Classifying 'fight or flight' response in facial superficial blood vessels: comparing evolutionary algorithm, neural network and other techniques

Yingjia Cai[1]

Australian National University
u6592822@anu.edu.au

Abstract. Brain-reading technology has been researched for many years. Based on physiology theory, detecting the intellectual 'fight or flight' responses via blood flows in facial superficial vessels has become a reasonable task. I implemented and tested four mature techniques, evolutionary algorithm, decision tree, maximum likelihood and neural network on a facial thermal dataset and compared the performance of them. The result indicated that evolutionary with appropriate parameters achieved best accuracy among them. I also validated models with relatively strong feature selection ability would be more suitable for this task.

Keywords: Fight or flight response · Neural network · Network physiology.

1 Introduction

Physiology of mind-body interactions is a subject of researching neural activities in the central nervous system which take responsibility for the relationship between brain and body. Cannon [1939] established the groundwork of this subject by introduced an influential theory called 'fight or flight'. This has been the fundamental statement in the area of nuance studies. It states that animals react to threats with a general discharge of the sympathetic nervous system, preparing the animal to either fight the threat or flee away from the danger. Selye [1956] enriched this theory. When an individual animal encounters an incident, its brain perceived stress. Then the brain sends messages to the body so the body can respond to the incident. Depending on the judgment towards the threat done by that animal, it will fight against the trouble if the problem is solvable, or run away if the animal realizes it requires capability exceeding its own power.

The amygdala as a part of the limbic system is the central controller of physiological responses. It would instantiate the thoughts about 'fight or flight' in creatures' mind. Lacroix et al. [2000] pointed out when the amygdala sending information to the hypothalamus, it would activate the sympathetic nervous system as the channel. Nestler et al. [2001] illustrated the hypothalamus could control the adrenal cortex by a set of complex interactions between organs called

hypothalamic–pituitary–adrenal axis. The corticosterone secretions which is controlled by adrenal cortex takes the responsibility to regulate vasoconstriction and vasodilation of the body is indicated by Buijs and Van Eden [2000].

The knowledge in physiology inspired us that the 'fight or flight ' response can be recognized based on the observing of facial superficial blood vessels. This is an unpopular classification task in the machine learning field. Derakhshan et al. [2019] made a crucial contribution to it. They designed a delicate experiment and collected relevant data that are sufficient to handle this problem. They also used some technologies like the Granger causality method to detect the physiological pattern of deceptive anxiety on the faces. I used four different techniques, evolutionary algorithm, neural network, decision tree and maximum likelihood, on their dataset and solve the same problem as a validation. Comparing these four and the result is given by Derakhshan et al. [2019], the accuracy of evolutionary algorithm and neural network are close to theirs and the others are lower. Fine-tuned evolutionary algorithm has the ability to achieve the best performance. The result of my experiments indicated that a well-designed evolutionary algorithm is suitable for this task.

2 The Data

Derakhshan et al. [2019] collected the experimental data using a mock crime protocol in Tehran. The researchers enlisted 42 health youth volunteers as participants under codes of academic ethics. The participants were divided into two groups, deceptive and truthful. The deceptive participants are asked to perform criminal acts while the truthful does not. Then the interviewers would ask questions about their behaviours and the deceptive group should deceive the interviewers. Meanwhile, a thermal camera would capture the movement of the blood on participants' facial cutaneous vasculatures. To simplify the question, we only concern the blood flow between five regions of interest, periorbital, forehead, perinasal, cheek and chin. Each region can be the origin and the destination so there are 20 directions of flow. The average flows over time are the 20 features of the data. The group the participant belongs to, deceptive or truthful, is the label of each sample. All values has been normalized to [0, 1] for data uniformity.

For some reasons, I got only 31 samples with features and labels as described above. I took 21 of them randomly as a training set and others as a test set for my methods.

3 Method

Considering the dataset I have obtained containing only not too many samples and 20 features, I decided to use four simple and classic methods, decision tree, maximum likelihood, neural network, and evolutionary to accomplish the classification task. They are tried-and-true and famous for generalization abilities. All methods below used the same random seed for PyTorch, NumPy, Pandas and Random so the results are reproducible. Derakhshan et al. [2019] presented the

results of five different classifiers with and without feature selection technique. SVM with linear kernel reached the best accuracy in both situations. I take it as the control group of my experiments. Since we used different data splitting, the results cannot be directly compared.

