predict stock returns may not be able to deal with the actual process in the financial data that is unpredictable and lacks a well-defined physical content. More powerful models should be applied that will be able to extract the hidden knowledge in the financial data that cannot be detected by either linear or well-defined non-linear models.

One consequence of the rapid development of computer power in the 1980s was the development of computer-intensive classification algorithms. These include among others neural networks, decision trees and rule induction techniques. Neural networks are likely to be most superior to other methods if the data is fuzzy, chaotic, or exhibits unpredictable non-linearities that cannot be detected by linear models or other models that are based on strictly defined non-linear functional forms. Furthermore, neural networks are likely to have better generalisation performance than other statistical methods if the data incorporates human judgement or other qualitative factors because they detect patterns in data in a manner analogous to human thinking. This excellent performance of artificial neural networks should not be considered surprising if we take into account that these algorithms have been built on strong theoretical foundations. However, neural networks have several drawbacks as well. The major drawback of neural networks is that the learning process is very slow. Another drawback of neural networks is their inherent inability to explain in a comprehensible form the process by which a given decision or output generated by the model has been reached.

Decision tree algorithms are powerful for classification and prediction. These algorithms are able to model a wide range of data distributions since only a few assumptions are made about

the model and the data distribution. In addition, they are based on the hierarchical decomposition which implies better use of available features and computational efficiency in classification. Therefore, they are very able to handle complex interactions between variables. A major advantage of decision trees is that they perform classification by a sequence of simple and easy to understand tests whose semantics are intuitively clear to domain experts. On the other hand, the main disadvantage of decision trees is that they tend to grow very large for realistic applications and are thus difficult to interpret by humans. In response to this limitation, there has been some research in transforming decision trees into other representations using rule induction techniques. Although a variety of other representations have also been used in machine learning, a great deal of research has focused on rule induction for the following reasons: first, rules are often easier for people to understand; second, certain types of prior knowledge can be easily incorporated in the learning process; and third, rule induction techniques overcome the use of the limited-knowledge propositional logic formalism and they can be easily extended to the first order logic. However, one disadvantage of rule induction methods is that they scale poorly for large data sets. In addition, despite the fact that rule induction techniques offer interpretable rules, they are not expert systems. The knowledge engineer has still a substantial amount of work to perform in order to generate rules that perform well and are also sensible so that they can enhance the knowledge of domain experts. Despite this weakness, however, rule induction systems result in simple rules that are more preferable than other machine learning representations.

We investigated the distributional properties of financial ratios and we compared three heterogeneous classifiers namely, LDA, PNN, and RRI in terms of their ability to predict long-term bond ratings. Our target data were 132 rated bond issues, whereas our predictor variables were five performance ratios and three growth indicators. We found that the PNN which is a non-linear classifier and the RRI which is a rule induction classifier not only significantly outperform the LDA, but they are also more robust to different distributional assumptions compared to LDA which is affected from the assumption of multivariate normality. The final outcome of this experiment is a new quantitative system to predict long-term bond ratings using probabilistic neural networks and rule induction techniques. Overall, the findings of this experiment support the evidence for the existence of non-linearities and other complex processes in the financial data. On the other hand, no model is found perfectly robust and small variations in the sample size affect the overall performance of the models. Therefore, rather than applying a single non-linear model for any specific financial application, it is more preferable and wise to use a variety of different models to deal with different sample sizes, unpredictable non-linearities, and inconsistent patterns in the financial data.

Considering the empirical evidence but also the "strong" and "wild" capabilities of supervised classification rules, we explored the possibility of using these algorithms to address the problems of stock return predictability and stock selection. For this purpose, we selected a small number of heterogeneous classifiers namely, LDA, PNN, LVQ, OC1, and RRI and we applied them to classify and predict which shares are likely to have exceptional returns in the future. These classifiers have been chosen as representatives of five different model families: the LDA is a well-known linear classifier, the PNN is a kernel density non-linear classifier, the LVQ is a vector quantization classifier that also gives a non-linear partitioning of the data, the oblique OC1 classifier is a recursive partitioning classifier, and the RRI is a rule induction algorithm that depends on the induction of logical rules. We used these classifiers to explore the potential for identifying high performing shares on the London Stock Exchange. The model inputs we used were 38 accounting ratios for around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. We compared and contrasted the classifiers in terms of classification accuracy and profitability in the target years 1993-97. Overall, the classification results suggested that the non-linear classifiers except RRI outperform on average the LDA. On the other hand, the financial results suggested that the improvements in classification that are achieved by the non-linear models are not reflected to the financial returns in the same degree and there are only minor inconsistencies in the profitability of the classifiers from one year to the next. Despite these differences, however, we found that all methods produce consistent excess returns and outperform the benchmark.

There are several aspects we considered in evaluating the importance of our trading system. The first aspect was the trade-off between predicted returns and risk. We observed that greater excess returns are achieved in years where the index rises more sharply and we indicated that high returns tend to be achieved only at the expense of high risk. On the other hand, we mentioned that the actual high portfolios and to a lesser degree the predicted high portfolios contain a disproportionate number of small capitalisation stocks. Despite these considerations, however, we emphasised that the strong advantage of our trading system is that all classification methods produce consistent returns and excess returns for the target years 1993-97 even though the theoretical measures of risk make the results vulnerable to the accusation that high returns are achieved only at the expense of high risk.

In view of the success of forecast combination in more conventional forecasting exercises such as the one reported from Makridakis et al. (1982), we investigated if we could improve the performance of our trading system by combining the predictions of the five classifiers using majority voting (MV) and unanimous voting (UV) schemes, whereby a share is not assigned to the high performing portfolio unless the majority of classifiers or all classifiers, respectively,

agree on their decisions. The attraction of combining heterogeneous models is very obvious. If the predictions of the component classifiers are exactly the same, then a combination of these predictions will not improve the predictive accuracy of the composite classifier. On the other hand, if the individual components make some different predictions, then there is a hope that combining their predictions using the appropriate mechanisms might improve the predictive accuracy of the composite architecture. Furthermore, a composite architecture based on heterogeneous component models will be more flexible to capture the structure of the data set if there are sub-areas with different underlying processes. For example, if the data set is non-linear, then a composite architecture that combines linear components may not perform well. On the other hand, a composite architecture that combines both linear and non-linear components will be more flexible to capture the underlying structure of the data. Under these considerations, we compared and contrasted our five classifiers namely, LDA, PNN, LVQ, OC1, and RRI in terms of classification accuracy, profitability, and trading volume. Our target data were total returns on all shares traded on the London Stock Exchange in the years 1993-97, whereas our predictor variables were 38 accounting ratios drawn from published accounting statements. After experimentation, we found that using MV to combine the classifiers does not improve classification accuracy and profitability, whereas using UV to combine the classifiers improves substantially overall accuracy and profitability and this improvement is achieved with lower transaction costs. However, we emphasised that findings like this may be vulnerable to the conclusion that high returns are achieved at the expense of high risk. After adjusting for market risk using the CAPM, we found that LDA and PNN are less attractive in terms of risk-adjusted returns compared to OC1 which achieves a better deal of risk adjusted returns.

We extended our methodology to predict high performing shares by combining the five statistical classifiers namely LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes and using accounting information, economic information, past share and index returns information as well as information about the industrial classification of around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. We performed two experiments: in the first experiment, we compared the five algorithms namely, LDA, PNN, LVQ, OC1 and RRI, and the two voting methodologies, namely MV and UV in terms of classification accuracy, profitability, and trading volume for the target years 1993-97. We implemented the five classifiers and the two voting methodologies using three different sets of information: first, using accounting information only (AI); second, using economic information, past share and index returns information, and information about the industrial classification of the companies only (ERIIC); and third, using all the available information (ALL). We found that all classification methods produce consistent excess returns. However, greater gains result from UV where a share is not classified as high performing share unless all classifiers agree.

The UV principle not only produces significantly greater returns than the other methods, but it also results in substantial reductions in the number of shares traded. Our results also suggested that there are substantial gains of using all available information rather than using subsets of information only especially for classifiers such as OC1, LVQ, and RRI. These results should not be considered surprising, however, if we take into account the way these classifiers form their hypothesis for different clusters of the sample data. In the second experiment, we applied the UV methodology over two parallel implementations of the classifiers using accounting and non-accounting information, respectively. According to this implementation, a share is not assigned to the high performing portfolio unless the five classifiers from the first implementation based on accounting information as well as the same classifiers from the second implementation based on non-accounting information agree unanimously on their decisions. After performing this experiment, we found greater gains in profitability and substantial reductions in the trading volume.

We further applied our methodology to different industrial sectors and we examined the benefits of applying our trading system to homogeneous industrial sectors. Under this consideration, we examined the potential of identifying outperforming shares in homogeneous U.K. industrial sectors by combining our five heterogeneous statistical classification algorithms namely, LDA, PNN, LVQ, OC1, and RRI through MV and UV schemes and using accounting information of around 700 companies with shares traded on the London Stock Exchange in the years 1991-97. Our target data were total returns on all shares traded on the London Stock Exchange in the years 1994-97. Our input variables were 38 accounting ratios drawn from published accounting statements. After applying the LDA and the PNN, we found that both classifiers produce consistent excess returns after restricting the sample to service companies, whereas additional benefits arise after adding utility, financial and property companies. On the other hand, the high performing portfolios that result after restricting the sample to manufacturing and extractive companies are not particularly profitable and fail to produce consistent excess returns. After implementing the five classifiers using service companies only, we found that all classification methods produce consistent excess returns. However, greater gains result from UV. The UV principle not only produces significantly greater returns than the other methods, but it also results in substantial reductions in the number of shares traded. Our results from this experiment provide substantial evidence for the ability of statistical classification methods to identify high performing shares if the sample is homogeneous. There are three main benefits that result for our trading system after restricting the sample to homogeneous industrial sectors: first, less data are required for the implementation of the models and therefore less time and effort are required to optimise the classification methods; second, the trading volume is reduced substantially and this results in more efficient trading

strategies; and third, the transaction costs are minimised because less shares are traded for each particular year.

We have to emphasise that our classification methods should be treated with caution due to their large number of parameters. To avoid the possibility of overfitting, we applied sophisticated data pre-processing techniques in order to eliminate the effect of outliers and increase the robustness of the models. More specifically, we investigated alternative methodologies to reduce the dimensionality of our data in ways other than ad hoc stepwise variable elimination procedures. The alternative methods we investigated were PCA as well as linear and non-linear dimensionality reduction techniques based on neural networks. PCA can be used as a dimensionality reduction technique within some other type of analysis such as discriminant analysis, cluster analysis, canonical correlation analysis etc. However, this procedure may be unsatisfactory for two reasons: first, the within-group covariance matrix may be different for different groups; and second, there is no guarantee that the separation between groups will be in the direction of the high-variance PCs. Another problem with PCA is that the relative sizes of the elements in a variable weight vector associated with a particular PC indicate the relative contribution of the variable to the variance of the PC. Therefore, the patterns of variable weights for a particular PC are used to interpret the PC. A problem is identified, however, if more than a few variables have a significant contribution to the variance of a particular PC. In this case, the interpretation of this PC is extremely difficult. Finally, we should consider that PCA is a linear method and most real problems are non-linear. It has been shown that if PCA is applied in non-linear problems, minor components might contain important information. Therefore, if minor components are discarded important information is lost. It is therefore proposed that a NLPCA should be applied to deal with these problems. NLPCA uncovers both linear and non-linear correlations among variables without restriction on the character of the non-linearities presented in the data.

As an alternative to PCA, we applied dimensionality reduction techniques based on neural networks and we examined the ability of the resulting PCs to predict which shares are likely to have exceptional returns in the future by applying our five heterogeneous classifiers namely, LDA, PNN, LVQ, OC1, and RRI. We examined the effectiveness of the neural network dimensionality reduction techniques by applying them using the minimum number of available observations and we compared them with ad hoc procedures to identify the best subset for variables that require more observations in order to avoid the problem of overfitting and they are more time-consuming than the state-of-the-art methodology. After experimentation, we found that neural network linear PCA (NN-PCA) and neural network non-linear PCA (NN-NLPCA) explain a higher proportion of variation in the original set of variables than the

common PCA methodology and their resulting PCs are competitive to other dimensionality reduction techniques in maintaining important discriminating power to identify which shares are likely to have exceptional returns in the future. Furthermore, we found that NN-PCA and NN-NLPCA dimensionality reduction techniques can be used as an alternative to ad hoc methodologies for variable selection because they require less effort and their resulting PCs classify as well as or even better than the optimal subsets of variables that we found after applying stepwise variable elimination procedures. The results of this experiment provide evidence for the existence of non-linear correlations in the financial data that cannot be detected properly by linear dimensionality reduction techniques such as the PCA methodology.

