Research on BP neural network model based on feature engineering

Lei Yang*, Longqing Zhang, Yong Fan, Liwei Tian, Yungui Chen
School of Computer Science, Guangdong University of Science and Technology, Dongguan, China

ABSTRACT

This research used the machine learning algorithm of the BP neural network to predict a data set. In order to verify the performance of the prediction model, we introduce the confusion matrix and F1 score to evaluate the effect of machine learning. In order to optimize the BP neural network model, we use feature engineering to process the data set and apply the BP neural network model to this new data set. The experimental results show that the machine learning performance of the BP neural network model based on feature engineering is improved.

Keywords: BP neural network, feature engineering, machine learning, model optimization

1. INTRODUCTION

The university is very interested in the enrollment intention of freshmen. Based on this, we collected the freshmen data set of a university, including the registration data of this university in recent years.

BP (back propagation) neural network is a concept proposed by scientists led by Rumelhart and McClelland in 1986[1]. It is a multilayer feedforward neural network trained according to the error back propagation algorithm. It is one of the most widely used neural network models[2]. BP network adds several layers (one or more layers) of neurons between the input layer and the output layer. These neurons are called hidden units[3]. They have no direct connection with the outside world, but the change in their state can affect the relationship between input and output. Each layer can have several nodes[4].

In this paper, we first deal with the data set so that it can be used by machine learning. Then we use BP neural network model to machine learn the data set. We use the F1 score to measure the prediction performance.

Feature engineering refers to the process of transforming the original data into the training data of the model[5]. Its purpose is to obtain better characteristics of the training data and make the machine learning model approach this upper limit. Some scholars call feature engineering attribute selection. It refers to selecting some features from the existing features to optimize the prediction of the system[6]. In order to improve the performance of machine learning, we design an optimization scheme.

We use feature engineering to help extract important features of data sets, so as to generate a new data set. This data set is much smaller than the original data set. After experimental testing, all performance indicators are basically the same as the original data set, but the time efficiency has been significantly improved.

Through the feature engineering processing of the data set, and then the new data set is used for machine learning using BP neural network model again, the results show that the performance has been improved.

2. THE DATASET AND FEATURE ENGINEERING

A university located in Guangzhou, China provided the data set studied in this thesis. We collected the data on this university in recent years. These data are from the official information of the college entrance examination admission system and the internal data of the University's own information management system.

2.1 Features

Part of the data in the dataset comes from the college entrance examination student database of Guangdong Enrollment Office. These data include all the information of students, including the national unified number and name of students, the volunteers of colleges and universities applying for the examination, the name of parents, the contact number of

* yanglei@gdust.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125062Y
 · 0277-786X · doi: 10.1117/12.2662000

parents, home address, the name of high school, the contact information of high school headteachers, etc. An add up to 38 columns of information are appeared in Table 1.

Table 1. Official external raw dataset.

NO.	Name	SN.	Sex No.	Sex	ID	...	Score	Nation
522**	Y**	094408**	2	Girl	4452241992**	...	5	Han
088**	X**	094407**	1	Boy	4452241990**	...		Han
051**	Z**	094403**	1	Boy	4452241989**	...		Han
...	...	...	...	...	...	...	...	...

2.2 Dataset preprocessing

Before machine learning these collected data, the first thing we need to do is preprocess these data. Because most of the collected raw data are missing, some data are noisy, and even some data will be repeatedly collected, these collected raw data can not be used directly. A large amount of processing work needs to be carried out on these raw data to make them become data that can be recognized and used by the program[7].

In the new dataset we constructed, the values of some feature attributes are null[8]. For example, in the art score column, some students do not have this score, so they need to change the null value to 0. There are many similar feature attributes. We have done the same treatment, that is, change the null value to 0.

Then there are some inconsistent data[9]. For example, the name of a major in a university often changes. For example, there is a major called computer network technology, which was changed to computer network a few years ago. These majors are essentially the same major, we have standardized their names in the past few years.

In addition, some noise data need to be removed[10]. In this new dataset, some data are not related to our research content, such as the names of students' parents, headteachers, etc. another additional piece of data that needs to be paid attention to is the number of students from other provinces, which is very small, less than 0.5% of the cases, and these students from other provinces are not within the scope of our research, it's also necessary to remove these data.

