Predicting Seizure Onset with Intracranial Electroencephalogram (EEG) Data - Project Report

Alex Greaves, Arushi Raghuvanshi, Kai-Yuan Neo

December 2014

1 Abstract

Epileptic patients have little to no warning about an oncoming seizure, and would benefit from knowing if they are about to have one because it allows them more time to find a safe place. The primary challenge in seizure prediction is differentiating between the preictal (pre-seizure) and interictal (baseline) states. In this paper, we developed a method to address the following goal:

Can we use EEG signals to accurately classify the preictal state in human and dog patients with naturally occurring epilepsy?

Using over 3,000 training examples, we investigated multiple feature extraction and classification algorithms including discrete wavelet transform (DWT), short-time Fourier transform (STFT), principal component analysis (PCA), k-Nearest Neighbors, logistic regression, and support vector machines. Using STFT, PCA, and logistic regression we developed a feature extraction and classification algorithm to classify EEG signals as preictal or interictal with an Area Under Curve (AUC) score of .75.

2 Introduction

2.1 Motivation

Over 65 million people worldwide currently live with epilepsy and 1 in 26 Americans will develop epilepsy in their lifetime. Patients with epilepsy are susceptible to spontaneous seizures, which are particularly dangerous if they occur in a potentially hazardous environment, such as behind the wheel of a car. Currently, the majority of epilepsy research funding goes towards anticonvulsant medications which are largely ineffective for about 40% of patients, leaving them just as susceptible to spontaneous seizures [1]. With recent developments in the wearable space, there is an increase in the usefulness of wearable devices that can take EEG readings [2]. For the first time, it is possible to track a patient's brain activity on a daily basis. If we can develop an algorithm that uses these EEG signals to predict when a seizure is going to occur, patients can be warned ahead of time that they are about to have a seizure, allowing them to lead a safer life.

2.2 Past Work

Until the last decade, seizure prediction was only possible by visual analysis of signals by certified neurologists. Since the advent of machine learning techniques and modern computing, seizure prevention researchers have focused efforts on deciphering EEG messages as a means of predicting seizures.

There were some early optimistic papers on seizure prediction using EEG data. However, these results were not tested in a rigorous, statistical fashion, and didn't perform better than random guessing when reproduced in the wild [4]. Following this 2007 review, there has been some progress in research in the space, however no algorithms have come close to the necessary accuracy needed for practical use [5]. In a response to the need to differentiate between preictal and interictal states, the American Epilepsy Society posed this problem as a Kaggle competition challenge [3].

One notable paper by Subasi et al. demonstrates that DWT combined with PCA and SVMs are most effective at classifying their test data. Subasi's experiments have 2 shortcomings: they only train and test on a small data set of 1600 samples, and their SVMs only use 1 kernel function, the radial basis function. The former is prone to over-fitting, and the latter means that there are potentially better-performing kernel functions and parameters to tune this algorithm [6].

Maiwald et al. test 3 non-linear prediction methods: effective correlation dimension, dynamical similarity index, and accumulated energy. These 3 prediction methods analyze distinct features of EEG data,

but do not aggregate over multiple features like a machine learning approach would. Maiwald et al. prove that the 3 proposed methods perform better than random or periodic methods of classification [7].

2.3 Our Work

With the current state of the field as a baseline, we will try different feature extraction and classification algorithms with the goal of developing a classification model with low variance and bias. Given roughly 100 gigabytes of data provided through the Kaggle competition, we seek to develop a high performing algorithm for classifying EEG signal segments as preictal or interictal.

3 Dataset

The data represents 10-minute clips of preictal and interictal training and test readings for a fixed number of electrodes on 5 epileptic dogs and 2 epileptic humans [3]. The following number of preictal, interictal, and test clips are provided for each:

Patient	# Preictal Training Clips	# Interictal Training Clips	# Test Clips
Dog 1	24	480	502
Dog 2	42	500	1000
Dog 3	72	1440	907
Dog 4	97	804	990
Dog 5	30	450	191
Human 1	18	50	195
Human 2	18	42	150

Figure 1: Number of preictal and interictal training and test clips

Each training example consists of a matrix, a vector, and some metadata. The matrix represents electrodes by time, and has samples of electrode values for the patient over 10 minutes. The rows contain the values of the signal at all time intervals for a given electrode. The columns contain electrode readings at a particular time segment. The names of the EEG electrodes are given in a vector. The time between interval, and total number of samples are given as scalars.

