Feature Selection for Computer-Aided Polyp Detection using MRMR

Xiaoyun Yang, Boray Tek, Gareth Beddoe and Greg Slabaugh

[a]Medicsight PLC, Kensington Centre, 66 Hammersmith Road, London, UK

ABSTRACT

In building robust classifiers for computer-aided detection (CAD) of lesions, selection of relevant features is of fundamental importance. Typically one is interested in determining which, of a large number of potentially redundant or noisy features, are most discriminative for classification. Searching all possible subsets of features is impractical computationally. This paper proposes a feature selection scheme combining AdaBoost with the Minimum Redundancy Maximum Relevance (MRMR) to focus on the most discriminative features. A fitness function is designed to determine the optimal number of features in a forward wrapper search. Bagging is applied to reduce the variance of the classifier and make a reliable selection. Experiments demonstrate that by selecting just 11 percent of the total features, the classifier can achieve *better* prediction on independent test data compared to the 70 percent of the total features selected by AdaBoost.

Keywords: CAD, Adaboost, Minimum Redundancy Maximum Relevance (MRMR), bagging

1. INTRODUCTION

Colorectal cancer ranks as the second leading cause of cancer related death in western countries.[1] Most colorectal cancers begin from benign polyps in the colon that develop into cancer over time. Early detection and treatment can significantly improve the prognosis of colorectal cancer patients. In recent years there has been much interest in CT Colonography (CTC), which scans the whole abdomen after cleansing and air insufflation of the colon and provides an interior view of the colon.[2] Polyps can then be detected by a clinical reader who examines the CT data using advanced visualization software that provides both 3D endoluminal views of the colon as well as 2D multi-planar reformatted slices, as shown in Figure 1.

Although CTC has been demonstrated to be an effective colorectal screening approach,[3] it is possible for the clinical reader to fail to detect lesions, due to the large quantity of data generated (typically 800 - 2000 images per patient). In recent years a number of computer-aided detection (CAD) prototypes[4–6] have been developed and demonstrated potential to identify suspicious colonic lesions (polyps and masses) with clinically acceptable detection rates. Typically these schemes start from a colon segmentation and then generate candidate regions around the colon wall based on geometric measures. From the candidate regions, various extracted features are then sent to a classifier to remove false regions to produce the final detection results presented to the user. The number of features is usually quite large. Some features are highly discriminative, while others are less relevant, useless, or even harmful. Therefore, selecting a set of good features is crucial to improve the classifier performance by reducing computational complexity and enhancing generalization to new data.

Recently, genetic algorithms (GAs) have been proposed for feature selection in CAD.[7–9] By representing a subset of features as a chromosome, a GA converts the problem of searching subsets in the feature space to an optimization problem to find the best gene. Each gene is evaluated by a fitness function determined by the classifier performance.[10] These studies show that the models trained from the resultant subset of features can outperform the models trained from the full feature set.

In this paper, we propose a feature selection scheme employing Minimum Redundancy Maximum Relevance (MRMR),[11] an approach based on information theory. The features that have strong relevance on the target class and are minimally redundant are recursively selected to generate a ranking of the features. The data is projected into the selected subspaces and fed into an Adaboost classifier. The performance of the classifier on

Further author information: (Send correspondence to Xiaoyun Yang)
Xiaoyun Yang: E-mail: xiaoyun.yang@medicsight.com, Telephone: +44 (0) 207 605 7976

