Analysis of the New Model of Academic Education Management of College Students in the New Era Based on Data Mining Analysis

Shan Wu[1*]

[1]*Qingdao No. 5 middle school, Qingdao 266071, Shandong, China*
wushan1997@yeah.net

Abstract

With the combination of the Internet and the education field, online learning has become the preferred choice for learners, and also provides the convenience for learners to watch anytime and anywhere, as well as to watch recorded lectures, without being restricted by location and time. Since the entire online learning process takes place online, this allows all the learning behavior data of students to be kept intact in the learning management system, providing a reliable source of data for studying early warning and teaching interventions for academic monitoring of students' online learning behavior. The most common method to establish an achievement early warning mechanism is to build a student achievement classifier, and common classifier models include decision tree, plain Bayesian, SVM, etc. In this paper, we propose a genetic algorithm optimized BP neural network for grade prediction after analyzing the above three classification models, which can fully consider the joint action between multiple features. Based on the BP neural network, the optimized BP network model using genetic algorithm can improve the generalization ability of the model, and the optimized BP neural network is used to solve the joint action of multiple features, which further improves the detection accuracy and precision.

Keywords: *Course evaluation; Association rules; Academic alert; Data-driven; Instructional management*

1 INTRODUCTION

At present, due to the strong support of the state and related departments for the construction of education informatization, most higher education institutions have introduced teaching management systems, which enable college students to study online and offline. These systems more or less save the data left by students in the process of using them, and as the number of students increases and the number of uses accumulates day by day, these data are growing in the form of exponential explosion, creating an opportunity for opportunities to analyse the online learning behavior of university students [1]. For different learners and researchers, the significance of educational data mining is very different: learners are interested in data mining to help them understand their learning level [2]. Learning efficiency and learning outcomes in a timely manner; educators are more interested in data mining to facilitate their work, such as more objective feedback on teaching and learning, learners' understanding of the content, as a reference to adjust their teaching plans, teaching content, and by predicting learning. The educators would like to see whether their teaching goals are accomplished by predicting learners' performance [3]. The mining of educational data not only promotes efficient learning, but also has a significant effect on the improvement of curriculum quality, which is important for the development of the education field [4].

2 C4.5 PERFORMANCE PREDICTION BASED ON DECISION TREE ALGORITHM

Decision tree is one of the common algorithms in the field of data mining research, which has the advantages of high accuracy, high efficiency, fast computing speed, and the ability to handle multiple types of data, etc.

2.1 Data acquisition and pre-processing age Setup

In this paper, the experiment is conducted on the business intelligence course taken by the 2017 students

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 329-334, 2023.
https://doi.org/10.2991/978-94-6463-040-4_50

of a university majoring in information management in the second semester of their junior year, and the data source is the records generated by students studying in the SuperStar Learning Platform, and some attributes involve students' privacy and are desensitized such as students' names, which are hidden and changed to Zhang**. The source data contains 12 dimensions, such as name, class, college, in-class discussions, in-class assignments, and in-class tests, but considering that some dimensions have little influence on students' performance, the experiments use Pearson's correlation coefficient to analyze the correlation of the original data. It can be seen that the correlations less than 0.1 are all static attributes, and these static attributes have little influence on grades and can be discarded when training the prediction model.

2.2 Experimental results and analysis

The model was trained in the experimental environment shown in Table 1, and the following experiments were conducted in this environment.

Tab.1 Lab environment

Projects	Value
Processor	i5-6300HQ (2.3 GHz) CPU
Memory	16GB
Operating System	Microsoft Windows 10
Programming Language	python3.6
Development software	Jetbrains Py Charm 2019.3

Precision, Recall, F1 value and confusion matrix are some of the commonly used model metrics for evaluating classification models. The evaluation effects of decision tree models are shown in Table 2 and Figure 1.

