An Algorithm of Constructing Maximal Consistent Block in Incomplete Information Systems

LIANG Ji-Ye WANG Bao-Li QIAN Yu-Hua Li De-Yu

(Key Laboratory of Computational intelligence & Chinese
Information Processing of Ministry of Education)
(School of Computer & Information Technology Shanxi University, Taiyuan, 030006)

Abstract The technique of maximal consistent blocks for knowledge acquiring in incomplete information systems was proposed in [3], and the better theoretical results than the ones using traditional methods can be obtained. In the present paper, the properties of maximal consistent blocks are studied and a hierarchical algorithm for constructing maximal consistent block is given, which will be helpful for acquiring knowledge efficiently in incomplete information systems.

Keywords Rough set, Incomplete information system, Maximal consistent block

1 Introduction

Rough set theory, as a mathematic tool for dealing with fuzziness and uncertainty, has been presented by Z. Pawlak in 1982[1,2]. Though the earlier studies have mainly been developed for completed information systems, some important results have recently been obtained for incomplete systems[2~7]. Several important extended rough set models for incomplete information systems have been established form the different relations induced by datum given in the systems in [4~7]. A new methodology, i.e., maximal consistent block technique, which is different from the models mentioned below, has been proposed to acquire knowledge in an incomplete system by analyzing the structure of a similarity class [3]. In paper [3], it has been proved that the maximal consistent blocks are the elemental knowledge units of the system, a higher approximation accuracy for a given set can obtained by using the maximal consistent block technique than the one using the similarity class of the similarity relation[4], and the size of discernibility matrix is reduced by adopting the technique, which can improve the efficiency of acquiring reduct of decision system.

All advantages of the methodology in paper [3] are based on acquiring the maximal consistent blocks decided by the given attributes, so it is important to design an efficient algorithm of constructing maximal consistent block. In this paper, based on the rough set model proposed in paper [3], more properties of maximal consistent blocks of incomplete information system are studied and a hierarchical algorithm for constructing maximal consistent block is presented. And in the end, an example is given to help to comprehend the algorithm.

2 Incomplete information system

Definition 2.1 Let $S=(U,AT,V,f)$ be an information system, where U is a non-empty finite set of objects, AT is a non-empty finite set of attributes, $V=\bigcup_{a\in AT}V_a$, V_a is domain of an attribute a , f: $U\times AT\rightarrow V$ is a information function, for $\forall x\in U, a\in AT$, $f(x,a)\in V_a$ .Generally, $S=(U,AT,V,f)$ is also denoted by $S=(U,AT)$, $a(x)$ denotes $f(x,a)$ 。

Any attribute domain V_a may contain special symbol “*” to indicate that the value of an attribute is unknown. Here, we assume that an object $x\in U$ possesses only one value for an attribute a. Thus, if the value of an attribute a is missing, then the real value must be from the set $V_a\setminus\{*\}$. Any domain value different from “*” will be called regular. If $\exists x\in U,\exists a\in AT, a(x)$ = “*”, the S is called incomplete information system. In this paper, we assume $\forall x\in U,\exists a\in AT,$ the $a(x)\neq *$ hold.

Definition 2.2 Let $S=(U,AT,V,f)$ be an incomplete information system, each subset of attributes $P \subseteq AT$ determines a binary similarity relation $SIM(P)$:

$$SIM(P) = \{(x, y) \in U \times U \mid \forall a \in P, a(x) = a(y)$$
$$\vee a(x) = * \vee a(y) = *\}.$$

From the definition $SIM(P)$, $\forall x \in U$, $S_P(x) = \{y \mid (x, y) \in SIM(P)\}$, $S_P(x)$ describes those objects from U which may be indiscernible to x, and is called the similarity class of x or the similarity block of x. Obviously, the binary relation $SIM(P)$ is a tolerance relation on U.

Definition 2.3[3] Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$ an subset of attributes and X an subset of objects U, we say X is consistent with respect to P, if $(x, y) \in SIM(P)$ for arbitrary $x, y \in X$. If there does not exist a subset $Y \subseteq U$ such that $X \subset Y$, and Y is consistent with respect to P, then X is called a maximal consistent block of P.

