Application Research of Data Mining in Grades Analysis and Course Correlation Analysis

Shaorong Feng
School of Information Science and Engineering
Xiamen University
Xiamen, China
shaorong@xmu.edu.cn

Abstract— Concerning the shortcomings of C4.5 algorithm, an improved one is put forward . This approach can improve the efficiency to generate the decision trees by reducing the computation complexity of information gain ratio. By applying the improved C4.5 decision tree algorithm to course grades analysis and course correlation analysis, provide powerful references that data mining is beneficial to improve teaching plan, which is also able to improve teaching quality.

Keywords-data mining*; *decision tree*; *information gain rate*; *grades analysis*; *correlation analysis

I. INTRODUCTION

Improving the teaching quality is the goal of every university and college and the exam grade is one of the important indicators to evaluate the teaching quality. Every year there will be a glittering array of grade data in the course of educational management and the traditional approach is to simply check and count the data without deeply analyzing the key factors that influence the student exam grades. Actually, there are various factors accounting for their exam grades, and the traditional grade analysis method cannot acquire the information from the exam grade data. The emergence of data mining technology provides powerful technology support for solving such problems. The data mining is a process to extract the implicit, potential and useful data people don’t know in advance from countless, incomplete, noise, vulgar and random practical applied data[1]. By applying the data mining technology to the course grades analysis and course correlation analysis in the educational administration system, it can be obtained that the data cannot have the traditional approach. Then, analyze the factors which can analyze the student grades through the data so as to pertinently improve the teaching approach and help students overcome the difficulties to properly arrange and set the teaching course, effectively guide students to chose proper courses to finally provide scientific guarantee for teaching management and teaching reform[2-4]. In the premise of analyzing the classification approach C4.5, apply data mining technology to the course grades analysis and course correlation analysis to find out concrete factors that influence the grades which can pertinently help students overcome the unfavorable factors, improve grades as well as teaching material to enhance the talents level.

II. DECISION TREE C4.5 ALGORITHM AND ITS IMPROVED ALGORITHM

The aim of data mining is not only to extract valuable data or relevant information rule from mass data but also carry out analysis and application so that people can have a better understanding for the rules and knowledge being extracted which can be adopted to solve the practical problems. The typical task for data mining includes concept description, associative analysis, classification and forecasting as well as cluster analysis[5].

Decision tree is a kind of modeling approach which can carry out classification and forecasting and is used to solve the data classification[6]. Its operating principle is showing in Fig. 1.

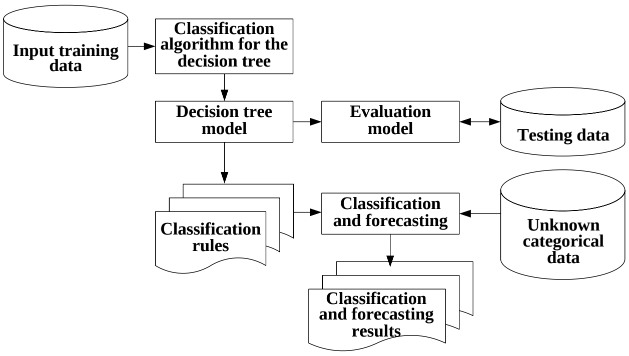
Input training data
Classification algorithm for the decision tree
Decision tree model
Evaluation model
Testing data
Classification rules
Classification and forecasting
Unknown categorical data
Classification and forecasting results

Figure 1. The work flow for decision tree

The generation algorithms for the decision tree include ID3[7],C4.5[8],CART[9],CHAID[10] etc and ID3 is suggested by J. Ross Quinlan which is a greedy algorithm, adopting a top-down decision tree. In inductive learning, it represents a method of decision tree. The following shows the demerits of ID3 algorithm.

Firstly, it is easy to forecast the discrete column but difficult to forecast the continuous column;

Secondly, the error ratio is in proportion to the increasing classification;

Thirdly, the calculation of information gain depends on property with more value.

