TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.4, April 2014, pp. 2843 ~ 2849
DOI: http://dx.doi.org/10.11591/telkomnika.v12i4.4732

Rough Set Extension Model of Incomplete Information System

Hou Yue
School of Electronic and Information Engineering, Lan Zhou Jiao Tong University
Lan Zhou 730070, P.R.China
email: houyue@mail.lzjtu.cn

Abstract

This paper investigated the incomplete information system in which situation of the more missing or absent unknown values. Based on the original characteristic relation, we proposed a new reflective relation which was controlled by the parameter alpha, beta. By analyzing illustrative examples and comparing to the original characteristic relation, this paper indicated the validity and practicability of the new-defined binary relation.

Keywords*: rough set, incomplete information system, tolerance relation, similarity relation, characteristic relation*

1. Introduction

Proposed by Poland mathematician Z. Pawlak in 1982, rough set theory is an emerging mathematical theory dealing with imprecise, uncertain and incomplete information [1, 2]. The main principle of rough set theory is to make approximate description to imprecise or uncertain knowledge utilizing the known repository [3]. Recently, rough set theory has achieved enormous success in the application of knowledge discovery.

The research object of classical rough set is the complete information system possessed with discrete attribute values, which means all the attribute values in the information system are known[4, 5]. However, the vast majority of information systems are incomplete in real problems. In general, concerning about unknown attribute values in incomplete information system there exist two explanations: 1) all the unknown attribute values exactly exist only to be missed, 2) all the unknown attribute values are considered to be lost but not be compared with. Kryszkiewicz has put forward the tolerance relation of the incomplete information system based on missing semantic and moreover make study on knowledge reduction [6, 7]. Stefanowski has proposed asymmetric similarity relation based on absent semantic [8]. In reality, however, the common situation always happened is that both missing semantic and absent semantic exist in incomplete information system at the same time, so using the above models would result difficulties. In order to use rough set to deal with the incomplete information system possessed with missing and lost unknown attributes, Grzymala-Busse has brought forward the characteristic relation, which is a generalized form combined with tolerance and similarity relation.

Yet, derived from characteristic relation, characteristic set has two unreasonable circumstances: 1) it is possible to make false judgment to classify two objects without any the same clear attributes into a set; 2) it is possible to seperate two objects with large the same known attributes. This paper, aiming at the problems with more missing and absent values, has obtained more reasonable and realistic characteristic set under the analysis and discussion respectively based on new binary relation.

2 Basic Concepts

2.1. Incomplete Information System

An incomplete information system is a four-tuple class: S=<U, AT, V, f>. In this equation, U is a non-empty finite object set called the domain; AT is a non-empty finite attributes

Received September 5, 2013; Revised November 2, 2013; Accepted November 21, 2013

set; regarding $\forall a \in AT$, there is $a : U \to V_a$. V_a is the range of attributes $a \in AT$ (including absent unknown attributes ? and missing unknown attributes *), V is the set of range of all attributes.

$$V = \bigcup_{a \in AT} V_a \tag{1}$$

Define f as the information function, for $\forall a \in AT$, $\forall x \in U$, there is $f(x,a) \in V_a$.

Table 1 is an incomplete information system analyzed in literature 6. Among the analysis, U={1,2,3,4,5,6,7,8},AT={a,b,c}={Temperature,Headache,Nausea}, V_a ={Very_High,High,Normal}

Table 1. Incomplete Information System

U	Temperature	Headache	Nausea
1	High	?	No
2	Very-High	Yes	Yes
3	?	No	No
4	High	Yes	Yes
5	High	?	Yes
6	Normal	Yes	No
7	Normal	No	Yes
8	*	Yes	*

2.2. Characteristic Relations

For the incomplete system shown in Table 1, no matter it is the tolerance relation or similarity relation, the domain can't be classified because the tolerance relation only considers that all the unknown attributes are the missing type and the similarity relation is considered that all the unknown attributes are the absent type. Therefore, concerning that the incomplete system has both the missing and absent type, Grzymala-Busse has constructed the characteristic relation [9] as the following.

