International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems
Vol. 10, No. 1 (2002) 95–103

THE ALGORITHM ON KNOWLEDGE REDUCTION IN INCOMPLETE INFORMATION SYSTEMS

JIYE LIANG

Department of Computer Science, Shanxi University
Taiyuan, 030006, People's Republic of China
jyliang2000@yahoo.com

ZONGBEN XU

Institute for Information and System Science
Faculty of Science, Xi'an Jiaotong University
Xi'an, 710049, People's Republic of China
zbxu@xjtu.edu.cn

Received January 2001
Revised December 2001

Rough set theory is emerging as a powerful tool for reasoning about data, knowledge reduction is one of the important topics in the research on rough set theory. It has been proven that finding the minimal reduct of an information system is a NP-hard problem, so is finding the minimal reduct of an incomplete information system. Main reason of causing NP-hard is combination problem of attributes. In this paper, knowledge reduction is defined from the view of information, a heuristic algorithm based on rough entropy for knowledge reduction is proposed in incomplete information systems, the time complexity of this algorithm is $O(|A|^2|U|)$. An illustrative example is provided that shows the application potential of the algorithm.

Keywords: Rough sets; incomplete information systems; knowledge reduction; complexity of algorithm.

1. Introduction

Rough sets theory, introduced by Pawlak[1,2], is emerging as a powerful tool for reasoning about data, it has been one of the important approach for data analysis[3-6]. It provides techniques to reduce knowledge in database, by which the irrelevant or superfluous knowledge (attributes) can be eliminated according to the learning task without losing essential information about the original data in the databases. As a result of the knowledge reduction, set of concise and meaningful rules are produced.

As is well known that an information system may usually have more than one reduct[1,5]. This means the set of rules derives from knowledge reduction is not unique. In practice, it is always hoped to obtain the set of the most concise rules.

Therefore, people have been attempting to find the minimal reduct of information systems, which means that the number of attributes contained in the reduction is minimal. Unfortunately, it has been proven that finding the minimal reduct of an information system is a NP-hard problem [7]. Similarly, an incomplete information system may usually have more than one reduct, and finding the minimal reduct of an incomplete information system is a NP-hard problem too.

Main reason of causing NP-hard is combination problem of attributes. General method to solve this class problem is using heuristic search in artificial intelligence. In this paper, first, rough entropy of knowledge is defined in incomplete information systems, by which we can analyze the significance of every attribute, and regard it as heuristic information in order to decrease search space. Then, we define knowledge reduction from the view of information. Based on these, a heuristic algorithm for knowledge reduction is proposed, and the complexity of the algorithm is analyzed. To illustrate this algorithm, a running example is presented.

2. Incomplete Information Systems

Information system is a pair $S = (U, A)$, where:

(1) U is a non-empty finite set of objects;
(2) A is a non-empty finite set of attributes;
(3) for every $a \in A$, there is a mapping a, $a : U \to V_a$, where V_a is called the value set of a.

Each subset of attributes $P \subseteq A$ determines a binary indiscernibility relation $IND(P)$, as follows

$$IND(P) = \{(x, y) \in U \times U | \ \forall a \in P, a(x) = a(y)\}.$$

It is easily shown that $IND(P)$ is an equivalence relation on the set U.

The relation $IND(P)$, $P \subseteq A$, constitutes a partition of U, which we will denote by $U/IND(P)$.

It may happen that some of attribute values for an object are missing. For example, in medical information systems there may exist a group of patients for which it is impossible to perform all the required tests. These missing values can be represented by the set of all possible values for the attribute or by the domain of the attribute. To indicate such a situation a distinguished value, a so-called null value is usually assigned to those attributes.

If V_a contains null value for at least one attribute $a \in A$, then S is called an incomplete information system [8-10], otherwise it is complete. Further on, we will denote null value by $*$.

Let $P \subseteq A$, we define tolerance relation:

$$SIM(P) = \{(x, y) \in U \times U | \ \forall a \in P, a(x) = a(y) \ \text{ or } \ a(x) = * \ \text{ or } \ a(y) = *\}.$$

It is easily shown that

$$SIM(P) = \bigcap_{a \in P} SIM(\{a\}).$$

Let $S_P(x)$ denote the object set $\{y \in U | (x, y) \in SIM(P)\}$. $S_P(x)$ is the maximal set of objects which are possibly indiscernible by P with x.

