A THRESHOLD-BASED SIMILARITY RELATION UNDER INCOMPLETE INFORMATION

Xuri Yin
Simulation Laboratory of Military Traffic, Institute of Automobile, Management of PLA, Bengbu, 233011, China, yinxuri@163.com

Abstract: The conventional rough set theory based on complete information systems stems from the observation that objects with the same characteristics are indiscernible according to available information. Although rough sets theory has been applied in many fields, the use of the indiscernibility relation may be too rigid in some real situations. Therefore, several generalizations of the rough set theory have been proposed some of which extend the indiscernibility relation using more general similarity or tolerance relations. In this paper, after discussing several extension models based on rough sets for incomplete information, a novel relation based on thresholds is introduced as a new extension of the rough set theory, the upper-approximation and the lower approximation defined on this relation are proposed as well. Furthermore, we present the properties of this extended relation. The experiments show that this relation works effectively in incomplete information and generates rational object classification.

Key words: rough sets, incomplete information, tolerance relation, similarity relation, constrained dissymmetrical similarity relation.

1. INTRODUCTION

Rough set theory (Pawlak, 1982), which has been developed by Z.Pawlak and his co-workers since the early 1980s, has recently received more and more attention as a means of knowledge discovery. Rough set is a kind of mathematical tool, which is used to depict incompletion and uncertainty of the information. We can discover the connotative knowledge and the

underlying rules through its analyzing and reasoning for the data (Yin et al., 2001).

In the classical rough set theory the information system must be complete. However, in the real world some attribute values may be missing due to errors in the data measure, the limitation of data comprehension as well as neglects during the data registering process. Therefore, several generalizations of the rough sets theory have been proposed to deal with the incomplete information systems. The LERS system first transforms an incomplete information system into complete information system, then generate rules (Chmielewski et al., 1998). Kryszkiewicz proposed a new method which produces rules from incomplete information system directly, he extended some concepts of the rough set theory in the incomplete information system, studied the tolerance relation in his papers (Kryszkiewicz, 1998; Kryszkiewicz, 1999). Stefanowski presented an extended rough set theory model based on similarity relations and tolerance relations (Stefanowski et al., 1999). In the paper (Yin et al., 2006), the constrained dissymmetrical similarity relation is introduced and showed that it is between the tolerance relation and the similarity relation. Another model based on constrained similarity relations was defined in the paper (Wang, 2002).

In this paper, several present extension models of rough set under incomplete information systems are discussed. Then the concept of threshold based similarity relation as a new extension of rough sets theory is introduced, and the upper-approximation and lower-approximation are redefined. Furthermore, the properties of this relation are discussed also. The experiments show that the proposed threshold based similarity relation can effectively process incomplete information and generate rational object classes. By threshold man can easily control the partition of universe in some way.

The rest of this paper is organized as follow: Section 2 introduces several extension models based on rough sets theory under incomplete information systems. Section 3 presents a concept of threshold based similarity relation as a new extension of rough sets theory and discusses its properties. Section 4 shows some examples. Finally, we conclude the paper with a summary in Section 5.

2. SEVERAL EXTENSION MODELS

Knowledge representation in rough set theory is done via information systems, which are a form of data table.

Definition 1. *Structure* $I = <U, \Omega, V_q, f_q>_{q\in\Omega}$ *is called an information system where:*

1. U is a nonempty and finite set of a group objects (or instances), called the universe of discourse, Assume the number of the objects is n, then U can be denoted to : $U=\{x_1, x_2, ..., x_n\}$,

2. Ω is a nonempty and finite set which contains finite attributes, Assume the number of the attributes is m, then it can be denoted to : $\Omega=\{q_1, q_2,..., q_m\}$,

3. For each $q \in \Omega$, V_q is enumerated domain of the attribute q ,

4. For each $q \in \Omega$, f_q is a information function, $f_q: U \rightarrow V_q$, such that $\forall x \in U, \exists y \in V_q, f_q(x)=y$.

The information system with such a domain V_q that contains missing values represented by "*" is an incomplete information system.