A. The algorithm for decision tree C4.5

In order to solve the existing problems of ID3, Quinlan put forward C4.5 algorithm which is another classified decision tree algorithm in machine learning algorithms. Besides, it is an important algorithm and is improved based on ID3. Comparing to ID3, it has the following features:

Firstly, select the property based on the information gain ratio. ID3 selects the property based on the information gain ratio of subtree, and at this point, define the information with lots of approaches. ID3 uses entropy (which is an impurity level measurement criterion), or we can say the change value while C4.5 adopts information gain ratio.

The information gain ratio is obtained from the proportion of information gain to information segmentation. Concerning T data, suppose A has S values with different discrete attribute, and there are v subsets from S_1, S_2, to S_v. The algorithm for information gain $Gain(S, A)$ which is obtained by diving sample data based on A is same to ID3 and the algorithm to divide the information quantity shows as follows:

$$Split(S,\ A)\ =\ -\sum_{i=1}^{v}\frac{S_i}{S}\log_2(\frac{S_i}{S}) \tag{1}$$

The algorithm for information gain ratio shows as follows:

$$Gain-Ratio(S,A)=\frac{Gain(S,A)}{Split(S,A)}=\frac{I(S,A)-E(A)}{Split(S,A)} \tag{2}$$

Secondly, carry out pruning in the course of constructing the decision tree because some nodes with fewer elements may make overfitting so the result should be better if these nodes are not taken into consideration.

Thirdly , it can also deal with the non discrete data.

Fourthly, it can deal with the incomplete data.

```
Algorithm C4.5(D)
Input: an attribute-valued dataset D
Tree={ }
if D is "prue" OR stopping criteria met then
    terminate
end if
for all attribute a∈D do
    Compute information-theoretic criteria if split on a
end for
abest=Best attribute according to above computed criteria
Tree=Create a decision node that test abest in the root
D_v=Induced sub-datasets from D based on abest
for all D_v do
    Tree_v=C4.5(D_v)
    Attach Tree_v to the corresponding branch of Tree
 end for
 return Tree
```

B. *The improvement principle for decision tree C4. 5 algorithm*

In the course of calculating the information quantity, adopt logarithmic function and the approach of higher mathematics to simplify our computational complexity. In the following, put forward an advanced approach to calculate information content which can be adopted to simplify the measures and complexity.

C4.5 algorithm firstly calculates the information gain ratio of the node property and then compares the gain ratio to finally have the best value as the property of this node. While calculating the information gain ratio, adopt logarithmic function and the generative process for the entire decision tree is created by recursively calling function in the database which undoubtedly will cost much time. Adopt the approach of advanced mathematics to reduce the amount of computation and simplify the attribute selection so that improve the efficiency of generating decision trees.

Suppose the information content for n counter example, and p positive attributes is

$$I(p,n)=-\frac{p}{p+n}\log_2\frac{p}{p+n}-\frac{n}{p+n}\log_2\frac{n}{p+n} \tag{3}$$

The information impurity is

$$E(A)\ =\ \sum_{i=1}^{v}\frac{p_i\ +\ n_i}{p\ +\ n}\,I(n_i,\ p_i) \tag{4}$$

The results are based on Formula 3

$$I(n_i,p_i)=-\frac{n_i}{n_i+p_i}\log_2\frac{n_i}{n_i+p_i}-\frac{p_i}{n_i+p_i}\log_2\frac{p_i}{n_i+p_i}$$

After simplifying the Formula 4,

$$E(A)\ =\ \frac{1}{(p\ +\ n)ln2}\sum_{i=1}^{v}(-p_i ln\,\frac{p_i}{n_i\ +\ p_i}\ -\ n_i ln\,\frac{n_i}{n_i\ +\ p_i}) \tag{5}$$

The value of $1/((p+n)ln2)$ is constant which can be ignored, $E(A)$ can be obtained,

$$E(A)\ =\ \sum_{i=1}^{v}(-p_i ln\,\frac{p_i}{n_i\ +\ p_i}\ -\ n_i ln\,\frac{n_i}{n_i\ +\ p_i}) \tag{6}$$

