Analysis of Brainwave (EEG) Signals for Confusion

Hemkesh V Kumar, Jeevan Nagaraj, Pooja Balusani, Pratik Chatterjee
(01FB14ECS083, 01FB14ECS088, 01FB14ECS145, 01FB14ECS161)
Team Aayana (Team Number 5)
Department of Computer Science and Engineering, PES University

ABSTRACT

Massive Open Online Courses (MOOC) have widely evolved over the years. MOOC videos don't support immediate feedback like classroom learning, and hence it is indeed hard to detect the student's reaction to the video. There are several situations in which the sentiment of confusion arises which can't be solved immediately on watching the video of the lecture. Hence, MOOCs face the shortcomings of not being able to detect the confusion levels. In order to overcome this, electroencephalogram test is conducted on the student while watching the video and the data is recorded and analysed to study the confusion suffered by the student while watching MOOC videos.

We trained and tested classifiers in order to detect whether the student is confused while watching the video of the course material. We used Machine learning algorithms such as Decision Tree, C5.0, KNN, Naive Bayes, Random Forests, Boosting Algorithm and SVM. For a 60-40 random division of the dataset, used as training and testing data respectively, Random Forest gave us the highest accuracy of ~66% followed by Support Vector Machines giving us an accuracy of ~63%.

1. INTRODUCTION TO THE CONTEXT

An electroencephalogram (EEG) is a test which is used to detect abnormalities with respect to the brain's electrical activity. [7] EEG is a procedure which tracks and records brain wave patterns. Electrodes are small metal disks having thin wires which are placed on the scalp of the head of the person and then signals are sent to the computer to record the results.

MOOC are widely taken up by many people around the globe. Though it has several advantages, it has a few disadvantages too. Students taking up these courses may get confused while watching the videos. This confusion cannot be easily detected by the makers of the videos as it lacks immediate feedback.

Hence, in order to bridge the gap between the students and the teachers and to enable immediate feedback, the brain activity of a certain number of students were recorded while watching these videos and they were later asked if they found the video confusing. Hence, the dataset consists of the measure of brain signals obtained (through EEG) and the feedback given by the student.

The EEG dataset for this student has the following features:

Parameter 1: Subject ID
Parameter 2: Video ID
Parameter 3: Attention (Measure of mental focus)
Parameter 4: Mediation (Measure of calmness)
Parameter 5: Raw (Raw EEG signal)
Parameter 6: Delta (Power spectrum - 1-3 Hz)
Parameter 7: Theta (Power spectrum - 4-7Hz)
Parameter 8: Alpha 1 (Power spectrum - Lower 8-11 Hz)
Parameter 9: Alpha 2 (Power spectrum - Higher 8-11 Hz)
Parameter 10: Beta 1 (Power spectrum - Lower 12-29 Hz)
Parameter 11: Beta 2 (Power spectrum - Higher 12-29 Hz)
Parameter 12: Gamma 1 (Power spectrum - Lower 30-100 Hz)
Parameter 13: Gamma 2 (Power spectrum - Higher 30-100 Hz) [8]
Parameter 14: predefined label (if student is expected to be confused)
Parameter 15: user-defined label (if student was actually confused) [6]

2. SUMMARY OF THE LITERATURE SURVEY REPORT

The literature survey report introduces the proposed plan of action which is to study and investigate the effectiveness of MOOC (Massive Open Online Courses) by analysing the mental state of students using a single channel EEG headset. The basic goal was to train classifiers on the obtained dataset in order to detect whether a student was confused or not while watching the course video and also to justify if Electroencephalography (EEG) can be used in this sort of real-time analysis.

The primary inspiration for this is derived from Paper 2.1 (Using EEG to Improve MOOC Feedback Interaction [1]) in the Literature Survey. Ten students were shown a set of videos across a wide array of topics of varying levels of confusion. The students were then asked to rate each video on a scale of 1-7 based on their levels of confusion. The students' body language was monitored by a set of human observers. Classifiers were trained and tested to detect the confusion levels of a student and its performance was analogous to that of the human observers. A correlation of 0.3 was found between the predefined labels for a video and its user defined labels.

Paper 2.2 (Brainwave Recognition of Words [2]) talks about the usage of data analytics techniques to predict a word being processed by a student associated with the brainwave frequency band of his/her brainwave. The obtained accuracy ranges from 34% to 90%.

