Final Report
of
EEG Brain Wave for Confusion

by

Shiyi Tan
Min Chen
Thanawut Ananpiriyakul

7 December 2016
University of San Francisco

Overview

Understanding is the key factor in successful education. Educators have been trying so hard to find ways to measure understanding. Exam is one of them and it has been used for decades. Now, in this era of technology, it might exist a better way to measure understanding.

There is no traditional way to measure understanding in real-time, however, it exists like in a sci-fi movie. The answer is simple, we just need to crack the brain since it has all information we want. This approach is very easy to understand but it is very hard to implement.

When the brain is active, brain activities can be described by brain wave. It is electrical signals that can be measured by electrode. The problem is that if we want a clear signal, we must place electrodes very close to the surface of the brain. Yes, I mean that we need to open the skull and place the electrodes inside. Obviously, this is not practical.

Electroencephalogram [1] (EEG) represents brain activity. In current state, most of the devices will be placed at the skin surface of the head. The number of electrode varies, which results in number of channels of the brain that we can capture. Brain wave captured from outside of the skull is very weak and noisy. We cannot do better than this unless we drill the holes and put electrodes inside of the skull.

To sum up, we want to detect confusion based on EEG data. In our case, the confusion is 0 or 1. So, this is binary classification problem.

Tools and Frameworks

Rapidminer is used for exploratory analysis. Because it is powerful and easy to use; It needs just drag and drop with a little coding. It could help us visualize the dataset and get the models' accuracies as the thresholds before we develop our own models.

R programming language is used to develop our own models. It is powerful and flexible; It needs a lot of coding and takes time to develop.

Experiment

According to the dataset provider [2], Haohan Wang, the experiment was conducted 3 years ago. The EEG data was recorded while 10 college students were watching 10 MOOC video clips. The length of each video clip is about 2 minutes, which was chopped from the middle of a complete topic. A single-channel wireless headset was used to record EEG of students.

Each video clip was labelled by professionals to see whether the subject is expected to be confused. Half of them were labelled as “confused”. Moreover, they also asked participants to label confusion based on their own perspective.

The table below show a summary of this experiment.

Year of Experiment	2012-2013
Device	Wireless Single-Channel MindSet
Number of Electrodes	1
Electrode Placement	Frontal Lobe
Number of Subjects	10
Subject Occupation	College Student
Number of Video	10
Length of Video	1-2 minutes

Table 1. Summary of the Experiment

Dataset

We use the dataset from [3]. It contains 3 categories of data: lecture videos used in the experiment, demographic of participants, and EEG data recorded while participants were watching lecture videos.

For EEG data, there are 12811 data points. Each data point represents 0.5 seconds. The target is confusion in 0 and 1 format. There are two columns of targets: the first one was labelled by professional (PreDefinedConfusion), and the second one was labelled by participants depending on how they did feel during experiments (SelfDefinedConfusion).

Demographic data contain age, ethnicity and gender. There are 10 participants in this experiment.

EEG data was recorded only 1 channel at the frontal lobe. The data was separated into delta, theta, alpha, beta, and gamma based on spectrum frequency. The dataset also includes attention and meditation. Of course, subject id, video id, and confusion labels are included in the dataset.

The dataset provider does not provide sample rate of the device used in the experiment. However, it should be at least 100 Hz since the Gamma is included in the list of features. No further explanation of each feature was provided by the dataset provider.

Our assumption is that, the raw EEG were recorded at over 100 Hz. Then, power spectrum analysis algorithm was applied. One of the most popular power spectrum analysis algorithm is Fourier Transform. We have no idea how attention and meditation were computed.

The table below shows a summary of the EEG dataset.

Number of Data Points	12811
Pre-Aggregated	Yes
Length of Aggregation	0.5 sec / data point
Number of Features	13
Number of EEG Features	11
Number of Targets	2

Table 2. Summary of the EEG dataset.

Problem Definition

The objective of this work is to predict confusion of videos based on EEG data. A video could be classified as “normal” or “confusing”. So, the target could be "0" or "1" which means "normal" and "confusing" respectively.