To be convenient, we denote the set of all maximal consistent blocks determined by the attribute set P as $C(P)$, and the set of all maximal consistent blocks of P which includes object x is denoted as $C_P(x)$.

Example 2.1 **Table 1** depicts an incomplete table $S = (U, AT)$ containing information about cars, $U = \{1, 2, 3, 4, 5, 6\}$, $AT = \{$ Price, Mileage, Size, Max-Speed $\})$, assume $P = \{$ Mileage，Max-Speed $\}$.

Table 1 an incomplete information system about car

Car	Price	Mileage	Size	Max-Speed
1	High	Low	Full	Low
2	Low	*	Full	Low
3	*	*	Compact	Low
4	High	*	Full	High
5	*	*	Full	High
6	Low	High	Full	*

The similarity blocks determined by the attribute set P are:

$$S_P(1) = \{1, 2, 3\}, S_P(2) = S_P(3) = \{1, 2, 3, 6\},$$
$$S_P(4) = S_P(5) = \{4, 5, 6\}, S_P(6) = \{2, 3, 4, 5, 6\}.$$

The collection of all maximal consistent blocks determined by P is the follows:

$$C(P) = \{\{1, 2, 3\}, \{2, 3, 6\}, \{4, 5, 6\}\}.$$

The collections of maximal consistent blocks with respect to each object determined by P are:

$$C_P(1) = \{\{1, 2, 3\}\}, C_P(2) = C_P(3) = \{\{1, 2, 3\}, \{2, 3, 6\}\},$$
$$C_P(4) = C_P(5) = \{\{4, 5, 6\}\}, C_P(6) = \{\{2, 3, 6\}, \{4, 5, 6\}\}.$$

Definition 2.4 [4] Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, and $X \subseteq U$, define the approximation operators:

$$\underline{Apr}_P(X) = \{x \in U \mid S_P(x) \subseteq X\},$$
$$\overline{Apr}_P(X) = \{x \in U \mid S_P(x) \bigcap X \neq \varnothing\}.$$

Definition 2.5 [3] Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$ and $X \subseteq U$, define the approximation operators:

$$\underline{apr}_P(X) = \bigcup\{Y \in C(P) \mid Y \subseteq X\},$$
$$\overline{apr}_P(X) = \bigcup\{Y \in C(P) \mid Y \bigcap X \neq \varnothing\}.$$

In paper [3], it has been proved that a better approximation can obtain by using maximal consistent blocks as the elemental knowledge granular in definition 2.5 than by definition 2.4. Furthermore, using the maximal consistent blocks as the elemental units can construct the smaller discerniblity matrix and improve the efficiency of acquiring reduct. All these advantages are based on obtaining the maximal consistent blocks $C(P)$ determined by P.

3 Properties of maximal consistent blocks

Property 3.1 Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, $x \in U$, if $\forall a \in P$, $a(x)$ is a regular value , then $S_p(x) \in C_P(x)$, and $|C_P(x)| = 1$.

Proof. $\forall a \in P, a(x)$ is a regular value, then $\forall y \in S_P(x)$, $\forall a \in P$, $a(y) = a(x)$ or $a(y) = *$ holds. For any $y, z \in S_P(x)$, there exist the following 4 cases:

$$a(y) = *, a(z) = *; a(y) = *, a(z) = a(x);$$
$$a(y) = a(x), a(z) = *; a(y) = a(x), a(z) = a(x).$$

Under the four conditions, y, z are indiscernible with respect to a, that is $(y, z) \in SIM(P)$, so $S_x(P)$ is consistent with respect to P. $\forall y' \in U \setminus S_P(x)$, $(x, y') \notin SIM(P)$, so $S_P(x)$ is a maximal consistent block, so $S_P(x) \in C_P(x)$ hold.

$\forall X \in C_P(x)$, $X \subseteq S_P(x)$, according to the maximal characteristic of X , $C_P(x) = \{S_P(x)\}$ is hold, that means $|C_P(x)| = 1$. This completes the proof.

Property 3.2 let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, $x \in U$, if $|C_P(x)| > 1$, then there exist $b \in P$ and $b(x) = *$.

Proof. It can easily follows from the property 3.1.

Property 3.3[3] Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, $X \in C(P)$ is the maximal subset of U such that $SIM(P)$ is transitive on it.

