Seizure Forecasting with Intracranial EEG recordings

Pavankumar Reddy Muddireddy
Department of Electrical and Computer Engineering
University of Illinois at Urbana Champaign
net-id: muddire2
pavankumarreddy.93@gmail.com

Abstract

There has been an increasing amount of interest in personal healthcare monitoring in recent times. When it comes epilepsy, the patients suffers not only from the disease but also anxiety. Recent research has suggested that these epileptic seizures could be predicted by analysing the intracranial Electroencephalographic data. This paper explores and tries to address this problem by using different machine learning algorithms and comparing them.

1 Introduction

1.1 Epilepsy

An epileptic seizure is a brief episode of signs caused by an unexpected and abnormal amount of synchronous neuronal activity in brain. It is commonly referred to as a fit and is characterized by an uncontrollable jerking movement and usually followed by a loss of awareness for a brief period. Epilepsy is a brain disorder characterized by recurrent and unpredictable interruptions to the brain function in the form of epileptic seizure. Epilepsy affects nearly 1% of the world population and there is a chance of 40-50%[4] of experience a recurring seizure in those who have experienced it. Medication often is not very helpful when it comes to treating this disorder. More than the seizure itself, the dangerous part of the disease is the unpredictable nature of the seizure which can lead to persistent anxiety among the patients. A seizure setting in when the patient is performing activities such as driving, swimming etcetera can be fatal. So giving out predictions about when the seizure can set in the patient can prove to be extremely helpful.

Recent research suggests a variety of electrophysiological changes to the brain before an actual seizure sets in. There has been an increasing amount of interest in the area of seizure forecasting and prediction in the recent times. This accompanied by an increasing interest in the industry with the rise of smart devices including health and fitness bands along with a variety of watches. So a successful seizure prediction model with appropriate technology could pave way into these devices where one gets notified ahead time if he might going to suffer a seizure along with the medical practitioner who could perform tests and provide appopriate medication to the patient.

1.2 Prediction of Epileptic Seizures

So, the next obvious question would be whether this can be done i.e. if an epileptic seizure could be predicted ahead of time using any statistical or computational techniques. It has been suggested that around 50% of patients have anticipated a feeling that they would be getting a seizure prior to getting one suggesting some kind of change in activity in the as one progress towards a seizure. So, there has been an increasing investigation among the various data recordings of the brain from different sensory sources for this change in activity. One that holds promise for this kind of study is the intercranial EEG data. EEG stands for Electroencephalography where one records the electrical activity of the brain by attaching a wide number of electrodes to the scalp

of a person for a short period of time usually 20-40 minutes. But this kind of recordings from the scalp are not of particularly great resolution and are usually smeared with cerebrospinal fluid and skull. So, to perform analysis on patients with epilepsy, the recording are usually taken from an electrode array usually in the form of strips of grids under the skull (dura matter), to get a much higher resolution EEG recordings commonly referred to as intracranial (meaning within the skull) encephalography.

1.3 Problem at hand

It has been suggested that the brain goes through four major stages before and after seizure and these are referred to as *inter-ictal*, *pre-ictal*, *ictal* and *post-ictal* indicating the time from seizure. interictal refers to the state far from seizure while preictal refers to one before seizure. Ictal and postictal refer to states during and after seizure. There has been an increasing evidence pointing to existence of these states and this is the fundamental assumption on which most the current techniques to predict epileptic seizures are based on.

Many statistical approaches have been studied to predict the states of a signal including time-series analysis of the data. Approaching the problem from a machine learning point of view, it boils down to differentiating a given signal to interictal (far away point) to perictal (just before) states. Hence, the machine learning task is a binary classification problem. The definition of preictal state among the full EEG recording is differnt in different studies and the data used in performing experiments referred to in this paper uses the recordings one hour prior to seizure as preictal segments with a 5 minute horizon. So essentially they refer to the interval 65 minutes before the seizure to 5 minutes before the seizure.

