MOOC Learning Behavior Analysis and Teaching Intelligent Decision Support Method Based on Improved Decision Tree C4.5 Algorithm

https://doi.org/10.3991/ijet.v14i12.10810

Qiang Wu
Anhui Jianzhu University, Anhui, China
rmgyrrezptpv7@163.com

Abstract—To better carry out Massive Open Online Courses (MOOC) teaching evaluation and improve teaching effect, firstly, a teaching decision support system with evaluation function is designed by analyzing the actual situation of the college. Secondly, the decision tree data mining algorithm is introduced in the subsystem of student score analysis and evaluation. Finally, the decision tree model of student score analysis evaluation is constructed according to the decision tree algorithm. Through the practical exploration of applying the decision tree algorithm to the MOOC teaching evaluation management system of higher vocational colleges, it is found that the application of data mining technology to the construction of digital campus is not only reflected in the theoretical feasibility, but also reflected in its technical feasibility.

Keywords—Decision tree C4.5 algorithm, MOOC learning behavior, intelligent decision support

1 Introduction

As a new teaching method that has developed rapidly in recent years, online teaching technology has been widely welcomed by society and colleges. Online teaching has abundant education resources and provides learners with various kinds of knowledge information with rich and colorful pictures and texts. Different from traditional teaching, with the increasing abundance of online teaching resources, it is more and more difficult for online learners to obtain objective learning scores. On the one hand, online teaching is an open environment. Learners often fail to get proper learning results according to their own learning behaviors, resulting in a waste of time and energy; on the other hand, summative scoring methods (such as final exams) can’t fully reflect the learner’s entire learning process. It is the above reasons that make it impossible to effectively determine the effect of online teaching during the application of online teaching. Therefore, building a more objective and comprehensive intelligent decision support system for online teaching has become the focus of current thinking. In this regard, based on the understanding of learners’ e-learning behavior, MOOC learning behavior analysis and intelligent decision support system for teaching is established through the construction of decision tree model. With the help of C4.5 decision tree algorithm, a teaching decision support system suitable for higher vocational colleges and universities is designed. From the perspective of student score analysis and evaluation data, valuable information hidden in the student score data is mined to help the college more effectively apply strategies to discover knowledge on education teaching. It is also hoped that the construction of this system will provide more favorable references for the current large number of online teaching systems, so as to provide references for the adjustment and improvement of online teaching systems, so as to further improve the effect of online teaching.

2 Literature Review

The study of e-learning behavior is mainly to apply the research on human behavior to the field of computational education. Since the emergence of e-learning, a new learning model, the research on e-learning has quickly become the focus of many scholars. The main contents of the research are as follows:

Firstly, it describes the connotation of e-learning behavior, summarizes the rules of learning behavior and the function of learning behavior in building personalized network learning environment. The definition of e-learning behavior is the basis of in-depth study of learners' learning behavior. Lakshmi (2013) et al. believed that learning behavior refers to the response of learners to specific stimuli in a specific learning environment, which can be observed [1]; Muralidharan (2013) et al. pointed out that learning behaviors are composed of learning elements, which are continuous and correlated [2]; Gokgoz (2015) et al. believed that learning behavior is triggered by learning motivation, which is an interactive behavior brought by learners using learning tools provided by learning environment to achieve a certain learning goal in the learning process, including the behavior of learning process and the result of learning [3]. According to most views, network learning takes place in the network environment, which is composed of observable learning elements, and there is a relationship between different learning behaviors. Therefore, in the network learning environment, including the MOOC learning environment, the learner's learning behavior can be effectively tracked and collected, so that the collected data can be deeply researched and utilized.

After analyzing learners' e-learning behavior, the learning style of learners is obtained. Kumar (2013) et al. proposed that the learning style is a relatively stable learning behavior structure represented by the learner in a specific learning environment, reflecting the preference of the learner to choose the learning mode in the learning process [4].