The features are EEG data which can be broken down into many channels. Since the target is included in the dataset, we are doing supervised learning. And since we only have two labels: “normal” and “confusing” so it is binary classification.

This dataset is special because it has 2 targets: self-defined confusion and pre-defined confusion. We will only focus on self-defined confusion which evaluated by subjects.

Data Visualization

Goal

Cleaning and plotting EEG signals

Reference

Eryk Walczak: Cleaning and plotting EEG signals [4]

Library

Dplyr, ggplot2

Overview Plotting

```
# add index/sample column
    eeg$Sample = ave( 1:nrow(eeg), eeg$SubjectID, factor(eeg$VideoID), FUN=function(x) 1:length(x))

# plot all subjects
p1 = ggplot(eeg, aes(x=Sample, y=Raw, color=as.factor(SubjectID)))
p1 + geom_line()
```

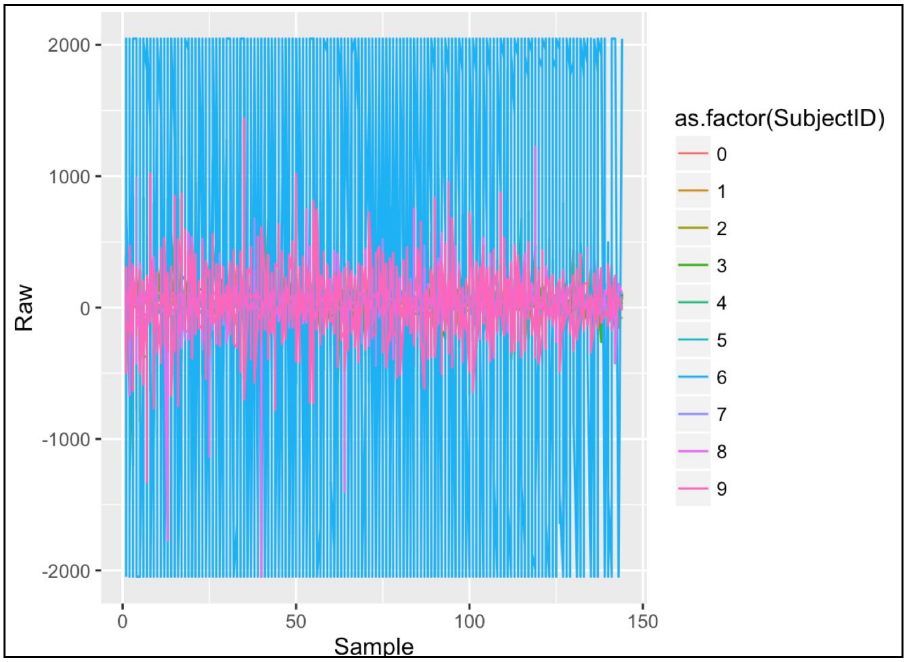


Figure 1. Plot of all subjects, Raw

```
# plot all videos
p2 = ggplot(eeg, aes(x=Sample, y=Raw, color=as.factor(VideoID)))
p2 + geom_line()
```

