Recognizing Depression from Physiological Features based on GIS and Genetic Algorithm

Siyu Fang[1]

[1] Research School of Computer Science, Australian National University, Canberra Australia
u7153099@anu.edu.au

Abstract. We can recognize people suffering from depression and neural networks (NNs) also have this ability by analyzing people's physiological features when they observe depressed individuals. 12 people's physiological features, from the Depression Recognition Challenge (AVEC 2014) dataset, were used to train and test our neural network models. One of our models, without the GIS technique, could reach 35% for overall accuracy. Then, we compared performances of models with and without the GIS technique, and found the technique made overall accuracy worse. Besides, we found that for the model without GIS, this model could not generalize well on the testing set. To alleviate overfitting, we utilized genetic algorithm (GA) to select an optimal subset of physiological features and overall accuracy reached 47%. Eventually, after optimizing features, we also employed GA to optimize learning rates and final overall accuracy reached 49%, but this value was less than that of another research paper.

Keywords: Depression Detection, Neural Networks, Learning Rate, GIS Technique, Genetic Algorithm

1 Introduction

Depression is a type of mood disorder, and it is common among adults, particularly those who frequently suffer from high pressure. It is sometimes related to feelings such as sadness, anger, loss of interest and passion. Depression can affect some chronic health conditions such as cardiovascular disease and cancer [1]. These conditions may be worse when patients gradually become depressed. In consequence, it is crucial to effectively diagnose and cure people who are negatively influenced by depression so that patients can have a high-quality life.

Normally, diagnosis of depressed patients is mainly based on some questionnaires which are often filled in by these patients [2]. To be honest, this process is not objective and sometimes wastes a large amount of valuable time. Furthermore, as depressed patients tend to take a negative attitude towards their daily life, they are unwilling to reveal their real mental states [3]. Consequently, current methods of diagnosis are not convincing enough most of the time and it is significant to find a more objective way of diagnosing depression.

Recent advances made it possible to diagnose depression based on some physiological features of individuals such as Galvanic Skin Response (GSR) [4]. These features can be more measurable and quantitative than people's mental states or feelings. If neural networks are applied on these physiological features, a more objective diagnosis of depression could be obtained.

Our goal here is to employ some features of observers to recognize levels of depression of real patients whom these observers watch. We used NNs to identify levels of depression based on observers' response to some other people's depression. Besides, we combined NNs with some other techniques, such as GIS [5] and compared performances of NNs with and without the GIS technique. In addition, to alleviate the issue of overfitting, we utilized some algorithms, such as GA [6], to choose which features could assist the neural network model in obtaining better generalization performances. Eventually, as the learning rate is an extremely crucial hyperparameter when configuring our neural network models [7], it is necessary to optimize the learning rate of our model which has been optimized in terms of input features.

2 Method

2.1 The Dataset

The dataset we employed in this paper is from the Depression Recognition Challenge (AVEC 2014) dataset [8]. The dataset was from 12 participants including 6 males and 6 females. The range of participants' age is from 18 to 27 with the average age being 21.1 and the standard deviation being 2.8. These people have the almost normal vision and hearing [4].

The dataset we used contains 192 data points and 85 features in total which can be divided into three parts. These three parts are galvanic skin response (GSR), skin temperature (ST) and pupillary dilation (PD) [8]. Moreover, the three parts include 23, 39 and 23 features respectively. To be specific, GSR evaluates a person's electricity flow through this person's skin and this value is affected by how much sweat a person has on the skin [9]. ST is negatively influenced by some bad emotions of people, for example if people feel stressful, values of their ST are more likely to get lower than the values when they feel delighted [10]. PD is the representation of people's mental activities [11]. Furthermore, there are four levels of depression, including "None", "Mild", "Moderate" and "Severe". "None" denotes no or minimal depression; "Mild", "Moderate" and "Severe" indicate that depressed individuals suffer from mild, moderate and severe levels of depression respectively. In the dataset, the distribution of "None", "Mild", "Moderate" and "Severe" was shown in Fig. 1.