Tab.2 Decision tree algorithm experiment results

	Precision	Recall	F1-score	Support
1	0.67	0.60	0.67	0.19
2	0.60	0.83	0.70	0.33
3	0.67	0.45	0.856	0.26
4	0.78	0,59	0.67	0.22

From the above evaluation results, the 4.5 decision tree algorithm classified a total of 54 samples, of which 10 samples were classified as excellent, 6 samples were classified correctly (60%), 18 samples were classified as good, 15 samples were classified correctly (83%), 14 samples were classified as medium, 6 samples were classified correctly (43%), and 12 samples were classified as poor (67%). There were 12 samples with poor classification and 8 samples with correct classification, accounting for 67% (Xu,2022). In terms of the correct classification rate of the samples, those with good category have the highest correct classification rate, and those with poor category have the second highest classification rate, but for student performance warning, the students who need to be warned are mainly in the poor and medium categories, so the classification effect is not very satisfactory [5]. Overall, the classification result of decision tree for grade prediction is not very satisfactory, and the average correct rate of the overall sample test is 65%.

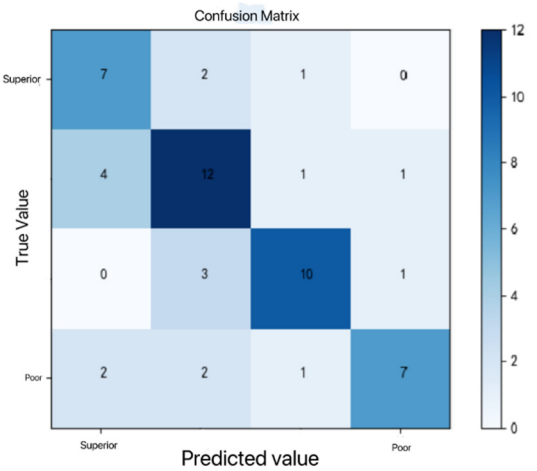

Fig. 1 Confusion matrix diagram of decision tree algorithm

3 PERFORMANCE PREDICTION BASED ON SVM ALGORITHM

The SVM algorithm mainly forms a hyperplane in a two-dimensional space based on a specific function. The hyperplane can maximize the separation of two types of data, and for multi-classification problems, the support vector machine algorithm needs to map it to a higher dimensional space to form a hyperplane to classify the data.or the mutual benefit and protection of Authors and Publishers, it is necessary that Authors provide formal written Consent to Publish and Transfer of Copyright before publication of the Book. The signed Consent ensures that the publisher has the Author’s authorization to publish the Contribution.

The copyright form is located on the authors’ reserved area [6].

The form should be completed and signed by one author on behalf of all the other authors [7].

3.1 Experimental results and analysis

SVM performs better in linear problems, but real-life problems are often nonlinear. sVM algorithms need to

transform the nonlinearity into another linear problem in high-dimensional space to solve the nonlinear data,and need to need to transform the nonlinear ratio into a high-dimensional linear problem [8]. kernel functions mainly include polynomial kernel functions, Gaussian kernel functions, etc. And in this experiment, the third-party library sklearn under Python language is used for the experiment, the kernel function is Gaussian kernel, and the model is also tested by ten-fold cross-validation, and the model evaluation is also evaluated using Precision, Recall, F1 value (score) and confusion matrix, and the evaluation results are shown in Table 3 and Figure 2.

Tab.3 SVM algorithm experiment results

	Precision	Recall	F1-score	Support
1	0.64	0.70	0.68	0.19
2	0.71	0.67	0.69	0.33
3	0.63	0.73	0.77	0.26
4	0.80	0,66	0.76	0.22

