Study on Comprehensive Quality Evaluation of Students Based on Data Mining Technology

Song Dechang, Yan Minjuan
School of Management, Wuhan University of Technology, Wuhan, P.R.China, 430070
(E-mail: songke@whut.edu.cn,

Abstract This paper firstly analyzes the researching actuality of the student result management and points out the problems that still exist. Secondly, a method of data mining is designed for solving the problems. Through the application of this method, the anthors construct the student classification models, which are mainly tested by student's score. Finally, the anthors do a prediction for a new student and give the result of the prediction, the analysis shows that the work is of great theoretical and realistic significance to assist managers in student result management programs.
Key words Evaluation of student; Student score; Management, Fuzzy data mining

1 Introduction

Generally speaking, there are two main aspects in the student result management, one is the record of score, and the other is statistical analysis of the score. Such works immediately concerns not only students' vital interests but also the future learning and development. Only we do the scientific，effective and objective management, then it will be able to better service the development of the students, teachers and school education.

Now there are many articles introduce the research achievements of student score management. Zhang Zhiyan, Li Junfeng(2009)[1]used association rules to study the student score and dug out credible information, which can provide scientific basis for the future score management. Kang Yanxia(2009)[2] applied the data mining techniques to the analysis of student score and found the association between curriculum, the reference was provided for choosing courses and teaching management. Gan Xiaoya,Liu Kan & Liu Ping(2009) [3] used data visualization technology to college students' score in an English test, in view of the visualization findings, the paper has put forward some advice to guide future English teaching. In conclusion, these articles from different angles carried on the theory to probe into to the structure and management of student score, and it has positive guidance meaning for students' learning in school and teachers' arranging. However, the study of students' comprehensive evaluation and student management need to make a thorough study which use student score as the key indicator.

With the increasing spread of lifelong education in every large university, there is more demand for student score during school study and after graduation, such as, work of scholarship and sent graduate students is linked tight with it. Because student score is given subjectively more or less, it was an arbitrary decision if we judge a student merely by high or low marks. Therefore, in real life other ability indexes such as innovation ability, research ability and presentation needed to be referred to help us accomplish such tasks.

2 Fuzzy Data Mining

Fuzzy data mining uses student score as the main evaluation index, and assess the student scientifically and rationally, the evaluation results can direct the work of student evaluation to accomplish.

Professor L.A. Zadeh, who was a cybernetics expert of California University in America has published one initiative thesis of "Fuzzy Sets" in *Information and Control*, and It marks the birth of Fuzzy Mathematics. Because the merely data mining may cause "acute boundary" problems, so we will introduce fuzzy data mining technology which combines data mining and fuzzy logic to the work of student score management. Thorough the application of fuzzy data mining, we can make fully use of the student score while consider other index, and avoid narrow-mindedness of the student evaluation wok which merely tested by student score.

3 Modeling

3.1 Choose index elements and objects mainly by student score

The object to be classified is called sample, such as: u_1, u_2 , u_3,…,u_n， and we call U={ u_1, u_2 , u_3,…,u_n } Samples. If there are ten objects will be divided, then domain U={studen1, studen2,…,

studen10}. In order to make rational classification, the specific attribute of sample should be quantified (table1), quantified attribute is called sample index, Set each sample has four indicators, each sample can be described by four-dimensional vector, that is:

$$u_i = (u_{i1}, u_{i2}, \cdots, u_{i4}) \qquad (\text{i=1, 2,…,10}) \tag{1}$$

Table 1 Samples of Students and Quantification Table of Attributes

sample number	score of specialized course 6	score of selected course 6	innovative ability 4	writing ability 5
student1	high5	good 2	not very strong 3.5	not too bad 3
student 2	fairly good 4	passable 1.5	not weak 2.5	not too bad 3
student 3	pretty good 3	pretty good 3	not weak 2.5	passable 2.5
student 4	good 2	pretty good 3	not weak 2.5	passable 2.5
student 5	very high 6	fairly good 4	very strong 4	not weak 3.5
student 6	not too bad 1	pretty good 3	not too bad 2	not too weak 2
student 7	pretty good 3	passable 1.5	not weak 2.5	not too weak 2
student 8	not very high4.5	high 5	strong 3	passable 2.5
student 9	good 2	very high 6	weak 0	weak 1
student 10	not very hig4.5	very high 6	not very strong 3.5	not too bad 3