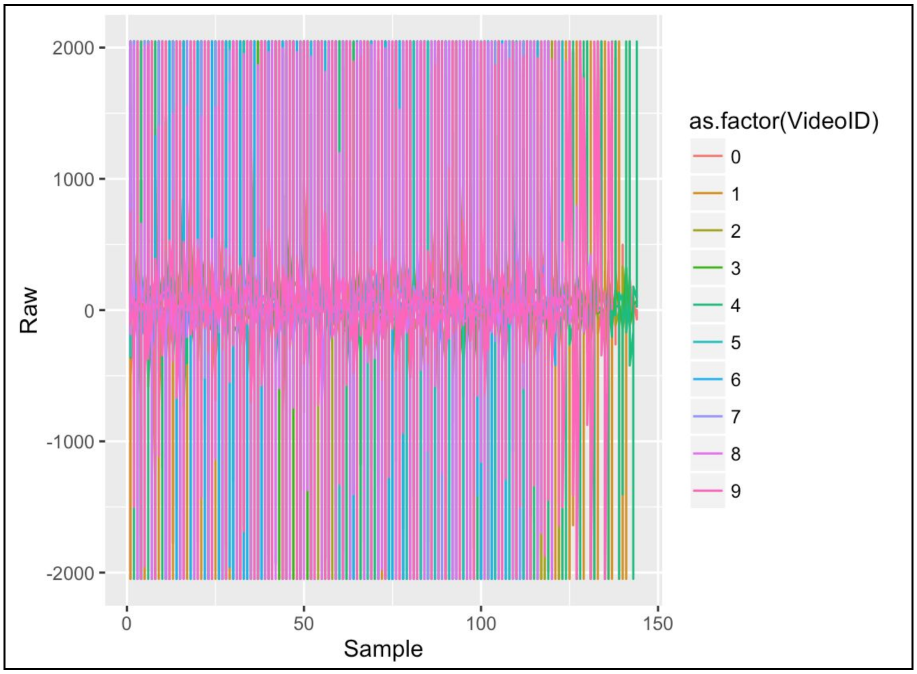


Figure 2. Plot of all videos, Raw

This creates pretty rainbow which shows that data is heavily contaminated.

Remove Outliers

In order to find out which participant/condition had a messy data, I plot every single participant and condition (VideoID).

```
par(mfrow=c(2,2))
for (i in unique(eeg$SubjectID)) {
        for (j in unique(eeg$VideoID)) {
                subject_df = subset(eeg, SubjectID == i & VideoID == j)
                print(paste("Participant", i, "-", "Video", j))
        plot(subject_df$Raw, type = 'l', ylim = c(-200, 200), xlab = "Sample", ylab = "mV")
        }
}
```

