The Influence of Threshold Technology and Use Genetic Algorithm for Feature Selection on the Classification Effect of Single Hidden Layer Neural Network on Emotional Reactions Data to Videos of Depression

Yangfei Gao[1]

[1]Research School of Computer Science,
Australian National University, Canberra Australia
u7151160@anu.edu.au

Abstract. Questionnaire diagnosis of depression is very inaccurate because of the patient's sense of shame. Some physiological signals like pupils change, heart rate up and other physiological response of normal people when they see the depression patients that we can't control can be used as a more objective basis for judgment. We use these physiological signal changes of people as data. By choosing appropriate hyper-parameters, our neural network can stabilize the classification effect at 33.66%. We use threshold technology to adjust test results, and found this method can correct the abnormal classification value to a slight extent and improve the classification effect. The best improvement can reach to 9.09%. Because the data set with too many features is very small, we use genetic algorithm to find the most representative feature combination. When using different fitness for genetic algorithm, the classification accuracy of the network can reach 44.62% and 63.89% respectively, which is 1.87% and 21.14% higher than the original. This shows that in the case of small size samples with too many features, the reasonable use of genetic algorithm for features election can improve the classification effect of the network.

Keywords: Neural network, PCA, Threshold, Genetic algorithm, Feature selection

1 Introduction

As a mental illness, depression has a great impact on personal life. Most patients will suffer from low self-esteem, low desire and weight loss. Severe patients may even have suicidal tendencies. With the development of society and the acceleration of the pace of urban life, depression has gradually become an epidemic among working people. For the whole society, the impact of depression is gradually increasing.

The first step in preventing disease is diagnosis. At present, the difficulty of diagnosis of depression is that it requires patients to have a certain degree of consciousness. A common diagnosis of depression requires patients to fill out questionnaires. But patients may hide something when filling in the questionnaire out of shame, which affects the accuracy of diagnosis. So we want to use some data that don't lie to doctor to make a more objective diagnosis of depression. When some one see a face of depression, physiological signals might fluctuate correspondingly[1]. The datasets we used included three signals including Pupillary Dilation (PD)[2], Galvanic Skin Response (GSR)[3] and Skin Temperature[4]. These physiological signals have been shown to reflect emotional changes. At the same time, these signals are easy to be digitized, which is convenient for us to use neural network to study and predict.

This paper uses a simple single-layer hidden layer neural network, through the adjustment of hyperparameters, PCA[5] dimension reduction, threshold technology and genetic algorithm for feature selection to adjust the network and classification results. Our goal is to find the best combination of algorithms for classification of different level depression and observe the effect of threshold technology and genetic algorithm on neural network classification.

1.1 About Data Set

The data set we used was derived from an experiment on the rating of depression. The experiment involved 12 people watching videos of people with four levels of depression. Pupillary Dilation(PD), Galvanic Skin Response(GSR) and Skin Temperature of these people were recorded. The change of the three indicators will reflect the response of the subjects to patients with different degrees of depression to a certain extent. From these three basic data, we extracted a total of 85 features. We processed the GSR, PD, and ST data using normalization and filtering respectively, and then calculated the mean, variance, standard deviation, maximum, minimum, root mean square, means of the absolute values of the first and second difference of them. Then we apply a very low pass Butter worth filter with a cut-off frequency of 0.08 Hz to the normalized GSR to obtain a very low pass signal (VLP). We eliminate the piece wise-linear features of VLP and LP, the following seven characteristics are obtained. Numbers of SCR occurrences for VLP, LP and normalized signal, amplitudes of SCR occurrences for VLP, LP and normalized signal Ratio of SCR occurrences in VLP to occurrences in LP. We use the similar method to process LP and VLP of PD and ST, and get the corresponding seven features. In all, we can get a total of 85 features. Depression levels is divided into four, and our goal is to adjust the neural network to classify the data into four categories. Amount of data is 192. We randomly scrambled the data, with 80% as the training set and 20% as the test set