In real work, different types of data might be different in nature and dimension, the initial data should be processed in order to meet the need of calculation, such as the data standardization, and then the initial data is given in the interval [0,1], at last fuzzy matrix will be constructed. There are two steps of processing method:

① Translation—standard error transformation

$$u''_{ik} = \left| \frac{u'_{ik} - u'_k}{S_k} \right|, \quad (\text{i}=1,2,\ldots,10,\ \text{k}=1,\ \ldots,4) \tag{2}$$

in which:

$$u'_k = (u'_{1k} + u'_{2k} + \cdots + u'_{nk})/n = \frac{1}{n}\sum_{i=1}^{n} u'_{ik} \tag{3}$$

$$S_K = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u'_{ik} - u'_k)}\ , \quad \text{k}=1,2,\ \ldots,\text{m} \tag{4}$$

in these formulas n=10, m=4.

Average is given according to formula (3):

$$u'_k = \{3.5,\ 3.5,\ 2.6,\ 2.5\},\ \text{k}=1,\ldots,4$$

Standardized vector is given according to formula (4):

$$s_k = \{1.563,\ 1.700,\ 1.101,\ 0.707\},\ \text{k}=1,\ldots,4$$

Standardized matrix (A) is given according to formula (2):

$$A = \begin{bmatrix} 0.96 & -0.88 & 0.82 & 0.71 \\ 0.32 & -1.18 & -0.09 & 0.71 \\ -0.32 & -0.29 & -0.09 & 0.00 \\ -0.96 & -0.29 & -0.09 & 0.00 \\ 1.60 & 0.29 & 1.27 & 1.41 \\ -1.60 & -0.29 & -0.55 & -0.71 \\ -0.32 & -1.18 & -0.09 & -0.71 \\ 0.64 & 0.88 & 0.36 & 0.00 \\ -0.96 & 1.47 & -2.36 & -2.12 \\ 0.64 & 1.47 & 0.82 & 0.71 \end{bmatrix}$$

Yet if still have some $u''_{ik} \notin [0, 1]$ after translation—standard error transformation, then the translation—range transformation is also needed, that is:

② Translation—range transformation

$$u_{ik} = \frac{u''_{ik} - u''_{\min k}}{u''_{\max k} - u''_{\min k}} \tag{5}$$

Here u''_{maxk} and u''_{mink} denote respectively the maximum and minimum of $u''_{1k}, u''_{2k}, \cdots, u''_{nk}$. Standardized matrix (B) is given according to formula (5):

$$B = \begin{bmatrix} 0.80 & 0.11 & 0.88 & 0.80 \\ 0.60 & 0.00 & 0.63 & 0.80 \\ 0.40 & 0.33 & 0.63 & 0.60 \\ 0.20 & 0.33 & 0.63 & 0.60 \\ 1.00 & 0.56 & 1.00 & 1.00 \\ 0.00 & 0.33 & 0.50 & 0.40 \\ 0.40 & 0.00 & 0.63 & 0.40 \\ 0.70 & 0.78 & 0.75 & 0.60 \\ 0.20 & 1.00 & 0.00 & 0.00 \\ 0.70 & 1.00 & 0.88 & 0.80 \end{bmatrix}$$

3.2 Establish fuzzy relations R

R can be expressed as similar matrix, and its general form can be described as

$$\mathrm{R}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{in} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \quad 0 \le r_{ij} \le 1 \ \ \mathrm{i}=1,2,\ldots,\mathrm{n} \quad \mathrm{j}=1,2,\ldots,\mathrm{n} \tag{6}$$