Definition 1: Assuming S is an incomplete information system, for $\forall A \subseteq AT$, the expression of characteristic relation decided by A is K (A),

$$K(A) = \{(x,y) \in U^2 : \forall a \in A \wedge f(x,a) \neq ? \tag{2}$$

$$f(x,a) = f(y,a) \vee f(x,a) = * \vee f(y,a) = *\} \tag{3}$$

Definition 2: Assuming S is an incomplete information system, and $A \subseteq AT$, so for $\forall X \subseteq U$, thelower-approximation and upper-approximate set of X based on characteristic relation (A) are respectively regarded as $\underline{A}_K(X), \overline{A}_K(X)$ and:

$$\underline{A}_K(X) = \{x \in U : K_A(x) \subseteq X\} \tag{4}$$

$$\overline{A}_K(X) = \{x \in U : K_A(x) \neq \varnothing\} \tag{5}$$

In the equation, $K_A(x) = \{y \in U : (x,y) \in K(A)\}$.

In the incomplete information system, characteristic relation not only can deal with missing type, but also can process absent unknown attributes incomplete information system. If all the unknown attributes are considered as missing type, the characteristic relation K (A) degenerates to tolerance relation [6, 7]; from the other view, if all the unknown attributes in the incomplete system are considered as lost type, the characteristic relation K (A) degenerates to

asymmetrical similarity relation [9]. Therefore, the characteristic relation has preserved the relevant characters of the tolerance and also has preserved the relevant characters of asymmetrical similarity relation.

3. New Binary Relation

For the characteristic relation K (A), because it has inherited the relevant characters of tolerance and similarity relation, it would make objects without any the same known attributes classify into a set or seperate majority objects with known same attributes easily [10]. For instance, in Table 1 there is $(1,8) \in K(A)$, $(4,5) \notin K(A)$. Hence, aiming at incomplete information system has relatively more absent and missing values, this paper has constructed a kind of new binary relation with parameters respectively.

3.1. Binary Relation with more Missing Values

Definition 3 [11]: Assuming S is an incomplete information system with more missing values, regarding $\forall A \subseteq AT$, new binary relation decided by A is represented as $R^{\alpha,\beta}(A)$ and:

$$R^{\alpha,\beta}(A)=\{(x,y)\in U^2:(\forall b\in B(f(x,b)=f(y,b)\vee f(x,b)=*\vee f(y,b)=*))\wedge (|B|/|N_A(X)|\geq\alpha)\wedge(|C|/|B|\geq\beta)\} \quad (6)$$

With: $B = N_A(x)\bigcap N_A(y)$

$$C = M_A(x)\bigcap N_A(x)\bigcap M_A(y)\bigcap N_A(y)$$

$$N_{AT}(x) = \{a \in A: f(x,a) \neq ?\}$$

$$M_{AT}(x) = \{a \in A: f(x,a) \neq *\}$$

$|X|$ represents the cardinal number of set X, $\alpha \in [0,1]$, $\beta \in [0,1]$.

3.2. Binary Relation with more Absent Values

Definition 4: Assuming S is an incomplete information system with more absent values, regarding $\forall A \subseteq AT$, new binary relation decided by A is represented as $K^{\alpha,\beta}(A)$ and:

$$K^{\alpha,\beta}(A)=\{(x,y)\in U^2:(\forall b\in B(f(x,b)=f(y,b)\vee f(x,b)=*\vee f(y,b)=*))\wedge (|B|/|N_A(X)|\geq\alpha)\wedge(|C|/|B|\geq\beta)\} \quad (7)$$

With $B = M_A(x)\bigcap M_A(y)$

$$C = M_A(x)\bigcap N_A(x)\bigcap M_A(y)\bigcap N_A(y)$$

$$N_{AT}(x) = \{a \in A: f(x,a) \neq ?\}$$

$$M_{AT}(x) = \{a \in A: f(x,a) \neq *\}$$

$|X|$ represents the cardinal number of set X, $\alpha \in [0,1]$, $\beta \in [0,1]$.

