Neural Networks with different initialization methods for depression detection

Tianle Yang[1]

[1] School of Computing in the College of Engineering and Computer Science, Australian National University, Australia

u6512077@anu.edu.au

Abstract. As a kind of a psychological mood disorder, depression is an important factor leading to suicide. Therefore, there is an urgent need for an effective way to diagnose depression for early treatment. Traditional methods not only cost a lot of human and financial resources, but also have the phenomenon of misdiagnosis. As a powerful tool, neural network can mine the inner information of data and find out the relationship between data. Since some studies have shown that some physical characteristics play an important role in the diagnosis of depression, which inspired us to predict depression through the construction of a neural network model based on physical characteristics. This paper uses the neural network models with two initialization methods which are Xavier initialization and Kaiming initialization. We found 3-layers neural network with Kaiming initialization achieved the highest accuracy which is 83%.

Keywords: depression, Neural Network, Xavier initialization, Kaiming initialization

1 Introduction

Clinical depression is a psychotic mood disorder, which is caused by the individual's difficulty coping with stressful life events. The patients will show constant sadness and negativity [2]. Depression is considered as one of the risk factors for suicide [1]. The World Health Organization (WHO) ranks it as the fourth leading cause of disability in the world and predicts that it will become the second leading cause of disability by 2030[3].

As depression has become more common in the general population [4] and a major burden for the health-care system worldwide [6]. Effective depression diagnosis and treatment techniques attracted great attention. However, the diagnosis of depression is very difficult. The diagnosis of depression is mostly given by general practitioners (GPs), although they can rule out depression in most people who are not depressed; the modest prevalence of depression in primary care means that misidentifications outnumber missed cases [6].

Nowadays, diagnosis based on the observed behavioral signals or some physiological indicators is gaining in popularity [7]. The study of objective biological, physiological and behavioral markers can not only improve the accuracy of psychological diagnosis and treatment of many mental diseases, but also help reduce the huge social and economic costs associated with these diseases [8].

As a data analysis tool, machine learning has been continuously developed in recent years, and the prediction of depression based on machine learning has become possible. As a common method in data mining tools, decision tree classifiers are already used in the discovery for the diagnosis of type II diabetes [12]. Also, Ahmad[11] built decision tree classifiers to make breast cancer diagnosis. As a

powerful tool of machine learning, neural network has already been used in the field of predicting depression [9].

In this work, we use observed behavioral signals and some physiological indicators to train depression prediction models, including neural network and tree-based models. The dataset and data preprocessing methods we used are both from [9]. The final subjects of this paper are 12 students composed of 6 men and 6 women, aged between 18 and 27, without prior knowledge of depression. The experiment collected students' physical sensors records, including galvanic skin response (GSR), skin temperature (ST) and pupillary dilation (PD) while they watching videos. The galvanic skin response (GSR) results from the spontaneous activation of sweat glands. Hand and foot sweating are caused by emotional stimuli: whenever we are emotional, GSR data show unique patterns visible to the naked eye and can be quantified statistically. Skin temperature (ST) is the temperature of the outermost layer of the body, [14] shows that skin temperature reveals the intensity of acute stress. Pupillary dilation (PD) provides signs of changes in mental state and mental activity intensity, and pupil size is found to constitute a response to emotional participation stimuli. With data preprocessing method mentioned in [9], we have obtained 23 Galvanic Skin Response (GSR) features, 39 Pupillary Dilation (PD) features and 23 Skin Temperature (ST), totaling 85 features.

2 Method

In this paper, we use neural network models with different layers and focus on the effect of initialization methods. We use two initialization methods: Xavier initialization [13] and Kaiming initialization [10]. We use these two initialization methods to build the neural network model with the same structure and choose the optimal hyper-parameters suitable for the current model through experiments.

2.1 Neural Network

Artificial neural network can be composed of several layers, each layer of the neural network consists of multiply neurons, which can be seen as composed of three elements: connection weight, adder and activation function. The connection weight is represented by the weight on each connection. The adder is used to sum the corresponding synapse weights of the input signal to the neuron, and the activation function limits the input signal to a certain value within the allowable range.

We design three kinds of neural network architectures, which are single-layer neural network, 2-layer and 3-layer neural networks. For 2-layer neural network, the number of hidden layers we use is 50 which is the same as[9]. For 3-layer neural network, the number of first hidden layers is 50 and the number of second hidden layer is 20.

In this paper, we mainly focus on how different initialization methods influence the neural networks. For all models, the learning strategy we use is SGD with momentum. After trying with different hyper-parameters, for MLP with Xavier initialization, the batch size we set is 24, lr is 0.0001 and m is 0.6. And for MLP with Kaiming, we found when batch size is 36, the model performs better. The hyper-parameters we chose for 2-layer neural network with Xavier initialization is batch size set as 24, lr set as 0.006 while m is 0.7. For 2-layer neural network with Kaiming initialization, the batch size is 36, lr is 0.003 while m is kept same as 0.7. The batch size, lr and m we use for 3-layer neural network with Xavier initialization

is 36, 0006 and 0.7 respectively. For 3-layer neural network with Kaiming initialization, the learning rate is changed from 0.006 to 0.0002 while others are kept the same.

