Seizure Detection and Prediction

Mingzhe Hu, Min Huang, Chenbin Huang

September 2021

1 Survey of Background Literature

1.1 Background and Importance

Epilepsy is a serious neurological disorder with unique characteristics, tending of recurrent seizures[1]. This is a worldwide disease which is not only happened on human but also some mammals. Epilepsy affects approximately 50 million people all over the world. What's more, 100 million people will be affected at least once in their lifetime[2][3]. The patients in this disease will experience a sudden breakdown or unusual activity in the brain which will last from seconds to a few minutes. This is dangerous because this situation will accompany with fractures, burns, and even death[4].

To detect seizures, the electroencephalogram (EEG) is used to record the overall electrical activity of the brain in a waveform. These signals store important information about seizures[5]. However, it is still difficult to obtain apparent difference in EEG activity between epileptic and non-epileptic seizures, so it is hard to predict when the patients will be attacked[6][7]. Therefore, it is vital to apply machine learning in catching the features of epileptic seizure, which will help the prevention of epileptic seizure.

1.2 Contribution

It is important to predict seizure precisely with the increase of epilepsy patients. A major challenge is to select suitable model for features and classifiers. In this project, we are trying to use machine learning to analyze EEG signal to predict epileptic seizure. We plan to compare different models and evaluate their sensibility and accuracy of prediction to find the most suitable model. This project will help with implement of prediction of seizure.

2 Proposed Methodology

2.1 Datasets

2.1.1 Seizure Prediction Data Set

The EEG data for the prediction of seizure is available through kaggle/seizure-prediction. The data were captured from five dogs and two patients. However, only the data of three dogs and one patient are available currently.

Data of each subject are composed of three different types of data which are *interictal*, *preictal*, and *test* data. Interictal data are the baseline data as the non-seizure training data; preictal data are the data for the preictal phase as the positive seizure prediction training data; test data are the data for the testing and scoring in the original competition; the labels for the test data are unknown.

Data are segmented every 10 minutes, and each segment has five fields which are:

1. data: the $channel \times signal$ matrix, the core data, usually is $16 \times (data\_length\_sec \times sampling\_frequency)$.
2. data_length_sec: the time span of the data, usually is 600 seconds.
3. sampling_frequency: the frequency of the signal; number of data per channel per second, usually is 399.6098 which is approximately 400 for dogs, but may vary in patients' data.
4. channels: the names of 16 channels.
5. sequence: the number of the segment in one hour, range from 1 to 6.

Among data segments, both interictal and preictal data have time span of 60 minutes and are segmented into 6 sub-segments of 10 minutes each. The order of sub-segments are recorded in the *sequence* field of the data.

2.1.2 Seizure Detection Data Set

The EEG data for the prediction of seizure is available through kaggle/seizure-detection. The data are captured from 4 dogs and 8 patients.

Data of each subject are composed of three different types of data which are *interictal*, *ictal*, and *test* data. Interictal data are the baseline data as the non-seizure training data; Ictal data are the data for the ictal phase as the positive seizure detection training data; test data are the data for the testing and scoring in the original competition; the labels for the test data are unknown.

Data are segmented every 1 minute, and each segment has three fields which are:

1. data: the $channel \times signal$ matrix, the core data, usually is $16 \times (data\_length\_sec \times sampling\_frequency)$.

	Interictal	Preictal	Test
Dog2	500	42	1000
Dog3	1440	72	907
Dog4	804	97	990
Patient1	50	18	195

Table 1: File Counts of Available Subjects in Seizure Prediction Data Set. Each file is a sub-segment of 600 seconds. Some file numbers are not divisible by 6, indicating the existence of missing or extra data. Also, the number of interictal and preictal data are strongly unbalanced.

2. freq: the frequency of the signal; number of data per channel per second, usually is 399.6098 which is approximately 400 for dogs, but may vary in patients' data.

3. channels: the names of 16 channels.

4. latency: (ictal data only) the time point of first data point for the segment.

All training data are sequentially arranged, but interictal data has no latency information while ictal data has. Test data are randomly sampled without order.

2.2 Data Augmentation

The data provided are unbalanced as is shown in table 1 for prediction data and table 2 for detection data. In the data of each subject, the number of preictal/ictal segments is much smaller than the number of interictal segments which may affect the training accuracy.

Furthermore, for prediction data, since each segment of one hour is sliced into six sub-segments of 10 minutes, the number of files are expected to be the multiple of 6. However, the number of files are not necessarily dividable by 6, which indicates that data may not be reassembled into one-hour-long segments. The feature extraction is one major problem to be solved in future work.

For detection data, the segments are of one minute which vary from the length of prediction data, which may lead to issues in unifying the feature extraction methods with prediction data.

2.3 Model Selection

In this project, we will attempt several different classification models and compare their performance. The criteria for choosing these models is that we must ensure these models are:

1. These models are mentioned or will be mentioned in the CS534 course. So we can apply and practice what we have learned in this class.

