Knowledge Dependency in Expanded Incomplete Information Systems

Chen Wu, Enbin Wang, Xibei Yang

School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang, 212003, China

Abstract
Incomplete information systems are expanded with the significance of objects in order to combine factors such as decision maker's preferences and prior domain knowledge, etc. In the expanded incomplete information systems, information granules have their own significance, some basic concepts in rough set theory such as accuracy measure, rough entropies of knowledge and a set are both established and proved to be monotonous. Weightily, the concepts of knowledge dependency and knowledge dependency for decision are not only established but also deeply investigated. Furthermore, the measurement of knowledge dependency is described by mathematical formula strictly, some interesting results about the measurement of knowledge dependency are gained with the changing of knowledge.

Key words:
rough set, incomplete information system, knowledge dependency, entropy

1. Introduction

As one of the effective mathematical tools for intelligent data analysis[1], Rough Set Theory (RST)[2] has been received more and more recognitions by a lot of researchers in recent years. The traditional RST was proposed by Pawlak and it is on the assumption that all objects in the universe have complete values of attributes. Unfortunately, incomplete information systems (IIS) can be seen everywhere in actual world. There are also many researchers have done significant jobs to expand the indiscernibility relation that in traditional RST to some other relations such as tolerance relation[3], similarity relation[4], limited tolerance relation[5] or more common binary relation (reflective is needed) in order to use RST to deal with incomplete information systems directly and effectively.

Rough set models, whether in complete or incomplete information systems, are on the assumption that objects are equally important, ie., all objects in universe have same significance or importance. However, practical problems are not as simple as the assumption sometimes. For instance, the different sources of data, the prior domain knowledge in different fields and subjective preferences with different decision-makers[6], etc, are all able to make different objects have different levels of significance. From that, it is important to extend incomplete information systems by the significance of objects.

2.Expanded incomplete information system

An incomplete information system is a quadruple $S =< U,AT,V,f >$, where U is a non-empty finite set of objects

Table 1 An expanded IIS

	P	M	S	A	f'
1	H	H	F	D,E	0.8
2	L	*	F,C	E	0.7
3	*	*	C	D	0.5
4	H	*	F	D	0.8
5	*	*	F	D	0.9
6	L	*	F	*	1.0

and AT is a non-empty finite set of attributes, such that $a \in AT : U \to V_a$, where V_a is called the value set of *a*. Any attribute domain V_a may contain not only special symbol "*" to indicate that the value of an attribute is unknown but also set of values. V is regard as the value set of all attributes in S and then V should satisfies with $V=\cup_{a\in AT} V_a$. Define *f* as an information function in S and there will be $f(x, a) \in V_a$ for any $a \in AT$ and $x \in U$.

Definition 1 Let S be an incomplete information system and ψ_1, ψ_2 be two coverings on universe U. If $\forall \mu \in \psi_1, \exists \nu \in \psi_2$ holds that $\mu \subseteq \nu$ and if $\forall \nu \in \psi_2, \exists \mu \in \psi_1$ holds $\nu \supseteq \mu$, then covering ψ_1 is a refinement of ψ_2, or equivalently ψ_2 is a coarsening of ψ_1, denoted by $\psi_1 \preceq \psi_2$.

Definition 2 An incomplete information system with the significance of objects is an expansion of incomplete information system, denoted by

Manuscript received May 5, 2008
Manuscript revised May 20, 2008

$S^{'}$ =< U,AT,V,f,$V^{'}$,$f^{'}$ >, where $V^{'}$ is the value set of objects' significance and $f^{'}$ is a function such that for $\forall x \in$ U, $f^{'}$ (x) $\in$ $V^{'}$. Table 1 is an expanded incomplete information system in which the information function $f^{'}$ represents the significance of each object. In general, $f^{'}$(x) $\in$ [0, 1].