Data reduction is to minimize the amount of data on the premise of maintaining the original appearance of the data as much as possible[11]. Because the source data in the new data table is very complete and contains a lot of repeated information, in order to reduce the time of program calculation and improve the efficiency of data learning, after analyzing the admission information and registration information, we deleted part of the data to reduce the amount of data in machine learning[12]. For example, you only need to keep one of the registered permanent residences, mailing addresses, date of birth, and age.

During machine learning of datasets, many data types cannot directly participate in the calculation of the program, such as text data such as major name, student class, and student native place. Therefore, we need to convert the data types of these data to enable the program to calculate[13].

For example, one of the feature attributes is called fixed telephone, which was originally a digital string. We changed it to numbers 0 and 1. The number 0 indicates that the fixed telephone is not installed in the student's home, and the number 1 indicates that the fixed telephone is installed in the home. For another feature attribute mobile telephone, we did the same data conversion[14]. In addition, we also convert some other text numbers into numerical numbers.

After the above processing methods, the dataset becomes standardized and complete, and the volume is reduced to 18 columns. The dataset after processing is displayed in Table 2.

2.3 The feature engineering

The characteristics of data are the useful information extracted from the data for the result prediction. feature engineering is a process that uses professional background knowledge and skills to process data so that features can play a better role in machine learning algorithms[15].

Table 2. Final dataset.

	y	a_1	a_2	...	a_{14}	a_{15}	a_{16}
	Registration status	Gender	Score	...	School type	Examination type	Politic countenance
x_1	0	2	305	...	2	1	13
x_2	1	2	372	...	2	2	1
x_3	0	1	358	...	1	1	1
...	...	...	...	...	...	...	...

The significance of feature engineering lies in three points: first, better features mean greater flexibility, second, better features mean that only simple models are needed, and third, better features mean better results. When performing feature engineering processing on data sets, attention should be paid to the processing of redundancy and noise[16]. Redundancy means that the correlation of some features is too high, which will consume a lot of computing performance. Noise is part of the characteristic, which have a negative impact on the prediction results.

Common feature selection methods include filtering, wrapping, and embedding. Filtering is sorting the parts that leave the most relevant features by evaluating the correlation between a single feature and the result value[17]. Wrapping is to regard feature selection as a feature subset search problem, filter various feature subsets, and evaluate the effect with models. Embedding is to analyze the importance of features according to the model.

3. THE BP NEURAL NETWORK MODEL BASED ON FEATURE ENGINEERING

BP (backpropagation) neural network is a multilayer feedforward neural network trained according to the error backpropagation algorithm. It adds several layers (one or more layers) of neurons between the input layer and the output layer. These neurons are called hide cells, they are not specifically related to the exterior world, but their state changes can influence the relationship between input and yield. Each layer can have several nodes.

In this paper, we made a neural arrangement with three covered-up layers, each containing 10 neurons. The built BP neural arrangement appears in Figure 1.

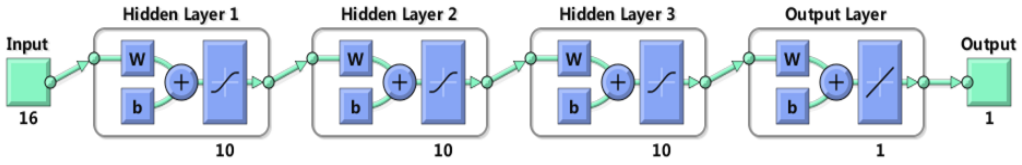

Figure 1. The BP neural network.

We created a neural network with three hidden layers, each containing 10 neurons, the number of iterations of the network is set to 1000, the training accuracy is 10^{-3}, the learning rate is 0.01, and Maximum validation failures is 10 times. The training process is displayed in Figures 2 and 3.

