Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Application of Data Mining in Student Score Analysis

Shuyu Mei [1], Guozhu Chen[1,] *, Xinyue Wang[2], Hui Pan[1], Liang Zhou[1]

[1]College of Hydraulic & Environmental Engineering, China Three Gorges University, Yichang, 443002, China

[2] School of Materials Science and Engineering, Hubei University, Wuhan, 430062, China

*Corresponding Author: guozhuc@foxmail.com

Abstract: In recent years, with the emergence of big data as a popular vocabulary in the Internet information technology industry, education is increasingly recognized as an important application area where big data can make a difference. Especially in school education, data has become the most significant indicator of teaching improvement. Nowadays, big data analysis has been successfully applied to education and has become an important force in teaching reform. By analyzing big data, we can discover some important information and use them to provide personalized services to improve student achievement. In this paper, based on the application of data mining in student performance analysis, establish fuzzy comprehensive evaluation and related rules, to rank the comprehensive strength of the class and judge the correlation between the topics.
Keywords: Fuzzy comprehensive evaluation; association rules.

---

1. INTRODUCTION

Data mining is also translated into data mining and data mining. It is a step in database knowledge discovery. It generally refers to the process of searching for information hidden in it from a large amount of data, usually related to computer science, and through statistics, online analytical processing, information retrieval, machine learning, expert systems and Pattern recognition and many other methods to achieve the above objectives. Data mining technology is application-oriented from the beginning. It analyzes, analyzes, synthesizes and reasoning data from micro to macro, knows the solution of practical problems, finds the interrelationship between things and makes predictions, in scientific research and market. Marketing, financial market analysis and forecasting, fraud screening, healthcare, modern education and communication network management have been widely used. At present, data mining has become a hot spot in computer science and engineering research. China's education sector has been exploring how to accelerate the modernization of education, information construction, how to strengthen the quality education of students, and how to provide candidates with more humanized services and people-oriented modern education. In recent years, as big data has become a popular vocabulary in the Internet information technology industry, education has

gradually been recognized as an important application area where big data can make a difference. Some people boldly predict that big data will bring revolutionary changes to education. More and more exam information is processed and stored by computers, which greatly reduces manual processing, reduces storage space, and provides storage security and convenience. In this way, there are a large number of data of various examinations. How to find out the inevitable connection and potential relationship in the examination data according to the requirements and characteristics of different examinations has become an inevitable requirement of various examination management institutions.

In education, especially in school education, data has become the most significant indicator of teaching improvement. Usually, these data mainly refer to test scores. Nowadays, big data analysis has been successfully applied to education and has become an important force in teaching reform. By analyzing big data, we can discover some important information and use them to provide personalized services to improve student achievement.

2. DATA MINING FUNCTION DESCRIPTION

Data mining makes proactive, knowledge-based decisions by predicting future trends and behaviors. The goal of data mining is to find hidden and meaningful knowledge from the database. There are three main types of functions :

(1) Association analysis

Data association is an important class of discoverable knowledge that exists in the database. If there is some regularity between the values of two or more variables, it is called association. Associations can be divided into simple associations, temporal associations, and causal associations. The purpose of association analysis is to find out the associated networks in the database. Sometimes the association function of the data in the database is not known, even if it is known to be uncertain, so the rules generated by the association analysis have credibility.

(2) Cluster analysis

Records in the database can be divided into a series of meaningful subsets, clusters. Clustering enhances people's understanding of objective reality and is a prerequisite for concept description and deviation analysis. Clustering techniques mainly include traditional pattern recognition methods and mathematical taxonomy. In the early 1980s, Mchalski proposed the concept clustering technique. The main point is that not only the distance between objects is considered when dividing objects, but also the classified classes have some connotation description, thus avoiding some one-sidedness of traditional technology. For example, applicants are divided into high-risk applicants, moderate-risk applicants, and low-risk applicants.

(3) Concept description

Concept description is to describe the connotation of a certain type of object, and to summarize the relevant characteristics of such objects. Concept descriptions are divided into characteristic descriptions and distinctive descriptions. The former describes the common features of certain types of objects, while the latter describes the differences between different types of objects. Generating a characteristic description of a class involves only the commonality of all objects in that class of objects. There are many ways to generate a distinctive description, such as a decision tree approach.