Binary relation $R^{\alpha,\beta}(A)$, $K^{\alpha,\beta}(A)$ in definition 4, 5 has introduced two parameters αand β which only satisfy reflexive. Setting parameter α is to prevent the separation of two objects with enormous the same known attributes because of the existence of unknown attribute “?”. Setting parameterβis to prevent two objects without or with only little the same known attributes are classified into a set because of the existence of unknown attribute “*”.

3.3. Algorithm Process

Algorithm of improved characteristic relation is shown as the following:

Step 1: Given an incomplete information system involving both missing and absent attributes.

Step 2: Through observation, the incomplete information system is divided into two situations:

If the system has more missing values, the characteristic relation $R^{\alpha,\beta}(A)$ is calculated according to Equation (4).

If the system has more absent values, the characteristic relation $K^{\alpha,\beta}(A)$ is calculated according to Equation (5).

Step 3: The improved characteristic relation between any two objects among all sample objects are obtained through the set to the threshold αand β.

The algorithm process chart is demonstrated in Figure 1:

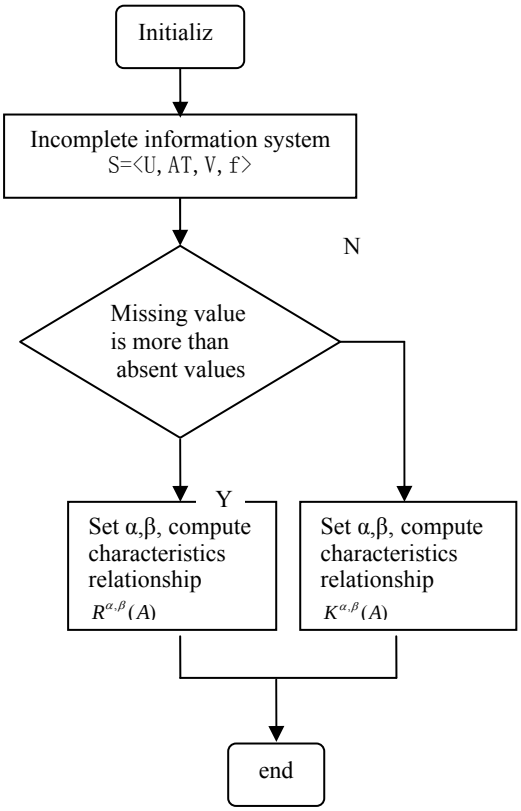

Figure 1. Algorithm Process

4. Case Analysis

The incomplete information system with more missing values shown in Table 2, U={O_i }(1≤i≤12), AT={a,b,c,d} is the set of all attributes.

According to the characteristic relation demonstrated in Definition 2, following characteristic sets can be obtained: $K_{AT}(O_1)=\{O_1,O_{11},O_{12}\}$, $K_{AT}(O_2)=\{O_2,O_3\}$, $K_{AT}(O_3)=\{O_2,O_3\}$, $K_{AT}(O_4)=\{O_4,O_{11},O_{12}\}$, $K_{AT}(O_5)=\{O_4,O_5,O_6,O_8,O_{11},O_{12}\}$, $K_{AT}(O_6)=\{O_6\}$, $K_{AT}(O_7)=\{O_7,O_8,O_9,O_{11},O_{12}\}$, $K_{AT}(O_8)=\{O_8\}$, $K_{AT}(O_9)=\{O_9,O_{11},O_{12}\}$, $K_{AT}(O_{10})=\{O_4,O_8,O_{10},O_{11}\}$, $K_{AT}(O_{11})=\{O_1,O_4,O_9,O_{11},O_{12}\}$, $K_{AT}(O_{12})=\{O_1,O_4,O_9,O_{11},O_{12}\}$.

