Research on Students' Performance Portrait Method Based on Multidimensional Data

ChangLin Ren, XueLong Wang
Department of Information Center, Xi'an Shiyou University, Xi'an 710065 China, rcl123@163.com

Abstract. With the construction and operation of the university's intelligent application system, it also accumulates a large amount of data for colleges and universities. Many of these data are student performance data. It is of certain significance to analyze its performance characteristics through data mining technology, to realize teaching in accordance with the aptitude, intelligent management, improve the quality of education and teaching, and work efficiency. We collect data such as wireless user trajectory data, card data, and book borrowing data to analyze student performance. Construct a classification model of student performance based on cluster analysis, and analyze the performance indicators to find the rules between performance and achievement, and improve the learning efficiency of students.

1. Introduction
With the continuous development of educational informatization and the in-depth application of Internet technology in the education industry, under the drive of “Internet +” education, the construction of smart campuses in colleges and universities is constantly improving. The construction and operation of intelligent application systems has also accumulated a large amount of data for colleges and universities. Many of these data are student performance data. How to effectively use student performance data, analyze its performance characteristics through data mining technology, realize the teaching of students in accordance with their aptitude, intelligent management, improve the quality of education and work efficiency, has important practical significance. As a new type of data analysis technology, “Big Data” technology is an indispensable part of the “smart campus” construction. It can deeply mine and analyze different types of data. Through big data analysis, it can mine user life learning habits and performance rules, predict user performance, mine hotspots according to user group performance, and cluster and analyze similar users. It not only provides strong decision support for university management, but also provides personalized and intelligent recommendation services for teachers and students, such as forecasting services, service push, precision poverty alleviation, performance warning, academic early warning, employment analysis, and subject planning. This paper intends to use the various campus data information of college students, according to the students' basic attributes, learning performances, consumption performances, online performances, mobile movement trajectories and other information, label each student, and objectively and comprehensively describe the student performance, abstracting out A tagged student model that builds a student portrait model based on multidimensional student data.

2. Data collection
Data collection seems easy and unobtrusive, but it is the basis and key to the study of campus wireless user group performance. Valuable information is mined from massive data, so to mine value, you must
collect a certain amount of data. Collect wireless user data as the basis for the analysis of campus wireless group performance. At present, for the collection of campus WLAN data, there are two methods: active acquisition and passive acquisition. The raw data collected by any method is generally unstructured data, and there are usually problems such as noise, missing values, error values and inconsistencies. It cannot be directly used as a data source for data analysis, and necessary data is needed. Preprocessing, the main processes of data preprocessing include data cleaning, data integration, data transformation and data specification. There are two methods of data acquisition: active acquisition and passive acquisition. Considering the cost and efficiency, we adopt passive acquisition. The passive acquisition methods can be summarized into four types: SYSLOG-based acquisition, SNMP-based acquisition, and terminal-based simulation. The collection of telnet, based on the collection of network sniffing tools. The passive acquisition method does not require any acquisition software for the wireless user terminal, nor does it require the active participation of the wireless user. It can be said that it has good scalability.

The student basic information form contains a large number of fields, including all basic information before and after the student enters the school. Only the names, student numbers, genders, departments, majors, and classes are required in the student performance information. Student achievement scores. There are many courses for students in the school. The grade point average is the comprehensive score of the students in one semester or one school year calculated by the GPA rules which Collected from the educational system database.

The Card Consumption Schedule (T_CARD_PAY) contains all the credit card records of undergraduate students, graduate students and school teachers. These include: ID, consumer account, spending amount, spending time, consumption port, transaction type ID, transaction type, balance, etc. The consumption time is recorded once every time the card is used; the transaction type is the place of consumption (for example, restaurant, supermarket, water room, bath center, etc.), and the data is collected from the card system.

The library borrowing table mainly records the students' borrowing of books in the library. Including the borrowed book, the borrower's logo, the time of borrowing, the time of return, the time of returning, the date of renewal, the number of renewals, and the time of overdue. Among them, the basic information of the students is identified by the borrower's logo.

The library access control table mainly records the data of students when they enter and leave the library. Through this form, students can observe the situation of students entering and leaving the library to read books and study. These include: ID, credit card identification, credit card time, school year, semester field.

