Study on Prediction of Course Failure Based on Improved Bagging-C4.5 Algorithm

Jing Luo
Guangzhou Nanyang Polytechnic College, Guangzhou, 510925, China
e-mail: jingluo@gznylg.edu.cn

Abstract. At present, the research of academic early warning system needs to know the situation of course failing. There are many research methods of course failing, but there are defects of low accuracy and long prediction time. In order to improve the accuracy and accuracy of course failing prediction, the improved Bagging-C4.5 algorithm is designed to predict the course failing. Firstly, the research progress of current social talent demand and course failure prediction is analyzed, then the test data set is determined, and the feature combination with the most influence is selected as the feature set of course failure prediction. Finally, the improved Bagging-C4.5 algorithm is used to realize the failure prediction, and the comparison test is conducted with other algorithms. The results show that the algorithm in this paper has high accuracy and accuracy in predicting course failure.

1. Introduction

With the development of socialist market economy, the talents trained by schools are closely combined with the needs of the society, but there are also some situations where supply and demand are not completely consistent. In order to solve the problem that the school training talents is consistent with the social demand, it is imperative to use the academic warning and monitoring model, among which the prediction of course failure is the primary problem of the research.

Course failure prediction refers to the statistics of students' examination results of various subjects, and the results obtained from the first courses can predict the failure of the relevant follow-up courses. The prediction model of course failure can be obtained by selecting a suitable algorithm in the research. Through this model, students' subsequent academic performance can be predicted, even to the specific risk of failing a course. By predicting the results, remind students to strengthen their study, pay attention to the course, and avoid failing. Supervise students, improve students' performance, discover students' professional ability, targeted training.

However, the existing failure prediction models mainly focus on K-means failure index feature mining. This algorithm is to study the characteristics of failing courses, to mine the characteristics of courses, to study the correlation between them, and to analyze the courses that may fail through the study of association rule algorithm. With this method, the prediction time is long and the prediction accuracy is not high when the tuple with incomplete data is encountered. Therefore, high precision and high speed course failure prediction still needs to be studied.

In order to solve the problems in the prediction of course failure, this paper proposes an optimized Bagging-C4.5 algorithm. This algorithm is based on the first-generation C4.5 algorithm with the addition of Bagging algorithm. This algorithm avoids a single classifier and uses random sampling to obtain n data sets, and then obtains n learners from the data sets. Then combine these learners to get a strong classifier. Bagging-C4.5 algorithm can improve the accuracy of prediction, but too many
learners are designed, resulting in slow operation and low efficiency. Finally, we decide to add genetic algorithm to Bagging-C4.5 algorithm. The algorithm can reduce the occupied storage space, improve the prediction accuracy, improve the prediction speed, can better achieve the course failure prediction.

2. Theoretical basis
In recent years, most researchers have studied data mining algorithms to analyze data. This paper involves three kinds of algorithms, namely decision tree algorithm C4.5, Bagging-C4.5 algorithm and the improved Bagging-C4.5 algorithm. The specific contents are as follows:

2.1. Course failure prediction system
The failure prediction system consists of the following parts: using students' grades in the first course, predicting grades in related subsequent courses, and analyzing which subjects are likely to fail. In the course analysis, courses related to the pre-learning courses, course average score, make-up average score and failure rate are selected. Firstly, the four kinds of data sets are analyzed. Secondly, the data are divided into training group, verification group and test group. Finally, the prediction model of students' course failure is formed through the testing of bagging algorithm and genetic algorithm.

2.2. Bagging - C4.5 algorithm
C4.5 algorithm can process incomplete data and is effective for both discrete data and continuous data. However, the predicted course accuracy of this kind of classifier is not high, the speed is general, and the operation process is too complicated. Therefore, Bagging algorithm is integrated on C4.5 algorithm. Bagging algorithm Bootstrap the training data set for N times to get N subsets of training data. The same algorithm is used to build decision trees for each subset, and the final classification result is the majority vote or average of the results of N decision trees[1]. The combination of Bagging algorithm and C4.5 algorithm can improve the accuracy and stability of course failure prediction. Overfitting is avoided by reducing the variance of the results. In the experiment, Bagging-C4.5 algorithm is used to predict the failure of a course of Computer Application Major of School of Information Engineering. The specific operation process is as follows:

2.2.1. According to the actual situation of the college professional to get the data information, obtain A data set of size n, determine the number of iterations is required t, at the same time, can be increasing number starting from one.

