Hindawi
Mobile Information Systems
Volume 2022, Article ID 6542300, 10 pages
https://doi.org/10.1155/2022/6542300

Research Article

Achievement Prediction of English Majors Based on Analytic Hierarchy Process and Genetic Algorithm

Guannan Li[1] and Wenyu Gao[2]

[1]*Department of Foreign Languages, Hebei Normal University, HuiHua College, Hebei 050000, China*
[2]*School of Foreign Languages, Jiangsu Commercial Vocational College, Nantong 226019, China*

Correspondence should be addressed to Wenyu Gao; 2014089@jsbc.edu.cn

Received 16 April 2022; Revised 24 May 2022; Accepted 30 May 2022; Published 7 July 2022

Academic Editor: Wen Zhou

The prediction and analysis of student achievement aim to realize personalized guidance for students and improve student achievement and teacher's teaching achievement. Student achievement is affected by many factors such as family environment, learning conditions, and individual performance. Traditional prediction methods often ignore that different factors have different effects on the same student's score, and different students have different effects on the same factor, so the model constructed cannot realize personalized analysis and guidance for students. Therefore, this paper proposes a prediction model based on the analytic hierarchy process and genetic algorithm. Firstly, according to the relationship among different levels, the analytic hierarchy process (AHP) model is established. Then, a k-means clustering algorithm is used to process the experimental data. Secondly, in order to get rid of the negative impact of the randomness of the initial threshold and weight on model prediction accuracy, which leads to the prediction result falling into a local minimum, a genetic algorithm is proposed to find the optimal initial threshold and weight of model first. Finally, a prediction model based on the BP neural network is established to predict students' scores, which proves that the prediction effect is good. The experiment was conducted with English major students in a university as the research object. Experimental results show that compared with traditional data mining methods, the proposed method has better prediction accuracy.

1. Introduction

Education is the foundation of a nation and the foundation of a strong nation. With the rapid development of Internet technology, it becomes more convenient and fast to collect education-related data. The analysis, mining, and application of education big data are an important demand and inevitable trend [1]. Student achievement prediction, also known as student academic achievement prediction, refers to the use of students' relevant information to predict their future academic performance. It includes course scores, comprehensive scores at the end of the semester, and whether there is a risk of dropping out. With the help of student achievement prediction technology, teachers can have a clear insight into students' learning status and quality, carry out differentiated teaching based on this, meet students' personalized learning needs, and truly achieve the goal of "promoting learning through evaluation." The prediction technology of students' scores is also helpful for colleges and universities to carry out academic early warning, especially to establish dynamic early warning mechanism according to the real-time prediction results of students' scores, so as to timely find the students who may not be able to finish their studies normally, guide them out of trouble, and successfully achieve the goal of talent training. Therefore, no matter from the perspective of improving teaching effect or strengthening student management, student achievement prediction technology has important research value and practical significance [2].

In recent years, the prediction of student achievement has been widely concerned by scholars at home and abroad, and a series of fruitful research work studies has emerged.

Most of the early studies focused on pedagogy and psychology, trying to explore the key factors affecting students' academic performance, such as personality composition, learning motivation, family environment, etc. [3]. This kind of research is mainly based on the self-assessment reports provided by some students, which has great defects in sample size and timeliness, and the conclusions are also susceptible to the influence of the subjective consciousness of the interviewed individuals. In some studies, students' performance information in the learning process is used to predict students' final grades [4], such as attendance, homework completion, and stage test scores. Because there is a strong correlation between the performance information of the course learning process and the final score, the obtained model can often achieve better predictive performance [5]. However, such research can be carried out after the course is carried out for some time, so it is impossible to predict students' learning performance at the initial stage of the course, leading to a certain lag in the predicted results.

At present, the prediction and analysis of student achievement and the research on the key influencing factors have attracted the attention of scholars at home and abroad. In terms of student performance prediction, literature [6] selected several typical learning behavior characteristics from many behavioral characteristics of MOOC learners. And use the selected characteristics to predict whether learners can successfully complete the learning task and obtain the certificate to find out the potential serious learners. Literature [7] selects 8 important attributes by calculating the information gain rate of each attribute characteristic from 18 attributes that affect students' performance, and uses the selected 8 important attributes to construct a decision tree to predict students' performance. In terms of the mining of influencing factors of students' scores, literature [8] conducted a study on the scores of 300 students in an Indian university and found that students' scores were greatly affected by such factors as home address, annual family income, mother's education, living habits, and students' historical scores. Literature [9] proposed that students' sociodemographic characteristics (such as race, gender, and economic status) and academic characteristics (such as school type and school performance) are closely related to their academic performance. Although the above work has made a good performance, but there are still two problems. (1) Only considers the current work has the characteristics of the selected influence on student achievement and ignores the influence of the selected features. (2) The current work assumes that the key factor to the influence degree of all the students is the same, ignoring the students' individual differences. In fact, different factors have different effects on the same student's score, and different students have different effects on the same factor.

Educational data mining aims to discover the internal connections and rules hidden in massive educational data and provide some help for students' learning, teachers' teaching, and the management of education managers [10]. Student achievement prediction can help teachers timely and effectively intervene and guide students' learning process, such as identifying students at risk so as to provide timely intervention measures. In addition, it can also be used in the online evaluation, cognitive diagnosis, student portrait construction, and recommendation system, which has important research significance and application value.