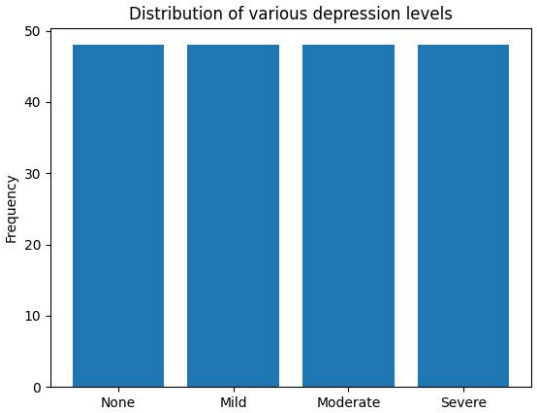

Fig. 1. Distribution of different levels of depression in the dataset

From Fig. 1, we can find that the distribution of various depression levels is well-balanced and each level has 48 data points. That means each level of depression has the same distribution.

Apart from depression levels, features of GSR include minimum, maximum, mean normalized GSR as well as some filtered GSR. Features of ST include minimum, maximum, mean normalized skin temperature and several filtered temperatures. For PD, some of features can be divided into left and right two parts.

2.2 Preprocessing

When we look through the dataset used in this paper, the range of different features is so wide that if we do not perform normalization on these features, results of our models are less likely to be robust [12]. On the dataset, we can find that the maximum value for features can reach 65, but the minimum value closes to zero. Apart from that reason, normalization can also assist our models in speeding up training [12], which saves a larger amount of time. Consequently, we normalized all features employed and to find which method of normalization could have better performances, we tested two ways of normalization. One was that values of a feature were firstly subtracted by the mean of the values and then divided by the variance. This way makes the mean close to zero. Another one was that values of a feature were firstly subtracted by the minimum value and then divided by the range of original values. After testing many times, we found that the second method could assist models in generalizing well on the testing set. Consequently, we eventually adopted the second way of normalization.

The whole dataset was split into three parts. The first part was used as the training set, and it accounted for 70% of the whole dataset. Then, 10% data belonged to validation set and the remaining 20% was assigned to the testing set. Training set and validation set were used when we utilized GA to optimize the input features and learning rates. Besides, training set and testing set were used when we train and test our neural network models.

2.3 Neural Networks

As we have few data points, the network topology needs a small number of parameters in order to avoid overfitting. Furthermore, the network topology is a fully connected network since this network has been applied to identify depression levels [4]. We confirmed the numbers of input neurons and output neurons. Then we needed to decide how many hidden layers we required and for each hidden layer, we should also determine the number of hidden layer neurons. To confirm these hyperparameters, we tested some different combinations and finally we realized that due to small number data points, we should make the number of hidden layers to the minimum value. Therefore, the neural network model we employed for this task is a three-layer network with a 4 nodes output layer which represents four different levels of depression. Apart from that, we also tested various numbers of hidden nodes from 10 to 100 with every 10 nodes. After that, we found that when the number of hidden nodes was 60, overall accuracy on the test dataset was optimal. In consequence, we selected 60 as the number of hidden nodes.

Another significant hyperparameter was the learning rate of the network model. This hyperparameter is so crucial that it greatly influences the generalization performance [7]. Consequently, based on my experience, I firstly chose a common value (0.01) for the learning rate of our models. Then, we employed GA to optimize this hyperparameter to obtain an optimal value so that our models can have better overall accuracy. Eventually, after tests were taken, 0.027 was the best value for the learning rate of our models.

For the optimizer of this network, we tested several types of optimizers with various parameters respectively, such as Adadelta [13], Adam [14] and Adagrad [15]. Parameters included different values of weight decay. After testing these optimizers, we realized that when we employed Adam as the optimizer and weight decay was set to zero, the average performance was optimal on the testing set. For this reason, we chose Adam as the optimizer of our models.

In order to avoid overfitting, apart from selecting simple model, we also tested some other methods, such as early stopping, using regularization and dropouts. After tests of some combinations of the above methods were taken, we found that overall accuracy was optimal when we utilized early stopping and made the network randomly drop features which came into the hidden layer with a 50% probability in each training epoch.

2.4 Evaluation Measures

We employed precision, recall and F1-score as ways of measuring performances for network models. For a level, the precision is the number of correct predictions divided by the number of predictions for this level; the recall means that, for this level, the number of correct predictions divided by the number of data points whose targets are this level; the F1-score is calculated by $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$ [4].

As a multiclass classification task, we also averaged evaluation measures so that we could have a better generalization performance. In addition, we calculated overall accuracy in order to measure the whole performance. Overall accuracy is computed by the sum of correct predictions for all levels divided by the number of data points [4].