1.2 Data Preprocessing

Although our data is derived from three basic biological information, in order to avoid the influence of different feature measurement, we still normalize the data. We linearly transform the data for each feature by subtracting the mean and dividing by the standard deviation like Formula (1). In this way, the discrete degree of each feature can be controlled, and the gradient updating of the neural network is facilitated.

$$X_{normalization} = \frac{X - \mu}{\sigma} \tag{1}$$

1.3 Basic Structure of the Network

We use the most basic three-layer neural network, including the input layer, a hidden layer, and the output layer. The number of neurons in the input layer is the same as the number of features, the number of neurons in the hidden layer is obtained by empirical formula, and the number of output layer neurons is 4, which represents the four levels of depression.

1.4 Measurement of Classification Result

We think the percentage of all correctly classified data in the total data is a good indication of the quality of the classification results. Therefore, in my original model and improved model by threshold in section 2, we will use this criterion to explain and compare the classification results. In addition, we need to compare the classification results with the papers related to the data set in section 3. For the convenience of comparison, we will convert the original measurement results to facilitate the comparison with the original classification results. In the articles related to data sets, the author uses precision, recall and F1-score[6] to measure the accuracy of prediction. Among them, precision of depression level L represents the proportion of people who are actually depressed among those who are predicted to be depressed level L. Recall of depression level L represents the proportion of people on a depression scale of L who were successfully predicted to be on a scale of L. F1-score is defined by both precision and recall. The F1-score corresponding to the depression level L is obtained by twice the product of the corresponding precision and recall and then divided by the sum of these two indexes.

2 Possible Ways to Improve Neural Network Performance

2.1 Adjust Epoch and Add Dropout Layer

We started by running a 500 epoch on the network whose hidden neurons are 13, using the entire dataset. We found that with the increase of epoch, our training accuracy gradually increased, reaching more than 99% at 500 epoch, but the performance on the test set was very poor, the accuracy was generally only about 30%. This is an obvious overfitting phenomenon. We hope to suppress the over-fitting phenomenon by adjusting the epoch and adding the dropout layer. After we add dropout layer for neural network, we get the result showed below.

Table 1. With the Decrease of Epoch, the Classification Accuracy on Training and Test Set

Epoch	Train Data Accuracy	Test Data Accuracy
500	100.00%	27.27%
400	100.00%	23.68%
300	100.00%	37.25%
280	100.00%	35.00%
250	100.00%	35.29%
230	100.00%	43.33%
220	100.00%	33.33%
200	100.00%	40.82%
180	99.30%	30.00%
150	100.00%	38.46%
100	98.71%	32.43%
50	85.71%	26.32%

From the performance of the training set and the test set after changing the epoch, it is easy to see that in the Table 1 the best performing epoch in the test set is around 230. So we use 230 as the epoch of our neural network. But we will find another phenomenon, on the training set, our test accuracy is still very high. This shows that the phenomenon of over-fitting still exists even though we have added the dropout layer.

2.2 Selecting the Appropriate Number of Hidden Layer Neurons

In addition to reducing the epoch, we can also try to improve the performance of the neural network by adjusting the number of hidden layer neurons. In fact, the selection of hidden layer neurons is uncertain. Our usual method is to guess a possible number of neurons according to the performance of some possible hidden layer neurons in the number less than feature. In the original paper on the data set, the authors chose 50 as the number of neurons in the hidden layer, and we chose a wider range of neurons to model and compare the classification results.