2. The classification must be proved effective on general classification problems.

Sample	Interictal	Ictal	Test	Total
Dog_1	418	178	3181	3777
Dog_2	1148	172	2997	4317
Dog_3	4760	480	4450	9690
Dog_4	2790	257	3013	6060
Patient_1	104	70	2050	2224
Patient_2	2990	151	3894	7035
Patient_3	714	327	1281	2322
Patient_4	190	20	543	753
Patient_5	2610	135	2986	5731
Patient_6	2772	225	2997	5994
Patient_7	3239	282	3601	7122
Patient_8	1710	180	1922	3812

Table 2: File Counts of Available Subjects in Seizure Detection Data Set. Each file is a sub-segment of 600 seconds. Some file numbers are not divisible by 6, indicating the existence of missing or extra data. Also, the number of interictal and preictal data are strongly unbalanced.

3. One advanced model beyond the content of this course will be implemented and compared with other models.

We will implement logistic regression, SVM (Support Vector Machine), Decision Tree, and an advanced LSTM (Long Short Term Memory) model.

2.3.1 Logistic Regression

Logistic regression is one of the most basic binary classification models. It is simple to implement and widely used. We could also observe the probability score easily. However, when the feature space is large, the performance of logistic regression will not be good, which is prone to overfit. Feature conversion is required for nonlinear features. We will implement logistic regression as the most basic model in our project.

2.3.2 SVM

SVM is an excellent algorithm for classification problems. Before the wide deployment of ensemble learning models and neural networks, SVM basically occupied the dominant position of the classification models. SVM is very effective in solving the classification problem with high-dimensional features. It still has a very good performance when the feature dimension is greater than the number of features. Moreover, there are a large number of kernel functions that can be used, which could be very flexible in solving various nonlinear classification problems. However, there is no universal standard for the choice of the kernel functions, and SVM is sensitive to missing data.

2.3.3 Decision Tree

The decision tree model is a tree structure composed of nodes and directed edges. There are two types of nodes: internal nodes and leaf nodes. Internal nodes represent a feature or attribute. The leaf node represents a class. The amount of calculation required by the decision tree is relatively small, and it is easy to transform into classification rules. No domain knowledge or parameter assumption is required to implement the decision trees. However, decision trees tend to overfit and ignore the correlation between attributes.

2.3.4 LSTM

The long-short-term neural network is a special kind of recurrent neural network. The so-called recurrent neural network is the prediction that the network can solve time series problems. As a special kind of recurrent neural network, LSTM can solve long-term time series problems. For our seizure detection and prediction tasks, the data we have are time series, and there might be important information in our data sequence. We presume that the LSTM would achieve the best performance among all the models we have chosen.

2.3.5 Evaluation Metrics

We will use the following metrics to assess the performance of our models: Type-1 Error, Type-2 Error, Specificity, Sensitivity, Precision, Accuracy, F1 score, ROC AUC score. We will also generate the confusion matrix, ROC curve, and precision-recall curve to evaluate our models visually.

3 Key Aims

3.1 Problem formulation

1. find proper windows for the detection and prediction data after downloading, cleaning the data.
2. balance the positive and negative data for training.
3. implement the models, and find proper parameters.
4. compare the models' performance, analyze the underlying reasons.

4 Research Plan/Schedule

Our research plan is shown as table 3.

Timeline	Tasks	Person in charge
Week1	Import Data & Data Visualization	MZH, MH, CBH
Week2	Data Cleaning & Preprocessing	MZH, CBH
Week3	Model Training & Evaluation	CBH, MH
Week4	Model Training & Evaluation	MZH, CBH
Week5	Performance visualization & Comparison	MZH, MH, CBH
Week6	Preparing report & slides	MZH, MH, CBH

Table 3: Project Timeline & Assignments: The first week starts at October 6, and the last weeks ends at December 1. The plan might be changed in the future as needed

5 Resources

`https://www.kaggle.com/c/seizure-detection`
`https://www.kaggle.com/c/seizure-prediction`

References

[1] World Health Organization. *Neurological disorders: public health challenges*. World Health Organization, 2006.

[2] Gonzalo Alarcón and Antonio Valentín. *Introduction to epilepsy*. Cambridge University Press, 2012.

[3] Mohammad Khubeb Siddiqui, Md Zahidul Islam, and Muhammad Ashad Kabir. A novel quick seizure detection and localization through brain data mining on ecog dataset. *Neural Computing and Applications*, 31(9):5595–5608, 2019.

[4] Jean A Hannah and Martin J Brodie. Epilepsy and learning disabilities—a challenge for the next millennium? *Seizure*, 7(1):3–13, 1998.

[5] Horia-Nicolai L Teodorescu, Abraham Kandel, and Lakhmi C Jain. *Fuzzy and neuro-fuzzy systems in medicine*, volume 2. CRC Press, 1998.

[6] Abdulhamit Subasi, Jasmin Kevric, and M Abdullah Canbaz. Epileptic seizure detection using hybrid machine learning methods. *Neural Computing and Applications*, 31(1):317–325, 2019.

[7] Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Xiaodi Huang, and Nasir Hussain. A review of epileptic seizure detection using machine learning classifiers. *Brain informatics*, 7:1–18, 2020.