To verify further the effectiveness of the neural network dimensionality reduction methodologies, we applied them to homogeneous subsets of financial ratios and used the derived PCs to assess the long-term credit standing of U.K. debt issuers. The results of this experiment confirmed the findings from the first experiment. More specifically, we found that both NN-PCA and NN-NLPCA explain a higher amount of variation in the original set of variables than the common PCA methodology and the derived PCs are easier to interpret if extracted from homogeneous groups of financial ratios. Although the NN-NLPCA explains a higher proportion of variation in the original set of variables than the NN-PCA, the PCs extracted from NN-PCA discriminate better than the PCs extracted from NN-NLPCA when the PNN model is applied to classify debt issuers into boundary rating classes. On the other hand, the PCs extracted from NN-NLPCA discriminate better than the PCs extracted from NN-PCA when the BPNN model is applied. In terms of classification accuracy, however, the PNN model performs better. The results of this experiment provide evidence that NN-PCA and NN-NLPCA architectures can be successfully implemented as a preliminary step to assess the credibility of U.K. debt issuers and at the same time provide an alternative solution to overfitting.

The key findings of our research are:

First, we confirm for U.K. data the results of previous studies that report predictable patterns in stock returns using firm specific variables, economic variables, as well as past share and index returns information.

Second, we extend previous research. Earlier studies typically examined the predictability of share returns mostly under restricted forms of linear and non-linear models. Our results provide evidence for the relatively greater ability of non-linear classification methods over the linear model to identify high performing shares out of sample. Non-linear models are more flexible to deal with the complex relationships that are evident in the financial data compared to the linear

model that is able to handle more simple linear patterns. The main advantage of our methodology is model flexibility that is essential for the complex financial processes that are chaotic and inconsistent in the time-scale.

A corollary is the observation for a certain degree of correlation between classification accuracy and profitability. Therefore, it is obvious that further improvements in classification accuracy may result in further improvements in profitability. Maybe more sophisticated data preprocessing techniques are required to improve the performance of our classification methods.

Third, our results also confirm findings of previous studies which demonstrated that combinations of individual forecasts improves forecasting accuracy. In response to previous studies that support combinations of homogeneous component classifiers rather than heterogeneous classifiers, we provide substantial evidence that a combination of heterogeneous classifiers using the unanimous voting scheme is also successful and produces impressive results over the individual component classifiers. The benefit of applying voting procedures to combine the individual component classifiers is that combining can be done with little or no increase in cost.

Fourth, the results from our dimensionality reduction procedures support the view for the existence of strong non-linear correlations in the financial data. The non-linear dimensionality reduction techniques based on neural networks significantly outperform the linear equivalents as well as the standard PCA analysis methodology. The practical importance of dimensionality reduction techniques based on neural networks becomes of particular interest if we consider that the same techniques can be applied in a variety of other applications that incorporate large data matrices. Finally, these techniques can be used as an essential data preprocessing technique for a variety of other models that suffer from the curse of dimensionality if the data is small.

Whether these findings together indicate market inefficiency is, as we discussed above, a contentious and undecideable issue. The excess profits made by the non-linear rules are substantial, and arguably exceed any reasonable estimates of risk premia. On the other hand the techniques we are applying are novel, high-cost (in terms of investor knowledge). Hence it is also arguable that investors could not realistically be expected to have uncovered the profit opportunities identified by our high-tech statistical models.

We have to reiterate that our conclusions are conditional on the information sets we have used, the way we implemented the models, and the specific trading rules we have applied. Certainly,

more financial applications are required to prove the general validity of our methodology. Despite these considerations, however, we have to emphasise that our methodology establishes a very novel and new way in stock predictability. It is a committee of experts rather than a single model. It is classification rather than regression.

10.2 SUGGESTIONS FOR FUTURE RESEARCH

Possible extensions to this research fall into two categories – academic and commercial.

In academic terms, the robustness of our findings needs to be confirmed by repeating the modelling process on other stock and bond markets, and possibly credit, currency and commodity markets.

The precise classification rules chosen are also to some degree arbitrary (though representative of the range of techniques currently on offer). As research on the performance of these techniques in other fields develops, some of the methods used here may prove to be suboptimal. Similarly, more complex voting and combining rules, possibly themselves based on neural networks, can be examined.

In commercial terms, the research reported here has obvious applications as the basis for a stock or bond trading system. For such purposes, it is desirable to "tune" the model to optimise some return: risk objective function. Parameters which could be adjusted include

- the fraction of the sample classified as "high"
- the range of input variables
- the trading rule

For example, many hedge funds practice long/short strategies, where not only are potentially high-performing shares bought, but also low-performing shares are short-sold. Most traders put stop-loss and limit orders in place, and these might be made conditional on the degree of confidence in the model predictions. Another improvement would be to incorporate into our trading system a Value-at-Risk framework for marking-to-market daily. And so on.

However, the description of the fully functioning commercial trading system goes well beyond the ambitions of this thesis. Indeed, we have deliberately been a little arbitrary in some of the features of the work reported here, because of the commercial sensitivity of the project.

REFERENCES

ABHYANKAR A., COPELAND L.S., and WONG W. (1997). Uncovering non-linear structure in real-time stock market indexes: The S&P 500, the DAX, the Nikkei 225, and the FTSE-100. Journal of Business and Economic Statistics 15: 1-14.

ACTON, F.S. (1959). Analysis of straight-line data. (New York: Dover Publications).

AGNEW, C.E. (1985) Bayesian consensus forecasts of macroeconomic variables. Journal of Forecasting 4: 363-376.

ALBANIS G.T., LONG J.A., and HISCOCK M. (1997). A quantitative system to predict bond ratings using probabilistic neural networks - A preliminary step for financial distress prediction. In: THE 5TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL FINANCE, Proceedings, London Business School, London 1997.

ALBANIS, G.T. (1998a). Using rule induction techniques to predict long-term bond ratings. In: THE BNP FORECASTING FINANCIAL MARKETS: ADVANCES FOR EXCHANGE RATES, INTEREST RATES AND ASSET MANAGEMENT CONFERENCE, Proceedings, London 1998.

——— (1998b). Using probabilistic neural networks and rule induction techniques to predict long-term bond ratings. In: THE 2ND WORLD MULTICONFERENCE ON SYSTEMICS, CYBERNETICS AND INFORMATICS AND THE 4TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, ANALYSIS AND SYNTHESIS (SCI'98/ISAS'98), Proceedings, Orlando.

ALBANIS, G.T. and BATCHELOR, R.A. (1999a). Five algorithms to predict high performance stocks. In: THE 6TH FORECASTING FINANCIAL MARKETS: ADVANCES FOR EXCHANGE RATES, INTEREST RATES AND ASSET MANAGEMENT INTERNATIONAL CONFERENCE, Proceedings, London 1999.

——— (1999b). Assessing the long-term credit standing using dimensionality reduction techniques based on neural networks - An alternative to overfitting. In: THE 3RD WORLD MULTICONFERENCE ON SYSTEMICS, CYBERNETICS AND INFORMATICS AND THE 5TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, ANALYSIS AND SYNTHESIS (SCI'99/ISAS'99), Proceedings, Orlando

——— (2000a). Five classification algorithms to predict high performance stocks. In: Advances in Quantitative Asset Management ed. by DUNIS C. (Boston: Kluwer Academic Publishers).

——— (2000b). 21 methodologies to beat the market. In: THE FORECASTING FINANCIAL MARKETS 2000/COMPUTATIONAL FINANCE 2000 (FFM2000/CF2000),

CONFERENCE, Proceedings, London, May.

——— (2000c). Combining non-linear classifiers for stock selection. In: THE FORECASTING FINANCIAL MARKETS 2000/COMPUTATIONAL FINANCE 2000 (FFM2000/CF2000), CONFERENCE, Proceedings, London.

——— (2001a). 21 non-linear ways to beat the stock market. In: Developments in Forecast Combination and Portfolio Choice, ed. DUNIS C. MOODY J., and TIMMERMAN A., (forthcoming), (Chichester: John Wiley & Sons).

——— (2001b). Predicting high performance stocks using dimensionality reduction techniques based on neural networks, ed. DUNIS C. MOODY J., and TIMMERMAN A., (forthcoming), (Chichester: John Wiley & Sons).

——— (2001c). Combining heterogeneous classifiers to predict stock returns in U.K. industrial sectors. In: THE FORECASTING FINANCIAL MARKETS 2001 CONFERENCE, Proceedings, London, May.

ALI, K.M. and PAZZANI, M.J. (1996). Error reduction through learning multiple descriptions. Dept. of Information and Computer Science, Technical Report 95-39. University of California, Irvine, CA.

ALPAYDIN, E. (1993). Multiple networks for function learning. In: THE IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, Proceedings, San Francisco, CA, 1: 9-14.

——— (1998). Techniques for combining multiple learners. In: THE ENGINEERING OF INTELLIGENT SYSTEMS CONFERENCE, Proceedings, ICSC Press 2: 6-12.

ALTMAN, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporation bankruptcy. Journal of Finance 23: 589-609.

ALTMAN, E.I. and EISENBEIS, R.A. (1978). Financial discriminant analysis: a clarification. Journal of Financial and Quantitative Analysis: 185-195.

ALTMAN, E. and KATZ, S. (1976). Statistical bond rating classification using financial and accounting data. In: THE CONFERENCE ON TOPICAL RESEARCH IN ACCOUNTING, Proceedings, ed. by SCHIFF, M. and SORTER, G. (New York: New York University School of Business).

ANANDALINGAM, G. and CHEN, L. (1989). Linear combination of forecasts: a general Bayesian model. Journal of Forecasting 8: 199-214.

ANG, J.S. and PATEL, A.K. (1975). Bond Rating Methods: Comparison and Validation. The Journal of Finance Vol. XXX (2): 631-640.

ARIEL, R.A. (1987). A monthly effect in stock returns. Journal of Financial Economics 18: 161-174.

——— (1990). High stock returns before holidays: Existence and evidence of possible causes. Journal of Finance 45(5): 1611-1626.

ARMSRONG J.S., LUSK E.J., GARDNER E.S. JR., GEURTS M.D., LOPES L.L., MARKLAND R.E., MCLAUGHLIN R.L., NEWBOLD P., PACK D.J., ANDERSEN A., CARBONE R., FILDES R., NEWTON H.J., PARZEN E., WINKLER R.L., and MAKRIDAKIS S. (1983). Commentary on the Makridakis time series competition (M-Competition). Journal of Forecasting 2: 259-311.

ARMSTRONG, J.S. (1986). The ombudsman: research on forecasting: a quarter century review, 1960-1984. Interfaces 16: 89-109.

ASCH, S.E. (1951). Effects of group pressure on the modification and distortion of judgement. In: Groups, Leadership and Men, ed. by GUETZKOW, H. Pittsburgh, Carnegie Institute of Technology Press.

ATLAS L., COLE R., MUTHUSAMY Y., LIPPMAN A., CONNOR J., PARK D., EL-SHARKAWI M., and MARKS R.J. (1990). A performance comparison of trained mutli-layer perceptrons and trained classification trees. IEEE Proceedings 78: 1614-1619.

ATTANASIO, O. and WADHWANI, S. (1990). Does the CAPM explain why the dividend yield helps predict returns ? Working Paper No. 104, London School of Economics.

BAILLIE, R.T. and DEGENNARO, R.P. (1990). Stock returns and volatility. Journal of Financial and Quantitative Analysis 25: 203-214.

BAILY W., STULZ R.M., and YEN S. (1990). Properties of daily stock returns from the Pacific-Basin stock markets: Evidence and implications. In: Pacific-Basin Capital Markets Research, ed. by RHEE, S.G. and CHANG, R.P. VOL 1, North-Holland, Amsterdam.

BALDI, P. and HORNIK, K. (1989). Neural networks and principal component analysis: learning from examples without local minima. Neural Networks 2: 53-58.

BALL, R. (1978). Anomalies in relationships between securities' yields and yield-surrogates. Journal of Financial Economics 6: 103-126.

——— (1992). The earnings-price anomaly. Journal of Accounting and Economics 15: 319-345.

BALL, R. and KOTHARI, S.P. (1989). Nonstationary expected returns: Implications for tests of market efficiency and serial correlation in returns. Journal of Financial Economics 25: 51-74.

BALL R., KOTHARI S.P., and SHANKEN J. (1995). Problems in measuring portfolio

performance: An application to contrarian investment strategies. Journal of Financial Economics 38: 79-107.

BALVERS R.J., COSIMANO T.F., and MACDONALD B. (1990). Predicting stock returns in an efficient market. Journal of Finance 45: 1109-1128.

BANSAL, R. and VISWANATHAN, S. (1993). A new approach to international arbitrage pricing. Journal of Finance 48: 1719-1747.