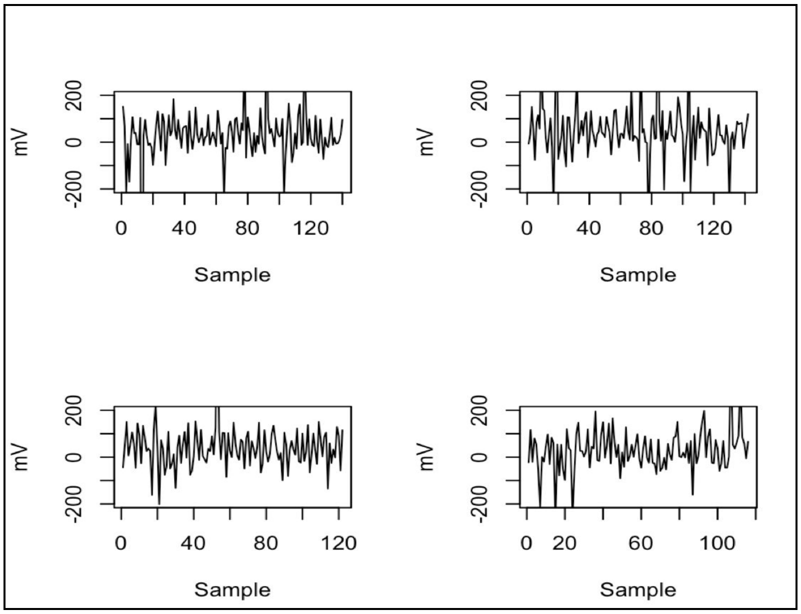


Figure 3. Sample plot of normal subjects, mV

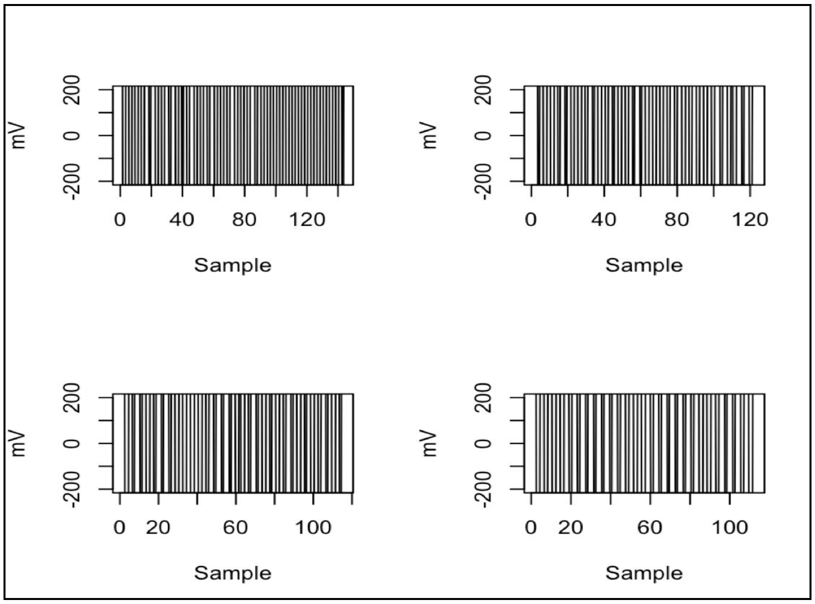


Figure 4. Sample plot of subject 6, video 3-6, mV

# That clearly shows that participant 6 has to be removed:

```
eeg_clean = subset(eeg, SubjectID != 6)
```

# While it is weird that after removing participant 6, the accuracy of logistic regression, support vector machine dropped below 50%.

More visualizations that we plot can be found in [5].

Data Preprocessing

First of all, we only focus on EEG dataset. We ignore demographic dataset since we want to know whether it is possible to predict confusion based solely on EEG data or not. There are 4 things we did about data preprocessing including Feature Selection, Data Normalization, Feature Generation, and Data Aggregation.

Feature Selection

The original features are channels of brain waves: Attention, Meditation, Raw, Delta, Theta, Alpha1, Alpha2, Beta1, Beta2, Gamma1, Gamma2, SubjectID, VideoID.

Only SubjectID, VideoID got removed. We did try best subset selection but there was no significant improvement from removing more features.

Data Normalization

The features are not in the same range. They varies from 10 to over 100000. So we take traditional approach which is normalizing all of them into range of [0,1]. Figure 5 shows the plot of Delta before normalizing and figure 6 shows the result after normalization. As you can see, the shape of the plot is still the same but the range has been changed to [0,1].

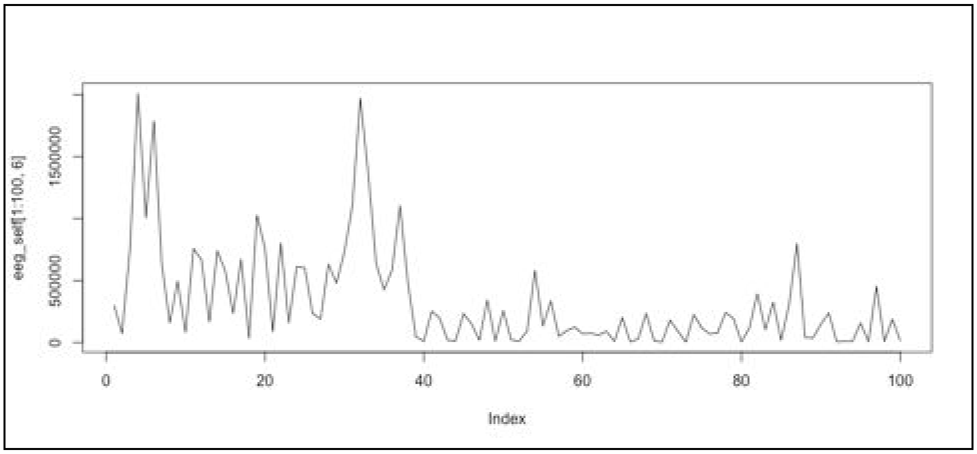


Figure 5. Plot of Delta before normalization.

