Bimodal Distribution Removal and Genetic Algorithms in Neural Network for Identifying Posed and Genuine Anger

Jinghan Zhang
Research School of Computer Science
Australian National University
u6832190@anu.edu.au

Abstract. The purpose of the experiment in the paper is to design a neural network that reads data on the change of people's pupil diameter to recognize facial expressions and determine the authenticity of emotions. In this paper, a simple feed-forward neural network is built with an overall accuracy of 91.02%. Subsequent experiments focus on how to improve the accuracy. Data quality can seriously affect the performance of feed-forward neural networks, Bimodal Distribution Removal (BDR) is attempted to identify and remove outliers and redundant features. However, experiments show that although BDR successfully removes outliers, and effectively eliminates erroneous bimodal distribution, it cannot accommodate redundant features and get over killed, the final accuracy drops to 89% with overfitting. Genetic Algorithm (GA) that can generate the optimal subset of features is used in the paper to obtain meaningful attributes, which successfully increases the accuracy to 95.23%. The final result indicates that the machine facial classification with 95.23% accuracy is more accurate in judging emotions than human recognition which only has 65% accuracy.

Keywords: Feed-Forward Neural Network, Bimodal Distribution Removal, Genetic Algorithm, Facial Expressions.

1 Introduction

Facial emotion recognition has always been a very important factor in human-human interaction, and when Human-Computer Interaction (HCI) is gradually becoming a development trend, machine recognition of human facial emotion changes has also become a hot topic. Accurate facial recognition and human emotion judgment can effectively improve the functionality and practical value of HCI. This has been explored by a very large number of scholars, such as Chen and Gedeon, who trained classifiers to identify whether a person's smile and anger are real or impersonated [1].

The purpose of the experiment in this paper is to design a neural network that takes data on the degree of change in the pupil diameter of participants to recognize facial expressions and determine the authenticity of emotions. As well as to verify whether the accuracy of the judgments based on physiological signals is higher than the results of the participants' verbal expressions.

We constructed a feed-forward neural network to train the dataset, which does not need to determine the mathematical equations of the mapping relationship between inputs and outputs in advance, but only learn some rules through their own training to get the closest result to the desired output value at a given input value.

However, the experimental environment of the participants is not perfect, so the physiological data of the participants we processed will contain a lot of noise as well as outliers or be influenced by the environment. Outliers in the experimental data can have an impact on the performance of the neural network, so in our experiments we used BDR [2] to remove the outliers during training. However, redundant features are also considered by BDR as outliers removed altogether and there is a risk of excessive removal.

We also used Genetic Algorithm (GA) [3] is for feature selection. The results generated by GA indicate the most likely best subset that optimizes the performance of the neural network, GA-based feature selection is used to train the neural network and optimize the label prediction accuracy of the neural network and reduce the computation time and cost.

2 Methodologies

2.1 data processing

The source data is from Chen and Gedeon’s research [1]. The data recorded the physiological reactions of 20 participants of different ethnicities, who were Chinese, Fijian and Caucasian. They watched 20 short videos from YouTube, 10 of which were genuine angry and 10 were posed. In total, 390 experimental data were available. The real anger videos were from documentaries and live news, while the posed anger videos were from movie clips. The raw data were divided into two parts, left eye and right eye, and recorded pupil diameter data during the participants' viewing of 20 videos. Each column is the pupil diameter measurement data for the specified participant watching the specified video, recorded every 1/60th of a second. The length of the videos varied, and therefore the number of seconds measured and the amount of data recorded varied. The longest video has 186 records.

We consider the time series data of each sample as the input vector of the neural network, so we applied data transformation to the raw data. We used the average of the left-eye and right-eye data as the recorded data and used zero padding to pad each feature vector, so that all vectors kept the same length as the longest vector (length 186). Besides, all missing values will be imputed 0. Then we added the 187th feature to record the labels, 1 indicates Genuine anger and 0 indicates posed. Since some participants did not participate in the full experiment, we ended up with a data set of 390*187, where 186 columns are pupillary measurement, 1 column of label, and 390 rows of individual experiments.

