2010 Fifth International Conference on Internet Computing for Science and Engineering

A Null Value Estimation Method Based on Similarity Predictions in Rough Sets

Jing Yang, Ze Jiang, Jianpei Zhang, Lejun Zhang
College of Computer Science and Technology
Harbin Engineering University
Harbin, China
yangjing@hrbeu.edu.cn, jiang_ze_789@hotmail.com

Abstract—**In this paper, we utilize rough set theory as a tool to deal with the problem of null-value estimation in an incomplete information system, a rating mechanism in collaborative filtering technology is introduced into this paper for the weakness of null value estimation based on similar relational algorithm (SIM-EM), such as no sparse degree process and low accuracy, and an improved null value estimation method, which based on SIM-EM is proposed from the perspective of similarity. The null value data is predicted and filled through the similarity of objects; otherwise a dual feature weight method is proposed according to the attribute's feature in rough set, which improves the accuracy in similarity calculation; the improved algorithm is good at dealing with sparse rough set, and the accuracy and the mean absolute error is better than the original method.**

Keywords-rough set; incomplete information system; null-value; similarity; weight

I. INTRODUCTION

Rough set theory was first put forward as a data analysis theory by Polish scientists Paw-lak[1] in 1982, which is used to reflect the capability of dealing with undistinguished phenomena by incomplete information or knowledge. Presently, machine learning, soft computing, decision analysis, inductive reasoning, expert systems, pattern recognition, finance has been applied in many areas. Classical rough set theory is based on incomplete information systems theory. However, in reality, the information system is usually an incomplete information system, thus, there is some uncertain information in the system which is shown as a non-value data table (it contains one or more non-value). Four uncertain factors are analyzed in literature [2]: discrete treatment, non-precise data, missing values, multiple descriptors. Missing value is null value which indicates the corresponding attribute value unknown or unavailable. The current null value estimation methods, including special attribute value method, method on statistical analysis, Roustida[3], model on the Bayesian forecasting, method on rough set theory.

Null value estimation method based on similar relation model is given by literature [4] (short for SIM-EM), which gives full consideration to the compatibility of data and the depend relationship of attributes. However, there are also some shortcomings in SIM-EM: ① there is no compatible class in some objects, voting strategies and other ways have to be relied on to complete the null value estimation; ② Voting strategy will fail when objects are not unique in some compatible object, it is actually mode problem; ③ It is not suitable for dealing with sparse data information system. ④ the impact of attribute value in the predicted object is not taken into account. The direction-area of attributes and simple-majority-rule ratio of objects are defined by literature [5], which corrects the problem of no compatible class and mode in the SIM-EM algorithm. Based on the correctional SIM-EM algorithm, we make further improve from the view of similarity in this paper, namely the null attribute values in a object are predicted by the similarity between objects with similarity relation, which instead of voting strategy. It not only overcomes the deficiencies of SIM-EM, but also is good at the accuracy of predicting result

II. RELATED CONCEPTS

Definition 1: Incomplete information system. Given an information table $I =< U, A, V ,f >$, where, $U = \{e_1, ..., e_n\}$ is a non-empty finite set of objects; A is a non-empty finite set of attributes; $V = \bigcup_{a_i \in A} V_{a_i}$ is a value set of attributes; for every $a \in A$, there is a mapping a, $f(x,a) \in V$,where V_a is called the value set of a, and at least one attribute of object is null (shown as "*") in the system, denotes $f(x,a) = *$ [6].

Definition 2: Sparse degree. Given an information table $I =< U, A, V ,f >$, it means the ratio of the unknown attributes to all attributes in a rough set of data tables, the formula is:

$$\tau = 1 - \frac{card(not\ null\ values)}{card(all\ values)} \quad (1)$$

Definition 3: Estimation value. The calculation formula proposed in literature [5] is as following:

$$v^{'}(x,a) = \frac{v(b(a))}{card(x \mid x \in b(a))} \times \gamma(x) \quad (2)$$

978-0-7695-4339-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICSE.2010.22

IEEE computer society

III. IMPROVED RATING PREDICTION ALGORITHM

A. Predicting value calculation

Given a known attribute set of object x which is represented by A_x , then A_{ij} represents:

$$A_{ij} = A_i \bigcup A_j \quad (i \neq j)$$

According to the obtained set of attribute A_{ij} , three methods of similarity measure are introduced from literature [7] to calculate the similarity between object i and object j , the formulas are as following

