Dynamic Reduction Based on Rough Sets in Incomplete Decision Systems

Dayong Deng[1,2] and Houkuan Huang[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, PR China, 100044
dayongd@163.com, hkhuang@center.njtu.edu.cn
[2] Zhejiang Normal University, Jinhua, Zhejiang Province, PR China, 321004

Abstract. In this paper we investigate the dynamic characteristics in an incomplete decision system while information is increasing. We modify the definition of reduction of condition attributes in this case, and present algorithms of reduction in order to deal with increase information.

Keywords: Rough Sets, Incomplete Decision, Increase Information, Attribute Reduction.

1 Introduction

When collecting information about a given topic in a certain moment in time, it may happen that we do not exactly know all the details of the issue in question. This lack of knowledge leads to an incomplete information system. Rough set theory is a valid mathematical tool, which deals with imprecise, vague and incomplete information[10,11,12]. In general, rough set theory deals with information in complete information systems. But in recent years, there are many people, who disposed incomplete information with rough set theory, and presented several methods of dealing with missing attribute values[3]. M.Kryszkiewicz[1,2] proposed a tolerance relation between objects, which is reflexive, symmetric but not transitive. J.Stefanowskis' model[6,7] is based on similarity relation, which is reflexive, transitive, but not symmetric. J.W.Grzymala-Busse's model[4,5] is based on characteristic relation, which is only reflexive. S.Greco's model[8] is based on similarity relation, which is transitive, but not reflexive or symmetric. G.Y.Wang[9] extended M.Kryszkiewicz's model so that his model fits real world more. But all of indiscernibility relations in these models are static, and they could not fit the case of increase information in incomplete information system. In the paper we consider the case of incomplete decision systems with increase information based on M.Kryszkiewicz's model. We investigate its dynamic properties and present new algorithms to get reducts while information is increasing. The method of reduction preserves their positive regions as well as other important information in these incomplete information systems such that the reducts are not put at a disadvantage when information is increasing.

The rest of the paper is organized as follows. In section 2 we introduce the basic concepts of incomplete information systems. In section 3 we investigate

J.T. Yao et al. (Eds.): RSKT 2007, LNAI 4481, pp. 76–83, 2007.

their dynamic properties. In section 4 we present new algorithms of dynamic reduction. In section 5 an example is given to show the ideas of new algorithms. At last, we draw a conclusion in section 6.

2 Information Systems

An information system is a pair $IS = (U, A)$, where U is the universe of discourse with a finite number of objects(or entities), A is a set of attributes defined on U. Each $a \in A$ corresponds to the function $a : U \to V_a$, where V_a is called the value set of a. Elements of U are called situation, objects or rows, interpreted as, e.g., cases, states.

With any subset of attributes $B \subseteq A$, we associate the information set for any object $x \in U$ by

$$Inf_B(x) = \{(a, a(x)) : a \in B\}$$

An equivalence relation called B-indiscernible relation is defined by

$$IND(B) = \{(x, y) \in U \times U : Inf_B(x) = Inf_B(y)\}$$

Two objects x, y satisfying the relation $IND(B)$ are indiscernible by attributes from B. $[x]_B$ is referred to as the equivalence class of $IND(B)$ defined by x. The equivalence classes of $IND(B)$ are denoted by

$$U/B = \{[x]_B : x \in U\}.$$

A minimal subset B of A such that $IND(B) = IND(A)$ is called a reduct of IS.

Suppose $IS = (U, A)$ is an information system, $B \subseteq A$ is a subset of attributes, and $X \subseteq U$ is a subset of discourse, the sets

$$\underline{B}(X) = \{x \in U : [x]_B \subseteq X\}, \overline{B}(X) = \{x \in U : [x]_B \bigcap X \neq \phi\}$$

are called B-lower approximation and B-upper approximation respectively. The lower approximation is also called positive region, denoted by $POS_B(X)$.