There are many learning style models proposed at present. Among them, Felder-Silverman learning style model is widely applied. This model points out that learning style is a specific way for learners to acquire knowledge, process information and understand knowledge in the learning process. Jiang (2013) et al. used the naive Bayesian tree combined with the binary correlation classifier to realize the establishment of the Felder-Silverman learning style model, which classifies learners according to their learning interests [5]. The research on learning behavior mainly focuses on describing behavior characteristics, summarizing and discovering behavior rules, and modeling the learning process of learners. The theory of centralized student model is proposed, such as rule-based method, fuzzy logic, case-based reasoning, Bayesian network and decision tree, etc., and the data is analyzed and processed through tools and methods such as cluster analysis and logical reasoning of artificial intelligence.

In terms of modeling and analysis of learners' learning behaviors, Barros (2014) et al. build models based on learners' learning behaviors, which can better describe learners' learning behaviors, and then accurately predict learners' learning behaviors [6]. In the process of supervising the learning behavior model, the most widely used techniques are high-order colored Petri net, Bayesian network and automata theory. High-order colored Petri net is proposed by Pandey (2013) et al. [7]. And the acquired learner behavior model can effectively predict their subsequent learning behavior. Using the Bayesian network is a relatively simple way to model, Barros (2013) et al. all successfully used Bayesian network to establish a learning behavior model and applied it to predict learners' learning behavior in actual scenarios [8].

At present, the study of learners' learning behavior in online courses is more inclined to the study of learners' learning behavior, instead of the correlation between learning behavior mode and learning effect. In addition, the data used in the study are mostly from WEB logs and questionnaires, and the learning data generated by learners in the network platform is not used. Therefore, it is difficult to truly and effectively reflect the whole learning process of learners. At the same time, the current studies on learning behaviors in MOOC are all one-sided, or focus on the analysis of different learning effects brought by different learning behaviors, or the statistical analysis of influencing factors, or the study of different learning styles; it doesn't start from analyzing the influence of learning behavior on learning effect, deeply find the deep correlation among these factors, and make a deep analysis of the correlation. In this study, all the research works are connected in a series, and the practical application value is explained, which provides better suggestions for learners, teachers, and the design and development of MOOC platform.

The first part introduces the research background, and the second part describes various researches on online learning behavior. The third part introduces the research method and the principle of decision tree algorithm, and then constructs the model of online learning scoring process. The fourth part applies the decision tree algorithm to the MOOC teaching evaluation management system of vocational colleges and tests the system. The innovation of this research lies in the introduction of c4.5 algorithm. On the one hand, the traditional algorithm can only process nominal data, but not continuous data, which limits the processing of the algorithm to a certain extent. On the other hand, the traditional algorithm is poor in data fitting degree. Therefore, the introduction of C4.5 algorithm not only improves the efficiency of data processing, but also eliminates the previous over-matching problem by means of pruning, which greatly improves the accuracy and quality of prediction. However, limited by the number of training data, there are still some problems in the analysis of learning behavior, which need to be further studied and modified to reflect the advantages of this model more objectively.

3 Methodology

3.1 Analysis of decision tree algorithm

In a variety of algorithms, the commonly used data mining algorithms are analyzed and compared as follows: in clustering algorithms, the class used to analyze and process data objects is unknown. Object sets are grouped into clusters of similar objects. For all grid clustering algorithms, there is a problem of quantitative scale. If the division is too rough, it will increase the possibility of different clustering objects being divided into the same unit, that is, insufficient quantization; if the division is too detailed, many small clusters will be obtained. Similarly, for the teaching evaluation data mining system, if the division is too rough, a reasonable classification can't be obtained, if the division is too detailed, it will lead to the overfitting of samples.

For Apriori algorithm, it is an algorithm to mine frequent item sets of association rules. It has great influence and is simple and easy to understand. However, it also has disadvantages: on the one hand, when candidate item set complement is transformed into frequent item set, in this process, transactions in the database need to be scanned repeatedly. At the same time, for each candidate set, its support degree needs to be calculated, so that the overhead of scanning the database will become larger. On the other hand, when the number of frequent item sets becomes large in the transaction database, the number of candidate item sets also becomes large. The large number of courses in colleges and large number of students make the workload of teaching evaluation large. The system needs to process a large amount of data when running, which requires good performance algorithm, so it is not suitable to select Apriori algorithm.

Genetic algorithm is an optimization method that can bionic the global. The disadvantage is that when applied, the algorithm is more complicated, and there are still problems that have not been completely solved. The advantage is that the data mining field involved is not deep, and it is relatively easy to combine with other models, which is different from many traditional methods and has the nature of implied parallelism.