Let $U/SIM(P)$ denote classification, which is the family set $\{S_P(x) | x \in U\}$. Any element from $U/SIM(P)$ will be called a tolerance class or the granularity of information. Tolerance classes in $U/SIM(P)$ do not constitute a partition of U in general. They constitute a covering of U, i.e., for every $x \in U$ we have that $S_P(x) \neq \emptyset$ and $\bigcup_{x \in U} S_P(x) = U$.

Let $P,\ Q \subseteq A$.

$U/SIM(Q) = U/SIM(P)$ denotes $S_Q(x) = S_P(x)$ for every $x \in U$.

$U/SIM(Q) \subseteq U/SIM(P)$ denotes $S_Q(x) \subseteq S_P(x)$ for every $x \in U$.

$U/SIM(Q) \subset U/SIM(P)$ denotes $S_Q(x) \subseteq S_P(x)$ for every $x \in U$ and $S_Q(x) \subset S_P(x)$ for at least one $x \in U$.

Definition 2.1. Let $S = (U, A)$ be an incomplete information system, $P \subseteq A$ and $a \in P$. We will say a is dispensable in P if $U/SIM(P) = U/SIM(P - \{a\})$; otherwise a is indispensable in P. P is independent if each $a \in P$ is indispensable in P; otherwise P is dependent.

Definition 2.2. Let $S = (U, A)$ be an incomplete information system. $P \subseteq A$ is a reduct of A if P is independent and $U/SIM(P) = U/SIM(A)$.

Obviously A may have many reducts.

The set of all indispensable attributes in A will be called the core of A, and will be denoted by $core(A)$. In fact, $core(A)$ is the intersection of all reducts of A.

In this paper, incomplete information system $S = (U, A)$ be regarded as knowledge representation system $U/SIM(A)$ or knowledge A.

3. Knowledge and Rough Entropy

The concept of rough entropy has been introduced in information systems[11,12]. Now we introduce a definition of the rough entropy of knowledge in incomplete information systems.

Definition 3.1. Let $S = (U, A)$ be an incomplete information system, $P \subseteq A$. We define the rough entropy of knowledge P as

$$E(P) = -\sum_{i=1}^{|U|} \frac{|S_P(x_i)|}{|U|} \log \frac{1}{|S_P(x_i)|},$$

where $U = \{x_1, x_2, \cdots, x_{|U|}\}$, $|U|$ is cardinality of set U and $\log x$ denotes $\log_2 x$; $\frac{|S_P(x_i)|}{|U|}$ represents the probability of tolerance class $S_P(x_i)$ within the universe U, $\frac{1}{|S_P(x_i)|}$ denotes the probability of one of the values in tolerance class $S_P(x_i)$.

Property 3.1. (Cardinality) Let $S = (U, A)$ be an incomplete information system and $P,\ Q \subseteq A$. If there exists a one-to-one, onto function $h: U/SIM(P) \to U/SIM(Q)$ such that

$$|h(S_P(x_i))| = |S_P(x_i)|, \quad i = 1, 2, \cdots, |U|,$$

then

$$E(P) = E(Q).$$

Property 3.1 states that the rough entropy of knowledge are invariant with respect to difference the set of tolerance classes that are size-isomorphic.

Property 3.2. (Monotonicity) Let $S = (U, A)$ be an incomplete information system and $P,\ Q \subseteq A$. If $U/SIM(Q) \subset U/SIM(P)$, then $E(Q) < E(P)$.

Proof. Since $U/SIM(Q) \subset U/SIM(P)$, we have that $S_Q(x_i) \subseteq S_P(x_i)$ for every $x_i \in U$ and $S_Q(x_j) \subset S_P(x_j)$ for at least one $x_j \in U$ (where $|S_P(x_i)| \geq 1$ and $|S_Q(x_i)| \geq 1$ for every $x_i \in U$). Hence,

$$\frac{1}{|U|}\sum_{i=1}^{|U|}|S_Q(x_i)|\log|S_Q(x_i)| < \frac{1}{|U|}\sum_{i=1}^{|U|}|S_P(x_i)|\log|S_P(x_i)|.$$

i.e.

$$-\sum_{i=1}^{|U|}\frac{|S_Q(x_i)|}{|U|}\log\frac{1}{|S_Q(x_i)|} < -\sum_{i=1}^{|U|}\frac{|S_P(x_i)|}{|U|}\log\frac{1}{|S_P(x_i)|}.$$

Thus, $E(Q) < E(P)$. □

Property 3.2 states that the rough entropy of knowledge decreases monotonously as the granularity of information become smaller.

