On the Application of Data Mining Algorithm in College Student Management

Xiaofei Sun[1(✉)] and Yunan Zeng[2]

[1] Nanjing University of Finance & Economics, Nanjing 210023, China
[2] Nanjing Niuding Technology Co., Ltd, Nanjing 210023, China
9220080008@nufe.edu.cn

Abstract. With the rapid development of computer technology and Internet technology, information technology has been widely used in all walks of life. Big data, Internet of things, artificial intelligence and intelligent manufacturing have become the pronouns of the times. This paper mainly expounds the use of data mining for scientific and convenient management in college student management, which is more conducive to the healthy growth of students.

Keywords: Data mining · College student management · Application analysis

1 Introduction

With the development of information technology, information technology has shown its advantages in the management of colleges and universities. Under the background of information technology, the management of colleges and universities will be more standardized and scientific, Therefore, education authorities at all levels have regarded information technology as an important manifestation of the level of running a university. However, a new management model will always face many opportunities and challenges instead of the old one. The main problems are as follows: some older managers often have rich management experience and always think that they have experience [1]. The new management model is not easy to use, They will find all sorts of reasons to stop new management models. At present, most of the students in Colleges and universities are post-95 students. They are active in thinking, do not accept too restrictive management, and have a strong ability to accept new things and new technologies. In particular, they are good at computer and mobile phone operation, and the information management mode is relatively easy for them to accept. Although the school attaches great importance to information construction, it needs to invest a lot of human, material and financial resources, which affects the speed of information process.

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2021, LNICST 392, pp. 248–257, 2021.
https://doi.org/10.1007/978-3-030-87903-7_32

2 The General Process of Data Mining

2.1 The Role of Data Mining in University Management

The subjective factors in the management of students are still too heavy and lack of scientific analysis and basis. On the one hand, the managers manage the students according to the provisions of various documents, but the documents only stipulate various conditions, procedures or quota limits. If the students who meet the conditions exceed the quota limit, How to select the truly excellent students objectively and reasonably depends on the subjective judgment of the managers because there is no specific operation method in the management documents.

2.2 The General Process of Data Mining in University Management

Data cleaning and integration eliminate noise or inconsistent data, combine multiple data sources together, select and transform data, retrieve data related to analysis from database, and transform data into a suitable form for mining. For example, data mining is the basic step by summarizing or aggregating operations, and data patterns are extracted by intelligent methods. According to a certain interest measure, pattern evaluation can identify the really interesting patterns that represent knowledge [2]. Knowledge representation uses visualization and knowledge representation technology to provide users with mining knowledge. The general process of data mining can be shown in Fig. 1.

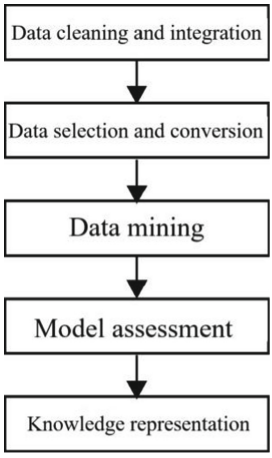

Fig. 1. Basic process of data mining

3 Fuzzy Mining Algorithm

In the process of data mining, data mining algorithm is the most important. By using fuzzy mining algorithm and data extracted from data warehouse, we can find the individual types existing in the current organization. In addition, we can judge which of these types an individual belongs to.

3.1 Pattern Discovery

The sample set to be classified is established on all data records of data warehouse. The object to be classified is called sample, such as sample set. The specific attributes should be quantified. The quantified attributes are called sample indicators, and there are m indicators. This is the case:

$$u_i = \{u_{i1}, u_{i2}, \cdots u_{im}\}\ i = 1, 2, \cdots, m \tag{1}$$

Therefore, these original data should be standardized, and the standardized value of each data can be calculated according to the following formula.

$$u_{ik} = \frac{w_{ik} + u_{ik}}{S_k} \tag{2}$$

Then the standard deviation of the original data is calculated according to the following formula:

$$S_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_{ik} - u_k)^2} \tag{3}$$

