Application in Student Management of Similarity

Dai Yubiao
School of Computer Science and Engineer
Qujing Normal University
Qujing,China
abiaodai@163.com

Ren Xueli
School of Computer Science and Engineer
Qujing Normal University
Qujing,China
oliveleave@126.com

Abstract—The learning achievement is the important index of students appraising and award grants, therefore, the scores of courses become the core content of the student management. A system is established to estimate scores to improve the learning effect based on a large number of students' grade data bases in the educational management system, and combined with similarity technology. The four methods to compute similarity are used to estimate scores, and the result shows that the mean absolute errors are less than 5 using these methods; the missing values are filled by listwise deletion, zero imputation, one imputation and mean imputation, and the result shows that the mean imputation is better than the other methods.

Keywords-similarity; score; missing value; student management;MAE

I. INTRODUCTION

With the continuous enrollment expansion of colleges and universities, the number of students is increasing; the quality of the students is declining, which brings a severe test to the management and teaching of college students. Every year, some students are stayed down or dropped out of school due to failing in some courses, which produces the serious influence on not only the student management of the school and but also the future life of students. The school as a training place for students, have the obligation to help the students learn the course, and avoid the final repetition and dropping out, which not only reflects the humanistic education to students, and but also raises the level of student management and teaching management. Therefore, it is necessary to establish system to estimate scores in order to give early warning, which help students understand their status for courses, render some help for these students to make them learn the next stage courses better. The score estimation are realized in the paper to guide students to choose courses and make reasonable teaching plan, which has a basis on the grade data base of the students in student management system, and combined with similarity technology.

II. THE METHOD TO COMPUTE SIMILARITY

Similarity quantifies the similarity of two objects. Although no single definition of a similarity measure exists, usually similarity measures are in some sense the inverse of distance metrics: they take on large values for similar objects and either zero or a negative value for very dissimilar objects. The common methods to compute similarity are Euclidean Distance, Cosine Similarity, Adjusted Cosine, Jaccard Coefficient and Pearson correlation [4-10].

A. Euclidean Distance

If uses rating look as the points in Euclidean space, then the distance in the points is similarity for them. If the common item set is Ij which include the items rated by user i and user j, $R_{i,c}$ and $R_{j,c}$ are the rate which are rated separately by user i and j, then the distance between user i and user j is computed used Formula 1 and the similarity between user i and user j is computed used Formula 2.

$$dist(i, j) = \sqrt{\sum_{c \in I_j} R_{ic} - R_{jc}} \quad (1)$$

$$sim(i, j) = \frac{1}{1 + dist(i, j)} \quad (2)$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j.

B. Cosine

If uses rate look as the vectors in n space, then the similarity between one user and the other user is defined as cosine between one vector and the other vector. If $\vec{x}$ and $\vec{y}$ are rating vectors by user i and user j, then the similarity between user i and user j is computed used Formula 3.

$$sim(X,Y) = \cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|x\| \cdot \|y\|} \quad (3)$$

C. Adjusted cosine

As the different user's rating scale does not considered in the cosine similarity, the modified cosine similarity is used to improve the defect by minus the average score of user rating for the project. If I_{ij} is the common item set that are rated by user I and user j, I_i and I_j are separately the rate which is rate by user I and j, then the

similarity between user i and user j is computed used Formula 4.

$$sim(i,j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_j} (R_{j,c} - \bar{R}_j)^2}} \quad (4)$$

Where $R_{i,c}$ is the rate of item c by user I; $\bar{R}_i$ and $\bar{R}_j$ are respectively the average rate for the whole items by user i and user j.

D. Jaccard Coefficient

The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Given two objects, X and Y, each with n binary attributes, the Jaccard coefficient is a useful measure of the overlap that X and Y share with their attributes. Each attribute of X and Y can either be 0 or 1. The Jaccard coefficient between object X and object Y is computed using Formula 5.

$$Jaccard(X,Y) = \frac{X \cap Y}{X \cup Y} \quad (5)$$

E. Pearson correlation

If the common item set is Iij which include the items rated by user i and user j, then $sim(i,j)$ of the Pearson correlation similarity of two users I and j is defined as Formula 6.

$$sim(i,j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_{ij}} (R_{j,c} - \bar{R}_j)^2}} \quad (6)$$

Where $R_{i,c}$ is the rate of item c by user I; $R_{j,c}$ is the rate of item c by user j. $\bar{R}_i$ and $\bar{R}_j$ are average value of rate for the whole items by both user x and user y.