Modified vector related similarity formula: different objects-scale of the attribute value problems is not taken into account in the cosine similarity measurement method, the modified cosine similarity measurement method improved the above defect by subtracting the average attribute value, the $sim(i, j)$ of object i and object j is shown as the following formula:

$$sim(i,j) = \frac{\sum_{a\in A_{ij}} (R_{i,a}-\overline{R_i})\cdot(R_{j,a}-\overline{R_j})}{\sqrt{\sum_{a\in A_i} (R_{i,a}-\overline{R_i})^2}\sqrt{\sum_{a\in A_j} (R_{j,a}-\overline{R_j})^2}} \qquad (3)$$

This method can not only solve the problem of lesser common known attribute value effectively, but also effectively solve the problem that all unknown attribute values are the same in cosine similarity measurement method and modify cosine similarity measurement method. It makes the calculated attribute values more accurate, thus effectively improves the quality of null value estimation. In this paper, the three basic steps (similarity calculation, setting neighbor sets, generate prediction value) in literature [7] are used to estimate the unknown attribute values of object i in the attribute set A_{ij} . The object i in attribute set A_{ij} which attribute value is unknown is represented by N_i , namely:

$$N_i = U_{ij} - U_i$$

For any attribute $a \in N_i$, the next steps are used to estimate the attribute predicting value a on object i :

Step 1: To calculate similarity between attribute a and others attribute;

Step 2: The neighbor set of attribute a is consisted by some objects with the highest similarity, that is, search out a set of attribute $M_p = \{I_1, I_2, ..., I_v\}$ in neighbor set of attribute a , and $sim(a, I_1)$ is the highest similarity between attribute I_1 and attribute a , $sim(a, I_2)$ is the second highest one and so on.

Step 3: After M_p is obtained, prediction value of attribute a on object i , shown as formula:

$$f(i,a) = \overline{R_a} + \frac{\sum_{I\in M_a} sim(a,I)\cdot(R_{i,I} - \overline{R_a})}{\sum_{I\in M_a} (|sim(a,I)|)} \qquad (4)$$

$\overline{R_a}$ Is the average value of attribute a , $R_{i,j}$ is the attribute value of I on object i .

After filling up attribute value of A_{ij} , the following formula is used to calculate the similarity between objects. In the same manner, the rating prediction value of attribute a on object i is:

$$\Pr e_{i,p} = \overline{R_p} + \frac{\sum_{j\in M_p} sim(i,j)\cdot(R_{j,p} - \overline{R_p})}{\sum_{j\in M_p} (|sim(i,j)|)} \qquad (5)$$

B. Feature weighting of prediction value

Null value prediction is a set of predicted objects which is similar with target object, null values estimation Based on similarity predictions is a kind of method like this. Thus the set of predicted objects can be feature of target object. Now, information theory is widely applied in lots of aspects, such as science, project, business, etc. In the process of feature selection and category learning, method related on information theory is also used, like information gain, mutual information. Information theory is a measure of overall, it can grasp the properties' relevance globally and importance degree of feature associated with category [8]. Using feature weight method to process these features can effectively improve the accuracy of prediction value, which means enhance the positive impact of "good object" on the predicting results, meanwhile reduce the negative impact of "bad object" on the predicting results. Since information table is a two-dimensional data table with vertical attribute, horizontal object, so we can consider two aspects about the feature of information table:

On the one hand, considering weight between the attributes In the vertical, if the value of all the objects' feature are dispersive in the entire range, the difference attribute value is obvious, and the predicted object will preference for the attribute. Entropy is usually used to measure the uncertainty of random variable in rough set, namely, entropy can be used as a measure of different attributes' value. Entropy is used to measure the weight between attributes [9]. Entropy value is calculated as follows:

$$\omega_a = \frac{H_a}{H_{a,\max}} \quad where\ H_a = -\sum_V p_{v,a} \bullet \log_2 p_{v,a} \qquad (6)$$

Where H_a represents entropy of attribute a , $p_{v,a}$ represents the probability that attribute value v appears in attribute a , $H_{a,\max}$ represents the biggest entropy of attribute a while probability distribution of attribute value is the same. We can tell that larger ω_a is, more important the attribute a. Given a sparse degree information table, weight based on entropy is more obvious if the number of attribute whose probability of values distributes disperse.

