Automated Information Extraction from Free-Text EEG Reports

Siddharth Biswal[1], Zarina Nip[1], Valdery Moura Junior[1],
Matt T. Bianchi[1], Eric S Rosenthal[1] M Brandon Westover[1], MD PhD

***Abstract*— In this study we have developed a supervised learning to automatically detect with high accuracy EEG reports that describe seizures and epileptiform discharges. We manually labeled 3,277 documents as describing one or more seizures vs no seizures, and as describing epileptiform discharges vs no epileptiform discharges. We then used Naïve Bayes to develop a system able to automatically classify EEG reports into these categories. Our system consisted of normalization techniques, extraction of key sentences, and automated feature selection using cross validation. As candidate features we used key words and special word patterns called elastic word sequences (EWS). Final feature selection was accomplished via sequential backward selection. We used cross validation to predict out of sample performance. Our automated feature selection procedure resulted in a classifier with 38 features for seizure detection, and 23 features for epileptiform discharge detection. The average [95% CI] area under the receiver operating curve was 99.05 [98.79, 99.32]% for detecting reports with seizures, and 96.15 [92.31, 100.00]% for detecting reports with epileptiform discharges. The methodology described herein greatly reduces the manual labor involved in identifying large cohorts of patients for retrospective neurophysiological studies of patients with epilepsy.**

I. Introduction

Over recent decades the medical field has generated large archives of free text reports, which contain a vast store of potential knowledge. In the past 20 years our institution (Massachusetts General Hospital, MGH) alone has generated over 100,000 free text reports describing the features of electroencephalogram (EEG) recordings from patients evaluated for epilepsy and other neurological conditions. These reports contain a wealth of untapped neurophysiological information in patients of all ages and across numerous neurological conditions, including epilepsy, delirium, neurodevelopmental disorders, stroke, migraine, and others.

Important categories of EEG findings that are frequently the subject of neurophysiological research include the presence or absence of seizures, epileptiform discharges (e.g. "spikes" and "sharp waves"), generalized periodic discharges, lateralized periodic discharges, and rhythmic delta activity [1]. To derive useful knowledge from these free-text reports, researchers typically have to perform weeks to months of intensive manual work to review and categorize these reports[7][8]. Automated algorithms offer the advantages of saving human labor, increased speed and the ability to scale the process to larger datasets [11]. A key challenge in creating automated classifiers for free text EEG reports lies in the wide syntactic variation inherent in natural language. Specifically, there are usually many ways to describe the same EEG pattern, whether normal or abnormal. Nevertheless, the field of text mining provides tools for coping with the richness and variety of natural language. We therefore sought to adapt text mining tools for categorizing free text EEG reports.

II. Materials and Methods

Figure 1 shows an overview of our process for creating the automated EEG report classification system. The process involves (1) creation of labeled to support feature selection, training, and classifier performance evaluation; (2) steps to normalize documents by reducing irrelevant complexity and heterogeneity; (3) extraction of informative features from normalized documents; (4) computing a classifier model for each document class from labeled data, and (5) testing the final model on held-out data using cross validation to evaluate classifier performance. We describe each of these steps in turn.

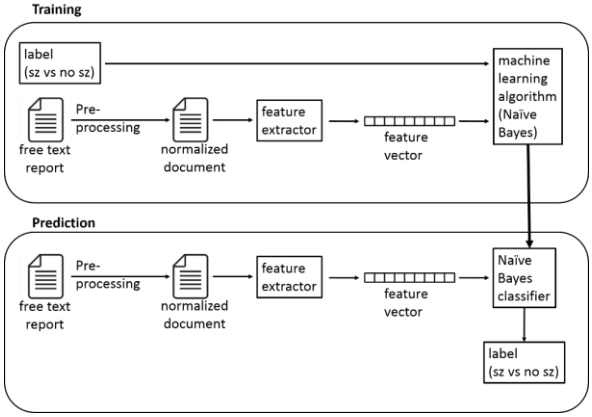

Fig. 1. Schematic of testing and training procedures for developing an automated report classification system