Paper 2.3 (I Think, Therefore I Am: Usability and Security of Authentication Using Brainwaves [3]) illustrates the use of EEG data to authenticate users based on their brainwaves. It involves collection of data by subjecting human subjects to a set of 7 tasks and studying the feasibility of the data obtained to authenticate users. It was found that this authentication was possible with a considerable degree of accuracy.

Paper 2.4 (Classification of EEG Signals from Four Subjects During Five Mental Tasks [4]) further strengthens the cause on analysing EEG data for mental tasks and provides a perspective for analysis using the frequency bands (Delta, Theta, Alpha and Beta) as present in our dataset. Accuracy ranging from 38% to 71% was obtained.

Paper 2.5 (Toward Exploiting EEG Input in a Reading Tutor [5]) was used for analysing the complexity of sentences being read by students to predict easy, hard, pseudo words and unpronounceable words. EEG data collected from adults and children were used to train binary logistic regression classifiers on a reader-specific and reader-independent basis. Accuracy varied from 43% to 69% for the reader-specific classifier and 41% to 65% for the reader-independent classifier.

It was proposed to use a set of appropriate classification methods to further strengthen EEG Analysis on the dataset at hand to obtain a considerable improvement of accuracy.

3. PROBLEM STATEMENT

The primary goal of this endeavour is to study the benefits/shortcomings of MOOC (Mass Open Online Courses) using Electroencephalography (EEG) data gathered by analysing EEG brainwaves of the students watching online videos and to detect if they find the video confusing.

The dataset consists of data for 10 students for each video watched, along with a set of features which primarily encompass EEG bandwidth as well as attention and meditation levels.

Assumptions made while conducting the study include presence of minimal noise in the tabulated observations as well as absence of any classification bias introduced either by the students or the observers. It was presumed that students were well versed with the rating scheme allocated to them so as to ensure correctness of the user defined confusion levels as well as ensure that all students had an approximately similar understanding of the video topics which they were tested upon.

4. APPROACH

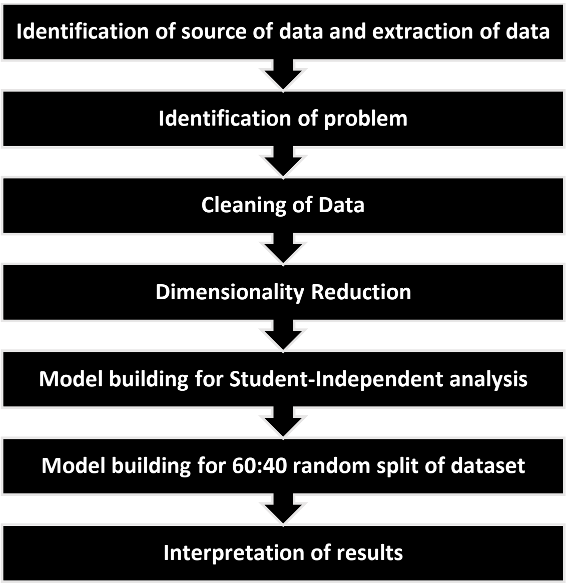
Identification of source of data and extraction of data
Identification of problem
Cleaning of Data
Dimensionality Reduction
Model building for Student-Independent analysis
Model building for 60:40 random split of dataset
Interpretation of results

Figure 1 **System Flow Chart**

Our hunt for a dataset to work on, began with the exploration of numerous datasets of various genres online. During our quest, we came across the EEG dataset. Only a handful of data analytics had been performed on this dataset and the dataset seemed extremely challenging. Enhancement of student-teacher interaction has been a persistent challenge ever since the bloom of e-education. Hence, we narrowed down to the EEG dataset to try and test out, if we could determine if a student was confused or not based on the electric brain potentials recorded when watching an educational video.

With the dataset in hand, the problem that we were going to be tackling was quite evident – predicting if a student found a lecture video confusing or not based on the EEG brain activity data of the student while watching the video. The dataset obtained was not exactly the values of electric brain potentials recorded from the electrodes, but instead was a more processed and reformed data, in terms of delta, theta, alpha, beta and gamma values. Hence, very little pre-processing had to be done.