A special type of information system is called decision system $DS = (U, A \cup \{d\})$, where $\{d\} \cap A = \phi$, A is a set of condition attributes, and d is a distinguished attribute called conclusion attribute. In a decision system the positive region of the decision attribute corresponding to the condition attributes is denoted by $POS_A(d)$:

$$POS_A(d) = \bigcup_{Y_i \in U/\{d\}} (POS_A(Y_i))$$

It may happen that some values of attributes for objects in information systems are missing. These information systems are called incomplete information systems. The missing values are called null values, which are denoted by *. Therefore, a similarity relation could be defined as follows[1,2]:

$$SIM(B) = \{(x, y) \in U \times U : \forall a \in B(a(x) = a(y) \; or \; a(x) = * \; or \; a(y) = *)\}$$

Let $S_B(x)$ denotes the object set $\{y \in U : (x, y) \in SIM(B)\}$, where $B \subseteq A$. The lower and upper approximation of a concept $X \subseteq U$ are defined as follows respectively:

$$\underline{B}(X) = \{x \in U : S_B(x) \subseteq X)\}$$

$$\overline{B}(X) = \{x \in U : S_B(x) \bigcap X \neq \emptyset\}$$

If there is not confusion, we will also denote the set of tolerance classes $S_B(x)$ by U/B, and the B-lower approximation of X is also called the positive region, denoted by $POS_B(X)$.

For $B \subseteq A$, $C \subseteq A$, we call the cover of U/B is finer than that of U/C, denoted by $U/B \subseteq U/C$, if for any tolerance class $S_B(x)$ in U/B there exists a tolerance class $S_C(x)$ in U/C such that $S_B(x) \subseteq S_C(x)$.

In incomplete decision systems, we assume that values of conclusion attributes are usually complete in the sequel.

3 Incomplete Systems with a Monotonic Increase of Information

In [13] G.Cattaneo and D.Ciucci defined three ways of increasing the knowledge in incomplete information systems. In this paper we are only dealing with the first case. Its definition is formalized in the following way.

Definition 1. Let $IS^{(t_i)} = (U_i, A_i)$ and $DS^{(t_{i+1})} = (U_{i+1}, A_{i+1})$, with t_i, $t_{i+1} \in R$, $t_i \leq t_{i+1}$ be two incomplete information systems, where $U_i = U_{i+1}$. The attributes in A_i are the same as that in A_{i+1}. We will say that there is a monotonic increase of information in the information system IS: For $\forall x \in U_i$ and $\forall a^{t_i} \in A_i$, $a^{t_i}(x) \neq *$ implies $a^{t_i}(x) = a^{t_{i+1}}(x)$. In such a case, we will denote by $IS^{(t_i)} \preceq_1 IS^{(t_{i+1})}$.

$IS^{(t_i)} \preceq_1 IS^{(t_{i+1})}$ means that, in the information system IS the universe of discourse and the attributes do not change, but the values of attributes may be changed from unknown to known. Because we only investigate this case, we will denote U_i by U in the sequel.

Definition 2. Let $IS^{(t_i)} = (U, A_i)(t_i \in R)$ be a series of incomplete information systems with a monotonic increase of information, i.e. $IS^{(t_i)} \preceq_1 IS^{(t_{i+1})}$. We say the information system IS is a complete information system if it satisfies the condition:

$$IS = \lim_{i \to \infty} IS^{(t_i)}$$

From definition 2, there are two types of complete information systems corresponding to a series of incomplete information systems with a monotonic increase of information(complete information systems, in short): (1) All of values of attributes are known. (2) Some values of attributes will be unknown forever. Without generality we assume that all of values of attributes are known in complete information systems.

We will investigate properties of incomplete information systems with a monotonic increase of information in the sequel.

Proposition 1. Suppose $IS^{(t_i)} = (U, A_i) \preceq_1 IS^{(t_{i+1})} = (U, A_{i+1})$, with t_i, $t_{i+1} \in R$, $t_i \leq t_{i+1}$ be two incomplete information systems, Then for $\forall a \in A_i$ and $\forall x \in U$, we have

$$S_{\{a\}}^{t_{i+1}}(x) \subseteq S_{\{a\}}^{t_i}(x)$$

Proof. In terms of the definition $S_{\{a\}}(x) = \{y \in U : (x, y) \in SIM(\{a\})\}$, we have $\forall y(y \in S_{\{a\}}^{t_{i+1}}(x) \Rightarrow y \in S_{\{a\}}^{t_i}(x))$. Therefore $S_{\{a\}}^{t_{i+1}}(x) \subseteq S_{\{a\}}^{t_i}(x)$.