A. *Creation of training data*

We extracted 42,972 EEG reports from the MGH Neurophysiology database spanning 10 years (2001-2010), a diverse set of neurological diseases, reporting physicians, and patient ages. The research in the paper was performed with approval of the Institutional Review Board. In order to create a training dataset, we manually assigned class labels to 3,277 reports according to whether the text described seizures vs no seizures. We also labeled the same reports as describing epileptiform discharges vs no epileptiform discharges. This step of creating a gold-standard dataset was performed by one of the authors who is an experienced clinical electroencephalographer (MBW). From the 3,277 reports, 284 were identified as describing seizures, and 874 as describing epileptiform discharges.

Funding: NIH-NINDS K23 NS090900(MBW), Rappaport Foundation(MBW), Andrew David Heitman Neuroendovascular Research Fund (MBW,ESR).

Corresponding author: MBW. mwestover@mgh.harvard.edu. [1] Massachusetts General Hospital, Boston, MA 02114

B. Cross validation

To estimate of out-of-sample performance, we performed 500 rounds of repeated random sub-sampling cross validation. For each round of cross validation, the training set consisted of a random selection of 50% (142 of 284) of the labeled cases with seizures and approximately 50% (1496 of 2993) of the cases labeled as being without seizures. The remaining cases were held out as testing data. Similarly, training a classifier to detect cases with epileptiform discharges, in each round we selected 50% (437 of 874) of the cases with and 50% (1202/2403) of the cases without epileptiform discharges. All of the preprocessing, feature selection, and classifier training steps described below were performed solely using the training data to obtain a working classifier. After training we used the trained classifier on the testing data, and tabulated performance statistics as described below.

C. Preprocessing

We applied the following preprocessing steps to normalize all documents before classifier training. In the first phase of preprocessing we applied the following steps.

1. Removal of new line characters, punctuation marks and numbers

2. Tokenization, i.e. splitting the text into individual words

3. Spelling correction which was done using an edit distance based method with a custom dictionary [2]

4. Stemming, by application of the Porter stemmer algorithm [3]

Following document preprocessing, we compiled a 'dictionary' of all words from the entire corpus of 42,972 normalized EEG reports. The words that occurred in less than 1% of the EEG reports were removed from the dictionary. This removal of rare words further reduced the size of the dictionary and effectively removed the 'noise' from the feature set which was subsequently used to train the classifier.

D. Feature Extraction

Extraction of impression section: To minimize irrelevant information / noise, we developed a method to automatically extract "key sentences sequences" for the purposes of classification, as follows. In this method we first extract the "impression" section of the EEG (see Supplemental Figure S1). The impression section is standard in all EEG reports, and always contains a brief, few-sentence summary of any pathological findings, including the presence of seizures or epileptiform discharges.

E. Synsets and Key sentences

Next, from the impression we searched any instance of synonyms of either "seizure" or "epileptiform discharge". For this purpose we developed synsets (exhaustive lists of synonymous words or phrases) for each of these classes. If none of the keywords on this list was detected, then the report was automatically classified into either the "no seizures" or "no epileptiform discharges" category.

Next, we parsed the extracted impression section of each report into sentences, and serially searched each sentence for words or phrases from the relevant synset, stopping at the first sentence that contained a 'hit' (see Supplemental Figure S2). If none of the sentences contained a hit, then this sentence was automatically classified as part of the negative class (i.e., as a report with either "no seizures" or "no epileptiform discharges"). All further operations were performed solely on these key sentences. The remainder of each document was non-contributory.

F. Elastic word sequences

The EEG reports describing seizures vs no seizure are usually not distinguishable from single word frequencies. For example, "seizure" occurs in majority of EEG reports, regardless of seizure occurred or not. We therefore attempted to develop more discriminative features by generating word sequences [10]. We defined an "elastic word sequence" (EWS) as a sequence of any two key words separated by a gap of no more than 7 intervening words. For example, the EWS made from word1 (w1) and word2 (w2), denoted w1…w2, consists of w1 followed by up to 7 other words (with no restriction on what they might be), followed by w2. For example, if w1 = "no" and w2="seen", then the following phrases are both instances of w1…w2: "no electrographic seizures were seen" and "no definite seizure activity was seen".

