Forecast of Student Achievement Variation Trend Based on C4.5 Decision Tree

Li Li
School of Electronics and Information Engineering
Tongji University
Shanghai, China
lili@tongji.edu.cn

Shaoming Yao*
School of Electronics and Information Engineering
Tongji University
Shanghai, China
1333705@tongji.edu.cn

Zhou Ou
School of Electronics and Information Engineering
Tongji University
Shanghai, China
18817366097@163.com

Qijun Chen
School of Electronics and Information Engineering
Tongji University
Shanghai, China
qjchen@tongji.edu.cn

***Abstract*—In the work of educational management, student achievement is one of the most important evidences to evaluate the quality of a student. With respect to many factors that may affect student achievement, a new method based on C4.5 decision tree is proposed to predict student achievement variation trend. Firstly, reasonable attributes from student historical data of campus activities are selected. Secondly, the samples which have few records of campus activities are removed, and the attributes of remainder are discretized. Finally, a prediction model is established by C4.5 decision tree method to predict the variation trend of student achievement. The simulation results demonstrate that the prediction accuracy achieves 80.84%. As a result, the prediction model can effectively help educational management departments find the bad behavior of students and offer guidance to the students in time.**

Keywords-data mining; student achievement; C4.5 decision tree; prediction model

I. INTRODUCTION

Computer network is developing at an unprecedented speed, which leads to trend changes in society and much more convenient in life. In the meantime, the development of database technology makes it possible that a mass of user' information is stored. With the development of the campus information, colleges gradually start to build their own data warehouse, which makes it easier to reach the campus data and improves the data's quality. Data has played a significant role in both scientific research and commercial activities. The colleges are seeking for related data mining methods to analysis the campus data. Zhong RW [1] and Qian Yuan [2] applied data association rule to the analysis of student performance and does association analysis of the result of data mining to provide decision support for college administrators. Mirwais Tanai [3] applied model-based clustering algorithms on student data, and found that the students whose mothers are educated, are in the group whose grades are higher than grade C and the students whose medium instruction is other than their mother language are in group of students whose grades are less than grade B. Liming Xue [4] analyzed the data of College Forum by Fuzzy C-Means algorithm and provided reference suggestions for administration department.

This paper mainly introduces C4.5 decision tree algorithm, and uses data of some college as sample to illustrate the validity and correctness of C4.5 decision tree algorithm in forecasting student achievement variation trend.

The remainder of this paper is organized as follows: related theories of C4.5 decision tree algorithm are described in Section II. In Section III, the application of C4.5 decision tree algorithm in the variation trend of student achievement is introduced in detail. And the simulation results are provided in Section IV. At last, section V concludes the work done and future work.

II. C4.5 DESCION TREE

Decision tree is a one of the most popular and commonly used approach of classification algorithms, which is a fast and efficient in classification and prediction. First, after comparing attribute values of objects, the optimal attribute is selected as the root node, which is the main feature. Each internal node represents the test attributes. Each branch represents the outcome of the test. And each leaf node represents the class labels. One of the biggest advantages of the decision tree algorithm is that it needs little background knowledge in the learning process, as long as the training examples can be expressed by attribute that is the conclusion of the model. The decision tree is usually in two steps: The first step is to train the sample datasets which are used for learning. After that a classification model is built. The second step is to use the model identification to classify sample datasets. Plenty of different decision tree algorithms are used as the data mining technique, such as ID3, C4.5 and Random Forest [5]. Many applications are done with decision tree learning algorithms. For example: Robert Burduk [6] successfully applied decision tree model to medical, and the result of computer simulation show that it can achieve good results in diagnosing diseases of upper abdominal pain with the user of decision tree classifier. Moon, Sung Seek [7] employed a decision tree algorithm to

conduct classification analysis to find the relationship between the average numbers of cigarette use per day, and showed that the age when first started smoking cigarettes every day, education level, and psychological distress are the top important explanatory. These results can recommend that social workers need to provide more customized and individualized interventions for older adults. Lin, Chun Fu [8] developed a personalized creativity learning system that is based on the data mining technique of decision trees to provide personalized learning paths for optimizing the performance of creativity.