3.2 Cluster Algorithm Analysis and Simulation

This paper uses the maximum tree method for clustering analysis, that is to construct a graph with all the classified objects as vertices. When $\neq 0$, the vertices can connect an edge. The method is to draw one of the vertices first, and then connect the edges in the order of row from large to small. It is required that there is no loop until all the vertices are connected. In this way, a maximum tree can be obtained. Each edge of the tree can be weighted by a certain number, but due to the different connection methods, the maximum tree can not be unique, and then the maximum number is taken into the cut set, that is, those weights are removed. In this way, a tree is cut into several disconnected subtrees [3]. Although the largest tree is not unique, the subtree is the same after taking the cut set. These subtrees are the patterns of inductive discovery in data warehouse. We use the cluster algorithm to simulate the manager system which is shown in Fig. 2. We can get thus results, with the increase time, the synchronous rate is increase. This also shows that our management system is safe and stable.

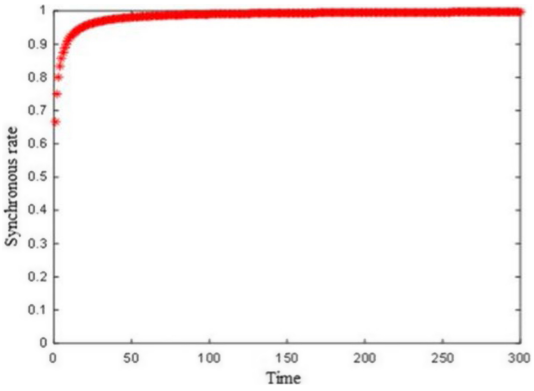

Fig. 2. Cluster algorithm and simulation for the manager system

3.3 Forecast

The first mock exam is used to get the average index of each mode and the average index is obtained by pressing the formula:

$$M_{ij} = \sum u_{kj}/P(i = 1, 2, \cdots, s, j = 1, 2, \cdots, m) \tag{4}$$

The total number of patterns. K is the total number of records that the pattern (i.e. the second pattern) is pushed out by in the warehouse. The first mock exam sample X and the first mock exam set the sample close to which model, and predict the whole situation from the overall situation of the model.

4 The Inevitability of Applying Big Data Technology to Student Management

In all colleges and universities, teaching is the center, and student management is the basis to ensure the work of teaching center. If the student management can not keep up with the teaching work, it can not be guaranteed, so the student management is related to the healthy and safe growth of students. The traditional management of students is often based on the experience of the management staff. The top-down meeting is used to manage the students. This way is often a long cycle and inefficient. As time goes on, it is not suitable for the management of the post-95 and post-2000 students. In the Internet age, contemporary college students acquire knowledge faster, and the way to acquire knowledge is also more. The way in the past is certainly not good. Through big data technology, students' life and learning situation in school are recorded, and each student's life and learning habits are analyzed through data mining, so as to better understand students' learning situation, mental health status and living habits, and achieve more accurate help, so that each student can live and study healthily and happily in school. When we add the noise to the system, we find the ARI has some changes which is shown in Fig. 3.