1.4 Limitations with current techniques and current approach

Most of the techniques pointed to by literature uses univariate analysis and most importantly stick to simple binary classification algorithms such as linear SVMs. Furthermore, as pointed to by Piotr W. Mirowski et al that most of these methods adopt unneccessary feature contraction i.e. reduce the number of features by some sort of averaging to a very small number (some times even 2) and use a simple linear classifier where the decision boundary is usually a line (in 2D case) or a hyperplane. The assumption that underlies such unnecessary feature reduction would be that the these very small number of features of different data segments encompasses the critical information which would help linearly separate the data. Such assumptions are usually proven wrong with low accuracy and unpredictable performance on a differnt data set from a different source such as another animal or a patient. Higher dimensional feature space could result in a linearly separable data owing to large distances between points in a high dimensional space. These problems are overcome to some extent in this paper by usage of non-linear classifiers and a large number of features as explained in the following sections. A very large feature space might not be very optimal as unnecessary features in the data could reduce the accuracy but choosing the right features in itself could be a very challenging task especially on data such as EEG recordings on which there are no established models prior to this. So, an empirical/experimental approach has been adopted in choosing an optimal number of *right features* which could help classify the data the best.

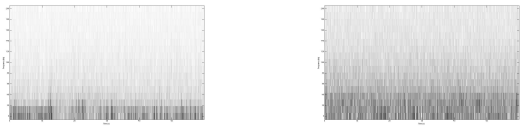
1.5 Forecasting using classification

The experiments have been performed on data acquired from intercranial EEG data taken from dogs with naturally occuring epilepsy. The dataset used is a mutli-channel EEG with upto 16 channels for acquiring the data. Firstly appropriate features have been extracted from the data. Then different classifiers including kernel SVM and Random Forests classifier have been used.

2 Feature Extraction

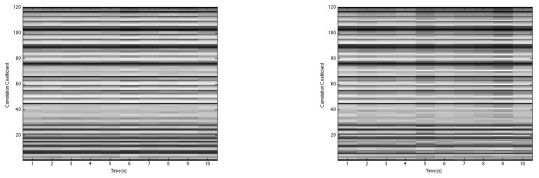
2.1 Frequency Domain Features

Feature extraction for data such as EEG where the model is completely unknown is as mentioned earlier a difficult task. Appropriate feature choice can prove to be critial in such a setting. Since, the time domain features of both preictal and interictal are visually quite random, visual inspection of data in time domain donot provide any indications of good features. So, frequency domain characteristics of the data have been examined. The data segments



(a) Preictal Segment (b) Interictal Segment

Figure 1: Frequency content of a preictal segment and an interictal segment



(a) preictal (b) interictal

Figure 2: Correlation features of preictal and interictal data

both preictal and interictal acquired were sampled at 400Hz and were split into groups of 1 hour data (as mentioned above). Each group further had six 10 minutes segments of the data. For further performing analysis, each segment was split into 10 pieces of 1 minute each. To put it all together, the data finally consisted of groups of 1 minute segments for each preictal and interictal data segment. An average spectrogram on one of the 1 minute segment of a preictal and interictal segment is shown below.

A marked different can be seen from the preictal data segment to the interictal data segment. After experimentation with various features related to frequency, power in frequency bands (spectral band power) as described are chosen as a set of features for each of the 1 minute interval - 10 bands in the first 50 Hz, 10 bands in the next 50 i.e 50Hz to 100Hz, 5 bands in 100-200Hz have been chosen as features. Small number of features were chosen at higher ranges owing to the fact that the amount of power in those ranges is quite low compared to the lower ranges and it proved to be a correct choice while experimentation. The accuracy didnot change very much when the FFT features from these bands were appended to the feature set (it actually decreased in some cases too). The 0 Hz frequency which is DC component (average value) in the input data is also added. These frequency domain features are extracted by performing a 32768 point FFT on each of the 1 minute epoch and then power spectral density for each of the band described above is calculated.

2.2 Time Domain Features

The time domain features include the variance across the channels and cross-correlation between the channels. So, for datasets with 16 channels, a cross-correlation matrix of size 16×16 has been constructed for every minute and removing the redundant features meant that only 120 were left. The cross-correlation between channels i and j could be calculated by,

$$CVar_{ij} = \frac{1}{N}\sum_{k=1}^{N}(X_{ik} - \mu_i)(X_{jk} - \mu_j)$$

$$CCor_{ij} = \frac{CVar_{ij}}{\sqrt{C_{ii}}\sqrt{C_{jj}}}$$

where $CVar$ is the cross-covariance matrix and $CCor$ is the cross-correlation matrix. There has been many suggestions in literature of using a maximum cross-correlation as features but since the number of features from the cross-correlation matrix small, all the co-efficients have been used as features. The cross-correlation features provide any time-domain linkange cross-channel behaviour in the time-series data before the onset of seizure. The co-efficients are calculated for each of the 1 minute segments and appended to the feature set.

So, the total feature size here would be 25×16 frequecy features per minute and $120 + 10$ (cross-correlation + variance) time-domain features per minute giving a total of 530 features for the data.

