Confidence in Medical Decision Making: Application in Temporal Lobe Epilepsy Data Mining

Shobeir Fakhraei[1,2]
shobeir@wayne.edu

Hamid Soltanian-Zadeh[2,3]
hamids@rad.hfh.edu

Farshad Fotouhi[1]
fotouhi@wayne.edu

Kost Elisevich[4]
nskoe@neuro.hfh.edu

1. Department of Computer Science Wayne State University Detroit, MI, USA

2. Image Analysis Lab. Dept. of Radiology Henry Ford Health System Detroit, MI, USA

3. CIPCE, School of Elec. and Comp. Eng. University of Tehran Tehran, Iran

4. Department of Neurosurgery Henry Ford Health System Detroit, MI, USA

ABSTRACT

Prior to neurosurgical resection of abnormal brain tissues in mTLE patients, focal points of the seizure should be identified via a set of examinations. Once decisive evidence is not present in noninvasive clinical profile of mTLE patients, extraoperative Electrocorticography (ECoG) is required which is the practice of using electrodes placed directly on the exposed surface of the brain. Through classification techniques on a dataset of mTLE patients, we have studied the possibility of reduction of such requirement and shown significant results. Furthermore, we compared the performance of six well known classifiers using the area under receiver operating characteristic (ROC) curve (AUC) and a proposed measure of decision confidence. We have shown that in critical domains such as medicine, use of AUC does not provide sufficient information about the confidence of the classification and further measures are needed.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-data mining; I.2.6 [**Artificial Intelligence**]: Learning; I.5.2 [**Pattern Recognition**]: Design Methodology-feature evaluation and selection

General Terms

Algorithms, Performance, Experimentation

Keywords

classification; decision confidence; performance evaluation; missing values; temporal lobe epilepsy; lateralization; AUC; confident prediction rate

1. INTRODUCTION

Epilepsy is a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiological, cognitive, psychological and social consequences of this condition [1]. Mesial temporal lobe epilepsy (mTLE) is the most commonly investigated and operated form of localization-related epilepsy. Clinical study of this disorder has become increasingly more elaborate, particularly through electrographic and imaging applications. Neurosurgical resection of the abnormal brain tissues in patients suffering from mTLE is a way of eliminating and reducing the occurrence of epileptic seizure onsets. Prior to such operation, focal points of the seizures should be identified via a set of examinations.

Availability of several diagnostic methods from multiple sources results in creation of high-dimensional spaces where data analysis and decision making become intricate tasks without the aid of appropriate tools. To lateralize the seizure focus in mTLE patients, several noninvasive clinical attributes are investigated. Such attributes include semiology, neuropsychological profiles, pathology, electrographic features, and magnetic resonance (MR) and single photon emission computed tomography (SPECT) imaging.

Once decisive evidence is not present in noninvasive clinical profiles of mTLE patients, extraoperative Electrocorticography (ECoG) is required. ECoG is the practice of using electrodes placed directly on the exposed surface of the brain to record electrical activity from the cerebral cortex. Besides the financial burden of this procedure, ECoG imposes further distress on patients and their families. In our paper, patients whose standard noninvasive evaluations are sufficient for their lateralization are referred to as phase I patients and those who require ECoG are referred to as phase II patients.

Since data mining techniques have been successfully applied in various biomedical domains to study complex diseases [2], such approach is applied in this study to provide decision assistance in lateralizing focal epileptogenicity. The goal of this paper is to reduce the need for ECoG via data mining techniques and finding the best classifier for this purpose. However, since decision making is highly critical in medical domains, classifiers that result in higher decision confidence are preferred. To be able to evaluate such confidence in different classifiers, we propose a new measure and compare it with well known area under receiver operating characteristic (ROC) curve (AUC) measure.

Six classifiers are applied to the lateralization task with preoperative data of patients to assess possibility of predicting the correct side of abnormality. These data exclude the invasive ECoG measurements. In the following sections, different features of the system, preprocessing stages, and classification tasks are explained. Furthermore, more details about the confidence measure are provided.

2. DATASET

To integrate several clinical attributes of TLE patients from various sources and subsystems, human brain image database system (HBIDS) [3], which is a clinical and imaging database of

KDD-DMH'11, August 21, 2011, San Diego, California, USA.

TLE patients, is developed at the radiology research department of Henry Ford Health System in Detroit Michigan. Details about the attributes and patient cohort used in this study are provided below.