Theorem 1: When the probability of a attribute value distributes more disperse, the ω_a is bigger, namely objects are preferred to the attribute. Otherwise, it distributes more concentrated, the smaller objects prefer to the attribute.

Proof: Given information system S, universe of discourse $U = \{e_1, ..., e_n\}$, set of attribute $A = (a_1, ..., a_m)$, according to simple majority rule, the probability distribution about attribute value v_a of attribute a is divided into two part, where x represents the number of different probability distribution, and y represents the number of same, well then

$$\lim_{x\to n} H_a = -\sum \frac{1}{n} \bullet \log_2 \frac{1}{n} = \log_2 n$$

And because $H_{a,\max} = \log_2 n$

Thus we have $\omega_a = H_a / H_{a,\max} = 1$

Similarly we can obtain that

$$\lim_{y\to n} H_a = -\sum 1 \bullet \log_2 1 = 0 \text{, then } \omega_a = 0$$

When attribute values are no equal, the probability distribution is the most dispersed, ω_a is the biggest, objects are preferred to this attribute most at this time, similarly when attribute values are equal, that is, when y is close to n, then ω_a close to 0, the objects are preferred to the attribute smallest.

On the other hand, due to entropy weight is to consider the features of its own attribute, and do not relate to the relationship between the object and the object, so if the prediction where object j on the target object i is very important, object j can be given a higher weight, thus improving the quality of prediction. Based on the idea, using mutual information method to measure the relationship between different objects, and as feature weight [10], the following formula:

$$\omega_{i,j} = I(V_i; V_j) \tag{7}$$

$$I(V_i;V_j) = H(V_i) + H(V_j) - H(V_i, V_j) \tag{8}$$

Where V_i and V_j are respectively attribute value of object i and j, $H(V_i, V_j)$ is joint entropy of there two object. As the object's attribute values are not all exist, so calculation is carried out on the attribute which attribute values exist in both two objects.

Based on these two aspects, this paper proposes a double feature weight method. From the "horizontal" and the "vertical" two considerations of the data sets respectively, it is necessary to consider feature weight of attribute own, but also consider the correlation between objects which makes up the failure of entropy weight method when the difference of attribute values is small. Thus, the similarity formula (5) is improved as following:

$$sim(i,j) = \frac{\sum_{a\in Aij} \omega_{i,j} \cdot (\omega_a \cdot R_{i,a} - \overline{R_i})(\omega_a \cdot R_{j,a} - \overline{R_j})}{\sqrt{\sum_{a\in Ai} (\omega_a \cdot R_{i,a} - \overline{Ri})^2} \sqrt{\sum_{a\in Aj} (\omega_a \cdot R_{j,a} - \overline{R_j})^2}} \tag{9}$$

Where $\omega_{i,j}$ represents mutual information weight between object i and object j, ω_a represents the entropy weight of attribute a.

C. Null values estimation Based on similarity predictions in Rough Sets

Prediction-EM Algorithm (Null values estimation Based on similarity predictions in Rough Sets)

```
Input：An incomplete information table
Output：A complete information table
1) Input an incomplete information table $S =< U, A, V, f >$;
2) Calculate sparse degree $\tau$ of informaiton table;
3) If $0 < \tau \le k$, then according to formula (2) to estimate the unknown attribute value $\Pr e_{i,p}$ of object $i$, go to (8), else $k < \tau \le 1$ go to 4);
4) Using formula 4 ot calculate null attribute value where $i$, $j \in A \wedge j \ne i$ in $A_{ij}$, according to formula 6 to get entropy weight $\omega_a$;
5) According to formula 7 to get all mutual informatiom weight $\omega_{i,j}$ where $j \in A$ and $sim(i,j)$ relate object $i$ to object $j$, where $i \ne j$;
6) Construct a set of $M_p = \{I_1, I_2, ..., I_v\}$ which is sorted from big to small by the value of $sim(i,j)$;
7) According to formula 9 to predict $\Pr e_{i,p}$ for null value of object $I$;
8) If $\Pr e_{i,p} \ne \varnothing$, then $a(i) \leftarrow \Pr e_{i,p}$, else $a(i) \leftarrow *$;
9) If an uncertain attribute is exist in object $i$, in which $p' \in N_i$, go to 7); else go to 10);
10)If an incomplete object is exist in information table $S$, go to 2); else go to 11);
11)Output a complete information table，End.
```

In the algorithm, the *k* which describes the threshold value is sparse or not, if we see the information table as a sparse matrix, the *k* is sparse factor of the sparse matrix, according to the actual information system requirements to set the size of the *k*.