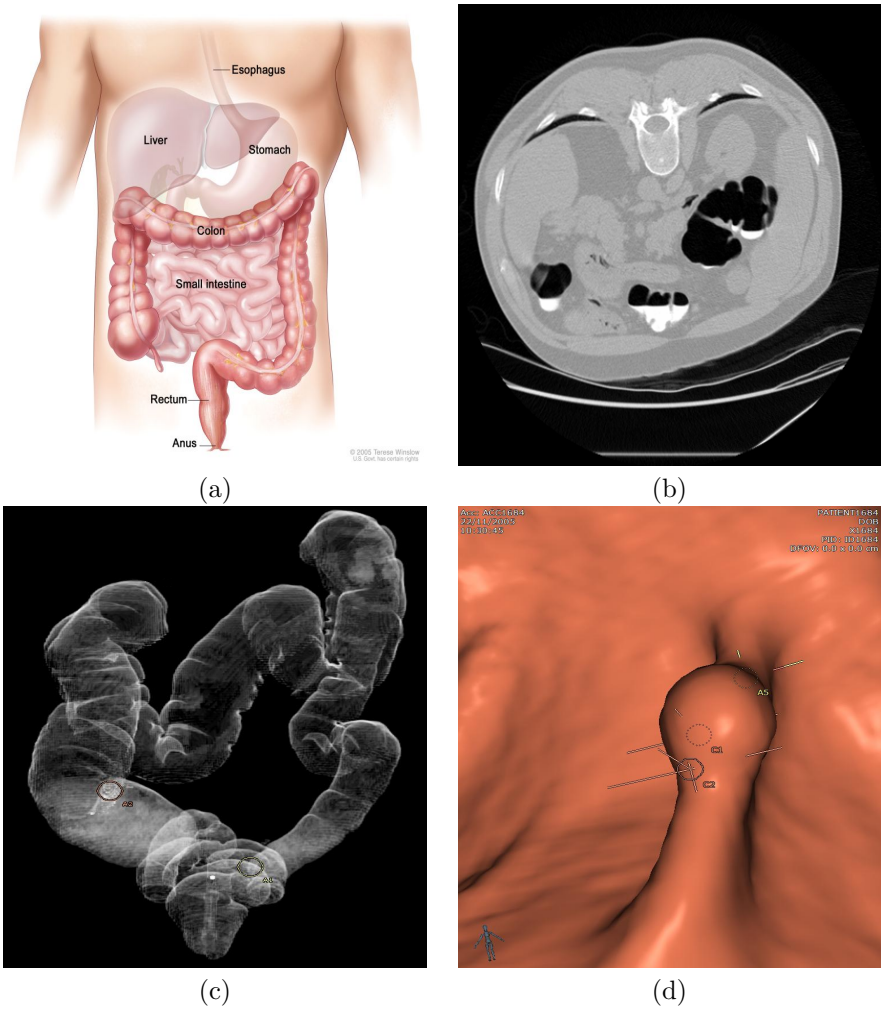

Figure 1. (a) Illustration of the small bowel (intestine) and colon. Notice that the small bowel is framed by the colon in this coronal view. Public domain image courtesy of the U.S. government. (b) Illustration of axial section image of CTC. (c) Colon segmentation, with overlaid marks displayed in CAD. (d) Visual example of detected polyp.

validation data and the number of selected features are used to estimate a fitness value as a metric to determine the optimal number of features in a wrapped forward search. A bagging approach is applied to address the variance of data and make a reliable feature selection. Results demonstrate better performance using merely 11 percent of the total features when evaluated on independent test data.

The content of this paper is organized as follows. Section 2 describes the AdaBoost classifier and the MRMR feature selection method. Section 3 describes our scheme for feature selection and the classifier. Section 4 presents our experimental results. We then conclude the paper in Section 5.

2. BACKGROUND

2.1 AdaBoost

Adaptive boosting (AdaBoost), introduced in 1995 by Freund and Schapire,[12] is a serial ensemble approach that builds an additive model. It begins by training a weak learner on a data set to generate a hypothesis h_1. The distribution of weights of the training samples is updated by a function of the classification error. This ensures that misclassified samples have larger weights so that the classifier can focus to separate the difficult examples. A next hypothesis h_2 is generated by training a weak classifier on the same set of samples again but with the

updated weight distribution. This process continues iteratively until a target error bound or maximum number of rounds has been reached. The final hypothesis H is formed by linearly combining the set of hypotheses (h_1,h_2, ... h_t) generated at each round with their weighted votes.

