Research on Course Recommendation Based on Rough Set

Xueli Ren[1, a *]and Yubiao Dai[1, b]

[1]School of Information Engineering Qujing Normal University Qujing, China

[a]oliveleave@126.com, [b]abiaodai@163.com

Keywords: Course recommend; Similarity; Rough set; Cosine; MAE

Abstract. A credit system is the inexorable trend of higher education development. Course selection is the basis and core. Therefore, it is necessary to establish a reasonable course recommendation system. A method to recommend the course is used to guide students to choose the right course based on a large number of grades in the educational management system, and the K nearest neighbors are chosen to estimate score based on similarities between the student and the others. Reducing the attribute of datasets is one of the core contents in rough set theory. Remove the attributes that are not as important or redundant in knowledge property to improve the efficiency. The method is applied to the prediction of student grades using 3 different methods to discrete scores, the results show that the equal frequency algorithm is better than the others methods.

Introduction

With the continuous expansion of colleges and universities, higher education has been transformed from outstanding mode into the quality and the mass mode, the scale of the school continues to expand, the students have a large difference in the level and the starting point, therefore the implementation of the credit system in colleges and universities meet not only the requirements of the times, but also the needs of higher education development and law of personnel training [1-3]. It is the most important how to guide the students to choose courses which are suitable for both the profession and their own. The similarities between students are computed in the paper, and then the scores are estimated to recommend these courses. The rough set is used to improved efficiency.

Similarity Computing and Rough Set

Similarity Computing. The common methods to compute similarity are Euclidean Distance, Cosine Similarity, Adjusted Cosine and Pearson correlation [4-7].

Rough Set. Rough set first described by Polish computer scientist Zdzisław Pawlak to deal with imprecise or vague concepts. In recent years we witnessed a rapid growth of interest in rough set theory and its applications, worldwide. Here, the basic notation is introduced only from rough set approach used in the paper [8-11].

An information system is denoted as S=(U, A, V, f) where U={ U1,U2,U3,…,U |u|} denotes the set of all objects in the system, A={a1,a2,a3,…,a |A|} is the set of all attributes. C is the set of conditional attributes and D is the set of decision attributes. C and D are mutually exclusive and C∪D = A, C∩D=φ, then S is viewed as a decision table. V=∪Va where a∈A， Va is the range of the attribute a; f; U×A→V is an information function, if q∈A, x∈U, then f(x, q)∈Va is the attribute value of the object in U.

f(x, q) denotes the value of attribute q ∈ A in object x ∈ U. f(x, q) defines an equivalence relation over U. With respect to a given q, the function partitions the universe into a set of pairwise disjoint subsets of U:

$$R_q = \{x : x \in U \wedge f(x,q) = f(x_0,q) \qquad \forall x_0 \in U\} \tag{1}$$

Assume a subset of the set of attributes, $P \subset A$. Two objects x and y in U are indiscernible with respect to P if and on

$$f(x,q) = f(y,q) \qquad \forall q \in P.$$

IND(P) denotes the indiscernibility relation for all $P \in A$.$U / ind(P)$ is used to denote the partition of U given IND(P) and is calculated by formula 2.

$$U / IND(P) = \otimes\{q \in P : U / IND(q)\} \tag{2}$$

Where $A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y = \phi\}$

The lower and upper approximation of a set P $\subseteq$ U (given an equivalence relation IND(P))is defined as:

$$\underline{PY} = \bigcup\{X : X \in U / IND(P), X \subseteq Y\} \tag{3}$$

$$\overline{PY} = \bigcup\{X : X \in U / IND(P), X \cap Y = \phi\} \tag{4}$$

Rough Sets involve the approximation of traditional sets using a pair of other sets, named the Negative or Positive Region. The positive region contains all objects in U that can be classified in attributes Q using the information in attributes P. The negative region is the set of objects that cannot be classified this way.

