Design of Early Warning Platform for College Students' Achievement Based on Data Mining

Yuanfu Mao1,*, Zhifeng Yu1
1Nanchang Institute of Science and Technology, Nanchang, 330108

*Corresponding author e-mail: 1722521178@ncist.edu.cn

Abstract. With the acceleration of the application of information technology in Colleges and universities in China, the efficiency of higher education is constantly improving. The focus of teaching management in Colleges and universities is to continuously improve the teaching level of colleges and universities, and the key is to strengthen the management of students' performance. Performance warning is a form of student performance management. In recent years, data mining technology is more and more mature, and its application is also very wide. Many students have applied data mining technology to university management. In this paper, we apply data mining technology to college students' performance early warning, and use Apriori algorithm in association rules to design and build college students' performance early warning platform, and select two classes of students as the research object to verify. In this study, we choose the English course scores of two classes as the test data, and define the performance warning, which is based on the score below 60. The results show that six students in class a will be subject to performance warning, while seven students in class B will be subject to performance warning. In addition, the performance early warning platform designed by this method, the early warning accuracy rate is as high as 92.85%, the accuracy rate is high, has certain application advantages.

Keywords: Data Mining, Performance Warning, Association Rules, Apriori Algorithm

1. Introduction
In recent years, the popularization of higher education has not only caused a sharp increase in the number of college students, but also caused a continuous decline in the quality of college students. It is an indisputable fact that a small number of students fail to pass many courses in the final examination or even fail to retake due to their maladjustment to college study and life [1-2]. How to conduct scientific, reasonable, timely and effective monitoring and supervision on students' academic performance has increasingly become a thorny and very important problem faced by colleges and universities, especially the teaching management department of colleges and universities [3]. In order to ensure the normal teaching order of the school, prevent the occurrence and continuous deterioration of the poor performance of a few students, and achieved certain results [4-5].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
PUBLISHED UNDER LICENCE BY IOP PUBLISHING LTD
We know that education big data is usually generated in the whole education activities, which can be said to be all the data collection collected according to the needs of education, which can be used for the development of education and create great potential value [6]. At present, data mining technology has developed more mature, and is widely used in the information management system. It mainly processes a large number of data, obtains useful information from the data, and discovers the internal rules and operation mode of data, so as to provide decision-making information for data users [7-8]. Applying data mining technology to college students' performance early warning, we can calculate the possibility of different performance levels through the early warning mechanism. In this way, we can use the initial performance data to infer the future performance trend, and then we can find out the association rules corresponding to this course as early warning factors to establish an early warning system. It is very convenient and timely to evaluate the students' academic performance, which has very important theoretical significance and application value [9-10].

This paper first introduces data mining, association rules and Apriori algorithm, and designs an early-warning platform for college students' achievement by using Apriori algorithm of association rules. In order to verify the effectiveness of this method, we choose two classes of computer major as the research object, and take their relatively weak English courses as test data. In the research, in order to better research, we defined the performance early warning, taking the score below 60 as the boundary to carry out the performance early warning. After verification, the results show that the performance early warning platform designed by this method, the early warning accuracy rate is as high as 92.85%, the accuracy rate is high, has certain application advantages.

2. Data mining and apriori algorithm of association rules

2.1. Data mining

Many subjects integrate and promote each other, which leads to data mining. The progress of data mining technology not only promotes the rapid development of data mining, but also makes data mining widely used in practice to solve the problems of all levels of society. The general definition of data mining is: the process of extracting hidden information that people do not know in advance but may be useful from a large number of noisy, incomplete, random and fuzzy practical application data. Data mining is a process of extracting previously unknown rules and knowledge from a large number of data, which may further support the enrichment of existing knowledge or decision-making. It mainly includes the following four steps.

(1) Determine mining target.
Data mining is based on the understanding of real business problems and business data. Only by understanding business, can we put forward problems according to business needs, make clear the goal of data mining, and carry out data mining blindly for the purpose of mining is meaningless and valuable.

(2) Data preparation and processing.
We know that data is the basis of data mining, so the data preparation and data cleaning stage is the key step of data mining, and also the key factor affecting the data results. It can be said that the preparation work in the first stage is very important, and the quality of the results will directly affect the accuracy and efficiency of data mining results. On the other hand, data preparation and processing are also the key links and important steps in the implementation process of data mining. The quality of the data sorted in this link is the prerequisite for data mining to get the correct results. The main links include data selection, preprocessing, conversion and so on.

1) Data selection. This step includes collecting the original data (including internal and external data) related to the studied business, selecting and integrating the data related to the studied business and suitable for data mining from the collected data, and establishing the data mining database.