2.1 Attributes

The HBIDS database contains several clinical attributes including risk factors underlying the condition, semiology, both pre- and postoperative neuropsychological profiles, location of surgery, pathology and outcome according to the Engel classification. Descriptive electrographic features include interictal waveforms, their location and predominance as well as ictal onset location. Both magnetic resonance (MR) and single photon emission computed tomography (SPECT) (ictal and interictal) imaging is included with the provision for quantitative, semi-automated assessment of compartmental volume, fluid-attenuated inversion recovery (FLAIR) mean signal and standard deviation, and texture analysis. Wada study results are also available.

2.2 Patient Cohort

Not all surgical resections result in complete relief from seizures due to various and possibly unknown reasons. Therefore, outcome of the epilepsy surgery is reported in the database according to Engel classification (class-I: free of disabling seizures, class-II: rare disabling seizures, class-III: worthwhile improvement, class-IV: no worthwhile improvement).

Cases with postoperative outcome of free of disabling seizures (Engel class-I) confirm a definitive laterality of focal epileptogenicity by human experts. To obtain ground truth for the classification evaluation, only such patients are selected from the database.

In this study, 79 patients with Engel class-I outcome are selected (31 males, 48 females) with 197 medical features. The patients have an average age of 38y (S.D. 12.2). Temporal lobe epileptogenicity is found to be on the left side in 43 patients and the right side in 36 patients. In 46 patients, standard noninvasive evaluations lateralize the TLE sufficiently well to proceed with resection of the site of epileptogenicity directly, whereas, 33 patients require ECoG (41.7%).

The dataset contains missing values in different features due to various reasons such as inability to perform all medical tests for each patient. Missing values are identified for EEG features in 21% of cases, for Wada studies in 31%, for SPECT imaging features in 35%, and for FLAIR and volumetric imaging in less than 10% of cases. The missing values of the remaining features are found in about 20% of cases on average.

3. FEATURE SELECTION

It is known that the prediction accuracy of practical machine learning algorithms degrades when faced with many features that are not necessary for predicting the desired output. In our case, with 79 patients and 197 attributes, the need for feature selection is apparent. Feature selection is utilized in biomedicine and bioinformatics in diagnostic value evaluation of a medical test and discovery of biomarkers [4].

However, we have to consider the characteristics of our domain such as missing values and class imbalanced distribution. Although the number of patients are chosen to be proportionate in each class (side of abnormality), missing values are not randomly distributed among classes and elimination of them imposes class imbalance problems.

To deal with these issues, we have applied a heterogeneous ensemble of single variable classifiers to rank the medical features based on their individual predictive performance. Only the patients with properly recorded values are considered in each feature evaluation. Feature performances are evaluated based on the area under the receiver operating characteristic (ROC) curve (AUC) to address any class imbalance problem due to the missing value elimination. As reported in our previous papers [5, 6] with more details, final score of each feature is calculated using the average AUC over multiple classifiers:

$$P(f_i) = \sum_{c_k \in C} AUC(c_k, f'_i)/|C| \quad (1)$$

where all missing values of a feature f_i are removed to generate f'_i and c_k is the classifier that belongs to the classifier pool C of the classifiers. $|C|$ indicates the number of classifiers.

Using this method, imaging, EEG and Wada attributes are ranked as the most discriminating features with regards to lateralization. Table 1 summarizes the performance of the top ranking attributes. We have reported more detailed clinical attribute rankings based on different patient cohorts in [7].

Table 1. Discriminative power of top ranking diagnostic features. Locations of EEG activities are numbered according to their dominance, as 1 being the highest amount of activity and 3 being the lowest.

Attribute	Avg. AUC
FLAIR standard deviation ratio (R/L)	0.914
Compartmentalized ictal SPECT subtraction (R-L)	0.912
Ictal EEG location (R/(R+L))	0.905
Interictal sharp wave EEG location (1)	0.878
Hippocampus volume ratio (R/L)	0.807
FLAIR mean signal intensity ratio(R/L)	0.803
5 Texture ratios (R/L)	0.790
Interictal sharp wave EEG location (2)	0.785
1 Texture ratio (R/L)	0.779
Interictal slow wave EEG location (1)	0.778
Wada subtraction number of correct answers (R-L)	0.698

To build a classifier, a feature subset from the top ranking features must be selected. We use a difference based cut off method described in [8]. Starting from an empty subset, features are added to the subset and their contributions to the improvement of the classification performance are measured. When improvements are not substantial, more features are not added to the subset.

