ATLANTIS PRESS

Advances in Computer Science Research, volume 80

International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018)

The Application of Data Mining Techniques in College Students Information System

Yang Liu[1], Jingwei Chen[2*], Jun Liu[1], Wei He[1] and Tingting Li[3]

[1]Institute of information science and engineering, Chongqing Jiaotong University, China

[2]Chongqing Key Laboratory of Automated Reasoning and Cognition, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China

[3]School of Mathematics and Statistics, Southwest University, China

[*]Corresponding author

***Abstract*—Nowadays, with the expanding of database application, every fields have accumulated huge amounts of data including the College students' activities records. These records are very meticulously reflecting the status of the students' learning and life by analysis their relationships using data mining techniques. The traditional methods of choosing Excellent students and Outstanding Class Leader and Postgraduate Recommendation and Poor students is manual manipulation. But, in this paper, we develop a system which brings in data mining techniques with decision tree algorithm and association rules mining algorithm. Through analyzing the data from college student library records and consumption records and student score and psychological test done by the students, this information system can automatically show the results under data mining algorithm.**

Keywords—date mining; decision tree algorithm; association rules mining algorithm; database application

I. INTRODUCTION

Data mining is one of the most important areas in Database. The current situation of higher education represents the development of the country. It's a good way to manage college students with data mining technology instead of traditional mining manipulation [1-5]. At present, the campus almost all use electronic management by campus card system. This system records all the students' activities including shopping on campus and dining in university students' dining halls and the time when they are in and out of the library. Of course, it also records the student score of every course.

Presently, some of researchers have studied parts of the relationship between the data [6-8], but, they don't analyze it from different aspects of college students which are the special community. The function modules are dispersive. In other words, the date in the college management system database is existence, but, we pay no attention to the relationships during them. So, when we want to select excellent students or outstanding class leader or postgraduate recommendation or poor student, we are accustomed to manual analysis. Furthermore, because of the academic stress, parts of college students get psychological problems. Unfortunately, the advisers can't find it timely. They always know the students who has psychological problems until it's very serious. As it often happens, this leads to that the student gets mentally ill or suicide.

According to this situation, this paper provides a view to help management the students' life and study. We use decision tree algorithm and association rules mining algorithm to analyzing the huge amounts of data in the system database. By analyzing the relationship between the data, we can more and more conveniently manage the students.

A. Decision Tree Algorithm

Decision tree algorithm is a kind of methods approximating discrete function value. C4.5 algorithm [9-10] is an important kind of classification decision tree algorithm in machine learning and it's an improved algorithm of the ID3 algorithm [11-13]. In this system, we use C4.5 algorithm to analyze the scores and psychological states and consumption of the college students and build the decision tree of students' comprehensive evaluation.

The standard is based on the average score. It divides into five grades that greater than 90 and 80~90 and 70~80 and 60~70 and less than 60. At first, we compute the information entropy H(X) [14] of students' score sample based on formula(1) and the conditional entropy H(X|Y) [15] based on formula(2).

$$H(X) = -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i) \tag{1}$$

X represents the sample data set;

n represents all the possible symbolic number of X;

C_i represents the different possible value of the i kind sample;

$P(C_i)$ represents the probability of the sample data belonging to i;

The information entropy is used to compute the expectation of information. And the conditional entropy is used to compute the uncertainty for the random variable X when received the random variable Y.

$$H(X \mid Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} P(C_i, T_j)\log_2 P(C_i \mid T_j) \quad (2)$$

C_i represents the signal source from X;

T_j represents the signal source from Y;

$P(C_i \mid T_j)$ represents the probability when Y is T_j and X is C_i;

The difference of the *H(X)* and *H(X|Y)* is the information gain[15] based on formula(3).

$$Gain(X|Y)=H(X)-H(X|Y) \quad (3)$$

B. *Association Rules Mining Algorithm*

In this system, the association rules mining algorithm used by us mainly learn the Apriori algorithm[16-17] through the degree of support and the confidence. The evaluation result in this comprehensive evaluation system is given by this two standard.