F. The model of score estimation based on similarity

According to the whole course grades of students, the model to predict the scores of the follow-up courses is established based on similarity, and which provides a useful guidance for the students to select courses and learning. The processes are shown in Fig. 1:

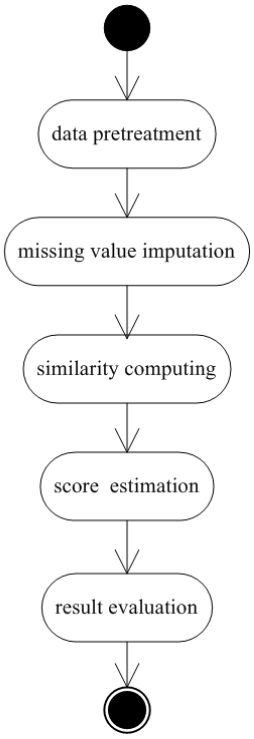

Figure 1. The process of score estimation

G. Data pretreatment

Since each metric has different value range, this first step normalizes values of metrics so that the value range becomes [0, 1]. There are non-quantitative values in the set of attributes of projects, such as Boolean, numeric, so non-quantitative values should be processed at first. The following methods are applied in non-quantitative values.

H. If the score gi is for numeric, then formula (7) is used.

$$nor(g_{ij}) = \frac{g_{ij} - \min(G_j)}{\max(G_j) - \min(G_j)} \quad (7)$$

In that, G_j is range of attribute j for the whole students, $\max(G_j)$ and $\min(G_j)$ denote the maximum and the minimum in the value domain of attribute j.

If gi is for fuzzy value, then two steps are used to normalize. Firstly, the fuzzy value is converted to number start from 1 based on the level from low to high. Secondly, the method for numeric is used.

In addition, some treatments are needed to compute easily. They are listed in the following:

The student ID will be transformed into a sequence from the beginning of S1, the curriculum will be transformed into a sequence from the beginning of C1.

I. Missing value imputation

One of the practical problems in using the estimation methods is that the historical grade data base usually contains substantial numbers of missing values. Especially, process metrics contain larger numbers of them since they are collected by hand. The missing values can give bad

influences to the accuracy of estimation, so some complementary techniques have been developed for dealing with missing values. The techniques are: listwise deletion, mean imputation and some types of hot-deck imputation [11-15]. Listwise deletion is the simplest technique to ignore data sets that have missing values. Mean imputation is a technique to fill the missing values on a variable with the mean of data sets that are not missing. Hot-deck imputation is alternative forms of imputation that are based on estimates of the missing values using other variables from the subset of the data that have no missing values. The four methods to deal with missing value are used in the paper which are listwise deletion, min imputation,max imputation and mean imputation.

- listwise deletion: these students with missing values are deleted from the grade data base directly.
- min imputation: the minimum value 0 is used to fill for all the missing values in the grade data base.
- max imputation: the maximum value 1 is used to fill for all the missing values in the grade data base.
- mean imputation: the average value is computed for not missing data, and it is used to fill for all the missing values in the grade data base.

J. *Similarity computing*

Jaccard Coefficient measures similarity only considering for Boolean, and the scores in grade data base are numerical, so in this step, the similarity $sim(s_a, s_i)$ of the target student S_a and another student S_i is computed respectively based on Euclidean Distance, Cosine, Adjusted Cosine and Pearson correlation except Jaccard Coefficient.

K. *Score estimation*

A score is estimated for the target student S_a after $sim(s_a, s_i)$ is computed. The steps are as following: Firstly, the k-nearest students are chosen based on similarity. Then the weighted sum is employed to compute score which is computed as the sum of the metrics' values given by the other students similar to S_a. Each value is weighted by the corresponding the $sim(s_a, s_i)$ between S_a and S_i. Formally, the value is defined using formula (8).

$$\hat{g}_{ab} = \frac{\sum_{i \in k-nearest} g_{ib} \times sim(s_a, s_i)}{\sum_{i \in k-nearest} sim(s_a, s_i)} \quad (8)$$

Where k-nearest students denotes set of k students chosen (called neighborhoods) that have highest similarity with S_a.

L. *Score estimation*

The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes[16], so it is used to measure forecast error in the paper. MAE is given by formula 9.

$$MAE = \frac{\sum_{i=1}^{n} |g_i - \hat{g}_i|}{n} \quad (9)$$

In that, g_i is the actual value of course i, $\hat{g}_i$ is the evaluation value of course I, and n is the number of courses evaluated.

