Predict Nationality of M-Learning Tester Based on Survey Result via Neural Network with Techniques

Tingwei Zeng
College of Engineering and Computer Science, The Australian National University, Canberra, Australia
u6225609@anu.edu.au

Abstract. Different people with distinctive culture background will have their own opinions, and habits. This paper mainly demonstrates how the neural network can be implemented to predict the nationality of M-learning testers with different answers in the survey. This paper also discusses and analyzes the result and compared to the theoretical result. Moreover, to deal with more complex contexts, a processing data measure is introduced to the model, aiming the better performance and reduce over-fitting. The results between them will be compared, analyzed, and figure out whether this technique is suitable in the model. This paper will also refer the previous result with smaller dataset and compare their results.

Keywords: M-Learning, Neural Network, Pre-processing Data, Dropout

1 Introduction

M-learning (mobile learning) is a new, advanced, and convenient way of providing study resources. There are many efficient methods of providing those resources, including texts (like e-books, slides), videos, audios, and combined forms. Based on the previous investigation, it seems different people with different personality traits will have their own preference [1]. In terms of personality traits, there is a popular theory, "The Big Five", which demonstrates mainly five distinctive personality traits of human beings. In each main kind of personality traits, people will also have different degrees [3]. In the previous research, people with different personality traits will prefer distinctive methods of m-learning. For example, high emotional-stability people and high agreeableness people will prefer not to do m-learning when with others familiar, including family, classmates, and friends rather than other personality traits [1].

Personality traits are not the only one factor which influence the preference of m-learning. The culture background also plays an important role. Based on the research, take Saudi Arabian people and Australian people as an example. when they are in the same context (like busy time with someone familiar), they will have different attitudes to the m-learning [1]. This variety may come from the culture from two countries.
Based on the survey data from the research [1], there comes an idea of whether the neural network can be used to classify people based on their survey data or preference.

Neural network and deep learning are a present widely used machine learning model. They can do highly non-linear models with neurons [5]. Neural network contains three

parts: input neurons (features), hidden layers (with hidden neurons), and output neurons. Neurons in continuous two layers relate to a weight and a weight is used to calculate the input of the active function. Deep learning means this neural network model has at least two hidden layers. Generally, neural network is very good at classification. And in this simplified context, neural network can be able to predict the nationality with proper training.

In the previous research, the paper showed the result based on the smaller dataset [2]. In this paper, the more comprehensive data containing more columns, features and answers of testers to the survey. In the previous paper, it assumes that there is some overfitting. Hence, it is important to implement some technique and pre-process the dataset to get better performance.

1.1 Analyzing and processing data

The survey data is huge and contain 5 tables with 95 columns and more than three hundred rows. This dataset contains all the survey answers of testers with some summarization columns. Hence, in order to build the neural network, the requirement should be clarified first.

The question is: predict the nationality of testers based on their preference (survey answers) in all situations. This statement is generally theoretical and correlated supported in previous research [1]. If this is correct, the model should be able to predict the nationality with high accuracy. And in fact, with more data and features, the performance of model should be better than previous one. Unlikely only selecting some columns, this time all survey answers will be selected.

The first step is to select the most suitable dataset as a training set and even testing set. After investigation, BIG-5 is the most significant and suitable dataset. The reason is that, compared to other four sheets, this table contains key features (key action in different contexts without duplicated answers). In these five tables, each table has some unique data while some of it is not important. For example, for the BIG-5 theory, there is no need to show each sub-persona, while the selection can reveal the persona of testers.

Because the “Nationality” column contains only two kinds of variables: “Saudi”, and “Australian”. I set a label encoder for my own to translate these two strings into number of 0 and 1 to make the model understand what the result means.

In the BIG-5 sheet, except for “Nationality” column, there are some columns with strings instead of numbers. Hence, those columns should be label encoded. Fortunately, in those features, the potential variables are: “high”, “low”, and “medium”. Hence the transformed data would just be in the range of [0, 1, 2].

Not all the data is included. There are some columns regarded not important, including “gender” and “marriage”. This doesn’t mean these two have no impact on the result, but gender and marriage status should be the result of training similar to “Nationality” instead of training data.

In conclusion, there are 88 columns in this table with 87 training columns and 1 target output. With this huge amount of data, it took time to process and modify it to avoid bad performance.

1.2 Build and Train the neural network model

The data has been processed and there is no need to regularization or normalization for the unit of this survey data has already been unified. The data should still be split into two parts: training data, and testing data. Then a suitable model should be selected to begin the predict.

