Extending the ExplainGrades Rule Extraction Method for High Dimensional Data

Ian Oxborrow

Research School of Computer Science, Australian National University, Canberra
Australia

Abstract. It is extremely important to be able to understand and simply the inner workings of a neural network. In doing so, the broader structure that data holds can be understood and explained. This report explores how the ExplainGrades method can be improved to be able to work on time series data with high dimensionality. ExplainGrades generates a value called a characteristic value which is what a typical input would look given a certain output. This paper expands on ExplainGrades by using a genetic algorithm to create a simplified version of the characteristic input. This simplified version of the characteristic input proves to be a useful tool that allows for insight into the neural network that is much more human friendly. It should be noted that this methodology would work best on time series data as the methodology that simplifies the line requires there to be correlation between adjacent values.

Keywords: ExplainGrades, evolutionary algorithm, machine learning, data simplification, rule extraction

1 Introduction

With neural networks being a highly useful tool for finding complex patterns in data, it begs the question how do we understand these networks. As the inner machinations of a neural network are enigmatic, methodologies such as the ExplainGrades were developed to extract rules from a neural network. The extracted rules are supposed to be easier for people to read, however, when this technology was applied to the Smiles dataset, due to the high dimensionality, the results were not easy for a human to understand at all, defeating the purpose of the technology. This leads to the main goal of this paper which is to build on, and improve the validity of the ExplainGrades method by making the result more human readable. This has been done by utilizing an evolutionary algorithm to develop a more simplified version of the "characteristic output" produced by ExplainGrades. While this solution is relatively effective on the dataset being used, it would likely have trouble being generalized to datasets that are not time series.

The technology that is being extended is ExplainGrades. This is a rule extraction method that creates something called the "characteristic input". The characteristic input is what the typical input that gives a certain output when fed into a network looks like. For example, in the original test of ExplainGrades, all the inputs that produced a false value were averaged and this averaged input would be the characteristic false input. Similarly, all the inputs that produced a true value were averaged and this was the characteristic true input.

The dataset used is the full version of the Smiles_v2 dataset. This dataset contains raw information on the pupil dilation of the left eye and right eye of 12 participants when they perform a real smile and a fake smile. This produced a dataset with 408 samples with a dimensionality of 1191. When compared to the Smiles_v1 dataset, it is clear that the dataset had undergone much modification to be more usable. Similarly, the raw data has been manipulated to be usable for this paper. The modifications that the data has undergone is documented more thoroughly below.

2 Data Inspection

As the data used was raw data, there were many issues with it that needed to be fixed before it could be made usable. It was immediately clear that each sample was of a different length, there were empty values all throughout the data and there were outliers all throughout the data.

	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	Unnamed: 11
0	NaN	0.004545	0.003921	0.005064	0.002959	0.002671	0.004116	0.004592	0.004020	0.003162	0.003406	NaN
1	NaN	0.004573	0.003975	0.005080	0.002947	0.002664	0.004243	0.004556	0.004009	0.003196	0.003506	NaN
2	NaN	0.004470	0.004104	0.005093	0.002959	0.002680	0.004226	0.004562	0.004008	0.003200	0.003516	NaN
3	NaN	0.004433	0.004069	0.005115	0.002939	0.002672	0.004181	0.004564	0.004025	0.003213	0.003504	NaN
4	0.004458	0.004345	0.004147	0.005100	0.002937	0.002699	0.004292	0.004581	0.004035	0.003204	0.003521	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
1186	NaN	NaN	0.004182	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1187	NaN	NaN	0.004148	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1188	NaN	NaN	0.004221	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1189	NaN	NaN	0.004136	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1190	NaN	NaN	0.004146	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 1. The raw data of one of the csv files. It shows that there are many nan values and that each sample has a different length.