Classification and prediction models are often used in decision tree algorithms, and in order to find potentially valuable information, a large amount of data needs to be classified. Simple description and fast classification are its main advantages, which are especially suitable for large-scale data processing. Its explicit rules are easy to extract, have relatively small calculations, can display important decision attributes, and have high classification accuracy.

The main reasons for the introduction of decision tree algorithm are: firstly, the rules generated by the decision tree method are still easy to understand for users without data mining knowledge background, such as teachers or teaching managers. Therefore, for the more professional data mining process, the simple explanation is very important for the practicability of the product. Since the final classification result of the decision tree is represented by the tree structure and the rules of the If-Then form, it is very close to the way that the real thing is recognized and represented by people. Secondly, the calculated amount is not large, which is the characteristic of the decision tree method. The system focuses on practical applications, which can shorten a large amount of calculation time, improve execution efficiency, and have relatively high work efficiency. Thirdly, the characteristic of decision tree method is that both continuous data and discrete data can be processed. This method can process the continuous time data well. More importantly, many methods can't process the discrete data effectively, but this method can still achieve good results. Fourthly, the importance of attributes can be clearly displayed with the decision tree. In this respect, since it chooses the split attribute, the attribute is determined by the computational information entropy that directly reflects the importance of the attribute. For a decision tree node, the more important the attribute it represents, the higher the level of its node. For nodes of the same level, there is no significant difference in size, so their functions are basically the same.

Combined with the advantages and disadvantages of the above algorithms and the characteristics of the data in the score database, the data in the student score database is analyzed and the decision tree method is selected, which is conducive to finding out the factors that affect students' scores and better discovering the relationships among the factors.

3.2 Construction of decision tree based on C4.5 algorithm

Data mining is a repetitive process, which can support decision, assist people to process data, or guide calculators to process data. Therefore, data mining technology is a kind of deeper data information analysis method in a certain sense. It is very meaningful to use data mining technology in test evaluation and analysis. Compared with traditional evaluation and analysis methods, data mining technology can find various hidden, relevant and important attribute features from the test results, and find out the inevitable connection between them through further analysis. This advantage can't be achieved by traditional methods.

In the system, data mining divides the given data to be processed into two independent sets, the training set and the test set. And after combining the data mining process with actual engineering practice, a detailed explanation should be provided based on the specific technical process of data mining. The test evaluation system should include requirements capture, analysis of design, implementation, and testing.

Firstly, the collected data information of students' exams is summarized to form the data source of examination information. These data structures require uniform specification, and then the data is processed so that the algorithm moves on to the next stage. The algorithm needs an indicator that repeatedly solves the best split, and then uses the appropriate pruning algorithm until the pruning generates the initial decision tree.

Construct a simple tree, which is why the C4.5 algorithm is chosen. The attributes of the decision tree are improved, the interpretable aspects of the metric are selected, and the corresponding pruning is performed, the empty branches are removed, and the unused branches are reduced to avoid excessive fitting. The C4.5 algorithm contains characteristics related to classification, and its information source is the decision tree.

Set T as the set label and Ci as the sample class label and initialize them. Set i=1, information entropy is a measure of the uncertainty of the corresponding attribute of an event. It is used in the algorithm to represent the amount of information waiting for data. The larger the indicator value, the larger the potentially uncertain information, and the better the classification of the data. The expression of information entropy is shown in formula (1):

$$O(\mathrm{T})=\sum_{i=1}^{m}\frac{frequ(C,\tau)}{|\tau|}\times log_2\frac{frequ(C,\tau)}{|\tau|} \tag{1}$$

The basic meaning is for the case in the set T, which is classified as the number in the class Ci. The number of samples in set T is described with the absolute value of T, and it is very convenient to calculate the probability of each data.