Since our dataset contains many missing values, using more features introduce more missing values. To avoid including unnecessary features, the ones that have high correlations with any other feature already in the subset are eliminated from the final selection. Nine of the attributes listed in Table 1 excluding texture features that are highly correlated with other imaging features are selected in the final subset.

4. CLASSIFICATION

Since our goal is to predict the side of abnormality in patients suffering from mTLE and reduce ECoG requirements, a classifier could be built only on the patients who underwent such operation. However, there are two reasons to include all the patients in the classifier. First, this decision support system should also be able to predict side of abnormality in phase I patients, so it could provide reassurance for the experts decisions. Secondly, there are limited number of phase II patients in the dataset and elimination of phase

I patients reduces the classifier's learning power. Furthermore, having 9 features in the final subset increases the risk of over fitting with limited number of samples. Therefore, we have performed the classification task using all patients with good surgical outcomes. Figure 1 demonstrates the relative placement of phase I and phase II patients in a scatter plot of FLAIR mean signal and standard deviation ratios. It is seen that phase I patients populate the space and helps with the classification task.

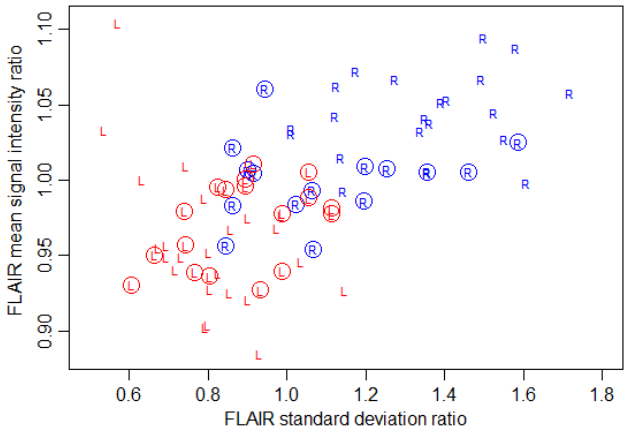

Figure 1. Patients in scatter plot of FLAIR standard deviation and FLAIR mean signal intensity ratios. Side of abnormality in patients is shown with "R" and "L" letters, respectively. Phase II patients are outlined.

5. PERFORMANCE AND CONFIDENCE EVALUATION

To evaluate the performance of the classifiers, AUC which is constructed by plotting the true positive rate versus the false positive rate by changing the decision threshold or boundary is calculated [9]. Using leave-one-out cross validation, the probability of "right" class membership is calculated for each patient and the ROC plots are generated using these probabilities.

However, since our domain has zero tolerance for invalid decision, although the classification is binary, two thresholds should be used, and the final classification response should be "left", "right", or "undecided". This is to ensure that predictions are only provided for samples that could be lateralized with certainty by the classifiers and avoid predicting the laterality for the ones with lower probabilities. To achieve such classifier, any chosen thresholds have to be set on points where no mistakes are made when the side of abnormality is predicted. The limits of the thresholds are invalid predictions with the highest predicted probabilities. As an example, when using the probability of patients having abnormality in their right side, threshold limits are α and β ($\alpha > \beta$) where all patients with predicted probabilities higher than α and lower than β are correctly classified and the ones right below α or right above β are predicted incorrectly. The system response for patients between α and β is "undecided".

Using this method, there will be no incorrect classifications, but only undecided cases. The number of predicted cases will be a measure of preference for the classifiers. This measure penalizes the classifiers that predict an incorrect side of abnormality with high probability. However, since the actual predictions thresholds could be chosen more conservatively than the very limits, performances of the final classifiers have to be evaluated after the thresholds are set. Two sets of sample holdouts are required for evaluation and final test sets for this reason. However, the α and β limit are the upper bounds for the classifier performance in this fashion and could indicate the classifier's potential in such classification. We refer to this measure as "confident prediction rate" (CPR):

$$\text{CPR} = \frac{\#\text{ of possible confident predictions}}{\text{Total \# of samples}} * 100 \qquad (2)$$

where samples with predicted probabilities more extreme than α and β are possible confident predictions.

Compared to AUC, the later performance measure will provide another detailed insight into the classification. In our experimental results, we show that a high AUC is not correlated with CPR, and despite high AUC, the number of possible confident predictions can be extremely low.

