Study on Effect and Value of Result Analysis with Decision Tree Algorithm in Colleges

Fengxian Deng

Hainan College of Software Technology, Qionghai, 571400, China

Keywords: college; decision tree algorithm; result analysis; effect and value

Abstract: in efficient management teaching, students' results are important basis to evaluate teaching quality and teaching effect. Factors influencing students' results are complex and diversified. It is necessary to reasonably utilize data mining technology and adopt decision tree to analyze and predict students' results, correct bad behaviors influencing students' results in time in allusion to predication results and change teaching strategies. This paper studies application and value of decision tree algorithm in analyzing students' results in colleges.

Teaching objective of higher education is to cultivate high-quality elites and inter-disciplinary talents and improve teaching quality. However, in teaching, students' results are an important indicator to evaluate students' knowledge mastery and also important basis to evaluate teaching quality. Reasonable analysis and prediction of students' results can provide important basis for enhancing teaching management, improving teaching environment, boosting teaching quality and deepening teaching reform. Data mining technology is the foundation and precondition of further deep-level data analysis in decision-making process. So, it is very significant to apply data mining technology in result analysis. It can comprehensively analyze the relationship between exam results and factors influencing results. When data mining technology is used to analyze students' results, relevant results can be gained in time and students, and bad behaviors can be corrected in time, too.

I. Importance of analyzing students' results in colleges

In current education, students' results are a basic standard to measure students' knowledge mastery and important basis of evaluating teaching quality. In actual teaching process, teachers will generally accumulate a large quantity of data, utilize relevant technology to rationally analyze and mine data, transform classification rules, and carry out quantitative analysis of data from many aspects so as to make sure the relationship between various factors and exam results can be displayed clearly. Analysis of students' results with certain technology, application of data mining technology in analysis of students' results and rational expression of the problems contribute to formulating corresponding strategies and measures by teachers and relevant departments and improving teaching quality and teaching effects.

Main significance of data mining lies in rational analysis of huge data knowledge among massive data and mining unknown data with potential value and influence on decisions as powerful basis for relevant decisions. Decision tree algorithm is a relatively important algorithm in data mining. The tree-shaped structure is used to express results so as to better understand the data. Besides, in actual data mining process, data are regarded as the main research object. Through comparing traditional analysis methods and combining traditional technology, fuzzy mathematics and visualization technology, new mining technologies and methods form. The new mining technologies mainly include association rules algorithm, genetic algorithm, artificial neural network, decision tree algorithm, visualization technology and rough set theory etc.

In teaching management process of colleges, students' results are important data which not just reflect teachers' teaching level to some extent, but also can evaluate students' learning situations. In teaching management, decision tree method can be used to analyze and comprehensively mine results of students from different majors. To specify relations among courses can greatly help teachers improve their teaching level. In actual teaching, more rational method should be chosen to better boost students' results and teaching level and provide guarantee for improving teaching quality.

II. Basic concept of decision tree algorithm

Decision tree is actually a visual and common classification algorithm and is also a tree-shaped structure chart similar to flow chart. According to different levels and nodes, it can be classified into three types: internal node, root node and leaf node. In addition, every node owns a corresponding sample. The highest point of the tree-shaped chart is the root node and can correspond to the whole sample data. Internal node can correspond to a class mark. Internal node and root node can test sample attribute. In accordance with actual testing results, samples can be classified into two or more subsets. Each subset can be classified into branches which are expressed with attribute value. Class mark corresponding to leaf node corresponds to types of sample data. Generally, rectangle is used to express the middle node of the decision tree. Oval node expresses the leaf. Basic structure of the decision tree mainly includes two steps: 1) decision tree generation; 2) reasonable pruning. Basic generation of decision tree starts from the root node and from top to bottom. The decision tree can continuously divide samples into subsets for construction. Besides, attribute testing can be carried out from root node. The testing results are used to confirm other nodes until the leaf node. Decision tree pruning aims to rationally prune structure of the decision tree and delete redundant branches so as to gain the decision tree with the expected minimum.

Basic causes why decision tree algorithm is applied are as follows: ① the main purpose of result analysis is to clearly know factors influencing results and carry out rational classification. The decision tree is the best method to solve such problem. ② During common result analysis, discrete variables are mainly used for evaluation. The decision tree is the best method for discrete analysis. ③ The decision tree can visualize data, which will not take too much time. It owns such characteristics of rapid convergence and modeling as well as clear structure.

