Selecting influential physiological features with genetic algorithms

Andrew Law

Research School of Computer Science
The Australian National University
Canberra ACT 2601
u6673406@anu.edu.com

Abstract. When someone observes depression, measurable physiological responses are provoked. We produce a neural network to classify subject depression level based on an observer's physiological responses. A subset of features from the physiological data set are selected using genetic algorithms. We then leverage the full causal index method to extract rules from compressed representations of the training set in the form of characteristic input patterns. Finally using these rules we produce a naive decision tree based model to classify depression level in an explainable way. Genetic algorithms seem to be effective at selecting appropriate features although we could not fully validate this. Both the neural network and decision tree model still perform significantly worse than the published paper for the same dataset. Thus we are unable to determine the effectiveness of the characteristic input method for explaining the mechanisms behind detecting depression.

Keywords: Depression Detection · Physiological Signals · Neural Networks · Explainable AI.

1 Introduction

1.1 Background

Depression is a mood disorder in which the individual experiences an ongoing state of low mood and aversion to activity. It is often characterized by sadness, loss of motivation and difficulty coping with everyday tasks [1]. Current diagnosis relies primarily on the person's reported experiences or their responses to interview style examination. These methods are inherently biased as they rely on the patient's ability to report their symptoms which is often hampered by their low mood leading them to be unwilling to express themselves [1]. This issue is further confounded by variations in training and practice of health-care providers which can result in spotty diagnosis. Therefore more objective diagnosis methods are required to improve the accuracy at which clinicians can diagnose depression.

Prior study has demonstrated the possibility for quantitatively measured physiological responses to be used in signalling psychological disorders [2]. In this report we analyse how physiological responses from observers towards individuals with depression could be used to identify depression. For this task we leverage a dataset of emotional reactions to videos of depression to train a neural network to recognize depression from an observer's physiological signals [3]. Genetic algorithms are used to select the subset of influential features from the dataset [4]. Crucially, the characteristic input method is then employed to generate rules to explain conclusions drawn by the neural network [5]. Through the analysis of these rules we hope to aid in the generation of objective measures for depression and other psychological disorders which can overcome issues associated with self reporting of patient symptoms and reduce the burden on clinicians to analyse verbal responses when determining a diagnosis.

1.2 Explaining Model Outputs

Although neural networks have been proven to be highly effective at solving a variety of difficult and complex prediction problems, their utility is critically limited by the difficulty of interpreting these models [6]. As a result, a variety of different methods have been formulated to understand these so called 'black-box' models. In our case we use the differentiability of the activation function to determine the effect that changes on the input have on the output and thus explain the workings of our neural network [5]. From this causal index, rules can be extracted about which inputs are likely to cause which outputs in the model. Analogous to how a doctor might diagnose a patient by listing their symptoms, our method seeks to explain the output of a

neural network based on its inputs which is intuitive for humans to understand. To avoid fully computing the input or output derivatives for every single training entry which would be very computationally expensive, compressed representations of the training set as characteristic input patterns are used to extract rules from.

The characteristic input method classifies input patterns in terms of their effect on each particular output [5]. The set of inputs which turns on a particular output is described as the characteristic ON pattern for that output, characteristic OFF patterns are defined similarly. In our task, the output label depression level is classified into 4 different classes which we derive Characteristic ON patterns from to create rules for which we can explain how each depression level is reached.

Characteristic ON patterns are calculated by computing the mean of the set of inputs in the dataset that output each respective class. To extract the rules, we compute the derivatives of model outputs w.r.t inputs for each characteristic input pattern yielding a Jacobian matrix [7]. Given this matrix, we then select the input variables that have the highest absolute derivatives w.r.t the output of the characteristic input pattern to derive decision rules. Rules are derived from the sign of the derivative chosen and the characteristic input value, e.g. if for a characteristic input c, $\frac{dO}{di} > 0$, then the rule would be:

$$if\ input > input(c) \Rightarrow output = output(c) \tag{1}$$

2 Experiment Design

2.1 Dataset

The experiments in this report are applied to the Detecting emotional reactions to videos of depression [3] dataset, which contains a set of 192 responses towards videos with depressed individuals. The data was originally captured as time series data and processed into discrete responses. For each response, 85 different statistics were measured related to galvanic skin response, skin temperature and pupil dilation. Depression levels were measured from none (0) to severe (3) depression. There is also an equal count of each class present in the dataset, hence we use accuracy as our main evaluation metric throughout this paper. The statistics recorded included: min, max, mean, standard deviation, variance, root mean square and count where applicable. Exploratory data analysis was conducted to inform the starting point for our model and dataset preprocessing, a sample of which is shown in table 1.