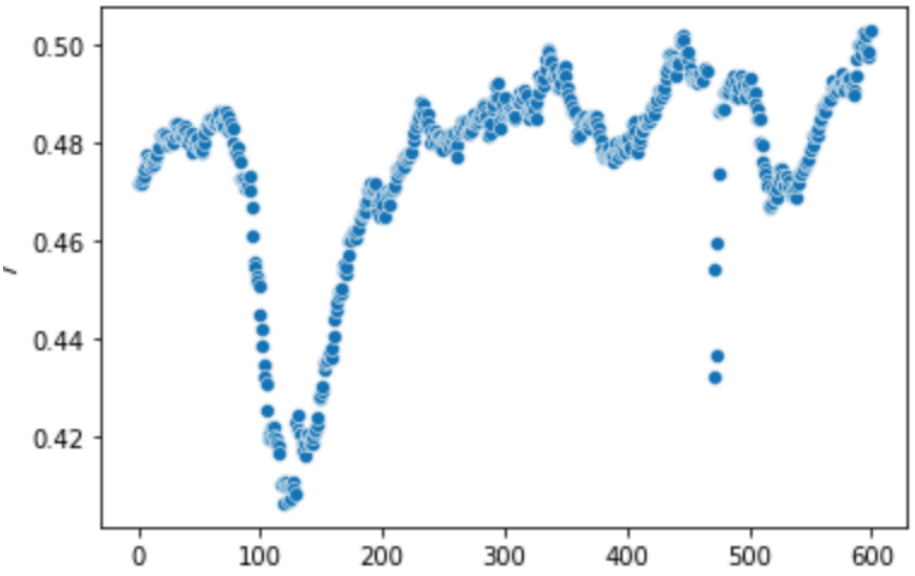


Fig. 2. An example of one of the pupillary dilation samples. It can be seen that there are outliers in the data as well.

To ensure that each sample is of the same length, each sample was trimmed from 1191 to 541. This number was selected because the modified data in Smiles_v1 dataset was of the same length. To identify outliers in the dataset, the average difference between adjacent points in a sample was calculated. If a point was over 35 times this distance from its neighbor, it was then replaced with a nan value. The value

35 was chosen simply because quick experimentation demonstrated that it isolated most outliers while leaving most inliers. To fill in these missing values, consecutive nan values were identified, and the boundaries of these consecutive values were used to linearly interpolate the missing data in that segment. An example of this is shown below.

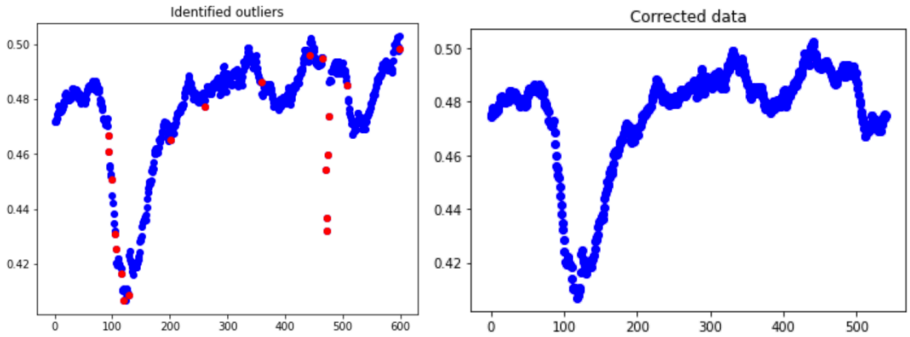


Fig. 3. A demonstration of the methodology used to remove outliers and identify missing values as well as the corrected data.

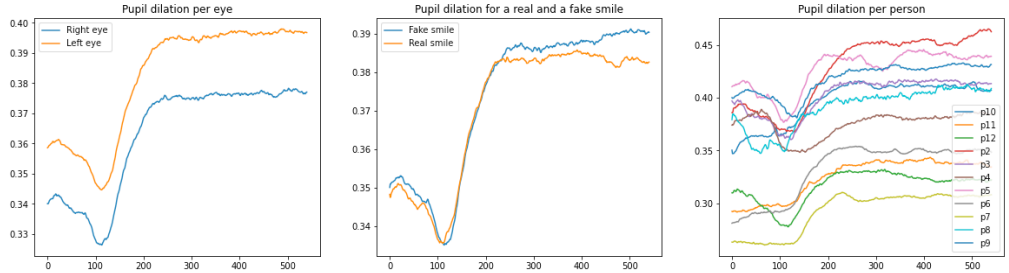


Fig. 4. A figure demonstrating the patterns in the data before data correction processes were applied

This image shows the basic patterns in the data after interpolation was performed. It is clear from the third image that some participants' eyes dilate much less than others. This sort of behavior could lead to some people always being marked as a real smile. To account for this, the data was scaled on a per person basis.