III. EXPERIMENT

Two experiments are done to show the method feasible.

A. *Score estimation based on similarity*

As an example, some grades of students of a class in specialized in computer for 1 year are taken to show the method is feasible in score prediction. These scores are processed using formula (1); and then listwise deletion is used to process missing scores in grade table. A part of results are shown in Table 1 that scores are processed:

TABLE I. THE GRADE TABLE PRETREATED

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
S1	0.613	0.619	0.857	0.400	0.542	0.733	0.793	0.550	0.750	0.761	0.652	0.500	0.432	0.698
S2	0.806	0.476	0.000	0.350	0.627	0.500	0.586	0.400	0.583	0.775	0.261	0.522	0.135	0.209
S3	0.613	0.381	0.571	0.600	0.356	0.367	0.828	0.450	0.500	0.634	0.319	0.000	0.270	0.372
S4	0.323	0.238	0.714	0.050	1.000	0.933	0.759	1.000	0.750	0.113	0.797	0.652	0.405	0.744
S5	0.516	0.143	0.714	0.525	0.847	0.567	0.207	0.550	0.417	0.775	0.884	0.848	0.459	0.419
S6	0.968	0.238	0.286	0.675	0.695	0.367	0.931	0.250	0.750	0.775	0.725	1.000	0.378	1.000
S7	0.903	0.286	0.286	0.425	0.424	0.800	0.483	0.250	0.917	0.775	0.696	0.152	0.243	0.395
S8	0.613	0.905	0.714	0.700	0.898	0.667	0.897	0.350	0.833	0.845	0.841	0.522	0.649	0.628
S9	0.516	1.000	0.857	0.425	0.695	0.067	0.310	0.400	0.667	0.718	0.652	0.587	0.514	0.512
S10	0.774	0.476	0.714	0.850	0.593	0.700	0.966	0.650	0.833	0.831	0.681	0.543	0.432	0.395

The similarities between students are computed by Euclidean distance, cosine, modify cosine, and Pearson correlation. The 10 nearest neighbors are chosen to estimate score based on the results of similarity computing, and the scores of 14 courses for one student are predicted and shown in Fig. 2:

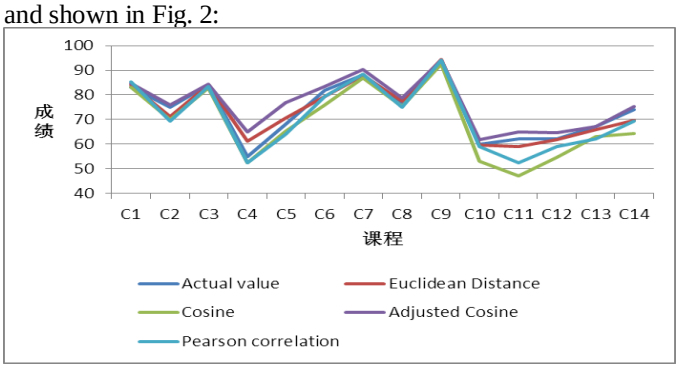

Figure 2. the result of score estimation

The MAEs are computed for every method to compute similarity using the data of Fig. 2, and the result is shown in Fig. 3.

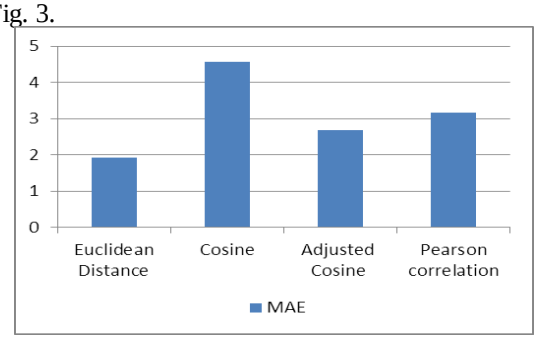

Figure 3. MAEs of the 4 similarities

M. Score estimation with missing value

In addition, four different methods to deal with missing values are evaluated with the same data in experiment 1. The missing scores are 4.6% in this grade table, the S1 as the target student , the listwise deletion, imputation 0, imputation 1 and mean imputation are used to deal with missing values, cosine is used to compute similarity, and their calculation results are shown in Fig. 4 The result shows that the mean imputation is better than the other methods.