Conventional AdaBoost combines two tasks: selecting features and building a strong classifier. However, a weak learner only considers a single feature by choosing an optimal threshold to minimize an error function. This greedy search strategy renders the classifier susceptible of being trapped by local minima, especially when the data is noisy.

2.2 MRMR

The MRMR approach introduced by Peng,[11] who extends work from Battiti[13] and Kwak and Choi,[14] solves the feature selection problem using information theory. In information theory, mutual information measures the statistical dependency between two random variables. The problem of feature selection can be interpreted as a search for a subset of features that jointly have the largest dependency on the target class (polyps, in our case). However, the difficulty of estimating multivariate densities accurately makes the computation difficult, especially for continuous features. Instead of seeking maximum dependency, the method seeks a subset of features that are *maximally relevant* on the target class but for which the features within the subset are *minimally redundant*.

According to information theory, the mutual information between two random variables x and y can be defined as

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{1}$$

where $p(x)$, $p(y)$ represents the probability density function of x and y respectively, and $p(x,y)$ is the joint probability density function. To find a best single feature among a set of features (x_1, ..., x_n) to separate class c, one can search for the largest mutual information $I(x_i;c)$ that represents largest dependency between feature variable x_i and target class c. The best m relevant features can be found using

$$max\ D(S,c),\ D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i;c). \tag{2}$$

where S represents the full feature set. This equation gives the best subset of features that are most relevant to the target class. However, it doesn't describe the redundancy among features. Although two features may have strong separability on the target class, it would be undesirable to include them if they are correlated. To determine redundancy between features, the following equation is used:

$$min\ R(S,c),\ R = \frac{1}{|S|} \sum_{x_i, x_j \in S} I(x_i;x_j). \tag{3}$$

By combining Equations 2 and 3, we can find a subset of features that are maximally relevant with minimum redundancy:

$$max\ \Phi(D,R),\ \Phi = D - R. \tag{4}$$

In practice, Equation 4 can be implemented incrementally as follows:

$$\max_{x_i \in X - S_{m-1}} [I(x_j;c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j;x_i)]. \tag{5}$$

in this equation, the method starts by seeking a single feature x_k that is maximally relevant to target class, and includes x_k in the set S_1. From the remaining features $X - S_1$, it then seeks next feature that is maximally relevant to target class but minimally redundant with features that have been selected in S_1. The selected feature with feature set S_1 forms feature set S_2. This process iterates until the feature number has reached m features, $S_1 \subset S_2 \subset \ldots \subset S_{m-1} \subset S_m$.

3. METHODS

The method aims for selecting a subset of features most discriminative for classification from a feature set with many potentially redundant or noisy features. To avoid exhaustive search, the Minimum Redundancy Maximum Relevance (MRMR) with a given feature number is applied on the data, and the performance of Adaboost on the projected feature data is used as a criteria in a designed fitness function to determine the optimal number of features. Once the optimal number is fixed, the subset of features is determined.

3.1 Fitness function

To determine the best number of features, a searching criterion has to be defined. Classification performance and the number of selected features are used to design a fitness function. Traditional metrics that consider classification errors as equally important may fail to accurately evaluate the classifier due to the imbalanced datasets of CAD. By reflecting the trade-off between the positive class and negative class accuracies the Receiver Operating Characteristics curve better suits the highly imbalanced problem. A ROC curve is a plot of the sensitivity on the y-axis against false positive rate on the x-axis as a decision function threshold is varied. The area under the curve (AUC) can be used as a performance metric to cope with cost-sensitive classification in imbalanced data.[15] Accordingly, we use the area under curve (AUC) to evaluate the performance of classifier on the given feature subset of validation data. The AUC is estimated by Wilcoxon-Mann-Whitney statistics.[16] Our fitness function is defined as follows:

$$fitness = W_{auc} \times AUC + \frac{W_f}{N_f} \tag{6}$$

where W_{auc} is the weight for AdaBoost classifier performance, W_f is the weight for the feature number, N_f is the feature number.