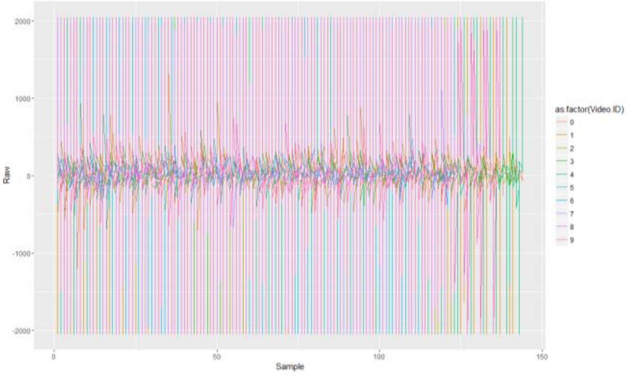

Figure 2 **Raw v/s Sample Graph before cleaning of data**

The next step in the process was cleaning of data. In order to clean the data, the dataset was initially grouped based on video ID and assigned consecutive sample numbers starting from 1, to every video group. Next, we plotted the value of ‘Raw’ parameter against the sample value. The results obtained (Figure 2) clearly proved that the dataset was noisy. So, we now had to determine the source of this noise. There were 2 possibilities here: the noise could have been due to a student or due to a video. Hence, individual plots for every video-student combination depicting ‘Raw’ value were made. On studying the nature of all the graphs, it was apparent that ‘Student 6’ had messy data as every graph of Student 6 had regular spikes [6] (Figure 3) in contrast to others (Figure 4). On removing student 6 from the dataset, the graph of Raw v/s Sample values looked comparatively noise-free (Figure 5).

Post the cleaning of data, dimensionality reduction was performed. As already mentioned, the dataset was already pre-processed and hence, there were not many parameters to be dropped. On performing Principal Component Analysis (PCA) on the dataset and looking at the contribution percentages, there was only one column which could be dropped which was ‘predefined.label’. Hence, all models were built on the dataset that did not include recordings with ‘subject.ID’ = 6 and without having ‘predefined.label’ as a parameter.

Next, a number of classifier models were built on this dataset, namely decision trees, C5.0, random forests, Naïve Bayes Classifier, boosting model, k-nearest neighbour and support vector machines. On performing the ‘Student-Independent’ analysis on the dataset, KNN, SVM and random forests gave the highest cumulative accuracy average.

Having obtained the accuracies of ‘Student-Independent’ analysis, we wanted to test and find out how these classifiers would probably perform for a random sample of the dataset. Hence, we divided the dataset so that 60% of the samples could be used to train the model while the rest could be used to test the same.