BANZ, R.W. (1981). The Relationship between return and market value of common stocks. Journal of Financial Economics 9: 3-18.

BANZ, R.W. and BREEN, W. (1986). Sample dependent results using accounting and market data: Some evidence. Journal of Finance 41: 779-794.

BASU, S. (1977). Investment performance of common stocks in relation to their P/E ratios: A test of the efficient market hypothesis. Journal of Finance 3: 663-681.

——— (1983). The relationship between earning's yield, market value, and the returns for NYSE common stocks: Further evidence, Journal of Financial Economics 12: 129-156.

BATCHELOR, R.A. (1990). All forecasters are equal. Journal of Business and Economic Statistics 8: 143-144.

BATCHELOR, R.A. and DUA, P. (1990a) Forecaster ideology, forecasting technique, and the accuracy of economic forecasts. International Journal of Forecasting 6: 3-10.

——— (1990b). Product differentiation in the economic forecasting industry. International Journal of Forecasting 6: 311-316.

——— (1992). Conservatism and consensus-seeking among economic forecasters. Journal of Forecasting 11: 169-181.

——— (1995). Forecaster diversity and the benefits of combining forecasts. Management Science 41 (1): 68-75.

BATES, J.M. and GRANGER, C.W.J. (1969). The combination of forecasts. Operational Research Quarterly 20: 451-468.

BATTITI, R. and COLLA, A.M. (1994). Democracy in neural nets: voting schemes for classification. Neural Networks 7 : 691-707.

BEENSTOCK, M. and CHAN, K.F. (1988) Economic forces in the London stock market, Oxford Bulletin of Economics and Statistics 50: 27-39.

BENEDIKTSSON, J.A. and SWAIN, P.H. (1992). Consensus theoretic classification methods. IEEE Transactions on Systems, Man, and Cybernetics 22: 688-704.

BENGIO, Y. (1996). Using a financial training criterion rather than a prediction criterion. Technical Report #1019. Dept. Informatique et Recherche Operationnelle, Universite de Montreal.

BERNARD, V.L. and THOMAS, J.K. (1989). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. Journal of Accounting Research 27 (Supplement): 305-340.

BERNARD, V.L. and WAHLEN, J. (1997). Accounting-based stock price anomalies: separating market inefficiencies from risk. Contemporary Accounting Research 14: 89-136.

BESSLER, D.A. and BRANDT, J.A. (1981). Forecasting livestock prices with individual and composite methods. Applied Economics 13: 513-522.

BESSLER, D.A. and CHAMBERLAIN, P.J. (1987). On Bayesian composite forecasting. OMEGA International Journal of Management Science 15: 43-48.

BHANDARI, L.C. (1988). Debt/equity ratio and expected common stock returns: empirical evidence. Journal of Finance 43: 507-528.

BILSON, J.F.O. (1983). The Evaluation and use of foreign exchange rate forecasting services. In: Managing Foreign Exchange Rate Risk, ed. by HERRING, R. Cambridge. Cambridge University Press, 149-179.

BISCHOFF, C.W. (1989). The combination of macroeconomic forecasts. Journal of Forecasting 8: 293-314.

BISHOP, M.C. (1995). Neural networks for pattern recognition. (Oxford: Oxford University Press).

BLACK, F. (1972). Capital market equilibrium with restricted borrowing. Journal of Business 45: 444-454.

BLACK F., JENSEN M.C., and SCHOLES M. (1972). The capital asset pricing model: some empirical tests. In: Studies in the theory of capital markets, ed. by JENSEN, M.C (Praeger, New York).

BLAKE D., BEENSTOCK M., and BRASSE V. (1986). The Performance of U.K. exchange rate forecasters. The Economic Journal 96: 986-999.

BLUME, M. and FRIEND, I. (1973). A new look at the capital pricing model. Journal of Finance March: 19-34.

BLUME, M.E and STAMBAUGH, R.F (1983). Biases in computed returns: an application to the size effect. Journal of Financial Economics 12: 387-404.

BODIE, Z. (1976). Common stocks as a hedge against inflation. Journal of Finance 31: May.

BODURTHA, J. and MARK, N. (1991). Testing the CAPM with time-varying risk and returns. Journal of Finance 46: 1485-1505.

BOLLERSLEV, T.R. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31: 307-327.

BOLLERSLEV, T.R (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. The Review of Economics and Statistics 72 (3): 498-506.

BOLLERSLEV, T.P. ENGLE, R.F. and WOOLDRIDGE, J.M. (1988). A Capital asset pricing model with time-varying covariances. Journal of Political Economy 96: 116-131.

BONELLI, P. and PARODI, A. (1991). An efficient classifier system and its experimental comparisons with two representative learning methods on three medical domains. In: THE 4TH INTERNATIONAL CONFERENCE ICGA-91, Proceedings, (San Matteo, CA: Morgan Kaufmann), 288-295.

BORDLEY, R.F. (1982). The combination of forecasts: a Bayesian approach. Journal of the Operational Research Society 33(2): 171-174.

———— (1986). Technical note - linear combination of forecasts with an intercept: a Bayesian approach. Journal of Forecasting 5: 243-249.

BOUDOUKH, J. and RICHARDSON, M. (1993). Stock returns and inflation: a long horizon perspective. The American Economic Review 83 (5): 1346-1355.

BOURLAND, H. and KAMP, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics 59: 291-294.

BOSSAERTS, P. and GREEN, R.C. (1989). A general equilibrium model of changing risk premia: Theory and tests. Review of Financial Studies 2: 467-493.

BRANCH, B. (1974). Common stock performances and inflation: an international comparison. Journal of Business 47 (January): 57-81.

BRANDT, J.A. and BESSLER, D.A. (1981). Composite forecasting: an application with U.S. hog prices. American Journal of Agricultural Economics 63: 135-140.

———— (1983). Price forecasting and evaluation: an application in agriculture. Journal of Forecasting 2: 237-248.

BREEN W., GLOSTEN L.R., and JAGANNATHAN, R. (1989). Economic significance of predictable variations in stock index returns. Journal of Finance 44 (5): 1177-1189.

BREEDEN, D.T. (1979). An intertemporal asset pricing model with stochastic consumption and investments opportunities. Journal of Financial Economics 7: 265-296.

BREIMAN L., FRIEDMAN J.H., OLSHEN E.A. and STONE C.J. (1984). Classification and regression trees. (Monterey, CA: Wadsworth and Brooks).

BREIMAN, L. (1994). Bagging predictors. Technical Report 421. Department of Statistics. University of California, Berkeley, CA.

BROCK W.A., LEBARON B., and LAKONISSHOK J. (1992). Simple technical trading rules and the stochastic properties of stock returns. Journal of Finance 47: 1731-1764.

BRODLEY, C.E. (1992). Dynamic automatic model selection. Technical Report 92-30. Dept. of Computer Science, University of Massachusetts, Amherst, MA. Pattern Recognition 26: 953-961.

BROWN P. A., KLEIDON W., and MARSH T.A. (1983) New evidence on the nature of size related anomalies in stock prices. Journal of Financial Economics 12: 33-56.

BROWN, S. J. and WEINSTEIN, M.I. (1983) A new approach to testing asset pricing models: the Bilinear Paradigm. Journal of Finance 38: 711-743.

BROWN D.E., CORRUBLE V., and PITTARD C.L. (1993). A comparison of decision tree classifiers with backpropagation networks for multimodal classification problems. Pattern Recognition: 953-61.

BRUNK C., and PAZZANI, M. (1991). Noise-tolerant relational concept learning algorithms. In: THE 8TH INTERNATIONAL WORKSHOP ON MACHINE LEARNING, Proceedings, Ithaca, New York, 1991.

BUNN, D.W. (1975). A Bayesian approach to the linear combination of forecasts. Operational Research Quarterly 26: 325-329.

———— (1977). A comparative evaluation of the outperformance and minimum-variance procedures for linear syntheses of forecasts. Operational Research Quarterly 28: 653-660.

———— (1978). A simplification of the matrix beta distribution for combining estimators. Journal of Operational Research Society 29: 1013-1016.

———— (1979). The suboptimality of composite forecasts derived from posterior probabilities. European Journal of Operational Research 3: 379-381.

———— (1985). Statistical efficiency in the linear combination of forecasts. International Journal of Forecasting: 151-163.

———— (1987). Expert use of forecasts: bootstrapping and linear models. In: Judgmental Forecasting, ed. by WRIGHT, G. and AYTON, P. (New York: Wiley), 229-241.

BUNN, D.W. and MUSTAFAOGLU, M.M. (1978). Forecasting political risk. Management Science 24: 1557-1567.

BURMEISTER, E. and WALL, K.D. (1986). The arbitrage pricing theory and macroeconomic factor measures. The Financial Review 21: 1-20.

CACCIATORE, T.W. and NOWLAN, S.J. (1994). Mixtures of controllers for jump linear and non-linear plants. In: Advances in neural information processing systems 6, ed. by COWAN J.D., TESAURO G., and ALSPECTOR J. (San Francisco, CA: Morgan Kaufmann): 719-726.

CACOULLOS, T. (1966). Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics 18(2): 179 - 189.

CAGAN, P. (1974). Common stock values and inflation: the historical record of many countries. Annual Report Supplement, National Bureau of Economic Research, Cambridge, MA.

CALVET, A. and LEFOLL, J. (1989). Risk and Return on Canadian capital markets: Seasonality and size effect. Finance 10: 21-39.

CAMPBELL, J.Y. (1987). Stock returns and the term structure. Journal of Financial Economics 18: 373-400.

CAMPELL, J.Y. and SHILLER, R.J. (1988). Stock Prices, Earnings, and Expected Dividends. Journal of Finance 43(3): 661-676.

CAMPBELL, J.Y. and HAMAO, Y. (1992). Predictable stock returns in the United States and Japan: a study of long-term capital market integration. Journal of Finance 47 (1): 43-69.

CAMPBELL J.Y., LO A.W., and MACKINLAY A.C. (1997). The econometrics of financial markets. (New Jersey: Princeton University Press).

CAPAUL C., ROWLEY I., and SHARPE W.F. (1993). International value and growth stock returns. Financial Analysts Journal 49 (1): 27-36.

CARBONE, R. and LONGINI, R.L. (1977). A feedback model for automatic real estate assessment. Management Science 24: 232-244.

CARBONE, R. in ARMSTRONG ET AL. (1983). Commentary on the Makridakis time series competition (M-Competition). Journal of Forecasting 2: 259-311.

CARTER, C. and CATLETT, J. (1987). Assessing credit card applications using machine learning. IEEE Expert: Intelligent Systems and their Applications 2: 71-79.

CHAN K.C., CHEN N. F., and HSIEH D. (1985). An exploratory investigation of the firm size effect. Journal of Financial Economics 14: 451-471.

CHAN, K.C. (1988). On the contrarian investment strategy. Journal of Business 43: 309-325.

CHAN, K.C. and CHEN, N.F. (1988). An unconditional asset-pricing test and the role of firm

size as an instrumental variable for risk. Journal of Finance 43: 309-325.

——— (1991). Structural and return characteristics of small and large firms. Journal of Finance 46: 1467-1484.

CHAN K.C., HAMAO Y., and LAKONISHOK J. (1991). Fundamentals and stock returns in Japan. Journal of Finance 46: 1739-1764.

CHAN, P.K. and STOLFO, S.J. (1995). A Comparative evaluation of voting and meta-learning on partitioned data. In: THE 12TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, Proceedings (San Mateo, CA: Morgan Kaufmann), 90-98.

CHANDRASEKHARAN R., MORIARTY M.M., and GORDON P.W. (1994). Testing for unreliable estimators and insignificant forecasts in combined forecasts. Journal of Forecasting 13: 611-624.

CHANG, K. (1985). Combination of opinions: the expert problem and the group consensus problem. PhD Dissertation, University of California at Berkeley.

CHANG, W.C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics 32: 267-275.

CHEN N.F., ROLL R., and ROSS S.A. (1986). Economic forces and the stock market. Journal of Business 59: 383-403.

CHEN, N. (1991). Financial investment opportunities and the macroeconomy. Journal of Finance 46: 529-554.

CHEN S., COWAN C.F.N., and GRANT P.M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks (2): 302-309.

CHENG, B. and TITTERINGTON, D.M. (1994). Neural networks: A review from a statistical perspective. Statistical Science 9(1): 2-54.

CHOLETTE, P.A. (1982). Prior Information and ARIMA forecasting. Journal of Forecasting 1: 375-384, 1982.

CHOPRA N.J., LAKONISHOK J., and RITTER, J.R. (1992). Measuring abnormal performance: do stocks overreact ? Journal of Financial Economics: 235-269.