3.1 Decision Tree Classification

Classification And Regression Tree (CART) algorithm which is invented by Breiman et al. [1984], is one of the most powerful tree methods. The main idea of it is to use Gini impurity as a measurement to find the best feature and the best split point. The feature with the largest Gini impurity contains the largest uncertainty and should is the best node to split. To find the feature that has the largest Gini impurity, Breiman et al. [1984] defined the Gini coefficient as:

$$Gini(t) = 1 - \sum_i [p(C_i|t)^2]$$

where $p(C_i|t)$ is the probability of a node t being the category i. Finding the node with the largest Gini impurity is equal to finding the node with the largest Gini coefficient. The Gini coefficient of the best split of node T:

$$Gini(T) = \frac{|t_l|}{|T|} Gini(t_l) + \frac{|t_r|}{|T|} Gini(t_r)$$

where t_l and t_r are the nodes on the left and right. The absolute value sign means the size. By finding the node with the largest Gini coefficient and split it recursively, we can build the optimal classification and regression tree.

3.2 Maximum Likelihood Classification

Maximum likelihood classification is a statistically based method introduced by Richards and Jia [2006]. For each sample x we want to know whether it belongs to class i, we can calculate the discriminant function $g_i(x)$:

$$g_i(x) = -ln|\Sigma_i| - (x - m_i)^T \Sigma_t^{-1} (x - m_i)$$

where m_i is the mean of class i, Σ_i is the covariance matrix for class i.
If $g_i(x)$ is larger than the threshold T_i:

$$T_i = -4.774 - \frac{1}{2} ln|\Sigma_i| + lnp(w_i)$$

Then we can say sample x belongs to class i. Where $p(w_i)$ is our prior knowledge of the probability of a sample belongs class i.

3.3 Neural Network Classification

Neural networks are widely used on different tasks in recent years. I tested a various number of hidden nodes and hidden layers. The best architecture contains 1 input layer with 20 nodes, 2 hidden layers with 15 hidden nodes each, 1 output layer with 1 node. The optimiser is set as a stochastic gradient descent method and the loss function is mean square error. Sigmoid function is used as an activation function for the input layer and hidden layers. The output of the output layer would be manually analysed so there is no activation function for the output layer. To avoid over-fitting, I used leave-one-out validation to decide the point to stop training.

The output of each sample in the test set is a value from -0.28 to 0.87. Then we need a threshold θ to determine which one belongs to the positive class and which one belongs to the negative class. Milne et al. [1995] configure their threshold by manually varying it in another classification task. I used a similar procedure by varying it from -0.2 to 0.7. The precision and the recall rate on the test set of different threshold are also reported.

3.4 Evolutionary Algorithm

The result of the experiments given by Derakhshan et al. [2019] indicated that a model with strong feature selection ability is more likely to have ideal performance. I chose evolutionary as the final method because it is suitable for this question. Evolutionary algorithm is a cutting-edge technology to solve optimization problem inspired by biological evolution. Its main idea is to create some candidates solutions as individuals, then keep mutating and eliminate the inadequate individuals until the appropriate solution is found.

I use Distributed Evolutionary Algorithms package (DEAP) to implement my evolutionary algorithm. I define the individual as a ternary string of 20, consists of 1, 0 and -1. 20 is the number of features. 1, 0 and -1 respectively mean that the feature on that position has a positive influence, no influence and a negative influence to whether a person is deceived. I believe the ternary design have a stronger fitting capability than binary. To evaluate an individual, I calculate the inner product between the individual and each sample in the training set. If the value is larger than a hyper-parameter threshold θ, the sample would be predicted as positive. Otherwise, the sample would be taken as negative. Thus the fitness can be define as the accuracy of the prediction made by the individual over the training set.

On the other side, I have prior knowledge that the information of whether some is deceiving is concentrated in a few features. Most elements in the individual should be zeros. So I use another hyper-parameter k to determine the maximum capacity of non-zero elements. The individual with more than k non-zero elements would get 0 in fitness. These two hyper-parameters played a very important role in the algorithm. I made a grid search, $\theta \in \{-2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0\}$ and $k \in \{1, 2, 3, 4, 5, 6, 7\}$, to find the best combination of them.