To process and analyze the incomplete information system, Kryszkiewicz proposed the tolerance relation T as follows (Kryszkiewicz, 1999):

$$\forall_{x,y\in U}(T_B(x,y) \Leftrightarrow \forall_{b\in B}((f_b(x)=*)\vee(f_b(y)=*)\vee(f_b(x)=f_b(y))$$

The tolerance relation T satisfies the reflexivity and symmetry, but not transitivity. The lower-approximation $\underline{X}_B^T$ and upper-approximation $\overline{X}_B^T$ can be defined as:

$$\underline{X}_B^T=\{x \mid x\in U \wedge I_B(x)\subseteq X\} \qquad \overline{X}_B^T=\{x \mid x\in U \wedge (I_B(x)\cap X\neq\phi)\} \quad (1)$$

Where

$$I_B(x)=\{y \mid y\in U \wedge T_B(x,y)\} \quad (2)$$

Stefanowski and others proposed a dissymmetrical similarity relation S.

$$\forall_{x,y\in U}(S_B(x,y) \Leftrightarrow \forall_{b\in B}((f_b(x)=*)\vee(f_b(x)=f_b(y))$$

Obviously, the relation S is dissymmetrical, but transferable and reflexive. Also, Stefanowski defined the lower-approximation $\underline{X}_B^S$ and the upper-approximation $\overline{X}_B^S$ of the set $X \subseteq U$ based on the dissymmetrical similarity relation S (Stefanowski et al., 1998):

$$\underline{X}_B^S=\{x \mid x\in U \wedge \underline{R}_B^S(x)\subseteq X\} \qquad \overline{X}_B^S=\bigcup_{x\in B}\overline{R}_B^S(x) \quad (3)$$

Where

$$\underline{R}_B^S(x)=\{y \mid y\in U \wedge S_B(x,y)\} \qquad \overline{R}_B^S(x)=\{y \mid y\in U \wedge S_B(y,x)\} \quad (4)$$

It can be proved that the lower-approximation and upper-approximation of the object set *X* which is based upon the dissymmetrical similarity relation S is an extension to which is based upon the tolerance relation T .

In the paper (Yin et al., 2006), we presented a constrained dissymmetrical similarity relation C, which is defined as:

$$\forall_{x,y\in U}(C_B(x,y)\Leftrightarrow\forall_{b\in B}(f_b(x)=*)\vee((P_B(x,y)\neq\varnothing)\wedge\forall_{b\in B}((b\in P_B(x,y))\rightarrow(f_b(x)=f_b(y))))$$

Where:

$$P_B(x,y)=\{b\mid b\in B\wedge(f_b(x)\neq *)\wedge(f_b(y)\neq *)\}$$

Obviously, relation C is reflexive, but not symmetric and transferable. The lower- approximation $\underline{X}_B^C$ and upper-approximation $\overline{X}_B^C$ based on the constrained dissymmetrical similarity relation C is defined:

$$\underline{X}_B^C=\{x\mid x\in U\wedge\underline{R}_B^C(x)\subseteq X\}\qquad\overline{X}_B^C=\bigcup_{x\in B}\overline{R}_B^C(x)\qquad(5)$$

Here,

$$\underline{R}_B^C(x)=\{y\mid y\in U\wedge C_B(x,y)\}\qquad\overline{R}_B^C(x)=\{y\mid y\in U\wedge C_B(y,x)\}\qquad(6)$$

Theorem 1. *Information system* $I=<U, \Omega, V_q, f_q>_{q\in\Omega}$, $X\subseteq U$, $B\subseteq\Omega$,

(1) $\underline{X}_B^T\subseteq\underline{X}_B^C$; $\overline{X}_B^C\subseteq\overline{X}_B^T$

(2) $\underline{X}_B^C\subseteq\underline{X}_B^S$; $\overline{X}_B^S\subseteq\overline{X}_B^C$

Theorem 1 shows that the constrained dissymmetrical similarity relation is just between tolerance relation and similarity relation (Yin et al., 2006).

3. THRESHOLD-BASED SIMILARITY RELATION

In the paper (Yin et al., 2006), the second kind of situation of relation C is one in which the object x and y must have the same definite attribute value in at least one attribute. But with the incensement of the attribute number in data sets, this kind of condition still appeared quite loosely .Therefore, we have the necessity to introduce a threshold value, ratio of the number of attributes that has the same definite attribute value for object x and y to the number of all attributes. By adjusting the threshold values, to a certain extent, man can flexibly determine the class of the objects to meet the needs for practical applications in the areas of data mining.

