Classification of Seizure Prediction Data: Reduction of Large Feature Sets with Small, Unbalanced Data

Mark Connolly, Michael (Mac) Merrit, Sang Park, Hyunkoo Chung

Up to 1% of the population is affected by epilepsy, a condition characterized by spontaneous seizures. While medications are effective for some patients, closed-loop neurostimulation has shown promising results. One way to improve these systems is by implementing a classifier that can predict the onset of a seizure and respond by delivering stimulation prior to the onset of the seizure.

The overall goal of this project is to develop an algorithm that can classify intracranial recordings as interictal (low seizure risk) and preictal (high seizure risk). One of the limitations of high-frequency, multi-channel intracranial recording data is that it suffers from *too much* information. On a sufficiently long timescale the data is stationary and lends itself to the extraction of spectral features, while on a short timescale it is non-stationary suggesting the use of non-stationary signal measures such as intrinsic mode functions. Additionally, as each data set contains up to 16 channels of simultaneous recordings from an interconnected network, features describing the correlation between recordings could also discriminate between preictal and interictal segments. Given the significant number of features and the variation in what they describe the task of selecting and extracting the correct features all without overfitting a specific model becomes non-trivial.

In this paper we selected three types of data describing the stationary, non-stationary and correlation properties of a signal. The features of each type of data were then extracted using either principal component analysis (PCA) or multidimensional scaling (MDS). The features were then used to train three classifiers - linear support vector machine (SVM), radial basis function SVM, and K nearest neighbors, for a total of 6 models (2 feature extraction x 3 classifiers).

In addition to the two parameters intrinsic to each model, it was necessary to determine the ideal number of features of each type to use in the classifion model, leading to a total of 5 parameters to be optimized. The performance of a particular set of parameters was assessed using a stochastic cross-validation approach where a preictal and interictal segment were randomly selected as the validation set. To optimize the parameters a genetic global optimization algorithm was employed to rapidly explore the parameter space while avoiding convergence in local extrema. The combination of a stochastic validation and stochastic optimization algorithm helped ensure that we did not overfit any particular model, given the number of possible features with a relatively small number of data segments.

The trained models were then tested against a holdout set to determine their performance. It was found that the RSVM classifier using features extracted using PCA performed the best when trained on an individual subject. However, these results were sub-optimal, likely do to overfitting, and insufficient standardization of data

2. Methods

Data

Kaggle provides data from five canine and two patients. Among these, we used four canine data sets to maintain consistency data. Each set has 16 channel recordings at 400 Hz. A one hour data set was already segmented into 10 minute blocks labeled as interictal or preictal. Preictal data sets are sequentially ordered from 1 hour and 5 minutes prior to the seizure onset to 5 minutes prior to seizure onset. Interictal data sets are randomly selected from recordings at least 4 hours from any seizure event. The raw data are shown in the following figures.

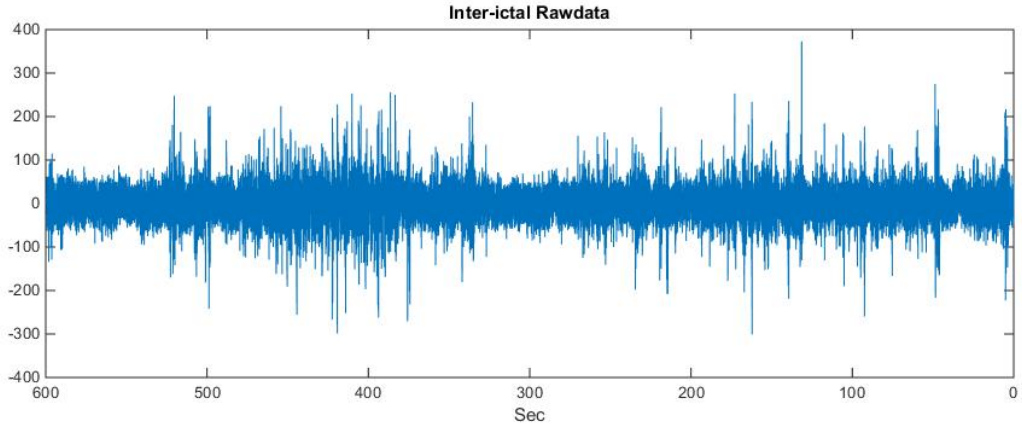

Fig 1. A 10 minute raw data of inter-ictal arbitrarily selected within the recording of at least 4 hours from any seizure events.

