Research on Prediction of Student Status' Change Based on Neural Network Algorith

Wenwen Yu^{1,a,*}, Wenxi Zheng^{2,b} and Min Wu^{1,c}

[1]School of Software Engineering, University of Science and Technology of China, Hefei 230051, China

[2]Mordern Educational Technology Center, University of Science and Technology of China, Hefei, 230051, China

[a] 15840367371@163.com, [b] wxzheng@ustc.edu.cn, [c] minwu@ustc.edu.cn

Keywords: the change of student status, neural network algorithm, prediction, decision support.

Abstract: In order to provide effectively assessment and decision support for student status management, a prediction model for the change of student status based on back propagation neural network algorithm is presented, compared with the traditional statistical methods, it achieves the prediction of student status changes for individual student as well as the accuracy assessment of status changes prediction for student groups with technical method. According to the basic principle and modeling method of BP neural network algorithm, the implementation process involves studying data feature of student related dataset, defining the topology of neural network algorithm, and adding cross-validation to train and validate the neural network model of student status changing. The result shows that the training model can effectively predict the change of student status, and the accuracy of prediction can reach 85%.

1. Introduction

Change of student status, which is non-procedural changes on students at the level of student status management, mainly includes transfer of majors, suspension of studies, returning to school, repeating the grade, dropping out of school, transferring to another school, extending graduation, going abroad[1]. Through investigation, it has been found that there are many changes of student status in every university every year. However, these situations are mostly recorded or statistically analyzed by tables or systems. The operations are tedious, and it is difficult to find the potential information and patterns. Some of the students are even unable to know whether they can finish their studies in school normally. Therefore, it is an important content to analyze and predict the change of student status timely and effective for university student management. Through choosing the student whose grade is in 2011 to 2015 in one university to analyze, we can find that the number of students who have a change in student status is increasing year by year. Therefore, to predict the change of student status timely, evaluate the accuracy of prediction, and realize the early warning the change of student status by technology means and methods are necessary, it can be more effective to prevent students from being unable to complete their study or some other events, as well as providing assistant decision support for student status management in universities.

Back Propagation neural network algorithm is one of the methods of data mining which is used for predicting tasks. The algorithm has high tolerance to noise data and excellent pattern classification ability for untrained data, which can be used even if the relationship between attributes and classes is lacking[2]. Therefore, a prediction model for the change of student status is proposed, it is based on the BP neural network algorithm, and using the cross-validation to assess the model at the same time. The construction of the model uses the actual data of students, it is constructed by data pre-processing, network topology design, related parameter definition, network training and verification process. At last, the model can predict the change of student status, calculate the accuracy of the prediction on the change of student status, and provide supportive decision for the educational administration staff.

2. The Construction of Prediction Model for the Change of Student Status

2.1 BP Neural Network

BP neural network is a multi-layer feed forward network trained by error back propagation algorithm. It is one of the most widely used neural network models at present[3]. Gradient descent method is generally used to train the network, and in order to minimize the sum of squared error of the network, the weight and the threshold of the network are constantly adjusted by back propagation. The BP neural network model topology includes an input layer, a hidden layer, and an output layer. Each layer is composed of a number of units. The unit of the input layer is generally called an input unit, the unit of the hidden layer and output layer is a neural node or called output unit. The topology of the network is shown in Figure 1.

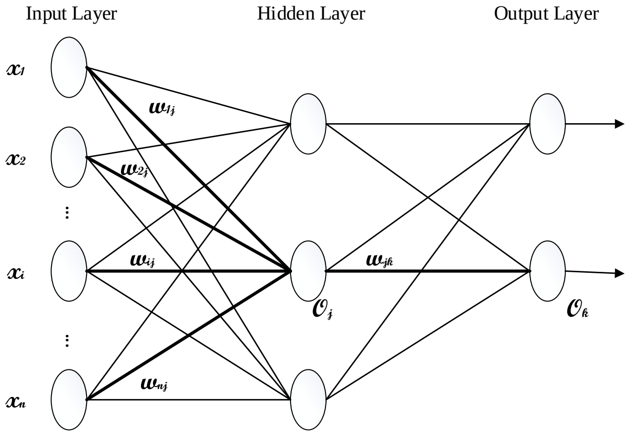

Fig.1 Neural network topology

The weighted sum of output for units of each layer is the input for the next layer, it is applied by using an active function. Given a sufficient number of training samples and hidden units, BP neural network can simulate one function according to arbitrary accuracy [4].

2.2 Construction and Verification of Neural Network Algorithm Model

To construct BP Neural Network Algorithm model, it usually need the following 6 steps：

1) Define data set characteristics and data collection.
2) Data preprocessing.
3) Define the topology.
4) Parameters initialization (weight, learning rate etc).
5) Training neural network under unsatisfied termination conditions.
6) Validate the effect of prediction model.