An example of the matrix for one training example is provided below:

-36	-44	-53	-69	-77	-86	-97	-114	-118	...
-26	-12	-6	-19	-31	-46	-49	-50	-44	...
24	24	19	3	-14	-25	-13	-15	-18	...
30	6	-28	-33	-19	1	17	8	11	...
17	24	23	27	15	1	-8	0	-3	...
-20	-20	2	5	9	3	-13	-9	-10	...
30	14	-3	-10	5	12	13	4	2	...
-90	-76	-39	-11	36	73	85	85	85	...
-60	-65	-72	-69	-67	-73	-79	-70	-52	...

Note that for each patient, there is no guarantee that the EEG sensors were placed in the exact same location on the patient's brains.

4 Methodology

4.1 Feature Extraction

The raw EEG data is over 100 gigabytes, so we use common EEG feature extraction methods to extract useful features. The raw amplitude data is not useful for prediction using traditional machine learning methods. The first challenge in classifying EEG data is to get it into a form in which we can apply traditional machine learning techniques. In the past this has been done using one of either Discrete Wavelet Transform (DWT) or Short-Time Fourier Transform (STFT), because these methods preserve time information while extracting frequency information and other features from EEG signals. Principal Component Analysis is also commonly used for dimensionality reduction. [6]

4.1.1 Discrete Wavelet Transform

DWT is implemented using what are called quadrature mirror filters (QMFs), which separate high and low frequency components of the signal. There are many variations of these QMFs, but a common choice, one that we used, is form 4 Daubechies filters (known as d4). When we pass the original signal through these filters we get two output signals called the approximation coefficients (low-frequency) and detail coefficients (high-frequency). We can further decompose the approximation coefficients into their own approximation and detail coefficients, the output of which is now $\frac{1}{4}$ the length of the original signal. We can repeatedly take the low-frequency filter output of this process and further decompose it to an arbitrary degree. A typical choice is 5 levels of decomposition and this is what we have implemented.

Therefore, DWT separates the signal into high and low frequency bands by repeatedly performing separation to arbitrary degree in a branching structure (Figure 2), allowing adjustment for more course or granular readings [8]. In Figure 2, h(n) represents

the high pass filter and g(n) represents the lowpass filter.

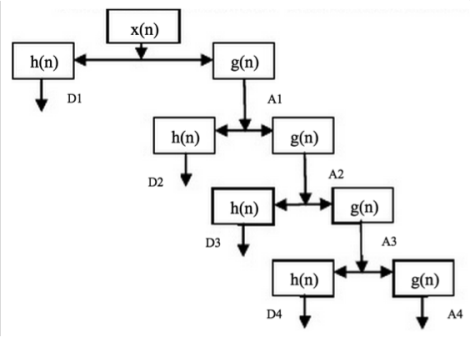

Figure 2: Branching structure of DWT

From the DWT coefficients at each level, we took as features the mean magnitude, mean power (magnitude squared), and standard deviation. In addition, we used as features the ratios of the mean magnitudes of adjacent levels.

4.1.2 Short Time Fourier Transform

STFT is used for the determination of sinusoidal frequency and phase content of local sections of signal. This feature extraction method was most effective on the Kaggle data. We used the Discrete Time - STFT as given in Figure 3.

$$X_{STFT}[m,n]=\sum^{L-1}_{k=0} x[k]w[k-m]\, e^{-j2\pi nk/L}$$

Figure 3: Equation for computing STFT

Where $x[k]$ denotes a signal and $w[k]$ denotes an L-point window function. The STFT can be defined as the Fourier transform of the product $x[k]w[k-m]$ [9].

In order to extract features from the EEG data, we performed STFT twice on each 10 minute segment (the second time with a time offset) to obtain frequency spectra for each 30-second segment and then for each 60-second segment. From this, we extracted features by computing mean power of the magnitude at each frequency within 1Hz bands up to 120Hz.