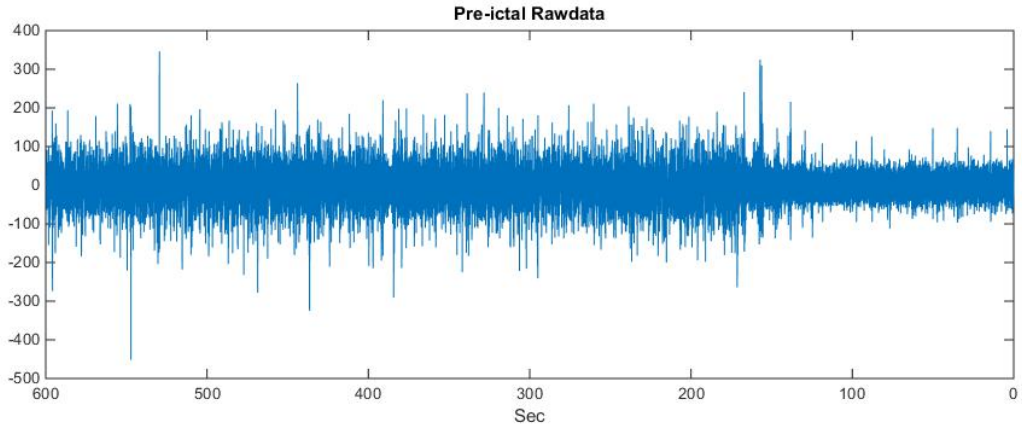

Fig 2. A 10 minute raw data of pre-ictal. The data is from 5 minute before to 15 minute before the seizure event onset.

Stationarity

In order to know how long of segments to use for stationary vs. non-stationary metrics, it was necessary to determine at what length the seizure data was stationary. The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [2] was used to investigate if the data is stationary. A few

segment length between 2-400s were chosen for KPSS test. The whole data was split for each segment and analyzed. The test rejected the null hypothesis that the seizure data segment is stationary if it shows more non-stationary characteristic. The rejected ratio was obtained through dividing the number of rejected segments by the number of total segments.

Table 1: KPSS Test

Segment Length(s)	Rejected Ratio
2	0.7298
5	0.7368
10	0.8066
20	0.8465
30	0.8283
40	0.8098
100	0.7123
400	0.4554

It was obvious that the segment shorter than 100s shows non-stationary characteristic. However, the longer segments were rather quasi-stationary. This result affected our choice of the length of segments.

Feature Selection

Stationary

Band limited power in each channel was estimated using stationary spectral estimation methods. The built in matlab spectrogram function was used for this extraction with a single window set to the full ten minute length of each preictal and interictal segment. The ten minute segment was multiplied by hamming window, and then the fourier transform of the autocorrelation function was used to estimate the power at various frequencies. All preictal segments were selected to include only those segments that were immediately preceding a seizure.

Non-Stationary

FFT assumes that signal is stationary, which is not true for shorter time windows. Thus, we also conducted the non-stationary analysis via using the Hilbert-Huang Transform (HHT). For the segment length, 20 second was used because it shows the most non-stationary characteristic. The last 20 second data of each seizure was used for the analysis.

The first step is to do empirical mode decomposition (EMD) to decompose the data into some intrinsic mode functions (IMFs). An IMF has a variable amplitude and frequency as a function of time so that HHT can deal with the non-stationary data. Next, we applied the HHT to each IMF. After that, the total powers within each IMF were calculated.

Correlation

A study [1] showed that correlation between the channels increases gradually before the seizure. Correlations between 16 channels were estimated in matlab. The correlation matrix was obtained; redundant and auto correlation coefficients were not considered in the feature extraction step. Thus, the cross correlations of 16 channels resulted in 120 correlation coefficients for each 10 minute segment. Analyzing the correlation may provide information to distinguish interictal from preictal data.

Feature Extraction

Principal Component Analysis

Principal component analysis was conducted on each of the data sets independently (ie stationary, nonstationary, and correlation data were not combined). The analysis was conducted by simply using matlab's built in PCA function which operates using singular value decomposition (SVD) to produce a low dimensional space where each dimension is the correlation between the data points is effectively zero for all dimensions

Multidimensional Scaling (MDS)

Multidimensional scaling was also used for feature extraction in order to compare to the PCA results. This procedure was conducted using the simple euclidian distance between each of the data points, and then matlab's mdscale algorithm was used to try to retain the similarities between points while mapping into a lower space.

