Analysis of Higher Education Teaching Data Based on Data Mining Technology

Wang Zhi Qi*

Criminal Investigation Police University of China, Liaoning Shenyang 110854, China

709060958@qq.com

Abstract. In order to effectively mine and utilize the large amount of valuable data stored in the academic management system of universities, a data mining technology based method for analyzing higher education teaching data has been proposed. This article uses data mining technology to deeply mine and analyze the grade data in the school's academic affairs system. Firstly, the data in the academic affairs system is collected and preprocessed, and then factor analysis is used to comprehensively evaluate student grades. Then, the decision tree improvement method of K-means clustering algorithm and C5.0 algorithm is used to predict the target grades. Finally, the above method is compared and analyzed with other methods. The results show that the estimation accuracy of the improved decision tree method is the highest 64.8%, and the generated decision tree has the smallest depth and the smallest number of leaf nodes. This indicates that the decision Tree model generated by it is more accurate and robust than the other two methods, and can try to avoid over fitting.

Keywords: Data mining; Higher education; C5.0 algorithm; Factorial analysis

1 Introduction

With the development of computer technology, the state has put forward a clear plan for the development of educational informatization, and colleges and universities have gradually integrated computers and corresponding information technologies into the education and teaching management system by building a "internet plus" platform, which greatly improves the ability of the education management system to generate, collect, store and process data, and the data resources are also increasing. Massive data resources are a huge treasure house for monitoring the teaching quality in colleges and universities, and using data mining technology to turn the original data information with thin value into useful resources to guide educational management decisions has become an important measure for colleges and universities to realize information development [1]. Information mining technology is of great significance to the optimization and innovation of teaching quality evaluation in colleges in the development of education and teaching. First, it is advantageous to improve teaching quality in colleges and promote the setting up and development of teaching supervision system in colleges. The application of data mining technology can collect data from various

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_22

aspects, do research and maintenance on multi-source data in level, timely and accurately grasp the actuality of education and teaching in various colleges, use the data to establish the appraisal system, . supervise the development of teaching activities in various colleges, and improve the quality of education and teaching. Secondly, it is necessary to establish a fair and reasonable evaluation system and encourage colleges to continue to increase the efficiency of teaching.

2 Methods

2.1 Data Acquisition and Pretreatment

All the sample data involved in this paper are from the educational administration system of our university, and the final examination results of 17 courses of 186 computer science and technology students in 2018 are taken as the research object. In order to standardize the ownership of the research data and protect the privacy of the data owner, the following related data will be virtualized by means of numbering [2-3].

2.2 Comprehensive evaluation of students' performance based on factor analysis

In this paper, based on the four basic ability requirements for the cultivation of computer professionals put forward by the Computer Science and Technology Teaching Steering Committee of the Ministry of Education (hereinafter referred to as the "Teaching Steering Committee"), the scores of 17 compulsory courses are selected as the research objects, and the students' scores are comprehensively evaluated according to the basic principles and steps of factor analysis. These 17 courses are: computer organization principle (X1), computer circuit foundation (X2), data structure (X3), computer network (X4), discrete mathematics (X5), microcomputer interface technology (X6), operating system (X7), C++ object-oriented programming (X8), database principle and technology (X9). algorithm design and analysis (X10), JSP/ASP WEB technology (X11), JAVA language programming (X12), hardware curriculum design (X13), computer information security (X14), software engineering (X15), programming language curriculum design (X16), C programming foundation (X17) [4-6].

(1) KMO and bartlett sphericity test.

The sampling suitability of KMO and bartlett sphericity test is 0.957. According to the standard given by statistician Kaiser, it can be considered that the data object is suitable for factor analysis.

(2) Extracting common factors.

From the commonality of factor analysis, it can be seen that the extracted values are all between 0.5 and 0.8, and it can be considered that factors can basically explain the variance of various professional courses. In addition, from the explanation of total variance, we can see that the variance contribution rate of the first factor after rotation is

43.347%, and the cumulative variance contribution rate of the three factors is 69. 916%, that is, the three factors explain 69. 916% of the original 17 variables [7].

2.3 Based on the improved decision tree method of student achievement prediction

(1) Introduction of improved decision tree method.

Decision tree is a supervised machine learning method, and the decision tree algorithm used in this paper is C 5.0. The entropy boxing method of MDL P (minimum description length principle) is used in the discretization of numerical variables in C5.0 algorithm, and its core measures are information entropy and information gain. This method is not flexible enough, because it can't customize the boxing rules according to the degree of difficulty and the number of boxing in each course. Therefore, this paper improves it, using K-means clustering algorithm to discretize the performance data of each course, and then predicting the target course performance.

(2) Determine the forecast target.

The prediction goal of this paper is to sort out the professional basic courses and professional core courses in the teaching plan according to the order of starting courses, and then select the course scores offered earlier as explanatory variables to predict the academic performance of higher-level professional courses related to them, find out the students who may fail the courses, and intervene them in advance, so as to achieve the purpose of early warning. For example, taking "computer information security" as the prediction target, comprehensively considering the order of starting semester and course category, and taking the scores of nine courses such as C programming foundation, operating system and discrete mathematics as explanatory variables, the target course scores are predicted.