Property 3.4[3] Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, then $X \in C(P)$, if and only if $X = \bigcap_{x \in X} S_P(x)$.

Property 3.5 Let $S = (U, AT)$ be an incomplete information system, then for any $a \in AT$, $C(\{a\}) = \bigcup_{i=1}^{|V_a \backslash \{*\}|} \{\{x \mid a(x) = v_i \vee a(x) = *\}\}$.

Proof. we let $X_i = \{x \mid a(x) = v_i \vee a(x) = *\}$, then it is clear that for $\forall x, y \in X_i$, x and y is indiscernible with respect to a , so X_i is a consistent block, for any $z \in U \setminus X_i$, $a(z) \neq v_i$, so z and the objects of X_i which information value under a is v_i are discernible, so X_i is a maximal consistent block determined by the attribute set $\{a\}$. The right of the equation include all values of $V_a \backslash \{*\}$, so the right are the collection of all maximal consistent blocks with respect to the attribute set $\{a\}$. That completes the proof.

Property 3.6[3] Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, if $P \subseteq Q \subseteq AT$, for arbitrary $X \in C(Q)$, there exists $Y \in C(P)$ such that $X \subseteq Y$.

Property 3.7 Let $S = (U, AT)$ be an incomplete information system, $P \subseteq AT$, for arbitrary $X \in C(P)$, there exists $Y \in C(P \backslash \{a\})$ such that $X \in C^Y(\{a\})$, where $C^Y(\{a\})$ denotes all maximal consistent blocks determined by the attribute set $\{a\}$ under the universe Y .

Proof. We prescribe that the collection of all maximal consistent blocks with respect to $\varnothing$ is $C(\varnothing)$, and $C(\varnothing) = \{U\}$.

According to property 3.6, $\forall X \in C(P)$, there must exist $Z \in C(P \backslash \{a\})$ such that $X \subseteq Z$.

Assume that $Z_1, Z_2, \cdots Z_m \in C(P \backslash \{a\})$ and $X \subseteq Z_i$, then for each Z_i , we can obtain $C^{Z_i}(\{a\})$ by means of the method given in property 3.5, and let $D = \bigcup_{i=1}^{m} C^{Z_i}(\{a\})$.

We use apagoge to prove it.

Assume there doesn't exist a subset $Y \subseteq U$ satisfies the conclusion, then, $X \notin D$ and two cases for the collection D :

(1) $\exists Y' \in D$ such that $X \subset Y'$;

(2) $\forall Y'' \in D$ such that $Y'' \subset X$.

In the case (1), we can easily know that Y' is a consistent block determined by the attribute set P according to the process of constructing Y' , which is converse with the fact that X is a maximal consistent block with respect to the attribute set P .

In the case (2), for $\forall Y'' \in D, Y'' \subset X$ is hold, so in $\forall Z_i$, there always exist $x_i, y_i \in X$ such that $a(x_i) \neq a(y_i)$. Therefore, in the universe U , exist $x, y \in X$ such that $a(x) \neq a(y)$, which is also converse with the fact that X is a maximal consistent block with respect to the attribute set P .

That is complete the proof.

According to property 3.7, we can find the maximal consistent blocks determined by the attribute set P by hierarchical technique.

Above all, we can obtain the collection of the maximal consistent blocks $C(P \backslash \{a\}) = \{Z_1, Z_2, \cdots Z_m\}$, then we can use attribute a to gain the finer indiscernible set $C^{Z_i}(\{a\})$ for each Z_i of $C(P \backslash \{a\})$. Furthermore, $C(P) \subseteq \bigcup_{i=1}^{m} C^{Z_i}(\{a\})$ holds. However, in most cases, the relation $=$ does not hold, that is to say the element of $C^{Z_i}(\{a\})$ is a consistent set with respect to P , but may not be a maximal consistent set determined by P . There may exist $X_l \in C^{Z_i}(a), X_m \in C^{Z_k}(a),$ and $X_l \subseteq X_m$ holds, so X_l may not be maximal consistent block with respect to P .

So we can obtain the maximal consistent blocks by using hierarchical technique. And we should judge the blocks whether or not are maximal consistent blocks with respect to the current attribute set after decomposing the coarser maximal consistent blocks of the last lever. A particular operator is if X and the super set of X are in the same lever, X is not a maximal consistent block and should be rejected.