Pupil diameters varied considerably between participants, also, different ethnicities and genes have an effect on pupil diameter size. The average pupil diameter of Caucasians is larger than that of Chinese and Fijians. Therefore, the measured data were at different scales. This means that attributes contribute differently to the model learning and model fitting functions, which may introduce bias into the neural network training process. Therefore, data normalization is needed to address the comparability between data, and I used Min-Max Scale before training the model. The mathematical formula is:

$$x = \frac{X\text{i} - \min(X)}{\max(X) - \min(X)} \tag{1}$$

X_i represents a single vector of variables. All features will be transferred to the range [0, 1] after normalization. Compared to using the original unscaled data, the backpropagation of the neural network will be more stable, obtain more accurate results, and improve the computational speed when the input features are Min-Max scaled.

After normalization, we divide the dataset into 80% training set and 20% test set for training the model. When tuning hyperparameters, the dataset will be divided into 70% training set, 15% validation set and 15% test set. Considering the small size of our dataset, a random partitioning will result in unbalanced distribution of classes. Therefore, we use an algorithm to ensure that the distribution of classes in the obtained training and test sets is the same.

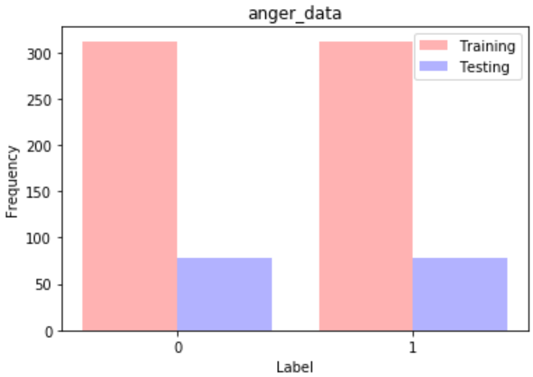

Figure.1. class distribution in training and test sets.

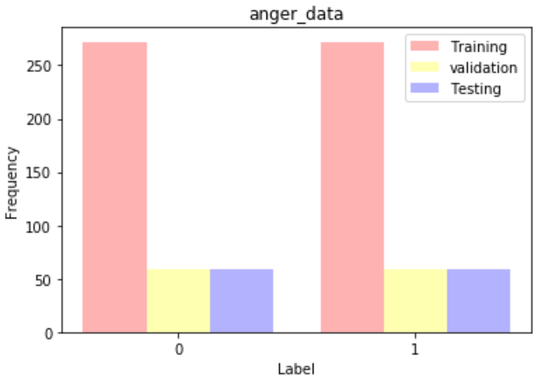

Figure.2. class distribution in training, validation and test sets.

2.2 Neural Network Architecture

The selected features are fed into the constructed neural network to train it and identify the genuine anger and posted anger. The architecture I used is a simple 3-layer feed-forward neural network, and train by a Back-propagation algorithm. The three layers are input layer, hidden layer and output layer. The Back Propagation algorithm basically replicates its input to its out-put via a narrow conduit of hidden units[3].

Each input vector has 186 features, so the input layer has 186 nodes. There are 2 nodes in the output layer, corresponding to the genuine and posed anger labels. We set up several experiments to compare the prediction accuracy of the validation set for hyperparameter tuning process. The final selection of hyperparameters achieves the optimal classification effect. The number of hidden neurons was set to 60 and epoch number is 120 to achieve optimal accuracy and to prevent overfitting. The learning rate was set to 0.01 to make the objective function converge to a local minimum in a suitable time. Since our experiment is a binary classification, we use Cross-Entropy Loss as the loss function.

ReLU is chosen as the activation function to solve the vanishing gradient problem that occurs in deep networks, making deep networks trainable and reduce calculation time. Finally, we choose to use Adam as the optimizer, because after bias correction, the learning rate of each iteration has a definite range, making the parameters relatively smooth. Adam helps to escape from the saddle point and reach global minimum.