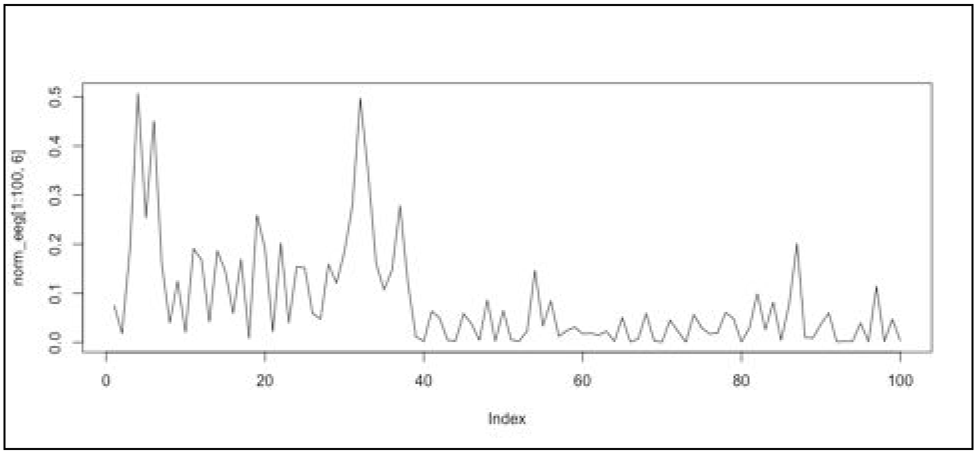


Figure 6. Plot of Delta after normalization.

Feature Generation

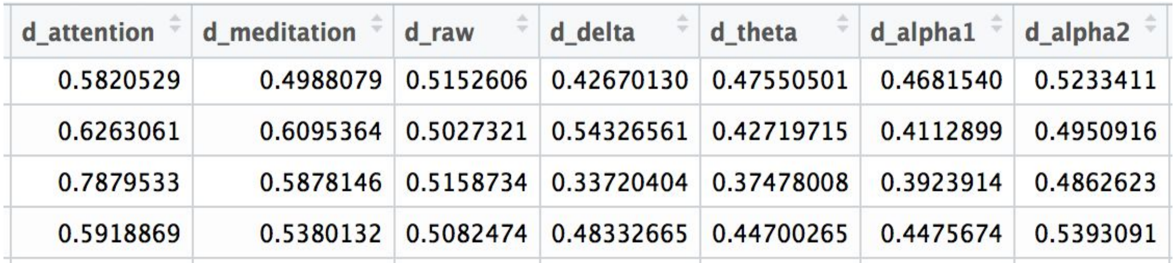
As we mentioned that we do not have many features since the EEG data were recorded with single-channel device. So, generating more useful features could help improve the result.

Since our data are time series, changes could be a nice feature. So we decided to add new features which are differences of each feature between each time frame. Table 3 shows an example of calculation of differences. For example, the difference of Attention of the first and the second row is -0.07.

Attention	ΔAttention	Meditation	ΔMeditation	Raw	ΔRaw
0.68		0.5		0.8	
0.61	-0.07	0.52	+0.02	0.67	-0.13
0.63	+0.02	0.51	-0.01	0.72	+0.05

Table 3. Example of difference calculation.

Table 4 shows some real values of new features that we have generated. As a result, we ended up having double numbers of features.



d_attention	d_meditation	d_raw	d_delta	d_theta	d_alpha1	d_alpha2
0.5820529	0.4988079	0.5152606	0.42670130	0.47550501	0.4681540	0.5233411
0.6263061	0.6095364	0.5027321	0.54326561	0.42719715	0.4112899	0.4950916
0.7879533	0.5878146	0.5158734	0.33720404	0.37478008	0.3923914	0.4862623
0.5918869	0.5380132	0.5082474	0.48332665	0.44700265	0.4475674	0.5393091

Table 4. Real values of features that have been generated.