2.5 The GIS Technique

According to [5], we know that, for a classification task, if we change the values of threshold, it is likely to adjust the numbers of false positive and false negative classifications so that we may adjust values of precision and recall to obtain a better overall performance. The paper [5] applied this technique on the binary classification tasks. In this paper, the GIS technique was used for the multiclass classification task.

For this task, there are four different levels of depression; hence, we added a fixed value to the probability of every level of the task. To find optimal values of thresholds, we varied each threshold from 0.4 to 0.7, and tested values on both the training set and testing set.

2.6 Genetic Algorithm

A typical GA contains five phases including initialization, selection, crossover, mutation as well as calculating fitness function [16]. We used GA to choose which combination of the original 85 features from GSR, ST and PD could get a better performance for overall accuracy. After optimizing input features, GA also assisted our model in selecting an optimal learning rate.

For the initialization phase, we normally initialize a population of randomly generated individuals. To be specific, when GA was applied to features, the length of every chromosome was set to 85 which indicated the number of features the dataset had. The order of features in a chromosome was the same as that in the training set. Besides, each chromosome was represented in binary as a list of "False" and "True" [17]. We utilized "True" to indicate that a chromosome used the corresponding feature and "False" to represent that a chromosome did not contain the feature. In addition, the initial population employed different combinations of the 85 features which means the corresponding lists are likely different from each other. Moreover, when we used GA to select an optimal learning rate, the value of the learning rate for neural networks is normally set to between 0.0 and 1.0 [7]. In consequence, the length of every chromosome was set to 14 so that we could partition 1.0 into 2^{14} parts. As a result, the resolution could be lower than 0.0001 and we were likely to obtain an appropriate value. In this case, each chromosome was represented in binary as a list of 0s and 1s [17]. Specifically, for a chromosome "00000010101011", if we convert it from binary counting into decimal counting, the result is $(1 / (2^{14}) * 171) \approx 0.0104$.

In GA, selection phase includes a few methods such as the roulette wheel selection, tournament selection and elitism selection [18]. In this paper, we tested the roulette wheel selection and elitism selection. When we compared the average performances of these two selection methods, we found that the elitism selection could assist our models in obtaining better results. In consequence, we chose the elitism selection for optimizing both input features and learning rates. Furthermore, we set overall accuracy on the testing set as the fitness function of GA. All parameters of GA are shown in TABLE 1 [4].

Table 1. Parameters of GA

GA parameter	Value	
	GA for features	GA for learning rates
Length of chromosome	85	14
Population size	20	20
Crossover rate	0.8	0.8
Mutation rate	1 / (length of chromosome)	1 / (length of chromosome)
Selection type	Stochastic uniform selection	Stochastic uniform selection
Mutation type	Uniform mutation	Uniform mutation
Crossover type	Uniform crossover	Uniform crossover

3 Results and Discussion

Based on the AVEC 2014 dataset, two models were tested; one is the fully connected network without the GIS technique and the other one is a network with the technique. Features which were used for testing were the same as 85 features of GSR, ST as well as PD. Each model was run 10 times and we averaged results of 10 times. The average performances on the training set and testing set are shown in TABLE 2 and TABLE 3.

Table 2. Performance measures for depression detection on the training set with and without GIS

Depression level	NN			NN+GIS		
	Precision	Recall	F1 score	Precision	Recall	F1 score
None	0.87	0.88	0.87	0.87	0.91	0.89
Mild	0.87	0.85	0.86	0.92	0.89	0.90
Moderate	0.87	0.88	0.87	0.91	0.87	0.89
Severe	0.87	0.86	0.86	0.87	0.91	0.89
Average	0.87	0.87	0.87	0.89	0.89	0.89
Overall Accuracy	0.87			0.89		

Table 3. Performance measures for depression detection on the testing set with and without GIS

Depression level	NN			NN+GIS		
	Precision	Recall	F1 score	Precision	Recall	F1 score
None	0.35	0.35	0.35	0.26	0.08	0.12
Mild	0.40	0.24	0.30	0.26	0.08	0.12
Moderate	0.27	0.24	0.25	0.28	0.49	0.36
Severe	0.39	0.59	0.47	0.26	0.44	0.32
Average	0.35	0.35	0.35	0.26	0.27	0.27
Overall Accuracy	0.35			0.27		

From TABLE 2, for the training set, when the GIS technique was applied to the NN, changes in the average performances of the precision, recall and the F1 score were moderately different among four depression levels. To be specific, for the "None" level, the technique had more influence on the average recall than the precision and the F1 score by increasing approximately 3%. There was also a slight increase in the F1 score. By contrast, the change in the average precision was insignificant between two models. The similar trend could be seen in the "Severe" level. The only difference was that the amount of increase was more. However, the average precision increased with 5% and 4% in the "Mild" and "Moderate" levels respectively. In summary, the average precision, recall and the F1 score were augmented after applied the GIS technique on the training set.