G. Feature relevance scoring:

We combined the dictionary of single words with the list of EWS to form a final set of candidate features. To select the most promising discriminative features from this large set, we calculated for each feature its Matthew's correlation coefficient (MCC) [4]

$$MCC = \frac{TP\ X\ TN - FP\ X\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC measures the correlation between the binary class labels and the best prediction of the class label based on the feature in question. It returns +1,0,-1 corresponding to if features are able to perfectly discriminate between two classes, no better than guessing at random and perfect anti-correlation with actual class labels. Since MCC is generally useful when there is a large imbalance between classes, as in the case of out classification problem. We elected to retain 300 features (single words or EWS's) with the highest MCC values as candidates for inclusion in the final classifier. This number was chosen empirically as being more than enough features to create accurately performing classifiers. The number is subsequently reduced by a pruning procedure (see below).

As stated above, in each stage of cross validation, feature selection based on MCC values was done entirely based on the training data, without making use of the held-out testing data. In this way we sought to avoid overfitting.

H. Naïve Bayes:

We trained Naïve Bayes classifiers with "bag of words" feature vectors (modified to include EWS) to distinguish between reports with and without seizures, and between reports with and without spikes [12]. Briefly, a Naïve Bayes document classifier is a probabilistic model which assumes independence among features (words). Let C_1 and C_2 be two classes of documents (e.g. "seizures" and "no seizures") and n features, f_1,f_2,…,f_n, (the EWSs and single words in

our final feature dictionary). Let P(C|D) be the probability that a document D belongs to class C.

Within the Naïve Bayes framework we calculate the class assignment for a document D by the following approximation to the maximum a posteriori (MAP) estimate,

$$C^* = \max_{C\in\{C_1,C_2\}} P(C)\prod_{i=1}^{n} P(f_i|C)$$

The prior probability $P(C_j)$ is estimated as

$$P(C_j) = \frac{Documents(C = C_j)}{Number\ of\ Documents}$$

The conditional probability for each feature belonging to a class is given by the following formula:

$$P(f_i|C_j) = \frac{Count(f_i, C_j) + 1}{\sum Count(f_i, C_j) + 1}$$

To avoid numerical rounding errors, in our implementation we converted the products of probabilities in these expressions to sums of log-probabilities. The "bag of words" terminology when Naïve Bayes classifiers are used in this way for text classification refers to the fact that word order is not taken into account in calculating the conditional probability.

I. Classification

Classification of a document D is accomplished by calculating the value of the Naïve Bayes classifier for each of the two classes in question, taking the difference of their logarithms,

$$L = L(D) = \log P(C_1|D) - \log P(C_2|D)$$

and comparing the difference to a threshold, θ. Documents for which $L(D) > \theta$ are classified as belonging to class C_1, while documents for which $L(D) \leq \theta$ are classified as belonging to class C_2. The threshold θ can be varied to trade off sensitivity against specificity or precision against recall (see below), but is typically set to equal to zero, as we do in this work.

J. Classifier performance assessment

We evaluated classifier performance by calculating sensitivity, specificity, receiver operating characteristic curves (ROC), and precision-recall curves [5]. These are defined as follows. Sensitivity (Se; also known as recall, Re), false positive rate (Fp), and precision (Pr) statistics depend on the choice of the threshold parameter θ used to define the classifier, and are defined as:

In these formulae, TP is the number of cases in the positive class for which L>θ; FP is the number of positive cases with L≤θ; TN is the number of negative cases for which L≤θ; and FN is the number of negative cases for which L>θ.

We generated ROC curves by plotting sensitivity vs false positive rate while varying the threshold value θ between the minimum and maximum value of L(θ) over all cases. We generated precision recall curves by varying the threshold in a similar fashion while plotting precision vs recall.