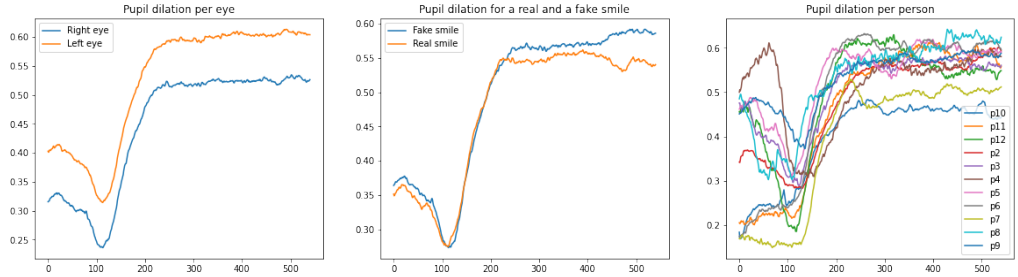


Fig. 5. A figure demonstrating that each person has a closer average value after data correction is applied.

Scaling the data per person is the last modification made to the data. It can be seen that there is now less variance in the pupil dilation of different participants, but the difference in pupil dilation for a real smile and a fake smile is preserved.

3 Methodology

Characteristic values

To produce characteristic values, a neural network was trained on the pupillary data. The network had three hidden layers, all of which used relu as an activation layer. The output layer had two output values and was passed through a softmax function producing an output that correlated to True/False values. To actually produce the characteristic values, all the samples in the dataset were fed back into the trained network. The samples were then sorted by the output they produced and were then averaged. These averaged values are the characteristic values. These two characteristic values can then be averaged to produce the "average characteristic value". This value sits somewhat in the middle of the two classes and is used in building a simplified model.

Simplified characteristic values

To produce the simplified characteristic values, the characteristic value was approximated with a polyline. To develop this polyline, the program tries to approximate the characteristic value with a given number of points. The program iteratively adjusts the position of these points and once done, straight lines are placed between these points to create a polyline. This polyline is then divided into segments, and the average difference between each point on a smile and its respective segment on the polyline is used to create a simplified version of the smile. This reduces the original smile line into significantly fewer points while retaining information about its structure. A decision tree classifier is then trained on these simplified smiles.

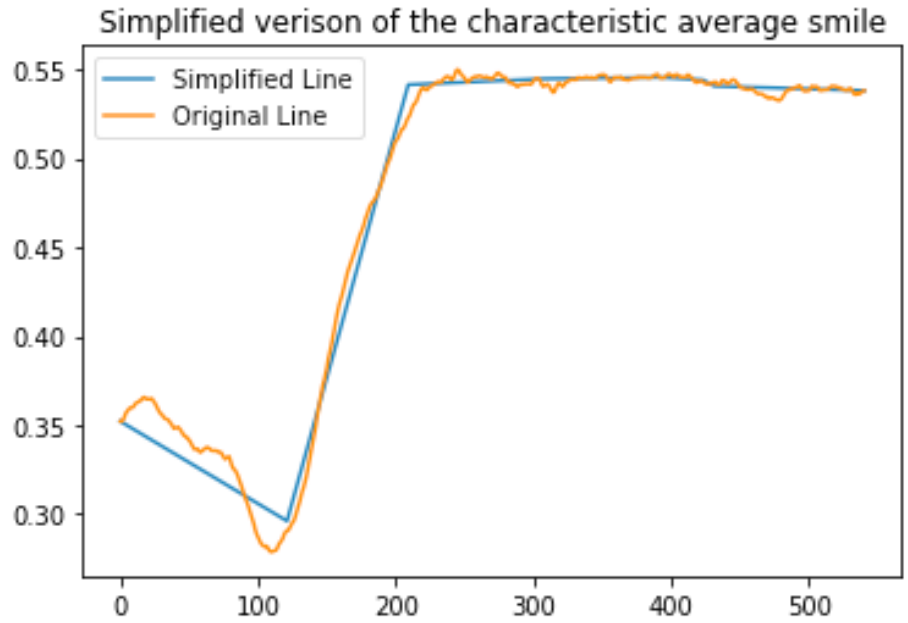
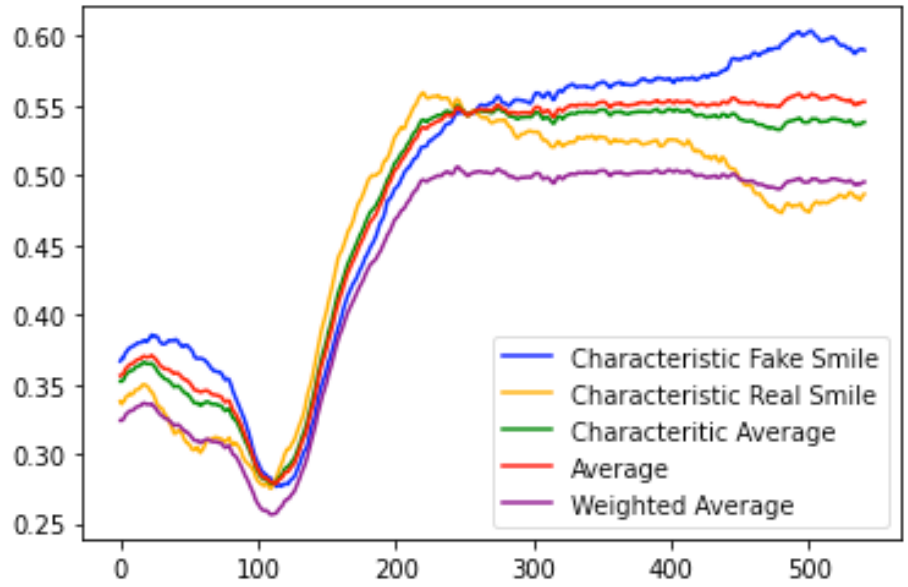