3. Data preprocessing

3.1. Vacancy value processing
Due to the relatively early construction of various systems in the school, the update of the system in recent years has led to confusion in its data structure, resulting in data collected from students who have dropped out of school or left school, as well as uncertainty due to lack of exams, suspension of school or cheating in exams. The data contributed by the factors are missing. In order to ensure the correctness and accuracy of the data, the author scores 0 points for the students who are absent from the exam and cheated in the exam. For students who have dropped out of school or are absent from school, the information is removed from the sample data. For book borrowing and access control data, all students who do not have the data will be recorded as 0 because some students who do not go to the library and have no borrowing records will have some data missing.

3.2. Statistical Analysis
The school's various business systems record all the performance information of the students at school. The huge amount of data is a big problem for the mining of knowledge. The author will transform the data of the mining target, and use the methods of statistics, clustering and classification to compress,
generalize and normalize the data. For the data of the campus card, this article first uses the statistical method to calculate the total amount of students' consumption for one semester in the semester period. The total number of consumption and total amount of one semester are counted in the morning, middle and dinner time periods; the total amount of supermarket consumption and the total number of times are counted in a semester, book borrowing is also a total number of borrowed books for a semester; the total number of entrances and exits to the library, as well as the total number of water and bathing, this process is also a cumbersome and complicated data compression, from the overall observation of the performance dynamic characteristics of a semester, it is also the process of data generalization. The student achievement data of the educational administration system only extracts the comprehensive grade points of the students per semester, which represents the overall level of the students' grades in this semester, and calculates the average grade points of different faculties and different majors per semester to evaluate the professional level of academic achievement between departments and departments.

3.3. Data discretization
After the data is processed as described above, the target data required by the research is basically formed, all the data are combined, and the data is further discretized. The fields included in the combined comprehensive data are digitally identified, and the tagged data is quantized to prepare for the subsequent correlation analysis.

4. Student performance portrait method based on cluster analysis
Reasonable classification and subdivision of the performance data of students during school, personalized management and service for different types of students, improve the accuracy of school management, and provide reasonable decision-making basis for management is the campus education One of the goals in data analysis. Classifying students according to their performance characteristics is the basis for managers to effectively and individually manage different types of students. Most of the traditional methods of classifying students are based on the students' grades, the student status files, and the teacher's evaluation of the students' class time. This classification method lacks dynamic and comprehensive, and is now a digital campus. In the era of big data, all the performances of students in school can be displayed through data. The existing big data analysis technology should be combined to comprehensively and reasonably classify students. According to the data in each system of digital campus, the construction can respond. Students have a valuable relevant indicator system for all aspects of performance, and then use the cluster analysis method in data mining technology to subdivide the students and get the relevant performance characteristics and distribution of different types of students. We mainly focus on the dynamic data in the student data. Analyze and mine, and divide the dynamic performance data into three evaluation index systems: consumption law, life law and learning effort level to describe and analyze the performance of students in school. According to the three evaluation indicators, the K-means algorithm is used. Data clustering results, the characteristics of different performances The categories are summarized, classified, labeled, and the labels of their indicators are combined with some of the basic information of the students. The student performances are portrayed, and the performance characteristics of different types of students are derived. Provide management decision-making basis, master the current performance status and overall trend of the whole school, and facilitate personalized management and effective early warning work for students.

5. Student performance portrait method and application of association rule analysis
Association rule mining is a process of analyzing and summarizing different items or different attribute values in a large number of data objects. It is useful to find out the data that frequently appears or the same attribute value. Interesting rules, which in turn can be described in the form of expressions. Through mining analysis, you can find that these useful rules can help relevant departments to provide better decision-making basis. Association rule mining is also called frequent
item set mining. One of the most classic examples is shopping basket analysis. This process is to analyze the customer's shopping habits by discovering the relationship between the products placed in the shopping basket, thus helping the business management. People develop better marketing strategies. Apriori algorithm is a classic association rule mining algorithm. Finding all frequent itemsets in the mining process is the focus of related data mining. The performance of the algorithm determines the performance of frequent item set mining. In the process of deep mining and use, the algorithm is found to have the following disadvantages: the time and space overhead of the algorithm is huge, and the efficiency of the algorithm is low.