Corollary 1. Suppose $IS^{(t_i)} = (U, A_i) \preceq_1 IS^{(t_{i+1})} = (U, A_{i+1})$, with t_i, $t_{i+1} \in R$, $t_i \leq t_{i+1}$ be two incomplete information systems, then $U/B_{i+1} \subseteq U/B_i$ for $B \subseteq A$.

Corollary 2. Suppose $IS^{(t_i)} = (U, A_i) \preceq_1 IS^{(t_{i+1})} = (U, A_{i+1})$, with t_i, $t_{i+1} \in R$, $t_i \leq t_{i+1}$ be two incomplete information systems, $IS = (U, A)$ their corresponding complete information system and $X \subseteq U$ a concept. Then

$$\underline{B_i}(X) \subseteq \underline{B}_{i+1}(X) \subseteq \underline{B}(X)$$

$$\overline{B}(X) \subseteq \overline{B}_{i+1}(X) \subseteq \overline{B}_i(X)$$

for $\forall B \subseteq A$.

Theorem 1. Suppose $DS^{(t_i)} = (U, A_i \bigcup \{d\}) \preceq_1 DS^{(t_{i+1})} = (U, A_{i+1} \bigcup \{d\})$, with t_i, $t_{i+1} \in R$, $t_i \leq t_{i+1}$, be two incomplete decision systems, Then

$$POS_{B_i}(\{d\}) \subseteq POS_{B_{i+1}}(\{d\})$$

for $B \subseteq A$.

Proof. It can be got directly from Corollary 2.

Corollary 3. Suppose $DS^{(t_i)} = (U, A_i \bigcup \{d\})$ is an incomplete decision system, $DS = (U, A \bigcup \{d\})$ is its corresponding complete decision system, Then

$$POS_{B_i}(\{d\}) \subseteq POS_B(\{d\})$$

for $B \subseteq A$.

From the above propositions, the positive regions in incomplete decision systems are increasing with increasing information in them. We should not delete any of condition attributes unless these condition attributes are confirmed not to influence the positive regions in the series of incomplete decision systems. In the next section we will investigate reduction of condition attributes in incomplete decision systems.

4 Dynamic Reduction

In decision systems with missing values almost all of existed methods are to get reducts in the criterion of the positive regions preserved, these methods don't consider dynamic increase of information. In this section we will investigate reduction in this case. The criterion of reduction, except for preserved positive region, is to delete condition attributes in which there are no null values in the negative positive region, i.e. all of elements with missing value are in the positive region. It is easy to prove that these deleted condition attributes are irrelative to the positive regions in incomplete decision systems with increase information in terms of above propositions. In terms of the criteria, the algorithm of reduction in an incomplete decision system is presented as follows:

Algorithm 1: Static reduction of incomplete decision system(SRIDS, In short).

Input: An incomplete decision system $DS^{(t_i)} = (U, A_i \bigcup \{d\})$
Output: A reduct of $DS^{(t_i)} = (U, A_i \bigcup \{d\})$
Step1: $U_1 = POS_{A_i}(\{d\})$, $B = A_i$
Step2: $U_2 = U - U_1$
Step3: For j=1 to $|A_i|$
{ $flag = 1$;
For k=1 to $|U_2|$
If $a_j(x_k) = *$ Then $flag = 0$;
If flag and $POS_{B-\{a_j\}}(\{d\}) = U_1$
Then $B = B - \{a_j\}$; }
Step4: Output the reduct B

In algorithm 1, $DS^{(t_i)}$ represents the state of the decision system DS at t_i. The symbol $flag$ is to decide whether any elements in U_2(it stands for the negative positive region) are missing values, if $flag = 1$ then there are no null values in the negative positive region, or else there are some null values. $|\bullet|$ denotes the cardinality of the set, and x_k is an element of U_2.