CHOU, S.-R and JOHNSON, K. (1990). An empirical analysis of stock market anomalies: Evidence from the Republic of China in Taiwan. In: Pacific Basin Capital Markets Research, ed. by RHEE, S.G. and CHANG, R.P. VOL 1, North-Holland, Amsterdam.

CICHOSKI, A. and UNBEHAUEN, R. (1993). Neural networks for optimisation and signal processing. (New York: John Wiley).

CLARE, A.C. and THOMAS, S.H. (1994) Macroeconomic factors, the APT and the U.K. stock market. Journal of Business Finance and Accounting 21: 309-330.

CLARK, P. and NIBLETT, T. (1989). The CN2 Induction algorithm. Machine Learning 3:261-183.

CLEMEN, R.T. (1984). Modelling dependent information: a Bayesian approach. PhD Dissertation, Indiana University.

———— (1986). Linear constraints and the efficiency of combined forecasts. Journal of Forecasting 5: 31-38.

———— (1987). Combining overlapping information. Management Science 33(3): 373-380.

———— (1989). Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5: 559-583.

CLEMEN, R.T. and MURPHY, A.H. (1986a). Objective and subjective precipitation probability forecasts: statistical analysis of some interrelationships. Whether and Forecasting 1: 56-65.

———— (1986b). Objective and subjective precipitation probability forecasts: some methods for improving forecast quality. Whether and Forecasting 1: 213-218, 1986b.

COATS, P.K. and FANT, L.F. (1993). Recognizing financial distress patterns using a neural network tool. Financial Management (November): 142-155.

COCHRAN, W.G. (1963). Sampling techniques. (John & Wiley Sons).

COHEN, W.W. (1993). Efficient pruning methods for separate-and-conquer rule learning systems. In: THE 13TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, Proceedings, Chambery, France.

———— (1995). Fast effective rule induction. In: THE 12TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, Proceedings, Morgan Kaufmann.

COHRANE, J.H. (1991). Production-based net asset pricing and the link between stock returns and economic fluctuations. Journal of Finance 46: 209-238.

CONNOR, G. and UHLANER, R. (1988). New cross-sectional regression tests of beta pricing models. Working Paper. University of California at Berkeley, School of Business Administration, Chicago.

CONNOR, G. and KORAJCZYK, R.A. (1988). Risk and return in an equilibrium APT: Application of a new test methodology. Journal of Financial Economics 21: 255-289.

———— (1995). The arbitrage pricing theory and multifactor models of asset returns. In:

Handbooks in operations research and management science, ed. by JARROW R.A., MAKSIMOVIC V., and ZIEMBA W.T. (Elsevier SCIENCE B.V.), Vol 9, 87-145.

CONRAD, J. and KAUL, G. (1988). Time variation in expected returns. Journal of Business 61: 409-425.

CONROY, R. and HARRIS, R. (1987). Consensus forecasts of corporate earnings: analysts' forecasts and time-series methods. Management Science 33(6): 725-738.

CONWAY, D.A. and REINGANUM, M.R. (1988). Stable factors in security returns: Identification through cross validation. Journal of Business and Economics Statistics 6: 1-15.

COOKE R., MENDEL M., OORTMAN G., STOBBELAAR M., and VAN STEEN J. (1989). Expert opinions in safety studies, Faculty of Mathematics and Informatics, Delft University of Technology, Appeldoorn/Delf, Netherlands.

COOPER, J.P. and NELSON, C.R. (1975). The EX ante prediction performance of the St-Louis and FRB-MIT-PENN econometric models and some results on composite predictors. Journal of Money, Credit, and Banking 7: 1-32.

COOPER L.N., ELBAUM C., and REILLY D.L. (1982). Self-organizing general pattern class separator and identifier. In: Prototype Selection for Composite Nearest Neighbor Classifiers, SKALAK, D.B. (1997). Dissertation Thesis. University of Massachusetts Amherst. Dept of Computer Science.

COPELAND, T.E. and WESTON, J.F. (1992). Financial theory and corporate policy. (Addison-Wesley Publishing Company).

CORHAY A., HAWAWINI G., and MICHEL P. (1987). The pricing of equity on the London Stock Exchange: Seasonality and size premium. In: Stock Market Anomalies, ed. by DIMSON E. (Cambridge University Press, Cambridge).

COULSON, N.E and ROBINS, R.P. (1993). Forecast combination in a dynamic setting. Journal of Forecasting 12: 63-67.

CUTLER D. M., POTERBA J. M., and SUMMERS L.H. (1991). Speculative dynamics. Review of Economic Studies 58: 529-546.

CRAGG, J. and MALKIEL, B. (1968). The consensus and accuracy of some predictions of the growth in corporate earnings. Journal of Finance 23: 67-84.

CRAGG, J.G. and MACDONANLD, S.G (1992). Testing and determining arbitrage pricing structure from regressions on macro variables. Working Paper. University of British Columbia, Vancouver, BC.

CRANE, D.B. and CROTTY, J.R. (1967). A two-stage forecasting model: exponential

smoothing and multiple regression. Management Science 13: 501-507.

CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematical Control, Signal and Systems 2: 303-314.

DAY, N.E. and KERRIDGE, D.F. (1967). A general maximum likelihood discriminant. Biometrics 23: 313-324.

DEAKIN, E.B. (1976). Distributions of financial accounting ratios: some empirical evidence. The Accounting Review January: 90 - 96.

DEBONDT, W.F. and THALER, R. (1985). Does the stock market overreact ?. Journal of Finance 40: 793-805.

——— (1987) Further evidence on investors overreactions and stock market seasonality. Journal of Finance 42: 557-581.

DHRYMES P.J., FRIEND I., and GULTENKIN N.M. (1984) A critical reexamination of the empirical evidence on the arbitrage pricing theory. Journal of Finance 39: 323-346, 1984.

DIACOGIANNIS, G.P. (1986). Arbitrage pricing model: A critical examination of its empirical applicability for the London Stock Exchange. Journal of Business Finance and Accounting 13: 489-504.

DICKINSON, J.P. (1973). Some statistical results on the combination of forecasts. Operational Research Quarterly 24: 253-260.

——— (1975). Some comments on the combination of forecasts. Operational Research Quarterly 26: 205-210.

DIEBOLD, F.X. (1988). Serial correlation and the combination of forecasts. Journal of Business and Economics Statistics 6 (1), January: 105-111.

DIEBOLD, F.X. and PAULY, P. (1987). Structural change and the combination of forecasts. Journal of Forecasting 6: 21-40.

DIEBOLD, F.X. and LOPEZ, J.A. (1995). Forecast evaluation and combination. In: Handbook of statistics volume 14: statistical methods in finance, ed by MADDALA, G.S. and RAO, C.R. (Amsterdam: North Holland).

DIETTERICH, T.G. and BAKIRI, G. (1995). Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2: 263-286.

DONG, D. and MCAVOY, T.J. (1995). Non-linear principal component analysis - based on principal curves and neural networks. Computer Chemical Engineering 20(1): 65-78.

DRUCKER H., CORTES C., JACKEL L.D., LECUN Y., and VAPNIK V. (1994). Boosting

and other ensemble methods. Neural Computation 6(6): 1289-1301.

DRUCKER, H. and CORTES, C. (1996). Boosting decision trees. In: Advances in neural information processing systems 8, ed by TOURETZKY, D.S., MOZER, M.C., and HASSELMO. M.E. MIT Press, Las Vegas, NV, 479-485.

DUNIS C., LAWS J., and CHAUVIN, S. (2000). The use of market data and model combination to improve forecast accuracy. Working paper. Centre for International Banking, Economics, and Finance, Liverpool Business School (http://www.cibef.com).

DUNIS, C. and JALILOV, J. (2001). Neural network regression and alternative forecasting techniques for predicting financial variables. Working paper. Centre for International Banking, Economics, and Finance, Liverpool Business School (http://www.cibef.com).

DUNIS, C. and HUANG, X. (2001). Forecasting and trading currency volatility: an application of recurrent neural regression and model combination. Working paper. Centre for International Banking, Economics, and Finance, Liverpool Business School (http://www.cibef.com).

DUTTA, S. and SHEKHAR, S. (1988). In: THE IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, Proceedings, July: II443-II450.

EDELMAN, S. (1995). Representation, similarity, and the chorus of prototypes. Minds and Machines 5: 45-68.

EDWARDS, W. (1968). Conservatism in human information processing. In: formal representations of human judgements, ed. by KLEINMUNTZ, B. (New York: Wiley).

EFRON, B. and TIBSHIRANI, J. (1993). An introduction to the bootstrap. (London: Chapman and Hall).

EGGERT, R.J. (1976). Blue Chip Economic Indicators. Arlington, VA: Capitol Publications Inc.

EISENBEIS, R.A. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. Journal of Finance 32: 875 - 900.

ELTON, E.J., GRUBER, M.J., and GULTEKIN M. (1981). Expectations and share prices. Management Science 27: 975-987.

ELTON, E. J. and GRUBER, M.J. (1995). Modern portfolio theory and investment analysis. (John Wiley & Sons, Inc).

ENGLE, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50: 987-1008.

ENGLE R.F. and BOLLERSLEV (1986). Modelling the persistence of conditional variances. Econometric Reviews 5: 1-50.

ENGLE R.F., GRANGER C.W.J., and KRAFT D.F. (1985). Combining competing forecasts of inflation using a bivariate ARCH model. Journal of Economics Dynamics and Control 9: 67-85.

ENGLE R.F., LILIEN D., and ROBINS R.P. (1987). Estimating time-varying risk premia in the term structure: The ARCH-M model. Econometrica 55: 391-408.

ENGLISH, T.M. (1996). Stacked generalization and simulated evolution. In: Prototype Selection for Composite Nearest Neighbor Classifiers, SKALAK, D.B. (1997). Dissertation Thesis. University of Massachusetts Amherst. Dept of Computer Science.

FAIRFIELD, P.M. and HARRIS, T.S. (1993). Price-earnings and price-to-book anomalies: Tests of intrinsic value explanation. Contemporary Accounting Research 9: 590-611.

FAMA, E.F. (1965). The behaviour of stock market prices. Journal of Business 38: 34-105.

——— (1970). Multi-period consumption - investment decisions. American Economic Review 60: 163 - 174.

——— (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance 25: 383-417.

——— (1981). Stock returns, real activity, inflation and money. American Economic Review 71: 545-565.

——— (1984). Term premiums in bond returns. Journal of Financial Economics 13: 529-546.

——— (1990). Stock Returns, Expected Returns and Real Activity. Journal of Finance 45: 1089-1108.

——— (1991). Efficient capital markets II. Journal of Finance 46: 1575-1617.

FAMA, E.F. and FRENCH, K.R. (1988). Dividend yields and expected stock returns. Journal of Financial Economics 22: 3-25.

——— (1989). Business Conditions and Expected Returns on Stocks and Bonds. Journal of Financial Economics 25: 23-49.

——— (1992). The cross-section of expected stock returns. Journal of Finance 47: 427-465.

——— (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33: 3-56.

——— (1995). Size and book-to-market factors in earnings and returns. Journal of Finance 1 (1): 131-155.

FAMA, E.F. and MACBETH, J.D. (1973). Risk, return and equilibrium: empirical tests. Journal of Political Economy 81: 607-636.

FAMA, E.F. and SCHWERT, G.W. (1977). Asset returns and inflation. Journal of Financial Economics 5: 115-146.

FARIA, A.E. and SOUZA, R.C. (1995). A re-evaluation of the quasi-Bayes approach to the linear combination of forecasts. Journal of Forecasting 14: 533-542.

FERSON, W.E. (1990). Are the latent variables in time-varying expected returns compensation for consumption risk ? Journal of Finance 45: 397-429.

FERSON, W.E. (1995). Theory and empirical testing of asset pricing models. In: Handbooks in operations research and management science, ed. by JARROW R.A., MAKSIMOVIC V., and ZIEMBA W.T. (Elsevier SCIENCE B.V.), Vol . 9, pp. 145-200.

FERSON, W.E. and HARVEY, C.R. (1991a). The variation of economic risk premiums. Journal of Political Economy 99: 385-415.

———— (1991b). Sources of predictability in portfolio returns. Financial Analysts Journal 47: 49-56.

———— (1993a). An explanatory investigation of the fundamental determinants of national equity market returns. NBER Working Paper No. 4595.

———— (1993b). The risk and predictability of international equity returns. Review of Financial Studies 6: 527-566.

———— (1994). Sources of risk and expected returns in global equity markets. Journal of Banking and Finance 18: 775-803.

FERSON W.E., FOERSTER S.R., and KEIM D.B. (1993). General tests of latent variables models and mean variance spanning. Journal of Finance 48: 131-156.

FERSON, W.E. and FOERSTER, S.R. (1994). Small-sample properties of Generalised Method of Moments in tests of conditional asset pricing models. Journal of Financial Economics 36: 29-36.