With the rise of data mining technology, a large number of data mining methods have been applied to the study of student achievement prediction. Existing research methods can be divided into two categories: one is to regard prediction problems as regression or classification problems. Data mining models such as linear regression [11], decision tree [12], support vector machine [13], deep neural network [14], and Bayesian network [15] are used in the literature [16]. On the other hand, the student prediction problem is likened to the user evaluation problem in the recommendation system, and the technology in the recommendation field is borrowed to solve the problem, including collaborative filtering, matrix factorization (MF) [17], and other methods. Compared with regression-based methods, recommendation-based methods are more widely used because of their higher prediction accuracy and interpretability.

However, recommendation-based approaches tend to perform poorly in the absence of historical data. Because this kind of method mainly relies on the historical record of students' scores to mine the similarity of courses, and then predicts the results. Therefore, when the number of history courses is small, additional information must be used to help accurately depict the similarity between courses. These background information are usually miscellaneous, have high requirements for data sources, and have limited mining of knowledge information. So far, there is no research that relies on knowledge information to predict students' performance. In view of this, this paper proposes a student achievement prediction model based on the analytic hierarchy Process and genetic algorithm.

The innovations and contributions of this paper are listed as follows:

(1) According to the relationship among different levels, the analytic hierarchy process model is established, and then the k-means clustering algorithm is used to process the experimental data

(2) Genetic algorithm is proposed to find the optimal initial threshold and weight of the model first

(3) A prediction model based on the BP neural network is established to predict students' scores

This paper consists of five main parts: the first part is the introduction, the second part is state of the art, the third part is a methodology, the fourth part is result analysis and discussion, and the fifth part is the conclusion.

2. State of the Art

Taking English major students in a university as the research object, this paper conducts a questionnaire survey on their academic performance. The design of the questionnaire was carried out in accordance with the principles of clear theme, reasonable structure, easy to understand,

appropriate control of the length of the questionnaire, convenient data verification, sorting, and statistics.

A questionnaire survey is conducted on the influencing factors of English major students' scores, which is divided into three levels, namely, individual, school, and family. The corresponding specific influencing factors are set in each layer with different numbers. The questions set in this questionnaire are mainly as follows

The influencing factors at the individual layer mainly include gender, class acceptance, taking notes or not, course interest, the length of study after class, and own health status

The influencing factors at the school layer mainly include the academic atmosphere of the school, teaching mode of teachers, faculty, and equipment resource sharing

The influencing factors of the family layer mainly include the education level of parents, family environment relationship, and family economic level

The questionnaire is mainly investigated from three aspects, each of which contains specific factors affecting English major students' scores. There are 13 question factors in total. A total of 202 questionnaires were collected for this survey, and 18 invalid papers were screened out after statistics and collation of the data in the later stage. Finally, 184 pieces of data could be used as data sources for this survey. Reliability and validity tests of the questionnaire data showed good performance.

3. Methodology

3.1. Establishment of the Hierarchy of Influencing Factors of English Major Scores. The influencing factors of English major scores can be divided into three levels: target layer A, criterion layer B_i, and subcriterion layer C_y. The target layer is the influencing factor of English major students' performance. The criterion layer is divided into three factors, namely, the influence factors of individual, school, and family on English major students' performance. The three criteria layers are decomposed into more subcriteria layers, for example, the individual layer considers gender, class acceptance, taking notes or not in class, interest in English, etc. The school layer considers the teaching mode of teachers, the academic atmosphere of the school, equipment resource sharing, and the faculty. At the family layer, the education level of parents, family economic level, and family environment relationship (see Table 1).

The factors in the criterion layer have an effect on the factors of the upper layer, and all the factors in the subcriterion layer have an effect on the target layer, but for the upper layer, only the factors belonging to the upper layer have an effect, but each factor in the subcriterion layer is independent of each other and does not affect each other.

3.2. Consistency Test and Hierarchical Single Ordering of Judgment Matrix. Through the analysis of English major student performance influence factors, namely, the five criteria layer individual, class, school, family, and society, and the corresponding criterion layer of the factor analysis, respectively, established principles of the target layer, layer of criterion of the judgment matrix, by using mathematical software Matlab to calculate the maximum eigenvalue and eigenvector of a judgment matrix. The feature vectors are normalized, and finally, the consistency test is done.

Table 1: Hierarchy table of factors influencing English major students' scores.

Target layer A	Criterion layer B_x	Subcriteria layer C_y
An influencing factor of English majors' achievement	B_1 individual factors	C_1 gender
		C_2 class acceptance
		C_3 take notes or not
		C_4 own health status
		C_5 course interest
		C_6 length of study after class
	B_2 school factors	C_7 academic atmosphere of the school
		C_8 teaching mode of teachers
		C_9 faculty
		C_{10} equipment resource sharing
	B_3 family factors	C_{11} education level of parents
		C_{12} family environment relationship
		C_{13} family economic level

(1) Judgment matrix A-B_x:

$$A - B_x = \begin{vmatrix} 1 & 3 & 4 & 6 & 5 \\ \frac{1}{4} & \frac{1}{3} & 1 & 4 & 2 \\ \frac{1}{6} & \frac{1}{5} & \frac{1}{4} & 1 & \frac{1}{2} \end{vmatrix}. \quad (1)$$

The maximum eigenvalue of the matrix λ Max = 5.1984, and its corresponding eigenvector is $M_1 = (0.8388, 0.2332, 0.0856)^T$.

The corresponding weight can be obtained by the normalization of M_1:

$M_A = (0.4769, 0.1326, 0.0487)^T$.