The degree of support computed by formula(4).

$$Support(X)=occur(X)/count(D)=P(X) \quad (4)$$

P(X) shows the probability of X appearing in D.

The degree of confidence computed by formula(5).

$$Confidence(X->Y)=Support(X \bigcup Y)/Support(X)=P(Y|X) \quad (5)$$

P(Y|X) reflects some relations between X and Y.

II. RELATED WORK

In our view, we divide the system into four functional models. The first one is the score analysis model. And the second one is the consumption analysis model. Then the third one is the psychological status testing model. The last one is the comprehensive analysis model as Figure 1.

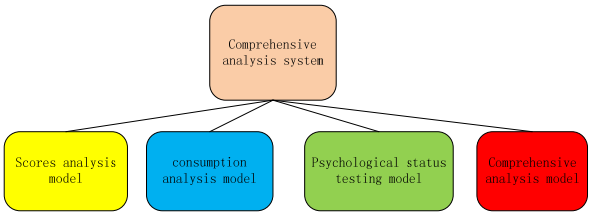

FIGURE I. THE FRAME OF THE COMPREHENSIVE ANALYSIS SYSTEM

Using decision tree algorithm, we classify the attributes of students' score and consumption and psychological status into different degrees. The different degrees of the attributes influence the results of the evaluation.

A. *Decision Tree Algorithm Applying in the System*

The first function model is the score analysis as figure 2. It has included every course score of every student in different class. We can use decision tree algorithm to get some regularities during different courses for one student. For example, if one student is good at Operating System, it also is good at Database Theory. If one student does well in fresh year and second year and third year, it also does well in the last year and Graduation Design. From the database, we can get an information entropy H(X) for each student X. At the same time, conditional entropy H(X|Y) can be computed. So, the information gain can be calculated.

The algorithm of building the decision tree is followed.

The tree begins with the single node which represents the samples.

1. If the samples have existed in the same category, then, this node is a leaf node marked with this category.

2. Or else, it will automatically generate node which choose the one occupying most of the attributes.

3. After analysis and conclusion, the information from the samples are divided into many sets. Every branch node can get the value of its sub-set. Every sub-set corresponds a branch. For every sub-set from the last step, it repeats the procedure. Then, it will produce a decision tree for every samples.

4. Once a kind of attribute appears in one node, it doesn't need to consider its descendant.

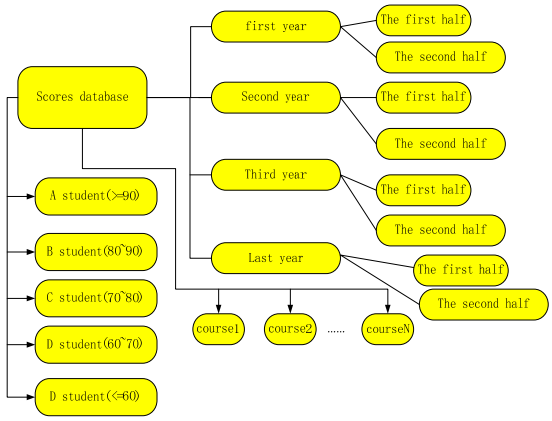

FIGURE II. THE SCORE ANALYSIS MODEL WITH DATABASE

This algorithm will stop when satisfy these conditions as follow:

1. All the samples of the nodes belong to the same category.

2. There is no left attributes used to divide. In this condition, the nodes of the tree will update after analysis and conclusion, and it can automatically generate the leaf nodes marked with the category which has the most elements.

3. If one branch doesn't have sample which satisfies this existed category, it will build a leaf node with the sample which has majority classes.

B. Association Rules Algorithm Applying in the System

Using association rules algorithm, we take support degree and confidence degree standard to support the results given by the system when we select excellent students or outstanding class leader or postgraduate recommendation or poor student. Based on Apriori[8] algorithm, our method is followed:

1. Build the initializing collections of students' scores and consumptions and psychology testing status from the database.

2. Give the association rules which we build in rule base.

3. If the confidence degree of the student meets the threshold value we need, then the student will be the candidate.

The holistic concept is as the figure 3.