Let $S^{'}$ be an expanded incomplete information system, for each subset of attributes $A \subseteq$ AT, A determines a binary relation R_A on U. R_A is not always an indiscernibility relation but some other binary relation. The binary relation represents the similarity between elements of a universe. It is reasonable to assume that a binary relation is at least reflexive, but not necessarily symmetric and transitive and as a result we call R_A is a reflective binary relation[7].

Furthermore, for any $x \in$ U, let us denote by $[x]_R$ the set of objects y for which R_A holds, in other words, $[x]_{R_A}$ is the maximal set of objects which have relation R_A with x. In expanded incomplete information system $S^{'}$, $[x]_{R_A}$ is a information granule of x with relation R_A, it is different from the information granule created by indiscernibility relation for the reason that any one element in the universe may belong to two or more different information granules.

Any one information granule in the expanded incomplete information system has an important property—-significance, that could be represented as F $([x]_{R_A})$= $\sum_{y \in [x]_A} f^{'}(y)$

Let U/R_A denote classification, which is the family set $\{[x]_{R_A}: x \in U\}$. What should be noticed is that a reflective binary relation in $S^{'}$ does not constitute a partition in general, but a covering on universe U. Obviously, $\cup U/R_A$ = U and $[x]_{R_A} \neq \varnothing$.

We know that if $A \subseteq B \subseteq$ AT, then $U/R_B \preceq U/R_A$.

Definition 3 Let $S^{'}$ be an expanded incomplete information system, $A \subseteq$ AT, then for $\forall X \subseteq$ U, the B-lower approximation and the B-upper approximation of X are defined as follows, respectively:

$$A_*(X) = \{x \in U : [x]_{R_A} \subseteq X\} \quad (1)$$

$$A^*(X) = \{x \in U : [x]_{R_A} \cap X \neq \varnothing\} \quad (2)$$

Definition 4 Let $S^{'}$ be an expanded incomplete information system, $A \subseteq$ AT and $X \subseteq$ U, then the significance of set X is denoted by F (X)= $F(X) = \sum_{y \in X} f^{'}(y)$, the significance of lower and upper approximation of X are defined as follows, respectively:

$$F(A_*(X)) = \sum_{y \in A_*(X)} f^{'}(y) \quad (3)$$

$$F(A^*(X)) = \sum_{y \in A^*(X)} f^{'}(y) \quad (4)$$

Using the concepts of significance of lower and upper approximation, we can define accuracy measure in expanded incomplete information systems as follows.

Definition 5 Let $S^{'}$ be an expanded incomplete information system, $A \subseteq$ AT, the accuracy measure of rough set $X \subseteq$ U is defined as:

$$\alpha_A(X) = F(A_*(X)) / F(A^*(X)) = \sum_{y \in A_*(X)} f^{'}(y) / \sum_{y \in A^*(X)} f^{'}(y) \quad (5)$$

Theorem 1 Let $S^{'}$ be an expanded incomplete information system and $X \subseteq$ U, if $A \subseteq B \subseteq$ AT, then $\alpha_A(X) \leq \alpha_B(X)$.

Proof For any x $\in$ U, $[x]_{R_A} \supseteq [x]_{R_B}$. According to the definition of lower approximation, $A_*(X) \subseteq B_*(X)$ holds. That is, $F(A_*(X)) \leq F(B_*(X))$. Similarly, it is easy to prove that $F(A^*(X)) \geq F(B^*(X))$. To sum up, we get $\alpha_A(X) \leq \alpha_B(X)$.

Definition 6 Let $S^{'}$ be an expanded incomplete information system, $A \subseteq$ AT, then the rough entropy of the knowledge A, denoted by E(A), is defined as:

$$E(A) = 1/F(U) \sum_{x \in U} F([x]_{R_A}) \log(1/F([x]_{R_A})) \quad (6)$$

Theorem 2 Let $S^{'}$ be an expanded incomplete information system, if $A \subseteq B \subseteq$ AT, then E(A) $\geq$ E(B) holds.