K. Pruning

We applied a sequential backward feature elimination ("pruning") method with cross validation to minimize model overfitting [6]. Pruning in each stage of cross validation consisted of the following procedure. First, using only the training data, we trained a Naïve Bayes classifier using the 300 most promising features (highest MCC values), and calculated the area under the ROC curve (AUC). Then, in the first round of pruning, we sequentially removed each of the 300 features and re-trained a Naïve Bayes classifier with the remaining 299 features. We then removed from the feature set the one whose elimination yielded the largest AUC value. We repeated this procedure until all but 10 features were eliminated. Each round of cross validation yields a sequence of classifiers, indexed by the number of training features, denoted c(i,300),c(i,299),…,c(i,1), where i=1,2,…,n are the indices of the n rounds of cross validation.

To estimate the out-of-training-sample performance, we evaluated each of these models on each round of cross validation using the held-out testing data, calculated the resulting testing AUC, and then averaged over the n folds of cross validation. That is, letting *AUC(i,j)* be the test-data AUC for the i'th round of cross validation in the model with j features, the estimated testing performance of the model is AUC(j). We define the optimal number of features m* as the number of features which maximizes the average cross validation AUC.

$$\widehat{AUC}(j) = \sum_{i=1}^{n} AUC(i,j); \qquad m^* = \operatorname*{argmax}_j \widehat{AUC}(j)$$

IV. RESULTS

A. Inadequacy of simple keyword searches

Before constructing an automated classifier, we first tested the performance of simple keyword searches. Using the keyword "seizure" as discriminant feature between reports that do and do not describe seizures produces excellent recall, very poor precision (1%), indicating that the vast majority of non-seizure EEG reports nevertheless contain the word "seizure". Other keywords such as "status epilepticus had moderate precision (61%) but poor recall. These results suggest that the classification problems under study are nontrivial, and justifies the effort required to develop a machine-learning based classifier.

B. Document normalization

Before document normalization we identified a vocabulary of approximately 12,000 distinct words by pooling across all 42,972 reports. This vocabulary shrank to 7,800 after customized spelling correction, to 5,414 after stemming, and to 271 after excluding the words that occurred less than 1 time per 100 documents. A total of 220,223 EWS were generated using the 271 words from these documents. The number of EWS shrank to 318 after discarding EWS that occurred less than 150 times within the labeled training data. Thus the total number of candidate features that could be used in training classifiers was 271+318 = 589. From the 3,277 labeled EEG reports, we identified by exhaustive manual review a synset of 7 words and phrases that always occurred in sentences relevant to determining the presence or absence of seizures.

C. Pruning

Figure 2 shows the results of our pruning experiments. Pruning yielded a maximum average cross validation area

under the ROC curve (AUC) of 99.78%, when the number of features reached 62. The minimum number of features for which the AUC was still within 1 standard deviation of the maximum was 38, with average cross validation AUC value of 99.73%. Further pruning below 25 features resulted in rapid performance deterioration. Based on these results, in the remainder of the experiments we use classifiers pruned back to 38 features.

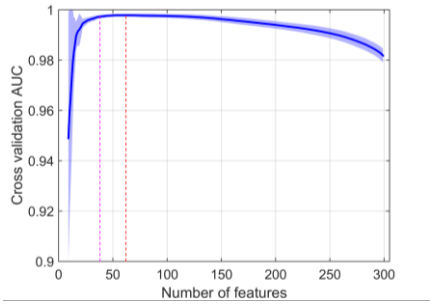

Fig. 2. Backward sequential feature selection performance curve calculated with respect to the testing data using 500 rounds of repeated random subset cross validation (CV). The average AUC value is shown as a solid blue line, and the 95% confidence intervals are shown as pale blue shading.

D. Performance

Figure 3 summarizes the performance analysis for our EEG report classification experiments with the final learned classifier for detecting reports with seizures. In ROC curve analysis, our final classifier achieved an average [95% CI] ROC AUC value of 99.22 [99.05, 99.38] % on training data, and 99.05 [98.79, 99.32] on testing data. Since class sizes are imbalanced in our case, we also performed precision-recall curve (PRC) analysis and calculated AUC for PR curves. The average PRC AUC was 97.89 [97.29, 98.49] % for training data, and 97.36 [96.30, 98.42] % for testing data.