(3) Data conversion using clustering algorithm.

According to the above, before using C5.0 algorithm to predict the achievement of the target course, the data should be transformed, that is, K-means clustering algorithm is used to discretize the course achievement.

(4) Using C50 algorithm to forecast and analyze.

Load the discretized data into SPSSModeler, and call C5.0 algorithm to model it. Ten-fold cross-validation algorithm is used as the evaluation method of the model, and Boosting algorithm is used as the method to improve the prediction accuracy, so as to obtain the best tree structure. The final result is shown in Figure 1 [8-9].

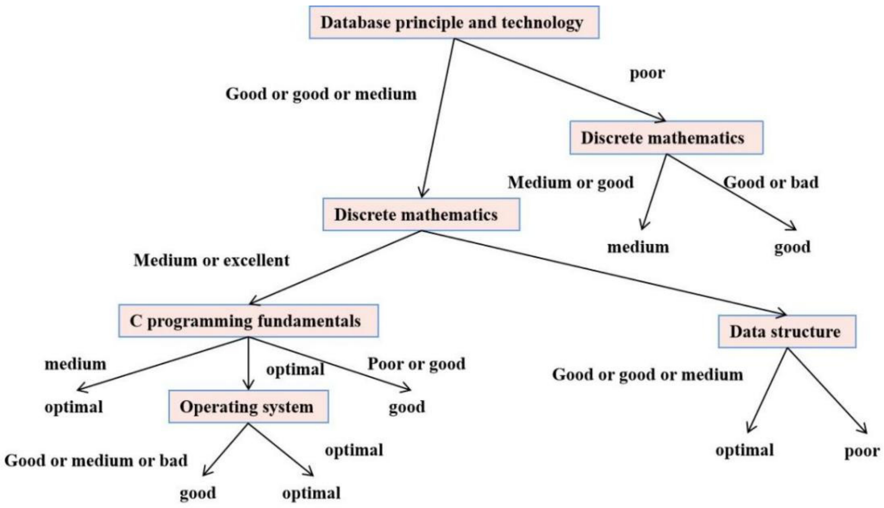

Fig. 1. Decision tree generated by improved method

As shown in Figure 1, "database principle and technology" is the root node of the model, which shows that it is the attribute with the strongest information entropy gain, followed by discrete mathematics, C programming basis and data structure. Therefore, among the nine courses as explanatory variables, "Database Principle and Technology" and "Discrete Mathematics" have the greatest contribution to the prediction of the target course "Computer Information Security", and students who want to achieve excellent results and have poor prediction results should strengthen their study of these two courses. In addition, we can extract rules from the decision tree, that is, the path from the root node to each leaf node of the decision tree is expressed by IF-THEN statement. Because there are many rules extracted according to the decision tree shown in Figure 1, only the rules with the "Computer Information Security" rating of "Poor" are listed below, as follows:

3 Analysis and discussion

3.1 Factor analysis and traditional comprehensive evaluation methods

The comparison of factors is shown in Table 1.

Table 1. Comparison Table of Comprehensive Evaluation

student number	F1 factor score	F1 ranking	F2 factor score	F2 ranking	F3 factor score	F3 ranking	F comprehensive score	F ranking	Total average score	Average ranking 6
f14011632	1.757	1	0.265	123	0.842	97	1. 073	1	92. 1	9
f14011507	1.542	4	1.047	41	1.187	56	1.025	2	95.8	1
f14011611	1.	3	0.801	67	0.578	115	0.992	3	92.4	6

...	...	...	...	...	...	...	...	...	...	...
f14011407	1.211	13	1.631	7	0.987	78	0.866	11	94.3	2

The research can not only sort the students' scores as a whole, but also know the details of each student's knowledge mastery, and know which professional abilities students have developed well and which ones are relatively weak. For example, in Table 1, the F comprehensive scores of two students, f14011632 and f14011507, are similar (First and second place respectively.), but the F1 factor score of f14011632 is higher. It means that their comprehensive ability and computational thinking are relatively strong, while the score of F2 factor is relatively low, which means that their algorithm design and analysis, system cognition, analysis, development and application, and program design and implementation ability are poor, and the situation of students with student number f14011507 is just the opposite. Thus, compared with the traditional average ranking method, factor analysis contains more information to help teachers guide students to adjust the direction of learning and development in time.

It can also be seen from Table 1 that the results of factor analysis are different from those of traditional average scores. For example, the factor analysis of f14011632 ranks first, but the average score ranks ninth. However, f14011407 ranked 11th in factor analysis, but ranked 2nd in average score. After careful observation, it is not difficult to find that f14011407' s scores are relatively uniform, while f14011632' s score of F1 factor is very outstanding, which corresponds to the fact that the weight of each course is the same when ranking by average score, while the weight of factor analysis F is relatively large. It can be seen that the traditional method of ranking by average score does not distinguish the importance of courses, while factor analysis highlights important factors and quantifies all aspects reasonably.

