Researcher: Dustin McAfee
Presentation Title: High Performance Clustering of Electroencephalogram (EEG) Data for Prediction of Seizure Events
Institution: University of Tennessee, Knoxville
Department: Department of Electrical Engineering and Computer Science

Abstract

Electroencephalography (EEG) is a noninvasive monitoring method that measures voltage fluctuations in the neurons of the brain due to ionic current. Certain events such as seizures tend to create abnormalities in EEG readings. The goal is to implement and run parallelized clustering algorithms for classification of seizure events from EEG readings. All programs are written in python 3.6 using the pyspark 2.2.0 library for parallelization. The EEG dataset consists of 11500 rows, each representing one second of EEG readings. Each one second observation has 178 voltage readings [1]. The data is first z-normalized and projected onto its first 40 principal components, which represent about 95.88% of the variance of the entire dataset. The projected data is then split into a training and validation dataset. The validation data represents about 23.5% (2700 observations) of the entire dataset.

Both K-Means and (Gaussian) Expectation Maximization (EM) are implemented with K = 2 and K = 3 clusters. Performance metrics such as confusion matrices, F1 scores, accuracy, precision, sensitivity, and specificity are calculated for each algorithm execution, and the implemented algorithms also give clustering evaluation metrics, such as the Dunn index for K-Means, and the Silhouette index for EM. Both algorithms pick the starting centroids via the K-Means++ method, which is an approximation algorithm for minimizing the intra-class variance [2].

Non-seizure activity is generally centered around 0 Volts on all attributes, while seizure activity generally surrounds the non-seizure clustering on the outside. This can be shown by projecting the data into its first 3 principal components, as in the following Figure (1).

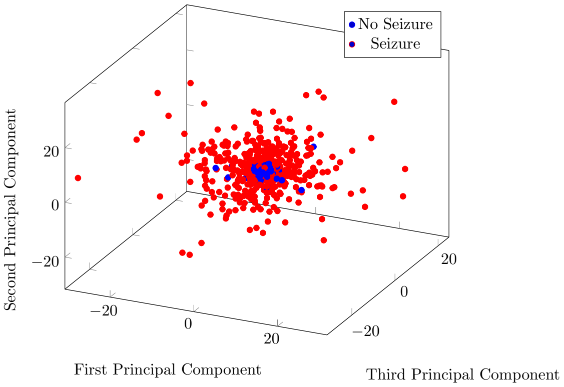

Figure 1: Standardized Testing Dataset Projected onto First 3 Principal Components

This clustering does not seem possible for the K-means algorithm, but was attempted anyway for contrast with the EM algorithm. K = 3 clusterings are attempted for both algorithms in hopes of classifying with high/medium/low risk categorizations, with the high risk category having a higher sensitivity rate than the K = 2 clusterings.

As expected, the K-means algorithm does not perform well, with less than 15% sensitivity on both K = 2 and K = 3 and for both the testing and validation datasets. EM with Gaussian clusters is implemented in an attempt to fit this data with higher performance. EM fits this clustering shape much better than K-Means, as expected, performing with much higher performance, overall. For K = 2, EM has 96.48% accuracy and 95.85% sensitivity for the training dataset, and 87.22% accuracy and 81.10% sensitivity for the validation dataset. The validation dataset projected onto the first three principal components and color coded from EM (K = 2) is shown in Figure 2, below. Note how these clusterings look very much like the clusterings from Figure 1.

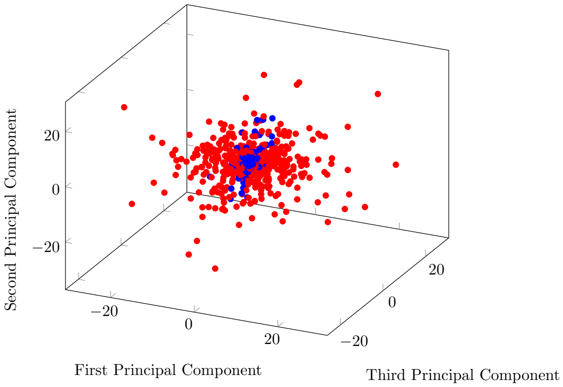

Figure 2: EM Clustering (K=2) of Standardized Testing Dataset Projected onto First 3 Principal Components

As expected, K = 3 yields a higher sensitivity rate for this dataset, however, the algorithm only classifies two groupings, as one of the clusters is deemed not likely on any observation. The accuracy and sensitivity for the training dataset is 67.32% and 100.0%, respectfully, and the accuracy and sensitivity for the validation dataset is 66.25% and 92.23%, respectfully.

The EM algorithm performs much better than K-means with this data, and though increasing K from 2 to 3 does in fact increase the sensitivity when categorizing seizures, it decreases accuracy, precision, and specificity at an even higher rate. Future work involves implementation of spectral clustering algorithms, which may fit these high dimensional clusters better, and a K-Folds cross-validation approach, which may help generalize better for fitting the testing data.

References

[1] Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE (2001) Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical

activity: Dependence on recording region and brain state, Phys. Rev. E, 64, 061907

[2] Arthur, David & Vassilvitskii, Sergei. (2007). K-Means++: The Advantages of Careful Seeding. Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms. 8. 1027-1035. 10.1145/1283383.1283494.