The algorithm of C4.5 is based on the improvement of ID3 algorithm. It uses the ratio of information gain to select the attribute and avoid the shortcomings of ID3 about choosing attribute with lots of values. The current attribute node has the maximum rate of information gain which has been calculated, and the each node of the decision tree is obtained in this way.

Suppose D is a set of D data samples. Assume class label attribute has m different values, defines m different categories of C_i (i =1, 2,…, m). The expected information needed by a given sample classified information, that is, the system's total information entropy is given by the following formula:

$$info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{1}$$

where p_i is the probability that random sample belongs to C_i, and it can be estimated by D_i / D. Attribute Information entropy of partition according to attribute A is:

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j) \tag{2}$$

Information gain of attribute A to D is:

$$gain(A) = info(D) - info_A(D) \tag{3}$$

The partition entropy of A is:

$$split\_info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} log_2(\frac{|D_j|}{|D|}) \tag{4}$$

By (3) and (4), the ratio of information gain of the attribute A is as follows:

$$gain\_ratio(A) = \frac{gain(A)}{split\_info(A)} \tag{5}$$

In the same way, the ratio of information gain is calculated for each attribute, and then we choose the attribute with the maximum rate of information gain as a division of attribute. After completing the division, the original data set was divided into several subsets of data. If the data under a same branch are already the same type, then it need not to be divided anymore, otherwise it need to repeat the method above to divided the subset of data until all the data within a subset of the data are the same type .

```
createBranch()
Detect the data under a same branch in the same
classification or reached the ratio preset.
If so
   return class label;
Else
   choose the attribute with the maximum rate of
    information gain as a division of attribute;
   data partitioning with the attribute;
   create a branch node;
      for each subset
         calling the function createBranch();
   return the branch node;
```

FIGURE I. THE PSEUDO-CODE OF THE FUNCTION OF CREATING BRANCH

The pseudo-code of the function of creating branch is shown in Fig. 1.

III. APPLICATION

A. Problem Description and Data selection

The variation trend of student's achievement in this paper means the changing of student achievement ranking. We establish prediction model by analyzing the historical data of student daily life in school. Its essence is to obtain classification rules by using C4.5 decision tree of data mining algorithm, analyzing the relationship between changes in the ranking of student achievement and student daily life.

In this paper, we will take a selected college as an example, which includes 1076 students. We create table by selecting and calculating the relevant attributes from the campus card consumer database, library database and information of student database. The meaning of attributes we selected and calculated is shown in Tab 1:

TABLE I. THE MEANING OF ATTRIBUTES

s_id	student ID
n_books	number of borrowing books from library
f_library	frequency of studying in library
f_consuming	frequency of consuming in mess
avg_s_mess	average spending in mess at a time
f_out	frequency of going out of school
avg_s_out	average spending in transportation at a time
f_printing	frequency of printing
avg_s_printing	average spending of printing at a time
tendency	variation trend of student achievement

s_id	n_books	f_library	f_consuming	avg_s_mess	f_out	avg_s_out	f_printing	avg_s_printing	tendency
1	35	36	634	3.31	0	0	80	0.04	up
2	8	41	474	3.96	1	4	5	0.01	up
3	7	16	0	0	0	0	0	0	up
4	0	1	0	0	7	1.43	2	0.98	up
5	11	63	18	4.30	0	0	0	0	down
6	0	0	0	0	0	0	0	0	up
7	0	8	602	3.83	2	2.5	5	3.04	down
8	35	54	0	0	1	1	133	0.138	down
9	0	21	605	4.19	1	4	5	1.74	up
10	7	38	47	2.77	0	0	11	1.48	up

TABLE II. A PORTION OF RAW DATA

The number of borrowing books from library and the frequency of studying in library is a reflection of status of student studying. The frequency of going out of school is a reflection of status of leaving school and the average spending at a time of going out of school reflects how far students go out of school. The other attributes are a mainly behavior of student campus life. We defined the tendency of student achievement changing that ranking of achievement moved up more than 10 is "up", and defined ranking of achievement moved down more than 10 is "down" in the same way, and the others is "no change". So we can create raw data table and a portion of the raw data are show in Tab 2.

B. Data Preprocessing and Sampling

There are many "dirty" data in raw data and we should clean them before data mining as a high accuracy prediction model requires high quality sample data.