FERSON, W.E. and KORAJCZYK, R.A. (1994). Do arbitrage pricing models explain the predictability of stock returns ?. Working Paper No. 115. Northwestern University.

FLANNERY, M. J. and JAMES, C. M. (1984). The effect of interest rate changes on the common stock returns of financial institutions. Journal of Finance 39 (4): 1141-1153.

FIRTH, M. (1979). The relationship between stock market returns and rates of inflation. Journal of Finance (June): 743-749.

FISHER, L. (1966). Some new stock-market indices. Journal of Business 39: 191-225.

FOERSTER, S. and KEIM, D.B. (1992). Direct evidence of non-trading and implications for daily

return autocorrelations. Unpublished Manuscript. University of Pennsylvania.

FOSTER G., OHLSON C., and SHEVLIN T. (1984). Earnings releases, anomalies and the behavior of security returns. Accounting Review 59: 574-603.

FRECKA, T.J. and HOPWOOD, W.S. The effects of outliers on the cross-sectional distributional properties of financial ratios. The Accounting Review: Vol LVIII(1): 115:127.

FRENCH K.R., and ROLL, R. (1986). Stock return variances: The arrival of information and the reaction of traders. Journal of Financial Economics 17: 5-26.

FRENCH K.R., SCHWERT W., and STAMBAUGH R. (1987). Expected stock returns and volatility. Journal of Financial Economics 19, pp. 3-30.

FRENCH, S. (1980). Updating of belief in the light of someone else's opinion. Journal of the Royal Statistical Society, Series A: 43-48.

———— (1981). Consensus of opinion. European Journal of Operational Research 7: 332-340.

FREUND, Y. and SCHAPIRE, R.E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In: THE 2ND EUROPEAN CONFERENCE ON COMPUTATIONAL LEARNING THEORY, Proceedings (Barcelona, Spain: Springer Verlag), 23-37.

FREUND Y., IYER R., SCHAPIRE R.E., AND SINGER Y. (1998). An efficient boosting algorithm for combining preferences, In: THE 15TH INTERNATIONAL CONFERENCE, Proceedings: Machine Learning.

FRIEDMAN, J.H. and STUETZLE, W. (1981). Projection Pursuit Regression. Journal of American Statistical Association 76: 817-823.

FRIEDMAN, J.H. (1991). Multivariate adaptive regression splines. Annals of Statistics 19: 1-141.

FRIEND I., LANSKRONER Y., and LOSQ E. (1976). The demand for risky assets and uncertain inflation. Journal of Finance 5: 1287-1297.

FRITZ R., BRANDON C., and XANDER J. (1984). Combining time-series and econometric forecast of tourism activity. Annals of Tourism Research 11: 219-229.

FRITZKE, B. (1994). Fast learning with incremental RBF networks. Neural Processing Letters (1): 2-5.

FURNKRANZ, J. and WIDMER, G. (1994). Incremental reduced error pruning. In: THE 11TH ANNUAL CONFERENCE ON MACHINE LEARNING, Proceedings, New Brunswick, New Jersey.

GARDNER, E.S in ARMSTRONG ET AL. (1983). Commentary on the Makridakis time- series competition (M-Competition). Journal of Forecasting 2: 259-311.

GEISEL, M. (1973). Bayesian comparisons of simple macroeconomic models. Journal of Money, Credit and Banking 5: 751-772.

GEURTS. M.D. in ARMSTRONG ET AL. (1983). Commentary on the Makridakis time-series competition (M-Competition). Journal of Forecasting 2: 259-311.

GESKE, R. and ROLL, R. (1983). The fiscal and monetary linkage between stock returns and inflation. Journal of Finance 28 (March): 7-33.

GIBBONS, M. and HESS, P. (1981). Day of the week effects and asset returns. Journal of Business 54: 579-596.

GIBBONS, M.R. and FERSON, W.E. (1985). Testing asset pricing models with changing expectations and an unobservable market portfolio. Journal of Financial Economics 14: 217-236.

GILBERT, E.S. (1969). The effect of unequal variance-covariance matrices on Fisher's linear discrimination function. Biometrics 25: 505-515.

GLOSTEN C.R., JAGANNATHAN R., and RUNKLE D.E. (1993). On the relation between the expected value and the volatility of the nominal excess returns on stocks. Journal of Finance 48: 1779-1802.

GOETZMANN, W. N. and JORION, P. (1993). Testing the predictive power of dividend yields. Journal of Finance 48 (2): 663-679.

GOLDBERG, L.R. (1965). Diagnostians versus diagnostic signs: the diagnosis of psychosis vs neurosis from MMPI. Psychological Monographs 79: 1965.

———— (1970). Man versus model of man: a rationale plus some evidence for a method of improving on clinical interfaces. Psychological Bulletin 73: 422-432.

GOLDING, A. R. and ROSENBLOOM, P. S. (1991). Improving rule-based systems through case-based reasoning. In: THE 9TH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, Proceedings, Anaheim, CA. American Association for Artificial Intelligence. 22-27.

GOODMAN D.A., PEAVY J.W., and COX E.L. (1986). The interaction of firm-size and price-earnings ratio on portfolio performance. Financial Analysts Journal. January-February: 9-12.

GORDON, K. and SMITH, A.F.M. (1988). Modelling and monitoring discontinuous changes in the time-series. In: Bayesian Analysis of Time Series and Dynamic Models, ed. by SPALL, J.C (New York: Marcel Dekker), 359-391.

——— (1990). Modelling and monitoring biomedical time series. Journal of the American Statistical Association 85: 328-337.

GORMAN, R.P. and SEJNOWSKI, T.J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. Neural Networks 1 (Part A): 75-89.

GOURIEROUX, C. and MONFORT, A. (1992). Qualitative threshold ARCH models. Journal of Econometrics 52: 159-200.

GRANGER, C.W.J. and NEWBOLD, P. (1975). Economic forecasting: the atheist's viewpoint. In: Modelling the Economy, ed. by RENTON, G.A. (London: Heineman), 131-147.

GRANGER, C.W.J. and RAMANATHAN, R. (1984). Improved methods of combining forecasts. Journal of Forecasting 3: 197-204.

GREENE M.N., HOWREY E.P., and HYMANS S.W. (1986). The use of outside information in econometric forecasting. In: model reliability, ed. by KUN, E. and BELSLEY, D.A. (Cambridge, MA:MIT Press).

GREIG, A.C. (1992). Fundamental analysis and subsequent stock returns. Journal of Accounting and Economics 15: 413-442.

GUERARD, J.B. (1987). Linear constraints, robust weighting and efficient composite modelling. Journal of Forecasting 6: 193-199.

GUERARD, J.B. (1989). Composite model building for foreign exchange rates. Journal of Forecasting 8: 315-329.

GUERARD, J.B. and BEIDLEMAN, C.R. (1987). Composite earnings forecasting efficiency. Interfaces 17: 103-113.

GULTEKIN, N.B. (1983). Stock market returns and inflation: evidence from other countries. Journal of Finance 38 (1): 49-65.

GULTEKIN, M. and GULTEKIN, B. (1983). Stock market seasonality: international evidence. Journal of Financial Economics 12: 469:482.

GUNTER, S.I. and AKSU, C. (1997). The usefulness of heuristic N(E)RLS algorithms for combining forecasts. Journal of Forecasting 16: 439-463.

GUPTA, S. and WILTON, P.C. (1987). Combination of forecasts: an extension. Management Science 33 (3): 356-372, March.

——— (1988). Combination of economic forecasts: an odds-matrix approach. Journal of Business and Economic Statistics 6: 373-379.

HAFER, R.W. and HEIN, S.E. (1985). On the accuracy of time-series, interest rate, and survey

forecasts of inflation. Journal of Business 58: 377-398.

HAN J., CAI Y., and CERCONE N. (1992). Knowledge discovery in databases: an attribute oriented approach. In: THE 18TH VLDB CONFERENCE, Proceedings, Vancouver, British Columbia, Canada 1992, 547-559.

HAND, D.J. (1983). A comparison of two methods of discriminant analysis applied to binary data. Biometrics 39: 683-694.

———— (1987). A shrunken leaving-one-out estimator of error rate. Computers and Mathematics with Applications 14: 161-167.

———— (1997). Construction and assessment of classification rules. (Chichester: John Wiley & Sons Ltd).

HANSEN, L.R. and HODRICK, R.J. (1983). Risk averse speculation in the forward foreign exchange market: An econometric analysis of linear models. In: Exchange rates and international economics, ed. by FRENKEL, J. (University of Chicago Press), Chicago 1983, IL.

HANSEN, L.K. and SALAMON, P. (1990). Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12:993-1001.

HARRIS, L. (1986). A transaction data study of weekly and intraday patterns in stock returns. Journal of Financial Economics 16: 99-117.

HARRISON, P.J. and STEVENS, C.F. (1976). Bayesian forecasting. Journal of the Royal Statistical Society, Series B, 38: 205-247.

HARVEY, C.R. (1989). Time-varying conditional covariances in tests of asset pricing models Journal of Financial Economics 24: 289-317.

HARVEY, C.R. (1991). The world price of covariance risk. Journal of Finance 46: 111-157.

HAWAWINI, G. and VIALLET, C. (1987). Sesonality, risk premium and the relationship between risk and return of French common stocks. Working Paper. INSEAD and the Wharton School of the University of Pennsylvania.

HAWAWINI G., MICHEAL P., and CORHAY A. (1989). A look of the validity of the capital asset pricing model in light of equity market anomalies: The case of Belgian common stocks. In: A reappraisal of the efficiency of financial markets, ed. by S. Taylor, NATO, ASI series. (Springer -Verlag).

HAWAWINI, G. and KEIM, D.B. (1995). On the predictability of common stock returns. In: Handbooks in operations research and management science, ed. by JARROW R.A., MAKSIMOVIC V., and ZIEMBA W.T. (Elsevier SCIENCE B.V.), 9, 497-544.

HAYKIN, S. (1994). Neural networks - a comprehensive foundation (New York: MacMillan

College Publishing Company).

HEATH D., KASIF S., and SALZBERG S. (1996) Committees of decision trees. In: Cognitive technology: in search of a human interface, ed. by GORAYSKA, B. and MEY, J. Amsterdam, the Netherlands, Elsevier Science, 305-317.

HENDRY, D.F. and MIZON, G. (1978). Serial correlation as a convenience not a nuisance: a comment on a study of the demand for money by the Bank of England. Economic Journal 88: 549-563.

HICKMAN, W.B. (1958). Corporate bond quality and investor experience. National Bureau of Economic Research, New York.

HINICH, M.J. and PATTERSON, D.M. (1985). Evidence of non-linearity in daily stock returns. Journal of Business and Economic Statistics 3: 69-77.

HO T.K., HULL J.J., and SRIHARI S.N. (1994). Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 16: 66-75.

HODRICK, R.J. (1992). Dividend yields and expected stock returns: alternative procedures for inference and measurement. Review of financial studies 5(3): 357-386.

HOLDEN, K. and PEEL, D.A. (1989). Unbiasedness, efficiency and the combination of economic forecasts. Journal of Forecasting 8: 175-188.

HOLDEN, K. and THOMPSON, J. (1997). Combining forecasts, encompassing and the properties of U.K. macroeconomic forecasts. Applied Economics 29: 1447-1458.

HOLLIFIELD, B. (1993). Linear asset pricing with time-varying betas and risk premia. Working Paper. University of British Columbia, Vancouver, BC.

HOLSEIMER, M. and SIEBES, A. (1991). Data Mining - the search for knowledge in databases. Report CS-R9406. ISSN 0169-118X, CWI. P.O. Box 94079 GB, Amsterdam. The Netherlands.

HOLTHAUSEN, R.W. and LARCKER, D.F. (1992). The prediction of stock returns using financial statement information. Journal of Accounting and Economics 15: 373-411.

HORNIK K.J., STINCHCOMBE M., and WHITE H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks 2: 359-366.

HORRIGAN, J.O. (1966). The determination of long-term credit standing with financial ratios. Empirical research in accounting: selected studies. Journal of Accounting Research 4 (supplement): 44-62.

HUANG, W.Y. and LIPPMANN, R.P. (1987). Comparisons between neural net and conventional classifiers. In: THE IEEE 1ST INTERNATIONAL CONFERENCE ON

NEURAL NETWORKS, Proceedings, Piscataway, New Jersey 1987, 485-494.

JACOBS, B. and LEVY, K. (1988). Disentangling equity return regularities: New insights and investment opportunities. Financial Analysts Journal, May-June: 18-43.

JACOBS R.A., JORDAN M.I., NOWLAN S.J., and HINTON G.E. (1991a). Adaptive mixtures of local experts. Neural Computation 3(1): 79-87.