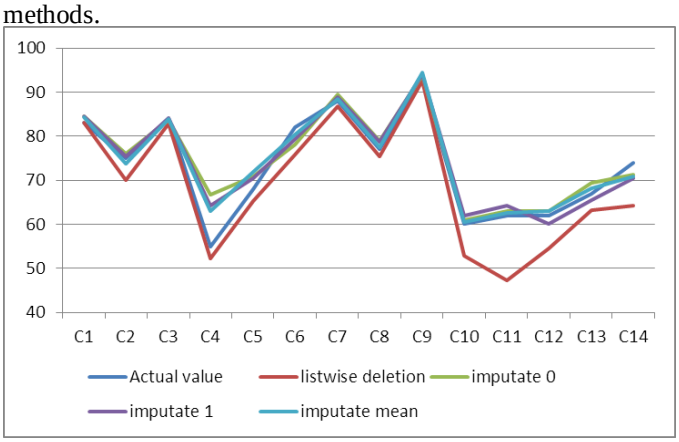

Figure 4. The prediction result with missing value

IV. CONCLUSIONS

These methods to compute similarity have been successfully applied in various fields. The score estimation are realized in the paper based on similarity, which may find the students who may not pass and give some help as soon as possible, so these improve the level of management of students in the school effectively and lay a solid foundation for improving the quality of teaching. On the basis of the data of students' achievement in educational administration management system, 4 kinds of similarity calculation are used to estimate scores; and the 4 methods are used to deal with missing values, and the results show that the score estimate based on similarity is feasible.

REFERENCES

[1] Liu Huihui. The practice and enlightenment of the total quality management of United States[J].Journal of Jiamusi College of Education.2013.10:190-192

[2] Qi Youran,Pan Zhieheng , Luo Jing.The Mathematical Model of the University Course Recommendation System[J].Acta Scientiarum Naturalium Universitatis ankaiensis.2011.8:50-52

[3] Zhou Lijuan, Xu Mingsheng, Zhang Yanyan.Model of recommended courses based on collaborative filtering[J].ApplicationR esearch ofComputers.2010.4:1315-1318

[4] Anonymous. Calculation of similarity [EB/OL]. http://wenku.baidu.com/link?url=ofsojlXw0bVKDzRl2VEwOHiCbK6GaUsP0YIBm7k-up6YvVvnzeksK3O2j_UwnOibZjlXvLwJNvJmIes9wl0yg2I9Ma6Udugsilwm7g1peue. 2014.12

[5] Brendan J. Frey; Delbert Dueck . Clustering by passing messages between data points [J]. Science 315, 2007: 972–976

[6] F. Gregory Ashby, Daniel M. Ennis. Similarity measures[EB/OL]. http://www.scholarpedia.org/article/Similarity_measures. 2015.11

[7] Guo, G.-D., Jain, A. K., Ma, W.-Y., & Zhang, H.-J.. Learning similarity measure for natural image retrieval with relevance feedback. IEEE Transactions on Neural Networks, 2002: 811-820.

[8] Ralph Bergmann. Introduction to case-based reasoning.http://www.dfki.uni-kl.de/~aabecker/Mosbach/Bergmann-CBR-Survey.pdf,2014.12

[9] Euclidean distance, https://en.wikipedia.org/wiki/Euclidean_distance, 2015.8

[10] Pearson Correlation Coefficients, http://baike.baidu.com/link?url=RfN3xBwLjuYgg6BFT2cQEKWHADgy_KUVjObXC0Kd6CcdOxTVF2hKT-scdPwjtK1UXYIYEcATIehBYsT3MP6hJa,2015.3

[11] Qin, Y.S. . Semi-parametric Optimization for Missing Data Imputation. Applied Intelligence, 2007, 27(1): 79-88.

[12] Zhang, C.Q.. An Imputation Method for Missing Values. PAKDD, LNAI 4426, 2007: 1080–1087.

[13] Zhang Shichao. Missing Value Imputation Based on Data Clustering[EB/OL]. http://link.springer.com/chapter/10.1007%2F978-3-540-79299-4_7,2015.10

[14] Anonymous. Imputation (statistics)[EB/OL]. https://en.wikipedia.org/wiki/Imputation_%28statistics%29, 2015.10

[15] Young, W. , Weckman, G. and Holland, W. A survey of methodologies for the treatment of missing values within datasets: limitations and benefits, Theoretical Issues in Ergonomics Science,2011:16- 30

[16] Mean absolute error[EB/OL]. https://en.wikipedia.org/wiki/Mean_absolute_error.2015.5