Definition 2. *Assume that information system* $I=<U, \Omega, V_q, f_q>_{q\in\Omega}$, $B\subseteq\Omega$ *and is a nonem*pty, *the threshold based similarity relation* A *can be defined as:*

$$\forall_{x,y\in U}(\mathrm{A}_B(x,y)\Leftrightarrow\forall_{b\in B}(f_b(x)=*)\vee((\frac{P_B(x,y)}{|B|}\geq\alpha)\wedge\forall_{b\in B}((b\in P_B(x,y))\rightarrow(f_b(x)=f_b(y)))$$

Where:

$$P_B(x,y)=\{b\mid b\in B\wedge(f_b(x)\neq *)\wedge(f_b(y)\neq *)\}$$

$$0\leq\alpha\leq 1$$

The lower-approximation and upper-approximation based on the threshold based similarity relation A can be defined in the following.

Definition 3. *Assume that information system* $I=<U, \Omega, V_q, f_q>_{q\in\Omega}$, $X\subseteq U$, $B\subseteq\Omega$ *and is a nonempty, the lower- approximation* $\underline{X}_B^{\mathrm{A}}$ *and upper-approximation* $\overline{X}_B^{\mathrm{A}}$ *based on the threshold based similarity relation* A *can be defined as:*

$$\underline{X}_B^{\mathrm{A}} = \{x \mid x \in U \wedge \underline{R}_B^{\mathrm{A}}(x) \subseteq X\} \qquad \overline{X}_B^{\mathrm{A}} = \bigcup_{x\in B} \overline{R}_B^{\mathrm{A}}(x) \qquad (7)$$

Here,

$$\underline{R}_B^{\mathrm{A}}(x) = \{y \mid y \in U \wedge \mathrm{A}_B(x, y)\} \qquad \overline{R}_B^{\mathrm{A}}(x) = \{y \mid y \in U \wedge \mathrm{A}_B(y, x)\} \qquad (8)$$

Theorem 2. *Information system* $I=<U, \Omega, V_q, f_q>_{q\in\Omega}$, $X\subseteq U$, $B\subseteq\Omega$ *and is a nonempty,* $0 \le \alpha \le 1$. *Then*

(1) If $\alpha = 0$, *then* $\mathrm{A}_B(x, y) = T_B(x, y)$;

(2) If $0 < \alpha \le 1/|B|$, *then* $\mathrm{A}_B(x, y) = C_B(x, y)$;

(3) If $1/|B| < \alpha \le 1$, *then* $\underline{X}_B^{\mathrm{A}} \supseteq \underline{X}_B^{C}$, $\overline{X}_B^{\mathrm{A}} \subseteq \overline{X}_B^{C}$;

(4) If $1/|B| \le \alpha_1 \le \alpha_2 \le 1$, *then* $\underline{X}_B^{\mathrm{A}_2} \supseteq \underline{X}_B^{\mathrm{A}_1}$, $\overline{X}_B^{\mathrm{A}_2} \subseteq \overline{X}_B^{\mathrm{A}_1}$

Proof. (1) According to the definitions of relation T and relation A, it is obvious that If $\alpha = 0$, then $\mathrm{A}_B(x, y) = T_B(x, y)$

(2)For any object x and y of U, if $0 < \alpha \le 1/|B|$, then $\frac{P_B(x,y)}{|B|} \ge \alpha \Leftrightarrow P_B(x,y) \ne \varnothing$. By the definition 2 and the definition of constrained dissymmetrical similarity relation we have

$$\mathrm{A}_B(x, y) = C_B(x, y)$$

(3)Let $1/|B| < \alpha \le 1$, for any object x and y of U,

if $\frac{P_B(x,y)}{|B|} \ge \alpha \Rightarrow P_B(x,y) \ne \varnothing$, so

$$\forall_{x,y\in U}(\mathrm{A}_B(x, y) \Rightarrow C_B(x, y)) \qquad \forall_{x,y\in U}(\mathrm{A}_B(y, x) \Rightarrow C_B(y, x))$$

$$\underline{R}_B^{\mathrm{A}}(x) \subseteq \underline{R}_B^{C}(x) \qquad \overline{R}_B^{\mathrm{A}}(x) \subseteq \overline{R}_B^{C}(x)$$