Fig. 6. The simplified version of the average characteristic line plotted on top of the average characteristic line

4 Discussion

The first issue that should be mentioned is that the dataset smiles_v2 still does not include the sex of each participant. Given that the paper mentioned that this was an important characteristic in predicting the pupil dilation behavior [4], it is very surprising that this information was kept from the raw data. The prediction model could likely be made much better by incorporating this information.



In contrast to what was found in the original paper, the weighted tensor does not appear to be a good alternative to an average tensor as is indicated visually. Given that it is much closer to the characteristic real smile than it is the characteristic fake smile, it is very surprising that the network considered it a fake smile. This is most likely reflective of the imbalance in the dataset toward fake smiles

To provide context to the issue with the previous iteration of ExplainGrades, the graphs are provided below. As is shown, all of the data in the original dataset could be divided in only a few divisions. However, each time a different attribute was used. The implication of this was that there were many attributes that could be used to divide the data. This meant that there was no definite way to divide the data, making a more human-friendly simplification of the patterns in the data very difficult to create.

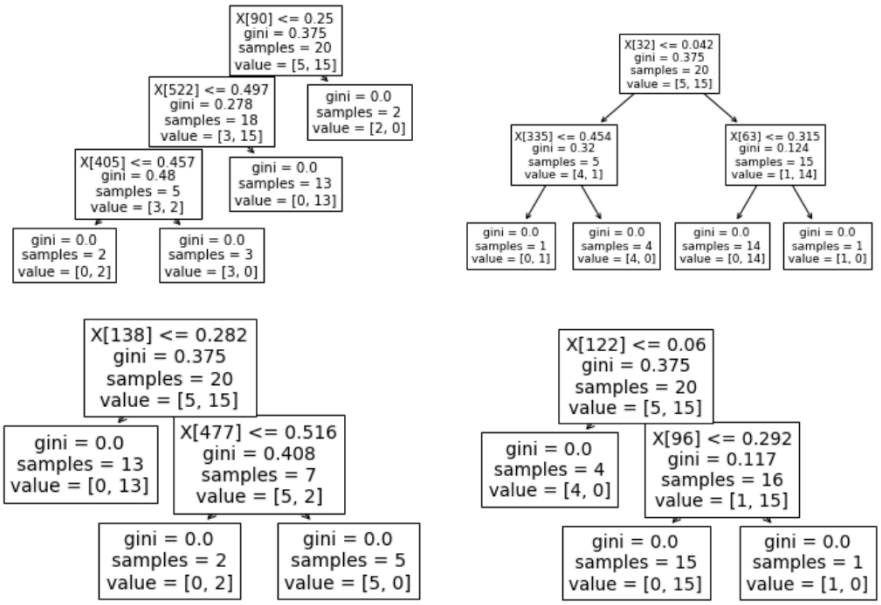


Fig. 7. Decision trees generated from the characteristic values

In contrast, the version that incorporates the simplified version of the line, consistently separates on the same values. The data used tends to have an upward curve when a fake smile is detected and this is demonstrated in the graph below. This tree is based on a polyline with seven segments, so it is clear that the main split is based on the values in the last portion of the graph. In the tree below X[6] corresponds to segment 6 in the graph.