From property 3.2 we can obtain immediately the following properties.

Property 3.3. (Equivalence) Let $S = (U, A)$ be an incomplete information system, $P \subseteq A$. Then $U/SIM(P) = U/SIM(A)$ if and only if $E(P) = E(A)$.

Therefore, $P \subseteq A$ is a reduct of A if and only if A is independent and $E(P) = E(A)$.

Property 3.4. (Maximum) Let $S = (U, A)$ be an incomplete information system, $P \subseteq A$. The maximum of the rough entropy of knowledge P is $|U|\log|U|$. This value is achieved only by the $U/SIM(P) = \{S_P(x) = U | x \in U\}$.

Property 3.5. (Minimum) Let $S = (U, A)$ be an incomplete information system, $P \subseteq A$. The minimum of the rough entropy of knowledge P is 0. This value is achieved only by the $U/SIM(P) = \{S_P(x) = \{x\} | x \in U\}$.

If X is a finite set then the Hartley measure ([13]) of uncertainty is

$$H(X) = \log_2 |X|.$$

We will now show the relationship between the rough entropy of knowledge and the Hartley measure.

$$\begin{aligned} E(P) &= -\sum_{i=1}^{|U|} \frac{|S_P(x_i)|}{|U|} \log \frac{1}{|S_P(x_i)|} \\ &= \sum_{i=1}^{|U|} \frac{|S_P(x_i)|}{|U|} \log |S_P(x_i)| \\ &= \sum_{i=1}^{|U|} \frac{|S_P(x_i)|}{|U|} H(S_P(x_i)). \end{aligned}$$

Thus the rough entropy of knowledge P is the sum of the weighted Hartley measures of the elements of $U/SIM(P)$.

4. Significance of Attributes

Using rough entropy of knowledge, we can analyze the significance of every attribute.

Definition 4.1. Let $S = (U, A)$ be an incomplete information system, $a \in A$. We define the significance of a in A as

$$sig_{A-\{a\}}(a) = E(A - \{a\}) - E(A).$$

In the special case where A is a singleton, $A = \{a\}$, we also denote $sig_{\emptyset}(a)$ by $sig(a)$:

$$sig(a) = sig_{\emptyset}(a) = E(\emptyset) - E(\{a\}) = |U| \log |U| - E(\{a\}).$$

We know the following:

Property 4.1.

(1) $0 \leq sig_{A-\{a\}}(a) \leq |U| \log |U|$.
(2) Attribute $a \in A$ is indispensable in A if and only if $sig_{A-\{a\}}(a) > 0$.
(3) $core(A) = \{a \in A |\ sig_{A-\{a\}}(a) > 0\}$.

Definition 4.2. Let $S = (U, A)$ be an incomplete information system, $C \subseteq A$. We define the significance of $a \in A - C$ about C as

$$sig_C(a) = sig_{(C \cup \{a\})-\{a\}}(a) = E(C) - E(C \cup \{a\}).$$

Obviously, attribute $a \in A - C$ in $C \cup \{a\}$ is indispensable if and only if $sig_C(a) > 0$.

Theorem 4.1. Let $S = (U, A)$ be an incomplete information system, $C \subseteq A$. Then C is a reduct of A if C satisfies:

(1) $E(C) = E(A)$;
(2) $sig_{C-\{a\}}(a) > 0$ for every $a \in C$.

Proof. Follows immediately from Property 3.3 and Property 4.1. □

Theorem 4.1 provides the definition of knowledge reduction from information prospective, which is theoretical foundation for our algorithm given in the next section.

5. Algorithm Based on Rough Entropy for Knowledge Reduction

In this section, a heuristic algorithm for finding reduct is presented.

Let $S = (U, A)$ be an incomplete information systems. Since core is the common part of all reducts, core can be used as the starting point for computing reduct. The significance of attributes can be used to select the attributes to be added to the core. This algorithm finds an approximately minimal reduct.

Algorithm on knowledge reduction in incomplete information systems:

Input: An incomplete information system $S = (U, A)$.

Output: One reduct P of A.

Step 1. Compute the rough entropy $E(A)$ of attributes A.

Step 2. Compute $core(A) = \{a \in A |\ sig_{A-\{a\}}(a) > 0\}$.