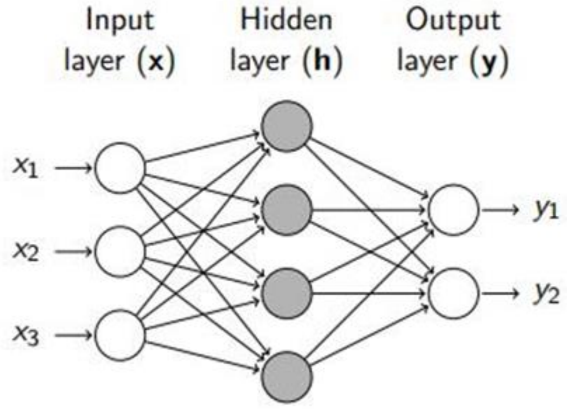

Figure.3. feed-forward neural network structure

2.3 Bimodal Distribution Removal

As the neural network learns the features of the data set, most of the data in the training set obtains very small errors, while outliers obtain large errors. When the neural network model has been generally well trained, the variance of error is smaller than 0.1, the error distribution will form a bimodal distribution as in Figure.4, with the higher peak symbolizing that the vast majority of the data is correctly classified. And the lower peaks represent the outliers being misclassified with larger errors. Then the model will then focus mainly on mining deeper features with lower peaks and outliers. But it is pointless to probe deeply into the features of outliers, it will waste computational cost and even lead to lower final accuracy. Therefore, we need to remove the outliers by using Bimodal Distribution Removal (BDR).

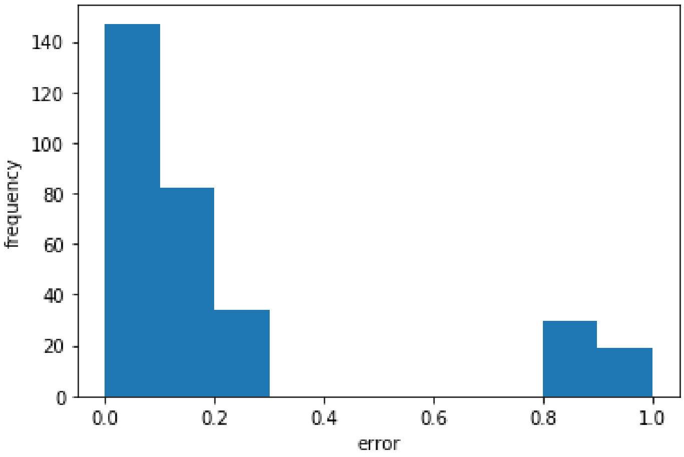

Figure.4. Error distribution when epoch=40 and variance=0.1.

Slade and Gedeon proposed BDR method [5], its main idea is performing outlier detection using a distribution of errors during training, selecting two thresholds to identify noise and outliers, and remove them. The two thresholds are determined according to the following method:

$$T_1 = \bar{\delta} \qquad (2)$$

The first threshold T1 is the average of all errors, then those records with error higher than T1 will be placed into a subset. Then we defined the second threshold:

$$T_2 = \overline{\delta_{ss}} + \alpha\sigma_{ss} \qquad (3)$$

Where $\overline{\delta_{ss}}$ is the mean error in the subset, and σ_{ss} is the standard deviation of the subset. α is a hyperparameter which is in range [0,1], and it is chosen as 0.9 in this paper by tuning and the analysis the result distribution.

To prevent deleting too much training data, BDR is activated only when variance is less than 0.1 and terminates when variance continues to decrease to another threshold VT, which we finally set to 0.01 by experiment.

2.4 Genetic Algorithm

During our experiments, we found that the BDR algorithm not only removes outliers, but also is less inclusive of redundant attributes. It will remove all the redundant attributes as noise, which may contain important features, which leads to a lower final accuracy. We therefore used a GA [3] based feature selection method [7] to train the model, removing redundant features and retaining important attributes, thus preventing excessive removal.

Originally developed by John Hollander, the GA draws on the biological idea that individual candidates with high fitness have a better chance of passing on good genetic traits to the next generation, and combines this with Darwin's theory of speciation, which is applied to algorithmic modeling, exploring all possible genomes combined in a randomized exploration of a given problem to obtain the best solution.

In GA, a gene is a binary value that represents a feature. A set of genes is a subset of traits, represented by chromosomes. The number of chromosomes in each evolutionary generation is called the population. In our experiments, there are 186 input features and therefore there will be 186 gene segments on the chromosome. The binary value of each feature gene represents whether the feature is selected or not, and it will be assigned randomly during the initialization phase of the GA model. The population value is set to be 20.

GA evaluates the quality of an individual by the value of the fitness function, the higher the value, the better the quality of the solution. Individuals with low fitness are less likely to be inherited in the next generation. We put the chromosomes into the simple feedforward neural network proposed in section 2.2 for training and use the test set accuracy as the criterion of fitness. During the training process, we still divided the dataset into an 80% training set and a 20% test set.