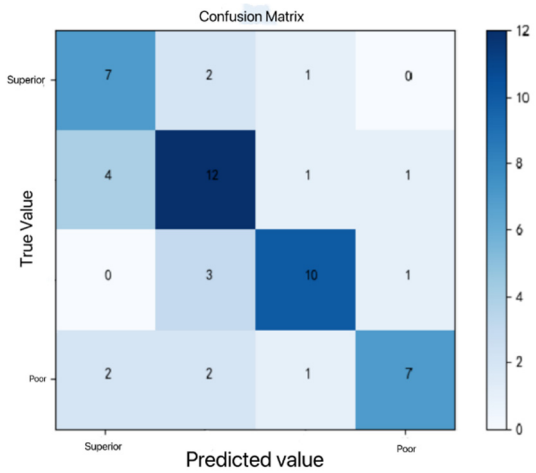

Fig. 2 SVM algorithm confusion matrix diagram

From the evaluation results in Table 4 and Figure 2, the SVM algorithm classifies a total of 54 samples in 4 categories, of which there are 10 samples with excellent category and 7 samples with correct classification (70%), 18 samples with good category and 12 samples with correct classification (67%), 14 samples with medium category and 10 samples with correct classification (71%), and 12 samples with poor category and 8 samples with correct classification (67%). In terms of the correct classification rate of the samples, the category of medium has the highest correct classification rate, and the category of excellent is the second, but for student performance warning, the students who need to be warned are mainly concentrated in the two categories of poor and medium, so the classification effect is not very satisfactory, but from the F1 value, the category of poor is a little better compared with the other three categories, followed by the overall average correct rate of 69% for the tested samples. In general, the classification effect of SVM algorithm is better than that of C4.5 decision tree, but the SVM algorithm is more influenced by the selection of kernel function and the selection of super parameter.

4 PERFORMANCE PREDICTION BASED ON THE PLAIN BAYESIAN ALGORITHM

NB is an effective classification method. It has important applications in a range of practical fields such as social sciences, economic and business activities, education, etc.

4.1 Analysis of the Plain Bayesian Algorithm

If Let the input space $a \in R^n$ be the set of n-dimensional vectors and the output space dimension class labeled set $\beta = \{C_1, C_2, C_{3,} \cdots, C_K\}$. x is a random variable defined on the input space α and Y is a random variable defined on the output space β.

The prior probability distribution is given in equation (1).

$$\mathrm{P}(\mathrm{Y}=C_k), k = 1,2, \ldots, k \tag{1}$$

The conditional probability distribution is given in equation (2).

$$P(X = x|Y = c_k) = P(X^1 = x^1, \ldots, X^n = x^n),$$
$$k = 1,2, \ldots K \tag{2}$$

Thus, the joint probability distribution P(X, Y) can be found from the conditional probability formula $P(X|Y) = \frac{P(X|Y)}{P(Y)}$.

4.2 Experimental results and analysis

NB is a generative model and converges faster than the discriminative approach. NB uses a combination of both a priori and a posteriori probabilities to avoid the subjective and overfitting problems that arise when using one probability alone. In the algorithm implementation, we use the plain Bayesian algorithm in the learn library for the experiments. For the model evaluation, we also use Precision, Recall, F1-score and confusion matrix to evaluate the prediction effect of different models, and the evaluation results are shown in Table 4 and Figure 3.

Tab.4 NB algorithm experiment results

	Precision	Recall	F1-score	Support
1	0.53	0.69	0.61	0.19
2	0.63	0.78	0.69	0.33
3	0.77	0.70	0.74	0.26
4	0.78	0,59	0.67	0.22

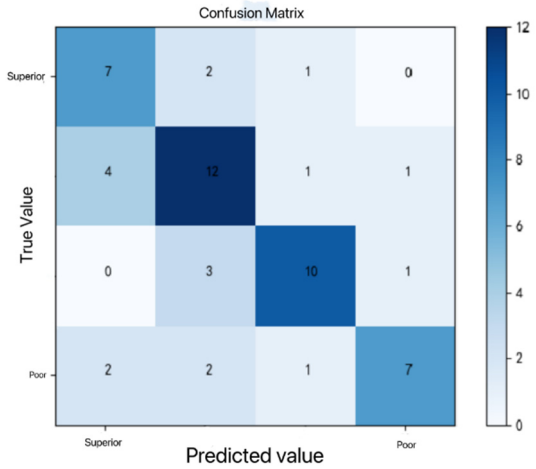

Fig. 3 Naive Bayes algorithm confusion matrix diagram

From the evaluation results in Table 5 and Figure 3, the plain Bayesian algorithm classifies a total of 54 samples in 4 categories, among which 10 samples with excellent category and 6 samples with correct classification account for 70%, 18 samples with good category and 12 samples with correct classification account for 67%, 14 samples with medium category and 10 samples with correct classification account for 71%, and 12 samples with poor category and 7 samples with correct classification account for 58%. There are 12 samples with poor classification and 7 samples with correct classification, accounting for 58%. In terms of the classification correctness of the samples, the classification correctness of the category is the highest. In the early warning mechanism, students who are in the category of medium can be warned, so as to improve the teaching effect.

