CSC 529 Fall 2016: Individual Data Analysis #1
Group: Dinosaur Rodeo
Matt Winkler
October 19, 2016

# Analyzing Aggregated EEG Data with Decision Tree, Random Forest, and SVM

## Introduction:

The motivation for this analysis is that the EEG dataset contains measurement data with more granularity than the target variable. Our goal is to predict subjects' self-assessment (represented by the variable "self-definedlabel" in this analysis) of whether or not they were confused by a given video, which is a binary output. The target variable is only assigned once for each combination of subjectid and videoid. However, the EEG features in the dataset are measured on half-second intervals, and give us between 112 and 144 observations for a given combination of subjectid and video id. The original dataset contains 11,536 observations, which the grouped data compresses this down to 90 observations (9 subjects who each watched 10 videos) The goals are to understand 1) which aspects of the raw data (central tendency, extreme values, skew) are most relevant to predicting a subject's confusion state at the conclusion of the video, and 2) what classifier (among Decision Trees, Random Forests and Support Vector Machines) will handle this data best.

## Testing Process:

To prepare the data for this analysis, all the EEG variables (attention, meditation, delta, etc.) were grouped at the subjectid/videoid level and aggregated using the mean, median, max, min, range, standard deviation, and skew. There are two hypotheses being tested: 1) Whether information useful for prediction is contained within certain aspects of the observation time series for each subject's video experience, and 2) That the information can be summarized in a useful way. Grouping the data and running it through each of the classifiers creates an 8

(feature groups) x 3 (classifiers) grid in which we can compare their performance. I also ran a Random Forest across all possible features in the dataset, where the idea was to compare the importance of each feature against all the others within and across feature groups. Each combination of the feature group / classifier was run through a 5-fold cross validation 90 separate times, with the average accuracy score computed at each iteration.

**Feature Groups Analysis**:

There were three main learnings from analyzing the individual feature groups across each classifier. The first is that, based on running the Random Forest with all available features, the demographic data provided does not influence the predictions by very much. Those features were at or near the bottom of the feature importance ranking, while the other features extracted from the EEG variables appeared to be more relevant (See Figure 1 in Appendix). Within the EEG feature groups, it appeared that training the models measures of central tendency (mean and median) provided more accurate predictions than the other aggregated metrics (See Figure 2 in Appendix). The best classifier (a ___ using the extracted medians) yielded ___ % accuracy. Finally, there did not appear to be significant improvements in classification accuracy from the initial runs with Random Forests and Support Vector Machines. However, it is possible that better accuracy could be attained by further experimentation with the parameters for those models.

