Rough Set Data Analysis Algorithms for Incomplete Information Systems

K.S. Chin[1], Jiye Liang[2], and Chuangyin Dang[1]

[1] Department of Manufacturing Engineering and Engineering Management,
City University of Hong Kong, Hong Kong
[2] Department of Computer Science, Shanxi University
Taiyuan, 030006, China
ljy@sxu.edu.cn

Abstract. The rough set theory is a relatively new soft computing tool for dealing with vagueness and uncertainty in databases. To apply this theory, it is important to associate it with effective computational methods. In this paper, we focus on the development of algorithms for incomplete information systems and their time and space complexity. In particular, by using measure of significance of attribute which is defined by us, we present a heuristic algorithm for computing the minimal reduct, the time complexity of this algorithm is $O(|A|^3|U|^2)$, and its space complexity is $O(|A||U|)$. The minimal reduct algorithm is very useful for knowledge discovery in databases.

1 Introduction

The rough set theory as proposed in [1, 2] provides a formal tool for dealing with imprecise or incomplete information. This approach seems fundamentally important to artificial intelligence and cognitive science.

In this paper, we develop rough set data analysis algorithms in [3] into incomplete information systems in [4]. Basic computational methods of rough set data analysis are given for incomplete information systems, their time and space complexity are analyzed. The minimal reduct algorithm which is proposed by us is very useful for knowledge discovery in databases.

2 Incomplete Information Systems

Let $S = (U, A)$ be an incomplete information systems [4, 5], we will denote a null value (i.e., missing values of attribute) by $*$.

Let $P \subseteq A$. We define tolerance relation:

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

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

Let $U/SIM(P)$ denote classification, which is the family set $\{S_P(u)|u \in U\}$. Any element from $U/SIM(P)$ will be called a tolerance class or the granularity of information.

G. Wang et al. (Eds.): RSFDGrC 2003, LNAI 2639, pp. 264–268, 2003.

3 Computing the Low Approximation and Upper Approximation

Let $S = (U, A)$ be an incomplete information system. Let $W \subseteq U$, and $a \in A$. For a classification $U/SIM(\{a\})$, $\underline{R_a}W = \{u \in U | S_a(u) \subseteq W\}$ is called the lower approximation to W from $U/SIM(\{a\})$; $\overline{R_a}W = \{u \in U | S_a(u) \cap W \neq \emptyset\}$ is called the upper approximation to W from $U/SIM(\{a\})$.

We now present an algorithm for computing the lower approximation.

Algorithm L

Let $S = (U, A)$ be an incomplete information system. Let $U/SIM(\{a\}) = \{S_a(u_1), S_a(u_2), \cdots, S_a(u_{|U|})\}$. Let $W \subseteq U$. This algorithm gives the lower approximation $\underline{R_a}W = \{u \in U | S_a(u) \subseteq W\}$ to W from $U/SIM(\{a\})$.

```
L1. Input $U/SIM(\{a\})$.
L2. Set $\emptyset \to L$.
L3. For $i = 1$ to $|U|$ Do
        If $S_a(u_i) \subseteq W$, then $L \cup \{u_i\} \to L$.
        Endfor
L4. Output $L$.
```

The time complexity of Algorithm L is $O(|U|)$, and its space complexity is $O(1)$.

Similarly, we can design an algorithm to compute the upper approximation as follows.

Algorithm H

Let $S = (U, A)$ be an incomplete information system. Let $U/SIM(\{a\}) = \{S_a(u_1), S_a(u_2), \cdots, S_a(u_{|U|})\}$. Let $W \subseteq U$. This algorithm gives the upper approximation $\overline{R_a}W = \{u \in U | S_a(u) \cap W \neq \emptyset\}$ to W from $U/SIM(\{a\})$.

```
H1. Input $U/SIM(\{a\})$.
H2. Set $\emptyset \to H$.
H3.  For $i = 1$ to $|U|$ Do
        If $S_a(u_j) \cap W \neq \emptyset$, then $H \cup \{u_j\} \to H$.
        Endfor
H4. Output $H$.
```

The time complexity of Algorithm H is $O(|U|)$, and its space complexity is $O(1)$.

4 Significance and Core

Definition 4.1 Let $S = (U, A)$ be an incomplete information system. Let X be a non-empty subset of A: $\emptyset \subset X \subseteq A$. Given an attribute $x \in X$, we say that x is significant in X if $U/SIM(X) \subset U/SIM(X - \{x\})$; and that x is not significant or non-significant in X if $U/SIM(X) = U/SIM(X - \{x\})$.

In the following we introduce a quantitative measure for significance.

Definition 4.2 Let X be a non-empty subset of A: $\emptyset \subset X \subseteq A$. Given an attribute $x \in X$, we define the significance of x in X as

$$sig_{X-\{x\}}(x) = \sum_{i=1}^{|U|} \frac{|S_{X-\{x\}}(u_i)| - |S_X(u_i)|}{|U| \times |U|}.$$

The overall time complexity for computing a significance is $O(|X||U|^2)$, and the space complexity for computing a significance is $O(|X||U|)$.