3.2 Feature selection scheme

Subset feature selection can be generated in a search strategy with certain criteria. We perform a sequential forward search in an incremental manner (as shown in Figure 2). MRMR was applied to the training data to obtain a series of feature sets $S_1 \subset S_2 \subset \ldots \subset S_{m-1} \subset S_m$. AdaBoost was trained on the training data with selected subset of features. The generated model is then applied to the validation data on the selected feature subset. The decision function value given by the model is then used as the confidence of prediction. The threshold can be adjusted to generate a ROC curve. With this decision function value and target class, we computed the AUC by Wilcoxon-Mann-Whitney statistics to determine the performance for the given feature subset.

Searching from the minimum number of features to the maximum, we can obtain AUC values for each given feature number. However, this selection is vulnerable to the variance of the data distribution resulting from data noise and random downsampling for training. To reduce the variance of estimation and make the selection more reliable, it is important to introduce a bagging approach. We repeatedly divided the data M (an odd number) times with random initializations and performed the same experiment as shown in Figure 2. The average of AUC for each number of features is then used to estimate fitness value in Equation 6. The one with best fitness value is chosen as the optimal feature number. The models generated on these different training data with the selected optimal feature subset will be combined together to predict future data by simple voting. The majority votes of class category will be the final prediction.

While Adaboost has a built-in feature selection mechanism, the scheme we propose makes use of MRMR's ability to reduce irrelevancy and redundancy makes the classifier focus on the most discriminative feature set. Bagging further produces a more reliable and stable feature selection. The selected feature subset is much smaller than the one selected by AdaBoost; this reduces the complexity of the generated models and achieves better generalization on the independent test data. The reduced set of features avoids computation of unnecessary features and thus improves the efficiency.