6. EXPERIMENTAL RESULTS

The codes to support different stages of the experiments are implemented in Java, WEKA, and R. Since most patients have at least one feature with a missing value, elimination of patients with missing values is not possible. Therefore, neutral values are used to impute the missing values, 1 for attributes representing a ratio and 0 for attributes representing a subtraction or deviation. For categorical attributes, missing values are replaces with "N/A".

The classifiers included in this study are those of naïve Bayes (NB), support vector machine (SVM), 3-nearest neighbors (3NN), multilayer perceptron(MLP), logistic regression (LR), and random forests (RF). In addition, several variations of the selected features subset are used to provide more insight into the dataset and feature sets predictive values. Experiments are conducted on all selected features, top four features, imaging features only, EEG features only, and Wada feature only. Leave-one-out cross validation is used to evaluate the performance of each classifier. The results are summarized in Table 2.

Table 2. Performance evaluation of different classifiers on different feature subsets.

(a) AUC of the classifiers using different feature subsets

Features	NB	SVM	MLP	3NN	LR	RF
All 9	0.993	0.959	0.978	0.964	0.986	0.968
Top 4	0.973	0.970	0.974	0.952	0.966	0.957
Imaging	0.982	0.981	0.957	0.929	0.975	0.916
EEG	0.951	0.951	0.943	0.964	0.925	0.958
Wada	0.609	0.660	0.599	0.596	0.660	0.689

(b) Confident prediction rate (CPR) of the classifiers using different feature subsets

Features	NB	SVM	MLP	3NN	LR	RF
All 9	84.8%	36.7%	72.2%	43.0%	44.3%	64.6%
Top4	65.8%	72.2%	54.4%	38.0%	41.8%	36.7%
Imaging	81.0%	75.9%	59.5%	0.0%	72.2%	0.0%
EEG	53.2%	65.8%	45.6%	73.4%	30.4%	62.0%
Wada	10.1%	13.9%	12.7%	22.8%	13.9%	12.7%

From Table 2a, it is seen that the best results are generated using all 9 features of imaging, EEG, and Wada. In this case, naïve Bayes generates the best performance results in terms of AUC. Using only the top 4 features, namely FLAIR standard deviation ratio (R/L), compartmentalized ictal SPECT subtraction (R-L), ictal EEG locations (R/(R+L)), and interictal sharp wave EEG location (1) comparably good performances are resulted suggesting the possibility of classification using only these features. Similarly, imaging only and EEG only features also produce good classification results. However, despite the general

confidence in Wada tests, our experiments suggest that such test does not provide sufficient information for reliable lateralization.

More interestingly is the dissimilarity between the AUC and the CPR measure that we discussed in the previous section. Table 2b shows that although some classifiers generate results with high AUCs but since they incorrectly predict some classes with high probability, they could not be trusted in a sensitive domain such as medical decision support. An example of such case is in using LR with all 9 features which generated the AUC of 0.986 and CPR of 44.3% while MLP results in lower AUC of 0.978 with higher CPR of 72.2%. In a medical domain such as this case, MLP should be preferred over LR despite the AUCs suggesting otherwise. Figure 2 demonstrates the ROC curves of six classifiers in this study using all features. It could be seen that although LR, SVM, and 3NN generate high AUCs but their ROC curves diverge from the vertical axis at the beginning of the curve, decreasing the CPR measure.