By the definitions above we can conclude:

$$\underline{X}_B^{\mathrm{A}} \supseteq \underline{X}_B^{C}, \overline{X}_B^{\mathrm{A}} \subseteq \overline{X}_B^{C}$$

(4) When $1/|B| < \alpha_1 \le \alpha_2 \le 1$, it is evident that

$$\forall_{x,y\in U}\left(\frac{P_B(x,y)}{|B|} \ge \alpha_2 \Rightarrow \frac{P_B(x,y)}{|B|} \ge \alpha_1\right)$$

So,

$$\forall_{x,y\in U}(\mathrm{A}_{2B}(x,y)\Rightarrow \mathrm{A}_{1B}(x,y)) \qquad \forall_{x,y\in U}(\mathrm{A}_{2B}(y,x)\Rightarrow \mathrm{A}_{1B}(y,x))$$

$$\underline{R}_B^{A_2}(x)\subseteq \underline{R}_B^{A_1}(x) \qquad \overline{R}_B^{A_2}(x)\subseteq \overline{R}_B^{A_1}(x)$$

According to the definitions above we have the following conclusion:

$$\underline{X}_B^{A_2}\supseteq \underline{X}_B^{A_1}, \overline{X}_B^{A_2}\subseteq \overline{X}_B^{A_1}$$

In the classical rough set theory, the lower-approximation and the upper approximation of the object set are defined by equivalence relation. In incomplete information systems, the tolerance relation and the similarity relation can be seen as an extension of equivalence relation. From theorem 2, we can see that the threshold based similarity relation proposed in this paper is an extension of the tolerance relation and the constrained dissymmetrical similarity relation.

4. EXAMPLES

We use two examples to analyze the threshold based relation proposed above, one of which is an incomplete information system from the paper (Stefanowski et al., 1999) and the other is a data set from the *UCI Machine Learning Repository*.

Table 1. An example of the incomplete information system

A	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
C_1	3	2	2	*	*	2	3	*	3	1	*	3
C_2	2	3	3	2	2	3	*	0	2	*	2	2
C_3	1	2	2	*	*	2	*	0	1	*	*	1
C_4	0	0	0	1	1	1	3	*	3	*	*	*
D	Φ	Φ	Ψ	Φ	Ψ	Ψ	Φ	Ψ	Ψ	Φ	Ψ	Φ

Firstly, the incomplete information system is given in Table 1, where U is the set of objects denoted as U = {a_1, a_2 . . . , a_{12}} and B is the set of condition attributes denoted as {C_1, C_2, C_3, C_4}, D is the decision attribute, "*" denotes the missing value. Assume that X = {a_1, a_2, a_4, a_7, a_{10}, a_{12}} (Stefanowski et al., 1999).

(1)If $\alpha = 0$, then we can conclude:

$$\underline{X}_B^{A}=\underline{X}_B^{T}=\phi \qquad \overline{X}_B^{A}=\overline{X}_B^{T}=\{a_1,a_2,a_3,a_4,a_5,a_7,a_8,a_9,a_{10},a_{11},a_{12}\}$$

(2)If $0<\alpha\le 0.25$, then we can conclude:

$$\underline{X}_B^{A}=\underline{X}_B^{C}=\{a_{10}\} \qquad \overline{X}_B^{A}=\overline{X}_B^{C}=\{a_1,a_2,a_3,a_4,a_5,a_7,a_9,a_{10},a_{11},a_{12}\}$$

(3)If $0.25<\alpha\le 0.5$, then we can conclude:

$$\underline{X}_B^{A}=\{a_1,a_{10},a_{11}\} \qquad \overline{X}_B^{A}=\{a_1,a_2,a_3,a_4,a_5,a_7,a_9,a_{12}\}$$

(4)If $0.5<\alpha\le 0.75$, then we can conclude:

$$\underline{X}_B^{A}=\{a_1,a_4,a_5,a_7,a_8,a_{10},a_{11}\} \qquad \overline{X}_B^{A}=\{a_1,a_2,a_3,a_9,a_{12}\}$$

(5)If $0.75 < \alpha \le 1$, then we can conclude:

$\underline{X}_B^A = \{a_1, a_4, a_5, a_7, a_8, a_{10}, a_{11}, a_{12}\}$ $\overline{X}_B^A = \{a_1, a_2, a_3\}$

Obviously, with $1/|B| < \alpha \le 1$, we have

$$\underline{X}_B^A \supseteq \underline{X}_B^C, \overline{X}_B^A \subseteq \overline{X}_B^C .$$

Moreover, if $1/|B| \le \alpha_1 \le \alpha_2 \le 1$, then $\underline{X}_B^{A_2} \supseteq \underline{X}_B^{A_1}, \overline{X}_B^{A_2} \subseteq \overline{X}_B^{A_1}$

Secondly, we choose a data set named shuttle-landing-control which is concerned about *Space Shuttle Autolanding Domain* from the *UCI Machine Learning Repository*. In order to validate its ability in dealing with practiced problems, we made some appropriate modification in it: replacing some real values with missing values randomly at the ratio of less than 15%. Just as the following Table.2.

Table 2. Modified shuttle-landing-control data set

A	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}
C_1	*	2	1	1	1	*	1	1	1	*	1	1	*	1	1
C_2	*	*	2	1	3	*	4	4	4	3	*	3	3	3	3
C_3	*	*	*	*	2	*	*	*	*	*	1	1	1	1	1
C_4	*	*	*	*	2	*	*	*	*	1	*	2	*	1	2
C_5	*	*	*	*	*	*	1	2	*	1	*	1	*	3	3
C_6	2	1	1	1	1	1	1	1	1	*	1	*	1	1	1
D	Φ	Ψ	Ψ	Ψ	Ψ	Ψ	Φ	Φ	Φ	Φ	Φ	Φ	Φ	Ψ	Φ

For this data set, let X⊆U, B⊆Ω and is a nonempty, by the calculation we can also draw all conclusions of theorem 2.

The experiment results show that the threshold based similarity relation proposed in this paper is an extension of the tolerance relation and the constrained dissymmetrical similarity relation, that is, the tolerance relation and the constrained dissymmetrical similarity relation are a special case of the threshold based similarity relation. This relation makes objects' classification more reasonable, and it is more practicable and flexible than the present.

5. CONCLUSION

Using standard rough set theory we may describe complete information systems. However many real-life data sets for data analysis usually contain a mass of missing values. So, the research how to acquire knowledge from such an incomplete information system has become a hotspot.

In this paper, after analyzing several present models based on rough sets for incomplete information systems, we propose an extended model under the threshold based similarity relation. From both the theoretically proof and experiments it can be seen that the rough set model based on the threshold

based similarity relation is classifies more reasonable and flexible than that based on tolerance relation or constrained dissymmetrical similarity relation. By threshold man can easily control the partition of universe in some way.

REFERENCES

Chmielewski, M.R., Grzymala-Busse, et al. (1998). The rule induction system LERS-A version for personal computers. *Found Compute Decision Sciences*, 3/4: 181~212.

Kryszkiewicz, M. Rough set approach to incomplete information systems. (1998). *Information Sciences*, 1-4: 39~49.

Kryszkiewicz, M. Rules in incomplete information system. (1999). *Information Sciences*, 4: 271~292.

Pawlak Z. Rough sets. (1982). International Journal of Information and Computer Science,, 5: 341~356.

Stefanowski, J., Tsoukias, A. (1999). On the extension of rough sets under Incomplete Information. In: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, Yamaguchi: Physica-Verlag, 73~81.

Wang G.Y. (2002). Extension of rough set under incomplete information systems. *Journal of Computer Research and Development*, 10: 1238~1243 (in Chinese).

Yin,X.R, Jia,X.Y., Shang,L. (2006). A New Extension Model of Rough Set Under Incomplete Information. In: Proceedings of the First International Conference on Rough Sets and Knowledge Technology, LNAI 4062, Springer-Verlag Berlin Heidelberg, 141~146.

Yin,X.R, Zhou,Z.H., Li,N.,Chen,S.F. (2001). An approach for data filtering based on rough set theory. In: Proceedings of the Second International Conference on Web-age Information Management. LNCS 2118, Springer-Verlag Berlin Heidelberg, 367-374.