According to relevant theoretical knowledge of Taylor formula and equivalent infinitesimal, conclude that if the x value is small, $ln(1+x)\approx x$ to finally have the following

$$ln\,\frac{p_i}{n_i\ +\ p_i}\ =\ ln(1\ -\ \frac{n_i}{n_i\ +\ p_i})\ \approx\ -\ \frac{n_i}{n_i\ +\ p_i} \tag{7}$$

$$ln\,\frac{n_i}{n_i\ +\ p_i}\ =\ ln(1\ -\ \frac{p_i}{n_i\ +\ p_i})\ \approx\ -\ \frac{p_i}{n_i\ +\ p_i} \tag{8}$$

Substitute Formula (7), (8) into (6), it can be obtained,

$$E(A)\ =\ \sum_{i=1}^{v}(-p_i ln\,\frac{p_i}{n_i\ +\ p_i}\ -\ n_i ln\,\frac{n_i}{n_i\ +\ p_i})\ \approx\ 2\sum_{i=1}^{v}\frac{n_i p_i}{n_i\ +\ p_i} \tag{9}$$

According to formula (1)

$$Split(S,A)=-\sum_{i=1}^{v}\frac{S_i}{S}\log_2(\frac{S_i}{S})=\frac{1}{Sln2}\sum_{i=1}^{v}S_i ln(\frac{S_i}{S})$$

The value of $1/(Sln2)$ is constant which can be ignored and according to $ln(1+x)\approx x$, it can be obtained,

$$Split(S,\ A)\ \approx\ \frac{1}{S}\prod_{i=1}^{v}S_i \tag{10}$$

Then,

$$Gain.ratio(S,A)=\frac{I(S,A)-E(A)}{Split(S,A)} \tag{11}$$

Based on the above analysis and treatment , adopt $2\sum_{i=1}^{v}\frac{n_i p_i}{n_i\ +\ p_i}$ to calculate the entropy value and while calculate the segmented information property, adopt $\frac{1}{S}\prod_{i=1}^{v}S_i$ which can simplify the calculation (sum, quadrature operation) with faster speed. Besides, adopt Formula (9) and (10) to replace $E(A)$ and $Split(S, A)$ of C4.5 which can reduce the calculation content and improve the efficiency of constructing the decision tree for C4.5.

C. *The improve algorithm for C4.5*

By advancing the C4.5 algorithm, simplify the entropy and take the maximum as the node by comparing to the information gain ratio.

Algorithm: Generate_decision_tree is created based on the given data.

```
Input: The samples are demonstrated by discrete attribute value and the attribute_list represents the candidate attributes set.
Output: a decision tree.
Approach:
{    Establish node N;
     if samples are in the same C then
          Return to N as the leaf mode and remark C;
     if attribute_list is empty then
          Return to leaf node N and remark as the common;
     Choose the maximum information gain ratio property Test_attribute in the attribute_list;
     Remark the node N as Test_attribute;
     for known bit $a_i$ in each Test_attribute
          The node generates branch of Test_attribute=$a_i$
          Suppose $S_i$ is the samples set of Test_attribute=$a_i$
          if $S_i$ is empty then
               to generate a leaf and to remark the samples as the most common one;
          else
               to generate a return node based on Generate_decision_tree($S_i$, attribute_list.Test_attribute)
}
```

III. THE APPLICATION OF IMPROVED C4.5 ALGORITHM IN GRADE ANALYSIS

Based on student test score data sets (Table 1), extract factors affecting students' test scores. Firstly, construct decision tree model and the property field in the student basic information table is abundant. So while determining whether the grade is the good decision tree mode or not, this paper takes the grades for computer science and technology professional 2012 level of C language programming design as examples, and select the fields which have big effects on the grades such as extra time to surf the internet, preparation before test, classing learning effect and common homework, and the property of whether it is good as the category property, the data is shown in Table 1.