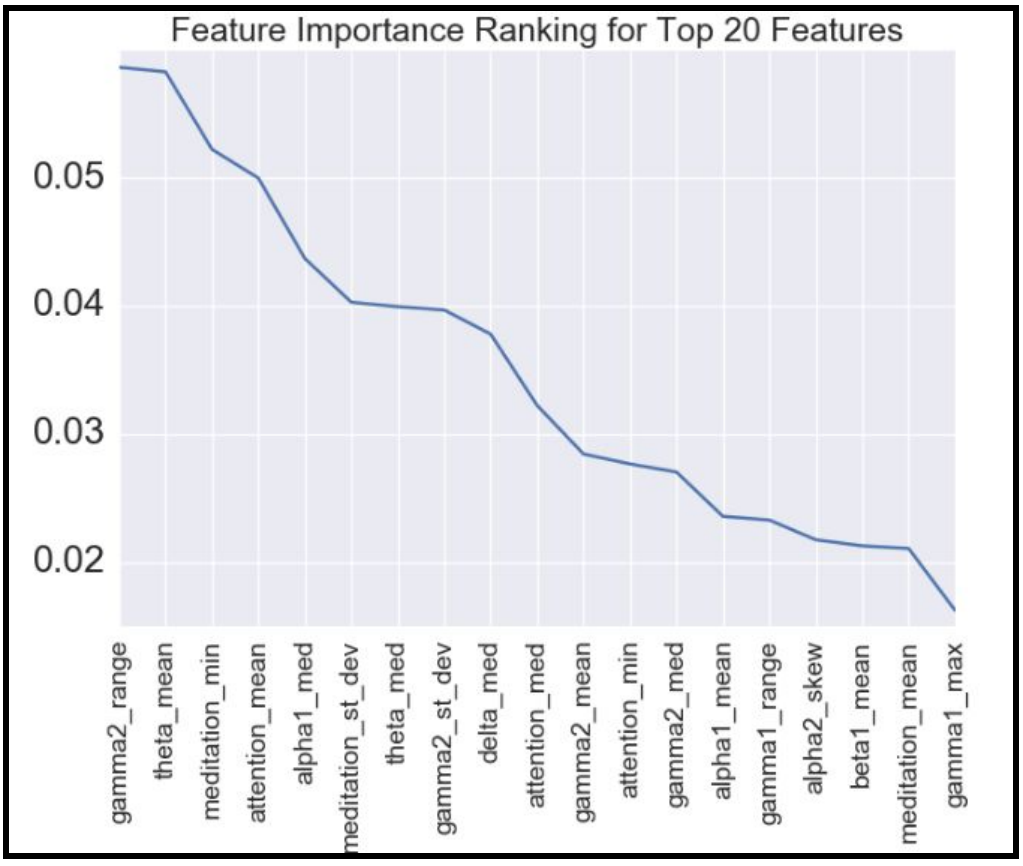
**Interpretation and Next Steps:**

With these initial results, it makes sense to explore ways to refine the summary measures, to tune some of the model parameters, and to reduce the dimensionality of the input data. For refining the summary measures, it appears that the spikes observable in the raw EEG measurements are not necessarily indicative of confusion, but that generally higher values within them do help predict confusion. It may be worth summarizing the raw data according to

deciles of values, to capture more data points at the higher end, or to summarize sub-intervals (e.g. the last 30 measurements). It seems plausible that a subject's state of mind at the end of the video would influence their self-assessed confusion level more than the beginning. Given that there are many possible features to extract from the raw data, it would probably be useful to experiment with PCA and kernel PCA as means of compressing the inputs and dealing with the strong correlations among many of them.

**Appendix**:

---- Figure 1 ----



---- Figure 2 ----

| | mean | st_dev | model |
|---|---|---|---|
| **max** | 0.583593 | 0.014013 | Decision Tree |
| **mean** | 0.725333 | 0.022296 | Decision Tree |
| **median** | 0.839333 | 0.012023 | Decision Tree |
| **min** | 0.641102 | 0.018544 | Decision Tree |
| **range** | 0.535019 | 0.022320 | Decision Tree |
| **skew** | 0.566861 | 0.017955 | Decision Tree |
| **st_dev** | 0.608806 | 0.027340 | Decision Tree |
| **max** | 0.590181 | 0.028784 | Support Vector Machine |
| **mean** | 0.755072 | 0.031228 | Support Vector Machine |
| **median** | 0.724130 | 0.033598 | Support Vector Machine |
| **min** | 0.682151 | 0.029213 | Support Vector Machine |
| **range** | 0.546112 | 0.029540 | Support Vector Machine |
| **skew** | 0.572040 | 0.029604 | Support Vector Machine |
| **st_dev** | 0.604112 | 0.029391 | Support Vector Machine |
| **max** | 0.587431 | 0.030824 | Random Forest |
| **mean** | 0.749434 | 0.030059 | Random Forest |
| **median** | 0.716765 | 0.030007 | Random Forest |
| **min** | 0.679401 | 0.031650 | Random Forest |
| **range** | 0.543115 | 0.028169 | Random Forest |
| **skew** | 0.573207 | 0.038402 | Random Forest |
| **st_dev** | 0.600292 | 0.030908 | Random Forest |