Besides, there are many other parameters in my evolutionary configuration. I use tournament with top-3 individuals can be alive as the select method, two-point crossover as recombination method, and shuffling the elements in the individual as mutate method. The function deap.algorithms.eaSimple is also used to solve the evolutionary question. The crossover probability, the mutate probability, and the number of generations are 0.8, 0.2 and 200. Changing these parameters will make an obvious reflection on performance. So I did not use grid search to find the optimal values of them.

4 Results and Discussion

Please allow me to show the result of the neural network and evolutionary algorithm with different hyper-parameters first. Then use the best configuration to compare with other models.

Table 1. The accuracy, precision and recall rate of neural network with different θ.

θ	Accuracy	Precision	Recall rate
-0.2	60.0%	55.6%	100.0%
-0.1	60.0%	57.1%	80.0%
0.0	60.0%	57.1%	80.0%
0.1	60.0%	60.0%	60.0%
0.2	80.0%	100.0%	60.0%
0.3	80.0%	100.0%	60.0%
0.4	80.0%	100.0%	60.0%
0.6	70.0%	100.0%	40.0%
0.6	70.0%	100.0%	40.0%
0.7	60.0%	100.0%	20.0%

As we can see in Table 1, the accuracy reached 80% at $\theta = 0.2$ and decreased after $\theta = 0.4$. Precision keeps increasing with the increasing θ as my expectation. It achieved 100% at $\theta = 0.2$. Meanwhile, the recall rate decreased from 100% to 20%. So in order to get the best accuracy, θ can be a value between 0.2 to 0.4.

We can get some other interesting conclusions. Firstly, there is a wide range for threshold θ to get the highest accuracy, from 0.2 to 0.4. Considering the maximum of the output is 0.87 and the minimum of the output is -0.28, the average of the maximum and the minimum is 0.295, which is pretty close to the centre of the best range. Using the mean of the maximum and the minimum as threshold might be a good choice for this and similar problems.

Secondly, there are two positive samples in the test set that have very low ($<$0.1) output, even lower than most negative samples. The meaning of a positive sample is this participant lied in the interrogation. So, we can say there are two people deceived in the interrogation but very hard to be identified. They

performed even better than most truthful participants. There are two possible reasons that may lead to this result: one is the blood flow in facial superficial cannot reflect 'fight or flight' decision for everyone; the other is some people may have congenital or acquired skills to be undetectable deceivers. No matter what, simply using facial thermal data to distinguish someone is lying or not is dangerous.

Thirdly, the accuracy and the precision reached the summit at the same value $\theta = 0.2$. But the recall rate is only 60% at that value. If we want to find out all of the deceivers, precision would down to 55.6%. This fact indicated that conservation in identifying deceivers is not harmful to accuracy. Detecting all liars may wrongfully accuse many innocent people.

Table 2. The accuracy of evolutionary algorithm with different θ and k.

k \ θ	2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
1	n/a	n/a	n/a	40.0%	50.0%	**90.0%**	n/a	n/a	n/a
2	n/a	50.0%	50.0%	30.0%	50.0%	60.0%	70.0%	50.0%	n/a
3	50.0%	50.0%	60.0%	60.0%	40.0%	60.0%	70.0%	70.0%	50.0%
4	60.0%	40.0%	40.0%	70.0%	**90.0%**	70.0%	70.0%	80.0%	50.0%
5	30.0%	50.0%	40.0%	40.0%	70.0%	70.0%	80.0%	70.0%	80.0%
6	50.0%	30.0%	40.0%	50.0%	50.0%	70.0%	70.0%	70.0%	70.0%
7	60.0%	30.0%	50.0%	40.0%	50.0%	50.0%	70.0%	70.0%	80.0%

Table 2 presents the accuracy of evolutionary algorithm with different θ and k. Since the data have been normalized to [0, 1], some thresholds cannot be reached because of the limitation of non-zeros elements. For example, if we only concern with one feature, the output value cannot be larger than 1. I use n/a in the table to denote these situations. There are two combinations that achieved 90% on the accuracy, which are remarkable and exciting. One of them only used 1 feature and the other used 4. This is confirmed to my expectation. The information concentrated in a few features. Increasing the number of non-zero elements does not directly lead to better performance.