3 Classification of Preictal and Interictal Data

Prior to classification, PCA has been employed to reduce the dimensionality. PCA - Principal Component Analysis orthogonalises the feature space by successively choosing creating new feature space with decreasing variance. The transformation is defined in such a way that the first principal component has the largest variance followed by the second and so on. The new feature space has uncorrelated features in the observations. So, by eliminating the

low variance features, the dimensionality of the data can be effectively reduced. Experiments are run with varying number of features and 100 features are chosen to represent the data.

3.1 Linear SVM

The first algorithm used to compare the results is the discriminative model classification algorithm, the linear SVM i.e. normal SVM with out any "kernel trick" employed. The linear SVM is employed on the input feature space with the standard 'l2' loss function. The problem is formulated below:

$$min \frac{1}{2} w^2 + C \sum_{i=1}^{n} \xi_i^2$$

$$\text{s.t.} \quad y^{(i)}(w \cdot x^{(i)} + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

The classification was performed after the parameter C was tuned to an appropriate value. By classifying with SVM with a linear kernel, the basic assumption that the data is nearly linearly separable is made and that the classification depends on few key important data points (the support vectors). These assumptions could be valid due to the high dimensionality of the data.

3.2 SVM with RBF kernel

The second algorithm used for the classification was the RBF kernel - Radial basis function kernel. The SVM formulation with "kernel trick" is described below (a dual version of the primal version described above has been shown below),

$$max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$K(x_i, x_j) = \phi(x_i).\phi(x_j)$$

$$\sum_i \alpha_i y_i = 0$$

$$C \geq \alpha_i \geq 0$$

In this formulation the function K is the kernel which transforms the input feature space x to a new feature space $\phi(x)$. The "kernel trick" lies in the fact that the inner product between samples in this new features space need not be explicitely calculated by rather could be done in the original space using the kernel function K. This is extremely important as the new feature space could be very large rendering it computationally difficult if not impossible (as in this case where the feature space consists of coutably infinite features) to calculate the inner product. The RBF kernel in particular is

$$K(x, x') = e^{(\frac{|x-x'|^2}{2\sigma^2})}$$

The $1/2\sigma^2$ is also called the γ parameter. It can be observed that the RBF kernel essentially estimates the similarity between points by their Euclidean distance. Points which are closer are give a score near 1 while points which are farther away give scores near 0. The infinite feature space of the RBF kernel is,

$$e^{(\frac{x-x'^2}{2})} = \sum_{j=0}^{\infty} \frac{(x^T x')^j}{j!} e^{(\frac{x^2}{2})} e^{(\frac{x'^2}{2})}$$

In an informal interpretations, using a polynomial kernel of degree 'n' is like using all functions whose n+1 derivative is a constant while using a RBF kernel is like using all functions which are infinitely derivable. So, the RBF kernel is expected to perform better than its linear couterpart when the data is not linearly separable. The paramter gamma is tuned to get ideal classification results.

3.3 Random Forest Classifier

The final classifier used was the random forest classifier. Random forest classifier is an ensemble learning methods where a number of decision trees are trained simultaneously and the decision is made based on the decisions of the individual trees. A simple decision tree used to learn the EEG feature data performed poorly. This poor performance can be attributed to traditional problem of overfitting in the decision trees. Pruning the tree sometimes increased the accuracy but not considerably. So, an ensemble of decision trees in the form of random forest was employed. Compared to normal bagging where differnt decision tree learners are grown with a random sampling of data (with replacements), the random forest classifier further adopts a different tree learning algorithm which is similar to traditional one except for splitting where at each candidate split dur-

ing training, a random subset of the features are selected. The tree bagging algorithm avoids correlation between trees by chosing random samples of training data but if one variable strongly predicts the outcome, then the trees all might end up with the variable. Random forests overcomes this by "feature bagging". Since, the number of training samples are not very large, the number of decision trees used were limited. The performance kept increasing till 11 trees and it decreased beyond that point. Hence, 11 decision trees were chosen for the ensemble learning.

3.4 Probability estimation

For each of the methods above the parameter $P(y|x)$ where y is the label - preictal or interictal and x is the 100 dimensional feature space is estimated. In many medical scenarios, probability estimates are much more important than hard estimation and evaluation is done by ROC AUC (Receiver operating characteristic - Area Under the Curve). This is further explained in the next section. The probability estimate for SVM is calculated by Platt Scaling[2] which is

$$p_i = \frac{1}{(1 + exp(A * f_i + B))}$$

[1]

For random forest however, instead of taking a majority vote to decide which label to assign (hard assignment), fraction of decision trees that estimate a particular label is estimated and and assigned to the example (soft assignment).