4.3 Genetic algorithm optimization based BP neural network for performance prediction

BP neural network is an effective tool for solving nonlinear problems by learning from a large amount of data and is one of the core models commonly used in deep learning. However, the selection of the initial parameters and the design of the network structure of BP neural network also have a great impact on the performance of neural network. Therefore, the global optimization feature of GA combined with back propagation can be used to optimize the neural network structure and determine the parameter selection.

4.3.1 Neural network analysis

The BP neural network first obtains the network error through the forward transmission of the input signal, and then passes the network error to each layer in reverse, and finally updates the network according to the error. In the forward propagation stage, the input signal needs to be transmitted to the hidden layer through the input layer, and finally reaches the output after being processed by the subsequent network layer by layer, if the result of the output layer fails to meet the function expectation, it needs to enter the backward propagation stage of the BP neural network; in the backward propagation stage, the neural network needs to propagate the error to each neuron according to the weights of each neuron node, to correct the neural network of each neuron's The back propagation phase requires the neural network to propagate the error to each neuron based on the weights of each neuron node to correct the connection weights and threshold values of each neuron of the neural network. The three-layer network structure is shown in Figure 4.

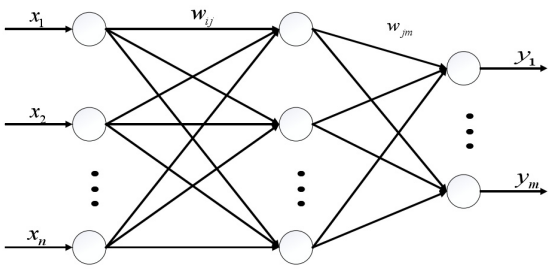

Fig.4 Structure diagram of three layers BP neural network

In BP neural network forward propagation, the input signal needs to be transmitted through the input layer to the hidden layer, and finally reaches the output after being processed layer by layer by the subsequent network. The output value of the neuron in the hidden layer is represented by, see equation (3).

$$h_j = f_1 \sum_{i=1}^{n} w_{ij} x_1 = \theta_i \tag{3}$$

In the forward propagation process of the BP algorithm, the input information is transmitted to the hidden layer after passing through the input layer, and finally reaches the output after being processed layer by layer by the subsequent network to obtain the output error of the network, and if the error does not reach the preset criteria, the error is passed to the neurons of each layer in reverse.

4.3.2 Analysis of BP Neural Networks Optimized by Genetic Algorithm

The basic idea of BP neural network based on genetic optimization is to combine BP neural network algorithm with genetic algorithm. The algorithm is mainly divided into two parts: the GA algorithm optimization part and the BP neural network part. The flow of the BP neural network based on genetic algorithm optimization is shown in Figure 5.

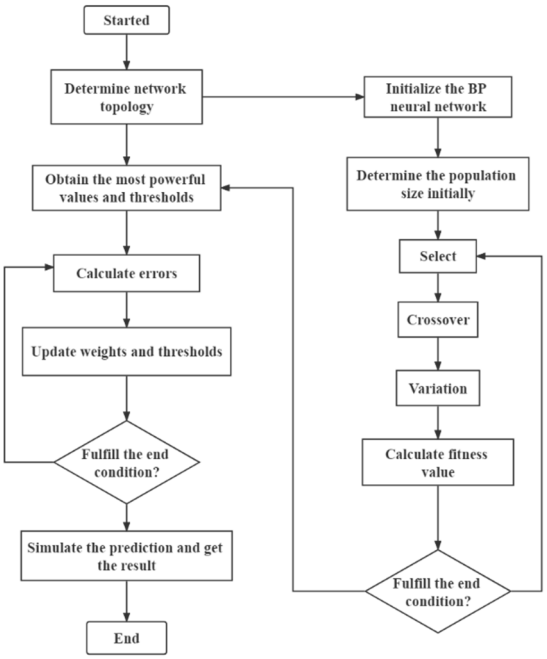

Fig. 5 The flow chart of BP neural network based on genetic algorithm

4.3.3 Experimental results and analysis

The prediction results of BP neural networks often depend on the original state of the network, and it is more effective to use BP neural networks in local solution search; genetic algorithms are better at global search. Therefore, this chapter uses the GA algorithm to optimize the initial weights and thresholds to find some better search spaces in the solution space, and then uses the BP neural network to search the optimal solution in these small solution spaces to train the model. The specific parameters are set as the number of population size 10, the maximum number of iterations 50, the crossover probability 0.4, and the variation probability 0.2 for training the model, and the evaluation results are shown in Table 5 and Figure 6.