According to the characteristic relation demonstrated in definition 3, assuming α = β =0.5, results can be obtained: $R^{AT}_{\alpha,\beta}(O_1)=\{O_1,O_{12}\}$, $R^{AT}_{\alpha,\beta}(O_2)=\{O_2,O_3\}$, $R^{AT}_{\alpha,\beta}(O_3)=\{O_2,O_3\}$, $R^{AT}_{\alpha,\beta}(O_4)=\{O_4\}$, $R^{AT}_{\alpha,\beta}(O_5)=\{O_5\}$, $R^{AT}_{\alpha,\beta}(O_6)=\{O_6\}$, $R^{AT}_{\alpha,\beta}(O_7)=\{O_7,O_9\}$, $R^{AT}_{\alpha,\beta}(O_8)=\{O_8\}$, $R^{AT}_{\alpha,\beta}(O_9)=\{O_7,O_9,O_{12}\}$, $R^{AT}_{\alpha,\beta}(O_{10})=\{O_{10}\}$, $R^{AT}_{\alpha,\beta}(O_{11})=\{O_{11},O_{12}\}$, $R^{AT}_{\alpha,\beta}(O_{12})=\{O_1,O_9,O_{12}\}$。

Clearly, $R^{AT}_{\alpha,\beta}(x)$ is better than $K_{AT}(x)$. For instance, object O_5 and O_8 have not a known equal attribute, but $O_8 \in K_{AT}(O_5)$, while $O_8 \notin R^{AT}_{\alpha,\beta}(O_5)$, the similar situation also has occurred at O_5 and O_{11}, O_5 and O_{12}. On the other side, O_7 and O_9 have greater possibility in similarity, but $O_7 \notin K_{AT}(O_9)$, while $O_7 \in R^{AT}_{\alpha,\beta}(O_9)$.

Table 2. Incomplete Information System with more Missing Values

	a	b	c	d
O_1	3	2	1	0
O_2	2	3	2	0
O_3	2	3	2	0
O_4	*	2	*	1
O_5	*	?	*	1
O_6	2	3	2	1
O_7	3	?	*	3
O_8	*	0	0	*
O_9	3	2	1	3
O_{10}	1	*	*	?
O_{11}	*	2	*	*
O_{12}	3	2	1	*

The incomplete information system with more absent values shown in Table 3, U={O_i }(1≤i≤12), AT={a,b,c,d} is the set of all attributes.

Table 3. Incomplete Information System with more Absent Values

	a	b	c	d
O_1	3	2	1	0
O_2	2	3	2	0
O_3	2	3	2	0
O_4	?	2	?	1
O_5	*	*	?	1
O_6	2	3	2	1
O_7	3	*	?	3
O_8	?	0	0	?
O_9	3	2	1	3
O_{10}	1	?	*	*
O_{11}	?	2	1	*
O_{12}	3	2	1	?

According to the characteristic relation demonstrated in Definition 2, following characteristic set can be obtained: $K_{AT}(O_1)=\{O_1\}$, $K_{AT}(O_2)=\{O_2,O_3\}$, $K_{AT}(O_3)=\{O_2,O_3\}$, $K_{AT}(O_4)=\{O_4,O_5,O_{11}\}$, $K_{AT}(O_5)=\{O_5,O_6\}$, $K_{AT}(O_6)=\{O_6\}$, $K_{AT}(O_7)=\{O_7,O_9\}$, $K_{AT}(O_8)=\{O_8\}$, $K_{AT}(O_9)=\{O_9\}$, $K_{AT}(O_{10})=\{O_{10}\}$, $K_{AT}(O_{11})=\{O_1,O_{11}\}$, $K_{AT}(O_{12})=\{O_1,O_9,O_{12}\}$.

