Seizure detection onset using Time, Frequency Correlation and Machine learning

Armand Hoxha
February 5th, 2018

Domain Background

Chronic epileptic seizure affects over two million Americans, and approximately 50 million worldwide. The need of a warning sign, to give patients enough time to stop activities that would otherwise be dangerous due to an epileptic attack is essential and can be provided with the present technology.

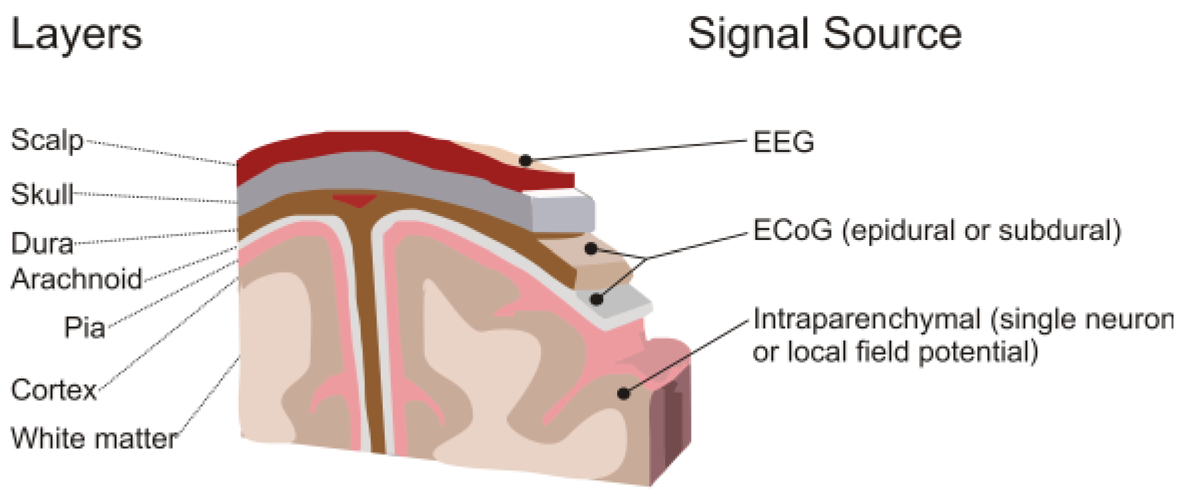
In this project, I created a classifier capable of predicting seizure development and occurrence. The classifier was trained using data provided by the Kaggle Competition UPenn and Mayo Clinic's Seizure Detection Challenge. The current project is a precursor step for successful intervention/ therapeutic device to prevent or warn patients of an oncoming seizure.

Problem Statement

Epilepsy is a common neurological disorder characterized by recurring seizures and abnormal brain activity in focal areas of the brain. Approximately 30% of patients suffering from focal seizure require surgical intervention. To determine where the seizures occur, surgeons use a combination of history, physical exam, as well as neuroimaging techniques such as EEG or intracranial EEG in cases where activity is hard to localize. Patient hospitalization can extend to weeks for enough seizures to be recorded using intracranial electrodes.

In cases of a brain lesion 80% of surgeries are successful in rendering patients seizure free, however in case of a lack of lesions, only 50% of the patients make it through to live the rest of their lives without lesions.

One of the reasons suspected for failure of surgical removal of the seizure piece of the brain is that epilepsy could be a network wide degenerative disease; the removal of an intricate network of brains would leave the patient with a diminished quality of life.


Layers
Signal Source
Scalp
Skull
Dura
Arachnoid
Pia
Cortex
White matter
EEG
ECoG (epidural or subdural)
Intraparenchymal (single neuron
or local field potential)

Recent advancements in the Brain to Machine Interface (BMI or BCI) have shown great success in the ability to read real time data, and apply machine learning algorithms to estimate the state of the underlying neural mechanisms of the brain.

The goal is to create an Ecog (intracranial encephalography) data classifier to predict whether a one second segment of data pertains to the first fifteen seconds of a seizure episode, after fifteen seconds within the seizure, or a non-seizure episode.