Formula (2) indicates that if the attribute X has n different values, the estimated value of information amount of the tree can be obtained by calculation. The expression of conditional gain entropy is shown in formula (2):

$$E_x=\sum_{i=1}^{n}\frac{T_i}{T}\times o(T_i) \tag{2}$$

The expression of gain amount is shown in formula (3):

$$\mathrm{gain}(\mathrm{X})=\mathrm{info}(\mathrm{T})-E_x \tag{3}$$

The corresponding information gain value can be obtained from the expected information and the entropy value, which is the C4.5 algorithm. It first selects the attribute X, finds the maximum value of the information gain and then tests the set T for the initial attribute. It then uses heuristic lookup methods to gradually test attributes, create nodes, comment tags, correspond to attribute values, and extend branches layer by layer. The calculation expression of the gain rate of information is shown in formula (4) :

$$\mathrm{Gainratio}=\frac{gain(X)}{Split(X)} \tag{4}$$

The expression of segmentation information amount is shown in formula (5):

$$\mathrm{Split}(\mathrm{X})=\sum_{i=1}^{K}\frac{|T_i|}{|T|}log_2\left(\frac{|T_i|}{|T|}\right) \tag{5}$$

The information gain rate is selected and adopted by the C4.5 algorithm. It can reflect the proportion of potential information, can be generated by branch operations, and can be used as a criterion for branch attribute selection. If the value of the information gain rate is large, it proves that there is more valuable information in the branch. The attribute with the largest value is selected as the split node of the decision tree, so that the attribute problem with more biased values can be effectively overcome. The form of IF-THEN can be directly expressed. The basic model of the decision tree is shown in figure 1:

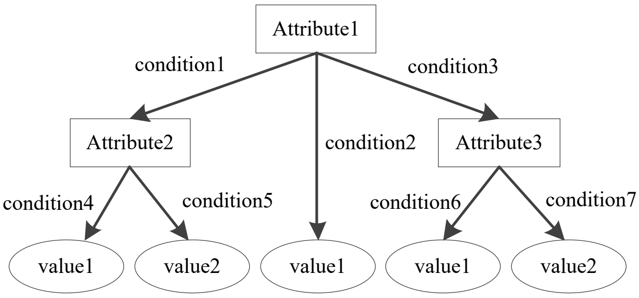

Fig. 1. Model of decision tree

Suppose that P represents a candidate attribute set and the discrete value attribute represents the training set sample S, a decision tree T is generated through the given training set.

The algorithm steps are as follows: step 1: create node N; step 2: judge whether the samples on the same class C are all samples of training set S; step 3: return N marked with class C as a leaf node; step 4: test attribute set P to determine whether it is null. If it is null, return to the most common class N in S as the leaf node; step 5: in S, select the attribute t with the highest information gain, and mark the node as t; step 6: for each known value ai in each t, a branch with a condition of t is generated by the node N; step 7: set Si as the sample set in S, and judge whether Si is null. If it is null, add a leaf node to it, mark the most common number in S, otherwise add T (si, S, t) to it.

4 Design and Implementation of Decision Support System

4.1 System design

Teaching decision support system structure diagram is shown in figure 2. The main module functions are as follows: data source: it contains information on courses, students, teachers, student scores, etc. These are taken from the college information management system. After cleaning, extraction, conversion, loading, etc., the data enters into the data warehouse. Model configuration: it can modify and add algorithms, excavate model graphical window configuration, and command line mode. Data visualization: it can report, graph, etc., present and implement data in various angles, export and store the mining results. Data warehouse: after processing the data in the data warehouse, a data mining system is established. It is based on decision support and is ultimately user-oriented for flexible querying. For example, the statistics and analysis of teaching and course setting can be carried out from various perspectives, and finally provide decision-making guidance. Control management: the configuration of the mining model is queried and set; user registration and login permission management are completed; the data visualization module is called, and the mining result is finally displayed.

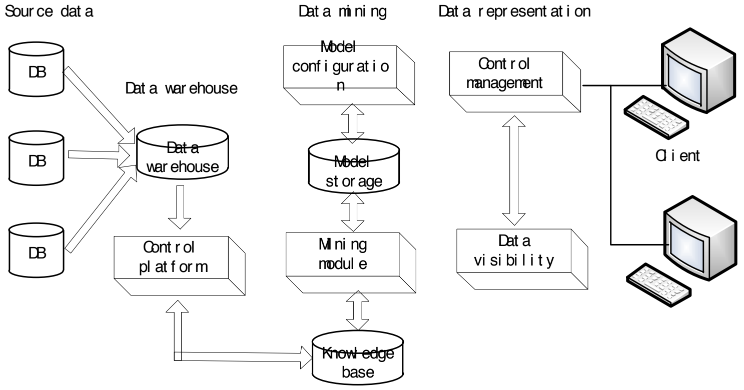

Fig. 2. Teaching decision support system structure diagram

4.2 Design of main modules

In the teaching decision support system, the emphasis is put on the design of the sub-system of student score analysis and evaluation, which is mainly to find out the main factors affecting the scores. The hidden information can help make decision makers to better set courses, arrange examination papers, improve teaching methods and improve teaching quality. The functional module diagram is shown in figure 3:

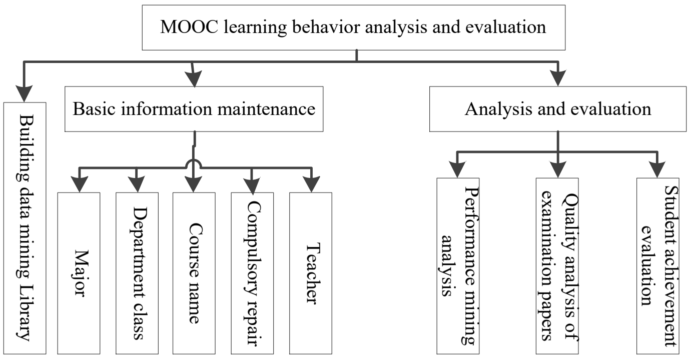
MOOC learning behavior analysis and evaluation
Building data mining Library
Basic information maintenance
Analysis and evaluation
Major
Department class
Course name
Compulsory repair
Teacher
Performance mining analysis
Quality analysis of examination papers
Student achievement evaluation

Fig. 3. The subsystem functional module of student score analysis and evaluation

For the mining function of analysis and evaluation control in the system, the Service component of Structured Query Language (SQL) Server Analysis is originally used. It can build multi-dimensional data sets for mining analysis, including data mining functions for cluster analysis and decision tree classification. In the design process, Clementine data mining tool developed by Statistic Package for Social Science (SPSS) is found, so that the data mining module is independent into a module, and the data storage and management only needed to be conducted in SQL Server. The implementation of the data mining algorithm in the application can be realized simply by calling the data mining tool. Script programming is used to invoke data mining tools for automated operations. The Clementine automation Clementine process is called by the system application, the data in the data mining library is called, the data mining flow is established, the imported data is processed, and the decision tree model diagram is displayed to the user. The result is stored back to the data mining library in tabular form, and the program is called to realize the data mining function in the system. The students' scores, examination paper quality and teaching quality are analyzed, the results are presented graphically and evaluated, hidden useful information is found, and the shortcomings are improved, which provided the basis for teaching decision support. The system flow chart is shown in figure 4:

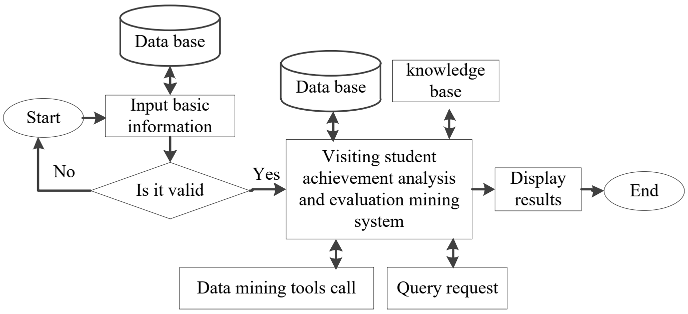

Fig. 4. Flow chart of the subsystem of student score analysis and evaluation

4.3 Decision tree model based on C4.5 algorithm and improved C4.5 algorithm

By using the C4.5 algorithm in the decision-making algorithm, the student score analysis and evaluation module in the subsystem is excavated, the score decision tree model is generated, and the student score information is analyzed and decided. Based on the analysis of student score, the function of the system can be analyzed with data mining technology. The implementation process of data mining is shown in figure 5: the first step is to input the sample data record of the information and proceed according to the students' grades. When building a model, it can accurately describe the score characteristics, establish score classification, and use the decision tree algorithm to construct a decision tree. The second step is to use the decision tree model to predict and classify the feature decision attributes of the evaluation indexes of students' scores according to their scores, and to form the condition rules. In the third step, the C4.5 algorithm in the decision algorithm mainly calculates the information gain for each attribute in the table.