The difference between algorithm 1 and other classical algorithms is whether the missing values in the negative positive region should be considered when condition attributes are reduced. The former considers the null values of condition attributes in order to avoid a disadvantage for the reduced condition attributes to the positive region in the future. The later only consider the positive region at a moment.

The time complexity of algorithm 1 is decided by that of counting positive region. Suppose we utilize the algorithm in literature [14] to compute positive region, whose time complexity is $O(|A_i||U|log|U|)$. Therefore, the time complexity of algorithm 1 is $O(|A_i|^2|U|log|U|)$.

Suppose the incomplete decision system $DS^{(t_i)} = (U, A_i \bigcup \{d\})$ is dynamically increasing its information, i.e. $DS^{(t_i)} \preceq_1 DS^{(t_{i+1})}$, with t_i, $t_{i+1} \in R$, $t_i \leq t_{i+1}$, and the maximum of i is equal to n. In this case we could call the above algorithm iteratively. The algorithm of reduction with respect to the dynamical incomplete decision system $DS^{(t_i)}$ is presented as follows:

Algorithm 2: Dynamical reduction of an incomplete decision system with increase information

Input: An incomplete decision system $DS^{(t_i)} = (U, A_i \bigcup \{d\})$ with increase information.

Output: A dynamic reduct with respect to the incomplete decision system DS

For i=1 to n
{ SRIDS($DS^{(t_i)}$);
$DS^{(t_i)} = (U, B)$;
}

B denotes a reduct of the incomplete decision system DS at t_i. Because the time complexity of algorithm 1 is $O(|A_i|^2|U|log|U|)$, the time complexity of algorithm 2 is $O(n|A_i|^2|U|log|U|)$. The Algorithm 2 counts the positive region iteratively. We could improve it by avoiding the iterative workload. The improved algorithm is presented as follows:

Algorithm 3: Improved dynamic reduction of an incomplete decision system.

Input: An incomplete decision system $DS^{(t_i)} = (U, A_i \bigcup \{d\})$ with increase information.

Output: A dynamic reduct with respect to the incomplete decision system DS

Step 1: $B = A$; $U_1 = \emptyset$;
Step 2: $U_2 = U$;
Step 3: For i=1 to n
{$U_3 = POS_B(\{d\})$;
$U_1 = U_1 \bigcup U_3$;
$U_2 = U_2 - U_3$;
$C = B$;
For j=1 to $|B|$
{ $flag = 1$;
For k=1 to $|U_2|$
If $a_j(x_k) = *$ then $flag = 0$;
If flag and $U_1 = POS_{C-\{a_j\}}(\{d\})$ then $C = C - \{a_j\}$;
} //End for j
Output C;
B=C;
} // End for i

In algorithm 3, U_1 denotes positive region of the incomplete decision system, U_2 negative positive region, U_3 the incremental positive region in the rest of the incomplete decision system, $POS_B(\{d\})$ the positive region of incomplete decision system $DS' = (U_2, B \bigcup \{d\})$, and $POS_{C-\{a_j\}}(\{d\})$ the positive region

of incomplete decision system $DS'' = (U, (C - \{a_j\}) \bigcup \{d\})$. The output value of C is a reduct at t_i. The rest of symbols are the same as that of algorithm 1.

The algorithm 3 could reduce the the iterative workload of computing positive region. In algorithm 3 we could only compute the additive positive region $POS_B(\{d\})$, not the whole positive region of the decision system, although the time complexity of algorithm 3 is the same as that of algorithm 2.

5 Example

Suppose the incomplete decision system $DS^{(t_i)} = (U, A_i \bigcup \{d\})$ is dynamically increasing its information, where $DS^{(t_1)}$ is denoted by Table 1, $DS^{(t_2)}$ is denoted by Table 2, and $DS^{(t_1)} \preceq_1 DS^{(t_2)}$, $A = \{a_1, a_2, a_3\}$. In $DS^{(t_1)}$ $a_1(x_3) = *$, but $a_1(x_3) = 2$ in $DS^{(t_2)}$. In terms of Algorithm 1 we could get the reduct $\{a_1, a_2\}$ at t_1. In table 1, although the attributes a_1 and a_3 could be deleted if we only preserve the positive region, there are some missing values of a_1 in the negative positive region, while there are no missing values of a_3 in the negative positive region. At t_2 we could get more elements in the positive region. For example, the element x_3 is not in the positive region at t_1, but it is in the positive region at t_2. It is easy to know the reduct of incomplete information DS at t_2 is also $\{a_1, a_2\}$ from table 2 in term of algorithm 1. That is to say, the condition attribute a_1 should not be reduced at t_1 in the incomplete decision system with increase information.