Pawlak defines the degree of dependency of a set Q of decision attributes on a set of conditional attributes P is defined as:

$$\gamma_p(Q) = \frac{\|POS_p(Q)\|}{\|U\|} \tag{5}$$

Where $\| \ \|$ is the cardinality of a set;γ gives a measure of the contradictions in the selected subset of the dataset. If γ = 0, there is no dependence; if 0 < γ < 1, there is a partial dependence. If γ = 1, there is complete dependence.

It is now possible to define the significance of an attribute. This is done by calculating the change of dependency when removing the attribute from the set of considered conditional attributes. Given P, Q and an attribute x $\in$ P:

$$\sigma_p(Q, x) = \gamma_P(Q) - \gamma_{P-\{x\}}(Q) \tag{6}$$

The higher the change in dependency, the more significant x is.

Our Method

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.

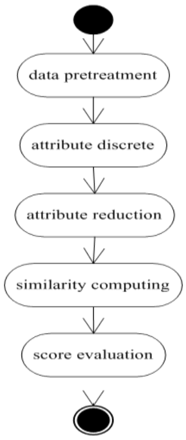

Figure 1. The process of score estimation

Data Pretreatment. Missing value in data tables should be processed firstly before computing. The techniques of missing value imputation are: listwise deletion, mean imputation and some types of hot-deck imputation [12-13]. The listwise deletion is used to deal with missing value in the paper.

Attribute Discrete. Rough set theory analytical requirements that data is in the form of categories, therefore, data must be discrete at first. Discrete results may reduce the accuracy of the raw data, but it will improve its general.Discrete in nature is that the issue of spatial conditions constitute property is divided using the selected breakpoints, dividing the n-dimensional space into a finite number of regions, so that the same decision values in each region of the object. These methods are commonly used: equal width algorithm, equal frequency algorithm, Naive Scaler methods and so on. The equal width algorithm is the simplest discretization method, which divides the numerical range into intervals according to number k by user specified, and each interval is equal to (max- min) / K. The equal frequency algorithm divides the numerical range into k intervals where the number of each interval is the same.

There are non-quantitative values in the set of attributes of grade table, such as Boolean, numeric, so the different methods are applied to discrete these values.

If the score gi is for numeric, then the two methods are used in the paper.

The equal width algorithm: Divide the grade into 4 intervals that are discrete by formula 7, the method discrete data using the same standard that don’t reflect the differences in each course. The other method is proposed in the paper which discrete grades of each course by formula 8.

$$g_i = \begin{cases} 1 & g_i \in [0,25) \\ 2 & g_i \in [25,50) \\ 3 & g_i \in [50,75) \\ 4 & g_i \in [75,100] \end{cases} \tag{7}$$

$$g_i = \begin{cases} 1 \quad g_i \in \left[\min, \left\lfloor \min + \frac{\max - \min}{4} \right\rfloor\right) \\ 2 \quad g_i \in \left[\left\lceil \min + \frac{\max - \min}{4} \right\rceil, \left\lfloor \min + 2 \times \frac{\max - \min}{4} \right\rfloor\right) \\ 3 \quad g_i \in \left[\left\lceil \min + 2 \times \frac{\max - \min}{4} \right\rceil, \left\lfloor \min + 3 \times \frac{\max - \min}{4} \right\rfloor\right) \\ 4 \quad g_i \in \left[\left\lceil \min + 3 \times \frac{\max - \min}{4} \right\rceil, \max\right] \end{cases} \tag{8}$$

Where max is the highest score and min is the lowest score.

The equal frequency algorithm: two steps are used to discrete. Firstly, the grades of each course are sorted from small to large; then discrete grades to 4 intervals.

If gi is for fuzzy value, then the fuzzy value is converted to number start from 1 based on the level from low to high.

Attribute Reduction. Reducing the attribute of datasets is one of the core contents in rough set theory. Remove the attributes that are not as important or redundant in knowledge property. The process is realized by algorithm in the paper [10].

Similarity Computing. The similarity $sim(s_a, s_i)$ is computed by cosine.

Score Estimation. Firstly, the K nearest students are chosen based on similarity;Then the score is estimated by the method in [3].

Experiment

An experiment is done to show the method feasible.