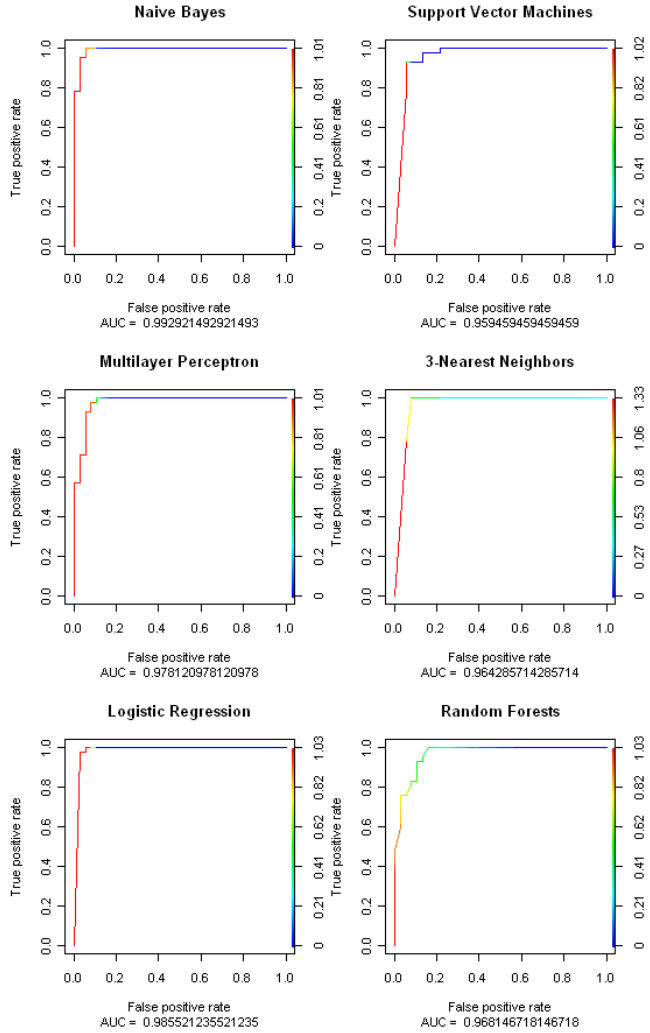

Figure 2. Receiver operating characteristic (ROC) curves of the six classifiers used in the study.

7. DISCUSSION AND CONCLUSION

Using six classifiers, we showed the possibility of using data mining techniques to build a decision support system that could potentially lateralize 84.8% of the patients with high confidence without the need for extraoperative Electrocorticography (ECoG). Lacking such system, only 58.2% of patients were lateralized by domain experts using noninvasive methods. Using this method, it is potentially possible to lateralized 78.8% of the phase II patients, while only 8.7% of the phase I patients will be undecided.

We also demonstrated that AUC does not provide sufficient information about the confidence of the classification and other measures such as our proposed "confident prediction rate" (CPR) are needed in domains such a medicine. Using the experiments, we demonstrated that classifiers that generate high AUCs might not be sufficiently confident for domains that require reliable predictions.

8. ACKNOWLEDGMENTS

This work was supported in part by NIH grant R01-EB002450.

9. REFERENCES

[1] C. P. Panayiotopoulos, A clinical guide to epileptic syndromes and their treatment: Springer Verlag, 2010.

[2] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International Journal of Medical Informatics, vol. 77, pp. 81-97, 2008.

[3] M.-R. Siadat, H. Soltanian-Zadeh, F. Fotouhi, A. Eetemadi, and K. Elisevich, "Data modeling for content-based support environment (C-BASE): Application on epilepsy data mining," in 17th IEEE International Conference on Data Mining Workshops, ICDM Workshops 2007, October 28, 2007 - October 31, 2007, Omaha, NE, United states, 2007, pp. 181-186.

[4] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, p. 2507, 2007.

[5] S. Fakhraei, H. Soltanian-Zadeh, F. Fotouhi, and K. Elisevich, "Consensus feature ranking in datasets with missing values," in 9th International Conference on Machine Learning and Applications, ICMLA 2010, December 12, 2010 - December 14, 2010, Washington, DC, United states, 2010, pp. 771-775.

[6] S. Fakhraei, H. Soltanian-Zadeh, F. Fotouhi, and K. Elisevich, "Effect of classifiers in consensus feature ranking for biomedical datasets," in 4th International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO'10, Co-located with 19th International Conference on Information and Knowledge Management, CIKM'10, October 26, 2010 - October 30, 2010, Toronto, ON, Canada, 2010, pp. 67-68.

[7] S. Fakhraei, H. Soltanian-Zadeh, K. Elisevich, and F. Fotouhi, "Attribute ranking for lateralizing focal epileptogenicity in temporal lobe epilepsy," in 17th Iranian Conference in Biomedical Engineering, ICBME 2010, November 3, 2010 - November 4, 2010, Isfahan, Iran, 2010, p. Isfahan University of Medical Sciences; Iranian Society of Biomedical Engineering; Medical Image and Signal Processing Research Center (MISP).

[8] A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benitez, "Empirical study of feature selection methods based on individual feature evaluation for classification problems," Expert Systems with Applications, vol. 38, pp. 8170-8177, 2011.

[9] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," Machine Learning, vol. 31, pp. 1-38, 2004.