2.2.2. The initial data set obtained in the first step is sampled from data set A in a random, uniform, and put-back manner. M subsets A of size n are selected as the new training set. By using algorithms such as classification and regression on these m training sets, M models can be obtained.

2.2.3. You can use the change iteration data to get the basic classifier. If the number of iterations is less than t, increase the number of iterations by 1 and return to the second step to continue to get the basic classifier[2].

2.2.4. Will be obtained before operation based classifier combination, through and through methods such as average, take majority for data sets, the output combination, get strong classifier, can get a Bagging - C4.5 results. The results of this experiment can be used to predict course failure.

2.3. Genetic Algorithm
GA is a computational model of biological evolution process which simulates natural selection and Genetic mechanism of Darwin's biological evolution theory. It is a method to search for the optimal solution through simulating natural evolution process[3]. The genetic algorithm takes all the individuals in a population as the object and uses the randomization technique to guide the efficient search of a coded parameter space. Among them, selection, crossover and mutation constitute the
genetic operation of genetic algorithm. The core content of genetic algorithm is composed of parameter coding, initial population setting, fitness function design, genetic operation design and control parameter setting[4].

2.3.1. The process of genetic algorithms
- Adopt binary coding scheme in the primary population.
- After the appropriate decoding process, a fitness assessment is made for each individual gene using the fitness function[5].
- The expected value method and sorting selection method were used to select the excellent individuals from all the individuals.
- Extract part of the structure of both sides of excellent individuals, and cross it to produce new individuals.
- To produce random changes in part of the structure of the individual, and eventually form a new individual.
- Repeat steps 2, 3, 4, and 5 until a new population is created, ending the cycle.

2.4. An improved Bagging-C4.5 algorithm
After the Bagging-C4.5 combined algorithm was tested, the same base classifiers were found, and these base classifiers occupied most of the storage space, which led to the slow test speed. In order to improve the accuracy of course failure prediction, the improved Bagging-C4.5 algorithm is added to the Bagging-C4.5 algorithm.

3. The algorithm process
Bagging-C4.5 model was selected to create the initial population, and genetic algorithm was used to optimize it. The course failure prediction final algorithm process is as follows:

3.1. Determine research objects and obtain training samples.
Data set \( A \) was formed by selecting the data of grades when entering the first year and those during the university.

\[
A = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\}
\]  
(1)

In Formula (1), Where \( i \) represents \( i \) students, and \( x_i \) represents course average score, make-up average score and failure rate. \( y_i \) stands for failing a related course. The M subtraining set was obtained by sampling data set \( A \) in a random, uniform and put-back manner.

3.2. Create the primary population.
C4.5 algorithm is used to train a basic classifier from each sub-training set \( A_i \), and \( B(1) \), \( B(2) \) and \( B(3) \) are obtained. Each basic classifier is encoded with binary code. If it is zero, it will be canceled, and if it is one, it will be selected, forming the first-generation population.

3.3. The fitness of the individual was assessed using a combinatorial predictive function.
Combination prediction model is one of the best methods to improve the prediction accuracy. The formula is as follows:

\[
f = \frac{\sum z_i}{n}
\]  
(2)

\[
z_i = \begin{cases} \ 1, & p(T_i) = p(y_i) \\ 0, & otherwise \end{cases}
\]  
(3)
The combined prediction model adopts the weighted average of the single prediction model. In Formula (2), \( n \) represents the number of samples, and \( f \) represents the average value. The part where the weighting coefficient is 1 in Formula (3) means that the classification result is equal to the actual result.

3.4. The best preservation strategy was used to obtain the next generation with the highest accuracy. The optimal preservation strategy is adopted to replace the individuals with the highest fitness in the current population with the individuals with the lowest fitness after crossover operation and mutation operation in the current generation population [6]. The individuals with the highest accuracy were eventually passed on to the next generation.