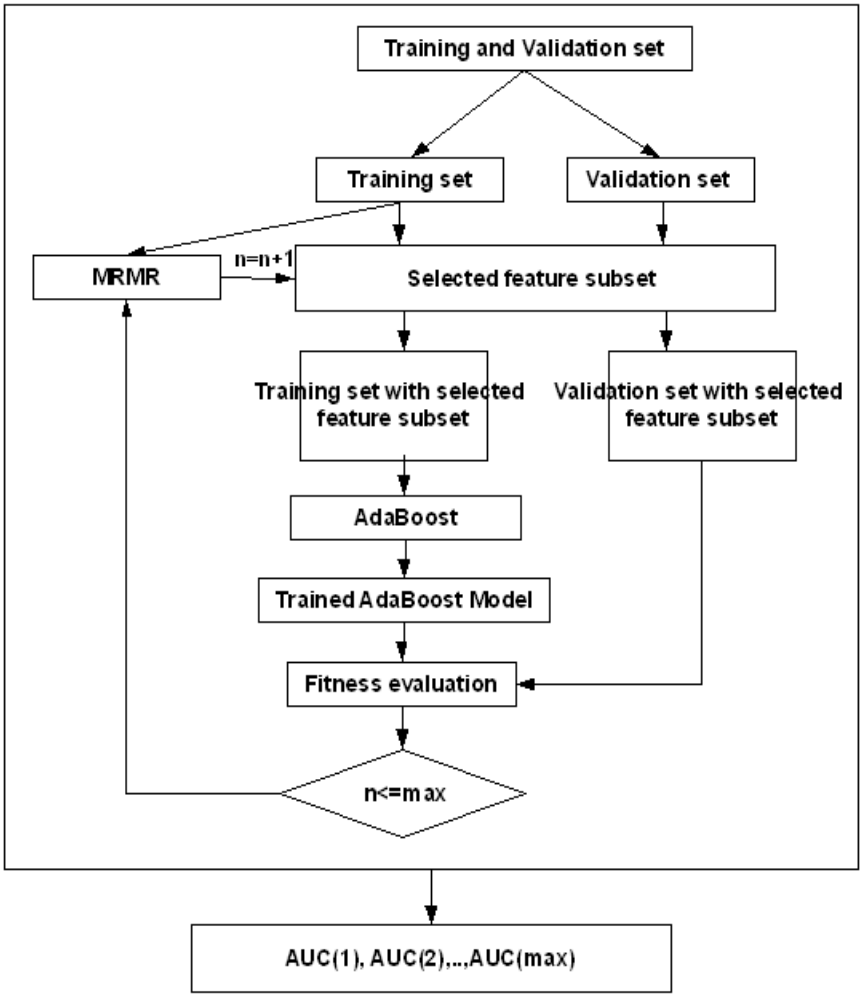

Figure 2. The flowchart of MRMR employed bagged AdaBoost algorithm.

4. EXPERIMENT

4.1 Data preparation

The CT colonography data in our experiments were obtained from 68 3D CT volumes containing prone and supine scans collected from seven hospitals. On each slice, the average pixel dimensions were 0.73 mm by 0.73 mm and the slice thickness ranged between 0.6 mm to 1.25 mm. A large number of cases were acquired using a fecal and/or fluid tagging protocol. Three medical experts annotated the scans and 106 polyps were identified. The polyps ranged in size from 5 to 20 mm and were marked by 111 polyp regions including few multiple detections. There were 171 features available for each sample; including the size of the structure, statistics (such as mean, standard deviation, entropy) of intensity, shape index map,[4] curvedness map, gradient concentration (GC) [4] , directional gradient concentration (DGC), Sato's shape features,[17] Law's energy texture.[18]

The data was randomly divided into two parts with 48 (70%) scans for training and validation and 20 (30%) scans for independent testing, ensuring the training and test scans were of different patients. In the training data and testing data, we randomly divided half of polyp regions and half the non-polyp regions into a training set and the other half are into a validation set as hold-out to estimate the performance for that selected feature subset. The non-polyp candidates in training data were downsampled so that the ratio between polyp regions and non-polyp regions was 1 : 5.

Method	Features selected	Sensitivity	Specificity	FPs	AUC
AdaBoost	120	0.7436	0.9049	6.85	0.82425
AdaMRMR	19	0.8718	0.8549	10.45	0.8633

Table 1. Performance of AdaBoost, AdaMRMR on independent set. Features selected from a total set of 171 possible features.

4.2 Results

We performed two experiments using AdaBoost and AdaBoost with MRMR (AdaMRMR). To reduce the variance of estimation as mentioned above, we generated training data and validation data 9 times randomly. For each of the 9 datasets of training and validation, MRMR was performed on training data, AdaBoost was trained on the training data with the selected subset and tested on the validation data. In the fitness function Equation 6, the AUC is set to be average of AUC values obtained for that number of feature subsets of different generated data using Wilcoxon-Mann-Whitney statistics. We set $W_{auc} = 0.9$ and $W_f = 0.1$. Figure 3 shows the fitness value for each number of feature set. The best number of features is the one with the maximum fitness value, as shown at the peak of curve.