Tournament selection [6] is based on fitness scores, and we select individuals with high fitness from the population to serve as parents (one father and one mother) for reproducing new individuals in the next generation of the population. In the pre-selection phase of the tournament, we set up the entry of a random half of the individuals in the population, allowing individuals to appear in different tournament groups.

Crossover is the exchange of some of the parents' genes in pairs under a random ratio, resulting in the formation of two new individuals. In our experiments we used a One-point Crossover, with only one crossover point in the individual coding strings, and then exchanged gene segments of paired individuals with each other at that point.

After chromosome crossover, we use mutation method to generate new individuals to enhance local search ability while maintaining population diversity. Mutation rate is a hyperparameter, and by the accuracy test of validation set, we set it to be 0.0001.

2.5 Evaluation Method

When analyzing the results of neural networks and machine learning, only use accuracy may produce large errors when the class distribution is unbalanced. Therefore, we use confusion matrix as a data visualization and evaluation method. [8,13] We also used evaluation metrics extended by the confusion matrix: Precision, Recall, Specificity, and F1-Score.

3 Result and Discussion

3.1 Activation Function

After determining the hyperparameters of the neural network (hiding neuron = 60, learning rate=0.01, epoch =120), we performed comparison experiments to determine the appropriate activation function.

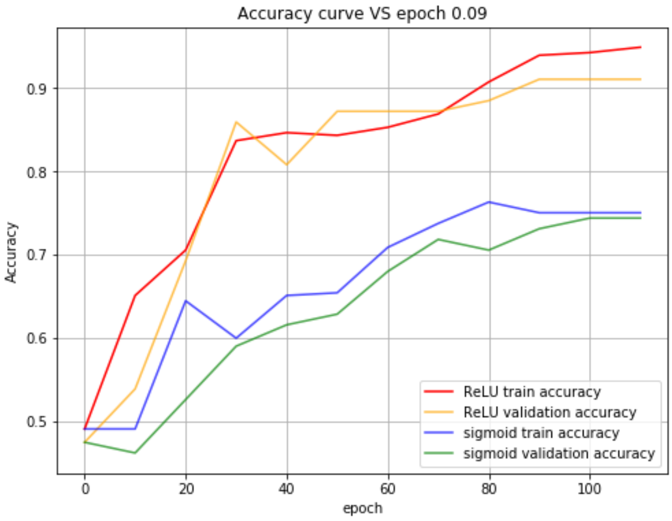

Figure.5. Validation set accuracy VS activation functions

Figure 5 shows the accuracy of validation set with different activation functions. It is clear from the images that ReLU outperforms Sigmoid. The results verify our initial conjecture that sigmoid has a vanishing gradient problem that causes information loss. While ReLU can better complete the training of the deep network. Also using ReLU will make some neurons to be zero, which causes the sparsity of the network and reduces the interdependence between parameters, alleviating the overfitting problem. This baseline model gives us an overall accuracy of 91.25%.

3.2 Bimodal Distribution Removal

We use BDR for outlier removal when the neural network is generally well trained, and variance is less than 0.1, and stop when variance decreases to the second threshold V_t. We set up comparison trials to find the appropriate V_t.

V_t	Pattern removed	Accuracy	precision	recall	specificity	F1-score
0.01	41	82.05%	1.00	0.71	1.00	0.85
0.05	25	89.74%	0.97	0.82	0.97	0.89
0.08	15	86.46%	0.97	0.75	0.97	0.87

Table.1. tuning hyperparameter threshold Vt

The accuracy results presented in Table 1 reject the hypothesis that BDR can improve network performance by removing outliers, as none of the settings had an accuracy higher than the accuracy of the baseline model (91.25%). The best outlier removal was achieved when Vt=0.05, but it was still worse than the baseline model.

When Vt= 0.01, the accuracy decreases significantly compared to the baseline model and Vt=0.05 model. The reason is that when the threshold is too small, 41 out of 186 features are removed, and the removal ratio is too large, resulting in excessive removal. Also, from the results of the confusion matrix measurements, it can be seen that Precision reaches 1, the classification is perfect for Genuine anger, while recall decreases too much, reflecting the overfitting of the model. When the threshold is set large (Vt=0.08), accuracy is still low. This is because when the threshold is set too high, certain important features are removed as outliers, while the true outliers are not in this range.