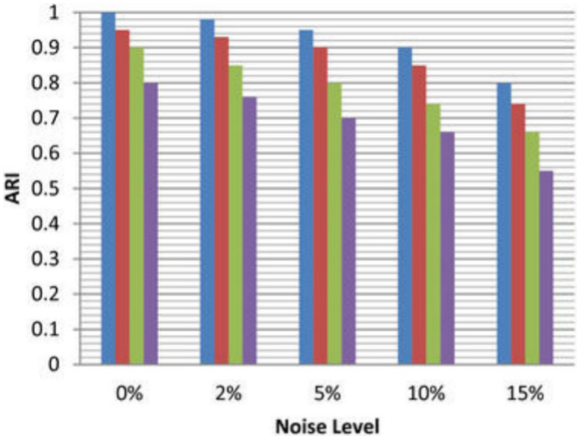

Fig. 3. Adding noise for the manage system

5 Application of Big Data Technology in Student Management

Colleges and universities are places where elites are concentrated, and all kinds of intellectuals are gathered. There are many kinds of data collected by the data center of colleges and universities, including not only relational data, but also non relational and semi relational data. These data must be stored in the data warehouse, The data platform is established to record the activity track of each student from getting up in the morning, brushing teeth, washing face, eating breakfast, reading early, doing morning exercises until the light off in the evening. The data stored in the data warehouse for a period of time is analyzed and modeled, Cleaning, integration, standardization, mining, finally mining out the students' learning situation every day, predicting when the students get up, when they go to the classroom, when they have three meals, when they turn off the lights to have a rest and so on [4]. No link is accurately recorded [5]. These information are used to classify the students, which students have good learning initiative and which students have good learning habits, Which students are fond of playing, which students are serious in class, which students have special skills, and so on. Through this information, we can help and guide students with unhealthy psychology through data mining, so as to make students healthy, Happy learning in school, the application of data mining technology in student management can accurately predict each student's learning situation, and the students' life performance from getting up to resting [6].

6 Establishment and Management of Data Warehouse

6.1 Data Extraction, Transformation and Loading

Data cleaning, transformation and loading is a very important stage in the process of data warehouse generation. In this paper, we extract the information of campus activities and academic achievements of grade 2005 students from the card information system,

library information system and educational administration system of Beijing Jiaotong University [7].

The processing of missing values and irregular values of data, such as the basic information of students, is imported by various early systems. Because of the non-standard among various systems, null values and irregular values appear in this section. Therefore, only the filtering function can be used to filter out null values and irregular values for manual correction. Data impurities and inconsistent data should be treated differently according to the situation, and can not be deleted. For example, the average score of the course should be queried for each student of the class or grade or major. If the student's score in the course is −0.1 or −0.2, it will be considered as disqualification or not attending the course, and the course will be calculated as zero; if all the courses of the student are empty, If it is determined that the student's studies have changed and cannot be compared with other students as valid data, all data of the student will be deleted. Repeat the results, the same person has more than one score in the same course, if there are make-up examination and re examination results, take the results of the first examination [8].

Because the object of our analysis is the students' scores of the courses in each college, and the courses with the course type of public courses are the courses with the unified proposition and examination of the whole school, the scoring standards of these public courses are consistent. If the course type is the achievement of professional basic courses and professional courses, and there are differences in the scores of various majors in each college, if we want to make a comparative analysis of the students' scores of the whole school, We must first normalize some grades, such as introducing Z value in statistics to compare students' course grades, and then increase the grade of field grades. The definition of Z value is to surpass different grades, colleges and inconsistent scoring standards, and replace students' course performance with a standard value, so that students' course performance can be judged by the same standard. Because this paper analyzes the course scores of the students of the same grade in the same college, there is no need for Z-value conversion, that is, the student scores in the original educational administration system are available [9].

6.2 Building a Cube

After loading the business data to be analyzed into the data warehouse, it lays the foundation for meeting the needs of students' management and decision-making. The future operations are based on the data warehouse with business data. However, the multidimensional analysis of data is not mainly for data warehouse, but for the subset extracted from data warehouse, such as data mart and multidimensional data set (also known as data cube). Before analyzing the data, you need to create a cube based on the analysis topic. There are two ways to organize multidimensional data sets: star and snowflake. In this paper, the organization of multidimensional data sets is mainly star. In the star model, business data are concentrated in the fact table, so as long as you scan the fact table, you can query, and there is no need to associate multiple huge tables. At the same time, complex queries can be completed through the comparison of various dimensions, drill up and drill down operations. So using star structure can improve the performance of query [10].

7 Application System Analysis and Design

7.1 Design Goal

The system can analyze and process a large number of data in the campus information system, find the information that needs attention, and realize the conversion of original data to valuable knowledge. The number of records in campus information system is usually large and complex [11]. Therefore, it is a key problem to extract, filter, transform and integrate a large number of data in order to discover knowledge. At the same time, to enable users to participate in the process of analysis and mining, the system should have a good interactive function and friendly interface.