3.5. Single-point crossover is used to obtain a new strong classifier. In binary coding, a random point is selected and bounded by the exchange of variables. After successful interchange, new individuals are formed.

3.6. Mutation Revisited was used to obtain the new strong learner. The two coding points are randomly selected in the previous data set, and then the part between them is reversed to carry out mutation operation. When the mutation succeeds, a new individual is formed.

3.7. At the end of operation, the optimal combination is obtained to form the final failure prediction model. The loop operation is carried out from 3.3 to 3.6, and the algorithm ends when the termination condition is reached. Through the improved Bagging-C4.5 algorithm, the optimal course failure test model is completed.

4. Performance test of improved Bagging-C4.5 algorithm

4.1. Acquire test objects and data collection
The test subjects were sophomores majoring in computer application in the School of Information Engineering. First of all, the basic information table of student scores is established, which includes student number, major, class, semester, course name, course attribute, total evaluation score and make-up examination score. The details are as follows:

| Student id | professional | The class | semester | Course name          | Class attribute | Overall performance | Make-up exam scores |
|------------|--------------|-----------|----------|----------------------|-----------------|---------------------|---------------------|
| 201901*01  | Computer     | 19th level| 2019-2020| Sports and health    | Public compulsory | 80                  | null                |
| 201901*02  | Computer     | 19th level| 2019-2020| Sports and health    | Public compulsory | 78                  | null                |
| 201901*03  | Computer     | 19th level| 2019-2020| Information technology| Public compulsory | 89                  | null                |
| 201901*25  | Computer     | 19th level| 2019-2020| Information technology| Public compulsory | 35                  | 65                  |
| 201901*26  | Computer     | 19th level| 2019-2020| Information technology| Public compulsory | 90                  | null                |

After obtaining the student's first semester grade, the student's admission test score information is obtained from the Admissions Office, with details as follows:
Table 2. Student admission information table

| Candidate number | gender | category | Total grade | city     | The examinee category |
|------------------|--------|----------|-------------|----------|-----------------------|
| 16*01            | female | science  | 430         | Guangzhou| The fresh             |
| 16*02            | female | science  | 460         | The fresh| The fresh             |
| 16*03            | male   | science  | 475         | shantou  | The fresh             |
| 16*25            | male   | science  | 300         | shantou  | The fresh             |
| 16*26            | female | science  | 500         | Guangzhou| The fresh             |

The school's enrollment scope is Guangdong Province, so the students' examination subjects, total scores are all the same, the type of data is consistent.

4.2. Data sorting

In the process of data collation, there will be field information loss, inconsistency and other problems, these problems need to be dealt with in time, to avoid information loss leading to the error of grade prediction.

4.2.1. Field information loss handling

The loss of field information is divided into individual loss and the loss of grades of all courses.

If individual scores are lost, first use the data of the teaching and research department to search and fill in the missing information. When the number of missing students in the same course is large, the average score of the same grade and course is used to fill in the missing information.

Students who have lost all grades in the course are required to confirm with the Department of Teaching and Research whether the student has withdrawn or joined the military. The information of such students can be deleted directly.

4.2.2. Handling of inconsistent tuple information

The inconsistency of information is mainly reflected in the inconsistency of course names, which is caused by the revision of teaching plans every year. For example: computer application base change to information technology, sports change to sports and health. These can use the system's find and replace functionality to ensure that the information is consistent.

4.3. Analysis of test results

The failure prediction of a course is to predict whether a subsequent course is likely to fail based on the results of the previous courses. Therefore, we chose four characteristic data of pre-course grades,

Table 3. Course failure predicts results

|                          | GA-Bagging-c4.5 | Baggg-c4.5 | c4.5 |
|--------------------------|-----------------|------------|------|
|                          | Correctness (%) | Precision (%) | Correctness (%) | Precision (%) | Correctness (%) | Precision (%) |
| Basic PHP                | 86.9            | 88.6       | 79.7   | 82.9   | 73.3          | 74.4          |
| Image                    | 82.9            | 84.6       | 79.3   | 83.3   | 74.3          | 76.3          |
| Basic Java               | 81.7            | 84.5       | 75.6   | 86.8   | 73.1          | 85.7          |
| Information              | 81.9            | 66.7       | 81.9   | 65.1   | 72.8          | 56.6          |
| C language               | 84.5            | 70.9       | 83.7   | 67.7   | 79.6          | 60.6          |
course average score, make-up average score and failure rate to analyze which feature has a greater impact on the prediction of failure.