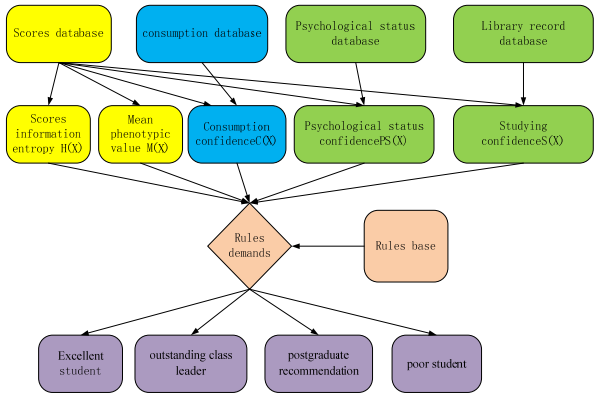

FIGURE III. THE HOLISTIC CONCEPT OF COMPREHENSIVE ANALYSIS MODEL

III. THE FUSION OF THESE TWO ALGORITHMS

Both of these two algorithms have advantages in solving specific problems. Using their respective advantage, we bring them in our system as figure 3. When we analyze students score, we use decision tree algorithm which the whole frame is as figure 4. We analyze their consumption and psychological status with association rules algorithm.

A. The Mean Phenotypic Value Analysis of Student

$$M(X)=\frac{\sum_{i=1}^{n} i.score(X)}{n} \tag{6}$$

$$M(X)=\frac{\sum_{i=1}^{n} i.score(X)}{n}\times 0.3+\frac{\sum_{j=1}^{m} j.score(X)}{m}\times 0.7 \tag{7}$$

$$M(X)=\frac{\sum_{i=1}^{n} i.score(X)}{n}\times 0.2+\frac{\sum_{j=1}^{m} j.score(X)}{m}\times 0.3+\frac{\sum_{k=1}^{t} k.score(X)}{t}\times 0.5 \tag{8}$$

$$M(X)=\frac{\sum_{i=1}^{n} i.score(X)}{n}\times 0.1+\frac{\sum_{j=1}^{m} j.score(X)}{m}\times 0.2+\frac{\sum_{k=1}^{t} k.score(X)}{t}\times 0.3+\frac{\sum_{l=1}^{s} l.score(X)}{s}\times 0.4 \tag{9}$$

Where i.score(X) represents the score of the course i of the student X. H(X) can reflect the stability of the student X. M(X) represents phenotypic value. If the student is just a freshman, then we consider mean phenotypic value M(X) as the first one in formula (6). If he is a second year student, then we take formula (7) into account. In the similar way, if he is a third year student, formula (8) will be used. Formula (9) will be taken, only when it is the last year student.

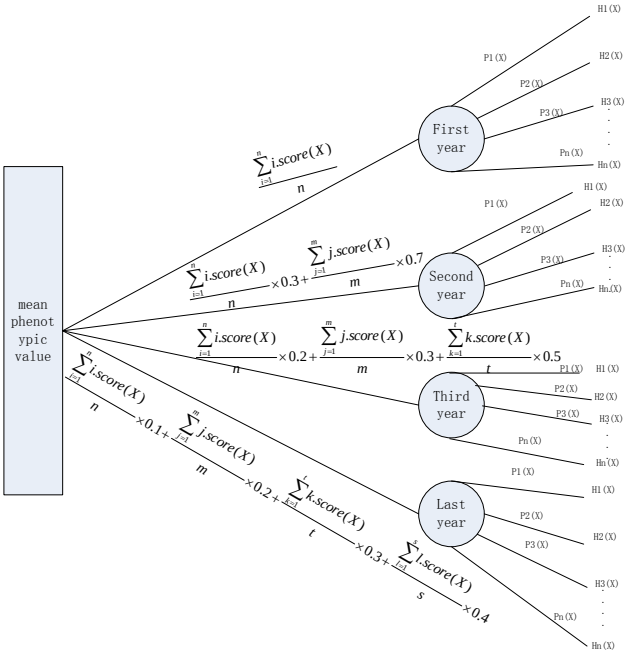

FIGURE IV. THE SKETCH FRAME DECISION TREE OF STUDENTS SCORE ANALYSIS

So, from H(X) of the student X, we can get hold of the studying status of the student X. If the H(X) becomes lower, we can use the second core model-psychological status testing model. Through testing, we can know the reason why the student gets a poor mark.