7.2 The Architecture of the System

The system architecture mainly includes the following five layers: data presentation layer: providing human-computer interaction interface, accepting the corresponding requests of users, and returning the corresponding query results to users [12].

Data analysis layer: process the corresponding data request of the upper layer. Multi-dimensional analysis module provides multidimensional data management environment and realizes OLAP function. Data mining module is used to discover the relationship between data and make model-based prediction. Data storage layer: realize data management, accept relevant data submitted by data processing layer, and save them to relevant data warehouse and multidimensional data set according to the logic set by the system. Other models support data mining. Data processing layer: it can extract, transform, clean and load data from different data sources. Data source: the data source of the system, mainly including the record data of students' dining, access control and Internet access in the all-in-one card system; the basic information of students in the educational administration system, course selection results and students' Graduation destination information; the borrowing record data of students in the library system.

7.3 Evaluation of Data Mining Results

A data mining system often obtains thousands of patterns or rules after running one (Group) mining algorithm. Association rule mining is a typical example. The execution result of association rule algorithm can get thousands of association rules even if it processes a small record set, such as tens of thousands of records. However, only a small part of these thousands of rules have practical application value. So how to evaluate the results of data mining effectively in order to get a practical model is very important. Usually, we can judge and screen from the subjective level of users and the objective level of the system. The subjective level mainly includes the following four criteria: easy for users to understand; potential value; novelty; and being able to determine the effectiveness of new data or test data. The objective criteria are mainly based on the structure and statistical characteristics of the mined patterns. For example, an objective evaluation criterion for association rules is confidence (probability), which indicates how trustworthy the rule is, that is, confidence (a = > b) = P (BA). It is usually necessary to combine the two to find the really valuable and interesting knowledge. The establishment

of data mining model generally includes two stages: the first stage is to establish the basic mining structure, in which the mining structure includes the mining algorithm and the input and output attribute set; the second stage is to optimize the parameters of the mining algorithm in the mining structure. The two phases use different data sets. Firstly, the training data set is used to try a variety of mining models, and then the test data set is used to evaluate these mining models to find the best mining model [13].

At present, the vendors of BI solutions include IBM, Oracle, Sybase, Microsoft and other database products. In addition, they also include Cognos, business objects and Brio, SAS, a company with data mining and advanced statistical tools, and NCR, a manufacturer focusing on large-scale data warehouse. Here is a brief introduction to the BI solutions of IBM, Oracle and Microsoft IBM:IBM The company provides a set of business intelligence (BI) solutions based on visual data warehouse, including isual warehouse (VW), Essbase/DB2 OLAP server, IBM db2udb, as well as front-end data presentation tools (such as Bo) and data mining tools (such as SAS) from third parties, VW is a powerful integrated environment, which can be used not only for data warehouse modeling and metadata management, but also for data extraction, transformation, loading and scheduling. Essbase/DB2 OLAP server supports dimension definition and data loading. Essbas/DB2 OLAP server is a hybrid of ROLAP and MOLAP. Strictly speaking, IBM itself does not provide a complete data warehouse solution, the company adopts a partner strategy [14]. For example, its front-end data presentation tools can be Bo of business objects, Roach of lotus, impromptu of Cognos or query management facility of IBM; Essbase supporting arborsoftware and DB2 OLAP server of IBM (jointly developed with arbor) in multidimensional analysis; SAS System in statistical analysis. Oracle: the architecture of oraclebi solution is divided into three layers: data acquisition layer. All ETL processes can be stored in Oracle10g database by the ETL script generated by Oracle Warehouse builder, which is a tool provided by Oracle data warehouse. According to the requirements of the data warehouse system, the data can be extracted and loaded into the data warehouse system regularly, Oracle 10g database realizes the centralized storage and management of various types of data in data warehouse system, and supports the storage of massive data by using partition technology [15]. In the data presentation layer, Oracle provides a new business intelligence solution, Oracle Bi EE, OLAP analysis and development tools (JDeveloper+bi beans) and data mining tools (Oracle data miner), which show the results of statistical analysis in various ways.