There are many ways to calculate r_{ij}, in order to make the calculation easier that we choose Euclidean distance method in this paper. That is:

$$\begin{cases} r_{ij} = 1 - E \cdot d(u_i, u_j) \\ d(u_i, u_j) = \sqrt{\sum_{k=1}^{m} (u_{ik} - u_{jk})^2} \end{cases} \quad (\mathrm{i}, \mathrm{j} = 1,2,\ldots,\mathrm{n}) \tag{7}$$

E is determinated constant which can meet all of $r_{ij} \in [0,1]$ （i，j=1,2,…,n）. According to this formula fuzzy similar matrix (C) is given as follows:

$$C=\begin{bmatrix} 1.00 & & & & & & & & & \\ 0.80 & 1.00 & & & & & & & & \\ 0.67 & 0.74 & 1.00 & & & & & & & \\ 0.58 & 0.67 & 0.88 & 1.00 & & & & & & \\ 0.68 & 0.52 & 0.50 & 0.41 & 1.00 & & & & & \\ 0.41 & 0.52 & 0.72 & 0.82 & 0.24 & 1.00 & & & & \\ 0.63 & 0.73 & 0.77 & 0.74 & 0.36 & 0.68 & 1.00 & & & \\ 0.58 & 0.51 & 0.67 & 0.60 & 0.64 & 0.47 & 0.49 & 1.00 & & \\ 0.05 & 0.12 & 0.34 & 0.35 & 0.00 & 0.44 & 0.25 & 0.34 & 1.00 & \\ 0.47 & 0.39 & 0.53 & 0.47 & 0.65 & 0.34 & 0.32 & 0.81 & 0.24 & 1.00 \end{bmatrix}$$

3.3 Cluster analysis and the establishment of classification model

There are three methods about clustering analysis: equivalent closed method, maximum tree method, netting method. But the most common is maximum tree method, and concrete account steps are

given in paper[3]. The maximum tree is not the only one if we use different connection methods, but it has the same result if gives the λ cut-set.

Through application of maximum tree method the maximum tree is gotten as follows:

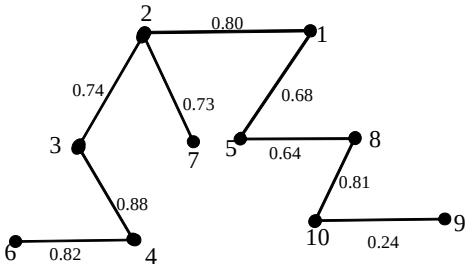

Figure 1 The Maximum Tree

If λ =0.64, it can be divided into two sections:
{student 9}, and the others belonged to the second group
If λ =0.73, it can be divided into four sections:
{student5} (typeA: students of fairly good comprehensive quality);
{student 8, student 10}(typeB: students of pretty good comprehensive quality);
{student 1, student 2, student 3, student 4, student 6, student 7}(typeC: students of good comprehensive quality);
{student9}(typeD: students of not good comprehensive quality)
If λ =0.80, it can be divided into seven sections:
{student 3，student 4，student 6},{ student 8，student 10}, the other were respectively.
If λ =0.24, all the students are bracketed together.

4 Forecast

In order to understand a student's comprehensive quality better, forecastion work should be done based on the analysis of cluster. The accomplishment of such work can provide the theoretical foundation to comprehensive evaluation of student, further more, and it can make contribution to provide scientific guidance for managers to implement the management of student score,

4.1 Calculate the average indexes of each mode

For each pattern, average index is evaluated according to formula as follows:

$$\mathrm{Mod}\, e_{ij} = \sum u_{kj} / p \quad i=1,2,\cdots,s \quad j=1,2\cdots,m \tag{8}$$

S used to represent the total number of models, k used to represent the mode derived by which few records, p represents the number of records.