Classification

Three different methods were trained and then used for all classifications. The basis of the first two methods were SVM, which attempt to find a maximum margin hyperplane to separate the data with an allowance for error that is controlled by a regularization parameter. The second classification method was based on K nearest neighbor which simply takes a weighted vote of the points in the training data with the smallest euclidian distance to the test point.

The seizure data set that was used for this analysis was extremely biased towards the interictal (negative) data. This imbalance creates a problem for most classification methods including both SVMs and KNN. In order to compensate in the SVM, we added an additional parameter alpha which compensates by increaseing the pentality associated with miscatagorizing positive data. The modified SVM equation is shown below, and as is shown, if alpha is equal to .5 then this system reduces to a classical SVM, but if alpha is very low, then the system will be biased in favor of the positive data and will thus compensate for the imbalance in the data see figure 1a.

$$min(\frac{1}{2}w \cdot w + C \cdot \alpha \sum_{i|y_i=+1}^{l} \xi_i + C(1-\alpha) \sum_{i|y_i=-1}^{l} \xi_i)$$

$$s.t. \quad y_i(w \cdot \Phi(x_i) + v) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \quad i = 1, \ldots, l$$

The two SVM methods that were used in this analysis were simple linear SVM and kernalized SVM with a gaussian radial basis function. The RBF is a kernel that allows for arbitrary decision boundaries that are ultimately based on the distance between the points.

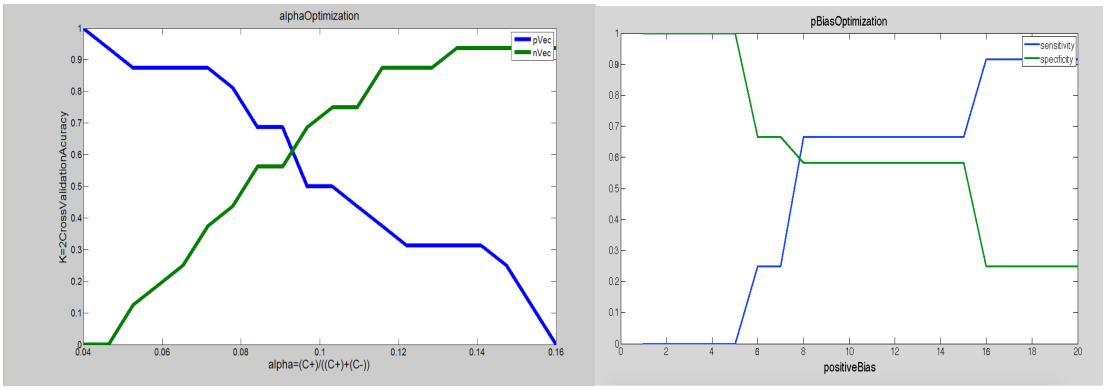

Figure 3: A trade off between sensitivity and specificity as a function of alpha and b as a function of positive bias.

The KNN algorithm was modified to compensate for the imbalance in a similar way to the SVM. Here, an extra parameter was added to the voting equation which multiplied all votes by a positive bias. Since the negative terms always vote for 0 by definition, the positive bias has no effect on their vote. On the opposite side, the positive votes are by definition equal to one and therefore they are scaled by the bias term (Figure 3b) .

These simple modifications to the SVM and KNN algorithms created classifiers that each had 2 parameters in need of tuning. The first parameter is roughly related to the degree of complexity to which the classifier is allowed to take (C for SVM and K for KNN). The second parameter is related the degree to which positive data is biased to compensate for the imbalance in the data set. because these two parameters depend on each other and then number of parameters, they were optimized together as described in the section below.

Optimization

Initially, the effect of each parameter was estimated by a simple parameter sweep. The contour plot below shows the relationship between number of features and positive bias in the PCA and RBF Kernel Method. Figures 4 and 5 are illustrations of three optimized parameters (alpha, number of components, and C value) for one of the models, the stationary features of one of the canine subject. The optimized parameters were used to estimate the minimum accuracy for

both linear and RBF kernel method where the yellow region illustrates the highest possible minimum accuracy. Note that the optimized parameter contour plots for K-nearest neighbors and MDS features are not shown in this paper. After initial analysis to determine the appropriate constraints on the three parameters, we moved into more sophisticated global optimization of all five parameters.