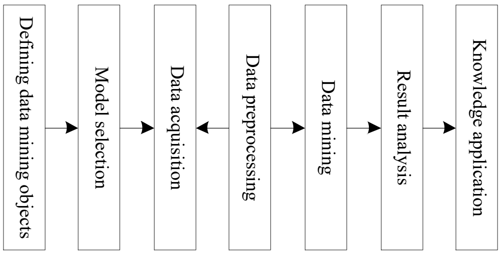

Fig. 5. The flow chart of the implementation of data mining

The following is the calculation flow chart, as shown in figure 6:

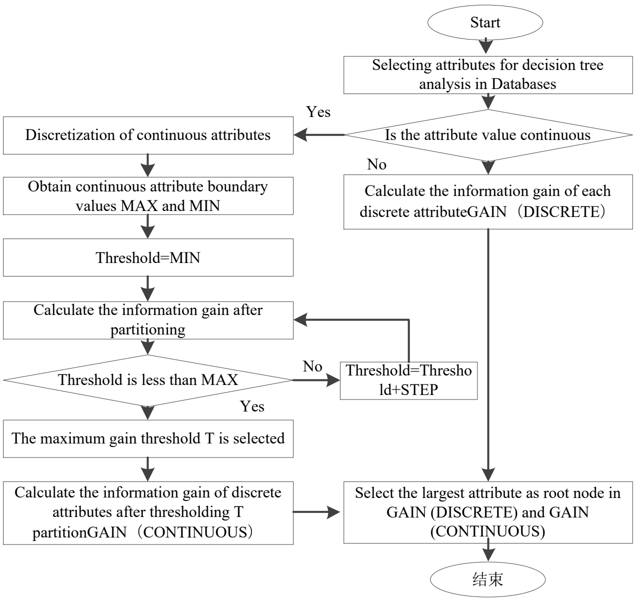
Start
Selecting attributes for decision tree analysis in Databases
Is the attribute value continuous
Yes
No
Discretization of continuous attributes
Obtain continuous attribute boundary values MAX and MIN
Threshold=MIN
Calculate the information gain after partitioning
Threshold is less than MAX
No
Yes
Threshold=Threshold+STEP
The maximum gain threshold T is selected
Calculate the information gain of discrete attributes after thresholding T partitionGAIN（CONTINUOUS）
Calculate the information gain of each discrete attributeGAIN（DISCRETE）
Select the largest attribute as root node in GAIN (DISCRETE) and GAIN (CONTINUOUS)
结束

Fig. 6. Information gain calculation flow chart

Taking the case sampling analysis of the online examination of the Computer Aided Design (CAD) course in the college as an example, the whole process of data classification mining is described in detail, such as the determination of object and target of data mining. The decision tree C4.5 is generated with the decision tree algorithm, and the workflow is introduced, including the data preprocessing technologies such as the definition of the problem, the preparation of the data, etc., and the decision tree model for completing the student score survey is established. The generated model generates classification rules, draws analysis conclusions, finds some valuable rules and information, and completes the establishment of decision tree model of whether students have passed the CAD online examination.

4.4 System implementation

The overall design of the system and the detailed design of each subsystem are continuously programmed to complete the design of the system. After repeatedly testing the system, it is found that the function of the system is basically normal, which is basically consistent with the assumption in the stage of demand analysis. The system has entered the trial stage, and there are many deficiencies in system research and development, and the system needs to be further improved. The design software of the system is a programming tool, which is implemented in the operating system environment. It makes a preliminary attempt in the general process of data mining and processing. The research results prove that the application of data mining technology in the construction of digital campus is not only manifested in the theoretical feasibility, but also in the technical feasibility.