According to the characteristic relation demonstrated in definition 3, assuming α = β =0.5, results can be obtained: $K^{AT}_{\alpha,\beta}(O_1)=\{O_1,O_{12}\}$, $K^{AT}_{\alpha,\beta}(O_2)=\{O_2,O_3\}$, $K^{AT}_{\alpha,\beta}(O_3)=\{O_2,O_3\}$, $K^{AT}_{\alpha,\beta}(O_4)=\{O_4\}$, $K^{AT}_{\alpha,\beta}(O_5)=\{O_5\}$, $K^{AT}_{\alpha,\beta}(O_6)=\{O_6\}$, $K^{AT}_{\alpha,\beta}(O_7)=\{O_7\}$, $K^{AT}_{\alpha,\beta}(O_8)=\{O_8\}$, $K^{AT}_{\alpha,\beta}(O_9)=\{O_9,O_{12}\}$, $K^{AT}_{\alpha,\beta}(O_{10})=\{O_{10}\}$, $K^{AT}_{\alpha,\beta}(O_{11})=\{O_{11},O_{12}\}$, $K^{AT}_{\alpha,\beta}(O_{12})=\{O_1,O_9,O_{11},O_{12}\}$.

Seen from the results, we can find $K^{AT}_{\alpha,\beta}(x)$ is better than $K_{AT}(x)$. For instance, object O_5 and O_6 have only one known equal attribute, but $O_6 \in K_{AT}(O_5)$. O_1 and O_{12} have greater possibility in similarity, but $O_{12} \notin K_{AT}(O_1)$, while $O_{12} \in K^{AT}_{\alpha,\beta}(O_1)$. O_7 and O_9 have only two same attributes which cannot be judged whether the two attributes can be distinguished or not intuitively, but it considered to be undistinguished under α = β =0.7. Therefore, by setting α and β, the new characteristic relation has better solved the unreasonable condition caused by classification of characteristic relation and also conformed the intuitive feeling dealing with data.

5. Conclusion

Application of rough set in the incomplete information system has become a study hot spot recently. In order to use rough set to deal with the incomplete information system possessed with missing and lost unknown attributes at the same time, Grzymala-Busse has proposed the concept of characteristic relation and characteristic set. However for different information system, according to the characteristic relation, it is possible that the classification would divide two objects without any known same attributes into a set or would separate two objects with large known same attributes. Therefore, aiming at two incomplete information system with more missing and absent values, this paper respectively has proposed a new binary relation with two parameters. New constructed rough set model is better than general rough set model as long as setting α and β properly.

Acknowledgements

This work was supported by the Youth Science Fund Project of Lanzhou Jiaotong Unive rsity (No. 2013006).

References

[1] Pawlak Z. Rough set theory and its applications to data analysis. *Cybernetics and Systems*. 1998; 29: 661-688.
[2] Pawlak Z. Rough sets and intelligent data analysis. *Information Sciences.* 2002; 147: 1-12.
[3] Zhang WX, Wu WZ. A summarize of survey and introduction on rough ret theory. *Fuzzy Systems and Mathematics.* 2000; 14(4): 1-12.
[4] Haiying Dong, Xiaonan Li, Zhanhong Wei. Substation Fault Diagnosis Based on rough Sets and Grey Relational Analysis. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2014; 12(2): 1162-1168.
[5] Chu Yan, Wang Haiguang, Chen Liang. Study on Fault Diagnosis of Circuit-breaker Based on Rough-Set Theory. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11(1): 296-301.
[6] Xuexia Liu. Study on Knowledge-based Intelligent Fault Diagnosis of Hydraulic System. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 10(8): 2041-2046.

[7] Marzena K. Generation of rules from incomplete information systems. *Lecture Notes in Computer Science*. 1997; 1263: 156-166.
[8] Jerzy S, Alexis T. On the extension of rough sets under incomplete information. *Lecture Notes in Computer Science*. 2004; 1711: 73-82.
[9] Diwakar S, Rahul S, Singh T. A New Imputation Method for Missing Attribute Values in Data Mining. *Journal of Applied Computer Science and Mathematics*. 2011; 5(10): 14-19.
[10] Shen JB. An extension model of rough set in incomplete information system. *Application Research of Computers*. 2009; 26(6): 2101-2103.
[11] Yu DJ, Yang XB, Yang JY. An extension model of rough set in incomplete information system. *Journal of Huaiyin Institute of Technology*. 2008; 17(1): 31-37.