Data Aggregation

Basically, data aggregation is a type of data mining process where data is combined and gathered in a summarised way. For this project, data was recorded every 0.5 seconds, which causes distribution very ugly, shown as the Delta distribution below.

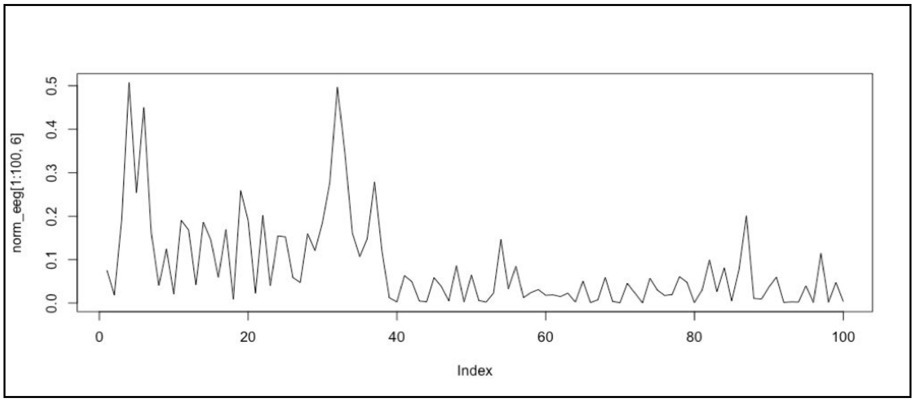


Figure 7. Delta distribution before aggregation

In order to reduce noise in the dataset, we aggregated data from 0.5 second to 5 seconds, then we get 2400 data points in total which is much less than before. But the data distribution goes more smoothly.

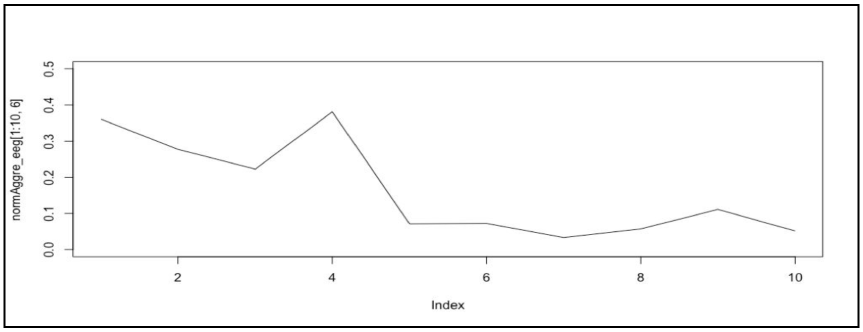


Figure 8. Delta distribution after aggregation

Evaluation Setup

We use two evaluation approaches which are 5-fold cross validation and leave-one-out cross validation. In the case of leave-one-out cross validation, we do not leave one data point out but we leave one subject-watch-video out. Since the final result of this work is to predict the confusion of video so instead of evaluating the result of each data point, we should evaluate it in each video of each subject. We will focus on the later approach.

We have 10 subjects with 10 video so we have 100 cases in total.

Model

We are doing binary classification so we tried the following 9 classifiers which fit the nature of our problem:

1) K-Nearest Neighbors
2) Logistic Regression
3) Linear Discriminant Analysis
4) Quadratic Discriminant Analysis
5) Decision Tree
6) Random Forests
7) Boosting
8) Support Vector Machines
9) Neural Network

We expected SVM to perform the best since all of features are numeric and SVM is very good in binary classification.