Consistency test: the indicators of consistency test are as follows:

$$CX_1 = \frac{\lambda_{\max} - t}{t - 1} = \frac{0.1984}{4} = 0.0496, \quad t = 5. \quad (2)$$

The average consistency index is

$$RX_1 = 1.12,$$

$$\mathrm{CR}_1 = \frac{CX_1}{RX_1} = \frac{0.0496}{1.12} = 0.0443 < 0.1. \quad (3)$$

If the consistency ratio $CR1 < 0.1$, the matrix A-B_x passes the consistency test. It shows that the matrix A-B_x is reasonably constructed without secondary construction.

According to the normalized weight M_A, it can be seen that the eigenvalue 0.0487 is the smallest; that is, family factors have the least influence on the scores of English majors. The characteristic value of 0.4769 is the largest; that is, personal factors have the greatest influence on the scores of English majors.

(2) Judgment matrix B_1-C_y:

$$B_1 - C_y = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{6} & \frac{1}{5} & \frac{1}{7} & \frac{1}{5} & \frac{1}{4} \\ 3 & 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{5} & \frac{1}{3} & \frac{1}{3} \\ 6 & 2 & 1 & \frac{1}{2} & \frac{1}{3} & 1 & \frac{1}{4} \\ 5 & 3 & 2 & 1 & \frac{1}{2} & 1 & 1 \\ 7 & 5 & 3 & 2 & 1 & \frac{1}{3} & 1 \\ 5 & 3 & 1 & 1 & 3 & 1 & 2 \end{bmatrix}. \quad (4)$$

It can be concluded from the above matrix that when λ Max = 7.6323, $M_2 = (0.0692, 0.1356, 0.2595, 0.3829, 0.5017, 0.5683)^T$.

After normalization, the resulting vector is as follows: $M_{B_1} = (0.0294, 0.0576, 0.1103, 0.1627, 0.2132, 0.2416)^T$; then, the consistency test index is obtained.

$$\begin{aligned} CX_2 &= \frac{\lambda_{\max} - t}{t-1} = \frac{0.6323}{6} = 0.1054, \quad t = 7, \\ RX_2 &= 1.36, \\ \mathrm{CR}_2 &= \frac{CX_2}{RX_2} = \frac{0.1054}{1.36} = 0.0775 < 0.1. \end{aligned} \quad (5)$$

If the consistency ratio $\mathrm{CR}_2 < 0.1$, the matrix B_1-C_y passes the consistency test, indicating that the matrix B_1-C_y is reasonably constructed and no secondary construction is required.

According to the normalized weight M_{B1}, it can be seen that the eigenvalue 0.0294 is the smallest; that is, gender factors have the least influence on the scores of English majors. The characteristic value of 0.2416 is the largest; that is, the length of study after class has the greatest impact on the scores of English majors. The second factor is course interest with a weight of 0.2132, and their health status also plays an important role in English major students' performance.

(3) Judgment matrix B_2-C_y:

$$B_2 - C_y = \begin{bmatrix} 1 & 3 & 4 & 6 & 6 & 6 & 7 \\ \frac{1}{4} & 1 & 1 & 3 & 2 & 2 & 2 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{2} & 1 & 1 & 1 & 1 \\ \frac{1}{7} & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 1 & 1 & 1 \end{bmatrix}. \quad (6)$$

It can be concluded from the above matrix that when λ Max = 7.1067, $M_3 = (-0.8703, -0.2807, -0.1357, -0.1159)^T$.

After normalization, the resulting vector is $M_{B2} = (0.4387, 0.1415, 0.0684, 0.0584)^T$.

Then, the consistency test index is obtained as follows:

$$\begin{aligned} CX_3 &= \frac{\lambda_{\max} - t}{t-1} = \frac{0.1067}{6} = 0.0178, \quad t = 7, \\ RX_3 &= 1.36, \\ \mathrm{CR}_3 &= \frac{CX_3}{RX_3} = \frac{0.0178}{1.36} = 0.0131 < 0.1. \end{aligned} \quad (7)$$

If the consistency ratio $\mathrm{CR}_3 < 0.1$, the matrix B_2-C_y passes the consistency test. It shows that the structure of matrix B_2-C_y is reasonable and no secondary structure is needed.

According to the aforementioned normalized weight M_{B2}, it can be known that $0.0584 < 0.0684 < 0.1415 < 0.4387$.

Therefore, it can be seen that the academic atmosphere of the school has the greatest weight, followed by the teaching mode of the teacher, followed by the faculty, and the equipment and resources of the school have the lowest impact on the scores of English majors.

(4) Judgment matrix B_3-C_y:

$$B_3 - C_y = \begin{bmatrix} 1 & 2 & 3 \\ \frac{1}{3} & 1 & 2 \\ \frac{1}{2} & 2 & 1 \end{bmatrix}. \quad (8)$$

From the above matrix, it can be concluded that when λ Max = 3.0889, $M_4 = (0.8500, 0.4287, 0.3061)^T$.

After normalization, the resulting vector is $M_{B3} = (0.5363, 0.2705, 0.1931)^T$. Then, the consistency test index is obtained as follows:

$$CX_4 = \frac{\lambda_{\max} - t}{t-1} = \frac{0.0889}{2} = 0.0445, \quad t = 3. \quad (9)$$

The average consistency index is

$$\begin{aligned} RX_4 &= 0.58, \\ \mathrm{CR}_4 &= \frac{\mathrm{CI}_4}{\mathrm{RI}_4} = \frac{0.0445}{0.58} = 0.0767 < 0.1 \end{aligned} \quad (10)$$

Table 2: Composite weight table of influencing factors of English major scores.