Table 2. With the Change of Hidden Neurons the Classification Accuracy on Training and Test Set

Hidden Neurons	Train Data Accuracy	Test Data Accuracy
140	100.00%	33.33%
130	100.00%	33.33%
120	100.00%	27.78%
110	100.00%	38.89%
100	100.00%	30.56%
90	100.00%	36.11%
80	100.00%	30.56%
70	100.00%	36.11%
60	100.00%	33.33%
50	100.00%	26.19%
40	100.00%	27.91%
30	100.00%	29.73%
20	100.00%	27.78%

Table 2 shows the effect of different numbers of hidden neurons on the training result. We found that when the number of neurons in the hidden layer is between 70 and 110, the classification performance of the training set is slightly better, and when the number of neurons continues to increase, the classification performance of the training set is basically stable at about 33.33%. But in fact, compared with the performance of the neural network when the number of hidden layer neurons is 13 in Section 2.1, we can find that adjusting the number of hidden layer neurons does not improve the classification ability of the neural network significantly. But we will use 110 as the number of hidden neurons.

2.3 Dimension Reduction

Adjusting the number of epoch and hidden layer neurons did not improve our classification results. This prompts us to think about the possible problems of data itself. We have a total of 192 pieces of data, each of which has 85 dimensions. In fact, too many dimensions sometimes do not lead to better classification results. In small data sets, too many features may lead to over-fitting of the data to the training set, resulting in weak generalization ability of the classification model. So we want to use PCA[5] to reduce the dimensionality of the data.

We want the dimensionality reduced data to contain as much information of the original data as possible, so we calculate the proportion of the first K largest eigenvalues to all eigenvalues' sum. To some extent, this ratio represents the amount of information we keep in the original data in the low-dimensional space. Therefore, whether the proportion of K largest eigenvalues to all eigenvalues is higher than 0.99 is taken as the benchmark for selecting low-dimensional space dimensions.

PCA can reduce the data from 85 dimensions to 39 dimensions. The following is the classification performance after dimension reduction

Table 3. With the Change of Hidden Neurons the Classification Accuracy on Training and Test Set

	Not use PCA	After using PCA
Average Test Accuracy	33.66%	29.40%
Highest Test Accuracy	43.18%	37.50%

For the same training set and test set, the results of training with neural network may be a little different every time. This is because the initial parameters of the neural network are randomly generated, which may cause slight differences in the final results. In order to eliminate the influence of randomness, we train 100 neural networks on the same data set, record the best test set classification results and calculate the average performance of all networks in the test set.

From the chart, we can see that after using PCA to reduce the dimension of data, the overall classification effect has decreased by 4.26%, and the best classification effect has decreased by 5.68%. Although PCA theoretically preserves the original information of the data as much as possible, the performance of PCA is not good on the data set we used this time.

2.4 Threshold for Adjustment

In [7], by setting thresholds for different categories, the authors adjusted the proportion of false negative and false positive data in the test results to some extent. Although the original paper only uses threshold as a tool to ensure the same classification results while adjusting the false negative and false positive, we still record the changes of classification tests after using threshold method on the new data set.

Table 4. With the Change of Hidden Neurons the Classification Accuracy on Training and Test Set

	After using Threshold
Highest Training Accuracy up	6.08 %
Average Training Accuracy up	2.28 %
Highest Testing Accuracy up	9.09 %
Average Testing Accuracy up	0.16 %

We got some interesting results from Table 4. Here, we also train 100 neural networks on the same training set and test set, record the highest improvement of classification accuracy after using threshold, and calculate the average improvement of classification results. We find that, although the threshold is only to adjust the proportion of false negative and false positive predictive data, the classification ability of the network also has a slight improvement.

We show in the following Table 5 the results of the test set with the largest improvement when threshold is used.

Table 5. The Adjustment of Threshold Value to Classification Result

	Not use threshold			Use threshold		
Deprssion Level	correct	false +ve	false -ve	correct	false +ve	false -ve
None	6	9	4	6	9	4
Mild	4	9	9	4	7	9
Moderate	0	5	12	4	5	8
severe	3	8	8	3	6	6

From the graph, we can see that the number of false negatives and false positives becomes close. Before the threshold adjustment, the difference between the two is 12, and after the threshold adjustment, the difference is 10 (sum the absolute value of the number difference of false positives and false negatives of each degree), and more importantly, for the classification of Moderate level depression, it is changed from 0 to 4. This is actually very important, which tells us that the threshold method not only adjusts the proportion of false negative and false positive, but also has a certain degree of corrective function on abnormal classification data.