Experiment Result

In order to improve the prediction accuracy, we modified the original dataset in 5 ways:

a) Data aggregation

b) Data normalization

c) Aggregation + normalization

d) Aggregation + difference

e) Aggregation + normalization + difference

Below is the accuracy plot we got based on 6 types of datasets:

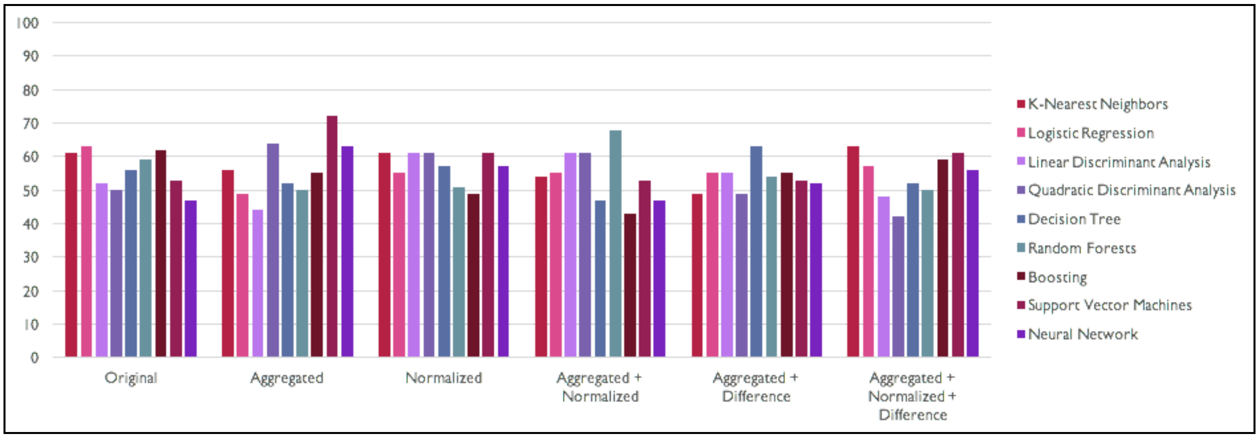


Figure 10. Accuracy of leave-one-out cross validation.

According to the plot, we got the highest accuracy at 72% on aggregated dataset, achieved by Support Vector Machines. Therefore, it seems that data aggregation helps improve the model accuracy the most.

Challenges

The first challenge is that the data is only one channel. Because 1-channel wireless headset was used in the experiment. It means that there are less features in the dataset that we could use to predict the confusion.

We solved this challenge by using the “delta” to get additional features, as the report says in Dataset Improvement part.

The second challenge is that the data itself is very noisy. Because during the experiment, the electrode is placed at students’ skins.

We solved this challenge by doing data aggregation, which made the data distribution more smooth and removed the noise of the data.

Timeline

After meeting with the instructor last week, we decided to update the timeline. Here is the new version.

Task	Due Date	Status
Team Formation	26 SEP	Complete
Project & Dataset Selection	30 SEP	Complete
Project Proposal	3 OCT	Complete
Data Normalization	10 OCT	Complete
Feature Selection - Correlation Analysis - All Possible Subset	17 OCT	Complete
Applying Model - Logistic Regression - Support Vector Machine	24 OCT	Complete
Data Visualization - Plot - Data Statistics	31 OCT	Complete
Data Transformation - Data Aggregation	3 NOV	Complete
Project Progress Report	3 NOV	Complete
Feature Generation	10 NOV	Complete
Reapplying 9 Models on 6 Different Datasets	17 NOV	Complete
Result Analysis	24 NOV	Complete
Wrapping Up	1 DEC	Complete
Project Presentation	5 DEC	Complete
Project Final Report and Code	7 DEC	Complete

Reference

1. Electroencephalogram: http://emedicine.medscape.com/article/1138154-overview
2. Wang, H., Li, Y., Hu, X., Yang, Y., Meng, Z., & Chang, K. M. (2013, June). Using EEG to Improve Massive Open Online Courses Feedback Interaction. In AIED Workshops.
3. Dataset: https://www.kaggle.com/erykwalczak/d/wanghaohan/eeg-brain-wave-for-confusion/cleaning-and-plotting-eeg-signals
4. Eryk Walczak: Cleaning and plotting EEG signals https://www.kaggle.com/erykwalczak/d/wanghaohan/eeg-brain-wave-for-confusion/cleaning-and-plotting-eeg-signals
5. Visualizations: https://github.com/JumpThanawut/Predicting_Confusion_using_EEG_Data/tree/master/result/visualization