Target layer A	Criterion layer B_x	Weight M_A	Subcriteria layer C_y	Weight M_B	Weight M_C
An influencing factor of English majors' achievement	B_1 individual factors	0.4769	C_1 gender	0.0294	0.0241
			C_2 class acceptance	0.0576	0.0476
			C_3 take notes or not	0.1103	0.1105
			C_4 own health status	0.1627	0.1065
			C_5 course interest	0.2132	0.1144
	B_2 school factors	0.1326	C_6 length of study after class	0.2416	0.2012
			C_7 academic atmosphere of the school	0.1549	0.0271
			C_8 teaching mode of teachers	0.1415	0.0162
			C_9 faculty	0.0684	0.012
			C_{10} equipment resource sharing	0.0584	0.0471
	B_3 family factors	0.0487	C_{11} education level of parents	0.5363	0.0143
			C_{12} family environment relationship	0.2705	0.013
			C_{13} family economic level	0.1931	0.0121

If the consistency ratio $CR_4 < 0.1$, the matrix B_3-C_Y passes the consistency test. It shows that the structure of matrix B_3-C_Y is reasonable and no secondary structure is needed.

According to the normalized weight M_{B3} above, the eigenvalue is $0.1931 < 0.2705 < 0.5363$.

It can be seen that at the family layer, parents' educational level has the highest impact on the scores of English major schools, followed by family environment relationship, while family economic level has a low impact on the scores of English major academic schools. Therefore, parents' educational level is also particularly important, which has a greater impact on students' academic performance.

Because the weight ranking of a single layer is relatively simple, it is not possible to comprehensively analyze the influencing factors of college students' higher mathematics academic performance. Therefore, it is necessary to intuitively understand the influence of various factors on English major students' performance through hierarchical overall ranking.

3.3. Hierarchical Total Ranking and Its Consistency Test. The pair comparison matrix is constructed, and the maximum eigenvalue and corresponding eigenvector of the matrix are obtained by Matlab. The eigenvector is normalized, and then the weight M is obtained. M_A is the weight of the criteria layer to the target layer, that is, the influence factors of English major students' scores. M_B is the weight of the subcriteria layer to its corresponding criteria layer. M_C is the weight of all subcriteria layers to the target layer. The synthetic weight of all factors at all levels to the scores of English majors was calculated, and the total ranking of all levels was carried out. The CR value of each matrix is less than 0.1, and the consistency test is passed. Table 2 shows the results of composite weight after the hierarchical total ordering.

It is found that the weight value of the length of study after the class is the highest. The weight value of the academic atmosphere of the school in the school layer is the highest. The weight value of parents' education level is the highest at the family layer.

3.4. BP Neural Network Algorithm. Due to the uncertainty of subjective factors, this paper only considers the influence of objective factors when building the prediction model. To effectively predict the real school English professional level of academic performance, according to the theory of the BP neural network model, select five parameters to build a neural network algorithm model, namely: spare time learning time p, curriculum interest, school academic atmosphere q, a teacher teaching model w, and parents education level x, as shown in Figure 1. In Figure 1, there are 5 input nodes in the model input layer, which are the 5 parameters selected above to participate in model prediction.

3.5. Data Cluster Analysis. According to the original data samples, the data of each dimension has its own change interval and unit; that is, the data of each dimension are data of different attributes. Therefore, in order to improve data quality and accuracy, it is very important to define data in the same interval and minimize the impact of data repetition and redundancy on model calculation while maintaining the original data relationship unchanged during model training. The methods of data clustering are divided into hard clustering and flexible partitioning. In this section, the k-means clustering algorithm in the hard clustering method is selected to preprocess experimental sample data. The original data set I with t objects was divided into K clusters to minimize the distance between each data point and the cluster center. In other words, the cluster analysis was completed and the overall redundancy

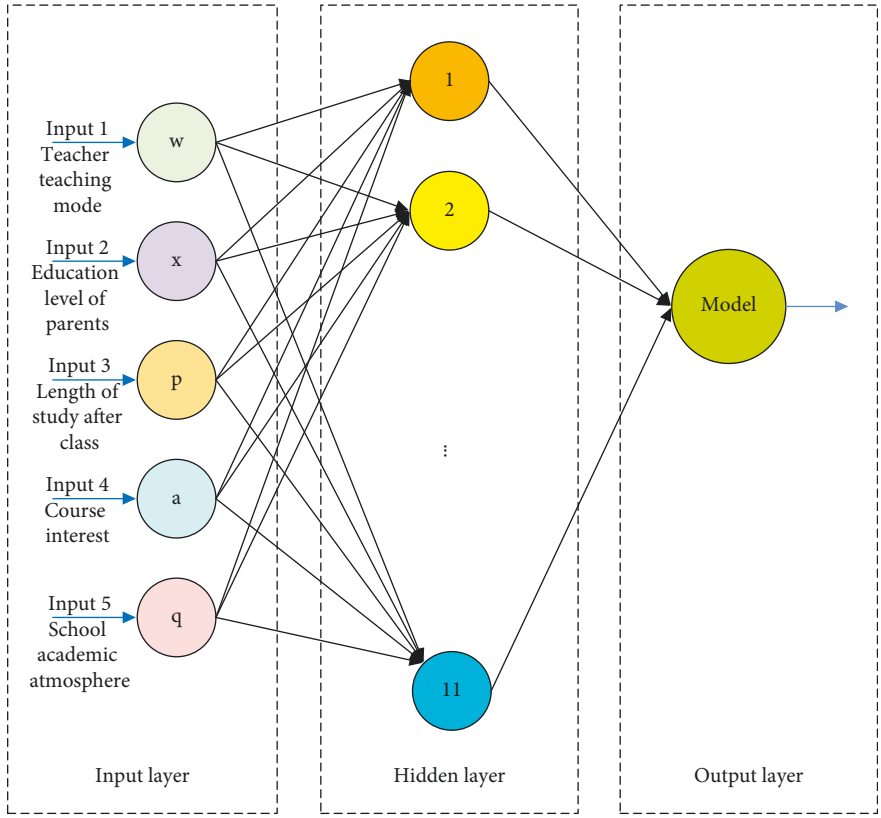

Figure 1: Schematic diagram of the BP neural network model.