III. ID3 algorithm

Among decision tree algorithms, ID3 algorithm is most influential. Its application scope is also wide. Basic process of ID3 algorithm is actually the recursion method from top to bottom to effectively collect samples. It is also a relatively classical greedy algorithm. In actual application process, information gain is utilized to judge the attribute of each node in the decision tree. In actual process of decision tree construction, large information gain should be selected as the node. In actual operation process, reasonable selection of the concept of entropy in information should be made according to node attribute.

(I) Information entropy

Information entropy is actually the expectation of different quantities of information. Information entropy is used to measure whether information source X is certain. Supposing sample data set is X, and symbolic number in information source is n, the possible value is ai. The value of ai probability is P (ai). So, the relational expression is:

$$H(X) = -\sum_{i=1}^{n} p(a_i)\log_2 p(a_i)$$

(II) Conditional entropy

Under different conditions, expected value of information entropy is conditional entropy. Signal source corresponding to Y is bi. Signal source corresponding to X is ai. P (ai/bi) is the expected probability. So, the relational expression is:

$$H(X) = -\sum_{i=1}^{n} p(a_i)\log_2 p(a_i)$$

(III) Average information gain

In actual operation process, information gain is used to express possible difference value between two quantities of information. During selecting classification attribute, the largest average information gain serves as the basic classification attribute. Average information gain relationship is:

$$G(X/Y) = H(X) - H(X/Y)$$

IV. Application of decision tree algorithm for result analysis

(I) To confirm mining target and object

The results of curriculum design in a college serve as the objects of data mining. The factors influencing students' results are gained through analysis of mining technology to provide powerful basis for teachers' teaching process. The result report is applied to analyze students' results and test papers to provide data information for evaluation by education departments.

(II) To select model

This paper selects decision tree for data mining and chooses classical mode of decision tree methods - ID3 to establish corresponding decision tree which mainly includes pruning and tree construction.

(III) Data cleaning and data collection

Result database mainly involves students' basic information, curriculum information and result data information. In combination of questionnaire survey, results are analyzed comprehensively according to students' results and questionnaire survey results to form decision tree of results. During establishing the decision tree, it is required to fully take into account of students' hobbies, interests, class attendance, test paper difficulty and assignment completion. Thus, questionnaire and some structures of exam results should be rationally extracted to establish basic structural table. Quantitative results of students' assignment completion and class attendance are shown in the following table.

Table 1 questionnaire survey results

Interests	Test paper difficulty	Class delay
A Yes	A High	A Often
B General	B Moderate	B Occasionally
C No	C Low	C Never

During data cleaning, information questionnaire form lacks some attribute values. For example, students have no result due to cheating or missing the exam. No similar result is given. Therefore, it is necessary to delete similar data. Through analysis and sorting, 12 pieces of records are gained.

Result analysis table forms after integrating the above three tables. However, since there are many result attributes in actual database, we select class attendance, students' inertest in curriculum, computer operation completion and test paper difficulty are chosen as the basic data to establish the decision tree. Through analysis and sorting, the following quantitative structure is gained, as shown in the following table.

(IV) Rational construction of result analysis decision tree with ID3 algorithm

Through data cleaning, training set of result distribution information of a curriculum in Table 4 can be gained. Exam results are mainly classified into three categories. So, data samples can be classified into fail, pass and excellent, expressed with C1, C2 and C3 respectively.

1. Calculate corresponding entropy. 12 pieces of records can output three kinds of results: excellent, pass and fail. Three pieces of records are "excellent". Six pieces of records are "pass". Three pieces of records are "fail". So, sample statistics is as follows: S1=3, S2=6 and S3=3.

According to entropy calculation formula $H(X) = -\sum_{i=1}^{n} p(a_i)\log_2 p(a_i)$, the following can be gained: $H(X) = -(3/12)\log_2(3/12) - (6/12)\log_2(6/12) - (3/12)\log_2(3/12) = 1.5$

Table 2 Result analysis

Student No.	Class attendance	Interest	Difficulty	Assignment	Result
	Good	Yes	Low	Good	Excellent
	Poor	No	High	Poor	Fail
	General	General	High	General	Pass
	General	Yes	Low	General	Pass
	Poor	No	Low	Poor	Fail
	Good	Yes	Moderate	Good	Pass
	General	General	High	Poor	Fail
	General	No	Moderate	General	Pass
	General	Yes	High	Good	Pass
	Good	Yes	Low	Good	Excellent
	Good	General	Moderate	Good	Excellent
	Good	Yes	Moderate	Good	Pass