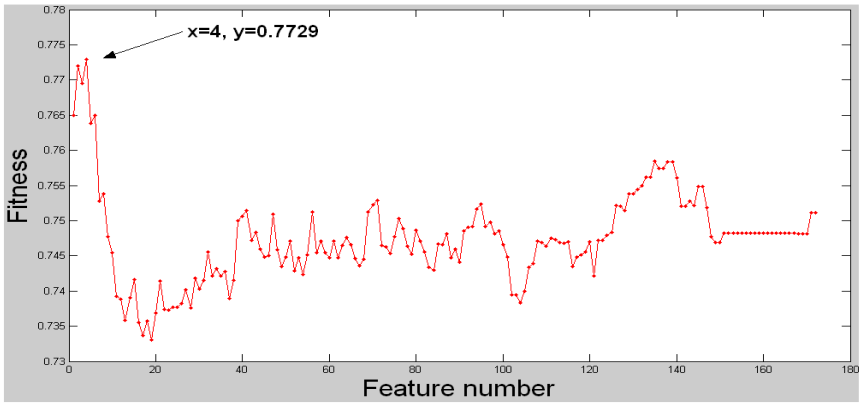

Figure 3. Fitness for each number of feature subset, performed by AdaMRMR.

Figure 3 has shown that the AdaMRMR has the maximum fitness value at feature number 4. AdaMRMR may find different features each time; in total, 19 features were used by 9 generated models as shown in Table 1. Figure 4 gives an example of scatter plot of four features selected by AdaMRMR and it shows a high degree of discrimination visually. We compared the result of this to without feature selection, where AdaBoost was applied directly on the whole feature set. AdaMRMR selected a much smaller feature set by merely 11 percent of the 171 total feature set comparing to 70 percent chosen by AdaBoost. We combine the generated 9 models for Adaboost and AdaMRMR respectively. Given the independent data set, each individual will make a prediction and the majority will be used as the final decision. As shown in the Table 1, the AUC of AdaMRMR on the independent data is 0.8633, outperforming AdaBoost's 0.82245.

The results show that AdaMRMR makes the classifier focus on the most discriminative feature set. Bagging them together makes the feature selection more reliable and stable. The selected feature subset is much smaller than the one selected by AdaBoost; this reduces the complexity of the generated models and achieves better generalization on the independent test data. Using significantly reduced feature set, Adaboost avoids computation of unnecessary features and thus improves the efficiency.

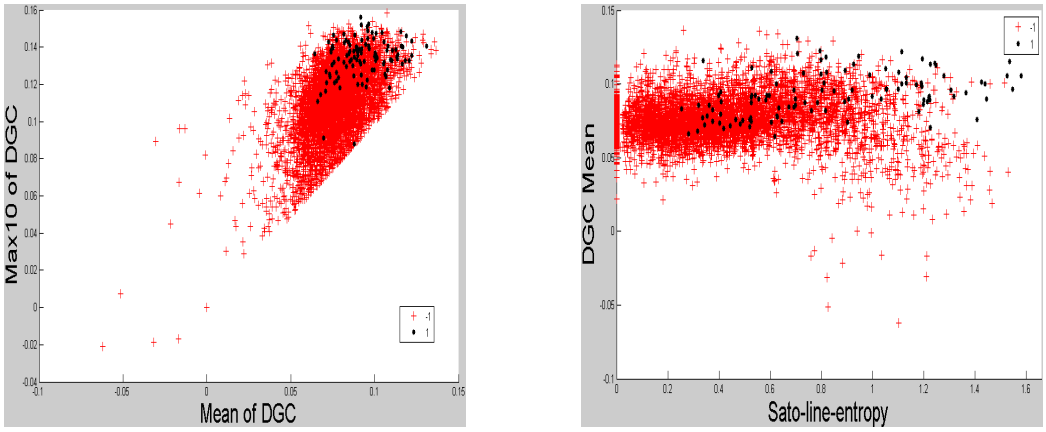

Figure 4. Scatter plot for features: mean of DGC,[4] mean of 10 maximum values of DGC and entropy of line map.[17]

5. DISCUSSION AND CONCLUSION

We have presented a novel feature selection scheme for the task of detecting colon polyps in CT colonography. The proposed method effectively combines mutual information feature selection with AdaBoost. In addition, a bagging approach was applied to reduce the variance and make the selection more stable and reliable. With the help of the measures of mutual information, AdaBoost can focus on important, non-redundant features for classification. A sequential forward search was conducted to determine the optimal feature subset by the proposed criteria. The result demonstrates the promise of using a much smaller feature set than the whole feature sets which gives better generalization performance on independent data.