The evaluation methodology has been presented in the next section.

4 Evaluation Criteria and Results

4.1 Dataset

The standard dataset for intracranial EEG was Freiburg dataset (Seizure Prediction Project Freiburg - University of Freiburg). But this dataset is now superseded by the new European Epilepsy Database which consists of 60 patient datasets for €6000. Since, it was not feasible, the dataset from the kaggle competition *American Epilepsy Society Seizure Prediction Challenge* [2] was used. Since, the key was not published for the test data, only cross-validation tests are performed on the data as discussed. The experiments are performed on one dog with a dataset size of 7.72GB.

4.2 Perfomance Metrics and Results

A 5-fold cross validation test has been performed to estimate the performance of each classifier. The input data consisted of 24 preictal segments and 480 interictal segments of 10 minutes each. Further splitting the dataset ot 1 minute segment gave 5040 data points with 530 raw features and 100 features after performing PCA. Performing 5-fold cross-validation involved splitting the dataset into 5 segments and testing on one these segments with classifier trained on the remaining 4 segments. So, the training data consisted of 4032 samples and testing on the remaining. However, for later analysing the performance at the resolution of each segment (i.e 10 minute segment of preictal and interictal data), groups of 10 rounding to the nearest 10 multiple of 1008 (which is 1000) have been chosen for testing the data. The training data set was highly unbalanced with 20 to 1 positive to negative labelled examples. This led to very biased output and was tackled by random undersampling of the majority class - interictal segments (labelled 0). Undersampling is not the best approach to treating unbalanced data but the majority class which is interictal segments are recorded at random time instances far before the seizure sets in. The indicators for seizure before it sets in are believed to be present in the preictal segments. Hence, removing the interictal data segments, which are a base class for prediction, is justified. The data has been undersampled to get to a ratio of 4:1 for preictal to interictal samples.

The results of the cross-validation are presented in the table below.

The ROC-AUC score is the Area under the curve for Receiver Operating Characteristics. This is plot between True positive rate against False positive rate as the threshold is changed. Ideally one would like high true positive rate with very low false positive rate. So, higher the curve away from $y = x$ line, the better. The area under the curve is an indication of

[1] This has been achieved by using python's sklearn library sklearn.svm.SVC which estimates the probabilities using Platt scaling

[2] currently the contest is expired

Sample	Lin SVM	Ker SVM	Ran Forest
1	68%	75%	83%
2	60%	80%	84%
3	55%	65%	83%
4	60%	69%	81%
5	59%	65%	79%

Table 1: Cross-validation Results.

Sample	Lin SVM	Ker SVM	Ran Forest
1	0.76	0.84	0.94
2	0.7	0.83	0.96
3	0.69	0.75	0.95
4	0.71	0.82	0.95
5	0.63	0.72	0.94

Table 2: ROC AUC Scores

how good the classifier is. The ROC curves for the above cases (2 samples are shown) have been plotted below (for Random Forest Classifier only),

Furthermore, the histogram depicting the distribution of probabilities among preictal and interictal datasegments is shown below (for one case),

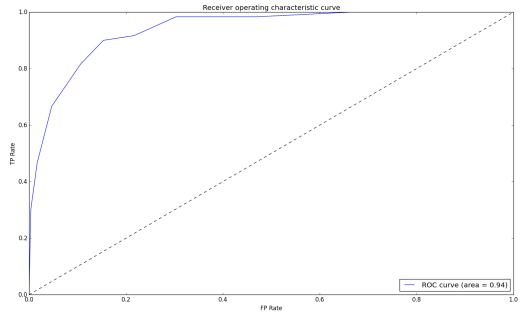
An approximately good demarcation could be seen in the histogram between the pre-ictal and interictal data. This explains the good ROC curves and high Area Under the Curve.

4.3 Discussion

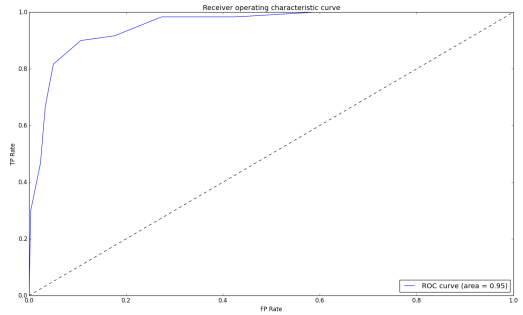
The above results clearly indicate that the given data is not linearly separable since the performance with linear SVM is very poor. It is seen that a kernel SVM with RBF kernel gives much better results but still is quite lagging. This might be due to absense of some unknown key features or even the higher order features are not fitting the data. Finally, random forests performed the best. This could be explained by non-linearity of the data and accurate learning (relatively) by the underlying decision trees. The ensemble method generalized much better compared to any single classifier explored in this paper which is expected from ensemble learning technique.