This influential feature data is input into the improved Bagging-C4.5 algorithm to test the accuracy and accuracy values. The details of the form are as follows:

4.4. Course failure prediction accuracy statistics

![Figure 1. Comparison of the prediction accuracy of course failure](image)

The accuracy of failure prediction of all courses was counted, as shown in Figure 1. It can be seen that the average accuracy of the algorithm in this paper is 84, and the prediction accuracy of Bagging-C4.5 algorithm and C4.5 algorithm are 80 and 75 respectively. The algorithm in this paper has a higher accuracy, which improves the accuracy of the prediction of course failure.

4.5. Course failure prediction accuracy statistics

![Figure 2. Accuracy comparison of course failure prediction results](image)

The accuracy of the failure prediction results of all courses was calculated, as shown in Figure 2. It can be seen that the mean accuracy of the algorithm in this paper is 79, and the prediction accuracy of Bagging-C4.5 algorithm and C4.5 algorithm are 77 and 70 respectively. The algorithm in this paper has a higher accuracy, which improves the prediction accuracy of course failure.

However, in terms of analysis on the basis of Java programming, Bagging-C4.5 algorithm optimized by genetic algorithm is slightly lower than Bagging-C4.5 algorithm and C4.5 algorithm, which may be because the failure rate of Java programming is too high, and many excellent students in other courses will fail, so its implicit rules are not obvious and its accuracy is reduced[7].
In terms of accuracy and accuracy, the improved Bagging-C4.5 algorithm is efficient in predicting course failure, which can provide a strong basis for the follow-up study of academic early warning system.

5. Conclusion
In order to improve the prediction results of course failing, the Bagging-C4.5 algorithm based on genetic algorithm is designed. The test results show that the proposed algorithm has high accuracy and accuracy in the prediction of course failing, which has a wide application prospect.

Acknowledgments
2020 university-level scientific research project of Guangzhou Nanyang Polytechnic Vocational College Research and implementation of academic early warning System based on Hadoop Platform （NY-2020KYYB-03）.Research on the Cultivation Mode of Skilled Talents in CIM-CDIO Private Colleges and Universities based on the Information Technology Education Steering Committee in 2018 (XXIZW2018059).2018 Guangzhou Nanyang Polytechnic Vocational College Innovation and External Practice Teaching Base Project (NY-2018 CQ2JD-03).2019 -2020 Vocational Education Teaching Reform Project of Teaching Committee of China Association of Vocational and Technical Education (Course No. : 1910160)

References
[1] Gong, F. (2017) Application of Data Mining in Academic Early Warning of College Students. Electronic Technology and Software Engineering, 04:202-203.
[2] Li, G.h, Wang, F.(2018)Data set optimization based on decision tree generation and pruning and its application. Computer Engineering and Design, 01:205-211.
[3] Wang,D.D.(2020) Evaluation of Scheme Quality Based on Genetic Neural Network Algorithm. Journal of Shanxi Normal University (Natural Science Edition). 04:150-155
[4] Zhang, X.F.(2018)Application of multiple linear regression in the analysis and prediction of student achievement ranking. Fujian Computer, 05:158-164.
[5] Cheng, L.(2019) A study on the pretest model of students based on machine learning. Electron World,24:73-74.
[6] Hu,J.T. (2017)Research and Implementation of Student Achievement Prediction Teaching System Based on C4.5 Decision Tree. Southwest Jiaotong University.
[7] Xing, Y.Y.(2018) Research on the Prediction Method of Academic Performance Ranking Based on Campus One-card Data. Information Technology in Education in China, 09:85-89.