Proof For any $x \in U$, $[x]_{R_A} \supseteq [x]_{R_B}$, that is, $F([x]_{R_B}) \leq F([x]_{R_A})$ and $-\log(1/F([x]_{R_B})) \leq -\log(1/F([x]_{R_A}))$. Based on the formation of entropy

in Definition 6, $\frac{F([x]_{R_B})}{F(U)}(-\log(1/F([x]_{R_B}))) \le \frac{F([x]_{R_A})}{F(U)}(-\log(1/F([x]_{R_A}))).$

Extending this inequality, $E(A) \geq E(B)$.

Definition 7 Let S' be an incomplete information system and $A \subseteq AT$, then the rough entropy of $X \subseteq U$ about knowledge **A**, denoted by $E_A(X)$, is defined as:

$$E_A(X) = E(X) + (1-\alpha_A(X)) \quad (7)$$

or $$E_A(X) = E(X) * (1-\alpha_A(X)) \quad (8)$$

Clearly, $A \subseteq B \subseteq AT$ means that U/R_A has a small level of granularity than U/R_A. It is not difficult to prove that $E_A(X) \geq E_B(X)$ if $A \subseteq B \subseteq AT$, no matter what kind of definition of rough entropy of X about knowledge.

The accuracy measure, rough entropy of knowledge and rough entropy of a set will be degenerated to the traditional concepts in incomplete information systems if all objects have the same significance. In other words, for any $x \in U$, $f'(x)$ is a constant. Besides, those basic concepts in expanded incomplete information systems are also monotonously changing, while the granularity level of the classification on the universe is changing monotonously.

It is easy to prove the following properties in expanded incomplete information system:

$$F(A_*(X)) \le F(X) \le F(A^*(X)) \quad (9)$$

$$F(A_*(\varnothing)) = F(\varnothing) = F(A^*(\varnothing)) = 0 \quad (10)$$

$$F(A_*(U)) = F(U) = F(A^*(U)) \quad (11)$$

$$F(A_*(X \cup Y)) \geq F(A_*(X) \cup A_*(Y))) \quad (12)$$

$$F(A_*(X \cap Y)) \le F(A_*(X) \cap A_*(Y))) \quad (13)$$

3. Knowledge dependency

Using classification, we can analyze dependencies between two subsets of attributes[8]. For an expanded incomplete information system S', let $x, y \in U$ and $A \subseteq AT$. We denote $(x, y) \in R_A$ if and only if $(x, y) \in R_a$ for all $a \in A$.

Definition 8 Let S' be an expanded incomplete information system, a knowledge dependency between subsets of attributes $A, B \subseteq AT$, is denoted by $A \to B$ which holds in the information system S' if and only if, for every $x, y \in U$, which have that $(x, y) \in R_A$ implies $(x, y) \in R_B$.

The partial dependency of knowledge means that reasoning between knowledge could be partially. In other words, part of the knowledge B could be reasoned by A and the partial reasonability could be represented by positive space of knowledge.

Definition 9 In an incomplete information system without considering the significance of objects, $A, B \subseteq AT$, the positive space determined by A with respect to B on universe U, denoted by POS(A, B), is defined as follows:

$$POS(A,B) = \cup\{A_*(X) : X \in U / R_A\} \quad (14)$$

Definition 10 Let S' be an expanded incomplete information system and $A, B \subseteq AT$. Knowledge B depends in degree k from knowledge A, denoted by $A \xrightarrow{k} B$, where $k \in [0, 1]$ and is defined as follows:

$$k = \frac{\sum_{x \in POS(A,B)} f'(x)}{\sum_{x \in U} f'(x)} = \frac{F(POS(A,B))}{F(U)} \quad (15)$$

Clearly, when $f'(x)$ is a constant, k is the traditional degree of knowledge dependency. If $A \xrightarrow{0} B$, then we can say that A, B are independent; if $A \xrightarrow{1} B$, then we can simply write $A \to B$.