REFERENCES

[1] "Cancer facts and figures," *American Cancer Society Annual Report* **12**(1), 11–12 (2007).

[2] Johnson, C. D. and Dachman, A. H., "Ct colonography: The next colon screening examination?," *Radiology* **216**(2), 311–319 (2000).

[3] Johnson, C. D., Chen, M., Toledano, A., Heiken, J., Dachman, A. H., Kuo, M. D., Menias, C., Stewert, B., Cheema, J. I., Obregon, R. G., Fidler, J. L., Zimmerman, P., Horton, K. M., Coakley, K., Iyer, R. B., Hara, A. K., Halvorsen, R. A., Casola, G., Yee, J., Herman, B. A., Burgart, L. J., and Limburg, P. J., "Accuracy of ct colonography for detection of large adenomas and cancers," *New England Journal of Medicine* **259**(12), 1207 – 1217 (2008).

[4] Yoshida, H. and Nappi, J., "Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps," *IEEE Transactions on Medical Imaging* **20**, 1261–1274 (December 2001).

[5] Summers, R. M., Johnson, C. D., Pusanik, L. M., Malley, J. D., Youssef, A. M., and Reed, J. E., "Automated polyp detection at CT colonography: Feasibility assessment in a human population," *Radiology* **219**, 51–59 (2001).

[6] Slabaugh, G., Yang, X., Ye, X., Boyes, R., and Beddoe, G., "A robust and fast system for ctc computer-aided detection of colorectal lesions," *Algorithms* **3**(1), 21–43 (2010).

[7] Miller, M. T., Jerebko, A. K., Malley, J. D., and Summers, R. M., "Feature selection for computer-aided polyp detection using genetic algorithms," *IEEE Transactions on Information Technology in Biomedicine* **10**, 504 – 511 (July 2006).

[8] Boroczky, L., Zhao, L., and Lee, K. P., "Feature subset selection for improving the performance of false positive reduction in lung nodule CAD," *IEEE Transactions on Information Technology in Biomedicine* **10**, 504 – 511 (July 2006).

[9] Zheng, Y., Yang, X., Siddique, M., and Beddoe, G., "Simultaneous feature selection and classification based on genetic algorithms: an application to colonic polyp detection," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference* **6915** (April 2008).

[10] Fröhlich, H., Chapelle, O., and Schölkopf, B., "Feature selection for support vector machines by means of genetic algorithms," in [*Proceedings of the 15th IEEE International Conference on Tools with Artificial Inteligence (ICTAI)*], 142–148 (3-5 Nov 2003).

[11] Peng, H., Long, F., and Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 1226–1238 (2005).

[12] Freund, Y. and Schapire, R. E., "A decision-theoretic generalization of on-line learning and an application to boosting," in [*EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory*], 23–37, Springer-Verlag (1995).

[13] Battiti, R., "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks* **5**, 537–550 (1994).

[14] Kwak, N. and Choi, C.-H., "Input feature selection for classification problems," **13** (2002).

[15] Hong, X., Chen, S., and Harris, C. J., "A kernel-based two-class classifier for imbalanced data sets," *Neural Networks, IEEE Transactions on* **18**, 28–41 (Jan 2007).

[16] Hanley, J. and McNeil, B., "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**(1), 29–36 (1982).

[17] Sato, Y., Westin, C.-F., Bhalerao, A., Nakajima, S., Shiraga, N., Tamura, S., and Kikinis, R., "Tissue classification based on 3d local intensity structures for volume rendering," *IEEE Transactions on Visualization and Computer Graphics* **6**(2), 160–180 (2000).

[18] Laws, K., *Textured Image Segmentation*, PhD thesis, University of Southern California (January 1980).