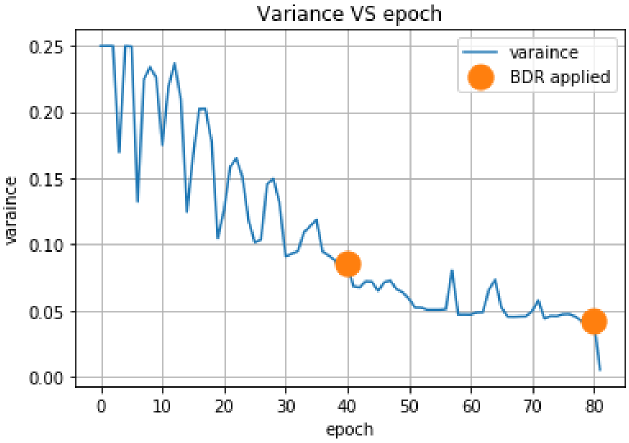

Figure.6. variance and BDR applied, terminate points

After tuning hyperparameters in BDR model, we further explore the functionality of the BDR model. BDR is applied from the 40th epoch, and terminates at the 80th epoch. We used the error distribution plot again to explore whether the BDR model effectively eliminates the bimodal distribution of errors.

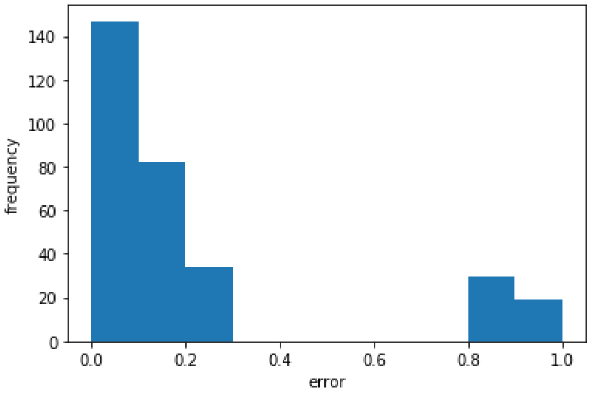

Figure.7. error distribution when epoch=40 without BDR

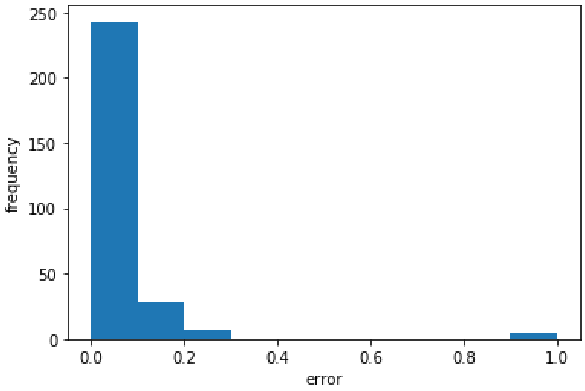

Figure.8. error distribution when epoch=80 with BDR

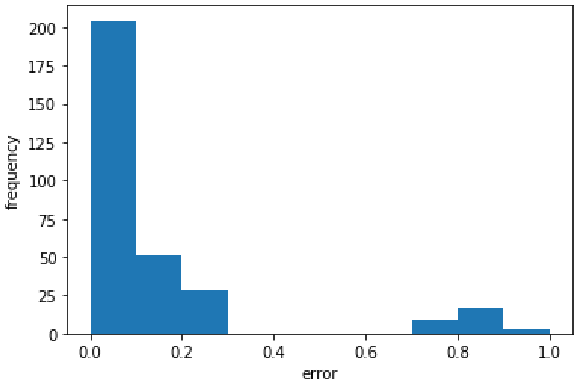

Figure.9. error distribution when epoch=80 without BDR

From the comparison of Figure9, Figure 8 with Figure 7, it can be seen that the bimodal error distribution is eliminated during the training process with or without BDR. However, it is obvious that the elimination effect is more obvious with the use of BDR in Figure 8 than without BDR in Figure 9. We conclude that the BDR is effective in detecting, determining, and removing outliers from the data, and achieve high performance in eliminating bimodal error distributions. However, BDR leads to a decrease in the accuracy of the model compare to the baseline model due to the fact that the BDR model is less inclusive of redundant attributes and removes them all as outliers. The excessive removal of BDR leads to a loss of features in the model.