This paper analyzes the historical data of student behavior in school to predict the variation trend of student achievement. So those samples with few records are beyond the scope of this paper. With data sampling, we select 814 samples from raw data table and calculate the distributions of all attributes and then discretize all consecutive data in new samples. For example, the number of borrowing books from library is more than 50 can be discretized into "highest". In the same way, 30-50 is "high", and 5-30 is "mid", and the number less than 5 can be discretized into "low". Distribution of each attributes and class label from discrete data is shown in Fig 2 and Fig 3.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We use new samples as the training data in C4.5 training, the confidence factor used for pruning is set to 0.25 and the minimum number of instances per leaf is set to 2. In the training, we use 10-fold cross-validation to evaluate training model. Then 814 samples are chosen respectively as unlabeled sample to classify. The decision tree model we get from experimental result is show in Fig 4, and the confusion matrix is show in Tab 3. Experimental results show that the C4.5 decision tree algorithm has a good precision rate, approximately 80.84%. We also can get the decision rules, and a part of decision rules are shown in Tab 4.

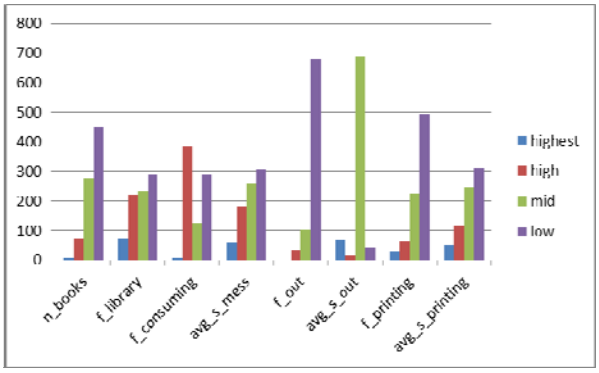

FIGURE II. DISTRIBUTIONS OF EACH ATTRIBUTES

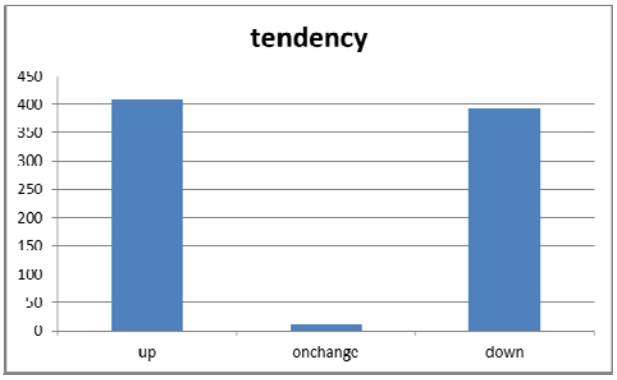

FIGURE III. DISTRIBUTIONS OF CLASS LABEL

TABLE III. CONFUSION MATRIX

	up	down	no change
up	299	94	0
down	61	349	0
no change	0	1	10

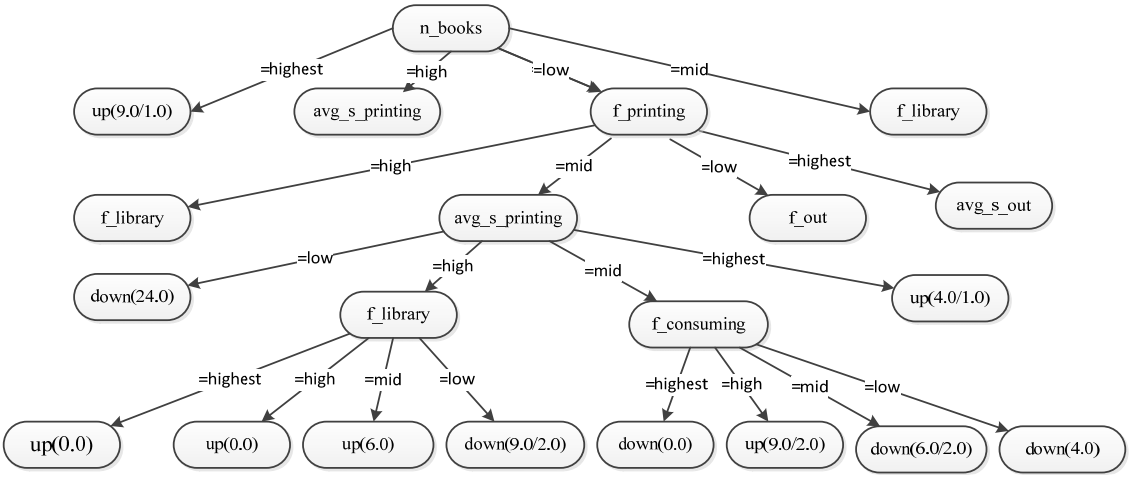

FIGURE IV. A PART OF DECISION TREE MODEL

TABLE IV. A PART OF DECISION RULES

1:if(n_books =highest) then **up**
2: if(n_books=low)and(f_printing=mid) and(avg_s_printing =highest) then **up**
3: if(n_books =low)and(f_printing =mid) and(avg_s_printing =mid) and(f_consuming =high) then **up**
4: if(n_books =low)and(f_printing =mid) and(avg_s_printing =mid) and(f_consuming =low) then **down**
5: if(n_books =low)and(f_printing =mid) and(avg_s_printing =high) and(f_library=low) then **down**

Fig 4 shows that different factors reflect the variation trend of student achievement to a different degree. The number of borrowing books from library, the frequency of studying in library, the frequency of printing and the average spending of printing at a time are the four main factors. The decision tree rules 1 in Tab 4 show that the larger number of borrowing books from library is, the greater the chance ranking of achievement moves up. As rule 5 shows, if someone does not like borrowing books from library, never goes to library and prints infrequently, but his average spending of printing at a time is high. There is a strong possibility of moving down in his ranking of achievement. Then management department should observation such students and remind them to adjust their academic lifestyles if needed. So a fundamental improvement is made in traditional management.