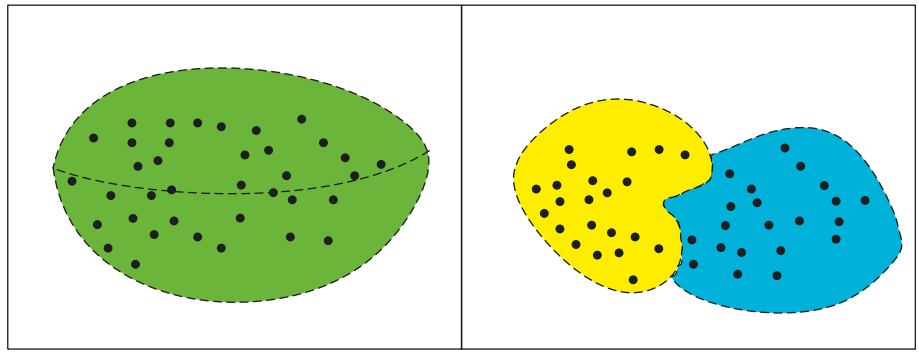

Figure 2: Schematic diagram of K-means clustering algorithm.

was reduced. The steps of this method to analyze indoor environmental data are as follows:

Step 1: initialize the K cluster center:

$$w_1^{(1)}, w_2^{(1)}, \ldots, w_k. \tag{11}$$

Step 2: allocate t data to the cluster set with the least square Euclidean distance from the cluster center, that is, complete the nearest neighbor cluster center. The classification principles are as follows:

$$S_x^{(n)} = \left\{ i_u : \left\| i_u - w_x^{(n)} \right\|^2 \le \left\| i_u - w_y^{(n)} \right\|^2 \forall y, \quad 1 \le y \le z. \right\} \tag{12}$$

Step 3: calculate the new sample center of the cluster set assigned to the node, as shown below.

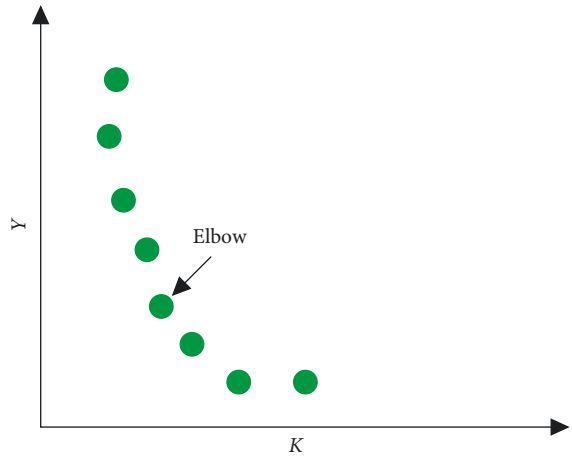

Figure 3: Schematic diagram of Elbow algorithm.