The first best combination is $k = 1, \theta = 0.5$. The best individual only consists of one rule: if the thermal flow from forehead to periorbital larger than 0.5, the sample deceived. This is an interesting and practical solution. Derakhshan et al. [2019] used four selection methods, including T-test, Relative entropy, ROC and Wilcoxon to estimate the Granger causality of each feature. All four methods take forehead to periorbital flow as the most informative feature. This is potent validation of my result. Based on my experiment results, I can make a conclusion that if someone has abnormal thermal flow from forehead to periorbital, there is a considerable probability that he or she deceived.

The best individual of the second best combination $k = 4, \theta = 0$ is the thermal flow from forehead to periorbital and from chin to cheek increased the probability of someone deceived while the flow from cheek to perinasal and from perinasal to periorbital decreased the probability. This is a collective solution of the informative features. By using these rules we can suspect someone deceived reasonably. Just remember to be cautious that the probability of misjudgment cannot be ignored.

By using the best accuracy of neural network and evolutionary, I can compare them with other techniques.

Table 3. The accuracy of four techniques and the control group.

Technique name	Accuracy
Decision Tree (CART)	50%
Maximum Likelihood	50%
Neural Network	80%
Evolutionary Algorithm	90%
SVM(l) with all features	58.9%
SVM(l) with selected features	87.1%

Table 3 shows the accuracy of my four models and the control group. We can see that CART, maximum likelihood and SVM with linear kernel and all features have about 50% accuracy and SVM(l) is slightly better. Neural Network and SVM(l) with selected features have about 80% accuracy and SVM(l) is slightly better. Evolutionary Algorithm has the best performance. It reached 90% on accuracy.

It is clearly that the results can be divided into two groups with a significant gap. CART and maximum likelihood used all features. It is very reasonable that their result is close to SVM(l) with all features. I expected that CART performs better because it used Gini coefficient to sort the information of the features but it did not. Neural network would automatically weigh the features in the backpropagation procedure. The weight of noise features may decrease in backpropagation, which has a similar effect with using Granger causality to select the most informative features. Evolutionary algorithm step further. It is forced to abandon most unproductive features. So evolutionary algorithm and neural network should be much better in this task. The result of this experiment validated that using weighted features or a small part of the features on this dataset may lead to better performance. And fine-tuned evolutionary algorithm is better than neural network, decision tree method and maximum likelihood classification method.

5 Conclusion and Further Work

I tested CART, maximum likelihood, neural network and evolutionary algorithm on the facial superficial blood flows dataset. By analysing the result and compar-

ing their performances with the control group, I got some meaningful conclusions. There is a wide range for the best threshold in a neural network. Outliers should be attended in the deceiver detection task. Be conservative on judgement might be a good choice. Four features chosen by evolutionary algorithm are most informative to detect the deceivers. The thermal flow from forehead to periorbital is the most important one.

Future research should focus on conducting more social experiments to enlarge the dataset. Now the lack of data made the result very unstable and not convincing. A rich dataset may lead to many other meaningful conclusions, which is helpful in many areas like interrogation.

And there is a flaw in my evolutionary algorithm. Using continuous variables to describe individuals may have better performance. Three dispersed numbers are more or less inflexible. Continuous variables can differentiate the importance of different relevant features. It might be a better choice.

Bibliography

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

R. M. Buijs and C. G. Van Eden. The integration of stress by the hypothalamus, amygdala and prefrontal cortex: balance between the autonomic nervous system and the neuroendocrine system. *Progress in brain research*, 126:117–132, 2000.

W. B. Cannon. The wisdom of the body. 1939.

A. Derakhshan, M. Mikaeili, A. M. Nasrabadi, and T. Gedeon. Network physiology of 'fight or flight'response in facial superficial blood vessels. *Physiological measurement*, 40(1):014002, 2019.

L. Lacroix, S. Spinelli, C. A. Heidbreder, and J. Feldon. Differential role of the medial and lateral prefrontal cortices in fear and anxiety. *Behavioral neuroscience*, 114(6):1119, 2000.

L. Milne, T. Gedeon, and A. Skidmore. Classifying dry sclerophyll forest from augmented satellite data: Comparing neural network, decision tree & maximum likelihood. *training*, 109(81):0, 1995.

E. J. Nestler, S. E. Hyman, and R. C. Malenka. *Molecular neuropharmacology: a foundation for clinical neuroscience*. McGraw-Hill Medical, 2001.

J. A. Richards and X. Jia. Remote sensing digital image analysis. *Remote Sensing Digital Image Analysis*, page 197, 2006.

H. Selye. *The stress of life*. 1956.