JACOBS R.A., JORDAN M.I., and BARTO A.G. (1991b). Task decomposition through competition in a modular connectionist architecture: the what and where vision tasks. Cognitive Science 15:219-250.

JAFFE, J. and MANDELKER, G. (1976). The Fisher effect for risky assets: An empirical investigation. Journal of Finance 31: 335-367, May.

JAFFE J., KEIM D., and WESTERFIELD R. (1989). Earnings yields, market values and stock returns. Journal of Finance 44: 135-148.

JAMES C., KOREISHA S., and PARTCH M. (1985). A VARMA analysis of the causal relations among stock returns, real output, and nominal interest rates. Journal of Finance 40 (5): 1375-1384.

JEGADEESH, N. (1992). Does market size really explain the size effect ? Journal of Financial and Quantitative Analysis 27: 337-351.

JOHN G., MILLER P., and KERBER R. (1996). Stock selection using Recon. In: THE 3RD INTERNATIONAL CONFERENCE ON NEURAL NETWORKS IN THE CAPITAL MARKETS, Proceedings, London Business School, London 1996.

JOLLIFFEE, I.T. (1986). Principal component analysis. (Springer-Verlag: New York).

JONES, C.P. and LITZENBERGER, R.H. (1970). Quarterly earnings reports and intermediate stock price trends. Journal of Finance 25: 143-148.

JONES, S.L. (1993). Another look of time-varying risk and return in a long-horizon contrarian strategy. Journal of Financial Economics 33: 119-144.

JORDAN, M.I. and JACOBS, R.A. (1993). Hierarchical mixtures of experts and the EM algorithm. Technical Report 1440. Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA.

JUTTEN, C. (1995). The ELENA project - enhanced learning for evolutive neural architecture. ESPRIT Basic Research Project Number 6891 (see http://www.dice.ucl.ac.be/neural-nets/ELENA.html).

KAPLAN A., SKOGSTAD A.L., and GIRSHICK M.A. (1950). The prediction of social and technological events. Public Opinion Quarterly 14: 93-110.

KAPLAN, S.R. and URWITZ, G. (1979). Statistical models of bond ratings: A methodological inquiry. Journal of Business 52(2): 231-261.

KARELS, G.V. and PRAKASH, A. (1987). Multivariate normality and forecasting of business bankruptcy. Journal of Business Finance and Accounting 14: 573-593.

KARHUNEN, J. and JOUTSENSALO, J. (1994). Representation and separation of signals using non-linear PCA type learning. Neural Networks 7(1): 113-127.

KEHAGIAS, A. and PETRIDIS, V. (1997). Predictive modular neural networks for time series classification. Neural Networks 10(1): 31-49.

KEIM, D.B. (1983) Size related anomalies and stock return seasonality. Journal of Financial Economics 12: 13-32.

———— (1988). Stock market regularities: A synthesis of the evidence and explanations. In: Stock market anomalies, ed. by DIMSON, E. (Cambridge University Press), Cambridge, 16-39.

KEIM, D.B. and STAMBAUGH, R.F. (1984). A further investigation of the weekend effect in stock returns. Journal of Finance 39: 819-835.

———— (1986). Predicting returns in the stock and bond markets. Journal of Financial Economics 17: 357-390.

KENDALL M.G., STUART A., and ORD J.K. (1983). The advance theory of statistics. Vol. 3: Design and analysis and time series. Chapter 44. (London: Griffin).

KIM M.J., NELSON C.R., and STARTZ R. (1991). Mean reversion in stock prices ? A reappraisal of the empirical evidence. Review of Economics Studies 58: 515-528.

KIRWOOD C., ANDREWS B., and MOWFORTH P. (1989). Automatic detection of gait events: a case study using inductive learning techniques. Journal of Biomedical Engineering 11(23): 511-516.

KLEIN, R.W. and BAWA, V.S. (1976). The effect of estimation risk on optimal portfolio choice. Journal of Financial Economics 3: 215-231.

KOHN, R. (1982). When is an aggregate of a time-series efficiently forecast from its past ? Journal of Econometrics 18: 337-349.

KOHONEN T., BARNA G., CHRISLEY R. (1988). Statistical pattern recognition with neural networks: Benchmarking studies. In: IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, Proceedings, San Diego 1998, CA, 161-168.

KOHONEN, T. (1989). Self-organization and associative memory (Berlin: Springer Verlag).

KOHONEN T., HYNNINEN J., KANGAS J., LAAKSONNEN J., and TORKKOLA, K.

(1995). LVQ_PAK - The learning vector quantization programme package. Version 3.1. LVQ Programming Team. Helsinki University of Technology. Laboratory of Computer and Information Science. FILAND.

KONG, E.B. and DIETTERICH, T.G. (1995). Error-correcting output coding corrects bias and variance. In: THE 12TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, Proceedings (San Francisco, CA: Morgan Kaufmann), 313-321.

KOTHARI, S.R. and SHANKEN, J. (1992). Stock return variation and expected dividends: A time-series and cross-sectional analysis. Journal of Financial Economics 31 (2): 177-211.

———— (1995). Book-to-market, dividend yield, and expected market returns. Working Paper 95-13. University of Rochester.

KOTHARI S.P., SHANKEN J., and SLOAN R. (1995). Another look at the cross-section of expected returns, Journal of Finance L(1): 185-224.

KOTON, P.A. (1988). Using experience in learning and problem solving. Ph.D. Dissertation. Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA.

KRAMER, A.M. (1991). Non-linear principal component analysis using autoassociative neural networks. American Institute of Chemical Engineers Journal 37(21): 233-243. February.

KRZANOWSKI, W.J. and MARRIOTT, F.H.C. (1995). Multivariate analysis. (London: Edward Arnold).

KUBAT, M. (1998). Decision trees can initialise radial-basis-function networks. IEEE Transactions on Neural Networks 9: 813-821.

LACHENBRUCH P.A., SNEERINGER C., and REVO L.T. (1973). Robustness of the linear and quadratic discriminant function to certain types of nonnormality. Communication Statistics 1: 39-56.

LACHENBRUCH, P.A. (1975). Discriminant analysis. (New York: Sage Publications).

LAKONISHOK, J. and SMIDT, S. (1988). Are seasonal anomalies real ? A ninety year perspective. Review of Financial Studies 1: pp. 403-425.

LAKONISHOK J., SHLEIFER A., and VISHNY R.W. (1994) Contrarian investment, extrapolation and risk. Journal of Finance 49: 1541-1578.

LAWRENCE, K.D. and REEVES, G.R. (1981). Consensus time-series forecasting. In: organisations: multiple agents with multiple criteria, ed. by MORSE, J. (New York: Springer-Verlag), 199-204.

LAWRENCE M.J., EDMUNDSON R.H., and O'CONNOR M.J. (1985). An examination of the

accuracy of judgmental extrapolation of time-series. International Journal of Forecasting 1: 25-35.

LEE, B.-S. (1992). Causal relations among stock returns, interest rates, real activity, and inflation. Journal of Finance 47 (4): 1591-1603.

LEHMANN, B.N. and MODEST, D. (1988) The empirical foundations of the arbitrage pricing theory. Journal of Financial Economics 21: 213-254.

LEHMANN, B.N. (1990). Fads, martingales, and market efficiency. Quarterly Journal of Economics 105: 1-28.

LEUNG M.T., DAOUK H., and CHEN A.-S. (2000). Forecasting stock indices: a comparison of classification and level estimation models. International Journal of Forecasting 16: 173-190.

LEROY, S.F. and PORTER, R.D. (1981). The present value relation: tests based on implied variance bounds. Econometrica 49: 555-574.

LESAGE, J.P. and MAGURA, M. (1992). A mixture-model approach to combining forecasts. Journal of Business and Economic Statistics 10(4): 445-452, October.

LEV, B. and THIAGARAJAN, S.R. (1993). Fundamental information analysis. Journal of Accounting Research 31: 190-215.

LEVIN, A.U. (1995). Stock selection via multifactor models. In: advances in neural information processing systems (San Francisco: Morgan Kaufmann).

LEVIS, M. (1989a). Market size, P/E ratios, dividend yield and share prices: The U.K. evidence. In: A Reappraisal of the Efficiency of the Financial Markets, ed. by GUIMARAES R.C., KINGSMAN B.G, and TAYLOR S.J. (Springer-Verlag).

———— (1989b). Stock Market Anomalies: A re-assessment on the U.K. evidence. Journal of Banking and Finance 13: 675-696.

———— (1995). Macroeconomic studies in size and value strategies. Working Paper. City University Business School, Department of Accounting and Finance.

LINDLEY, D. (1983). Reconciliation of probability distributions. Operational Research 31: 866-880.

———— (1985). Reconciliation of discrete probability distributions. In: Bayesian Statistics 2, Bernardo J. et al., North-Holland, Amsterdam, 375-390.

LINTNER, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47: 13-37.

LITTLESTONE, N. (1987). Learning quickly when irrelevant attributes abound: a new

linear-threshold algorithm. In: THE 28TH ANNUAL SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE, Proceedings, Washington, DC. IEEE Computer Society Press, 68-77.

LITTLESTONE, N. and WARMUTH, M. (1989). The weighted majority algorithm. In: THE 30TH ANNUAL SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE, Proceedings, Washington, DC. IEEE Computer Society Press. 256-261.

LIU, D.C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimisation. Mathematical Programming 45: 503-528.

LO, A.W. and MACKINLAY, A.C. (1988). Stock prices do not follow random walks: Evidence from a simple specification test. Review of Financial Studies 1: 41-66.

———— (1990a). Data Snooping Biases in tests of financial asset pricing models. Review of Financial Studies 3: 431-467.

———— (1990b). An econometric analysis of non-synchronous trading. Journal of Econometrics 45: 181-211.

———— (1999). A Non-random walk down Wall street. (New York: Princeton University Press).

LONGBOTTOM, J.A. and HOLLY, S. (1985). The role of time-series analysis in the evaluation of econometric models. Journal of Forecasting 4: 75-87.

LUTKEPOHL, H. (1984). Forecasting contemporaneously aggregated vector ARMA processes. Journal of Business and Economic Statistics 2: 201-214.

MACGREGOR, J. (1989). Multivariate statistical methods for monitoring large data sets from chemical processes. Paper 164a. American Institute of Chemical Engineers Meeting. San Francisco.

MACLIN, R. AND SHAVLIK, J.W. (1995). Combining the predictions of multiple classifiers: using competitive learning to initialize neural networks. In: THE 14TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, Proceedings, 524-530 (San Mateo, CA: Morgan Kaufmann).

MAKRIDAKIS, S. (1989). Why combining works ? International Journal of Forecasting 5: 601-603.

MAKRIDAKIS S., ANDERSEN A., CARBONE R., FILDES R., HIBON M., LEWANDOWSKI R., NEWTON J., PARZEN E., and WINKLER R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. Journal of Forecasting 1: 111-153.

MAKRIDAKIS S., ANDERSEN A., CARBONE R., FILDES R., HIBON M.,

LEWANDOWSKI R., NEWTON J., PARZEN E., and WINKLER R. (1983). The forecasting accuracy of major time-series methods (London: Wiley).

MAKRIDAKIS, S. and WINKLER, R.L. (1983). Averages of forecasts: some empirical results. Management Science 29: 987-996.

MALKIEL, B.G. (1996). A random walk down Wall street. (New York: W.W. Norton & Company).

MALKOVICH, J.F. and AFIFI, A. (1973). On tests for multivariate normality, Journal of the American Statistical Association 68: 176-179.

MALTHOUSE, C. (1996). Non-linear partial least squares. Ph.D. Thesis. Northwestern University. Graduate School of Management.

MANI, G. (1991) Lowering variance of decisions by artificial neural network ensembles. Neural Computation 3: 484-486.

MANKIEW N.G., ROMER D., and SHAPIRO M.D. (1985). Stock market forecastability and volatility: A statistical appraisal. Review of Economic Studies 58: 455-477.

MARDIA, K. (1971). The Effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrica 58: 105-121.

MARKOWITZ, H.M. (1959) Portfolio selection: efficient diversification of investments (Willey, New York).

MASTERS, T. (1993). Practical neural network recipes in C++ (New York: Academic Press).

———— (1995) Advanced algorithms for neural networks. A C++ Sourcebook (New York: Academic Press).

MCLACHLAN, G.J. (1992). Discriminant analysis and statistical pattern recognition. (New York: John Wiley).

MCNEES, S.K. (1987). Consensus forecasts: tyranny of the majority ? New England Economic Review 15-21, November/December, 1987.

———— (1992). The uses and abuses of consensus forecasts. Journal of Forecasting 11: 703-710.

MCQUEEN, G. (1992). Long-horizon mean-reverting stock prices revisited. Journal of Financial and Quantitative Analysis 28: 331-345.