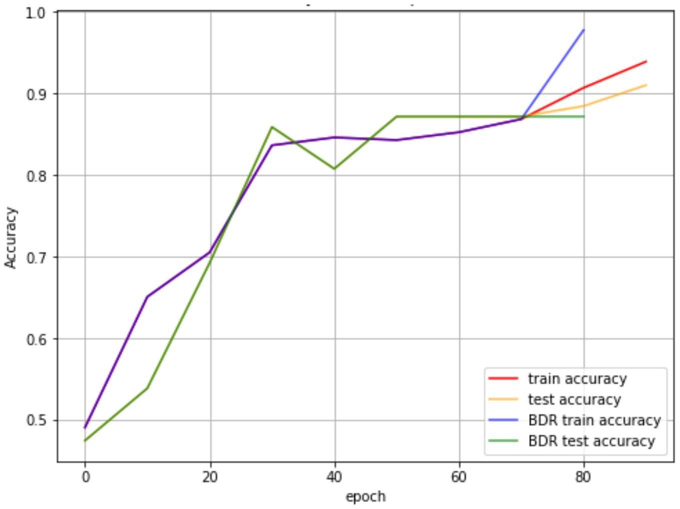

Figure.10. test set accuracy with BDR and baseline model.

Finally, we compare the results of the test dataset of the BDR with those of the baseline model. the training accuracy of the BDR model is improved significantly at the expense of the test set accuracy. This confirms our previous thesis that the BDR model over-fits the model by removing too many features.

3.3 GA-based Feature Selection

In order to achieve better feature selection and avoid killing too many important features, we use a GA-based feature selection method to train the neural network.[11] In the feature selection phase, we still divide the dataset into an 80% training set and a 20% test set. The test set is used to evaluate the fitness values.

According to the comparison test, it is concluded that the GA-based feature selection model runs best when the mutation rate is 0.05. When the mutation rate takes a larger value, it will destroy the original search mechanism, thus turning the GA into a random search [9]. However, a small mutation rate will make the algorithm's search in the population fall into a local optimal solution. We conclude that the mutation rate should be as large as possible without reducing the GA algorithm to a purely random search.[10]

Mutation rate	Pattern removed	Pattern Remain	Accuracy
0.2	96	90	93.20%
0.1	88	98	94.57%
0.05	85	101	95.23%
0.01	94	92	94.20%

Table.2. tuning hyperparameter mutation rate

As can be seen from the graphs, the GA-based algorithm is very effective in searching for the optimal solution to the problem, removing the redundant attributes and non-significant attributes. When mutation rate=0.05, the accuracy achieved is 95.23%, which is higher than the baseline model (91.25%) and the optimal BDR model (89.74%).

However, the high accuracy of the GA model also reflects the shortcomings in our experimental design. When the accuracy reaches 95.23%, there are 85 features removed out of 186 input features. When more than half of the input features are removed and higher accuracy is achieved instead, it indicates that our baseline neural network training is doing a lot of useless work and wasting a lot of computational cost to analyze the unimportant features. We should reflect on whether the input time series vector is reasonable. We analyzed the optimal solution provided by the GA-based feature selection model and tried to find the commonality of the removed feature columns. We found that a large fraction of the removed feature columns were concentrated in the tail of the data set, which is related to our data processing procedure. When there were only 100 pupil change records data in one time series and 186 records in another, we padded 86 zeros in the tail of the 100 short time data to make it a time series vector with 186 features. This resulted in a very large number of 0 data in the tail features, making these features redundant data. In addition, we found that there are many missing values in the removed features, which is also related to our use of zeros instead of missing values.

The flawed data processing method resulted in a large number of redundant features in our input data. The GA-based feature selection method can effectively identify the redundant features and remove them from the training set, thus improving the classification accuracy of the neural network.

4 conclusion and future work

In the paper, we used the simplest feedforward neural network to train time series data of individual pupil diameter variation and achieved 97.12% accuracy of training data and 91.25% accuracy of test data with only 100 epochs. Compared with the previous assignment1, which used the optimized overall pupil variability data, the final accuracy was only 90.31% in the training set and 73.75% accuracy in the test set after 1000 epochs training. The dataset Anger_V2 used in the article contains more data and is more informative. We believe that complex algorithms for neural networks are more suitable for handling large amounts of training data, while complex algorithms can lead to overfitting of the model when training small amounts of data like anger_V1. We believe that V2 achieves greater accuracy in the baseline model because the difference in the change in pupil diameter over time in the participating subjects is a better representation of changes in mood. High-quality datasets are the basis for training neural networks, and in the future, if we want to improve emotion recognition computation, we should first focus on the way data are collected and processed.