To compare the result to previous one, the Two Layer Network is chosen with the same structure of the last paper [2]. However, due to the different size of this data and more features, the hyperparameters should be changed accordingly. The learning rate should be lowered to avoid potential underfitting. Because the output is still two variables (two nationalities), output neurons remain unchanged while input neurons should increase to the 88 as the same to the number of features.

Another issue is to determine the epochs. Without proper training, the model might have underfitting. After calculating the relationship and testing, the epoch for this model is set with 2000.

Figure 1 shows the changing of training accuracy during epochs.

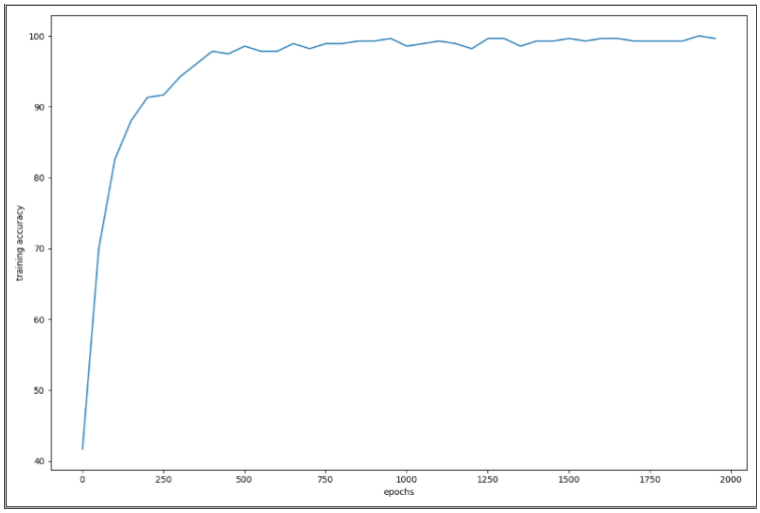

Figure 1. The loss of the model when epoch increases.

1.3 Result

During training, it seems the hidden neuron is not enough to support the model. The train accuracy can be around 99%. However, in the most of cases, the testing accuracy is only greater than the 70% and, in some cases, the testing accuracy is a big smaller

than 70%. The Figure 2 shows the distribution of testing accuracy of this model in 100 loops to get the most accurate performance.

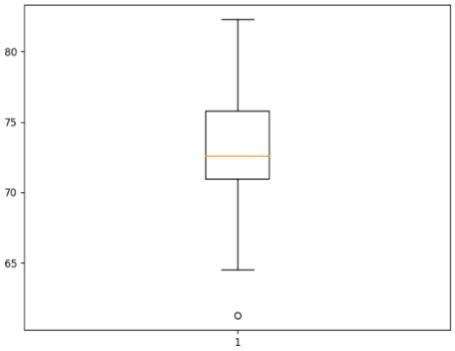

Figure 2. The distribution of testing accuracy.

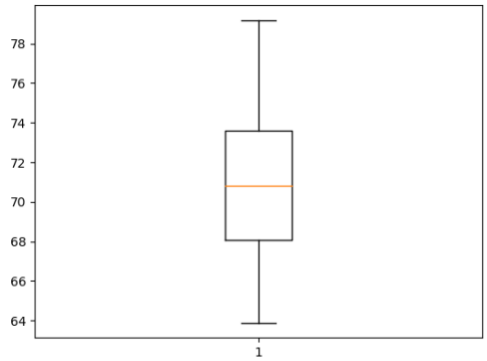

Figure 3. The previous distribution of testing accuracy.

1.4 Analysis and comparison of Result

The result reveal that there might be overfitting in this model with training accuracy much higher than the testing accuracy. To support the argument, the comparison is needed to for the present model and previous model.

Based on the previous research, the testing accuracy of model is shown in Figure 2. One argument of this paper is whether more features can reduce overfitting and improve performance compared to the previous model. Figure 3 shows the previous testing accuracy distribution.

As directly seen from two figures, the present testing accuracy is higher and more stable. For the present model, most of the result are in the range [72, 76], and median value is around 73 while the median of previous result is around 70. Though the difference is not significant, the issue is how to improve performance with high accuracy. This difference can reveal that in this model, with more data and features, the performance can be better. If there are more data, the testing accuracy may converge. However, the result of this model is not processed well enough. In the next section, a technique will be implemented to improve and compare the performance continuously.

2 Dropout Layer

As the result shown above, though the testing accuracy is improved with more data, the training accuracy is higher than the testing accuracy. There can be many factors affecting the performance. Generally, overfitting is a common problem in machine learning. Hence, in this section, this paper will introduce a popular technique called dropout used to reduce overfitting and compare the result above and the result with technique.