3.2 Comparison between the improved decision tree method and other student achievement prediction methods

The effects of three performance prediction methods are compared through experiments, that is, using C5.0 algorithm directly, using C5.0 algorithm after discretization by traditional methods, and improving decision tree. The comparison results of the three methods are shown in Table 2

Table 2. Comparison table of decision tree model generated by three methods

way	Accuracy of estimation	Tree depth	Number of leaf nodes	cross validation
C5.0 algorithm	40.8%	8	13	Average value: 14.4 Standard deviation: 2.9
Interval marking method+ C5.0 algorithm	56.7%	6	11	Average value: 37.8 Standard devia-

				tion: 3.6
Improved decision tree method	64.8%	5	9	Average value: 45.7 Standard deviation: 3.1

Among them, the traditional data discretization method is interval marking method, that is, a fractional interval is designated to mark it as the corresponding grade (in the experiment, 85-100 points are marked as excellent, 70-84 points are marked as good, 60-69 points are marked as medium, and 0-59 points are marked as poor), thus replacing continuous numerical values.

It can be seen from Table 2 that the improved decision tree method has the highest estimation accuracy, and the generated decision tree has the smallest depth and the least number of leaf nodes, which indicates that the decision tree model generated by it is more accurate and robust than the other two methods. In addition, the standard deviation of the improved decision tree method in the cross-validation results is also small, which shows that its model is stable and the fluctuation of prediction is small. On the whole, because the improved decision tree method establishes reasonable boxing rules when discretizing data, it distinguishes the examination difficulty of the course itself, and its effect is better than the other two methods [10].

4 Conclusion

The quality supervision of teaching is an important part of teaching quality management in university, which is of great significance to the improvement of teaching quality. With the development of university informationization, a lot of useful information has been produced in the educational administration of universities, but it has not been improved and used effectively yet. In response to this, this article expounds the application of information mining technology in the quality supervision of teaching in universities. The fuzzy comprehensive evaluation method is used to evaluate comprehensively and objectively the scores of students in computer major, and an improved decision tree is proposed to predict the grade of students. finally, the approach in this article is compared with other methods. The results show that, compared with the average ranking method, the factor analysis method contains more information and can provide a more comprehensive evaluation. In addition, it highlights important factors and can reasonably quantify all aspects. Compared with other performance prediction methods, the improved decision tree method has better stability, higher accuracy and less over-fitting. It has been proved that this method can play a certain role in the monitoring of teaching quality in colleges and universities. However, due to the limited data available at present, the results may have some limitations, which will be further studied in the future.

Acknowledgement

Liaoning Social Science Planning Fund Project (Project No L21ASH004); Liaoning Economic and Social Development Project Initiation Project (Project No 2022lslybkt-039); The Ministry of Education's Industry School Cooperation Collaborative Education Project (Project No 220906353200035)

5 References

1. Zheng, C. , & Zhou, W. . (2021). Research on information construction and management of education management based on data mining. Journal of Physics: Conference Series, 1881(4), 042073 (6pp).
2. Du, Y. , & Zhao, T. . (2021). Network teaching technology based on big data mining and information fusion. Security and Communication Networks, 2021(9), 1-9.
3. Lu, L. , & Zhou, J. . (2021). Research on mining of applied mathematics educational resources based on edge computing and data stream classification. Mobile Information Systems, 2021(7), 1-8.
4. Elatia, S. , Ipperciel, D. , Zaiane, O. , Bakhshinategh, B. , & Thibaudeau, P. . (2021). Graduate attributes assessment program. The International Journal of Information and Learning Technology, 38(1), 117-134.
5. Wu, X. . (2022). Research on the reform of ideological and political teaching evaluation method of college english course based on "online and offline" teaching. Journal of Higher Education Research, 3(1), 87-90.
6. Coonan, C. M. . (2023). Review of bower, coyle, cross & chambers (2020): curriculum integrated language teaching: clil practice:. Journal of Immersion and Content-Based Language Education, 11(1), 141-144.
7. Yuan, X. . (2021). Design of college english teaching information platform based on artificial intelligence technology. Journal of Physics Conference Series, 1852(2), 022031.
8. Zhou, W. , & Yang, T. . (2021). Application analysis of data mining technology in ideological and political education management. Journal of Physics: Conference Series, 1915(4), 042040 (7pp).
9. Denisova, O. A. , Kunsbaeva, G. A. , & Chiglintsiva, A. S. . (2021). Big data: some ways to solve the problems of higher education. Journal of Physics: Conference Series, 2001(1), 012021 (6pp).
10. Wang, Q. . (2021). Application of clustering algorithm in ideological and political education in colleges and universities. Journal of Physics: Conference Series, 1852(3), 032041 (6pp).