2.2.1 Define data set characteristics and data collection

The main purpose of this study is to build a prediction model for the change of student status. By studying the characteristics of data in the educational administration system database, the fields contained in the data set are identified, it includes 16 predictive variables and one categorical variable in all. The data comes from the database tables: the table of student basic information(STUDENT), the table of personal grade(STUDENT_SCORE), the table of course library information(COURSE _LIBRARY), and the table of student status(STUDENT_CHANGE). Because the data fields in dataset originate from multiple database tables, the sql statements are used to obtain the related data and rewritten it into a new table in the database to form the original research data set -student_change_traingData.

2.2.2 Data pre-processing

The number of student_change_traingData is about 2000, most of the extracted original data is incomplete, irregular and noisy. The neural network algorithm requires that all feature values must be coded in a standard way, and the value is generally between 0 and 1[5].Therefore, before the network training, data pre-processing mainly includes three aspects: data cleaning, data normalization, and data shuffle.

In original dataset, there are some scalar variables and their values are NULL, for example, the variable of student's major, because all the majors and colleges have their own fixed code, the code also has a certain correlation with each other, so for a certain student whose major code is NULL, we could firstly infer the code range of their major according to the college he's affiliation to, and then select the other student major code mode within the coding range of the major to replace the NULL value. For the variable of test score, the storage of the variable in the original database includes three criteria, like the 100-mark system, the five-grade marking system, and the two-level marking system. For specification, the three standards need to be unified into one standard, though the same kind of transformation rules to map it into the same type of data.

To normalized dataset after initial treatment, we will use the class of preprocessing.Min -MaxScaler in sklearn, the linear function is bellow:

$$y = \frac{2*(x - \min)}{\max - \min} - 1 \tag{1}$$

Normalization is about to scale the property to a specified maximum and minimum value ([-1,1]). As for the output, the neural network training result will always return a continuous value between [-1, 1]. This research is aimed at predicting the change of student status. Therefore, the change result of student status is divided into two points and defined as follows:

1) If the change type of student status is 0, it means there is no change of student status, and the output is marked as [1, -1,-1]. If there is a slight error in the output, it can be approximated, eg [0.99, -0.99,-0.99];
2) If the change type of student status is 1, it means there is one change type of student status, and the output is marked as [-1, 1,-1].
3) The rest change type of student status means the student have more than one change in the status, and the output is marked as [-1, -1, 1].

Finally, we will shuffle the dataset so that the neighboring samples hardly belong to the same class. Because disrupting dataset helps to increase the accuracy of the experimental results.

2.2.3 Topology design

The neural network model of the hidden layer is actually a linear or nonlinear regression model. It is generally believed that increasing the number of hidden layers can reduce network errors and make the network more complicated as well. This increases the training time of network and the tendency of "overfitting"[6]. Therefore, this study uses a 3-layer BP neural network. According to the characteristics of the data set, we will define the number of input layer nodes is 16, the number of hidden layer nodes is 6, and the number of output layer nodes is 3. The weights are random initialized within the defined interval. The activation function uses a hyperbolic tangent function, that is:

$$\tanh(\mathrm{x}) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2}$$

The range of this function is [-1, 1], which is the reason why the above normalization algorithm is normalized to this interval.

2.2.4 k-fold cross validation

Cross-validation is a technique used to assess the effect of a learning model applied to an independent dataset [7]. By adding training data and reducing feature attributes, the prediction model can obtain more accurate prediction results. At the same time, in order to reduce the impact of randomness on the results, multiple rounds of cross-validation are performed in different data. The final result comes from multiple rounds of aggregation[8]. Therefore, the k-fold cross-validation is raised, and the algorithm specific process description is summarized as follows:

1) The training set S is averaged into k disjoint subsets. Assuming that the number of training examples in S is n, then each subset has n/k training examples. The corresponding subset is called {s1,s2 ,...,sk};

2) Every time taking one for the test set, the other k-1 for training set;
3) The test set is applied to the model trained by the training set, and obtain the classification rate.
4) Calculate the average of the obtained classification rate for k times as the true classification rate of the model.

Generally, The higher the value of k is, the higher the accuracy of the result is, but the calculation cost also increases. Therefore, in this experiment, we set k=5, which means the training set data volume is 1600, and the test set data volume is 400. Taking into account about the general learning rate range of [0.0,0.1], the impulse value of [0.1,0.8], so through running the combination of two parameter values with 5 fold cross validation, the set of parameters with the highest average accuracy rate is used to obtain the best classification prediction model.

