The Application in Score Evaluation of Rough Set

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

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

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

Keywords: Analogy; Score; Rough set; Attribute reduction; MAE

Abstract. A credit system is the need of higher education development. The grade management is the basis and core of the credit system. Therefore, it is necessary to estimate score as soon as possible. Estimation by analogy is a common method to estimate effort; it is used to estimate scores in the paper. As the courses are not as important in grade table, or even some courses are redundant to estimate score. Reducing the attribute of datasets is one of the core contents in rough set theory. It is applied in reducing the redundant course to improve the accuracy of estimation. Firstly, the equal width method is used to discrete data; then the redundant attributes are reduced by rough set; finally the reduce set is used to estimate score. An experiment is done to show the method feasible. The two methods of without reduction and with reduction are used to estimate scores; the results show that the mean absolution error of the method with reduction is smaller.

Introduction

With the continuous enrollment expansion of colleges and universities, the number of students is increasing; the quality of students is declining, which brings a severe test to the management and teaching of college students [1, 2]. Scores are not only the results of the examination learn the courses after a period of time, but also the basis for the next stage to learn smoothly, so it is necessary to estimate scores ahead of time which may teach students to choose the correct courses.

The common methods to estimate effort fall into three main categories: expert judgment, algorithmic estimation and case-based reasoning. Expert judgment relies on the experience and the judgment of experts, so it is subjective and unrepeatable, accuracy of expert based prediction is erratic. Algorithmic estimation involves the application of mathematical models, which is mainly a data-driven method, so it is objective and repeatable; accuracy of result estimated depends on the model constructed. The idea of analogy-based estimation is to determine the factor value of the target project as a function of the known values from similar historical projects. Compared with the other two categories of estimation methods, It can be not only applied in the very early phase of a project when detailed information about the projects are not yet available, and can be but also later improved when more detailed information is accessible. Analogy based estimation has the potential to mitigate the effect of outliers in a historical data set, since estimation by analogy does not rely on calibrating a single model to suit all projects. As score estimation is similarity to effort estimation, these methods to estimate effort may be used to estimate score, the case-based reasoning is used to estimate score in the paper.

Estimation by Analogy and Rough Set

Estimation by Analogy. Estimation by analogy can be seen as a form of Case-based Reasoning (CBR), where the cases with the form of <Problem, Solution> are organized as a case base using <Attribute, Value> pairs for the problems and the solutions. Given a new problem, solutions are adapted from similar cases (analogies) retrieved from the case base using appropriate similarity measures [3].

The similarity of two objects is computed in estimation by analogy which takes 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 and Pearson correlation [3-8].

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 [9-13].

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 $C \cup D = A$, $C \cap D = \varphi$, then S is viewed as a decision table. $V = \cup Va$ where $a \in A$, Va is the range of the attribute a; $f: U \times A \rightarrow V$ is an information function, if $q \in A$, $x \in U$, then $f(x, q) \in Va$ is the attribute value of the object in U.

f (x, q) denotes the value of attribute $q \in A$ in object $x \in U$. 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:

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

For instance, drawing from Table 1:

$R_a = \{\{1,2,6\}, \{3,4\}, \{5,7,8\}\}$

$R_b = \{\{1,2\}, \{3,4,6,7\}, \{5,8\}\}$

$R_c = \{\{1,2,5\}, \{3,4,7\}, \{6,8\}\}$

$R_d = \{\{1,2,5\}, \{3,4,6,7,8\}\}$

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 only if

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

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 $A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y = \phi\}$ For instance, if P = {a, b}, objects 1 and 2 are indiscernible; 3 and 4 likewise; 5 and 8 likewise. The rest of the objects are not. This applies to Table 1 as follows:

$U / IND(P) = U / IND(a) \otimes U / IND(b) = R_a \otimes R_b = \{\{1,2\}, \{3,4\}, \{7\}, \{5,8\}\}$ The lower and upper approximation of a set $P \subseteq U$ (given an equivalence relation IND (P) is defined as:

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

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

Rough Sets involve the approximation of traditional sets using a pair of other sets, named the Negative or Positive Region (or lower/upper approximation of the set in question. 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. The complement of 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 ∈ 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.