Definition 11 Let S' be an expanded incomplete information system, an identity dependency between knowledge $A, B \subseteq AT$ is a statement, denoted by $A \leftrightarrow B$, which holds in a information system if and only if $A \to B$ and $B \to A$.

Lemma Let S' be an expanded incomplete information system, sets of attributes $A, B, C \subseteq AT$, then we have properties:

$$POS(A, B) \subseteq POS(A \cup C, B) \quad (16)$$

$$POS(A, B) \supseteq POS(A, B \cup C) \quad (17)$$

$$POS(A, B \cup C) \subseteq POS(A \cup C, B) \quad (18)$$

$$POS(B,C) = U \Rightarrow POS(A,B) \subseteq POS(A,C) \quad (19)$$

$$A \subseteq B \Rightarrow POS(A, C) \subseteq POS(B, C) \quad (20)$$

Proof For any $x \in U$, we have $[x]_{R_A} \supseteq [x]_{R_{A \cup C}}$. For any $W \in U / R_B$, if $[x]_{R_A} \subseteq W$, then $[x]_{R_{A \cup C}} \subseteq W$. Conversely, $[x]_{R_{A \cup C}} \subseteq W$ does not mean $[x]_{R_A} \subseteq W$.

So it is clear that POS(A, B)$\subseteq$POS(A$\cup$C, B). That is, formula (17) is held.

For any x$\in$U, we have $[x]_{R_B} \supseteq [x]_{R_{B\cup C}}$. For any W$\in U/R_A$, if W$\subseteq [x]_{R_{B\cup C}}$, then W$\subseteq [x]_{R_B}$. But W$\subseteq [x]_{R_B}$ does not mean W$\subseteq [x]_{R_{B\cup C}}$. So from the above discussed, it is clear that POS(A, B)$\supseteq$POS(A, B$\cup$C).

According to formula (16) and (17), we have POS(A, B)$\subseteq$POS(A$\cup$C, B) and POS(A, B) $\supseteq$POS(A, B$\cup$C). Then POS(A, B$\cup$C) $\subseteq$POS(A, B)$\subseteq$POS(A$\cup$C, B). This means that formula (18) is also right.

Suppose x$\in$POS(A, B). Then there exists W $\in U/R_B$ such that $[x]_{R_B} \subseteq W$. Owing to POS(B,C)= U, we have x$\in$POS(B,C) and there exists V $\in U/R_C$ such that W $\subseteq$ V. In other words, $[x]_{R_B} \subseteq W \subseteq V$. Therefore, x$\in$POS(A,C). Since x is arbitrary, POS(A, B)$\subseteq$POS(A,C). That is to say, formula (19) is out of question.

Suppose x $\in$ POS(A,C). That is, there exists W$\in U/R_C$ such that $[x]_{R_A} \subseteq W$. Owing to A$\subseteq$B, we have $[x]_{R_A} \supseteq [x]_{R_B}$. Thus $[x]_{R_B} \subseteq W$. In other words, x$\in$POS(B, C). Furthermore, POS(A,C)$\subseteq$POS(B,C), and then formula (20) is also okay.

Corollary 1 Let S' be an expanded incomplete information system and sets of attributes A, B, C$\subseteq$AT, then we have the following relationships:

(1) if $A \xrightarrow{k_1} B$ and $A\cup C \xrightarrow{k_2} B$, then $k_1 \le k_2$; (2) if $A \xrightarrow{k_1} B$ and $A \xrightarrow{k_2} B\cup C$, then $k_1 \ge k_2$; (3) if $A \xrightarrow{k_1} B\cup C$ and $A\cup C \xrightarrow{k_2} B$, then $k_1 \le k_2$; (4) if $A \xrightarrow{k_1} B$, $B \to C$, then $A \xrightarrow{k_2} C$, and $k_1 \le k_2$; (5) if $A \subseteq B$, $B \xrightarrow{k_1} C$, then $A \xrightarrow{k_2} C$, and $k_1 \ge k_2$.