If $E(core(A)) = E(A)$, then the algorithm terminates ($core(A)$ is the minimal reduct).

Step 3. (Create a subset C of attributes A by adding attributes)

Set $C := core(A)$.

While $E(C) \neq E(A)$ **Do**

(1) Compute significance of attribute $sig_C(a)$ for every attribute $a \in A - C$.

(2) Choose attribute a which satisfy equation:

$$sig_C(a) = max\{sig_C(a') |\ a' \in A - C\} \text{ and } C \cup \{a\} \to C.$$

(3) Compute $E(C)$.

Endwhile

Step 4. (Create a reduct P of A by dropping attributes)

Set $C' = C - core(A)$, $|C'| \to N$.

For $i = 1$ to N **Do**

(1) Remove the ith attributes a_i from C'.

(2) Compute $E(C' \cup core(A))$.

(3) If $E(C' \cup core(A)) \neq E(A)$, then $C' \cup \{a_i\} \to C'$.

Endfor

Let $C' \cup core(A) \to P$, the algorithm terminates (result P constitutes a reduct of A).

By using this algorithm, the time complexity to find one reduct is polynomial.

At the first step, the time complexity to compute $E(A)$ is $O(|U|)$.

At Step 2, we need to compute $core(A)$, i.e., compute $sig_{A-\{a\}}(a)$ for all $a \in A$. The time complexity for computing $core(A)$ is $O(|A||U|)$.

At Step 3, the time complexity to compute all $sig_C(a)$ is $(|A| + (|A| - 1) + \cdots + 1) \times |U| = |A| \times (|A| + 1)/2 \times |U| = O(|A|^2|U|)$. The time complexity to choose maximum for significance of attribute is $(|A| - 1) + (|A| - 2) + \cdots + 1 = (|A| - 1) \times |A|/2 = O(|A|^2)$. The time complexity to compute all $E(C)$ is $O(|A||U|)$. So the price of step 3 is $O(|A|^2|U|)$.

At Step 4, the time complexity to compute all $E(C' \cup core(A))$ is $O(|A||U|)$.

Thus the time complexity of this algorithm is $O(|A|^2|U|)$ (where we ignore the time complexity for computing tolerance classes).

Example 5.1. Consider descriptions of several cars as in Table 1.

Table 1.

Car	Price	Size	Engine	Max-Speed
u_1	low	compact	*	low
u_2	low	full	diesel	high
u_3	high	full	diesel	medium
u_4	high	*	diesel	medium
u_5	low	full	gasoline	high

This is an incomplete information system, where $U = \{u_1, u_2, u_3, u_4, u_5\}$, and $A = \{a_1, a_2, a_3, a_4\}$, where a_1-Price, a_2-Size, a_3-Engine, a_4-Max-Speed.

For Table 1, we compute an approximately minimal reduct by using our algorithm.

Step 1A. Compute $U/SIM(A) = \{S_A(u_1), S_A(u_2), S_A(u_3), S_A(u_4), S_A(u_5)\}$, where $S_A(u_1) = \{u_1\}, S_A(u_2) = \{u_2\}, S_A(u_3) = S_A(u_4) = \{u_3, u_4\}, S_A(u_5) = \{u_5\}$.

$E(A) = 0.8$.

Step 2A. Compute $sig_{A-\{a_1\}}(a_1) = sig_{A-\{a_2\}}(a_2) = sig_{A-\{a_4\}}(a_4) = 0$, $sig_{A-\{a_3\}}(a_3) = 3.2$; $core(A) = \{a_3\}$; $E(core(A)) = 7.521928$. Since $E(core(A)) \neq E(A)$, go to Step 3A.

Step 3A. Set $C := core(A) = \{a_3\}$.

Compute $U/SIM(C) = \{S_C(u_1), S_C(u_2), S_C(u_3), S_C(u_4), S_C(u_5)\}$, where $S_C(u_1) = \{u_1, u_2, u_3, u_4, u_5\}$, $S_C(u_2) = S_C(u_3) = S_C(u_4) = \{u_1, u_2, u_3, u_4\}$, $S_C(u_5) = \{u_1, u_5\}$.

Since $E(C) \neq E(A)$, we compute $sig_C(\{a_1\}) = 4.970951$, $sig_C(\{a_2\}) = 3.619973$, $sig_C(\{a_4\}) = 6.721928$.