B. The Psychological Status Testing Model Analysis Methods

This model includes eights parts as figure 5. Every part has lots of testing subjects which are automatically generated by system from the question database. Of course, this question database is from authority health psychological questions updated timely [9]. Here, we choose highcharts [10] to show the results which are from students' testing. Based on the relation setting, it will appear amazing effect. Difference

weighted method brings in counting the total values as formula (10).

$$Zx = W_1Z_1 + W_2Z_2 + W_3Z_3 + W_4Z_4 + W_5Z_5 + W_6Z_6 + W_7Z_7 + W_8Z_8 \quad (10)$$

Where Wi represents the number of the subject from one of the eight parts, and *Zi* represents the proportion in the subject. Through doing the test, we can master the dynamic psychological of the student. So, if the student appears some serious psychological crazy disease like depressive disorder, we can help him in time.

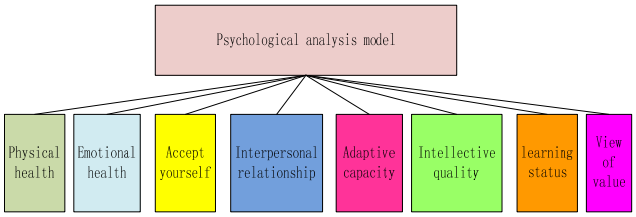

FIGURE V. THE COMPOSITION OF THE PSYCHOLOGICAL ANALYSIS MODEL

If the student's psychological status is right, we should check his record in and out of the library. The consumption records should be checked too if necessary.

C. The Comprehensive Selection System

TABLE I. PART OF EACH INDEX DATA OF THE STUDENT

Student NO.	Score information entropy H(X)	Mean phenotypic value	Consumption value	Psychological health value
060123	2.17	91	20	85
060102	4.76	65	25	74
060219	1.67	62	22	80
…	…	…	…	…
060114	1.13	47	46	57

This model includes selecting excellent students and outstanding class leader and postgraduate recommendation students and poor students.

At first, we must set up the standard value which we need and get the mean phenotypic value from the score database and consumption value and psychological health value. Then, we built the data source as table 1.

Each functional model of this part in comprehensive selection system has its own rules. So, different rules algorithms should be used.

IV. THE ALGORITHMS IN ANALYZING

When selecting excellent students, we should take two factors into refer, the mean phenotypic value and score information entropy of one student. The extra factor will be considered when we select outstanding class leader is that whether he or she is a class leader. the algorithm is as followed.

This algorithm is also suitable to selecting the postgraduate recommendation. But, when we award the poor students scholarship, the consumption value and psychological health value and library status confidence value are referenced. The algorithm is as followed next.

The Algorithm of Selecting Excellent Students

Input: $H_i(X)$ and $M_i(X)$of each student, i=1 to n
Output: the students set Z_1
1: filter the data, screening the $H_i(X)$ and $M_i(X)$ of each student in the same class;
2: sorting the $M(X)$ set of all the students from big to small ;
3: if more than one student has the same value equal $M_i(X)$
Then these students who have the same value $M_i(X)$ belong to one set $S(X)$;
Sorting $H_i(X)$ of all students in $S(X)$ from small to large;
The top of the queue belongs to Z_1.
Else
The top of the $M(X)$ queue belongs to Z_1.