2) Data preprocessing. According to the needs of data mining tasks, the data records in the database are sorted out, and the attributes not related to the mining tasks are cleared, and the incomplete data records are sorted out.
Data conversion. The main work of data conversion is to find the feature representation of data according to the task and goal of data mining, so as to meet the requirements of data format in data mining work. This is the key to the smooth implementation of data mining and the successful results.

(3) Data mining.
At this stage, it can be said that it is the work of mining the data. Before mining, we need to study and analyze the mining algorithm, and we need to select the appropriate mining algorithm according to the mining task and mining target, followed by the data mining algorithm. In this stage, the selection of algorithm is very important, which will directly affect the quality of the final mining results.

(4) Result analysis.
The purpose of data mining is to find useful data that is valuable to customers. These data can't be described by numbers, which is not intuitive and meaningful to customers. The last step of data mining is the expression and interpretation of data results. Generally, data can be described visually rather than symbolically, and rules can be used to explain customers' actual business. There will be many results of data mining. This process requires selecting the data that customers care about most to express, and removing the data that customers don't care about, so as to reduce the annoyance caused by too much data. If the customer is not satisfied with the final data result or has data missing, they need to repeat the above steps to carry out a new round of data mining on the data until the result is satisfactory.

2.2. Association rules and apriori algorithm
(1) Association rules.
Association rule analysis is one of the most active fields in data mining. The objective relationship between things is usually called connection. Association rules are used to describe the relationship between the data elements hidden in the database and analyze its rules. The analysis of association rules usually refers to the identification and analysis of the data that can meet the requirements in the transaction database D, the so-called meeting the requirements usually refers to meeting the set minimum support threshold (min)\_Support and minimum confidence threshold (min)\_Confidence of the association rules.

Suppose the transaction database is called transaction set D, and the elements of T are called items. Let the set of all items in d be I. obviously, transaction T belongs to I.

Let the itemset of D be A, and the support of itemset A is the percentage of all transactions of D that contain a in D.

\[
\text{Support}(A) = \frac{|\{T:A \subseteq T, T \in D\}|}{|D|} = P(A)
\]  

Association rules, in short, are implicative expressions in the form of $A \Rightarrow B$, in which the intersection of the two is an empty set. A is the condition of association rules, B is the conclusion of association rules.

Support describes the representativeness of rules in all cases. It is A measure of importance, while confidence is A measure of accuracy. Therefore, the higher the support, the more important the rules are:

\[
\text{Support}(A \Rightarrow B) = \frac{|\{T:A \cup B \subseteq T, T \in D\}|}{|D|} = \text{upsort}(A \cup B) = P(A \cup B)
\]  

The credibility of association rules includes both A and B transactions. The percentage of itemset A in the transaction is as follows:

\[
\text{Confidence}(A \Rightarrow B) = \frac{|\{T:A \cup B \subseteq T, T \in D\}|}{|\{T:A \subseteq T, T \in D\}|} = \frac{\text{Support}(A \Rightarrow B)}{\text{Support}(A)} = P(A|B)
\]  

(2) Apriori algorithm.
Apriori algorithm is the most influential classical algorithm for mining frequent itemsets of association rules. It uses the iterative method of layer by layer search to find all frequent itemsets.
The basic idea of Apriori algorithm is as follows:
1) The first step is to find all frequent itemsets.
2) In the second step, the frequent itemsets are used to generate strong association rules with minimum support and confidence.
   In this process, the first step is the key link.
   In short, the essence of Apriori algorithm is: if an item set is frequent, then all its non empty subsets should be frequent. In other words, if an itemset is not frequent, then all its supersets cannot be frequent. In other words, if element set B does not meet the minimum support, then B is not frequent. If a set is added to B, the result item set (BA) cannot appear more frequently than B. Therefore, BA is not frequent.

\[ P(B \cup A) < \text{Min}_\text{support} \] (4)

Apriori algorithm scans transaction database d many times to find frequent itemsets, and at the same time, it processes candidate itemsets by many times of join operation, which is the key to restrict the efficiency of the algorithm.

3. The design of experimental object and score division
(1) Subjects.
   This paper chooses two classes of computer major as the research object, which are class A and class B. There are 32 students in class A and 33 students in class B. In the process of investigation, we find that the College English public course of computer majors is relatively weak, and some students often fail to take it again. This paper will take the English course of these two classes as an example to warn the students' grades.
   (2) The design of students' grade division.
   In the selected student achievement, the student achievement is uneven, in order to more clearly express the student's achievement, we divide the achievement into four categories. Students' grades are usually based on the hundred point system. Then, we divide the grades into: if a student's grades are lower than 60, then the students' grades are divided into grade 4; if a student's grades are greater than or equal to 60 but less than 80, then the students' grades are divided into grade 3; if a student's grades are greater than or equal to 80 but less than 90 if a student's score is greater than or equal to 90, then the student's score is divided into Level 2; if a student's score is greater than or equal to 90, then the student's score is divided into level 1. If the student's score is in grade 4, then it is necessary to carry out performance warning.