Step 2) - 7) is the process of null value calculation, in which the sparse degree is big enough in Step 3), the concept of director-area is used to forecast the null value by formula 2), When a information table is sparse, entropy weight between attributes and mutual information between objects are used to predict the null values in Step 4) - 7). According to prediction value from Step 3 or Step 4) - 7), the corresponding null value is filled in Step 8), and then loop to the next prediction for null value. When null value does not exist in an object, the next incomplete object is going to be filled.

IV. INSTANCE AND DATA ANALYSIS

A. Instance Analysis

Incomplete information Table 1 from literature [11] is used in this paper. This table includes a attribute set (a_1, a_2, a_3, a_4), the range is (1,2,3,4), k is a decimal between 0 and 1, the accuracy rating of predicting result for information table show as a curve with peak at different k, thus choose three values around the peak as instances in this paper, they are $k = 0.4$ 、 $k = 0.5$ 、 $k = 0.6$. In the condition of these three k values, comparing the filling result of proposed algorithm predictoin-EM algorithm with SIM-EM algorithm, at the same time, getting a suitable threshhole for Prediction-EM algorithm. For information table each time to choose a non-null value instead of empty value, respectively using SIM-EM algorithm and Prediction-EM algorithm to estimate the selected data. In order to verify the valuation of the proposed algorithm, the information table shown in Table 1 is processed into a sparse information table which shown as Table 2, it is respectively estimated by SIM-EM algorithm and Prediction-EM algorithm.

The accuracy C in literature [5] and the average absolute error MAE in literature [7] are used to compare SIM-EM algorithm and the proposed Prediction-EM algorithm.

Accuracy rating: it refers to the ratio of the total number of correct estimation attribute values to the total number of non-null attribute values, denotes C:

TableI. INCOMPLETE INFORMATION TABLE

U	a1	a2	a3	a4	a5
x1	3	2	1	0	0
x2	2	3	2	0	0
x3	2	3	2	0	1
x4	*	2	*	1	0
x5	*	2	*	1	1
x6	2	3	2	1	1
x7	3	*	*	3	0
x8	*	0	0	*	1
x9	3	2	1	3	1
x10	1	*	*	*	0
x11	*	2	*	*	1
x12	3	2	1	*	0

TableII. INCOMPLETE SPARSE INFORMATION TABLE

U	a1	a2	a3	a4	a5
x1	3	*	*	0	0
x2	*	3	*	0	*
x3	2	*	*	0	*
x4	*	2	*	1	0
x5	*	2	*	1	*
x6	2	3	2	*	1
x7	3	*	*	3	*
x8	*	0	0	*	1
x9	3	2	*	*	*
x10	1	*	*	*	*
x11	*	2	*	*	1
x12	*	2	1	*	*

$$C = \frac{card(x|x \in U, a \in A, a(x) \neq \varnothing \wedge \Pr e(x,a) = a(x))}{card(x|x \in U, a \in A, a(x) \neq \varnothing)} \qquad (10)$$

In our study, the mean absolute error MAE (mean absolute error) algorithm is one of standards which can evaluate the quality of recommended algorithm, by calculating the deviation between the predicting score and actual score to measure the forecast accuracy.

$$MAE = \frac{\sum_{j=1}^{|CI_i|} |p_{ij} - r_{ij}|}{|CI_i|} \qquad (11)$$

Where p_{ij} is the predicting value of attribute i_k on object u_i, r_{ij} is the real value of attribute i_k on object u_i, $|CI_i|$ is the number of predicted attribute value on object u_i.