We used BDR for detecting and rejecting outliers. It is experimentally demonstrated that applying BDR can help the model to eliminate the erroneous bimodal distribution and successfully reject outliers. However, the BDR model had a negative effect, causing the accuracy of the model to drop to 89.74% and making the model more over-fitted. The reason is that the BDR model is insensitive to true outliers, redundant features, and features that require deep training, and will classify deep features as outliers and remove them. In summary, BDR has overkill effect. We chose variance as the criterion for BDR triggering and termination is not perfect, and the choice of threshold should be explored more. Future work can try more BDR triggering conditions, such as using cross-entropy error, to improve the situation of BDR over-deletion.

GA-based feature selection does improve the accuracy of classification (accuracy=95.23%) and removes the redundant input features, and successfully reduced the computational effort and computational cost. However, using the complex tournament selection model leads to a large amount of computation and makes the computation time longer, which makes the computation cost too high when applied to larger data. In future work, a balance between performance and cost should be found and the selection method should be optimized.

In the experiments, all the operations are costly because the data preprocessing method is not perfect. Using zero padding to process data of variable length leads to redundant attributes and reduces the flexibility of the dataset. For time series datasets with different lengths, we should find more suitable methods to process them and remove redundant attributes while preserving important features.

The experimental results we obtained are closer to those obtained in the paper 'Are you really angry'[1]. The accuracy of predicting genuine and posed anger using human physiological data in the article was 95%, and in our experiment, it reached 95.23%. In contrast, the experimenter's personal subjective judgment was only 65% correct. We conclude that the accuracy of the physiological signal response to pupil changes of the experimenter is much higher than the verbal judgment of the experimenter himself. In future research, we should focus more on the ability of neural network models to handle noise, because real-world environments are usually noisy.

References

1. Chen, L., Gedeon, T., Hossain, M.Z., Caldwell, S.: Are you really angry? detecting emotion veracity as a proposed tool for interaction. In: Proceedings of the 29th Australian Conference on Computer-Human Interaction. pp. 412–416 (2017)
2. Slade, P. and Gedeon, T.D., 1993, June. “Bimodal distribution removal.” In International Workshop on Artificial Neural Networks (pp. 249-254). Springer, Berlin, Heidelberg
3. De Jong, K. “Learning with Genetic Algorithms: An overview,” Machine Learning Vol. 3, Kluwer Academic publishers, 1988.
4. https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
5. Slade, P., Gedeon, T.D.: Bimodal distribution removal. In: International Workshop on Artificial Neural Networks. pp. 249–254. Springer (1993)
6. Yang, J., Soh, C.K.: Structural optimization by genetic algorithms with tournament selection. Journal of Computing in Civil Engineering 11(3), 195–200 (1997)
7. Leardi, R., Boggia, R., Terrile, M.: GENETIC ALGORITHMS AS A STRATEGY FOR FEATURE SELECTION. Journal of Chemometrics. Vol. 6, pp. 267-281(1992)
8. Milne, LK & Gedeon, Tom & Skidmore, AK. (1995). Classifying Dry Sclerophyll Forest from Augmented Satellite Data: Comparing Neural Network, Decision Tree & Maximum Likelihood.
9. Jonsson R,Malec J.Towards computing the parameters of the Simply Genetic Algorithms[C] ,Proceeding of the 2001 congress on Evolutionary Computation,2001:516-520
10. Thierens, D., Suykens, J., Vandewalle, J., De Moor, B.: Genetic weight optimization of a feedforward neural network controller. In: Artificial Neural Nets and Genetic Algorithms. pp. 658–663. Springer (1993)
11. Montana, D.J., Davis, L.: Training Feedforward Neural Networks Using Genetic Algorithms. Proceedings of the 11th International Joint Conference on Artificial intelligence - Volume 1 (1989)
12. Gen, M., Lin, L.: Genetic algorithms. Wiley Encyclopedia of Computer Science and Engineering pp. 1–15 (2007)
13. J. T. Townsend. (1971). "Theoretical analysis of an alphabetic confusion matrix." In: Attention, Perception, & Psychophysics, 9(1).