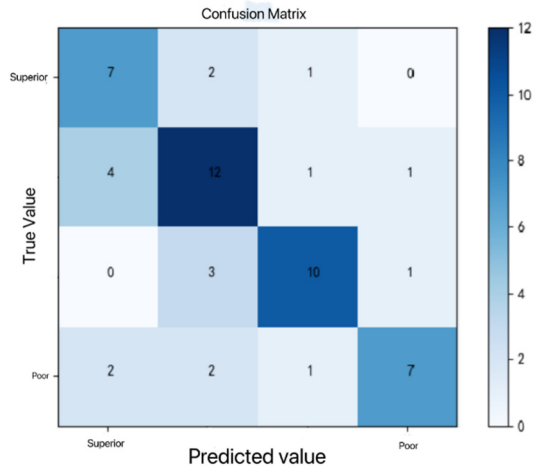

Fig.6 GA optimized BP algorithm confusion matrix iagram

Tab.5 GA optimized BP algorithm experimental results

	Precision	Recall	F1-score	Support
1	0.73	0.80	0.77	0.19
2	0.65	0.76	0.76	0.33
3	0.75	0.86	0.74	0.26
4	0.82	0,75	0.80	0.22

From the evaluation results in Table 6 and Figure 6, the BP neural network optimized by the genetic algorithm classified a total of 54 samples in 4 categories, of which 10 samples with excellent category and 8 samples with correct classification accounted for 80%, 18 samples with good category and 12 samples with correct classification accounted for 67%, 14 samples with medium category and 12 samples with correct classification accounted for 86%, and 12 samples with poor category and 9 samples with correct classification accounted for 75%. The number of samples classified as poor was 12, and the number of samples classified as correct was 9, accounting for 75%. In terms of the correct classification rate, both the classification with category as poor and the classification with category as medium, which is the main concern of the early warning model, have improved compared with decision tree, SVM and plain Bayesian algorithms.

5 CONCLUSIONS

Students' online learning behavioral characteristics contain rich information. In this chapter, the relationship between student behavioral characteristics data and grades is studied using C4.5 decision tree algorithm, SVM support vector machine algorithm, NB plain Bayesian and genetic algorithm optimized BP neural network. The results show that the genetic algorithm-optimized BP neural network outperforms the other three

classification algorithm models and can effectively classify students' performance based on their online learning behaviors, thus alerting students to adjust their learning status in a timely manner.

REFERENCES

[1] Ding Donghai. (2022). Application of humanistic care in the education management of students in colleges and universities [J]. Heilongjiang Science, 13(07):65-67.

[2] Jiang Wei, Song Wenzheng. (2022). Research on the influence of big data on modern university education management and improvement strategies[J]. Journal of Jilin Provincial College of Education,38(05):38-41.

[3] Liu Yuxin, Zhao Wenxin. (2022). Research on the use of red culture in college education management under the new situation [J]. PR World,(10):81-82.

[4] Shi Yuheng. (2022). The innovation path of university education management mode under the concept of people-oriented [J]. Journal of Huainan Vocational and Technical College, 22(02):111-113.

[5] Wu Mingsheng. (2022). Transformation of student education management mode and coping strategies in universities under the information environment [J]. Journal of Shanxi University of Finance and Economics,44(S1):118-120.

[6] Xie Chunxia. (2022). Discussion on the optimization path of student education and management work in universities under the Internet perspective [J]. Journal of Yanbian College of Education,36(02): 96-98.

[7] Xiong Guifang, Guo Runping. (2022). Analysis of the innovation path of university education management informatization in the era of big data [J]. Science and Technology Information, 2022, 20(08):16-18.

[8] Zhang Hongwei. (2022). The impact and strategy of the application of big data technology on the education management of colleges and universities [J]. Journal of Yanbian College of Education, 36(02):47-48+51.