TableIII. RESULTS OF SIM-EM AND PREDICTION-EM ON SPARSE INFORMATION TABLE

	Actual Value	**SIM-EM**	**Prediction-EM**		
			0.4	0.5	0.6
v(x1,a1)	1	1	1	1	1
v(x1,a4)	0	1	1	0	1
v(x2,a2)	3	3	3	3	3
v(x2,a4)	0	0	0	1	0
v(x2,a5)	0	1	1	0	1
v(x3,a1)	2	2	2	2	2
v(x3,a4)	0	0	0	0	0
v(x3,a5)	1	0	0	1	0

v(x4,a2)	2	1	1	1	1
v(x5,a2)	2	2	1	2	2
v(x5,a4)	1	1	1	1	1
v(x5,a5)	1	1	1	1	1
v(x6,a1)	2	1	2	2	1
v(x6,a2)	3	3	3	3	3
v(x6,a3)	2	1	1	1	1
v(x7,a1)	3	3	2	2	3
v(x7,a4)	3	2	3	2	2
v(x7,a5)	0	0	0	0	0
v(x8,a2)	0	1	0	0	1
v(x8,a3)	0	0	0	0	0
v(x9,a1)	3	3	3	3	3
v(x9,a2)	2	2	2	2	2
v(x9,a5)	1	1	1	1	1
v(x10,a1)	1	0	1	0	0
v(x10,a5)	0	0	0	0	0
v(x11,a1)	3	1	2	3	1
v(x12,a2)	2	2	2	2	2
v(x12,a3)	1	0	0	0	0
v(x12,a5)	0	0	0	0	0

As shown in Table 1, the sparse degree $\tau = 0.25$, because the information table is non-sparse, both SIM-EM algorithm and Prediction-EM use formula (2) to estimate, therefore we get the same accuracy rating C. As shown in Table 2, the sparse degree $\tau = 0.54$, the information table is sparse, SIM-EM algorithm and Prediction-EM algorithm is respectively used to estimate, the result shown as Table 3.

Due to the three different k value $k = 0.4$, $k = 0.5$, $k = 0.6$ is respectively used to train the information table which is shown in Table 2 with sparse degree 0.54 in this instance, Therefore, through accurate rating and the average absolute error MAE the two indicators to evaluate the training result, observate the different thresholds on the impact of experimental results, and finally get the suitable threshold for this information table. According to Table 3, the threshold of sparse degree are 0.4, 0.5, 0.6, the accuracy rating of filled result is respectively 68.9%、73.3%、65.6%. And the MAE is respectively 31.0%，24.1%，41.4%.

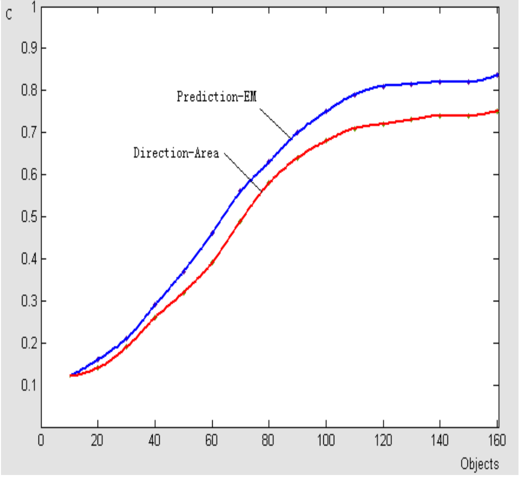

Figure1. Comparison of the Accuracy in filled Information Table

Based on accuracy rating C and the average absolute error MAE, $k = 0.5$ is a suitable threshold for the information table. So in the next section, we take Prediction-EM under $k = 0.5$ threshold to compare with SIM-EM algorithm.

B. *Application Instance*

In order to analyze the filling effect of the improved algorithm further, data sets MovieLens, which is provided by the U.S. GroupLens Research Group University of Minnesota is used as experimental source data in this paper. SIM-EM and Prediction-EM algorithm are used to fill the empty values. At last, accuracy rating and average absolute error of the two algorithms are calculatied.

As shown in Figure 1, the sparse information table is filled by Prediction-EM algorithm, which null value depends on attribute more, while the result of similarity calculation is modified by double weights method, which improves the forecast accuracy. After using Prediction-EM algorithm, accuracy rate is up to 83.6%, while the SIM-EM algorithm is only 75.1%, the Prediction-EM is 8.5% higer than the SIM-EM algorithm.