Datasets and Inputs

The competition data has 58,837 one second segments of data from 12 subjects (4 dogs, 8 human) who suffer from chronic seizures. Segments are denoted as ictal for seizure, interictal for non-seizure, and test for testing purpose. All data segments have the following fields:

Latency: how far into a seizure is the segment (as defined by experts clinicians)

__header__:platform info that saved data into segments

__globals__:empty

Channels:name of the channels by location

Freq:sampling frequency of the data

__version__: always 1.0

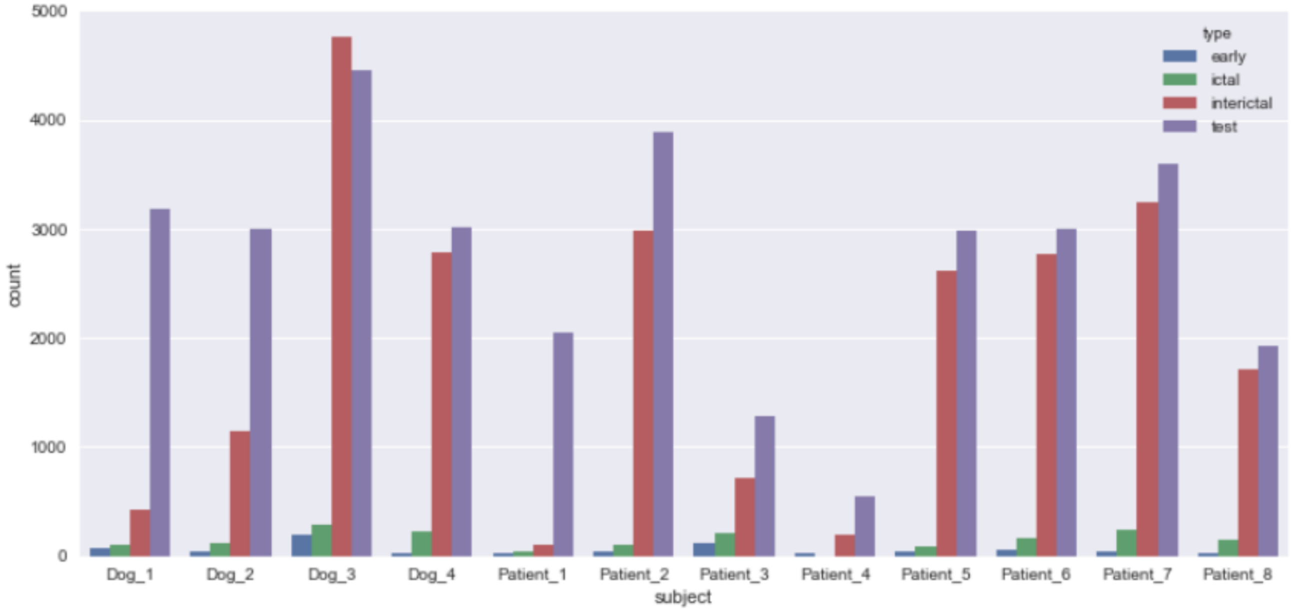
Data:all data in shape channels x sample

Most important field is Data, and the feature extraction is focused on this field only.

Most important field is Data, and the feature extraction is focused on this field only.
It must be noted that subjects have different sampling rates, as well as number of channels.

Subject	Sampling Frequency	Number of channels
Dog 1	399.6	16
Dog 2	399.6	16
Dog 3	399.6	16
Dog 4	399.6	16
Patient 1	500	68
Patient 2	5000	16
Patient 3	5000	55
Patient 4	5000	72
Patient 5	5000	64
Patient 6	5000	30
Patient 7	5000	36
Patient 8	5000	16

Glancing at the above table, clearly making a learner specific for each Subject is the best way approach the problem, as with different number of channels the number of features becomes hard to maintain to a single size, and little information is known about the condition of the seizure and their onset location.



Looking at the above countplot, majority of the data is interictal (non- seizure), and data distribution varies greatly; Patient_3 and Patient_4 have the least amount of data, with Patient_4 having no ictal (seizure) data.

Solution Statement

The objective of this project is to clearly identify seizure types of data from non seizure data from the provided time series in the competition. Schindler et al. Tzallas et al, have shown good results in seizure detection using time – frequency analysis of the data.