Table 1. Incomplete Decision system DS at t_1

U	a1	a2	a3	d
x1	0	0	1	1
x2	1	1	*	1
x3	*	0	1	0
x4	0	2	1	0
x5	0	0	1	1
x6	3	1	*	1
x7	3	2	1	0
x8	*	2	1	0

Table 2. Incomplete Decision system DS at t_2

U	a1	a2	a3	d
x1	0	0	1	1
x2	1	1	*	1
x3	2	0	1	0
x4	0	2	1	0
x5	0	0	1	1
x6	3	1	*	1
x7	3	2	1	0
x8	*	2	1	0

6 Conclusion

In the paper we investigate some properties of incomplete decision systems with increase information, and the reduction of condition attributes in this case. A new method of reduction is presented, in which we not only consider positive region in an incomplete decision system but also its potential influence on positive region in the future.

References

1. Kryszkiewicz, M.: Rough Set Approach to Incomplete Information Systems. Information Science1 112(1-4) (1998) 39-49.
2. Kryszkiewicz, M.: Rules in Incomplete Information Systems. Information Science 113(3-4) (1999) 271-292.
3. Grzymala-Busse, J.W., Hu, M.: A Comparison of Several Approaches to Missing Attribute Values in Data Mining. In: Proceedings of 2nd International Conference on Rough Sets and Current Trends in Computing(RSCTC2000), Canada (2000) 378-385.
4. Grzymala-Busse, J.W.: Characteristic Relations for Incomplete Data:A Generalization of the Indiscernibility Relation. In:Proceedings of 4th International Conference on Rough Sets and Current Trends in Computing(RSCTC2004), Sweden (2004) 254-263.
5. Grzymala-Busse, J.W.: Incomplete Data and Generalization of Indiscernibility Relation, Definability, and Approximation. In:Proceedings of 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSRDGrC2005), Canada (2005) 244-253.
6. Stefanowski, J., Tsoukis, A.: On the Extension of Rough Sets under Incomplete Information. In:Proceedings of 7th International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing(RSRDGrC1999), Japan (1999) 73 - 81.
7. Stefanowski, J., Tsoukis, A.: Incomplete Information Tables and Rough Classification. Computational Intelligence 17(3) (2001) 545-566.
8. Greco, S.,Matarazzo, B., Slowinski, R.: Dealing with Missing Data in Rough Set Analysis of Multi-attribute and Multi-criteria Decision Problems. In Decision Making:Recent developments and Worldwide Applications,ed. by S.H.Zanakis, G.Doukidis, and Z.Zopounidis, Kluwer Academic Publishers, Dordrecht (2000) 295-316.
9. Wang, G.Y.: Extension of Rough Set Under Incomplete Information Systems. Journal of Computer Research and Development(in Chinese) 39(10) (2002) 1238-1243.
10. Wang,G.Y.: Rough Set Theory and Knowledge Discovery(in Chinese). Xi'an Jiaotong University Press (2001).
11. Pawlak, Z.: Rough sets-Theoretical Aspect of Reasoning about Data. Kluwer Academic (1991).
12. Liu, Q.: Rough Sets and Rough Reasoning. Science Press(in Chinese) (2001).
13. Cattaneo, G., Ciucci, D.: Investigation about Time Monotonicity of Similarity and Preclusive Rough Approximations in Incomplete Information Systems. In:Proceedings of 4th International Conference, RSCTC2004, Uppsala, Sweden (2004) 38-48.
14. Liu,S.H.,Sheng,Q.J.,Shi,Z.Z.:A New Method for Fast Computing Positive Region. Journal of Computer Research and Development(in Chinese) 40(5) (2003) 637-642.