Score Estimation Based on Similarity. As an example, some grades of students in a class in specialized in computer for 1 year are taken. This decision table is constructed where courses are columns and students are rows; and then missing scores in grade table are processed. The scores are discrete separately based on the two equal width algorithms and the equal frequency algorithm in the previous paper. The attribute reduction sets are {C2, C4, C5, C6, C10, C11, C12, C13}, {C2, C4, C5, C6, C12}and {C1, C4, C5, C6, C8}.The similarity is computed by cosine, then the 10 nearest neighborhoods are chosen based on similarity to estimate score, and the result is shown in Fig. 2.

Result Evaluation. The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [14].The MAEs of the 4 methods are shown in Fig. 3., and it shows the equal frequency algorithm is better than the others methods.

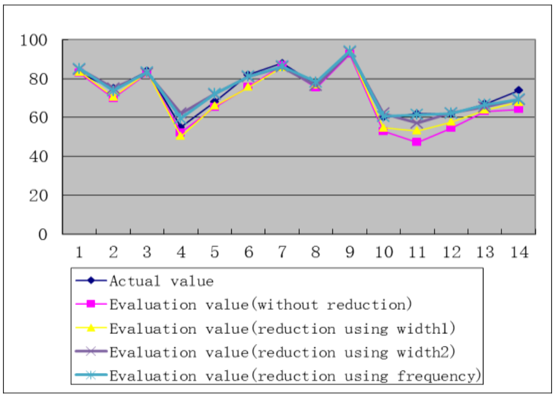

Figure 2. The results of scores estimated

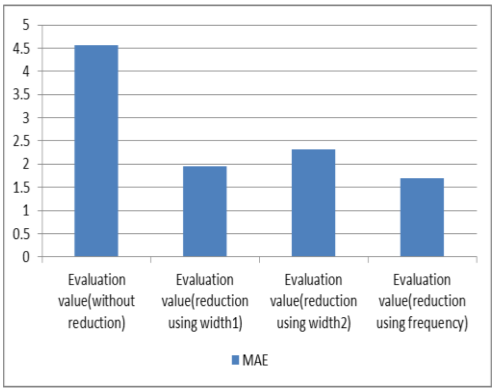

Figure 3. The MAEs of 4 methods

Conclusion

On the basis of the data of students' achievement in educational administration management system, the nearest neighborhoods are chosen based on similarity to estimate score. Rough set is used to redact attributes to improve the efficiency of score estimation, 3 kinds of methods to discrete attribute are used, and the results show the method with reduction is better than the method without reduction, the equal frequency algorithm is better than the others methods.

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] Ren xueli ,Dai yubiao.Course Selection of Students Based on Collaborative Filtering[C].emcs2016

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

[5] Calculation of similarity [EB/OL]. http://blog.csdn.net/wangzhiqing3/article/details/8293286. 2016.2

[6] Euclidean distance[EB/OL], http://blog.csdn.net/shiwei408/article/details/7602324，2015.8

[7] Pearson Correlation Coefficients[EB/OL], http://www.zhihu.com/question/21824291,2015.3

[8] Pawlak Z. Rough Set Theory and Its Application to Dat a Analysis[J] .Cyberneti cs an d Sy stems , 1998 , 9(5):661-668

[9] DING Jian-jie. Research of Software Project Risk Management Based on Rough Set Theory. Computer Science, 2010: 117-118.

[10] Rough set. http://wenku.baidu.com/link?url=rwZI1qa-HCmENenTWsjAhLjKu576XyMUH95qqQpGpnTSS5TDitB_L724vpFZ5eLHbyyK3QrkTSUIgr2o0-uVEXDruHSY5F1S5EJ-B4TiNCW,2015.12

[11] Ding Hao,Ding Shi-fei, Hu Lihua.Research Progress of At tribute Reduction Based on Rough Sets[J].COMPUTER ENGINEERING & SCIENCE.2010:93-94

[12] 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

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

[14] Mean absolute error [EB/OL]. http://baike.baidu.com/link?url=RJ14cUBYJRXo-s800M4EThtN6VEXN5g7s2PkygzTREojHKKeZ_D9SQBGn5mLatbS.2016.3