Table 1. Measures of spread from sample of feature data

Statistic	min normalised pupil left	max normalised pupil left	mean normalised pupil left
count	192.000000	192.000000	192.000000
mean	0.169293	0.652716	0.398758
std	0.092202	0.166815	0.09891
min	0.000000	0.310971	0.182056
25%	0.101014	0.518944	0.329394
50%	0.161945	0.634173	0.384433
75%	0.216800	0.753494	0.456805
max	0.548269	1.000000	0.752143

As the range of values in each feature variable varied considerably, the feature dataset was normalized using min-max scaling to avoid biasing the model to specific features.

2.2 Feature Selection

Since this dataset presents us with many different possible features, it is important that we take only the influential features from the dataset as input to avoid over-fitting. This is achieved through the employment of genetic algorithms which previous studies have used successfully to select the optimal subset of features from a large feature set [4]. Our dataset is also relatively small, leading to very fast training and evaluation times, even when using cross-validation which makes it ideal for use with genetic algorithms as they are typically

limited by the cost of fitness evaluation.

The set of features used in classification is represented as a chromosome with a binary list where each bit represents whether the feature is included or not. The population of chromosomes is initialized by randomly assigning 1 and 0 values for each bit in the chromosomes. If we were to evaluate fitness on neural network classification performance, the neuron count must be altered to account for differences in feature set sizes between individuals. Changing the neural network architecture could have an effect on classification performance which cannot be controlled for. Therefore we run the genetic algorithm with a logistic regression classifier model using 5-fold cross validation classification accuracy for fitness evaluation. Logistic regression is chosen as it is a deterministic model which would also overcome the effect that random model initialization has on the performance of non-deterministic models and is typically used as a baseline model for neural networks in the literature. The genetic algorithm hyper-parameters were also tuned manually on the logistic regression evaluation method to achieve optimal performance.

2.3 Model Architecture

A 3 layer neural network consisting of an input layer, hidden layer and output layer, was produced for the classification task. When evaluating models which have very limited amounts of data available, the cross-validation method is often used. However since our data contains multiple data points per response for a participant it is not appropriate to randomly split the data. Randomly splitting the data might lead to data from the same response ending up in the training and test set simultaneously, which would be akin to training on the test set and would seriously compromise the credibility of such results. Hence the leave-one-participant method as described in the original paper [3] is used to evaluate the model. Hyper-parameters were tuned manually to achieve the highest test accuracy. This included, experimenting with different network parameters, optimisers, loss and activation functions. Tests were run with extending the neural network to a 2 hidden layer and 3 hidden layer architecture however this did not yield any improvements in performance and so a single hidden layer architecture was used for simplicity. Weight decay and early stopping was also tested to solve the apparent over fitting of the neural network however experiments did not yield positive results.

2.4 Rule Based Model

To test our hypothesis, a naive model which outputs labels based on the rules generated in the previous section is developed to test against our neural network. Our rule based model first finds the characteristic input pattern that is most similar to the input, it then compares the input to the decision rules of the characteristic input pattern chosen prior. If the rules are satisfied the model returns the output of the characteristic input pattern. If the rules aren't satisfied the next most similar characteristic input pattern is selected and the process of checking the input against the rules is repeated and so on.
When tested against the dataset, the initial model created with this method was only able to classify a small number of instances, with some instances being unable to satisfy any of the rule sets. To handle this, a bias was attached to each rule so as to make each rule easier to satisfy. This was captured in the form of a variable x which adjusts the characteristic input value of the rule as follows:

$$input < c(input) \times x \Rightarrow output = c(output) \tag{2}$$

$$input > \frac{c(input)}{x} \Rightarrow output = c(output) \tag{3}$$

The number of rules per characteristic input pattern and size of the adjustment x were tested at various values to determine the optimal values which were a rule count of 4 and x of 1.7. Due to this issue of unknown classifications we were not able to evaluate this model using F1 score, precision or recall.

3 Results and Discussion

The neural network produced an accuracy of 0.276 which was only marginally better than a random selection and much worse compared to the results produced in [3] and the baseline logistic regression classifier. This is likely because we were provided with only a limited dataset to train our model, which contained only 192

Table 2. Performance of depression classifier models

Statistic	Logisitic AF	Neural Net AF	Rule model AF	Logisitic GF	Neural Net GA	Rule model GA
accuracy	0.427	0.237	0.189	0.531	0.276	0.234
f1 score	0.392	0.190	n/a	0.531	0.224	n/a
precision	0.427	0.190	n/a	0.531	0.232	n/a
recall	0.601	0.190	n/a	0.531	0.232	n/a

instances as opposed to the 250 described in the original paper.