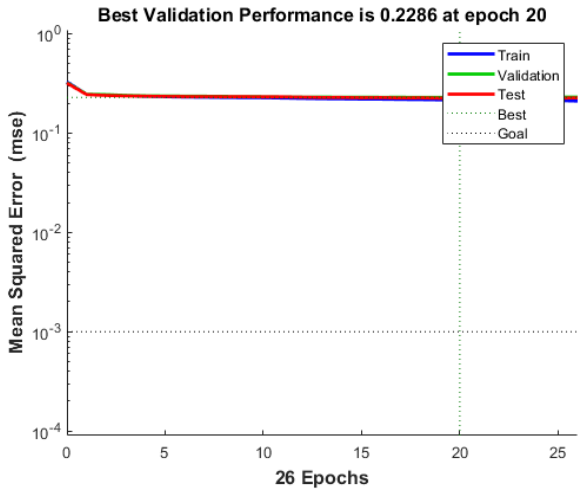

Figure 2. The performance of BP neural network.

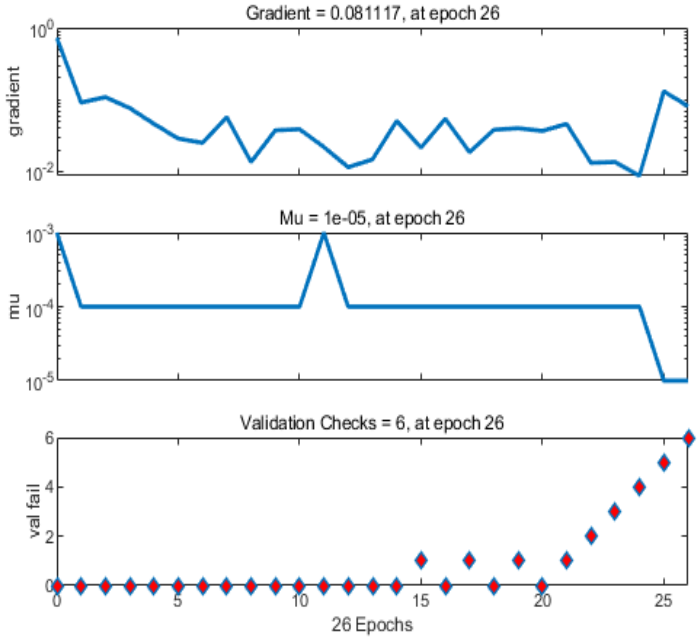

Figure 3. The training state of BP neural network.

The BP neural network model is used to predict the test set. The final predicted performance indicators are displayed in Table 3.

Table 3. Performance metric of BP neural network.

Recall	Precision	Accuracy	F1
68.57%	64.68%	61.07%	0.6656

After the data set is processed by feature engineering, we constructed the new BP neural network, but the input parameters changed from 16 features to 8 features, as displayed in Figure 4, the predicted performance indicators are displayed in Table 4.

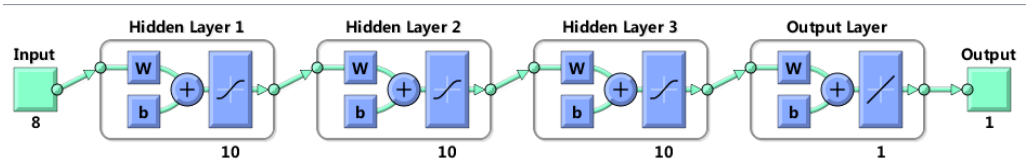

Figure 4. The new BP neural network.

In addition, we also made a neuron count of 10 × 10, the performance of the BP neural network on the new data set is poor, but when the number and number of neuron layers increase, the neural network shows better performance. The reason may be that a 10 × 10 neural network is too simple. When the structure of a neural network becomes complex, it can show better machine learning performance, which is consistent with the idea of deep learning.

Table 4. Performance metric of the new BP neural network.

Recall	Precision	Accuracy	F1
69.61%	66.23%	62.86%	0.6788

In addition, in terms of time cost, the time spent on the data set processed by feature engineering is about 80% of the original data set. Once again, it is proved that when the amount of data is large, the optimization scheme of feature engineering processing of data sets will bring great performance improvement to our machine learning.