4.5 System testing

The system test shows that the decision tree algorithm is effective at least for the evaluation of students in higher vocational colleges. In the follow-up work, the practical situation of the college will be further considered and the algorithm will be improved, and it is hoped that the system can better reflect the practical problems of the college, so as to guide the teaching practice. The problem has been solved: the application of data mining technology to the decision support system of higher vocational colleges has been initially realized. In the subsystem of score analysis and evaluation, the decision tree classification technology is applied to the score analysis to find out the main factors that affect students' scores of various subjects and whether they pass the final exam, as well as whether there is correlation between the factors. After making full use of the generated rules, it can provide guidance for the development of the school's teaching plan and keep students in a good learning state. The data mining tools are well combined with the system, so that the designed teaching decision support system has the data mining function, which can analyze and evaluate the students' scores, find the main factors affecting the students' scores, find out the relationship between various factors, make the graphical interface easy to understand, provide the decision basis for teachers, and improve teaching management.

5 Conclusion

Firstly, demand analysis is carried out according to the actual situation of the college, and a teaching decision support system with evaluation function is designed in a targeted manner. This system makes full use of the existing teaching data and network resources of the college and is designed based on the browser server architecture. Secondly, the decision tree data mining algorithm is introduced in the subsystem of student score analysis and evaluation. According to the decision tree algorithm, a decision tree model for student score analysis and evaluation is constructed, and corresponding rules are extracted. The purpose is to obtain the required indicators affecting MOOC learning behavior through the decision tree algorithm. The significance of this part is to make some practical exploration on the application of decision tree algorithm to MOOC teaching evaluation management system in higher vocational colleges; in addition, the obtained results can lay a foundation and provide references for the subsequent improvement of teaching effect.

6 Acknowledgement

This work was supported in part by the Dam safety monitoring system for reservoirs based on.net, major project of natural science foundation of education department of Anhui province, project approval number: ZD200906; Integrated research and demonstration of key technologies related to water environment safety monitoring system of key lakes and reservoirs, major science and technology special project of Anhui province in 2017 (project no. 16030801118).

7 References

[1] Lakshmi T M, Begum R M, Venkatesan V P. An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data. International Journal of Modern Education & Computer Science, 2013, vol. 5(5), pp. 18-27. https://doi.org/10.5815/ijmecs.2013.05.03

[2] Muralidharan V, Sugumaran V. Feature extraction using wavelets and classification through decision tree algorithm for fault diagnosis of mono-block centrifugal pump. Measurement, 2013, vol. 46(1), pp. 353-359. https://doi.org/10.1016/j.measurement.2012.07.007

[3] Gokgoz E, Subasi A. Comparison of decision tree algorithms for EMG signal classification using DWT. Biomedical Signal Processing and Control, 2015, vol. 18, pp. 138-144. https://doi.org/10.1016/j.bspc.2014.12.005

[4] Kumar A R S, Goyal M K, Ojha C S P, et al. Application of ANN, Fuzzy Logic and Decision Tree Algorithms for the Development of Reservoir Operating Rules. Water Resources Management, 2013, vol. 27(3), pp. 911-925. https://doi.org/10.1007/s11269-012-0225-8

[5] Jiang F, Sui Y, Cao C. An incremental decision tree algorithm based on rough sets and its application in intrusion detection. Artificial Intelligence Review, 2013, vol. 40(4), pp. 517-530. https://doi.org/10.1007/s10462-011-9293-z

[6] Barros R C, Basgalupp M P, Freitas A A, et al. Evolutionary Design of Decision-Tree Algorithms Tailored to Microarray Gene Expression Data Sets. IEEE Transactions on Evolutionary Computation, 2014, vol. 18(6), pp. 873-892. https://doi.org/10.1109/tevc.2013.2291813

[7] Pandey M, Sharma V K. A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction. International Journal of Computer Applications, 2013, vol. 61(13), pp. 1-5. https://doi.org/10.5120/9985-4822

[8] Barros R C, Basgalupp M P, Andre C. P. L. F. de Carvalho, et al. Automatic Design of Decision-Tree Algorithms with Evolutionary Algorithms. Evolutionary Computation, 2013, vol. 21(4), pp. 659-684. https://doi.org/10.1162/evco_a_00101

8 Author

Qiang Wu Doctor of Engineering, Associate Professor in the School of Electronic and Information Engineering, Anhui University of Architecture. His main research fields are: artificial intelligence, data mining, image recognition technology and its application (nvutzc@163.com).

Article submitted 2019-04-07. Resubmitted 2019-05-25. Final acceptance 2019-05-27. Final version published as submitted by the authors.