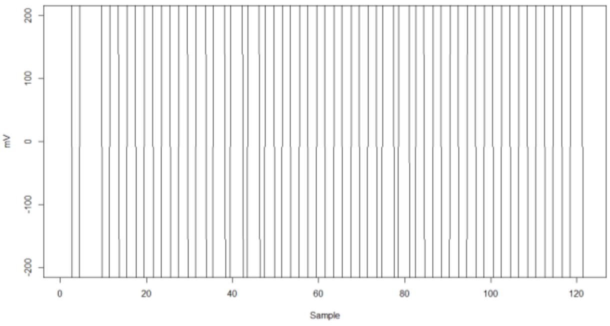

Figure 3 **Nature of graph of ‘Raw’ values for student 6**

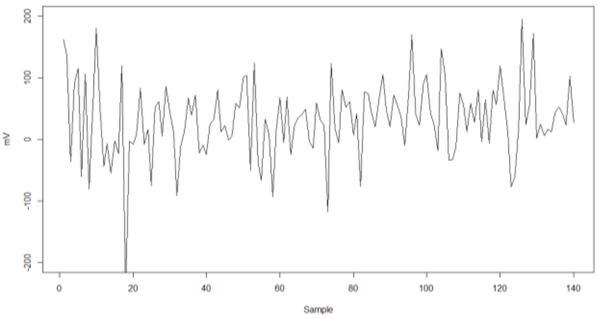

Figure 4 **Nature of graph of 'Raw' values for all subjects except student 6**

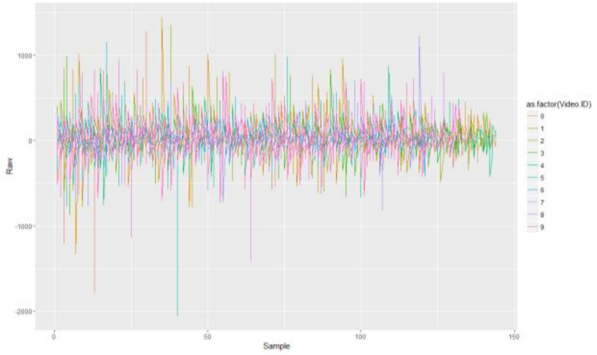

Figure 5 **Raw v/s Sample Graph after cleaning of data**

5. EXPERIMENTS AND RESULTS

The dataset was read and columns 'subject.ID', 'Video.ID' and 'Self.defined.label' were removed. Principal Component Analysis was performed on the dataset to the determine the most contributing factors. ‘predefined.label’ was not found to be a contributing factor. Hence, it was removed from the dataset while building models. Student 6 was removed from the dataset as he contributed to noise, as evident from the plot of value of raw against the value of sample for all subjects and videos.

Student Independent Analysis was performed on the dataset while building all models i.e. one student was removed from the dataset to be used as testing data while the rest of the students were used as training data. [1] All the models were built to predict ‘Self.defined.label’ parameter.

The following Machine Learning Algorithms were used to build models:

Decision tree uses a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's classification [9] (represented in the leaf nodes). Student-independent testing on the dataset with an Decision Tree Model built on ‘Self.defined.label’ gives us an accuracy ranging from 41.90% to 69.28 % with a mean of 56.36%.

A C5.0 model works by building a decision tree and splitting the sample based on the field that provides the maximum information gain. [10] The C5.0 model can predict only categorical target. Student-independent testing on the dataset with a C5.0 Model built on ‘Self.defined.label’ gives us an accuracy ranging from 43.12% to 59.48% with a mean of 51.83%.

Random forests are an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (for classification). [14] Student-independent testing on the dataset with a Random Forest Model built on 'Self.defined.label' gives us an accuracy ranging from 47.98% to 63.17% with a mean of 57.93%.

The k-Nearest Neighbours algorithm is a nonparametric method used for classification. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. [15] Student-independent testing on the dataset with k-Nearest Neighbours Model built on 'Self.defined.label' gives us an accuracy ranging from 47.26% to 68.26% with a mean of 58.89%.

Naive Bayes classifier is a simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. [13] Student-independent testing on the dataset with an Naïve Bayes Model built on 'Self.defined.label' gives us an accuracy ranging from 44.63% to 63.26% with a mean of 52.98%.

Boosting is an ensemble method in which the final predictions are made by aggregating predictions from a number of individual models, i.e applying past experiences to new iterations of tasks to be performed. [12] Student-independent testing on the dataset with a Generalized Boosting Model built on 'Self.defined.label' gives us an accuracy ranging from 48.05% to 67.63% with a mean of 56.69%.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. Given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. [11] Student-independent testing on the dataset with an SVM Model built on 'Self.defined.label' gives us an accuracy ranging from 48.37% to 67.56% with a mean of 58.11%.

We now have a rough idea of how these models performed on the dataset during the Student Independent Analysis. Random Forests, k-Nearest Neighbours and Support Vector Machine Models gave us the highest accuracies in Student-independent testing. To get a clearer view of how these models perform on the dataset, we carried out Random Sampling on the dataset. We bring about a 60-40 split on the dataset i.e. 60% of the data being used for training while the remaining 40% of the data being used for testing.

Random Sampling testing on the dataset with a Random Forest Model built on 'Self.defined.label' gives us an approximate accuracy of 66.87%.

Random Sampling testing on the dataset with a k-Nearest Neighbours Model built on 'Self.defined.label' gives us an approximate accuracy of 60.97%

Random Sampling testing on the dataset with an SVM Model built on 'Self.defined.label' gives us an approximate accuracy of 63.19%.

Support Vector Machines work exceedingly well on a two-class problem especially when the dataset is noiseless and balanced. In our dataset, the removal of the noisy component (student 6) enhanced the accuracy obtained when trained using SVM. Also, SVM avoids overfitting of data. This becomes primarily important for a dataset like EEG as overfitting of data in these kind of datasets would result in inaccurate results. For this particular dataset, linear kernel gave a higher accuracy over other non-linear kernels as the number of features is large and hence, mapping the data to a higher dimensional space would not have any substantial effect.