4.1.3 Principal Component Analysis

PCA is used to reduce the dimensionality of features by converting a set of possibly correlated variables into a set of linearly uncorrelated variables. It is an unsupervised projection that can project a high dimensional feature space onto a low dimensional hyperplane. The first principal component has the largest possible variance, and subsequent components have the next highest variances with the constraint of being orthogonal to the preceding components. We can truncate a list of features by only taking the first n principal components. These components have the largest variances, and likely have larger effects on the classification.

PCA improves efficiency of EEG classification because certain channels are placed on areas of the head that will be more relevant to seizure detection than others. PCA can reduce these noisy dimensions.

However, PCA is an unsupervised projection algorithm, so it can optimize out relevant information if the features are not pre-processed in a way that fits the model. When run on our DWT features, PCA resulted in poor performance.

Even for our STFT features dimensionality reduction didn't improve our performance with this data set, but the normalization and variance maximization of PCA led to better results. We find the principal components by maximizing the equation given in Figure 4 [10].

$$u^T \left(\frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)T} \right) u.$$

Figure 4: Closed form equation for finding the principal component

4.2 Classification

We used three standard classification algorithms: k-Nearest Neighbors, logistic regression, and support vector machines. We tuned the parameters for these algorithms by balancing overfitting and underfitting and maximizing the Area Under Curve (AUC) score.

4.2.1 k-Nearest Neighbor

As a baseline, we used k-NN to classify nodes based on their k closest neighbors in the feature space. To calculate the distance, we used an L2 norm, and we looked at $k = 22$ nearest neighbors, weighting preictal segments by a factor of $\frac{8}{3}$. k and the preictal segment weight are hyperparameters tuned to yield the best result. Out of the classification algorithms we used, this performed the worst, most likely due to the radically fewer number of preictal training examples than interictal examples. However, it provided a starting point for classification that brought us above the random guessing threshold.

4.2.2 Logistic Regression

Logistic Regression (LR) has been successfully used for EEG classification in literature. It is a flexible model that makes few assumptions on the prior structure of the data. In our implementation, LR is the best-performing model.

LR calculates probability of positive result based on features input into the logistic function. Our hypothesis function is described as follows [10]:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Figure 5: Our hypothesis and the underlying logistic function

We then used stochastic gradient descent to optimize the parameters θ. We trained one model on all 30-second samples and another on all 60-second samples, assigning double the weight for preictal segments. Then, for each test segment, we summed the output for all 30-second and 60-second samples. Classification was done by comparing this sum to a threshold, which was another hyperparameter.

4.2.3 Support Vector Machine

Support Vector Machines (SVMs) are models which assume a (mostly) linearly separable data set. This set can consist of features mapped to a high dimensional space in which they are linearly separable. By creating a margin which maximizes the functional and geometric margins between sets, SVMs have received attention in biomedical applications. We found that most kernel functions overfit our data using SVM. The linear kernel performed the best and provided an improvement over k-NN, but still overfit the data to some extent.

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

Figure 6: SVM objective function to calculate optimal margin with error penalty

5 Results

Correctness scores are generated by calculating the area under the receiving operator characteristic (ROC) curve, which takes into account both accuracy and sensitivity.

	SVM	LR	kNN
STFT	0.59	0.71	0.62
DWT	0.64	0.61	0.58
DWT + STFT	0.67	0.68	-
STFT + PCA	0.64	**0.75**	-

Figure 7: Correctness scores of each feature extraction method with each classification model

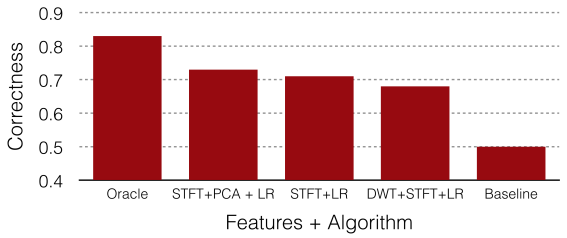

Figure 8: Correctness scores of the best feature extraction-classification model pairs