3. MODEL PREPARATION

Data mining is a complex multi-stage process that can be divided into the following main stages:

(1) Determining the mining object: understanding the data, asking questions, and having a clear definition of the mining target. Recognizing the purpose of data mining is an important step in data mining. In the data mining of student achievement, we are targeting the student achievement of each student.

(2) Data preparation: Data preparation for providing high-quality input data for mining is a prerequisite for ensuring the success of data mining, and it accounts for the largest proportion in the entire data mining process. Data preparation can be divided into three sub-steps: data selection, data pre-processing, and data conversion. The first step is data selection. The data required for the data mining process may be obtained from different heterogeneous data sources, so the first step is to obtain data from a variety of databases, files, and non-electronic data sources. Collect all internal and external data related to mining, select data suitable for data mining applications, and build a data warehouse. The second step is data preprocessing. Due to the variety of data sources, data types, and metrics, there may be some irregular data, and there are some different operations that are implemented simultaneously. Wrong data can be corrected or eliminated, but missing data must be supplemented or predicted to prepare for the next analysis. This is usually done using data mining tools. The data in the student's grade data warehouse is imported from the transaction database of the school's academic affairs office. The teachers of each department are not standardized in the input of the grades, resulting in the existence or omission of the format and structure of the data, so the student's score data It is very important to carry out the cleaning. For the lack of performance, the average score of the course is generally used. Format conversion is not uniform for the format. For the test scores to have a test, re-test and re-learning, the first test scores are used. The third step is the conversion of data. To facilitate data mining, data obtained from different data sources must be converted to a uniform format. Some data may need to be encoded or become a format that is easier to use. Data reduction may be required to reduce the number of data attribute values considered.

(3) Data mining: The core of data mining is pattern discovery. This step is to use data mining tools and related algorithms to analyze all the obtained transformed data to produce the desired mining results.

(4) Interpretation and evaluation: Analyze and verify the mining results to find valuable information. The excavated patterns and rules are presented to the user in an intuitive, easy to understand manner. How the results of data mining are submitted to users is a very important issue, because the usefulness of data mining results depends mainly on this step. In the final step of data mining, various visualization tools and graphical user interfaces are often used to present the results.

4. MODEL 1: FUZZY COMPREHENSIVE EVALUATION

In the school student management work, excellent class selection is an important and complicated task. Many schools simply select excellent classes based on the academic performance of each class. This is very unreasonable. Because, whether a class is excellent or advanced, the factors involved can be recorded in addition to the academic scores, which can be recorded by specific figures. Many of them cannot be recorded by a single number, such as the ideological and moral status of class students.

Basic civilization, study style, etc. Although some schools have taken into account these factors, they have set a fixed weight value for each factor from the beginning. It is obviously unreasonable to use the same assessment form for different professional and different academic classes. .

Therefore, it is necessary to use a method that can minimize the influence of various subjective and objective factors and make the evaluation results more reasonable and reliable. Among them, when judging certain subjective factors, it is easier and more objective to use vague language descriptions than to use specific score descriptions. For the class comprehensive ranking, we establish a fuzzy comprehensive evaluation model to determine the ranking of each class.

Through the analysis of the data given by the topic, in order to select the class with the strongest comprehensive strength, we establish a fuzzy comprehensive evaluation model to conduct a comprehensive and specific evaluation and analysis of the performance of each class. The model is as follows:

$$Z = \sum_{i=1}^{n} u_n * a$$

Assume that the evaluation of a class's performance mainly considers the following factors: mean, median, range, skewness, kurtosis, standard deviation, variance, excellent rate, good rate, pass rate, and fail rate. Then the main factor set for the evaluation can be determined as:

$$U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}\}$$

According to the size of the numerical values of different indicators and the degree of influence of the index on the comprehensive evaluation, we collect the evaluation coefficients for different factors according to the actual situation. The specific evaluation coefficients are as follows:

Because the mean and median are usually the most important when measuring the overall score of a class, we give the mean and median an evaluation coefficient of 10; the worst response is the range of a class score, the smaller the difference, the description The results are more concentrated, so we take a negative value for the extreme difference, and give the evaluation coefficient of 0.1; the degree of deviation of the skewness response data, because the value is small, it is not comprehensive enough in the evaluation index function, we give it the evaluation coefficient is 10; peak value of kurtosis reaction, calculated according to the original value in the function; standard deviation and variance response is the distribution of the scores of a classmate's grades. The standard deviation is calculated according to the original value. Because the variance is too large, the weight in the evaluation is over Large, so we give the variance evaluation coefficient of 0.1; excellent rate, good rate, pass rate, failure rate response of a class of performance distribution, set values and the proportion of different levels on the comprehensive evaluation of the impact of a class performance, we once The evaluation coefficient is given as 50, 20, 10, -20.

So we can get the specific comprehensive evaluation index function as follows:

$$Z = 10u_1 + 10u_2 - 0.1u_3 + 10u_4 + u_5 - u_6 - 10u_7 + 50u_8 + 20u_9 + 10u_{10} - 20u_{11}$$

The larger the index value is, the better the comprehensive score of the class get

5. MODEL 2: ASSOCIATION RULES

The association rule reflects the interdependence and association between a thing and other things. If there is such a relationship between two things or multiple things, one thing can be predicted by other things. Based on the categories of variables processed in the rules, association rules can be divided into Boolean and numeric. The values processed by Boolean association rules are discrete and categorical, showing the relationship between these variables. Numeric association rules can be combined with multi-dimensional associations or multi-level association rules to process numeric fields, dynamically segment them, or directly process raw data. Of course, numeric association rules can also contain category variables. Based on the dimension of the data involved in the rule, the association rules can be divided into single-dimensional and multi-dimensional. In a one-dimensional association rule, we only relate to one dimension of the data, such as the item purchased by the user. In a multidimensional association rule, the data to be processed will involve multiple dimensions. The motivation of the association rule is to study the customer's shopping behavior, which is called the shopping basket problem. By analyzing the sales record data of the goods, it tries to find out the customer's shopping habits, and then finds the relationship between different goods. In other words, the association rule refers to the implication of the form, in which two sets of disjoint items, such as bread and milk, mean that in a purchase, if purchased, they will also be purchased, therefore, There is an association between them. There is support and confidence in this association rule. Support is the probability that in all transactions, and at the same time, the support is required to satisfy the conditions similar to the intersection.

Confidence means that as many of the purchased transactions have been purchased by as many customers, as is the case with the famous Wal-Mart beer diapers. Therefore, the complete definition of the association rules is as follows:

The so-called association rule mining refers to finding the item set satisfying the user specified condition (the support degree is greater than the minimum support degree and the confidence is greater than the minimum confidence) from the transaction record library D composed of a large number of transaction records T, and the item set is concentrated. The interesting and meaningful relationship that exists is the association rule.

Since the beginning of the issue of the association rules between itemsets in the consumer transaction history database, more and more people have studied the problem. With the deepening of the research, many association rule algorithms have been born, including: search algorithms, Width-first algorithm, depth-first algorithm, data set partitioning algorithm, and so on. Among them, the most powerful algorithm that can be used to mine frequent itemsets of association rules-the algorithm is also a typical width-first algorithm.

The algorithm mainly consists of two steps: generating candidate sets and pruning, that is, combining the item sets from them, generating candidate sets, recording them, and then selecting the required frequent itemsets from the candidate set, that is, they will not meet the requirements. The minimum support or confidence is deleted, and all non-empty subsets of frequent itemsets must also be frequent. In order to reduce the computational L and improve the efficiency of generating association rules, the algorithm uses a layer-by-layer search method. The item set is obtained by item set connection and pruning. The first candidate set consists of all items. Then, the first frequent item set is obtained by

pruning, the first frequent item set is obtained by self-joining and pruning to obtain the second frequent item set, and so on, until the frequency item set is no longer found.