MEI, J. and SAUNDERS, A. (1994) The time-variation of risk premiums on insurer stocks. Journal of Risk and Insurance 61: 12-32.

MERTON, R.C. (1973) An intertemporal capital asset pricing model. Econometrica 41:

867-887.

——— (1980). On estimating the expected return on the market: An exploratory investigation. Journal of Financial Economics 8, 323-362.

MICHALSKI R. S., GARBONELL J.G., and MITCHELL T.M. (1986). Machine learning, an artificial intelligence approach. Volumes I, II. (San Mateo, California: Morgan Kaufmann).

MICHIE, D. (1989). Problems of computer aided concept formation. In: Applications of expert systems, ed. by QUINLAN, J.R. (London: Addison-Wesley). Volume 2: 310-333.

MICHIE D., SPIEGELHAFTER D., and TAYLOR C.C. (1994). Machine learning and statistical classification. Ellis Horwood (see www.amsta.leeds.ac.uk/~charles/statlog).

MILLS, T.C. and STEPHENSON, M.J. (1985). Forecasting contemporaneous aggregates and the combination of forecasts: the case of the U.K. monetary aggregates. Journal of Forecasting 4: 273-281.

——— (1987). A time series forecasting system for the U.K. money supply. Economic Modelling, July: 355-369.

MINSKY, M.L. and PAPERT, S.A. (1969). Perceptrons. (Cambridge, MA: MIT Press).

MOODY J., LEVIN U., and RDHFUSS S. (1993). Predicting the U.S. index of industrial production. Neural Network World 3(6): 791-794.

MORIARTY, M.M. and ADAMS, A.J. (1984). Management judgement forecasts, composite forecasting models, and conditional efficiency. Journal of Marketing Research 21: 239-250.

MORRIS, P. (1974). Decision analysis expert use. Management Science 20: 1233-1241.

——— (1977). Combining experts judgements: a Bayesian approach. Management Science 23: 679-693.

MOSSIN, J. (1966). Equilibrium in a capital asset market. Econometrica 34: 768-783.

MURPHY A.H., CHEN Y.-S, and CLEMEN, R.T. (1988). Statistical analysis of interrelationships between objective and subjective temperature forecasts. Monthly Whether Review 116: 2121-2131.

MURPHY, P.M. and AHA, D.W. (1994). University of California at Irvine Repository of Machine Learning Databases. (For information contact *ml-repository@ics.uci.edu*).

MURTHY, K. V. S. (1997). On growing better decision trees from data. Dissertation. John Hopkins University. Baltimore, Maryland.

NELSON, C.R. (1972). The prediction performance of the FRB-MIT-PENN model of the U.S. Economy. American Economic Review 63: 902-917.

———— (1976). Inflation and rates of return on common stocks. Journal of Finance 31: May.

NELSON, D. (1991). Conditional heteroscedasticity in asset returns: A new approach. Econometrica 59: 347-370.

NELSON, C. R. and KIM, M. J. (1993). Predictable stock returns: the role of small sample bias. Journal of Finance 48 (2): 641-661.

NEWBOLD, P. and GRANGER, C.W.J. (1974). Experience with forecasting univariate time-series and the combination of forecasts (with discussion). Journal of the Royal Statistical Society, Series A, 137: 131-149.

NEWBOLD P., ZUMWALT J.K, and KANNAN S. (1987). Combining forecasts to improve earnings per share prediction: an examination of electric utilities. International Journal of Forecasting 3: 229-238.

NG, L. (1991). Tests of the CAPM with time-varying covariances: a multivariate GARCH approach. Journal of Finance 46: 1507-1521.

NICHOLSON, S.F (1960). Price-earnings ratios. Financial Analysts Journal. July-August: 43-50.

NILSSON, N.J. (1990). The mathematical foundations of learning machines. (San Mateo, CA: Morgan Kaufmann).

ODOM, M.D. and SHARDA, R.A. (1990). A neural network model for bankruptcy prediction. In: THE IJCNN INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, Proceedings, California, San Diego 1990, Vol. II: 163-168.

OHLSON, D. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research Spring: 109-131.

OJA, E. (1992). Principal components, minor components, and linear neural networks. Neural Networks 5: 927-935.

OLDFIELD, G.S. and ROGALSKI, R.J. (1980). Treasury bills factors and common stock returns. Journal of Finance 36 (2): 337-350.

OPITZ, D.W. and SHAVLIK, J.W. (1995) Generating accurate and diverse members of a neural-network ensemble. In: Advances in Neural Information Processing Systems 8 ed. by Touretzky D.S.; Mozer M.C.; and Hasselmo M.E. MIT Press, Cambridge, MA, 535-541.

OU, J.A. and PENMAN, S.H. (1989a). Financial statement analysis and the prediction of stock returns, Journal of Accounting and Economics 11: 295-329.

———— (1989b). Accounting measurement, P/E ratios and the information content of security prices. Journal of Accounting Research 27 (Supplement): 111-144.

PAGGALLO, G. and HAUSLER, D. (1990). Boolean feature discovery in empirical learning. Machine Learning 5(1), 1990.

PARZEN, E. (1962). On estimation of a probability density function and mode. Annals of Mathematics and Statistics 33: 1065 - 1076.

PAWELZIK K., KOHLMORGEN J., and MULLER, K.R. (1996). Annealed competition of experts for a segmentation and classification of switching dynamics. Neural Computation 8(2): 340-356.

PEAVY, J.W. and GOODMAN, D.A. (1983). Industry relative price-earnings ratios as indicators of investments returns. Financial Analysts Journal 39 (4): 60-66.

PERRONE, M.P. (1993). Improving regression estimation: averaging methods for variance reduction with extensions to general convex measure optimization. Ph.D. Dissertation. Dept. of Computer Science, Brown University, Providence, RI.

PESARAN, M.H. and TIMMERMANN, A. (1994). Forecasting stock returns. An examination of stock market trading in the presence of transaction costs. Journal of Forecasting 13: 335-367.

———— (1995). Predictability of stock returns: robustness and economic significance. Journal of Finance 50: 1201-1228.

PETTENGILL G.N., SUNDARAN S., and MATHUR I. (1995). The conditional relation between beta and returns. Journal of Financial and Quantitative Analysis 30: 101-115.

PHILLIPS, L.D. and EDWARDS, W. (1966). Conservatism in a simple probability inference task. Journal of Experimental Psychology 72: 346-357.

PINCHES, G.E. and MINGO, K.A (1973). A Multivariate analysis of industrial bond ratings. The Journal of Finance Vol. XXVIII(1): 1-17.

———— (1975). The role of subordination and industrial bond ratings. The Journal of Finance Vol. XXX(1) (March): 201-206.

PINCHES, R.A. (1980). Pitfalls in the application of discriminant analysis in business, finance, and economics. Journal of Finance 3: 875-900.

POGUE, T. F. and SOLDOFSKY, R.M. (1969). What's in a bond rating ? Journal of Financial and Quantitative Analysis 4: 201-228, June.

POON, S. and TAYLOR, S.J. (1991). Macroeconomic factors and the U.K. stock market. Journal of Business Finance and Accounting 18 (5): 619-637.

POON, S.H. and TAYLOR, S.J. (1992). Stock returns and volatility: An empirical study of the U.K. stock market. Journal of Banking and Finance 16: 37-59.

POSKITT, D.S. and TREMAYNE, A.R. (1986). The selection and use of linear and bilinear time-series models. International Journal of Forecasting 2: 101-114.

POTERBA, J. M. and SUMMERS, L.H. (1988). Mean reversion in stock prices: evidence and implications. Journal of Financial Economics 22: 27-59.

QI, M. and MADDALA, G.S. (1999). Economic factors and the stock market: a new perspective. Journal of Forecasting 18: 151-166.

QUINLAN, J.R. (1983) Learning efficient classification procedures and their application to chess end games. In: Machine learning, an artificial intelligence approach, ed. by MICHALSKI R. S., GARBONELL J.G., and MITCHELL T.M., Volume 1. San Mateo, Califirnia, Morgan Kaufmann.

———— (1986). Induction of decision trees. Machine Learning 1:81-106.

———— (1996a). Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research 4: 77-90.

———— (1996b). Bagging, Boosting, and C4.5. In: THE 13TH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, Proceedings, AAAI Press/MIT Press, Menlo Park, CA, 275-730.

REEVES, G.R. AND LAWRENCE, K.D. (1982). Combining multiple forecasts given multiple objectives. Journal of Forecasting 1: 271-279.

REFENES, A.N. (1994). Stock Performance Modelling Using Neural Networks: A Comparative Study with Regression Models. Neural Networks 7(2): 375-388.

REID, D.J. (1968). Combining three estimates of gross domestic product. Economica 35: 431-444.

———— (1969). A comparative study of time-series prediction techniques on economic data. PhD Thesis, University of Nottingham, Nottingham.

REINGANUM, M.R. (1981a). A new empirical perspective on the CAPM. Journal of Financial and Quantitative Analysis 4: November.

———— (1981b). Misspecification of capital asset pricing: empirical anomalies based on earnings yields and market values. Journal of Financial Economics 12: 89-104.

———— (1982). A direct test of Roll's conjecture on the firm size effect. Journal of Finance 37: 27-35.

———— (1992). A Revival of the small-firm effect, Journal of Portfolio Management: 55-62.

REINMUTH, J.E. and GEURTS, M.D. (1979). A multideterministic approach to forecasting.

In: Forecasting, 12, TIMS Studies in the Management Sciences, ed. by MAKRIDAKIS, S. and WHEELWRIGHT, S.C., North-Holland, New York, 203-211.

RIISE, T. and TJOSTHEIM, D. (1984). Theory and practice of multivariate ARMA forecasting, Journal of Forecasting 3: 309-317.

RIPLEY, B.D. (1992). Statistical aspects of neural networks. In: Networks and chaos-statistical and probabilistic aspects, ed. by BARNDORFF- NIELSEN O.E., JENSEN J.L., and KENDALL W.S., Chapman and Hall, London.

———— (1993). Statistical aspects of neural networks. In: Chaos and networks - statistical and probabilistic aspects, ed. by BARNDORFF- NIELSEN O., COX D., JENSEN J., and KENDALL W. (Chapman and Hall).

RISSLAND, E.L. and SKALAK, D.B. (1991). CABARET: rule interpretation in a hybrid architecture. International Journal of Man-Machine Studies 34: 839-887, 1991.

RISSLAND E.L., DANIELS J.J., RUBINSTEIN Z.B., and SKALAK D.B. (1993). Case-based diagnostic analysis in a blackboard architecture. In: THE 11TH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, Proceedings, San Mateo, CA, AAAI Press/MIT Press, 66-72.

ROBERTS, H. (1965). Probabilistic prediction. Journal of the American Statistical Society 60: 50-62.

ROLL, R. (1977). A critique of the asset pricing theory tests. Journal of Financial Economics: March: 129-176.

———— (1981). A possible explanation of the firm size effect. Journal of Finance 36: 879-888.

ROLL, R. and ROSS, S. (1980). An empirical investigation of the arbitrage pricing theory. Journal of Finance 35: 1073-1103.

ROSE, D.E. (1977). Forecasting aggregates of independent ARIMA processes. Journal of Econometrics 5: 323-325.

ROSENBERG B., REID K., and LANSTEIN R. (1985). Persuasive evidence on market inefficiency. Journal of Portfolio Management 11(3): 9-17.

ROSENBLATT, F. (1962). Principles of neorodynamics. (Washington DC: Spartan Books).

ROSS, S. (1976). The arbitrage theory of the capital asset pricing. Journal of Economic Theory 13: 341-360.

ROSS, I. (1976). Higher stakes in the bond-rating game. Fortune (April): 133-142.

ROZEFF, M. (1984). Dividend yields and equity risk premium. Journal of Portfolio Management 11: 68-75.

RUBIO, G. (1988). Further international evidence on asset pricing: The case of the Spanish capital market. Journal of Banking and Finance 12: 221-242.

RUBISTEIN, M. (1976). The strong case for the generalized logarithmic utility model as the premier model of financial markets. Journal of Finance 31: 551-571.

RUMELHART D.E., HINTON G.E., and WILLIAMS R.J. (1986). Learning internal representations by backpropagation errors. Nature 323: 533-536.

RUSSELL, T.D. and EVERETT, E.A. JR (1987). An empirical evaluation of alternative forecasting combinations. Management Science 33(10): 1267-1276, October.

SAMUELSON, P.A. (1965). Proof that properly anticipated prices fluctuate randomly. Industrial Management Review: 41-49, Spring.

SCHAFFER, C. (1994). Cross-validation, stacking and Bi-level stacking: meta-methods for classification learning. In: Selecting models from data: artificial intelligence and statistics IV,ed. by CHEESEMAN, P. and OLDFORD, R.W. (New York: Springer Verlag), 51-59.