From TABLE 3, compared with the training set, there was an opposite trend for the testing set. There were different degrees of reduction among the three evaluation measures after the GIS technique was applied. Specifically, the network without the GIS had higher values in the average precision, recall and the F1 score than the one with the technique. Unlike the training set, the "None", "Mild" and "Severe" levels had the almost same trend except the degrees of reduction. However, almost all three measures experienced significant increase in the "Moderate" level. Furthermore, the average recall increased approximately 25% in the "Moderate" level. In summary, the GIS technique could not optimize the performances of this classification task.

Compared with performances of the training set, the results on the testing set drop significantly which indicates that overfitting existed. After analyzing the dataset, we found one problem could be that in comparison with the number of features used for this task, the number of data points was relatively few. However, it is difficult for us to augment the number of data points. For this reason, we needed to remove some features which were likely to contain redundant information. Then, we should decide which features had to be removed. To solve this problem, we employed GA since this algorithm has been utilized to choose features [10].

Over some iterations, GA found an optimal solution and the comparison among the NN without GIS, the result derived from NN+GA and the result from [4] is shown in TABLE 4.

Table 4. Performance measures for depression detection among three models

Depression level	NN			NN+GA			NN+GA (reference)		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
None	0.35	0.35	0.35	0.43	0.20	0.27	0.92	0.95	0.94
Mild	0.40	0.24	0.30	0.25	0.73	0.37	0.93	0.89	0.91
Moderate	0.27	0.24	0.25	0.44	0.77	0.56	0.88	0.90	0.89
Severe	0.39	0.59	0.47	0.90	0.09	0.17	0.95	0.95	0.95
Average	0.35	0.35	0.35	0.51	0.45	0.47	0.92	0.92	0.92
Overall Accuracy	0.35			0.47			0.92		

From TABLE 4, when GA was applied to the NN, changes in the average performances of the precision, recall and the F1 score were various among four depression levels. Specifically, for the "None" level, after applied GA, the average recall and the F1 score moderately decreased with 15% and 8% respectively, but the precision experienced a slight increase at the same time. In addition, the "Severe" level had witnessed a similar trend and the only difference was that the amount of change in the "Severe" level was more than that in the "None" level. By contrast, the opposite trend could be seen in the "Mild" level. However, all three evaluation measures of the "Moderate" level significantly augmented after combined with GA. In summary, overall accuracy considerably increased from 35% to 47% by employing GA and this algorithm truly optimized the performances of our models.

However, compared with the result from [4], almost all evaluation measures of the NN+GA model were relatively low. The closest values were the average precision of the "Severe" level with 95% and 90% respectively. After analyzing the results, we realized that one reason why difference between the two results was obvious could be that features selected by our GA were not an optimal subset of the original 85 features. Hence, our next work includes finding a more appropriate combination of the 85 features.

After we confirmed the numbers of input neurons, hidden neurons and output neurons, another vital hyperparameter, which should be optimized, is the learning rate of our models. To achieve this target, we also adopted GA to find a relatively optimal value for the learning rate. Over some iterations, GA found an optimal solution and the comparison among the NN with the learning rate being 0.01, the result derived from NN+GA and the result from [4] is shown in TABLE 5.

Table 5. Performance measures for depression detection among three models

Depression level	NN+GA (Learning rate=0.01)			NN+GA (Learning rate=0.027)			NN+GA (reference)		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
None	0.43	0.20	0.27	0.55	0.38	0.45	0.92	0.95	0.94
Mild	0.25	0.73	0.37	0.51	0.35	0.41	0.93	0.89	0.91
Moderate	0.44	0.77	0.56	0.43	0.63	0.51	0.88	0.90	0.89
Severe	0.90	0.09	0.17	0.51	0.62	0.57	0.95	0.95	0.95
Average	0.51	0.45	0.47	0.50	0.50	0.49	0.92	0.92	0.92
Overall Accuracy	0.47			0.49			0.92		

From TABLE 5, when the new learning rate was applied to our model, changes in the average performances of the precision, recall and the F1 score were different among four depression levels. To be specific, for the "None" level, after updated the learning rate, the average recall and the F1 score significantly decreased with 26% and 9% respectively, but the precision witnessed a slight increase at the same time. Further, the "Mild" level had experienced a similar trend and the only difference was that the amount of change in the "Severe" level was slightly less than that in the "None" level. In comparison to the above two levels, the opposite trend could be seen in the "Moderate" level. However, all three evaluation measures of the "Severe" level dramatically augmented after optimizing the learning rate. In summary, overall accuracy slightly increased from 47% to 49% by employing the new learning rate and the change of this hyperparameter truly made an impact on the performances of our model.