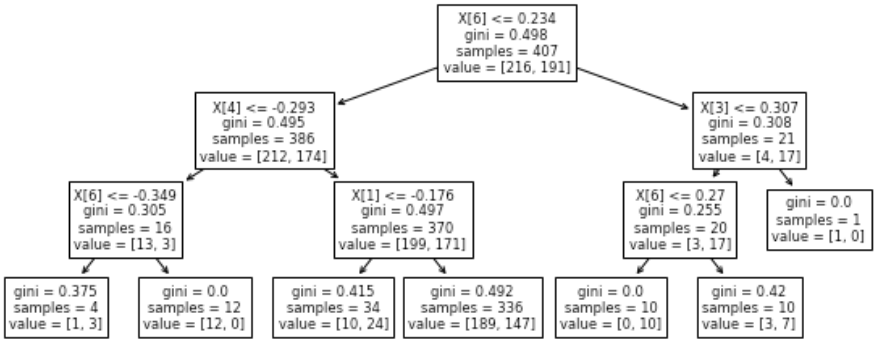


Fig. 8. Decision trees generated from the simplified characteristic values

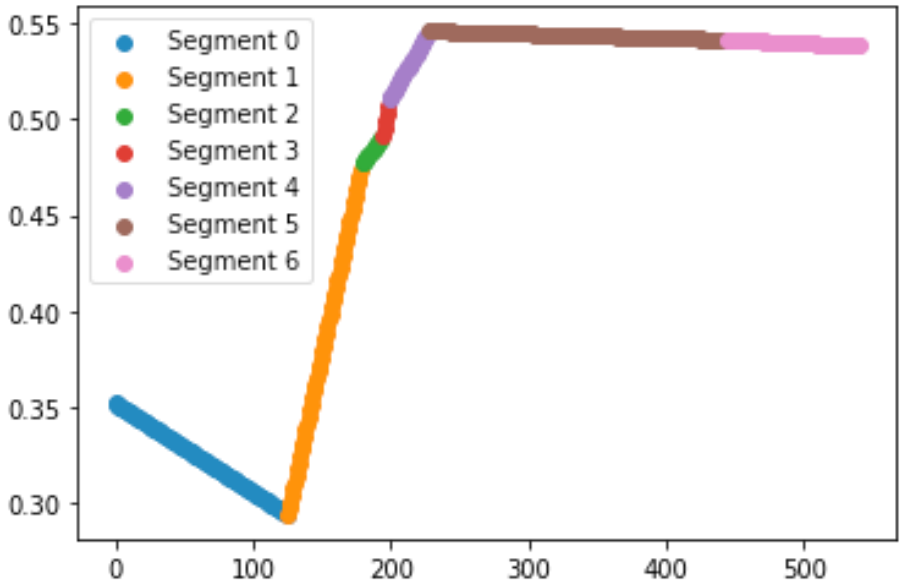


Fig. 9. The average characteristic line split into seven segments. Each segment corresponds to a decision in Fig 8

5 Conclusion

As has been demonstrated, a modification to the characteristic value can be used to simplify a neural network. This simplified characteristic not only preserves the approximate structure of the data, it enables the structure of the data to be displayed in a more human-friendly way, even if the original data has high dimensionality. It may be difficult to apply this to other situations outside of time series situations however, as it requires each point in the sample to be correlated with each other.

References

[1] Brownlee, Jason. "1D Convolutional Neural Network Models for Human Activity Recognition." Machine Learning Mastery, 20 Sept. 2018, machinelearningmastery.com/cnn-models-for-human-activity-recognition-time-series-classification/. Accessed 26 Apr. 2021.
[2] Gedeon, T. D., and H. S. Turner. Explaining Student Grades Predicted by a Neural Network.
[3] Hailesilassie, Tameru. "Rule Extraction Algorithm for Deep Neural Networks: A Review." IJCSIS, 7 June 2016.
[4] Hossain, M.Z., et al. Pupillary Responses of Asian Observers in Discriminating Real from Fake Smiles: A Preliminary Study. 27 May 2016.

[5] Konietschke, Frank, et al. “Small Sample Sizes: A Big Data Problem in High-Dimensional Data Analysis.” Statistical Methods in Medical Research, vol. 30, no. 3, 24 Nov. 2020, pp. 687–701, 10.1177/0962280220970228. Accessed 26 Apr. 2021.
[6] Oxborrow, I. (2021). Applying ExplainGrades Rule Extraction Methods to a Dataset with High Dimensionality.