2.2 Xaiver Initialization

The parameters need to be initialized before neural network training. The early parameter initialization methods are generally based on Gaussian distribution, but with the increase of neural network depth, this method cannot solve the problem of gradient disappearance. The idea that Xavier Glorot put forward in the paper [13] is that the variance of activation value is decreasing layer by layer, which leads to the gradient decreasing layer by layer in back propagation. In order to solve the problem of gradient vanishing, it is necessary to avoid the attenuation of the variance of the activation value. [13] proposed the ideal situation is the output value of each layer keeps Gaussian distribution in both forward and backward propagation.

In forward propagation:

$$Y = WX + B \tag{1}$$

$$w \in \mathbb{R}^{u \times d}\ x \in \mathbb{R}^{d}, y, b \in \mathbb{R}^{u}$$

In order to keep the forward calculation signal strength unchanged, it is necessary to meet the following requirements: $\mathrm{Var}\,(Y_i) = \mathrm{Var}\left(X_j\right)$

Based on the assumptions:

W, X, B are independent of each other

$$W_{ij} \text{ i.i.d. and } E\left[W_{ij}\right] = 0 \tag{2}$$

$$B_i \text{ i.i.d. and } \mathrm{Var}(B_i) = 0 \tag{3}$$

$$X_j \text{ i.i.d. and } E\left[X_j\right] = 0 \tag{4}$$

We can get:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(W_i X + B_i) \tag{5}$$

$$= \mathrm{Var}\left(\sum_{j=1}^{d} W_{ij} X_j + B_i\right) \tag{6}$$

$$= d \times \mathrm{Var}\left(W_{ij} X_j\right) \tag{7}$$

$$= d \times \left(E\left[W_{ij}^2\right] E\left[X_j^2\right] - E^2\left[W_{ij}\right] E^2\left[X_j\right]\right) \tag{8}$$

$$= d \times \mathrm{Var}\left(W_{ij}\right) \mathrm{Var}\left(X_j\right) \tag{9}$$

Then in order to realize $\mathrm{Var}\,(Y_i) = \mathrm{Var}\left(X_j\right)$, we need to satisfy $d \times \mathrm{Var}\left(W_{ij}\right) = 1$ which equals to $\mathrm{Var}\left(W_{ij}\right) = \frac{1}{d}$, then we get initialization methods:

In the Normal Distribution, $W_{ij} \sim \mathrm{Normal}\left(0, \frac{1}{d}\right)$.

In the uniform distribution, $W_{ij} \sim$ Uniform $\left(-\sqrt{\frac{3}{d}}, \sqrt{\frac{3}{d}}\right)$

2.3 Kaiming initialization

Although Xavier takes the variance of activation value into account, it does not take the activation function into account, which will change the distribution of data flow in neural networks. Kaiming initialization[10] is proposed to solve this problem.

In forward propagation:

$$Z = f(X) \tag{10}$$

$$Y = WX + B \tag{11}$$

$$f\ is\ relu\ function, w \in \mathbb{R}^{u \times d}\ x, z \in \mathbb{R}^{d}, y, b \in \mathbb{R}^{u}$$

Based on Xavier initialization, a new hypothesis is introduced in kaiming initialization: X_j has a symmetric distribution around zero which means $\mathrm{Var}\left(Z_j\right) = \frac{1}{2}\mathrm{Var}\left(X_j\right)$ if we want to keep $\mathrm{Var}\left(Y_i\right) = \mathrm{Var}\left(X_j\right)$.

Therefore, we come to the initialization methods:

In the Normal Distribution, $W_{ij} \sim \mathrm{Normal}\left(0, \frac{2}{d}\right)$.

In the uniform distribution, $W_{ij} \sim \mathrm{Uniform}\left(-\sqrt{\frac{6}{d}}, \sqrt{\frac{6}{d}}\right)$

3 Results and Discussion

We have 192 pieces of data, there are 16 experimenters and each experimenter own 12 records. We split 20% of the data as the test set, that is, we chose the data of three experimenters as our test dataset. For the data of the remaining 13 experimenters, we choose to leave one out method to divide the training dataset and the validation set.