TABLE 1 THE DATA SET TO DETERMINE WHETHER THE GRADE IS GOOD

Computer operation time after class/week	Preparation before class	Classing learning	Normal homework	Whether it is good
≤3	Basic understanding	Poor	Worse	No
3.5	Basic understanding	Basic grasp	Fairly good	Yes
3.5	Basic understanding	Basic grasp	Moderate	No
≥5	Understand some of them	Basic grasp	Moderate	No
3	Do not understand	Basic grasp	Fairly good	Yes
3	Understand some of them	Ordinary	Fairly Good	No
3	Understand some of them	Basic grasp	Fairly good	Yes
≤3	Understand some of them	Basic grasp	Worse	No
≤3	Basic understanding	Ordinary	Fairly good	Yes
3	Basic understanding	Totally grasp	Fairly good	No
……	……	……	……	……

There are 248 records(student) in the table 1, 72 positive example and 176 negative examples. By adopting the advanced decision tree C4.5, calculate the information gain ratio and get the maximum information gain ratio. Therefore, the daily grades can be regarded as the root nodes with 3 values, namely three branches. After that, calculate the branch information gain ratio and carry out comparison to get the decision tree showing in Fig. 2.

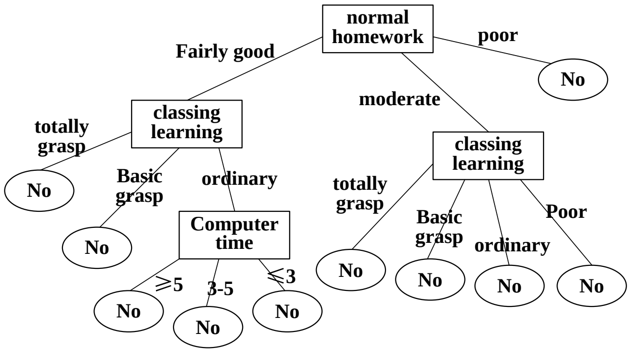

Figure 2. the decision tree demonstrating whether the grade is good

IV. THE APPLICATION RESEARCH OF IMPROVED C4.5 IN COURSES CORRELATION

A. *Application research for courses correlation*

By adopting the advanced C4.5, extract the data of professional course and proceeding course because students learn a dazzling array of courses. While setting their decision tree model, chose the grades for net program design and proceeding course and take the grades in the first term of junior year as examples to extract the relation between proceeding course and grades when the grade is excellent or good.

TABLE 2 GRADES FOR NET PROGRAM DESIGN AND PROCEEDING COURSE

Basics of Computer	C language programming	Java programming	Assembly language	Database technology	.Net programming
Ordinary	Good	Good	qualified	Good	Good
Excellent	Excellent	Excellent	Good	Good	Excellent
Excellent	Good	Good	Ordinary	Good	Good
Ordinary	Good	Good	Good	qualified	Good
Good	Excellent	Excellent	Good	Excellent	Excellent
Good	Ordinary	Good	Ordinary	qualified	Ordinary
Excellent	Good	qualified	Good	Good	Good
Poor	Ordinary	Ordinary	Poor	Poor	Poor
……	……	……	……	……	……

There are 248 records in the table 2 and 148 are positive examples (the grade is excellent and good for Net program design) while 100 are negative examples. By adopting the advanced decision tree C4.5, calculate the information gain ratio and the following classification rules can be concluded.

if Java programming is excellent then
    .Net programming is excellent(good)

if C language programming is excellent and database technology is good then

.Net programming is excellent(good)

if C language programming is good and Java programming is good then

.Net programming is excellent(good)

if C language programming is excellent and Java programming is excellent then

.Net program design is excellent

if C language programming is good and assembly language is good and database technology is good then

.Net programming is excellent(good)

if basis of computer is excellent and C language programming is excellent and Java programming is excellent and assembly language is good then

.Net programming is excellent(good)

By mining the course data for the pre and after term, generate the classification rules for the decision tree which is able to give us the information of the correlation and can help us to perfect the following work:

On the one hand, guide the students without taking good grades in preceding courses and help them overcome the learning difficulties to improve the grades.

On the other hand, provide powerful evidence for the scientific course schedule setting. While setting the course schedule, universities should take the continuity of proceeding course, professional course into consideration. In addition, references can be provided for the teaching plan which can not only help make the base for the proceeding course but also take the continuity among courses into consideration and is able to connect the correlation to improve the teaching effect and efficiency.