Take λ =0.73 for example, calculate the average indexes of each pattern according to formula (8):

Table 2 The Average Indexes of Each Pattern When λ =0.73

investigated objects	score of specialized course	score of selected course	innovative ability	writing ability
student of type A	6.00/6(1.00)	4.00/6(0.67)	4.00/4(1.00)	3.50/5(0.70)
student of type B	4.50/6(0.75)	5.50/6(0.92)	3.25/4(0.81)	2.75/5(0.55)
student of type C	3.00/6(0.50)	2.33/6(0.39)	2.58/4(0.65)	2.50/5(0.50)
student of type D	2.00/6(0.33)	6.00/6(1.00)	0.00/4(0.00)	1.00/5(0.20)

4.2 Judge the type of prediction sample

Definition 1 if A, B∈F(U), $A \odot B = \underset{u \in U}{\vee} (A(U) \wedge B(U))$ is called as inner product of fuzzy sets A and B.

Definition 2 if A，B∈F(U), $A \otimes B = \underset{u \in U}{\wedge} (A(U) \vee B(U))$ is called as exterior product of fuzzy sets A and B.

The Calculating Formula of proximity is as follows:

$$(X, Modei) = (1/2)[X \odot Modei + (1-X \otimes Modei)] \quad (9)$$

If a student x =(high5, fairly good4, not very strong3.5, not too bad3）is given, and it was compared with above modes, proximity between it and each mode will be calculated according to formula (9) and definition of inner product and exterior product.

First, X should be standardized after we get the standardized data X= (0.63, 0.51, 0.44, 0.38),

To mode A: inner product＝0.63, exterior product=0.67, proximity＝1/2(0.63+ (1-0.67))＝0.48

To mode B: inner product＝0.63, exterior product=0.55, proximity＝1/2(0.63+ (1-0.55))＝0.54

To mode C: inner product＝0.50, exterior product=0.50, proximity＝1/2(0.50+ (1-0.50))＝0.50

To mode D: inner product＝0.51, exterior product=0.38, proximity＝1/2(0.51+ (1-0.38))＝0.565

In light of approximately principle（X, Modei)＝max((X, Mode1), (X, Mode2),…,(X, Modes)), we discovered that the student has the highest identity with mode D, and the student's comprehensive quality belongs to the classification of not good.

In short, according to classification model and prediction result we found that the student with high score might not be the most outstanding one. Specific to the work of scholarship and sent graduate students, the probability that it will choose student A is high, because this student has not only high score but also good comprehensive quality. For student D, the score is all right but comprehensive quality is not the most outstanding one.

5 Conclusions

This article used fuzzy data mining technology to mine the data of student score, and classification was given to help the managers understand the students learning better, judge the type of prediction sample could not only help teachers to forecast the study of a student but also provide scientific guidance for managers. The student score management is the most important work of university teaching management, only fully use the data of score scientificly and effectively can we succeed in doing the job of student score management. On account of a good many of subjective factors, it seems too flexible for the mining wok of student score, according to this article the process of data's accuracy and reuse of data are the future direction of working hard.

References

[1] Zhang Zhiyan, Li Junfeng. Application of Association Rule in Analysis of Students' Score[J]**.** Science Technology and Industry*,* 2009, 9(5):120-122 (In Chinese)

[2] Kang Yanxia. Data Mining Applied to the Analysis of Student Achievement[J]. China Computer & Communication, 2009, (9):81-82 (In Chinese)

[3] Gan xiaoya, Liu Kan, Liu Ping. Visual Analysis of College Students' Score in English Test[J]. Proceedings of 2009 4th International Conference Computer Science & Education, 2009:1816-1819

[4] Wang Yilei, Li Tao, Guo Yinglin. The Application of Fuzzy Data Mining in the DSS[J]. Journal of Shandong Normal University (Natural Science)**,** 2006, 21(3):43-45 (In Chinese)