2.5 Genetic Algorithm for Feature Selection

In the case of only 192 samples, using 85 features to classify the samples is easy to appear over-fitting phenomenon. In the use of PCA dimensionality reduction, we found that the accuracy of classification has declined, which may be caused by the loss of information in the process of dimensionality reduction. However, considering that 85 features are calculated from three kinds of human information data, we believe that these features contain redundant information. Therefore, we try to use genetic algorithm for feature selection.

Genetic algorithm simulates the process of biological evolution. It encode the solution of the problem as biological chromosome, and generate multiple solutions by means of chromosome crossover and gene mutation simulation. In this process, we will calculate the suitability of each chromosome as a solution and use this suitability for presenting the probability of retaining it. Obviously, in this process, the appropriate solution is retained as the parent of a greater probability, so in general, the genetic algorithm will gradually close to the optimal solution which stands for the most representative combination of features.

We designed the genetic algorithm for this depression level classification problem. We use an array of 85 Boolean values as a chromosome, which can be used directly as a mask to select 85 features of the data. For each chromosome, we run the neural network five times, and use the average classification accuracy as the fitness of each chromosome. According to fitness, we select a new population with a certain probability of chromosome crossover and gene mutation. Crossover is the exchange of values at the corresponding positions of two parent chromosome arrays, and gene mutation indicates that the Boolean value of a certain position in the chromosome array may change.

We set the population size to 40, the crossover rate to 0.8, the mutation rate to 0.02, and the generation to 40, and run the program to get the following Figure 1. From the figure, we can see that after the 19th generation, fitness has remained basically unchanged. As a result of crossover and mutation, the curve begins to oscillate slightly. The highest classification accuracy is about 44.62%. The classification effect is improved by about 1.87%. The classification effect is only slightly improved.

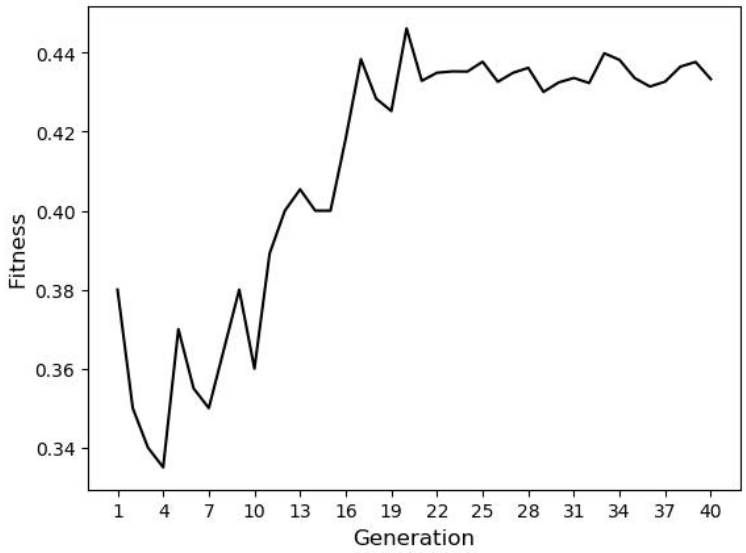

Fig. 1. This graph shows the highest fitness in each generation as the number of generations increases. Ordinate Fitness is the fitness of the best performing chromosome (feature combination) in each generation, and this fitness is also the best classification accuracy of the network under this feature combination. The curve began to oscillate slightly after the 19nd generation, and the overall trend tended to be stable.