SCHALLER, H. and VAN NORDEN, S. (1997). Fads or Bubbles ? Working Paper 97-2. Bank of Canada.

SCHAPIRE, R.E. (1990). The strength of weak learnability. Machine Learning 5: 197-227.

SCHMITT, J.R. (1954). An application of multiple correlation to population forecasting. Land Economics 30: 277-279.

SCHNAARS, S.P. (1986a). An evaluation of rules for selecting an extrapolation model on yearly sales forecasts. Interfaces 16: 100-107.

———— (1986b). A comparison of extrapolation models on yearly sales forecasts. International Journal of Forecasting 2: 71-85.

SELFRIDGE, O.G. (1959). Pandemonium: a paradigm for learning. In: THE SYMPOSIUM ON THE MECHANIZATION OF THOUGHT PROCESSES, Proceedings, Teddington, England. National Physical Laboratory, H.M. Stationery Office, London, 511-529.

SETHI, I.K. and OTTEN, M. (1990). Comparison between entropy net and decision tree classifiers. In: THE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS IJCNN-90, Proceedings, ARBOR A. MI. IEEE Neural Networks Council, 63-68.

SETIONO, B. and STRONG, N. (1998). Predicting stock returns using financial statement information, Journal of Business Finance and Accounting 25(5): 631-657, June/July.

SEWALL, M.A. (1981). Relative information contributions of consumer purchase intentions and management judgement as explanators of sales. Journal of Marketing Research 18: 249-253.

SHARPE, W.F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. Journal of Finance 19: 425-442.

SHARPE, W.F. and COOPER, G.M. (1972). Risk-return classes of New York Stock Exchange common stocks 1931-1967. Financial Analysts Journal 28 (2): 46-50.

SHAVLIK J., MOONEY R., and TOWELL G. (1991). Symbolic and neural learning algorithms: an experimental comparison. Machine Learning 6: 111-143.

SHECHACK, A. and MARTIN, J. (1987). The relative performance of the PSR and the PER investment strategies. Financial Analysts Journal: 46-56, March-April.

SHILLER, R.J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends ? American Economic Review 71: 421-436.

SHILLER, R.J. (1984). Stock prices and social dynamics. Brookings Papers on Economic Activity: 457-498.

SINGLETON, J.C. and SURKAN, A. (1991). Modelling the judgement of bond rating agencies: artificial intelligence applied to finance. Journal of the Midwest Finance Association 20: 72-80.

SINGLETON, J.C. and SURKAN, A. J. (1994). Bond ratings with neural networks. In: Neural Networks in the Capital Markets, ed. by REFENES A. P., John Wiley and Sons.

SKALAK, D.B. (1997). Prototype selection for composite nearest neighbor classifiers. Dissertation Thesis. University of Massachusetts Amherst. Dept of Computer Science.

SMIDT, S.A. (1968). New look at the random walk hypothesis. Journal of Financial and Quantitative Analysis: 235-262, September.

SMITH, A.M. and MAKOV, U.E. (1978). A quasi-Bayes sequential procedure for mixtures. Journal of Royal Statistical Society, Series B, 40: 106-112.

SMITH, D. (1989). Combination of forecasts in electricity demand prediction. Journal of Forecasting 8: 349-356.

SPECHT, D. (1990). Probabilistic neural networks. Neural Networks 3: 109-118.

SRINIVISAN, V. and KEIM, Y.W. (1987). Credit rating: a comparative analysis of classification procedures. The Journal of Finance 42: 665-681.

STAEL VON HOLSTEIN, C.-A.S. (1971). An experiment in probabilistic whether forecasting. Journal of Applied Meteorology 10: 635-645.

——— (1972). Probabilistic forecasting: an experiment related to the stock market. Organisational Behavior and Human Performance 8: 139-158.

STAMBAUGH, R.F. (1982). On the exclusion of assets from tests of the two-parameter model: a sensitivity analysis. Journal of Financial Economics 10: 237-268.

——— (1986). Discussion of Summer's paper. Journal of Finance 41: 601-602.

STANDARD and POOR'S (1991-97). Global ratings handbooks. 25 Broadway. New York, NY 10004(1) 212-2081810. Vol. 6 (8). August.

STATTMAN, D. (1980). Book values and expected stock returns. Unpublished MBA Honours Paper. University of Chicago.

STEPHANOPOULOS, G.N AND GUTERMAN, H. (1989). Pattern recognition in fermentation processes. Paper 163. ACS Meeting. Miami Beach. Florida.

STOBER, T.L. (1992). Summary financial statement measures and analysts' forecasts of earnings. Journal of Accounting and Economics 15: 347-372.

STOLL, H. and WHALEY, R. (1983). Transaction costs and the small firm effect. Journal of Financial Economics 12: 57-79.

SUMMERS, L.H. (1986). Does the stock market rationally reflect fundamental values? Journal of Finance 41: 591-602.

SUTTON, R.S. (1992). Introduction - the challenge of reinforcement learning. Machine Learning 8(3): 225-227.

TAYLOR, G.C. (1985). Combination of estimates of outstanding claims in non-life insurance. Insurance: Mathematics and Economics 4: 81-91.

THEIL, H. (1966). Applied Economic Forecasting. (New York: Rand McNally).

TIAO, G.C and GUTTMAN, I. (1980). Forecasting contemporal aggregates of multiple time-series. Journal of Econometrics 12: 219-230.

TIBILETTI, L. (1994). A non-linear combination of experts' forecasts: a Bayesian approach. Journal of Forecasting 13: 21-27.

TITMAN, S. and WARGA, A. (1989). Stock returns as predictors of interest rates and inflation. Journal of Financial and Quantitative Analysis 24 (1): 47-58.

TITTERINGTON D.M., MURRAY G.D., MURRAY L.S., SPIEGELHALTER D.J., SKENE A.M., HABBEMA J.D.F., and GELPKE G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. Journal of the Royal Statistical Society, Series A 144: 145-175.

TOLLEFSON, D. and JOY, O.M. (1978). Some clarifying comments on discriminant analysis. Journal of Financial and Quantitative Analysis: 197-200.

TRABELSI, A. and HILLMER, S.C. (1989). A benchmarking approach to forecast combination. Journal of Business and Economic Statistics 7(3) July: 353-362.

TRENKLER, G. and LISKI, E.P. (1986). Note - linear constraints and the efficiency of combined forecasts. Journal of Forecasting 5: 197-202.

TSAPTSINOS D., MIRZAI A., and JERVIS B. (1990). Comparison of machine learning paradigms in a classification task. In: THE 5TH INTERNATIONAL CONFERENCE, Proceedings, applications of artificial intelligence in engineering V, ed. by RZEVSKI, G. (Berlin: Springer-Verlag).

TYREE, E. and LONG, J.A. (1996). Assessing financial distress with probabilistic neural networks. In: THE 3RD INTERNATIONAL CONFERENCE ON NEURAL NETWORKS IN THE CAPITAL MARKETS, Proceedings, London Business School, London 1996.

UTGOFF, P.E. (1989). Perceptron trees: a case study in hybrid concept representations. CONNECTION SCIENCE 1: 377-391.

VIRTANEN, I. and YLI- OLLI, P. (1987). Forecasting stock market prices in a thin security market. OMEGA International Journal of Management Science 15: 145-155.

WALKER, R. (1992). An expert system architecture for heterogeneous domains. Ph.D. Dissertation, Vrije Universiteit te Amsterdam.

WALL, K.D. and CORREIA, C.A. (1989). Preference based method for forecast combination. Journal of Forecasting 8: 269-292.

WALZ, D.T. and WALZ, D.B. (1989). Combining forecasts: multiple regression versus a Bayesian approach. Decision Sciences 20: 77-89.

WARTON, R. (1999). Company profitability and finance. Office of National Statistics. Unpublished Manuscript, London.

WATERHOUSE, S. R. and ROBINSON, A. J. (1995). Non-linear prediction of acoustic vectors using hierarchical mixtures of experts In: Advances in neural information processing systems 7, ed. by TESAURO G., TOURETZKY D. S. and LEEN, T. K. MIT Press, 55 Hayward St., Cambridge, MA, 02142-1399, 835-842.

WATERHOUSE, S.R. (1997). Classification and regression using mixtures of experts. Ph.D. Thesis. Department of Engineering, University of Cambridge.

WEBB, W. and LOWE, D. (1990). The optimised internal representation multilayer classifier networks performs non-linear discriminant analysis. Neural Networks 3: 367-375.

WEI W.W. and ABRAHAM B. (1981). Forecasting contemporal time-series aggregates. Communications in Statistics A10: 1335-1344.

WEIGEND, A. S., MANGEAS, M. and SRIVASTAVA, A.N. (1995). Non-linear gated experts for time series - discovering regimes and avoiding overfitting. International Journal of Neural Systems 6(4): 373-399.

WEISS, S.M. and KULIKOWSKI, C.A. (1991). Computer systems than learn: classification and prediction methods from statistics, neural networks, machine learning and expert systems. (San Matteo CA: Morgan Kaufmann).

WELCH, B.L. (1939). Note on discriminant functions. Biometrika 31: 218-220.

WEST, R.R. (1973). Bond ratings, bond yields and financial regulation: some findings. Journal of Law and Economics 16: 159-168, April.

WEST, K.D. (1987). A specification test for speculative bubbles. Quarterly Journal of Economics 102(3): 553-580.

WEST, M. and HARRISON, J. (1989). Bayesian forecasting and dynamic Models. (New York: Springer-Verlag).

WHEATLEY, S. (1989). A critique of latent variable tests of asset pricing models. Journal of Financial Economics 23: 325-338.

WHITELAW, R.F. (1994). Time variations and covariations in the expectation and volatility of stock market returns. Journal of Finance 49: 515-541.

WILLIAMS, E. (1985). Learning internal representatives by error propagation. Institute for Cognitive Science Report 8506. San Diego, California.

WINKLER, R.L. (1971). Probabilistic precision: some experimental results. Journal of the American Statistical Association 66: 675-685.

——— (1981). Combining probability distributions from dependent information sources. Management Science 27: 479-488.

WINKLER R.L, MURPHY A.H., and KATZ R.W. (1977). The consensus of subjective probability forecasts: are two, three,...,heads better than one. In: 5TH CONFERENCE ON PROBABILITY AND STATISTICS, Preprint Volume. Las Vegas, Nevada (American Meteorological Society, Boston, MA) Nov. 15-18, 57-62.

WINKLER, R.L. and MAKRIDAKIS, S. (1983). The combination of forecasts. Journal of the Royal Statistical Society, Series A, 146: 150-157.

WINKLER, R.L. in ARMSTRONG ET AL. (1983). Commentary on the Makridakis time - series competition (M-Competition). Journal of Forecasting 2: 259-311.

WIPER, M. (1990). Calibration and use of expert probability judgements. PhD Thesis, Leeds University.

WIPER, M.P. and FRENCH, S. (1995). Combining experts' opinions using a normal-wishart model. Journal of Forecasting 14: 25-34.

WISE, B.M. and RICKER, N.L. (1989). Upset and sensor failure detection in multivariate processes. Paper 164b. American Institute of Chemical Engineers Meeting, San Francisco.

WOLF, M. (1997). Stock returns and dividend yields revisited: a new way to look at an old problem. Working Paper, UCLA, Los Angeles, CA.

WOLPERT, D. (1992). Stacked Generalization. Neural Networks 5: 241-259.

———— (1993). Combining generalizers using partitions of the learning set. In: Lectures in complex systems, ed by NADEL, L and STEIN, D. (Reading, MA: Addison-Wesley).

XU L., KRYZAK A., and SUEN C.Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics 22: 418-435.

YANG, Z. (1999). Probabilistic neural networks in bankruptcy prediction. Journal of Business Research 44: 67-74.

ZARNOWITZ, V. (1967). An appraisal of short-term economic forecasts. National Bureau of Economic Research, New York.

———— (1984). The accuracy of individual and group forecasts from business outlook Surveys. Journal of Forecasting 3: 11-26.

ZAROWIN, P. (1989). Does the stock market overreact to corporate earnings information ? Journal of Finance 44: 1385-1399.

ZEEVI A. J., MEIR R., and MAIOROV V. (1997). Error bounds for functional approximation and estimation using mixtures of experts. Technical Report CC-132. Faculty of Electrical Engineering, Technion, Haifa 32000, Israel.

ZHANG X., MERSIROV J.P., and D.L.WALTZ (1992). A hybrid system for protein secondary structure prediction. Journal of Molecular Biology 225:1049-1063.

ZHANG G., PATUWO B. E., and HU M. Y. (1998). Forecasting with artificial neural networks: the state of the art. International Journal of Forecasting 14: 35-62.