B. The meaning of relevant correlation research

By mining the student grade and course data, it can be found out that the correlation between student grades and courses. The data can provide valuable references for teaching department and is of significant meaning to guide the teaching activities:

1) Make educational behavior scientific

The existing education policy is one of the principles for educational theory and adopt analysis, inference, summary, and trail to make educational decision. Throughout the process of decision making, adopt traditional data statistics, sampling based on test data for a certain period of time to have a good knowledge of the limits. However, by adopting advanced computer technology, especially data mining, find out the rules among data by extracting the classification rules among historical education data which is able to make educational behavior more scientific.

2) Guide relevant course setting

In recent years, with the development of Chinese university the competition also becomes fierce and some of them open special courses so as to demonstrate the running features. However, newly open courses lack practical experience and the handling of new and traditional courses may be improper. Carrying out data mining among the historical data (including the courses opened in recent two years) can make it be possible to find a more scientific way which can make students be adapted to the studying and provide effective guidance.

3) Provide scientific references fro course setting

Existing course schedule is considered from the perspective of continuity of student learning, namely the preceding class is set first and then the appropriate follow-up courses. However, the continuity is not taken into consideration and the current course schedule may cause the condition that the time for the relevant course is longer and students may forget the contents which make it difficult to learn. If carry out data mining among historical data, people can find out the internal bounds of courses and then we can provide references for the following course schedule. Based on this, people do not only need to take the continuity into consideration but also the correlation, so that people can improve the teaching effect and efficiency.

By adopting the advanced C4.5, carry out data mining and create a decision tree to analyze the influencing rules for student grades. By layer digging it can be found out that the hidden relation among preceding courses and professional courses and generate correlation rules which can provide reliable references for the teaching department.

V. CONCLUSION

With the development of information technology and t he urgent need for information, the data becomes more and more complicated and how to extract valuable information in data mining technology is more and more challenging. In all industrial fields, managers are adopting advanced information technology to change the traditional pattern and the university information in recent years is developing fast. How to apply data mining technology to better serve the educational career is of significant meaning. This paper adopts decision tree C4.5 to carry out beneficial application and research on the student grades and provides perfected scientific basis for the teaching management.

With the deepening of research and application, the following problems have to be solved in the research and development work in the future:

Firstly, research on the data preprocessing method further improve data preprocessing method to reduce data processing handled by human resources so as to realize intelligent and ,automatic data preprocessing;

Secondly, deeply study the mining method for decision tree algorithm so as to improve the efficiency and quality of association rules mining.

These two aspects have significant meaning to improve the data mining quality and therefore, deeply research the above problems will be important for the following work and practice as well as application.

ACKNOWLEDGEMENT

This paper is financed by National Natural Science Fund Project(No.61303004), National Social Science Fund Project(No.13&ZD148); Natural Science Foundation of Fujian Province(No.2013J05099).

REFERENCES

[1] Han J, Kamber M., Fan Ming, Meng Xiaofeng. Data Mining: Concepts and Techniques [M]. Beijing: Machinery Industry Press 2007, 3-10.

[2] Liang Jinghua. Undergraduate Teaching Management Based on data mining technology [J]. Education Teaching Forum, Mar. 2014, No.13: 10-12.

[3] Song Jinping. Data Mining Technology and Its Application in Teaching Evaluation in colleges and universities [J]. Information and Computer : A database technology 2015, No.1: 85-86.

[4] Jiang Yun. Application of Data Mining technology in University Teaching Practice [J]. physical experiment, Mar. 2015, 35(3): 15-17, 20

[5] Zhenwei, Du Youfu, Qin Jianchao. Modern Data Mining technology and development [[J]. China science and technology information, 2007.2 (1): 7-8.

[6] Qian Chengdong. Data mining Based on Dynamic classification and its application in educational administration system [D]. National University of Defense Technology, 2006

[7] J. R. Quinlan. Induction of decision trees[J], Machine Learning, 1986, 1:81-106.

[8] J. R. Quinlan. C4.5: Programs for Machine Learning[M]. Morgan Kaufmann, 1993.

[9] L. Breiman, J. H. Friedman, R. Olshen, C. J. Stone. Classification and Regression Trees[M]. Chapman & Hall, New York, 1984.

[10] G. V. Kass. An exploratory technique for investigating large quantities of categorical data[J]. Applied Statistics, 1980, 29: 119-127