Try adjusting the fitness. For the feature combination corresponding to each chromosome, we select the best classification result instead of average classification result among the five neural network classification results as the fitness of the chromosome, and the result after 40 generations is shown in Figure 2

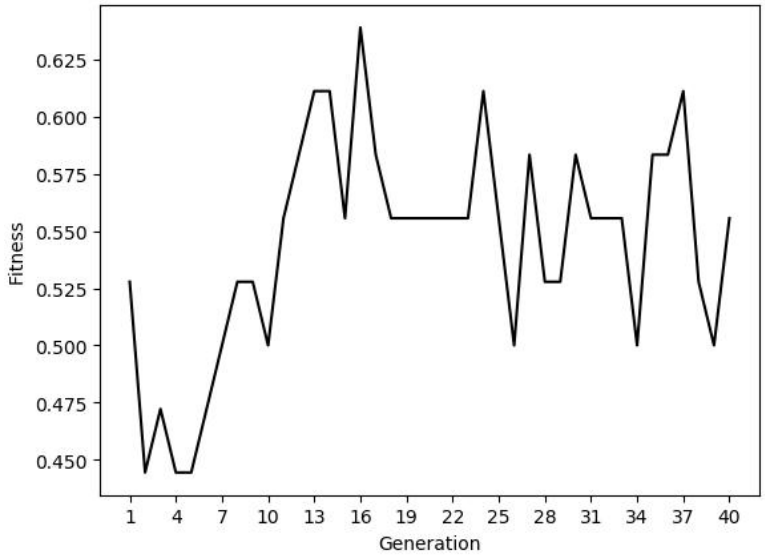

Fig. 2. This graph shows the highest fitness in each generation as the number of generations increases. Ordinate Fitness is the fitness of the best performing chromosome (feature combination) in each generation, and this fitness is also the best classification accuracy of the network under this feature combination. After changing the fitness, the oscillation amplitude of the curve increased significantly, but the overall classification effect became better.

Observing Figure 2, we find that the oscillations of the curve increase significantly because we always pick the best fitness for each feature combination rather than the average fitness. This makes the performance of each selected chromosome combination more

uncertain. But in the genetic algorithm, the uncertainty may find a better solution space. In fact, after replacing fitness, the classification effect of our network can reach 63.89% at the highest. The highest classification accuracy increased by 21.14%.

3 Results, Comparison and Discussion

In this section, we no longer use the overall prediction accuracy of each level of depression to measure the classification effect of the network. For the credibility of the comparison, we calculate precision, recall and F1score according to the confusion matrix to facilitate the comparison with the prediction effect in the paper of the data set source. In Table 6, P, R and F stand for precision, recall and F1 score respectively.

Table 6. Classification Effect of Different Model

Depression Level	NN designed by ourselves									NN designed by data set author[6]					
	NN			NN+ threshold			NN+ threshold+GA			NN			GA +NN		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
None	0.60	0.60	0.60	0.60	0.60	0.60	0.33	0.67	0.44	0.85	0.90	0.87	0.92	0.95	0.94
Mild	0.85	0.31	0.45	0.81	0.31	0.45	0.91	0.64	0.75	0.85	0.85	0.85	0.93	0.89	0.91
Moderate	0.60	0.00	0.00	0.78	0.33	0.46	0.70	0.60	0.65	0.92	0.88	0.90	0.88	0.90	0.89
severe	1.00	0.33	0.50	1.00	0.33	0.50	1.00	0.67	0.80	0.91	0.89	0.90	0.95	0.95	0.95
Average	0.76	0.31	0.39	0.80	0.39	0.52	0.74	0.65	0.66	0.88	0.88	0.88	0.92	0.92	0.92
Overall Accuracy	0.49			0.57			0.68			0.88			0.92		

Observing the data in Table 6, we found that the neural network we used was sensitive in detecting whether patients had depression or not. We can get this conclusion from the measurement standard of precision whose average accuracy is 76%. However, in the overall classification effect, the performance of our neural network is far inferior to that of the neural network in the relevant papers of the data set. Interestingly, in Section 2.4, we found that the threshold improved the classification accuracy of the model, which was also reflected in the new measurement standard. When we use threshold to adjust the prediction results, the improvement in precision, recall and F1score is 4%, 8% and 13%, respectively. This means that threshold does adjust some abnormal classification data while adjusting false negative and false positive, and the adjustment is most obvious in the classification of moderate level. In general, threshold improves the classification results by 8%. When using the genetic algorithm to find the appropriate combination of features, we find that recall increases and precision decreases. This means that after using the genetic algorithm, our network becomes more accurate in determining specific levels of depression, but slightly less accurate in determining whether a patient is depressed. Overall, the classification accuracy of the genetic algorithm is improved by 11% for the neural network using the threshold technique.