Nevertheless, in comparison with the result from [4], all evaluation measures of the NN+GA model were relatively low. The closest values were the average recall of the "Moderate" level with 63% and 90% respectively. After analyzing the results, we realized that apart from the reason that features selected by our GA were not an optimal subset of the original 85 features, another reason could be the value of the learning rate (0.027) was a local minimum instead of the global minimum. Consequently, our next work includes finding a more appropriate learning rate for our neural network model.

4 Conclusion and Future Work

We have employed physiological features from observers to discern various levels of depression on the training and testing sets. We applied neural networks with and without the GIS technique, as well as network with GA to the features and the learning rate. The network with the GIS produced the worst overall performances on the testing set. By contrast, after input neurons were optimized by GA, overall accuracy of the network model could reach up to 47% which is better than the result without GA. It proved that the impact of overfitting could be alleviated by utilizing GA to remove some redundant features. Besides, we also realized that our models could not benefit from the GIS technique. Eventually, after we updated the learning rate selected by GA, overall accuracy of the network model could reach up to 49% which is slightly better than the result with the learning rate chosen from my experience. However, this result was still lower than that of the paper [4].

The next stage of our work will record more observers' response to others who suffer from various levels of depression. Moreover, we may employ some other physiological features of observers to detect depression. Eventually, some hyperparameters including the learning rate will also be optimized by techniques.

References

1. Everything You Want to Know About Depression, https://www.healthline.com/health/depression
2. Beck, A.T., Steer, R.A., Brown, G.: Beck depression inventory–II. Psychological Assessment. 78, 490--498 (1996)
3. Josh, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G., Breakspear, M.: Multimodal Assistive Technologies for Depression Diagnosis and Monitoring. J. Multimodal. User. Interfaces. 7, 217--228 (2013)
4. Zhu, X., Gedeon, T., Caldwell, S., Jones, R.: Detecting emotional reactions to videos of depression. In: 2019 IEEE 23rd International Conference on Intelligent Engineering Systems, pp. 147--152. IEEE Press, Hungary (2019)
5. Milne, L.K., Gedeon, T.D., Skidmore, A.K.: Classifying dry sclerophyll forest from augmented satellite data: comparing neural network, decision tree & maximum likelihood. In: 6th Australian Conference on Neural Networks, ACNN'95, pp. 160--163. (1995)
6. Goldberg, D.E., Holland, J.H.: Genetic Algorithms and Machine Learning. Machine Learning. 3, 95--99 (1988)
7. Park, J., Yi, D., Ji, S.: A Novel Learning Rate Schedule in Optimization for Neural Networks and It's Convergence. J. Symmetry. 12, 660--675 (2020)
8. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In: 4th International Workshop on Audio/Visual Emotion Challenge, pp. 3--10, Association for Computing Machinery, New York (2014)
9. Stern, R.M., Ray, W.J., Quigley, K.S.: Psychophysiological Recording. Oxford University Press, USA (2001)
10. Sharma, N., Gedeon, T.: Modeling stress recognition in typical virtual environments. In: 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, pp. 17--24. IEEE Press, Venice (2013)
11. Laeng, B., Sirois, S., Gredebäck, G.: Pupillometry: a window to the preconscious?. Perspect. Psychol. Sci. 7, 18--27 (2012)
12. Why Data should be Normalized before Training a Neural Network, https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d
13. Zeiler, M.D.: ADADELTA: An Adaptive Learning Rate Method. ArXiv abs/1212.5701 (2012)
14. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. ArXiv abs/1412.6980 (2014)

15. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12, 2121--2159 (2011)
16. Evolutionary optimization: A review and implementation of several algorithms, https://www.strong.io/blog/evolutionary-optimization
17. Introduction to Genetic Algorithms — Including Example Code, https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3
18. Katoch, J., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. J. Multimed. Tools. Appl. 80, 8091--8126 (2021)