4. CONCLUSION

In this paper, we further process the data set through feature engineering, which is used in BP neural network algorithm. Our work shows that the performance of the BP neural network model based on feature engineering has been improved. In the future, we plan to work on the following topics. Use more feature engineering methods to process the data, and try to improve the performance of the model. Apply feature engineering to other machine learning algorithms to see if the performance of the model can be improved.

ACKNOWLEDGMENTS

This research was funded by Key scientific research platforms and projects of colleges and universities in Guangdong Province, special projects in key fields (natural science), grant number 2021ZDZX1075, Innovative and Strengthening Project of Guangdong University of Science and Technology NO. CQ2020062, and Natural Sciences Project of Guangdong University of Science and Technology NO. GKY-2020KYYBK-24, GKY-2020KYYBK-27.

REFERENCES

[1] Sadeghi, B. H. M., “A BP-neural network predictor model for plastic injection molding process,” Journal of Materials Processing Technology, 103(3), 411-416(2000).
[2] Jia, W., Zhao, D., Shen, T., Ding, S., Zhao, Y. and Hu, C. “An optimized classification algorithm by BP neural network based on PLS and HCA,” Applied Intelligence, 43(1), 176-191(2015).
[3] Lyu, J. C. and Zhang, J., “BP neural network prediction model for suicide attempt among Chinese rural residents,” Journal of Affective Disorders, 246(1), 465-473(2019).
[4] Zhang, Q., Guo, Y. and Song, Z. Y., “Dynamic curve fitting and bp neural network with feature extraction for mobile specific emitter identification,” IEEE Access, 9(1), 33897-33910(2021).
[5] Akhiat, Y., Manzali, Y., Chahhou, M. and Zinedine, A., “A new noisy random forest based method for feature selection,” Cybernetics and Information Technologies, 21(2), 10-28(2021).
[6] Zhang, Y., Zhu, Y., Li, X., Wang, X. and Guo, X., “anomaly detection based on mining six local data features and BP neural network,” Symmetry, 11(4), 571(2019).

[7] Beck, M. A., Liu, C. Y., Bidinosti, C. P., Henry, C. J., Godee, C. M. and Ajmani, M., "An embedded system for the automated generation of labeled plant images to enable machine learning applications in agriculture," PLoS One, 15(12), e0243923(2020).
[8] Bertsimas, D. and Dunn, J., "Optimal classification trees," Mach. Learn., 106(7), 1039-1082(2017).
[9] Han, T., Jiang, D.X., Zhao, Q., Wang, L. and Yin, K. "Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery," Transactions of the Institute of Measurement and Control, 40(8), 2681-2693(2018).
[10] Goh, Y. C., Cai, X. Q., Theseira, W., Ko, G. and Khor, K. A., "Evaluating human versus machine learning performance in classifying research abstracts," Scientometrics, 125(2), 1197-1212(2020).
[11] Chen, X. H., Yu, S. Y., Zhang, Y. F., Chu, F. F. and Sun, B. "Machine learning method for continuous noninvasive blood pressure detection based on random forest," IEEE Access, 9, 34112-34118(2021).
[12] Pande, M. and Mulay, P., "Bibliometric survey of quantum machine learning," Science & Technology Libraries, 39(4), 369-382(2020).
[13] Fletcher, S. and Islam, M. Z., "Decision tree classification with differential privacy: A survey," ACM Comput. Surv., 52(4), 33(2019).
[14] Zhang, L., Luo, J. H. and Yang, S. Y., "Forecasting box office revenue of movies with BP neural network," Expert Systems with Applications, 36(3), 6580-6587(2009).
[15] Rusli, R., Haque, M. M., Saifuzzaman, M. and King, M., "Crash severity along rural mountainous highways in Malaysia: An application of a combined decision tree and logistic regression model," Traffic Injury Prevention, 19(7), (2018).
[16] Mu, Y. S., Liu, X. D. and Wang, L. D., "A Pearson's correlation coefficient based decision tree and its parallel implementation," Inf. Sci., 435(1), 40-58(2018).
[17] Mu, Y. H., Qiu, B., Wei, S. Y., Song, T., Zheng, Z. P. and Guo, P., "Regression prediction of photometric redshift based on particle warm optimization neural network algorithm," Spectroscopy and Spectral Analysis, 39(9), 2693-2697(2019).