By comparison, we speculate that there are some internal and external factors that can not be eliminated in the detection and recording of physiological changes of subjects. Internal factors may be related to the emotional fluctuations of the subjects at a certain time. We can't ask humans to be as mechanical and stable as robots. External factors may include errors in measuring instruments. When these factors exist in the data, we can use the threshold method to adjust the classification outliers caused by these internal and external factors. When the amount of data is too small, too many features may affect the accuracy of classification. At this time, we can use genetic algorithm to select the most appropriate combination of features to improve the classification effect of the network. In addition, we use different fitness when we use genetic algorithm, and the results are very different. The use of fitness with more uncertainty has greatly improved the overall classification effect, which is a problem worthy of further study.

4 Conclusion and Future Work

In Section 2.4 and Section 3, we use different criteria to measure the classification results of our single-layer neural network, our classification accuracy can reach 68% at the highest.

Although epoch and hidden layer neurons have little effect on the classification effect of the model, we still compare the performance of the network in a wide range of values, and try to stabilize the network' performance at more than 30%. After the application of PCA dimension reduction, the classification effect of the network is reduced, which shows that the 85 features might all basically contribute to our classification, and the dimension reduction may lose some information. Then we use genetic algorithm to select the best combination of features and found that the effect of classification is improved by about 1.87%. Our experiments show that the threshold technology can indeed adjust the classification outliers to a certain extent, and the genetic technology can improve the classification accuracy by selecting the best feature combination in the case of fewer samples.

In fact, the biggest problem our neural network encounters is the amount of data. We have only 192 pieces of data available for 85 features. After setting the test set and the training set in a ratio of 2 to 8, the data we can use for training is reduced again. Compared with the method in the original paper of the data set, there are three directions for us to improve. Firstly, expanse data sets. Secondly, the comparison of the network using the improved k-fold cross-validation method. The author believes that the same person's response to patients with different degrees of depression may have a certain continuity, so he uses all the response data of a person for

testing. This is an improvement method worth trying. Thirdly, in the process of using genetic algorithm, the uncertainty of each step may directly affect the final classification results, how to use these uncertainties to make the final effect best is also a problem that can be studied in the future work.

References

1. Potvin,S.,Charbonneau,G.,Juster,R.P.,Purdon,S.,Tourjman,S.V.:Self-evaluation and objective assessment of cognition in major depression and attention deficit disorder: Implications for clinical practice. Compr. Psychiatry, vol. 70, pp. 53 – 64(2016)
2. Scherer, S.: Automatic behavior descriptors for psychological disorder analysis. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, pp. 1–8(2013)
3. Chen, Y.-T., Hung, I.-C., Huang, M.-W., Hou, C.-J., Cheng, K.-S.: Physiological signal analysis for patients with depression," in Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on, 2011, vol. 2, pp. 805–808.
4. Jain, F. A.: Heart rate variability and treatment outcome in major depression: a pilot study. Int. J. Psychophysiol., vol. 93, no. 2, pp. 204–210(2014)
5. Jolliffe, I. T.: Principal Component Analysis. Springer Series in Statistics. New York.(2002)
6. Zhu, X., Gedeon, T., Caldwell, S., Jones, R.: Detecting emotional reactions to videos of depression. 2019 IEEE pp.147-152(2019)
7. Milne, L.K., Gedeon, T.D., Skidmore, A.K.: Classifying Dry Sclerophyll Forest from Augmented Satellite Data: Comparing Neural Network, Decision Tree & Maximum Likelihood. Training(1995)