The algorithm of selecting poor students

Input: $H_i(X)$, $M_i(X)$, $PS(X)$, $C(X)$ and $S(X)$
Output: the students set Z_3
1: setting a threshold value T for $M_i(X)$ like 75, filter the students based on T;
2: through the above algorithm based on $H_i(X)$ and $M_i(X)$, we get the queue Z_1';
3: sorting the data $C(X)$ of all students in Z_1', then storing in queue R1 as order from small to large;
4: setting a threshold value C for $C(X)$, filter the students based on Z_1', the students who is less than C divide into Z_2';
5: sorting the data $PS(X)$ of all students in Z_2', then storing in queue R2 as order from big to small;
6: sorting the data $S(X)$ of all students in Z_2', then storing in queue R3 as order from big to small;
7: sorting the data $H((PS(X), S(X)) \mid (C(X), M(X), H(X))$ of all students in Z_2', then storing in Z_3 as order from small to large;

V. CONCLUSIONS

In this paper, we proposed a feature selection method for high education management when choosing special students each semester. In our proposal, the data is preprocessed firstly by the decision tree algorithm and association rules algorithm. Then, based on the previous step, we extract the data for further processing. The implement show that this method is a great improvement in college student management and study life. We can master the dynamic state of each student. If his study is declined, we can get his library record and information entropy to know whether he becomes slacker, or the consumption record to know whether he is too fun, or the psychology health test to know whether he is in a bad mood during this period. So, it's greatly effectively in students' study and life by this way.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Research Project of Chongqing Education Commission (under grant KJ1705121), and the Basic and Frontier Research Projects of Chongqing (under grant cstc2016jcyjA0285 and

cstc2016jcyjA0510). Also, the present work was partially supported by NSFC11501540.

REFERENCES

[1] L. Ming-Jiang, L. Yu, Y. Tang, et al. An Applied Research on Data Mining in Management of College Students with Financial Difficulties. Journal of Qiannan Normal College for Nationalities, 2014.

[2] B. Shannaq, Y. Rafael, V. Alexandro. Student Relationship in Higher Education Using Data Mining Techniques. Global Journal of Computer Science & Technology, 2010, 10(2010):45-59.

[3] X. J. Cai, Y. R.Jiang. Application of Fuzzy Data Mining in College Students Management. Journal of South China Agricultural University, 2006.

[4] H. R. Wang. Research on the application of data mining in teaching management. Electronic Design Engineering, 2013.

[5] Y. He, S. Zhang. Application of Data Mining on Students' Quality Evaluation// Intelligent Systems and Applications (ISA), 2011 3rd International Workshop on. IEEE, 2011:1-4.

[6] N. C. Pradeep, Planning I P, Customer S. Data Mining in Management, and Practice. LAP LAMBERT Academic Publishing, 2012.

[7] N. B. Peyman, E. Vahab. Introduction to quantitative researches in management (case study: applications of data-mining in management). 2011.

[8] D. F. Cao, Y. H. Wang. Application of data mining in management information system of comprehensive evaluation of employees. Hebei Journal of Industrial Science & Technology, 2012.

[9] J. R. Quinlan. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. 2014.

[10] J. R. Quinlan. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 1996, 4(1):77-90.

[11] S. Hong, S. Ratté, S. A. Prescott, et al. Building A Smart Academic Advising System Using Association Rule Mining. Computer Science, 2014, 32(4):1413-1428.

[12] S. H. Park, S. Y. Jang, H. Kim, et al. An association rule mining-based framework for understanding lifestyle risk behaviors. Plos One, 2014, 9(2):e88859-e88859.

[13] N. Dang, L. T. Nguyen, B. Vo, et al. A novel method for constrained class association rule mining. Information Sciences, 2015, 320:107-125.

[14] J. Liang, Z. Shi, D. Li, et al. Information entropy, rough entropy and knowledge granulation in incomplete information systems. International Journal of General Systems, 2006, 35(6):641-654.

[15] Q. S. Zhang, S. Y. Jiang. A note on information entropy measures for vague sets and its applications. Information Sciences, 2008, 178(21):4184-4191.

[16] A. Inokuchi and T. Washio and H. Motoda, “An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data”, Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Springer-Verlag, (1970):13—23.

[17] H. Sakai and M. Wu, M. Nakata, “Apriori-Based Rule Generation in Incomplete Information Databases and Non-Deterministic Information Systems”, Health Management Technology, 2014, 130(3):343-376.