5.1. Improved association rule mining algorithm
When the Apriori algorithm scans the transaction database, when the infrequent item set count of an item reaches, the transaction scan can be interrupted, the item set is set as an infrequent item set, and the item set is directly deleted, and the next item is entered. The scan of the item set, the scan of the data set is from the top to the bottom and from the bottom to the top middle of the data set.

Input: transaction database D, minimum support threshold min_sup

Output: Frequent itemsets in L

Method: (1) Infrequent_support=total number of transactions-min_sup+1;
(2) For(k=1;L_k!= Φ;k++) do begin ;
(3) C_{k+1}= generates a candidate set from L_k;
(4) Prune (C_{k+1}); (5) Scan transactions are performed from top to bottom and bottom to top, increasing the support count and infrequent item set support count for all candidate sets in CK+1.
(6) If support_count= min_sup or Infrequent_count= Infrequent_support Stop scanning;
(7) When support_count= min_sup ; the set is a frequent item set, retaining the set. Infrequent_count= Infrequent_support ; The item set is an infrequent item set, and the item set is deleted.
(8) (L_{k+1})=species in (C_{k+1});
(9) End; (10) Return (L_k); (11) End procedure;

5.2. Improved Apriori algorithm comparison test
Select the same data set and set the minimum support threshold to a different value to test the time required to generate frequent itemsets under the improved support and improved Apriori algorithm. The results of the comparison test are shown in Figure 5-1.
5-2 The total number of transactions scanned during each frequent item

Through the above experiments, the improved Apriori algorithm reduces the number of scans of the transaction database compared with the original Apriori algorithm, improves the efficiency of the algorithm, and reduces the execution time of the CPU.

5.3. Application of Improved Apriori Algorithm in Student Performance Analysis

Study the relationship between student achievement and student performance, combine performance data and performance data, and discretize the data obtained. In the comprehensive data table, there are many attributes. In order to convert these relational databases into transaction databases, the data needs to be discretized and the attributes are encoded and identified. After the above correlations are obtained between the acquisition of the association analysis rules and the results, the improved Apriori algorithm is used to screen out the more meaningful association rules. From the results, it can be seen that the academic performance has a certain relationship with the students' school performance. Based on the above research, the university can help the university to check the overall learning and living conditions of the students, and timely predict and warn the students' abnormal conditions for the school. Decision making provides data support. According to the level of students' efforts, we will carry out personalized guidance in a timely manner to improve students' learning level. According to the actual consumption situation of the students in the school, find out the students with invisible difficulties and improve the humanistic care of the school.

6. Conclusion

Based on the clustering analysis, the student performance classification method is studied. The performance data is divided into indicators, and the aggregation method is used to classify and construct the student portrait. Help our school to understand the performance characteristics and habits of different types of students, develop a personalized management model; analyze the development trend of students' performances, establish an early warning mechanism for students with abnormal performances, and respond and judge in time. The Apriori algorithm is improved by comparing the minimum support threshold of the frequent itemsets or the minimum support threshold of the infrequent itemsets, and stopping the scanning of the data transaction library to reduce the number of scans of the transaction database, thereby improving the running efficiency of the algorithm reduces the execution time of the CPU, and analyzes the performance and performance of the students. It is found that the students' daily consumption, life rules and learning effort are also factors that affect the performance. According to the category analysis of students' performance characteristics, The student behavioral portrait method based on multi-dimensional data provides scientific and personalized guidance and help for students of different behavior types, improves bad behavior habits, improves learning efficiency, and improves students' achievement.

References

[1] HaoWei, Xingyuan, Chen ChaoWang, User performance analyses based on network data stream scenario[J]. Communication Technology(ICCT), IEEE. 2012.
[2] YAO Wendi. Research of Data Mining by Association Rules and Its Application to the Analysis of Academic Achievements [D]. Chengdu: Southwest Jiaotong University, 2015.

[3] WU Xiping. The Analysis of College Student Achievement based on Association Rules Mining Technology [D]. Chengdu: Southwest Jiaotong University, 2010.

[4] WANG Xuefei. Application of College Students with Financial Difficulties on Campus Card Flow Data of Data Mining [D]. Changchun: Northeast Normal University, 2014.

[5] ZHANG Dongdong, LI Yulong, WAN GYuxin. The Application of Data Mining Technology in Identifying Needy College Students [J]. Journal of Xi'an University of Arts & Science (NatSciEd), 2013, 16 (4): 90-94.