The metrics we use here to measure the performance of different models are accuracy、precision、recall and F1 score. In the binary classification problem, it is assumed that the sample has two categories: positive and negative. True positive (TP) means that a positive sample is successfully predicted to be positive. True negative (TN) means negative samples are successfully predicted to be negative. False positive (FP) means false positive prediction of negative samples. False negative (FN) means the false prediction of positive samples is negative. The definitions of accuracy, precision, recall and F1 score are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$F1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{15}$$

As we can see from table 1, with Kaiming initialization, the precision and recall are both 0 on the moderate class which means that the performance of the model is extremely poor. Also, the predictions on the mild class seems poor too, since the precision is only 0.08. From table 2, we can see in 2-layer models, it is still the Xaiver initialization performs better. The 2-layer model with Xaiver initialization method achieves higher accuracy. Also, the models with 2-layers both perform better than model with only 1-layer. When we add layers to 3, we can see the model with Kaiming initialization performs better than model with Xaiver initialization, whose accuracy has achieved to 0.83 which is a substantial improvement compared with 2-layer models. And for 3-layer neural network with Xaiver initialization, though it performs better than the 2-layer neural network, the improvement of accuracy is limited.

Table 1. PERFORMANCE MEASURES FOR MLP DEPRESSION RECOGNITION MODELS

Depression level	MLP+Xaiver			MLP+Kaiming		
	Precision	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
None	0.38	0.36	0.37	0.89	0.3	0.45
Mild	0.38	0.4	0.39	0.08	0.32	0.13
Moderate	0.43	0.42	0.42	0	0	0
Severe	0.44	0.44	0.44	0.38	0.48	0.43
Average	*0.41*	0.41	0.41	*0.34*	0.28	0.25
Overall Accuracy	0.41			0.34		

Table 2. PERFORMANCE MEASURES FOR 2-LAYER NEURAL-NETWORK DEPRESSION RECOGNITION MODELS

Depression level	2-layer +Xaiver			2-layer +Kaiming		
	Precision	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
None	0.55	0.53	0.54	0.51	0.48	0.5
Mild	0.62	0.51	0.56	0.49	0.58	0.53
Moderate	0.52	0.49	0.5	0.5	0.47	0.49
Severe	0.45	0.67	0.54	0.57	0.56	0.57
Average	0.55	0.54	0.54	0.52	0.52	0.52
Overall Accuracy	0.54			0.52		

Table 3. PERFORMANCE MEASURES FOR 3-LAYER NEURAL-NETWORK DEPRESSION RECOGNITION MODELS

Depression level	3-layer+Xaiver			3-layer+Kaiming		
	Precision	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
None	0.54	0.52	0.53	0.74	0.78	0.76
Mild	0.66	0.61	0.63	0.89	0.85	0.87
Moderate	0.56	0.53	0.54	0.82	0.8	0.81
Severe	0.56	0.68	0.61	0.87	0.89	0.88
Average	0.58	0.58	0.58	0.83	0.83	0.83
Overall Accuracy	0.58			0.83		

According to the above analysis, we can find that although Kaiming initialization applies some improvements based on Xavier, it does not mean that its performance is better than Xavier initialization in all neural network models. For neural networks with the same hidden layers, initialization plays an important role in the final model performance. It also enlightens us that for the same data and the same neural network architecture, we can try different initialization methods to get better results.

Based on the accuracy achieved from 3-layer neural network with kaiming initialization, taking physiological signals of the observer as input, neural network will be a good tool to predict depression.

It also inspires us to introduce more neural network training strategies, and find the optimal neural network model architecture through neural architecture search, so as to achieve higher prediction accuracy, which is conducive to more objective diagnosis of depression and early use of effective treatment methods. It will be very beneficial to improve the national happiness and maintain social stability.

References

1. Isometsä, Erkki T., et al. "Suicide in major depression." The American journal of psychiatry (1994)
2. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F: A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10--49(2015)
3. Mathers, C. D., & Loncar, D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS medicine, 3(11), e442(2006)
4. Joyce, P.: Epidemiology of mood disorders. In: Gelder, M., Andreasen, N., Lo´ pez Ibor, J., Geddes, J. (Eds.), New Oxford Textbook of Psychiatry. Oxford University Press, Oxford; New York, pp. 645–650(2012)
5. Mann J J, Apter A, Bertolote J, et al. Suicide prevention strategies: a systematic review.J. Jama, 294(16): 2064-2074(2005)
6. Mitchell A J, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis.J. The Lancet, 374(9690): 609-619(2009)
7. Cummins N, Scherer S, Krajewski J, et al:A review of depression and suicide risk assessment using speech analysis.J. Speech Communication, 71: 10-49(2015)
8. JH Balsters M, J Krahmer E, GJ Swerts M, et al. Verbal and nonverbal correlates for depression: a review.J. Current Psychiatry Reviews, 8(3): 227-234(2012)
9. Zhu X, Gedeon T, Caldwell S, et al. Detecting emotional reactions to videos of depression,C,: 2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES). IEEE, 000147-000152(2019)
10. He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision(2015)
11. Azar, Ahmad Taher, and Shereen M. El-Metwally. "Decision tree classifiers for automated medical diagnosis." Neural Computing and Applications 23.7: 2387-2403(2013)
12. Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." 2011 International conference on innovations in information technology. IEEE(2011)
13. Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings(2010)
14. Herborn, Katherine A., et al. "Skin temperature reveals the intensity of acute stress." Physiology & behavior 152: 225-230(2015)