The rule based model produced an accuracy of 0.234, which is also worse than baseline. Despite the poor performance, some potentially interesting observations were made during model testing. Notably, measures involving the left eye only appeared more frequently than measures involving the right eye only among the derived rules. This might suggest some kind of bias towards physiological responses in the left pupil. In this case the so-called 'eyedness' or the tendency for an individual to prefer visual input from one eye over the other may be a factor for how individuals detect emotions from other humans [8]. This idea may be worth of further investigation should future studies become able to substantiate these observations.

After running the genetic algorithm we achieved an accuracy of 0.531 from the best individual using 46 features (GF) compared to 0.427 using all features (AF) with logistic regression classifier evaluation. Models trained on features selected by the genetic algorithm seem to perform better, however results of trained neural networks and rule models tended to fluctuate quite a bit which leaves this inconclusive.

4 Conclusion and Future Work

We find that the proposed method is ineffective in producing sufficiently reliable results to generate any meaningful inferences from. Furthermore, we cannot determine the effectiveness of the characteristic input pattern technique applied to problems of this sort as the base neural network from which rules were derived from performed so poorly. This rendered any attempts at downstream classification futile as any classifier formed from the basis of a poor classifier cannot possibly be expected to perform reasonably, which is confirmed in the results. The performance exhibited by the neural network may have resulted from insufficient training data or improper data preprocessing. Ideally, higher model accuracy in line with the performance of the models in [3] is required before we can make any reliable hypotheses about the observations in this report.

The genetic feature selection method was very effective in improving classification performance for the baseline logistic regression classifier, however we are yet to validate it's overall effectiveness due to poor performance of downstream classifiers. In future studies it may be worthwhile to extend this systematic approach to the way hyper-parameters are also selected. Manually tuning hyper-parameters can be quite time consuming and it can be difficult to find optimal values. An approach which utilizes a systematic grid search, randomized search or genetic algorithm might be more effective.

Our work was considerably limited by the poor performance of our base neural network classifier and any future work should seek to produce a better performing neural network from which the characteristic input pattern method can be properly applied and evaluated. To better leverage the effectiveness of neural networks, a larger dataset is needed so that the neural network can learn sufficiently. The rule loosening method we propose was quite simple and as our results demonstrate, not very effective. The increase in model strictness from adding rules was not commensurate with gains in accuracy as each additional variable is less influential than the previous, but imposes the same strictness. A more sophisticated method could be investigated which might help us leverage more rules to improve the performance of future models generated through this method. Since we are yet to prove the effectiveness of the characteristic input method on problems of this sort, future work may also focus on examining this in relation to other similar problems to create a broad picture of it's effectiveness. For example, it might be worthwhile trying to apply the method to anger recognition or the detection of other emotions or mental illnesses.

References

1. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Communication. 71, 10–49 (2015). https://doi.org/10.1016/j.specom.2015.03.004.
2. Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., Morency, L.: Automatic behavior descriptors for psychological disorder analysis. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–8 (2013). https://doi.org/10.1109/FG.2013.6553789.
3. Zhu, X., Gedeon, T., Caldwell, S., & Jones, R. (2019). Detecting emotional reactions to videos of depression. In INES'19: IEEE 23rd International Conference on Intelligent Engineering Systems (6 pp).
4. 1.Yang, J., Honavar, V.: Feature Subset Selection Using a Genetic Algorithm. In: Liu, H. and Motoda, H. (eds.) Feature Extraction, Construction and Selection: A Data Mining Perspective. pp. 117–136. Springer US, Boston, MA (1998). https://doi.org/10.1007/978-1-4615-5725-8_8.
5. Gedeon, T.D., Turner, H.S.: Explaining student grades predicted by a neural network. In: Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan). pp. 609–612 vol.1 (1993). https://doi.org/10.1109/IJCNN.1993.713989.
6. 1.Opening the black box of neural networks: methods for interpreting neural network models in clinical applications, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6035992/, last accessed 2021/04/26.
7. Engelbrecht, A., Viktor, H.L.: Rule Improvement Through Decision Boundary Detection Using Sensitivity Analysis. In: the International Work Conference on Neural Networks (IWANN'99. pp. 2–4. Springer-Verlag (1999).
8. 1.Chaurasia, B.D., Mathur, B.B.: Eyedness. Acta Anat (Basel). 96, 301–305 (1976). https://doi.org/10.1159/000144681.