Choose $a' = a_4$. Set $C := C \cup \{a_4\} = \{a_3, a_4\}$.

Compute $U/SIM(C) = \{S_C(u_1), S_C(u_2), S_C(u_3), S_C(u_4), S_C(u_5)\}$, where $S_C(u_1) = \{u_1\}$, $S_C(u_2) = \{u_2\}$, $S_C(u_3) = S_C(u_4) = \{u_3, u_4\}$, $S_C(u_5) = \{u_5\}$.

Compute $E(C) = 0.8$.

Since $E(C) = E(A)$, go to step 4A.

Step 4A. Set $C' = C - core(A) = \{a_4\}$, $|C'| = 1 \rightarrow N$.

Set $C' - \{a_4\} = \emptyset \rightarrow C'$.

Compute $E(C' \cup core(A)) = E(\{a_3\}) = 7.521928$.

Since $E(C' \cup core(A)) \neq E(A)$, $C' \cup \{a_4\} = \{a_4\} \rightarrow C'$.

Since $N = 1$, let $C' \cup core(A) = \{a_3, a_4\} \rightarrow P$, the algorithm is completed. Thus $P = \{a_3, a_4\}$ is one reduct of the set A.

In fact, $P = \{a_3, a_4\}$ is the minimal reduct of A, since the set A of all attributes has two reducts $\{a_1, a_2, a_3\}$ and $\{a_3, a_4\}$.

6. Conclusions

In this paper, we define knowledge reduction from the view of information and use rough entropy of attributes to define the significance of the attributes. A heuristic algorithm for knowledge reduction is proposed for finding an approximately minimal reduct in incomplete information systems. The time complexity of this algorithm is $O(|A|^2|U|)$. The importance of the minimal reduct is due to its potential for speeding up the learning process and improving the quality of classification. Furthermore, we are studying algorithms of knowledge reduction and knowledge discovery in incomplete decision table.

Acknowledgements

This work was supported by the national natural science foundation of China (No.69975016, No.69805004), the national 863 plan of China, the studying abroad foundation of Shanxi Province and the young science foundation of Shanxi Province, China.

References

1. Z. Pawlak, *Rough sets: Theoretical Aspects of Reasoning about Data* (Kluwer Academic Publishers, Dordrecht, 1991).
2. Z. Pawlak, J.W. Grzymala-Busse, R. Slowiński, and W. Ziarko, "Rough sets", *Comm. ACM* **38**(1995) 89-95.
3. Z. Pawlak, "Rough set theory and its application to data analysis", *Cybernetics and Systems: An International Journal* **29**(1998) 661-688.

4. I. Düntsch and G. Gediga, "Uncertainty measures of rough set prediction", *Artificial Intelligence* **106**(1998) 109-137.
5. J.W. Guan and D.A. Bell, "Rough computational methods for information systems", *Artificial Intelligence* **105**(1998) 77-103.
6. D.A. Bell and J.W. Guan, "Computational methods for rough classification and discovery", *Journal of the American Society for Information Science* **49**(1998) 403-414.
7. S.K. Wong and M. Ziarko, "On optimal decision rules in decision tables", *Bulletin of Polish Academy of Sciences* **33**(1985) 693-696.
8. M. Kryszkiewicz, "Rough set approach to incomplete information systems", *Information Sciences* **112**(1998) 39-49.
9. J.Y. Liang, Z.B. Xu, "Uncertainty measures of roughness of knowledge and rough sets in incomplete information systems", *Proceedings of the 3rd World Congress on Intelligent Control and Automation*, Press of University of Science and Technology of China, Hefei, P.R.China, 2000, pp.2526-2529.
10. H.Rybinski, M. Kryszkiewicz, "Data mining in incomplete information systems", In *Rough Set Method and Application*, eds. L.Polkowski, T.Y.Lin and S.Tsumoto (Physica-Verlag,Heidelberg, Vol.56, 2001) pp.567-582.
11. T.Beaubouef, F.E.Petry, G.Arora,"Information-theoretic measures of uncertainty for rough sets and rough relational databases", *Information Sciences* **109**(1998) 535-563.
12. D.Slezak,"Approximate reducts in decision table", *Pro. of IPMU'96*, Granada, July 1-5, 1996, pp.1159-1164.
13. R.V.L. Hartley, "Transmission of information", *The Bell Systems Technical Journal* **7**(1928) 535-563.