A variety of different learning models will be attempted, thus it is best to use the scikit-learn python toolbox, to allow changing and application of multiple learners with simple change modifications. Numpy and Scipy python toolboxes will be used to pre process the data, and extract necessary features in the time and frequency domain.

The results of the prediction will be measured using the AUC curve from a part of the learning set, after the final parameters for the learner have been selected through a GridSearch method, the learner will predict all test data, and the results will be submitted to the kaggle competition submissions.

Evaluation Metrics

The competition metric is the mean area under the ROC curve (AUC) of two predictions. First prediction must predict the probability that a clip is a seizure. Second prediction is the probability that a clip is within the first fifteen seconds of a seizure. Early clips are double counted because early detection is critical for intervention purposes.

Score=1/2(AUCseizure+AUCearly)

Project Design

Considering the size of the data (~50GB), a method to reduce the size is necessary to test different ideas effectively. A module called pickle, can be used to save extracted features from data, thus instead of loading 50GB of data to make predictions, one would have to load a much smaller amount of data.

Features to be extracted will be considered using the information provided by Schindler et al 2007, and Muller 2005, indicates that an important feature to be examined could be the correlation of EEG and Ecog channels in both time and the frequency domain. Correlation measures the mutual relationship or connection between two things, in this case the authors suggest to measure the correlation between the data channels. The two types of correlation discussed are frequency correlation and time correlation. Correlation is a measure between 1 (a=b) meaning completely identical relationship, and -1 meaning that there is an antagonistic behavior (b=-a). In time correlation we measure whether the channels are receiving similar data in the time series, where as frequency correlation measures “synchronous” brain activity in terms of frequency bands and their power. Thus an outline of the steps to follow would be:

1- Numpy will be used to normalize data across channels

2- Scipy to compute the Fourier Transform of the time series data

3- Use Numpy to compute the correlation matrix from he Fourier Transform data

4- Compute the Time series channel correlation

5- obtain only top triangle of the correlation matrix in a single dimension array NOTE about correlation matrices: A correlation matrix is mirrored across the diagonal, thus only half is needed)

After the features are extracted, they will be saved using pickle, to enhance the speed of loading the features for testing different learners, and debugging.

Benchmark Model

To objectively compare the progress of predicting models, the benchmark model learner will be a K-Neighbors classifier from sklearn:

sklearn.neighbors.KNeighborsClassifier(n_neighbors=10, weights='uniform', algorithm='auto', leaf_size=30)

According to Muller et al (2005) once onset of a seizure begins, channels start acting more alike, and during a complete seizure almost all channels act alike. Considering that our features measure the relationship between channels, it would be reasonable to assume that KNN would perform well.

References:

Link to competition: https://www.kaggle.com/c/seizure-detection

1. Rizzo G. (2013) Design and Evaluation of an Affective BCI-Based Adaptive User Application: A Preliminary Case Study. In: Carberry S., Weibelzahl S., Micarelli A., Semeraro G. (eds) User Modeling, Adaptation, and Personalization. UMAP 2013. Lecture Notes in Computer Science, vol 7899. Springer, Berlin, Heidelberg

2. Kaspar Schindler, Howan Leung, Christian E. Elger, Klaus Lehnertz; Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG, Brain, Volume 130, Issue 1, 1 January 2007, Pages 65–77, https://doi.org/10.1093/brain/awl304

3. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

4. Stéfan van der Walt, S. Chris Colbert and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30 (2011), DOI:10.1109/MCSE.2011.37

5. John D. Hunter. Matplotlib: A 2D Graphics Environment, Computing in Science &

Engineering, 9, 90-95 (2007), DOI:10.1109/MCSE.2007.55

6. Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

7.A. T. Tzallas, M. G. Tsipouras and D. I. Fotiadis, "Epileptic Seizure Detection in EEGs Using Time–Frequency Analysis," in *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 703-710, Sept. 2009.

8. Markus Müller, Gerold Baier, Andreas Galka, Ulrich Stephani, and Hiltrud Muhle, " Detection and characterization of changes of the correlation structure in multivariate time series", Phys. Rev. E **71**, 046116 – Published 14 April 2005