In example 1, let $P=$ {Mileage, Max-Speed}, $a=$ Max-Speed, the $C(P)\subset\bigcup_{i=1}^{m}C^{Z_i}(\{a\})$ is hold.

4 Hierarchical algorithm of constructing maximal consistent blocks

On the one hand, it is a NP-hard problem to obtain the maximal consistent blocks by using the tolerant relation. On the other hand, it is too complex to gain the maximal consistent blocks according to the property 3.4 and 3.5.

In this section, we design a hierarchy algorithm of constructing maximal consistent blocks according to the property 3.6 and 3.7.

Let $S=(U,AT)$ be an incomplete information system, if $U'\subseteq U, AT'\subseteq AT$, we call $S'=(U',AT')$ is the subsystem of the original system. The maximal consistent blocks with respect to AT' in S' are local maximal consistent blocks. And the maximal consistent blocks with respect to AT in S are public maximal consistent blocks.

Algorithm 1 Calculating the maximal consistent blocks with respect to the attribute a in $S=(U,\{a\})$.

```
Input: incomplete information system $S=(U,\{a\})$;
Output: $C(\{a\})$.
Initialize: $C(\{a\})=\varnothing$;
Step1: sorting the objects of $U$ on the value of them on $a$;
Step2: traversing sorted objects of the universe,
           $U/\{a\}=\{X_1,X_2,\cdots X_{|V_a|}\}$;
Step3: if $a(x)\neq *(x\in X_1)$, then $C(\{a\})=\bigcup_{i=1}^{|V_a|}\{X_i\}$,
        go to setp7;
Step4: if $|V_a|>1$ then go to step 5, else go to step 6;
Step5: for $i=2,3,\cdots|V_a|$, $X_i'=X_i\cup X_1$,
        $C(\{a\})=\bigcup_{i=2}^{|V_a|}\{X_i'\}$, go to setp 7;
Step6: $C(\{a\})=\{X_1\}$;
Step7: output $C(\{a\})$.
```

Algorithm 2 Finding the maximal consistent blocks $C(P)$ with respect to P.

```
Input: an incomplete information system $S=(U,AT)$,
        $P=\{a_1,a_2,\cdots a_r\}\subseteq AT$;
Output: $C(P)$.
Explanation: $\tilde{C}$ notes the sub-universe with respect to singular attribute set in the acquiring process, $X_{ij}$ is the j^th local maximal consistent block in $C(\{a_1,a_2,\cdots a_i\})$, $C^{X_{ij}}(\{a_{i+1}\})$ is the collection of local maximal consistent sets with respect to the attribute $a_{i+1}$ in the local subsystem $(X_{ij},\{a_{i+1}\})$.
Initialize: $C(P)=\varnothing, i=1, \tilde{C}=\varnothing$;
Step 1: input $S'=(U,\{a_1\})$ as the parameter of algorithm 1 to obtain $C(\{a_1\})$;
Step 2: if $i=r$, then go to step 9;
Step 3: $j=1$;
Step 4: if $j>|C(\{a_1,\cdots a_i\})|$ then go to step7;
Step 5: if $|X_{ij}|=1$ then
            $\tilde{C}=\tilde{C}\bigcup\{X_{ij}\}, C^{X_{ij}}(\{a_{i+1}\})=\varnothing$
        else
            input $S'=(X_{ij},\{a_{i+1}\})$ as the parameter
            of algorithm 1 to gain $C^{X_{ij}}(\{a_{i+1}\})$;
Step 6: $j=j+1$, go to step 4;
Step 7: $C(\{a_1,\cdots,a_{i+1}\})=\bigcup_{j=1}^{|C(\{a_1,\cdots a_i\})|}C^{X_{ij}}(\{a_{i+1}\})$;
Step 8: $i=i+1$, go to step 2;
Step 9: $C=C(\{a_1,\cdots,a_r\}\bigcup\tilde{C}$;
Step 10: $C'=\{X\in C\mid\exists Y\in C 且 X\subset Y\})$;
Step 11: $C(P)=C\setminus C'$
Step 12: output $C(P)$.
```

According to the algorithm given above, find the maximal consistent blocks with respect to the attribute set AT in the incomplete information system of example 1.