If there are 737 student scores corresponding to 20 questions, and the test paper has 8 multiple choice questions, 7 fill-in-the-blank questions, and 5 test questions, the association rules run as follows:

(1) Data extraction and preprocessing

There are 8 multiple-choice questions in the test paper, 7 fill-in-the-blank questions, and 5 calculation questions. In this regard, we establish an association rule algorithm model for specific correlation analysis. The algorithm is a frequent item set algorithm for mining association rules. The core idea is to mine frequent itemsets through two stages: candidate set generation and plot closed detection.

Because the algorithm can only process Boolean variables, in order to better analyze the data, we need to standardize the score data. First, the standard for each sub-question. The principle of standardization is to compare the scores of each sub-question with the mean of each sub-question. If it is greater than or equal to the mean, the score is set to 1, otherwise it is set to 0.

(2) Data integration

In order to integrate data from different data sources, it is also necessary to perform integration operations on the data in a certain way.

(3) Generation of frequent items

The 737 student scores in this question correspond to 20 questions, that is, the algorithm is used to find frequent itemsets in the transaction set. Set the minimum support count to 3 . First, the first candidate set is generated, and the specified minimum support degree is compared, and the frequent requirements are met, and no pruning is required, and then the second candidate set is generated by the self-joining method, and according to this method, the frequent item set is generated. Find a collection of frequent itemsets from the set of candidates that contain each item. The join is then generated using joins, and candidate sets of itemsets with infrequent subsets are deleted based on the nature of the algorithm. Next, the data that has been represented by a Boolean is scanned, and the supported count of the candidate set is counted to form a frequent set compared to the minimum supported count.

For each frequent item set, all non-empty subsets generated, if

$$\frac{\text{support\_count}(C)}{\text{support\_count}(s)} \geq \min\_conf$$

Then output the rules "$s \Rightarrow (C-s)$". $\min\_conf$ is the minimum confidence threshold.

(4) Confidence

Taking the generated frequent itemsets as an example. And finally obtain the confidence of each non-empty subset of the frequent itemsets according to the association rules.

6. THE ADVANTAGES OF THIS MODEL

Advantage 1: The computational complexity is not high and easy to explain and understand.

Advantage 2: The established fuzzy comprehensive evaluation model avoids the subjectivity of the target selection by experience, so that the final comprehensive strength ranking result is scientific and

accurate. At the same time, when evaluating certain subjective factors, it is easier and more objective to use vague language descriptions than to use specific score descriptions.

7. THE SHORTCOMINGS OF THIS MODEL

Disadvantage 1: In the question, we understand that the weights given are all our own presets. Such scientific basis is obviously inaccurate. We can consider using AHP to estimate the weights.
Disadvantage 2: The model used in this paper is relatively simple, and the results of the analysis may be significantly different from the actual ones.

8. CONCLUSION

This paper uses data mining technology to study the changes of students' performance, find the real reasons that affect students' performance, and formulate corresponding measures to improve the quality of education and teaching. It provides reference for college education management. Data mining technology can also be applied to other areas of school management, such as using data mining for student information management, student psychological management, student employment analysis, and teaching quality assessment. Data mining technology is also used in the banking industry. Data mining can discover or dig deeper and more detailed aspects of this relationship from a large number of historical records. For retail companies, information on product sales, customer information, inventory, and supermarket storefront information can be collected from supermarket sales management systems, customer data management, and other operational data. Data is collected from various application systems, classified according to different conditions, stored in a data warehouse, allowing management personnel, analysts, purchasing personnel, market personnel and customers to access, using data mining tools to analyze these data, providing management with Efficient scientific decision making tools.

ACKNOWLEDGMENTS

The authors express their appreciation to the College of Hydraulic & Environmental Engineering (Science and Technology Project of Undergraduate Students), China Three Gorges University for financial support of this work.

REFERENCES

[1] Liangbin Yang. Visualization Analysis of Research Status and Trends in Data Mining[J]. Library and Information Service, 2015, 59(S2): 142-147.
[2] Zhongmei Shu, Xiaodong Xu, Qiongfei Qu. Student Input Model and Learning Analysis Based on Data Mining[J]. Journal of Distance Education, 2015, 33(01): 39-47.
[3] Zhengyi Rao, Nianbai Fan. A Survey of Apriority Algorithms for Association Rules Mining[J]. Computer Age, 2012(09):11-13.