4. Analysis of the early warning platform of students' performance

4.1. Student performance analysis
The results are shown in Table 1 and Figure 1.

| Grades | Class A | Class B |
|--------|---------|---------|
| Level 1| 4       | 6       |
| Level 2| 8       | 7       |
| Level 3| 14      | 13      |
| Level 4| 6       | 7       |
Figure 1. Student achievement analysis

It can be seen from Table 1 and Figure 1 that the number of class A in Grade 1 is 4, and the number of class B in Grade 1 is 6. The number of class A in Level 2 is 8, and the number of class B in Level 2 is 7. The number of class A in Level 3 is 14, and the number of class B in Level 3 is 13. The number of class A in grade 4 is 6, and the number of class B in grade 4 is 7. That is to say, in class A, according to the rules, six students in class A will be subject to performance warning, while seven students in class B will be subject to performance warning.

4.2. Analysis of early warning results of student achievement

This paper uses the method of the above two classes of students' English course performance early warning, and compares it with the actual number of people who need early warning. The results are shown in Table 2 and Figure 2.

Table 2. Results of performance warning

| Class                        | Class A |
|------------------------------|---------|
| In this paper, the number of early warning results | 6       |
| Actual number of people in need of early warning | 6       |
| Early warning accuracy       | 100%    |

| Class                        | Class B |
|------------------------------|---------|
| Actual number in need of early warning | 7       |
| Early warning accuracy       | 85.7%   |

Figure 2. Result analysis of performance warning
It can be seen from Table 2 and Figure 2 that in this method, the number of grade early warning in class A is 6, and the number of grade early warning in class B is 6, while the number of grade early warning in class A is 6, and the number of grade early warning in class B is 7. The accuracy of this method is 100% in class A and 85.7% in class B. In general, the accuracy of this method is as high as 92.85%.

5. Conclusion
With the development of science and technology, data mining technology is more and more mature, and its application is more and more widely. At present, in order to better manage the students, many colleges and universities have set up a certain performance warning mechanism. Through the performance warning mechanism, they manage the students with poor performance and urge them to study better. On the basis of data mining, this paper uses Apriori algorithm in association rules to design and build the early warning platform of College Students' performance, and selects two classes of students as the research object to verify. The results of this paper show that the accuracy of the performance early warning platform designed by this method is as high as 92.85%, which has certain application advantages. The results of this paper will provide some reference value for the research of College Students' performance early warning.

Acknowledgments
Science and Technology General Project of Jiangxi Provincial Department of Education (No. GJJ191095) & (No. GJJ191101).

References
[1] Wong W K, Choy C W. Progress in Hong Kong's Tropical Cyclone Forecasting and Warning Services in Recent Decades [J]. Tropical Cyclone Research and Review, 2018, v.7(01):41-54.
[2] DD Heckathorn, Broadhead R S. Rational Choice, Public Policy And AIDS [J]. Rationality & Society, 2017, 8(3):235-260.
[3] Li X L, Ma L, He X D, et al. You Are How You Behave – Spatiotemporal Representation Learning for College Student Academic Achievement [J]. Journal of Computer Science and Technology, 2020, 35(2):353-367.
[4] Wang L, Shi Z, Qu J. Compose a new chapter of cardiopulmonary resuscitation training with Chinese characteristics [J]. Zhonghua Wei Zhong Bing Ji Jiou Yi Xue, 2018, 30(12):1117-1118.
[5] Qi T Y, Lei B, Wang R, et al. Solid-fluid-gas coupling prediction of harmful gas eruption in shield tunneling [J]. Tunnelling & Underground Space Technology, 2018, 71(jan.):126-137.
[6] Buczak A, Guven E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection [J]. IEEE Communications Surveys & Tutorials, 2017, 18(2):1153-1176.
[7] Chaurasia V, Pal S. A Novel Approach for Breast Cancer Detection using Data Mining Techniques [J]. Social Science Electronic Publishing, 2017, 3297(1):2320-9801.
[8] Yan X S, Zheng L. Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach [J]. Review of Financial Studies, 2017, 30(4):1382-1423.
[9] Fujimoto M, Kanou M, Hosomi K, et al. Angiotensin receptor blockers and the risk of cancer: data mining of a spontaneous reporting database and a claims database. [J]. International Journal of Clinical Pharmacology & Therapeutics, 2017, 55(4):295.
[10] CF Chien, Huang Y C, Hu C H. A hybrid approach of data mining and genetic algorithms for rehabilitation scheduling [J]. International Journal of Manufacturing Technology & Management, 2017, 16(1):76-100.