Random Forests is one of the most accepted machine-learning methods when it comes to dealing with datasets with categorical target data. It also has the ability to handle a generally prevalent level of noise. The bagging process ensures that there is very little bias in the training set thereby minimising any sort of overfitting. The Random Forest algorithm is therefore suited for EEG data as the data inherently contained a level of noise. Random forests associate a weight with every tree it builds and the result predicted is actually a weighted result. Hence, the use of this algorithm avoids overfitting and results in a sustainably viable model which is able to make predictions with an unparalleled level of accuracy.

6. CONCLUSION

In our study, we found that the Random Forest Classifier provided us with a classification accuracy of 66.87% against other algorithms used, whereas student-independent classifiers used in the paper achieve a classification accuracy of 51%. Yet, these are basic classification techniques. If a much more detailed study were to be performed, with sophisticated classifiers involved, we might possibly achieve a higher accuracy. Hence, electroencephalography could be deployed with Massive Open Online Courses to detect the confusion level of a student while watching an online video and obtain high accuracies of prediction of the confusion level. The deployment would unequivocally prove to be a breakthrough in enhancing the student-teacher interaction in e-education.

CITATIONS

[1] Haohan Wang, Yiwei Li, Xiaobo Hu, Yucong Yang, Zhu Meng, Kai-min Chang “Using EEG to Improve Massive Open Online Courses Feedback Interaction” Language Technologies Institute, School of Computer Science, Carnegie Mellon University

[2] Patrick Suppes, Zhong-Lin Lu, And Bing Han “Brain wave recognition of words” Center for the Study of Language and Information, Stanford University, Stanford, CA 94305; and Department of Psychology, University of Southern California, Los Angeles, CA 90089

[3] John Chuang, Hamilton Nguyen, Charles Wang, and Benjamin Johnson “I Think, Therefore I Am: Usability and Security of Authentication Using Brainwaves” School of Information, UC Berkeley Department of EECS, UC Berkeley Department of Mathematics, UC Berkeley

[4] Charles W. Anderson and Zlatko Sijercic “Classification of EEG Signals from Four Subjects During Five Mental Tasks” Department of Computer Science Colorado State University Fort Collins, CO 80523

[5] Jack Mostow, Kai-min Chang, and Jessica Nelson “Toward Exploiting EEG Input in a Reading Tutor” Project LISTEN, School of Computer Science, RI-NSH 4103, 5000 Forbes Avenue, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[6] Haohan Wang “EEG brain wave for confusion detection” | Kaggle

[7] WIKI “Electroencephalography”

[8] What are Brainwaves? | brainworks “http://www.brainworksneurotherapy.com/what-are-Brainwaves”

[9] Decision Trees | RDataMining “http://www.rdatamining.com/examples/decision-tree”

[10] C5.0 | Connor Johnson “http://connor-johnson.com/2014/08/29/decision-trees-in-r-using-the-c50-package/”

[11] SVM | SVM Tutorial “http://www.svm-tutorial.com/2014/10/support-vector-regression-r/”

[12] Boosting algorithm | Analytics Vidhya “https://www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-methods/”

[13] Naive Bayes | R-Bloggers “https://www.r-bloggers.com/revoscalers-naive-bayes-classifier-rxnaivebayes/”

[14] WIKI “Random forest”

[15] KNN | Analytics Vidhya “https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/”

[16] Principal Component Analysis: How to reveal the most important things in your data? | STHDA “http://www.sthda.com/english/wiki/principal-component-analysis-how-to-reveal-the-most-important-variables-in-your-data-r-software-and-data-mining”

CONTRIBUTION OF EACH MEMBER

- Hemkesh V
  Literature Survey of paper ‘Exploiting EEG Input in a Reading Tutor’, Implementation of Naïve Bayes Algorithm and Boosting Algorithm
- Jeevan Nagaraj
  Literature Survey of paper ‘Brain wave recognition of words’, Implementation of Principal Component Analysis, Cleaning of dataset and Support Vector Machines
- Pooja Balusani
  Literature Survey of paper, ‘I Think, Therefore I Am: Usability and Security of Authentication Using Brainwaves’, Implementation of Decision Tree and C5.0 Algorithm
- Pratik Chatterjee
  Literature Survey of paper, ‘Classification of EEG Signals from Four Subjects During Five Mental Tasks’, Implementation of Random Forests and k-Nearest Neighbour.

Each member of the team expressed themselves, influenced the group's decisions and made strong contributions to the discussion at meetings and to the work of the group project.