2. Calculate average information gain. In accordance with assignment completion, curriculum interest, class attendance and test paper difficulty as root nodes of the decision tree, corresponding average information gain is calculated. Four corresponding average information gain values are G (assignment completion) =0.643 1577, G (class attendance) =0.756 8638, G (curriculum interest) =0.686 831 and G (test paper difficulty) =0.486 54. The value of class attendance is the maximum. So, (class attendance is selected as the basic root node. Then, the samples are divided into three parts for rational calculation to gain a decision tree, as shown in Fig.1.

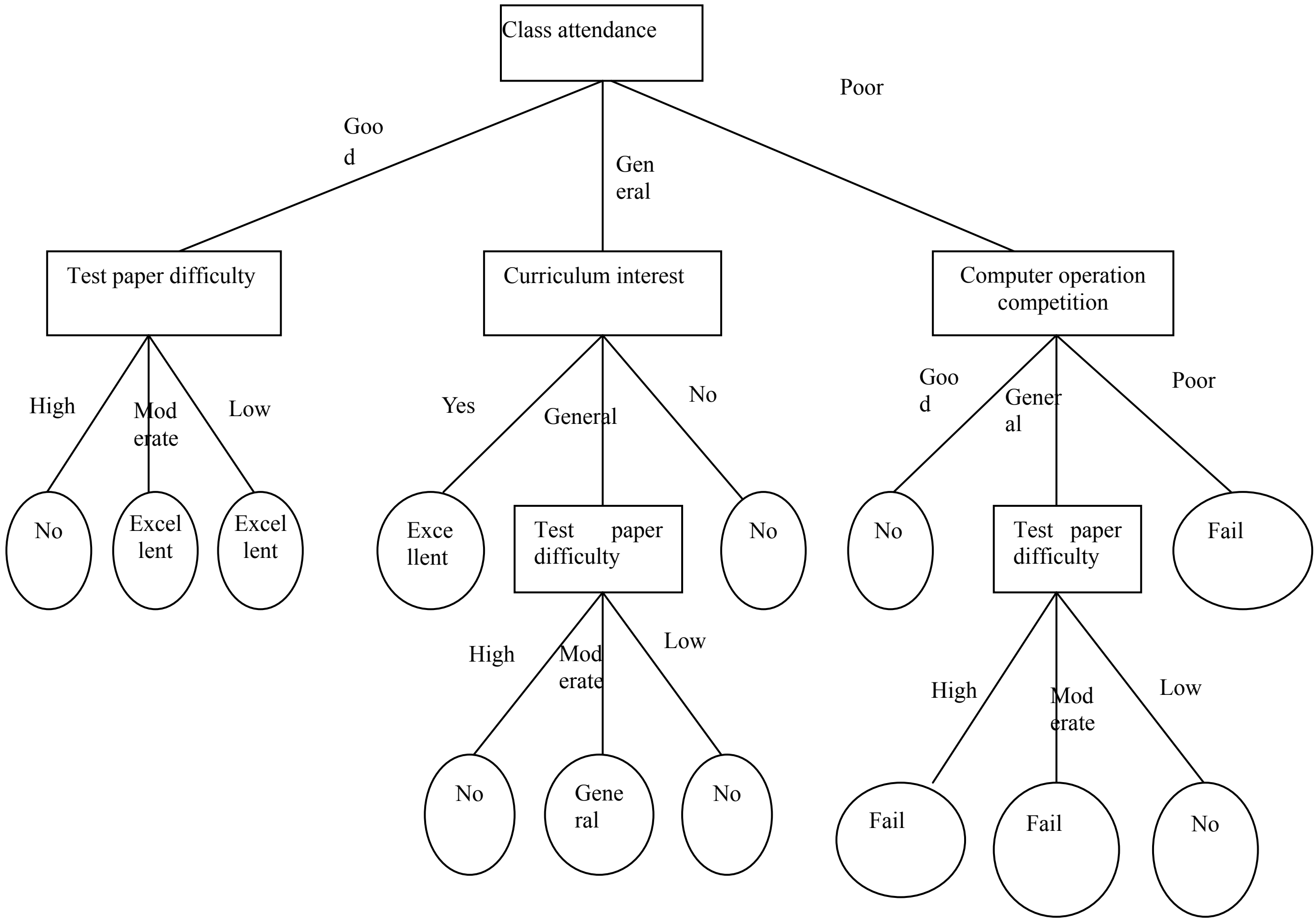

Fig.1 Result decision tree

(V) Revised decision tree

During establishing the decision tree, due to the influence of outlier and noise, data of many branches are ineffective or unusual. For final outcome, these redundant branches are unnecessary. This not just reduces data usability and understanding level, but also improves dependence on historical data. Therefore, certain pruning must be done. Mai methods include post-pruning and pre-pruning. Usually, post-pruning is selected to gain the decision tree as shown in Fig.2. The complicity of pruned decision tree reduces and the tree becomes small. So, it is more beneficial for observation.