Table 1 A dataset

object	a	b	c	d
x1	1	1	1	1
x2	1	1	1	1
x3	2	2	3	2
x4	2	2	3	2
x5	3	3	1	1
x6	1	2	2	2
x7	3	2	3	2
x8	3	3	2	2

Our Approach

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.

Decision Table Construction. A table is constructed including conditional attribute set and decision attribute set where attribute set are columns and objects are rows.

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 g_i is for numeric, then the equal width method is used in the paper. 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.

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

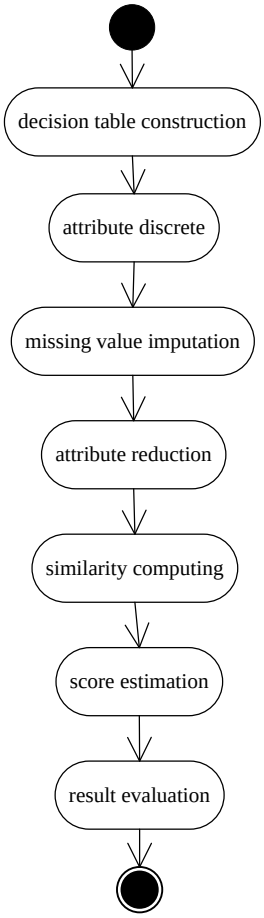

Figure 1. The process of score estimation

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

Missing Value Imputation. 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 [14-16]. The listwise deletion is used to deal with missing value in the paper.

Attribute Reduction. Reducing the attribute of datasets is one of the core contents in rough set theory. As the attributes are not as important in knowledge property, or even some attributes are redundant, attribute reduction removes redundant conditional attributes from nominal datasets, all the while making sure that no information is lost. The process is realized by the following algorithm.

Input: C={a1,a2,……}, the set of all conditional attributes; D={d}, the set of decision attributes.

Output: R, the attribute reduct, REDU

1:count the Positive Region D of the attribute set C:POSC(D);

2:For the attribute ai∈C, count the Positive Region D of the attribute set C exception ai C\{ai}: POSc\{ai}(D);

3:if POSc\{ai}(D) =POSC(D), it shows that the attribute ai is unnecessary for the decision attribute d, then C= C\{ai}, jump 2,or to 4.

4: output attribute reduction REDU=C.

Similarity Computing. The similarity $sim(s_a, s_i)$ of the target student S_a and another student S_i is computed by cosine.

Score Estimation. The k nearest students is chosen based on similarity, then, the score is computed by formula 8.

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

Result Evaluation. The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [17], 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.

Example

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. This grade table is denoted where courses are look as columns and students are rows; and then listwise deletion is used to process missing scores in grade table. The scores are discrete based on the equal width algorithm and the equal frequency algorithm. A part of results are shown in Table 1, that scores are discrete by formula 7.

Table 2 The results of discrete

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
1	4	4	4	3	3	4	4	4	4	3	3	3	3	3
2	4	3	4	3	3	4	4	3	4	3	2	3	3	3
3	4	3	4	3	3	3	4	4	4	3	2	2	3	3
4	3	3	4	2	4	4	4	4	4	1	3	3	3	4
5	4	3	4	3	4	4	3	4	4	3	4	4	3	3
6	4	3	4	3	4	3	4	3	4	3	3	4	3	4
7	4	3	4	3	3	4	4	3	4	3	3	2	3	3
8	4	4	4	3	4	4	4	3	4	3	4	3	4	3
9	4	4	4	3	4	3	3	3	4	3	3	3	3	3
10	4	3	4	3	3	4	4	4	4	3	3	3	3	3

The set of {C2, C4, C6, C7, C8, C10, C11, C12, C13} is formed after reduction. The similarity between students is computed by cosine based on the reduction set. 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 by the method without reduction and the method with reduction, and the results are shown in Fig. 2:

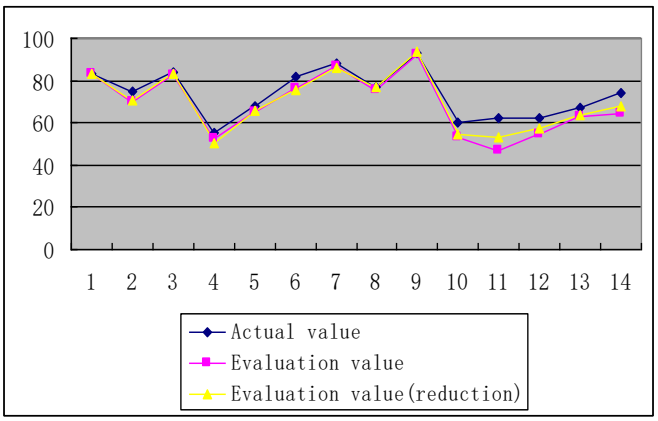

Figure 2. The result of score estimation

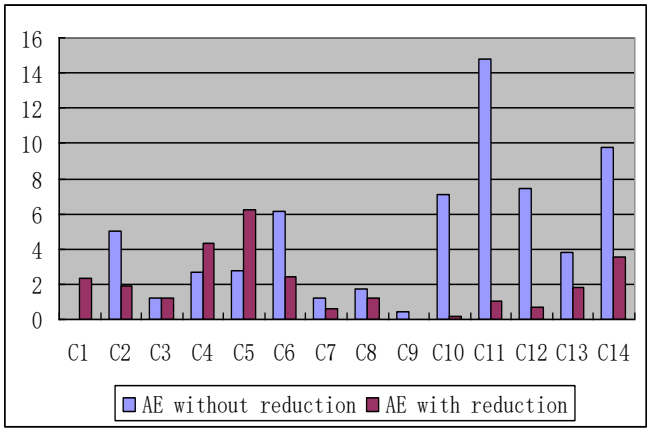

Figure 3. The absolution errors of every course

The MAEs are computed to compare the two methods, that are MAE without reduction is 4.572104, and MAE without reduction is 1.959013; and the result shows the attribute reduction improve the accuracy of estimation.

Summary

These methods based on analogy have been successfully applied in various fields. The score estimation are realized in the paper, 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, rough set is used to redact attributes to improve the accuracy of score estimation, cosine similarity calculation is used to estimate scores, and the results show that the score estimate based on rough set is feasible.

References

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

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

[3] Similarity measures on http://www.scholarpedia.org/article/Similarity_measures. 2015.11

[4] 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.

[5] Introduction to case-based reasoning on http://www.dfki.unikl.de/~aabecker/Mosbach/Bergmann-CBR-Survey.pdf,2014.12

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

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

[8] Pawlak. Z. Rough Sets: Theoretical Aspects of Reasonsing about Data. Boston, London, Kordrecht: Kluwer Academic Publishers, 1991

[9] Alexios Chouchoulas. A Rough Set Approach to Text Classification [EB/OL]. http://link.springer.com/chapter/10.1007%2F978-3-540-48061-7_16.2016.3

[10] A Review of Rough Set Models on http://wenku.baidu.com/link?url=UMxYXkHEH6KK_UlFdfFE2jf86_GpTpjUqnTyVUe44IP8Vqx16OK2F4suAiwnktxUkAfLj2VdvlL7PctJaaJGmzAUfpIslOiLIMZNHlfzrNm.2016.3

[11] Rough Sets on http://wenku.baidu.com/view/8cd094270722192e4536f6cd.html.2016.3

[12] The method and its application of Rough set on http://wenku.baidu.com/link?url=-7uUdm10FvZrEBee2Xjputj9HsnezEYwf_Ss9nUKSJQ0F1k2jAeL1XQqaLW-JS7ipWn16iUC3QZU-0SE84QQzyOyt44yg81DAvXSJxKzM7e.2016.3

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

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

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

[16] 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

[17] Mean absolute error on https://en.wikipedia.org/wiki/Mean_absolute_error.2015.5