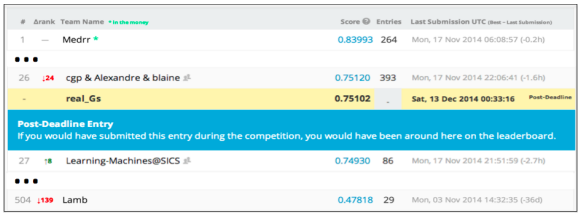

Figure 9: Our performance on Kaggle

6 Discussion

Short-time Fourier transform (STFT) feature extraction, principal component analysis (PCA) dimensionality reduction and normalization, and logistic regression (LR) model prediction provided the most accurate result.

STFT provided robust, domain-specific feature extraction because they provided frequency and phase information on every window of each 10-minute EEG clip. Taking the average of frequencies from 30- and 60-second window sizes provided an even more robust classification feature. The features were domain-specific because STFT is a method specifically utilized for time-series signal data.

PCA combined with STFT provided better results than STFT alone because of PCA's normalization step. Reducing dimension of the raw signals (performing PCA without STFT) reduced classification accuracy to below the baseline because it removed key latent elements of the data.

DWT, like PCA, removed key latent elements of the EEG data and thus did not perform as well as STFT.

We were surprised that LR performed better than SVM, but concluded it was because logistic regression made the fewest prior assumptions on the data set. In particular, logistic regression classifies with a probability based on a data point's distance from the regression line, whereas SVMs classify strictly and fail regularly when there is too much noise in the training data. To optimize SVM performance and prevent overfitting, we would like to tune our objective function to support more variance.

kNN was an effective baseline algorithm, but did not provide a nuanced enough definition of distance between data points to accurately classify test data.

Our classification results performed better than about 95% of competitors on Kaggle and would have been good enough to place us 27^{th} out of 504 entrants.

7 Future Work

First, we would like to improve our SVM model to perform better than logistic regression. The literature we have read shows that SVMs perform optimally, and we believe the Kaggle data set should not be an outlier.

In future work, we would like to attempt more complex classification algorithms such as neural networks and random forest walks. We attempted to apply neural networks for this project, however we experienced severe over-fitting and a classification of non-seizure for virtually all data points. We would like to tune neural networks and run them on faster processors to evaluate their potential with this data set.

The random forest model is a recently discovered, increasingly popular mode of classification and we would like to evaluate its performance in binary classification of EEG data.

References

[1] Epilepsy Facts (2014, December 9), *Citizens United for Research in Epilepsy* [Online], Available: http://www.cureepilepsy.org.

[2] Nine health wearables for your head (2013, August 5), *mobihealthnews* [Online], Available: http://mobihealthnews.com/24401/nine-health-wearables-for-your-head/

[3] American Epilepsy Society Seizure Prediction Challenge (2014, August 25), *Kaggle* [Online], Available: http://www.kaggle.com/c/seizure-prediction

[4] Mormann, Florian, Ralph G. Andrzejak, et al., "Seizure prediction: the long and winding road," in Brain, 2007, pp. 314-333.

[5] Binder, Devin K. and Sheryl R. Haut, "Toward new paradigms of seizure detection," Epilepsy & Behavior 26, 2013, pp. 247-252.

[6] Subasi, Abdulhamit and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," in Expert Systems with Applications, 2010, pp. 8659-8666.

[7] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H.U. Voss, A. Schulze-Bonhage, J. Timmer, "Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic," Physica D 194, 2004, 357–368.

[8] Paul, Manoranjan and Mohammed Zavid Parvez, "EEG Signal Classification using Frequency Band Analysis towards Epileptic Seizure Prediction," in IEEE International conference on Computer and Information Technology, Khulna, Bangladesh, 2013.

[9] Suleiman, Raouf Abdul-Bary and Toka Abdul-Hameed Fatehi, "Features Extraction Techniques of EEG Signal for BCI Applications," Faculty of Computer and Information Engineering Department College of Electronics Engineering, University of Mosul, Iraq, 2007.

[10] Ng, Andrew, Lecture Notes, 2014.