Definition 4.3 Let X be a non-empty subset of A: $\emptyset \subset X \subseteq A$. The set of attributes $x \in X$ which are significant in X is called the core of X, denoted by C_X. That is, $C_X = \{x \in X | sig_{X-\{x\}}(x) > 0\}$.

Also, we define $C_\emptyset = \emptyset$.

Algorithm C

Let $S = (U, A)$ be an incomplete information system. Let X be a non-empty subset of A: $\emptyset \subset X \subseteq A$. This algorithm obtains C_X of X.

```
C1. Input $S = (U, A)$, X.
C2. Set $\emptyset \to C_X$.
C3. For every $x \in X$, compute $SIM(\{x\})$.
C4. For $i = 1$ to $|X|$ Do
        Compute $sig_{X-\{x_i\}}(x_i)$. If $sig_{X-\{x_i\}}(x_i) > 0$, then $C_X \cup \{x_i\} \to$
        $C_X$.
        Endfor
C5. Output $C_X$.
```

The time complexity of Algorithm C is $O(|X||U|^2 + |X|^2|U|)$, and its space complexity is $O(|X||U|)$.

5 Reducts

Definition 5.1 Let $S = (U, A)$ be an incomplete information system. A subset A_0 of A is said to be a reduct of A if A_0 satisfies:

(1) $U/SIM(A_0) = U/SIM(A)$; i.e., $A_0 \leftrightarrow A$; and

(2) If $A^{'} \subset A_0$, then $U/SIM(A_0) \subset SIM(A^{'})$; i.e., if $A^{'} \subset A_0$, then $A^{'} \not\leftrightarrow A$.

From this definition, the time complexity to find all reducts is exponential. First, we need to consider all $|2^A| = 2^{|A|}$ subsets of A. And for every subset A_0, we need to compute $U/SIM(A_0)$. The time complexity to compute $U/SIM(A_0)$ for one subset A_0 is $O(|A||U|^2)$. So the total price is $O(2^{|A|}|A||U|^2)$.

We have the relationship between reducts and core as follows.

Theorem 5.1 Let $S = (U, A)$ be an incomplete information system. Then $C_A = \bigcap_{i=1}^{s} A_{0i}$, where $A_{01}, A_{02}, \cdots, A_{0i}, \cdots, A_{0s}$ are all reducts of A.

6 Minimal Reduct

Definition 6.1 Let $S = (U, A)$ be an incomplete information system, and $C \subseteq A$. We define the significance of $a \in A - C$ about C as $sig_C(a) = sig_{(C \cup \{a\}) - \{a\}}(a)$.

Algorithm M

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

```
M1. Input $S = (U, A)$.
M2. Compute $U/SIM(A)$, and $C_A = \{a \in A | sig_{A-\{a\}}(a) > 0\}$. Set
    $C_A \rightarrow C$.
M3. Compute $U/SIM(C)$.
M4. While $U/SIM(C) \neq U/SIM(A)$ Do
        (1)  Compute $sig_C(a)$ for $\forall a \in A - C$.
        (2)  Choose $a^{'} \in A - C$ such that

                 $sig_C(a^{'}) = max\{sig_C(a) | \forall a \in A - C\}.$

        (3)  Set $C \cup \{a^{'}\} \rightarrow C$, and compute $U/SIM(C)$.
        Endwhile
M5. Set $C^{'} = C - C_A$, $|C^{'}| \rightarrow N$.
        For $i = 1$ to $N$ Do
        (1) Remove the $i$th attributes $a_i$ from $C^{'}$.
        (2) Compute $U/SIM(C^{'} \cup C_A)$.
        (3) If $U/SIM(C^{'} \cup C_A) \neq U/SIM(A)$, then $C^{'} \cup \{a_i\} \rightarrow C^{'}$.
        Endfor
M6. Set $C^{'} \cup C_A \rightarrow C$. Output $C$.
```

The time complexity of Algorithm M is $O(|A|^3|U|^2)$, and its space complexity is $O(|A||U|)$.

7 Conclusions

In this paper, we develop rough set data analysis algorithms in [3] to incomplete information systems in [4]. Time complexity and space complexity of the algorithms have been analyzed. In particular, by using measure of significance of attribute which is defined by us , we present a heuristic algorithm for computing the minimal reduct, the time complexity of this algorithm is $O(|A|^3|U|^2)$, and its space complexity is $O(|A||U|)$. The importance of the minimal reduct is due to its potential for speeding up the learning process and improving the quality of classification.

Acknowledgements. This work was supported by the national natural science foundation of China (No. 60275019) and the natural science foundation of Shanxi, China.

References

1. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht (1991)
2. Pawlak, Z., Grzymala-Busse, J.W., Slowiński, R., Ziarko, W.: Rough Sets. Comm. ACM. **38**(1995) 89–95
3. Guan, J.W., Bell, D.A.: Rough Computational Methods for Information Systems. Artificial Intelligence. **105**(1998) 77–103
4. Kryszkiewicz, M.: Rough Set Approach to Incomplete Information Systems. Information Sciences. **112**(1998) 39–49
5. Kryszkiewicz, M.: Rule in Incomplete Information Systems. Information Sciences. **113**(1999) 271–292