Linear Kernel

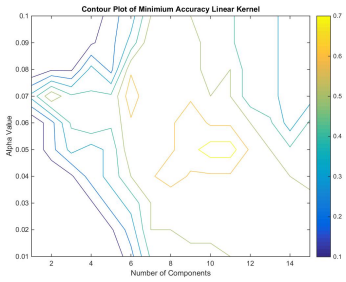

Figure 4a: Number of Components vs. Alpha value when C = 1.3

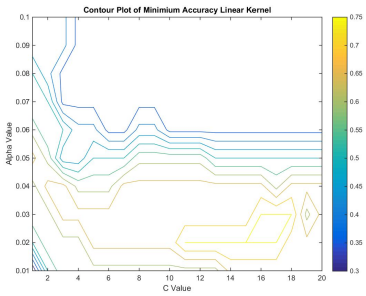

Figure 4b: C value vs. Alpha Value when Number of Components is 10

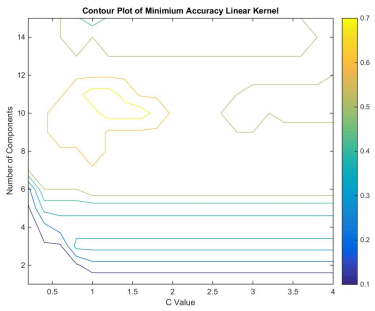

Figure 4c: C value vs. Number of Components when Alpha = 0.05

RBF Kernel

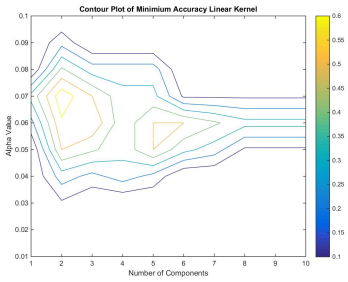

Figure 5a: Number of Components vs. Alpha value when C = 1.3

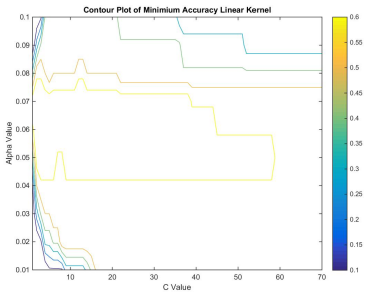

Figure 5b: C value vs. Alpha Value when Number of Components is 2

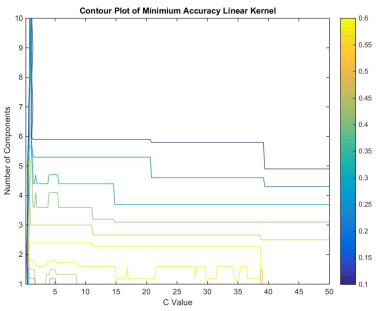

Figure 5c: C value vs. Number of Components when Alpha = 0.7

Each of the models had 5 parameters that could be varied and needed to be optimized including the intrinsic classifier parameters, and the number of features obtained from each of the feature classes.