5 Related Work

As mentioned in the introduction, eliminating features and combining them to result in a very small

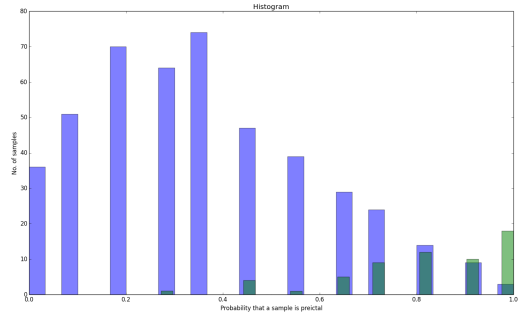


(a) Sample 1



(b) Sample 3

Figure 3: ROC curves



(a)

Figure 4: Histogram depicting the statistical probability distribution over different segments

dimensional space is one of the limitations which was overcome in this paper by choosing a large number of frequecy and time domian features. This kind of feature contraction could be seen in the work by Florian Mormann et al "On the predictability of epileptic seizures". Other techniques include using simple SVM or linear regression over preprocessed (filtered) frequency domain features as indicated by Temko et al in [3] and J. Jeffry Howbert et al in [3]. Although preprocessing and adding additional features might make the data linearly separable evidence suggest that non-linear classifiers such as Random Forest, Neural Networks have shown a better performance as indicated in [5].

6 Future Work and Conclusions

The input data set was heavily unbalanced. Although an argument was provided that undersampling the majority label would be a decent balancing technique, usage of sophisticated techniques would help improve the classification results. A better balancing algorithm such as such as SMOTE (SMOTE: Synthetic Minority Over-sampling Technique) might prove to be more effective than random undersampling as removing training exmaples is not the best solution. Further more additional time-domain and frequency domain features that encode the relationship between pairs of EEG channels, such as cross-correlation, nonlinear interdependence, difference of Lyapunov exponents and wavelet analysis-based synchrony such as phase locking as suggested by Piotr W. Mirowski et al. while exploring other classification algorithms such as neural networks could not only improve the accuracy but could also provide unique insights into the data itself and might pave way to understanding the intracranial EEG signals.

In this paper, three different classification algorithms have been presented for the purpose of seizure prediction. They have been applied on intracranial EEG data from a dog. Features extracted from the data were an aggregate from time-domain and frequency domain. The three algorithms have been analysed in this context and their performance reported. Kernel SVM performed modestly while random forest classifier outperformed both linear and RBF kernel SVM.

7 Implemetation

The algorithms and other programs including feature extraction has been implemented using MATLAB and Python. First various segments of both interictal and preictal data segments are loaded into MATLAB and the features are extracted and stored into a file. The classification and other analysis of the data has been performed in Python using the open source sklearn library. R has also been used for performing some of the analysis though not used for final results presented in this paper.

References

J. Jeffry Howbert, Edward E. Patterson, S. Matt Stead, Ben Brinkmann, Vincent Vasoli, Daniel Crepeau, Charles H. Vite, Beverly Sturges, Vanessa Ruedebusch,Jaideep Mavoori, Kent Leyde, W. Douglas Sheffield, Brian Litt, Gregory A. Worrell mail. 2014. *Forecasting Seizures in Dogs with Naturally Occurring Epilepsy*. 10.1371/journal.pone.0081920.

John C. Platt 1999 *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods MIT Press*, 61–74.

Temko, A and Thomas, E and Marnane, W and Lightbody, G and Boylan, G. 2011. *EEG-based neonatal seizure detection with Support Vector Machines*. *Clinical Neurophysiology*, 122(3):464–473

Fisher R, van Emde Boas W, Blume W, Elger C, Genton P, Lee P, Engel J. 2005. *Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)*. *Epilepsia* 46(4): 4702

Mirowski, Piotr W and LeCun, Yann and Madhavan, Deepak and Kuzniecky, Ruben. 2008. *Comparing SVM and convolutional networks for epileptic seizure prediction from intracranial EEG*. Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on, 244–249