$AT=\{a_1,a_2,a_3,a_4\}=$ {Price, Mileage, Size, Max-Speed}.

$C(\{a_1\})=\{\{1,3,4,5\},\{2,3,5,6\}\}=\{X_{11},X_{12}\}$.

For X_{11}, X_{12} we use a_2 find the local maximal consistent blocks :

$C^{X_{11}}(\{a_2\})=\{1,3,4,5\}$ $C^{X_{12}}(\{a_2\})=\{2,3,5,6\}$,

$$\begin{aligned}C(\{a_1,a_2\})&=C^{X_{11}}(\{a_2\})\bigcup C^{X_{12}}(\{a_2\})\\&=\{\{1,3,4,5\}\}\bigcup\{\{2,3,5,6\}\}\\&=\{\{1,3,4,5\},\{2,3,5,6\}\}\\&=\{X_{21},X_{22}\},\end{aligned}$$

For X_{21}, X_{22}, we use a_3 find the local consistent blocks as follows:

$$C^{X_{21}}(\{a_3\}) = \{\{1,4,5\},\{3\}\},$$

$$C^{X_{22}}(\{a_3\}) = \{\{2,5,6\},\{3\}\},$$

$$\begin{aligned} C(\{a_1,a_2,a_3\}) &= C^{X_{21}}(\{a_3\}) \bigcup C^{X_{22}}(\{a_3\}) \\ &= \{\{1,4,5\},\{2,5,6\},\{3\}\} \\ &= \{X_{31},X_{32},X_{33}\}. \end{aligned}$$

For X_{31}, X_{32}, use a_4 find the local maximal consistent blocks:

$C^{X_{31}}(\{a_4\}) = \{\{1\},\{4,5\}\}$, $C^{X_{32}}(\{a_4\}) = \{\{2,6\},\{5,6\}\}$,

because $\tilde{C} = \{\{x_3\}\}, C^{X_{33}}(\{a_4\}) = \varnothing$,

$$\begin{aligned} C(\{a_1,a_2,a_3,a_4\}) &= C^{X_{31}}(\{a_4\}) \bigcup C^{X_{32}}(\{a_4\}) \bigcup C^{X_{33}}(\{a_4\}) \\ &= \{\{1\},\{4,5\}\} \bigcup \{\{2,6\},\{5,6\}\} \bigcup \varnothing \\ &= \{\{1\},\{4,5\},\{2,6\},\{5,6\}\}. \end{aligned}$$

$$\begin{aligned} C &= C(\{a_1,a_2,a_3,a_4\}) \bigcup \tilde{C} \\ &= \{\{1\},\{4,5\},\{2,6\},\{5,6\},\{3\}\}. \end{aligned}$$

Here, $C' = \varnothing$,so the C is the $C(AT)$ we find.

5 Conclusions

There are several better theoretical results of the rough set model based on the maximal consistent technique than the traditional one. However, obtaining the maximal consistent blocks with respect to the given attribute set is the basis of the new rough set model. In this paper, we studied the properties of the maximal consistent block and designed a hierarchy algorithm of constructing maximal consistent blocks, which will helpful for acquiring knowledge efficiently in incomplete information systems.

References:

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publisher, 1991.

2 Liang J.Y., Li D.Y. Uncertainty and Knowledge Acquisition in Information Systems. Science Press, Beijing, China(2005).

3 Leung Y., Li D.Y. Maximal consistent block technique for rule acquisition in incomplete information systems. Information Science, 2003, 153: 85-106.

4 Kryszkiewicz M. Rough set approach to incomplete information systems. Information Science, 1998, 112(1-4):39-49.

5 Kryszkiewicz M. Rule in Incomplete Information Systems. Information Science, 1999, 113(3-4): 271-292.

6 Stefanowski J., Tsoukias A. On the extension of rough sets under incomplete information. In: N. Zhong, A Skowron, SOhsuga eds. Proc. of the 7th International Work shop on New Directions in Rough Sets, Data Mining, and Granular Soft Computing. Berlin: Springer-Verlag, 1999. 73-81

7 Wang G. Y. Extension of Rough Set under incomplete information systems. Journal of Computer Research and Development (in Chinese), 2002, 39(10):1238-1243