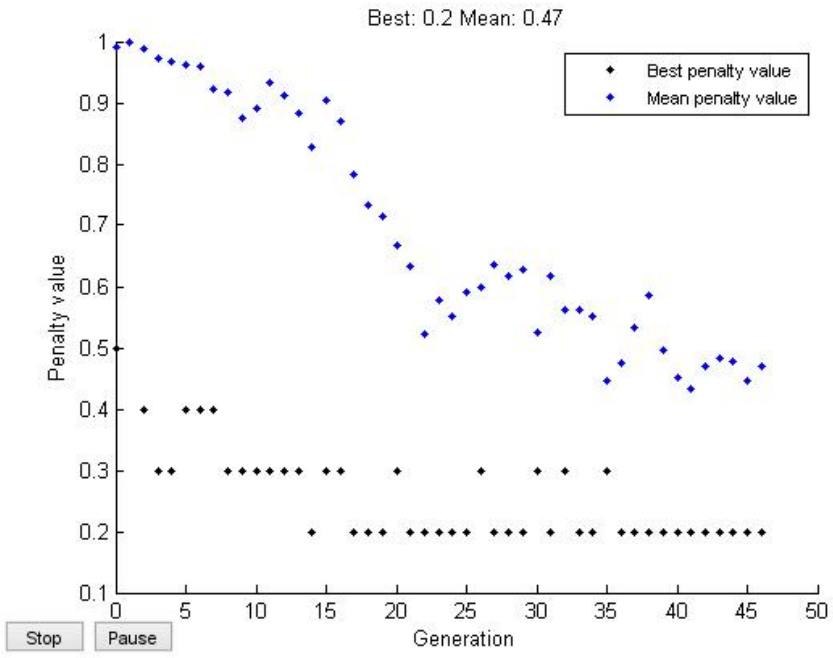

Figure 6: Result from the Genetic algorithm

The efficacy of a parameter set was determined by the stochastic cross-validation algorithm. Instead of a more typical k-fold approach, this algorithm randomly picked, without replacement, one preictal and one interictal segment as the validation set. Then the classifier was trained on the rest of the data and applied to the two validation points. This was repeated 10 times to determine the overall accuracy of the parameter set.

The parameters of the model were fit using a combination of stochastic cross-validation with a global optimization algorithm. Due to the high number of parameters, and the nonconvex cost-space, finding the local minimum using gradient descent would not have been feasible. Instead a genetic algorithm, where a large number of parameter sets were evaluated against a cost function (accuracy in cross-validation) and the most successful of the parameter sets were recombined to create a subsequent generation. This allowed us to explore the entire parameter space while avoiding local minima.

To test the performance of the classification models and rules, we reserved one of the canines with the largest data set to test against. Two different methods of testing were applied: global testing and individual testing.

In global testing the models were trained on the dataset containing three of the canines. This was used to create a rule based on one of the six models. This rule was then applied to the holdout set to determine the sensitivity and specificity of the classifier.

In the individual test, it was assumed that the specific model created using the global set would not perform well for a single, different set because the parameters of the model were in many ways functions of the amount of data and the balance between preictal and interictal segments. To counteract this, the model parameters were optimized on the data from one canine

with exactly same number of interictal and preictal segments as the holdout set, then applied to the holdout set. There were 134 interictal segments and 12 preictal segments.

3. Results

The results of the optimization classification method are listed in table 1. As a way of summary, the best method was RBF with PCA feature extraction method. The parameters that were used in the optimal method where C = 416.7, alpha = 0.06, and 1 stationary component and 3 correlation components

Table 2: Minimum accuracy results for all models

Training Data	Feature Selection	Linear SVM Min Accuracy	RBF SVM Min Accuracy	KNN Min Accuracy
Individual	PCA	.7	.75	.41
Individual	MDS	.4	.48	.31
Global	PCA	0	.06	0
Global	MDS	~failed to converge	0.05	0.04

4. Conclusion

This project involved a significant amount of time and energy in order to prove one of the fundamental lessons from day one of this course: there is a trade off between richness and the discrepancy between estimated and empirical risk. For our overall method selection, we attempted to optimize 5 parameters with one cross validation. We foresaw the potential issue of overfitting so we even added randomness to the cross validation set by not selecting all possible sets, but selecting a random subset of them. Our results display that the control for overfitting at the feature selection level was not sufficient. Our original results from the poster session showed a lower maximum error than many of the optimized methods. This situation could be remedied in future analysis by applying multiple cross validation layers, injecting more randomness to the validation, assuming more parameters so that less need to be selected, and or applying more stringent restrictions on the parameter space so that the parameters remain in a the pseudo convex region as displayed in the 2 D contour plot.

5. Citations

[1] K Schindler, H Leung, C Elger, K Lehnertz. Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG. Brain 2007

[2] Kwiatkowski, D.; Phillips, P. C. B.; Schmidt, P.; Shin, Y. (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root". *Journal of Econometrics* **54** (1–3): 159–178

[3] Ying Li, Yue-Loong Hsin, Wentai Liu. "Comparison Study of Seizure Detection Using Stationary and Nonstationary Methods"

[4] J. Martinerie, C. Adam, M. Le Van Quyen, M. Baulac, S. Clemenceau, B. Renault, and F. J. Varela, "Epileptic seizures can be anticipated by non-linear analysis.," Nat. Med., vol 4, no.10, pp. 11736, Oct. 1998

[5] Kaggle Dataset. American Epilepsy Society Seizure Prediction Challenge