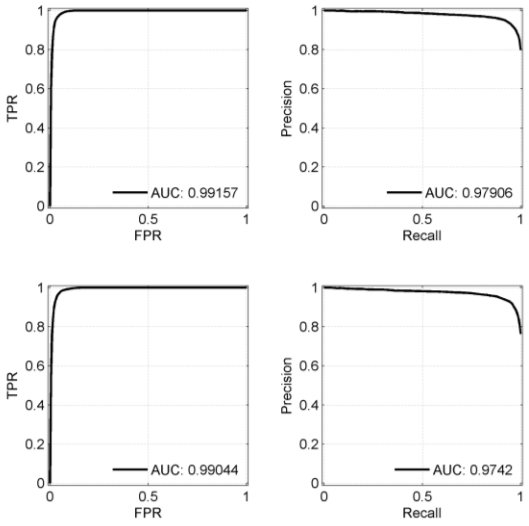

Fig. 3. Receiver operating characteristic (ROC) curves for training data (A) and testing data (B), and precision-recall (PR) curves for training data (C) and testing data (D).

Performance results were similar for identifying reports with epileptiform discharges. For this problem, the average ROC AUC was 98.79 [98.71, 98.97] for training data, and 97.53 [97.4, 97.69] for testing data. The average PRC AUC value was 98.66 [98.54, 98.80] for training data, and 98.39 [98.36, 98.53] for testing data. The closeness of average AUC values for the training and testing data suggests that our classifier training strategy largely avoids overfitting, and is thus likely to generalize well to new EEG reports, at least from the same institution.

III. Conclusion

We have described an extremely effective and efficient automated method for classifying free text EEG reports. We have demonstrated our method on the problems of identifying reports that describe seizures and/or epileptiform discharges. For these problems our method achieves nearly perfect discrimination. Our method succeeds despite the presence of strong dependence of word meaning on context, due to a specially tailored feature selection methodology. Because our solution is based on supervised machine learning techniques, it should be readily extensible to other patterns of interest in EEG reports, such as rhythmic and periodic patterns, and the so-called 'benign variant' patterns, by simply re-labeling the reports and re-running the training process. The methods described in this work thus open the way for large-scale mining of EEG report archives.

References

[1] P. W. Kaplan and S. R. Benbadis, "How to write an EEG report," Neurology, vol. 80, no. 1 Suppl 1, pp. S43–S46, Jan. 2013.

[2] P. Norvig, "How to write a spelling corrector," Online at: http://norvig. com/spell-correct. html, 2007.

[3] M.F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3, pp. 130–137, Mar. 1980.

[4] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta (BBA) - Protein Structure, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[5] Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159 (1997).

[6] R. Caruana and D. Freitag, "Greedy Attribute Selection," in In Proceedings of the Eleventh International Conference on Machine Learning, 1994, pp. 28–36.

[7] M. J. Schuemie, E. Sen, G. W.t Jong, E. M. van Soest, M. C. Sturkenboom, and J. A. Kors, "Automating classification of free-text electronic health records for epidemiological studies," Pharmacoepidemiol Drug Saf, vol. 21, no. 6, pp. 651–658, Jun. 2012.

[8] K. Yadav, E. Sarioglu, M. Smith, and H.-A. Choi, "Automated outcome classification of emergency department computed tomography imaging reports," Acad Emerg Med, vol. 20, no. 8, pp. 848–854, Aug. 2013.

[9] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, and R. A. Dudley, "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," J Am Med Inform Assoc, vol. 21, no. 5, pp. 871–875, Oct. 2014.

[10] I. Goldstein, A. Arzumtsyan, and Ö. Uzuner, "Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports," AMIA Annu Symp Proc, vol. 2007, pp. 279–283, 2007.

[11] M. Yetisgen-Yildiz, M. L. Gunn, F. Xia, and T. H. Payne, "A text processing pipeline to extract recommendations from radiology reports," J Biomed Inform, vol. 46, no. 2, pp. 354–362, Apr. 2013.

[12] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in IN AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.