3. Result analysis

Based on the above processing and setting, a neural network for prediction of student status' change can be constructed. The number of iterations of the process is set to 1000 times. Through repeat training of the neural network until reach the maximum number of iterations, we can get the predict result and analyzed the change of student status.

By cross-combining the parameter of learning rate and impulse, using the 5-fold cross validation, after continuous training verification, the average correct rate is compared as shown in Table 1 below.

Table 1 Parameter combination test results

Numble	Learning rate	Impulse	Average accuracy(%)
1	0.05	0.1	0.850252149
2	0.05	0.2	0.711613091
3	0.05	0.3	0.841652567
4	0.1	0.1	0.832164789
5	0.1	0.2	0.840565854
6	0.1	0.3	0.836513347
7	0.15	0.1	0.837204461
8	0.15	0.2	0.842541309
9	0.15	0.3	0.847088375

From Table 1, it can be seen that when the learning rate is 0.05 and the impulse is 0.1, the accuracy of model classification prediction is the highest, which can reach more than 85%. In this case, the cycle number-error diagram results are shown in Figure 2. From Figure 2, we can see that when training is less than 400 times, the error has a faster downward trend, and finally the error tends to be flat and basically does not change.

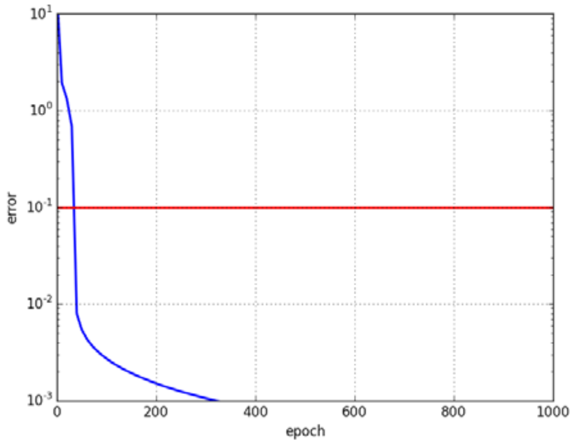

Figure 2 Neural Network training cycle-error map

Based on the above experimental results, we can get a predictive model for the change classification of student status. This model can effectively predict the change of student status and obtain the accuracy of the final overall forecast. The school administrators can make use of this model to manage the change of student status. Now to have an effective prediction for the change of student status is not only based on the students' existing score, but also based on the multi-dimensional analysis of the existing data.

4. Conclusions

After analyzing the existing problems of the educational administration management in student status in universities, the neural network model is used to analyze and predict the change of student status. Through the analysis of actual data, network topology design, and training model continually, the effectiveness of the BP neural network algorithm applied to predict the change of student status is verified. The constructed prediction model can predict changes in student status, and the accuracy of prediction can reach 85%. In this way, using scientific methods to achieve educational administration management, which can not only reduce the work pressure of the educational administrators and teachers, but also improve the quality of student teaching. If the model can further correlate other influencing factors with current data like the family background, the school will gain richer concrete analysis and forecasting conclusions.

Acknowledgments

This work was financially supported by provincial teaching research project (2016jyxm1144).

References

[1] Yang Bingyou, Du Xiaowei, Xie Hailong, et al. An analysis of the current situation of college student status and the management countermeasures. China Medical Herald. Vol. 7(2017), p. 134-137.

[2] Han Jiawei, Kamber M, et al. Data Mining Concepts and Techniques. Beijing: Mechanical Industry Press, 2001.

[3] Wang Lei, et al. Principles, classification and application of artificial neural networks. Science and Technology Information, (2014), No.3, p. 240-241.

[4] He Jiazhou, Zhou Zhihua, Gao Yang, et al. Fault Diagnosis Model Based on New Neural Network Classifier. Journal of Computer Research and Development. Vol 38 (2001) No. 1, p. 93-97.

[5] Daniel T. Larose, Chantal D. Larose, et al. Data mining and predictive analysis. Beijing: Tsinghua University Press, 2017.

[6] Zhang Yu, Yuan Xiaoxuan, Gong Xiaoqian, et al. Research on Prediction of Sports Performance Based on BP Neural Network Algorithm. Bulletin of Science and Technology. (2013) No. 6, p. 149-151.

[7] Donate J P,Li Xiaodong, et al. Time series forecasting by evolving artificial neural networks with genetic algorithms, differential evolution and estimation of distribution algorithm. Neural Compute & Applic. Vol.22 (2013), p.11-20.

[8] Bian Naizheng, Li Shuo, Chen Chucai, et al. Application of weighted cross validation neural network in water quality prediction. Computer Engineering and Applications. (2015) No.21, p. 255-258.