2.1 Introduction of Dropout

Dropout is a popular technique used in deep learning to reduce overfitting. The key idea is to randomly drop units (along with their connections) from the neural network during training [4]. Actually, this does not mean that dropout only randomly abandon data completely. There is a formula to describe this model [4]:

$$\begin{aligned} z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)}\mathbf{y}^l + b_i^{(l+1)}, \\ y_i^{(l+1)} &= f(z_i^{(l+1)}), \end{aligned} \tag{1}$$

Where f is any activation function and l is one layer of the neural network model.

Add this formula into the neural network model [4]:

$$\begin{aligned} r_j^{(l)} &\sim \text{Bernoulli}(p), \\ \widetilde{\mathbf{y}}^{(l)} &= \mathbf{r}^{(l)} * \mathbf{y}^{(l)}, \\ z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)}\widetilde{\mathbf{y}}^l + b_i^{(l+1)}, \\ y_i^{(l+1)} &= f(z_i^{(l+1)}). \end{aligned} \tag{2}$$

In each layer, there is a r^l which represents "random". Thus, under the guarantee of expectation unchanged, dropout will randomly remove some data in each layer and hence avoid overfitting.

2.2 Implement Dropout Layer

Implement dropout is not so difficult in coding. As a popular technique, there are packages with functions support this. In this project, a additional dropout layer is implemented instead of simple two layers neural network. The issue is that for dropout layer, the portion of drop data can be determined. Because there are 100 hidden neurons in hidden layer, the portion cannot be too large to avoid underfitting. Considering the epoch and learning rate, the drop portion is set 0.1.

2.3 Analyze Result

The result is shown as Fig 4. With 100 times training and the same dataset, as the Fig 4 shows, all the testing accuracy are above 70%. The range is in the [71, 80] with some outliers, though the train accuracy can achieve 99% with 2000 epochs.

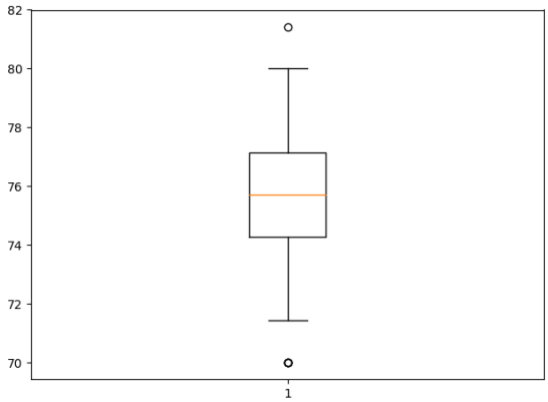

Figure 4. Testing accuracy with dropout layer.

2.4 Compare results with technique

Comparing the Figure 3 and Figure 4, the testing accuracy with dropout layers is better and more stable. The whole performance is improved as the median number is around 76%. And in some cases, the testing accuracy can achieve more than 80%. And the "box" is higher than the result without any processing.

Hence, the result reveals that one reason of unsatisfied performance is overfitting. In this model, the dropout technique can reduce the overfitting significantly and improve the overall performance. However, there are many other techniques which can help models reduce overfitting. For example, the early stopping technique can also be used to reduce overfitting, while this technique is not suitable in this model for the size of data is large enough.

Generally, the huge difference between training accuracy and testing accuracy contains various factors.

3 Conclusion

The paper has demonstrated the whole flow of pre-processing data, building neural networks, analyzing, and implementing techniques. Compared to the previous research, in the prediction of nationalities of testers, this paper shows better model which handle more features and data and demonstrates higher testing accuracy.

While the performance is still not satisfied, this paper lists a potential reason of overfitting. A technique, dropout layer, is implemented. Comparing two results with dropout layer or not, this paper demonstrates that dropout can reduce overfitting and indeed improve the performance.

With limitations, this paper cannot test all the possible techniques to reduce overfitting in this model. Generally, batch normalization might be available in this model and other techniques can be introduced for this unbalanced dataset (the portion of "Australian" testers and "Saudi" testers unbalanced).

References

1. Al-Ismail, M., Gedeon, T., & Yamin, M. (2017). Effects of personality traits and preferences on M-learning. International Journal of Information Technology, 9(1), 77-86.
2. Tingwei Zeng (2020). Predict Nationality of M-Learning Tester Based on Preference via Neural Network and Improvement with Technique
3. Goldberg LR (1993). The structure of phenotypic personality traits. Am Psychol. 1993 Jan, 48(1):26-34.
4. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. 15(56):1929−1958, 2014.
5. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Ann Transl Med. 2018 Jun, 6(11):216. doi: 10.21037/atm.2018.05.32.