Proof (1) Following the formula (16), POS(A, B)$\subseteq$POS(A$\cup$C,B) holds. So $\sum_{x\in POS(A,B)} f'(x) \le \sum_{x\in POS(A\cup C,B)} f'(x)$ is also valid. That is, $k_1 \le k_2$.

(2) According to formula (17), POS(A,B) $\supseteq$ POS(A,B $\cup$ C) holds. So $\sum_{x\in POS(A,B)} f'(x) \ge \sum_{x\in POS(A,B\cup C)} f'(x)$. That is, $k_1 \ge k_2$.

(3)Following formula (18), POS(A,B $\cup$ C) $\subseteq$POS(A$\cup$C,B) holds. So $\sum_{x\in POS(A,B\cup C)} f'(x) \le \sum_{x\in POS(A\cup C,B)} f'(x)$. That is, $k_1 \le k_2$.

(4) Owing to formula (19), we have POS(B, C)=U $\Rightarrow$ POS(A, B) $\subseteq$ POS(A, C). So $\sum_{x\in POS(A,B)} f'(x) \le \sum_{x\in POS(A,C)} f'(x)$. That is $k_1 \le k_2$.

(5) Due to formula (20), we have A$\subseteq$B implies POS(A,C) $\subseteq$ POS(B, C). So $\sum_{x\in POS(B,C)} f'(x) \le \sum_{x\in POS(A,C)} f'(x)$. That is, $k_1 \ge k_2$.

What have been discussed above is around the knowledge dependency in information systems, however, there are also knowledge dependency (for subset of decision) in decision systems. Reference [8] discussed the functional dependency (for subset of decision) and relational algorithms in complete decision systems. Naturally, it is useful to extend this concept in expanded incomplete decision systems.

Definition 12 An expanded incomplete decision system is an expanded incomplete information system $D' =< U,C\cup D,V,f,V',f' >$, where C is the set of condition attributes and D is the set of decision attributes such that C$\cap$D = $\varnothing$ and for any d$\in$D and x$\in$U, f(d, x) $\ne *\wedge |f(d)|=1$ in which $|T|$ represents the cardinality of set T.

Definition 13 A knowledge dependency (for a subset Y$\subseteq$D of decision attributes) between knowledge A, B$\subseteq$C is a statement, denoted by A$\to$B(Y), which holds in the decision system if and only if F (POS(A, Y)) $\ge$ F (POS(B, Y)), i.e., for $A \xrightarrow{k_1} Y$ and $B \xrightarrow{k_2} Y$, there must be $k_1 \ge k_2$.

Definition 14 Let $D' =< U,C\cup D,V,f,V',f' >$ be an expanded incomplete decision system and A, B$\subseteq$C, Y$\subseteq$D, that knowledge dependency $A \xrightarrow{k_1} Y$ has more strength than $B \xrightarrow{k_2} Y$ in degree *l*, denoted by

$$A \xrightarrow{l} B(Y), \qquad (21)$$

It means $l = k_1 - k_2$

$$= \frac{\sum_{x\in POS(A,Y)} f'(x) - \sum_{x\in POS(A,Y)} f'(x)}{\sum_{x\in U} f'(x)}$$

$$= \frac{F(POS(A,Y)) - F(POS(B,Y))}{F(U)}$$

(22)

What should be noticed is that the meaning of *l* is different from *k* in Definition 10. *l*=0 means that $A \xrightarrow{k_1} Y$ and $B \xrightarrow{k_2} Y$ have the same degree of knowledge dependency, while *l*=1 means that $A \to Y$ and $B \not\to Y$.