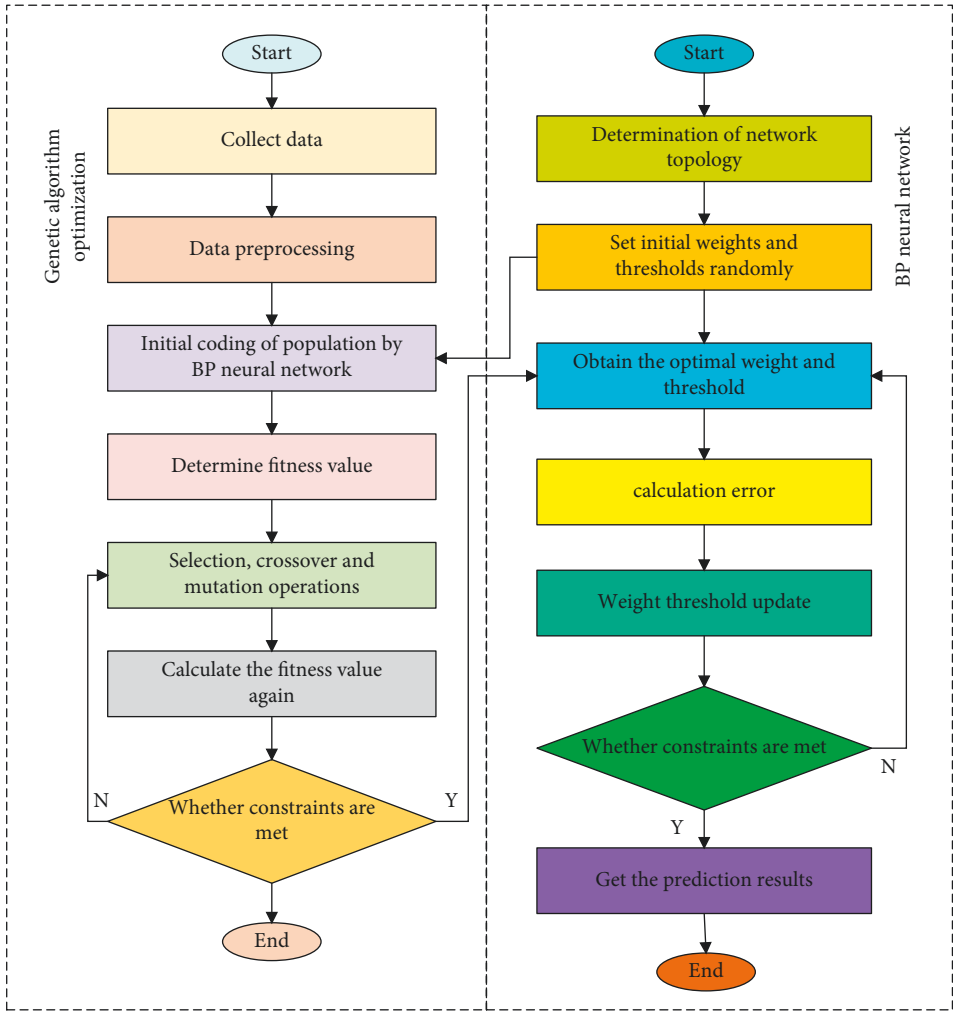

Figure 4: Flow chart of genetic algorithm optimizing BP neural network.

$$w_y^{(n+1)} = \frac{1}{\left|S_x^{(n)}\right|} \sum i_y \in S_x^{(n)} i_y. \quad (13)$$

Step 4: run steps 2 and 3 iteratively alternately until the clustering center does not change or reaches a certain number of iterations, as shown in Figure 2.

In the face of a large number of experimental data, it is difficult to determine the number of clustering K. In this paper, the Elbow method in the K value evaluation algorithm is used to determine it.

In the K-means clustering algorithm, the optimization objective is as follows:

$$Y\left(c^{(1)}, c^{(2)}, \ldots, c^{(w)}, \mu_1, \mu_2, \cdots, \mu_k\right) = \frac{1}{w} \sum_1^w \left(\left\| i^{(x)} - \mu_{c(w)} \right\|\right). \quad (14)$$

where $c^{(x)}$ is the corresponding subscript of the cluster center closest to $i^{(x)}$. μ_k is the clustering center. The optimization objective Y is the sum of the distance from each sample to the cluster center and also represents the clustering error. The smaller the Y value is, the smaller the clustering error is, and the better the clustering effect is.

When the K value is different, the Y value is also different. According to the idea of the elbow algorithm, the classification effect is the best when the value of K is the inflection point of the optimization objective function curve, see Figure 3.

In summary, the k-means algorithm was used to cluster the initial samples, the Y-cost function was used to evaluate the clustering effect, and the optimal K value was determined by the Elbow algorithm. The results show that there is an inflection point when $K = 5{,}928$; that is, there are 5,928 clustering centers.

3.6. Data Standardization Processing. After clustering, each variable is more dependent on its own specific unit property and change interval. In order to avoid this situation, after data clustering, this section conducts standardized data processing. That is, the clustering data is processed again, the weight of the same attribute is assigned, and mapped to the same change interval (0, 1). In this way, the data set quality can be improved again, which is more conducive to modeling, training, and analysis of data in the later period.

Table 3: Parameter table of genetic algorithm improving BP neural network model.

	Parameter category	Parameter value preset
1	Individual length of genetic algorithm	Lenchrom = 10
2	Number of genetic evolution	Maxgen = 90
3	Population size	Sizepop = 15
4	Crossover rate	Pcross = 0.5
5	Variation rate	pmutation = 0.1
6	Maximum number of iterations of neural network	net.trainParam.epochs = 90
7	Training and learning rate of neural network	net.trainParam.lr = 0.2
8	Allowable error range of neural network	net.trainParam.goal = 0.0005

Table 4: Raw data overview.

Basic data information	Mock exam 1 scores	Mock exam 2 scores	Mock exam 3 scores	Mock exam 4 scores	Final scores
The minimum	59	215	159.8	87.5	106
The quantile	328	306	307.5	321	313
The median	373	345.5	342.5	366.5	351.5
The mean	273	348.5	348.4	368	353.4
On the quantile	416.5	384	348.5	412	393
The maximum	550	528.5	535.5	575	522.5

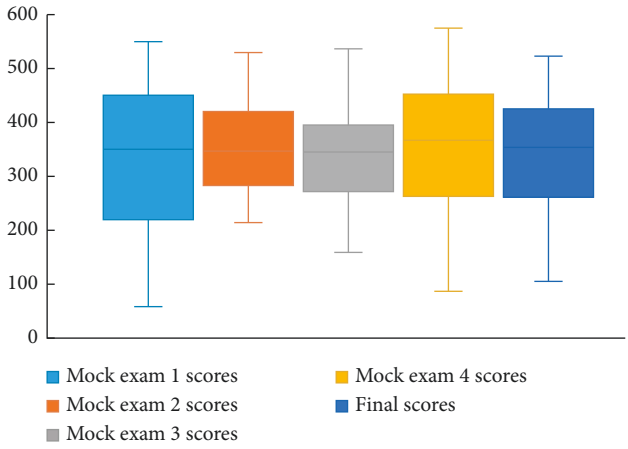

Figure 5: Boxplot of the original variable.