V. CONCLUSION

In this paper, a prediction model is established by C4.5 decision algorithm. From the decision tree rules in prediction mode, we can easily know the relationship between the variation trend of student achievement and their academic lifestyles. It is helpful to educational administrations' decision-making. However, attributes of samples used in paper are selected according to previous experience. We can attempt to use attribute-choosing algorithm instead, and dig up more reasonable attributes to establish prediction model. So, these tasks are considered as future work.

ACKNOWLEDGMENT

We would like to acknowledge the contribution of all the people to this article, and the university computing center has provided us an environment to analyze the student data.

REFERENCES

[1] Zhong, RW and Wang, HP, "Data AssociationRules in Analyzing Performance Level of College Students,"The 3rd International Conference on Computational Intelligence and Industrial Application, Wuhan China. December 2010, pp.454-458.

[2] Qian, Yuan and Shi, Quan, "Study on the Application of Data Mining based on Campus Card Platform," Advanced Materials Research, vol.846-847, September 2013, pp.977-980. doi: 10.4028/ AMR.846-847.977

[3] Mirwais Tanai, Jongwan Kim, and Joong Hyuk Chang. "Model-Based Clustering Analysis of Student Data," Convergence and Hybrid Information Technology, vol.6935, September 2011, pp. 669-676.

[4] Liming Xue, Zhihuai Li and Weixin Luan, "Data Mining Research Based on College Forum."Algorithms and Architectures for Parallel Processing, Dalian China, vol.8631, Auguest 2014, pp.525-532.

[5] Shiju Sathyadevan and Remya R. Nair, "Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest," Computational Intelligence in Data Mining, vol.1, Decmber 2014, pp. 549–562, doi: 10.1007/978-81-322-2205-7_51

[6] Robert Burduk and Michal Wozniak. "Different decision tree induction strategies for a medical decision problem," Central European Journal of Medicine, vol.7(2), April 2012, pp.183-193. doi: 10.2478/s11536-011-0142-x

[7] Moon, Sung Seek; Kang, Suk-Young; Jitpitaklert, Weerawat. "Decision tree models for characterizing smoking patterns of older adults," Expert Systems With Applications, vol.39(1), Jan 2012, pp.445-451. doi: 10.1016/j.eswa.2011.07.035

[8] Lin, Chun Fu; Yeh, Yu-chu and Hung, Yu Hsin. "Data mining for providing a personalized learning path in creativity: An application of decision trees," Computers & Education, vol.68, Oct 2013, pp.199-210. doi:10.1016/j.compedu.2013.0