Corollary 2 Let $D' = \langle U, C \cup D, V, f, V', f' \rangle$ be an expanded incomplete decision system and A, B, $G \subseteq C$, $Y \subseteq D$, then we have the following relations:

$$i) A \supseteq B \Rightarrow A \to B(Y) \quad (23)$$

$$ii) A \xrightarrow{k_1} B(Y) \Rightarrow A \xrightarrow{k_2} B \cup G(Y), k_1 \geq k_2 \quad (24)$$

$$iii) A \xrightarrow{k_1} B(Y) \Rightarrow A \cup G \xrightarrow{k_2} B(Y), k_1 \leq k_2 \quad (25)$$

$$iv) A \to B(Y), B \to G(Y) \Rightarrow A \to G(Y). \quad (26)$$

Proof $i)$ Following formula (20), we have POS(A,Y) $\supseteq$ POS(B,Y) because A $\supseteq$ B. That is, F (POS(A, Y)) $\geq$ F (POS(B, Y), which meets the condition in Definition 13, so we have $A \supseteq B \Rightarrow A \to B(Y)$.

$ii)$ According to formula (16), we have POS(B,Y) $\subseteq$ POS(B $\cup$ G,Y). That is,

$$\frac{F(POS(A,Y)) - F(POS(B,Y))}{F(U)} \geq \frac{F(POS(A,Y)) - F(POS(B \cup G,Y))}{F(U)}$$

and then $k_1 \geq k_2$.

$iii)$ From formula (16), we have POS(B,Y) $\subseteq$ POS(B $\cup$ G,Y). That is,

$$\frac{F(POS(A,Y)) - F(POS(B,Y))}{F(U)} \leq \frac{F(POS(A \cup G,Y)) - F(POS(B,Y))}{F(U)}$$

and then $k_1 \leq k_2$.

$iv)$ From condition, F(POS(A, Y)) $\geq$ F(POS(B,Y)) and F(POS(B,Y)) $\geq$ F (POS(G, Y)) imply F(POS(A,Y)) $\geq$ F(POS(C,Y)) and as a result we have $A \to B(Y),\ B \to G(Y) \Rightarrow A \to G(Y)$.

Clearly, formula (26) tells us that knowledge dependency (for a subset of decision) in expanded incomplete information system is of transitivity.

4. Conclusions

The expanded incomplete information system, in other words, incomplete information system with significance of objects is more suitable to meet the needs of actual researches. Owing to the information function f. being added in the information system, so many basic concepts of RST that mentioned in section 2 should be modified. Some properties studied in section 3 help us to understand the measurement about the knowledge dependency and the knowledge dependency for decision deeply and thoroughly, especially suitable to the dynamic changing of knowledge in the information and decision systems. In the future, other methods such as probability and conditional entropy[9] which can measure knowledge dependencies will be our keynotes.

References

[1] Pawlak Z. Rough set theory and its applications to data analysis. Journal of Cybernetics and Systems, 1998(29):661–688.

[2] Pawlak Z. Rough sets and intelligent data analysis. Journal of Information Sciences, 2002(147):1–12.

[3] Kryszkiewicz M., Rough set approach to incomplete information systems. Journal of Information Sciences, 1998(112):39–49.

[4] Jerzy S., Incomplete information tables and rough classification. J. of Computational Intelligence,2001(17):545–566.

[5] Guoyin. Wang, Extension of rough set under incomplete information system. Journal of Computer Research and Development, 2002(39):1238–1243.

[6] Xianghui Chen, Shanjun Zhu, Yindong Ji, Yongmin Li.Generalized rough set model and its uncertainty measure. Journal of Tsinghua University, 2002(42):128 –131.

[7] Y.Y.Yao,Information granulation and rough set approximation. International Journal of Intelligent Systems, 2001(16):87–104.

[8] D. A. Bell, J. W. Guan. Computational methods for rough classification and discovery. Journal of the American Society for Information Science, 1998(49):403–414.

[9] Y.Y.Yao,Probabilistic Approaches to Rough Sets. Expert Systems. 2003(20):287 –297.