Figure 2 shows the recommended result of information table filled by Prediction-EM relative to the SIM-EM has a certain improvement on MAE, in particular, when the users' rating data is extremely sparse, that is the users' rating items in 0-110, the improved algorithm have been greatly improved in the accuracy of prediction score for candidate projects. Although with the user-intensive, this advantage will gradually diminish, but improved algorithm for solving the problem of sparse recommendation system is effective.

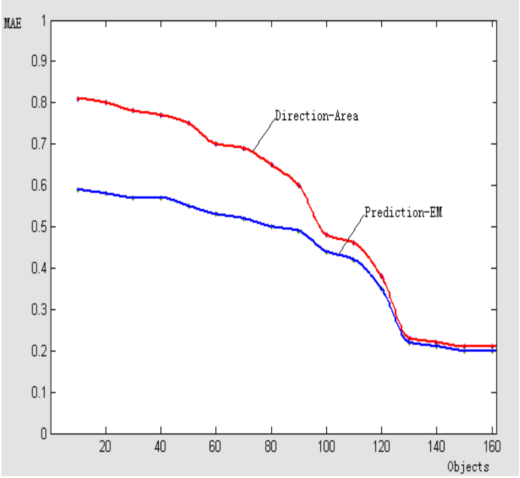

Figure2. Comparison of Recommended Error in filled Information Table

According to comparison between the accuracy rating C and the average absolute error MAE, two index curves can be drawn, from which we can tell the proposed algorithm is more efficient and accurate in the sparse information table.

V. CONCLUSION

The application of rough set theory in incomplete information systems is one of the key to make it into the practical, especially in premise of keeping the original information system, how to change the incomplete information system into a complete information system, the study of theory and method based on it are more and more urgent. For the null value estimation problem in incomplete system, an improved method is proposed in this paper from the perspective of predicting value in collaborative filtering technology, the sparse degree and the dual weight of information tabe are defined, which is combined with the prediction value of attribute, to estimate null values, so both in dense and sparse information table, the quality of estimation is guaranteed. The proposed method in this paper is feasible. But, the complexity of the proposed algorithm needs to be improved.

ACKNOWLEDGMENT

This work is sponsored by the National Natural Science Foundation of China under grant number 61073043, 61073041; the Fundamental Research Funds for the Central Universities of China under grant number HEUCF100602; the Natural Science Foundation of Heilongjiang Province under grant number F200901 and Harbin Special Funds for Technological Innovation Research under grant number 2010RFXXG002.

REFERENCES

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982,11(5):341-356P

[2] Tzung-Pei Hong, Li-Huei Tseng, Been-Chian CHien. Mining from incomplete quantitative data by fuzzy rough sets. Expert Systems With Applications. 2010, 37 (8):2644-2653P

[3] DUAN Peng, ZHUANG Hong, HE Lei, ZHANG Han-yun. Improved algorithm based on incomplete data analysis method. Computer Engineering and Design. 2009.30(7):1681-1684P

[4] YANG Shan-lin. Intelligent decision-making methods and Intelligent Decision Support System [M].Beijing: Science Press,2005

[5] LI Cong, LIANG Cang-yong, YANG Shan-lin. Null values estimation method based on rough set for incomplete information systems. Computer Integrated Manufacturing Systems. 2009.15(3):604-607P

[6] YANG Lian. Study on Attribute Reduction in Incomplete Information Systems Based Rough Set Theory. Sichuan Normal University. Master thesis. 2008.11.26

[7] GUO Yan-hong. On Clooaborative Filtering Algorithm and Applications of Recommender Systems. Dalian University of Technology , PHD thesis.2009.3.19

[8] WANG Wei-ling, LIU Pei-yu, CHU Jian-chong. Improved feature selection algorithm with conditional mutual information. Journal of Computer Applications, 2007.27 (2):433-434P

[9] MA Jian-min, ZHU Chao-hui, ZHANG Wen-xiu. Rough entropy under tolerance relation in set-valued information systems. Computer Engineering and Applications. 2010.46(7):29-31P

[10] K.Yu, Z.Wen, X.Xu and M.Ester. Feature Weighting and Instance Selection for Collaborative Filtering.2nd International Workshop on Management of Information on the Web, in conjunction with the 12th international Conference on DEXA' 2001, Munich, Gemeny, 2001:285-290P

[11] WANG Guo-yin. Extension of rough set under incomplete information systems, Journal of Computer Research and Development. 2002.39(10):1238-1243P