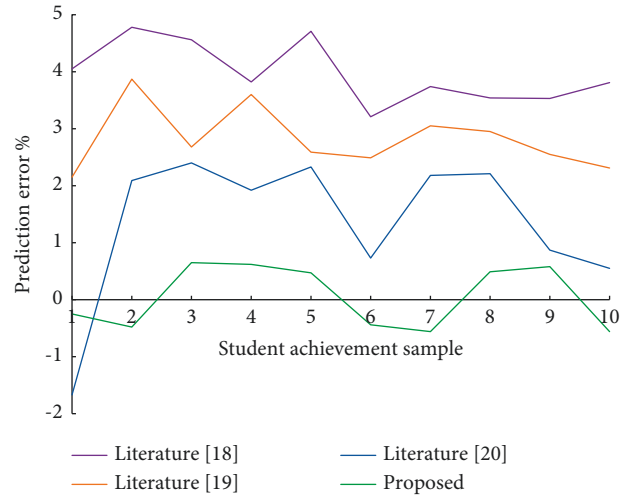

Figure 6: Comparison of prediction errors of various models.

3.7. BP Neural Network Model Training and Prediction Analysis. In order to solve the problems caused by the defects of the BP neural network mentioned above, this paper considers a genetic algorithm to directly use the fitness function as search information. The search process is not constrained by the continuity of a function and has good global search ability, which can overcome the problem that the BP neural network easily falls into the local minimum and find the optimal value of the BP neural network quickly and accurately. Therefore, the genetic algorithm is adopted to optimize the BP neural network first and strive to improve the accuracy of the prediction model, so as to realize an accurate grasp of the scores of English majors. The optimization process is shown in Figure 4.

Taking the initial threshold and weight of the BP neural network as the initial population, the MATLAB GA toolbox is used to optimize it. Parameter settings of the genetic algorithm optimization model are shown in Table 3.

4. Result Analysis and Discussion

Some factors in the above analytic hierarchy process modeling are obtained by calculating the synthetic weight of all factors at all levels to the scores of English majors in a certain university and making a total ranking of all levels. The empirical analysis data came from the results of four mock exams scores and the final exam score of English majors in this university, with a sample size of 669 and a data dimension of 669×5. The source of students in the school is medium level in the province, and the college entrance examination scores have been relatively stable for many years. The data are real and reliable and have strong representativeness. Table 4 provides an overview of the raw data. It can be clearly seen from Table 4 that the mean value of $X1$ is 273, and the mean value of the other three mock exams and the final exam is about 350. The median of the

Table 5: Comparison results of average relative errors of various models.

Prediction model	Test sample 1 (%)	Test sample 1 (%)	Test sample 1 (%)	Average relative error (%)
Literature [18]	15.63	16.81	5.56	12.67
Literature [19]	2.90	7.50	18.37	9.59
Literature [20]	17.44	0.11	14.01	10.28
Proposed	4.21	7.93	7.40	6.51

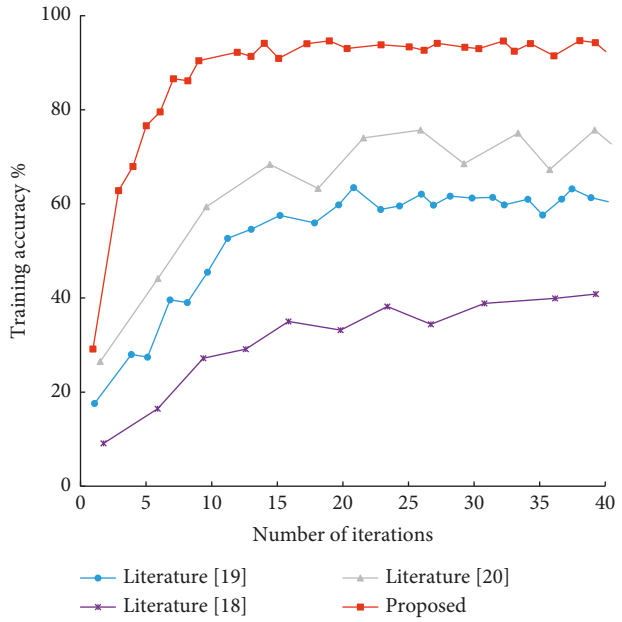

Figure 7: Training accuracy curve of each model.

five variables is around 360, their maximum value is around 540, and their variance is stable around 75. Furthermore, it was observed from the boxplot (see Figure 5) that the range of predicted variables and the final exam score data were basically the same, as well as the variance, so there was no need to standardize the variables during modeling.

Figure 6 shows the prediction results of 10 sample scores randomly selected from English majors in a university in [18], [19], [20], and the proposed model.

The smaller the average relative error is, the higher the accuracy of the model is and the better the prediction performance is. As can be seen from Table 5, among the four prediction models, the average relative error of the model in this paper is the smallest, which is 6.51%. This shows that the prediction results of the model are more consistent with the real value. Thus, the model in this paper can eliminate redundancy between data. The main reason is that genetic algorithm automatic parameter optimization can make support vector machines have better performance, so the prediction result is better.

This section also explores the influence of iteration times on the prediction accuracy of the model. As shown in Figure 7, the training accuracy of the algorithm in this paper tends to be stable after about 10 cycles of iteration, and the model begins to converge.