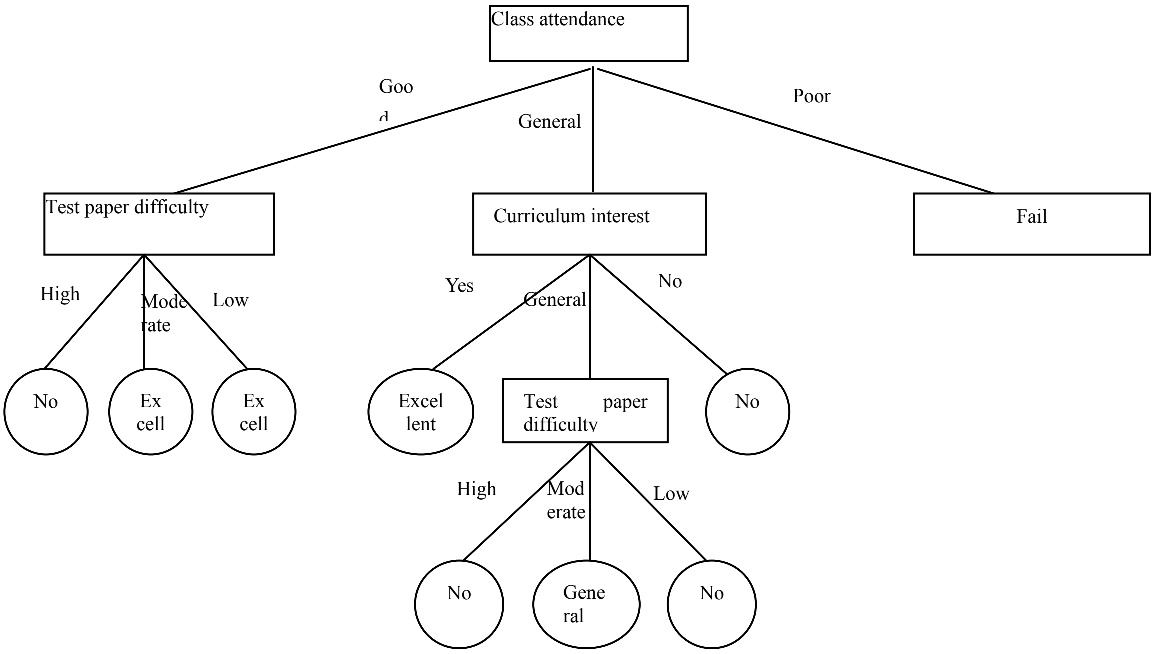

Fig.2 Revised decision tree

V. Conclusions

In short, data mining technology is a new data analyses method, but it has been widely applied in many fields. The decision tree method in data mining is used to analyze students' results in colleges and rationally construct a result decision tree so as to study effects of various curriculums on students' comprehensive quality. Teachers can gain rational solutions through data analysis, improve teaching quality and provide basis for cultivating high-quality inter-disciplinary talents.

References

[1] Kuang Tao, Applied study of decision tree technology in result analysis in colleges [J]. Journal of Pingyuan University (Natural Science), 2011, 28(1):49-51.

[2] Zhang Chunqin, Applied study of decision tree algorithm in result analysis in colleges [J]. Computer Programming Skills & Maintenance, 2010(12):112-113,118.

[3] Xuan Shili, Applied study of decision tree in analysis of English level exam results [J]. Computer & Digital Engineering, 2014,42(5):843-845,871.

[4] Cai Chunhua, Yang Liu, Application of improved decision tree in result analysis [J]. Microcomputer Information, 2010,26(36):284-286.

[5] Dong Huan, Applied study of decision tree in analysis of students results in colleges [D]. Xidian University, 2012:10.