8 Epilogue

This paper focuses on the research and implementation of university student management decision support system based on data mining. In this paper, firstly, the necessity and feasibility of applying data mining technology to students' management decision-making are expounded. Then, it introduces the concepts of data mining, data warehouse, online processing analysis and their relationship, as well as the related mining algorithm and its key parameters applied in this paper. Then, the architecture of decision support system based on data mining, the source of data, the data preprocessing process of data extraction, transformation and loading (ETL), and the construction, processing and access of multidimensional cube and mining structure are described.

Finally, the analysis results of multidimensional cube and data mining model are introduced in detail, and the practical significance of these results in college student management decision-making is analyzed. In this paper, data warehouse, online analytical processing and data mining technology are applied to the decision support system of student management, which has a certain reference significance for the construction and development of teaching management decision support system. The research of this paper is based on the data of students in the computer college, and the conclusions and rules are universal and special compared with the whole school. Among them, students' dining, students' borrowing and the correlation of students' scores belong to the common characteristics of various colleges; students' Internet access, students' courses, students' clustering and the correlation of students' Graduation destination have their special characteristics.

Due to the limitation of time and ability, there are still some deficiencies in this study, which need to be further studied and improved, mainly in the following aspects: 1. This study only takes the students of a certain term in the computer college as the research object, so it is necessary to further expand the scope of students, such as the horizontal expansion in the scope of colleges, and the vertical expansion in the scope of students' grades. 2. The selection of related attributes. In the attribute selection of data collection, there are many other factors that do not affect students' performance, such as the analysis of students' professional interests, the analysis of students' learning objectives, and so on. The corresponding questionnaire is needed to count the information in this aspect. Therefore, the data information used for data mining may not be the best data set. 3. The mining algorithm is improved. This study uses the mining algorithm of the mining tool, and does not improve the corresponding algorithm according to the actual mining needs.

References

1. Chen, W.: Data Warehouse and Data Mining Tutorial. Beijing Tsinghua University Press (2006)
2. Yu, L., et al.: Principle and Practice of Data Warehouse. Beijing People's Posts and Telecommunications Press (2003)
3. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kanfmann Publishing (2000)
4. Wei, Z.: SQL Server Development Guide OLAP Beijing Electronic Industry Press 2001 Shen Zhaoyang SQL sever20ap solution. Beijing Tsinghua University Press (2001)
5. Xu, H.: Data Warehouse and Decision Support System. Beijing Science Press (2005)
6. Jian, P., Wei, C., Chang, Z.: J. Data Cube Algebra Softw. OLAP **10**(6), 561–569 (1999)
7. Li, J., Gao, H.: A multidimensional data model for data warehouse. Acta Sin. Sin. **11**(7), 200908–200917
8. Jia, J., Kamber, M.: Translated by Fan, M., Meng, X., et al. Data Mining Concepts And Technologies. Beijing Machine Press (2003)
9. Ian H. Witten EIBE Frank, New Zealand, translated by Dong, L., Qiu, Q., Ding, X., Wu, Y., Sun, L., Practical Machine Learning Technology for Data Mining. China Machine Press, Beijing (2006)
10. Hand, D., et al.: Principles of Data Mining. Beijing Machinery Industry Press (2003)

11. Oliviaparrrud: Data Mining Practice, translated by Zhu, Y. Beijing Machinery Industry Association (2003)
12. Li, X., Jin, F., Yu, H.: The architecture of decision support system based on data warehouse. J. Hefei Univ. Technol. (Nat. Sci. Ed.) (2003)
13. Data Mining Tutorial HTP/mdn2 Microsoft. COM/zh – Cnllibrary
14. Zhang, J.: Data Mining Extension Language DMX.Windowsitpromagazine International Chinese Edition (3) (2007)
15. Lv, W., Huo, Y., Lv, B.: Introduction and Improvement of c-2005. Beijing Tsinghua University Press (2006)