5. Conclusion

Student achievement prediction is a research hotspot in the field of educational data mining in recent years and is also one of the important objectives of learning analytics. In view of the problem that the influence degree of different factors on the same student's score is not considered in the current relevant research, and the influence degree of different students by the same factor is also different, this paper proposes a student's score prediction model based on AHP and genetic algorithm. Firstly, a questionnaire survey was conducted on the factors influencing English major students' scores. According to the relationship between different levels, an analytic hierarchy process model was established, and then the k-means clustering algorithm was used to process the experimental data. Finally, BP neural network algorithm improved by the genetic algorithm is used to predict students' grades, and more accurate and reliable prediction data are obtained. The prediction results of English majors in a university proved to be effective. Experimental results show that compared with traditional data mining methods, the proposed method has better prediction accuracy. The prediction accuracy of the model needs to be further improved. Because there are many subjective and objective factors affecting students' scores, it is necessary to carefully screen and extract levels, so as to predict students' scores more accurately.

Data Availability

The labeled data set used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by (1): A Research on the Effective Cultivation of English Majors' Comprehensive Capacities from the Perspective of the Multimodality Teaching Design Based on POA 2020YYJG069 and (2) Research on the Way to Improve the Effectiveness of Cross-Cultural Communication Virtual Experiment Teaching Based on Network Platform (2021-R-92296) issued by Institute of modern Educational Technology, Jiangsu Academy of Educational Science, which were funded by National and Regional Institute, Nantong University, and Jiangsu Federation of Social Science.

References

[1] C. Romero and S. Ventura, "Educational data mining and learning analytics: an updated survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, Article ID e1355, 2020.

[2] Y. C. Yeh, O. M. Kwok, H. Y. Chien, N. W. Sweany, E. Baek, and W. A. McIntosh, "How college students' achievement goal Orientations predict their expected online learning outcome: the mediation roles of self-regulated learning strategies and supportive online learning behaviors," *Online Learning*, vol. 23, no. 4, pp. 23–41, 2019.

[3] O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, A. J. Q. Lin, H. Ogata, and S. J. H. Yang, "Applying learning analytics for the early prediction of Students' academic performance in blended learning," *Journal of Educational Technology & Society*, vol. 21, no. 2, pp. 220–232, 2018.

[4] M. Saqr, U. Fors, and M. Tedre, "How the study of online collaborative learning can guide teachers and predict students' performance in a medical course," *BMC Medical Education*, vol. 18, no. 1, pp. 1–14, 2018.

[5] A. Abu Saa, M. Al-Emran, and K. Shaalan, "Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques," *Technology, Knowledge and Learning*, vol. 24, no. 4, pp. 567–598, 2019.

[6] C. H. Yu, J. Wu, and A. C. Liu, "Predicting learning outcomes with MOOC clickstreams," *Education Sciences*, vol. 9, no. 2, p. 104, 2019.

[7] M. Zaffar, K. S. Savita, M. A. Hashmani, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students academic performance," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 5, pp. 541–549, 2018.

[8] R. Naylor, "Key factors influencing psychological distress in university students: the effects of tertiary entrance scores," *Studies in Higher Education*, vol. 47, no. 3, pp. 630–642, 2022.

[9] K. U. Lazarus, "Socio-demographic factors affecting Reading comprehension achievement among secondary school students with learning Disabilities in Ibadan, Nigeria," *IAFOR Journal of Education*, vol. 8, no. 1, pp. 145–158, 2020.

[10] Y. K. Salal, S. M. Abdullaev, and M. Kumar, "Educational data mining: student achievement prediction in academic," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 4C, pp. 54–59, 2019.

[11] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. Niakan Kalhori, "Predicting COVID-19 Incidence through analysis of Google trends data in Iran: data mining and deep learning Pilot study," *JMIR public health and surveillance*, vol. 6, no. 2, Article ID e18828, 2020.

[12] H. H. Patel and P. Prajapati, "Study and analysis of decision tree based classification algorithms," *International Journal of Computer Science and Engineering*, vol. 6, no. 10, pp. 74–78, 2018.

[13] M. Mohammady, H. R. Pourghasemi, and M. Amiri, "Assessment of land subsidence susceptibility in Semnan plain (Iran): a comparison of support vector machine and weights of evidence data mining algorithms," *Natural Hazards*, vol. 99, no. 2, pp. 951–971, 2019.

[14] B. K. Francis and S. S. Babu, "Predicting academic performance of students using a hybrid data mining approach," *Journal of Medical Systems*, vol. 43, no. 6, p. 162, 2019.

[15] S. Hussain, N. Abdulaziz Dahan, F. M. Ba-Alwi, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, p. 447, 2018.

[16] A. V. Manjarres, L. G. M. Sandoval, and M. S. Suárez, "Data mining techniques applied in educational environments: literature review," *Digital Education Review*, vol. 33, pp. 235–266, 2018.

[17] I. Fernández-Tobías, I. Cantador, P. Tomeo, V. W. Anelli, and T. Di Noia, "Addressing the user cold start with cross-domain collaborative filtering: exploiting item metadata in matrix factorization," *User Modeling and User-Adapted Interaction*, vol. 29, no. 2, pp. 443–486, 2019.

[18] S. U. Damuluri, K. Islam, P. Ahmadi, and N. S Qureshi, "Analyzing Navigational data and predicting student Grades using support vector machine," *Emerging Science Journal*, vol. 4, no. 4, pp. 243–252, 2020.

[19] S. Sahoo, M. Mohanty, and S. Sabut, "Automated ECG beat classification using DWT and Hilbert transform-based PCA-SVM classifier," *International Journal of Biomedical Engineering and Technology*, vol. 32, no. 3, p. 287, 2020.

[20] Y. Cao, K. Yin, C. Zhou, and B. Ahmed, "Establishment of landslide groundwater level prediction model based on GA-SVM and influencing factor analysis," *Sensors*, vol. 20, no. 3, p. 845, 2020.