# Attribute Reduction Algorithm Based on Incomplete Decision Table

Du Yue
Sixth Department, Army Officer Academy
Liaoningduyue@163.com

Wang Jian
Eleventh Department, Army Officer Academy, Hefei, Anhui

Zhang Xu
Eleventh Department, Army Officer Academy, Hefei, Anhui

***Abstract*—The paper describes the basic concepts of rough set theory and discernibility matrix and presents an attribute reduction algorithm based on reduced discernibility matrix, which aims at resolving the inadequate of the existing attribute reduction based on incomplete decision table. There only contain useful elements for the algorithm in the reduced discernibility matrix, which obtain one reduction of incomplete decision table by iteration and set operations. The experimental results show that the algorithm can not only obtain reduced attribute, but also reduce the computation time and storage space greatly.**

***Keywords-incomplete decision table; attribute reduction; rough set; discernibility matrix; algorithm complexity***

## I. INTRODUCTION

Rough set theory is a mathematical tool of dealing with ambiguity and imprecision knowledge, which is presented by Polish mathematician Pawlak Professor in 1982 [1].At present, rough set theory has make great progress in theory and application, which has been widely used in machine learning, knowledge acquisition, inductive reasoning, pattern recognition and other fields[2-4].

Attribute reduction is an important concept of rough set for data analysis. The main idea is to obtain decisions or classifications of problems on the condition s of maintaining the classification ability of the knowledge base. So attribute reduction in rough set theory is the core content, but it has been proved that looking for minimum attribute reduction of decision table is a NP-hard problem. The complete decision table is treated as the object of study in classical rough set theory.However in real life, due to the measurement error in the extraction process of raw data or the restrictions of accessing process make the raw data incomplete. The phenomenon of incomplete decision table is widespread and greatly limits the development of rough set to the direction of practical. So in recent years, there have been lots of literatures on the problem of attribute reduction in incomplete decision table. [5-7].Although the ideal of attribute reduction based on discernibility matrix is intuitive and the algorithm is simple and efficient, the required storage space of discernibility matrix may be large while the decision table may be not large. For example, when the number of condition attributes is 100 units and the number of objects is 1000000 units. In the worst case, the storage of the difference matrix requires 100 * 1000000 * (1000000-1) / 2 = 5 * 1013 units, which will increase the computing time and storage space undoubtedly.This paper presents an attribute reduction algorithm based on the dependence between attributes, which only contains the useful elements for the algorithm. The experiments show that the algorithm is effective and make great improvement in time and space complexity compared with the original algorithm.

## II. INCOMPLETE INFORMATION SYSTEM

An information system can be expressed as a four-tuple, as followings= (U, A, V, f), where U denotes finite sets of objects. A denotes attribute collection, where A=C∪D, C∩D=∅ .C denotes condition attribute collection and D denotes decision attribute collection. V is defined as the value field of attribute "a". "f" is defined as a mapping from the attribute to the value field, which denotes as the form of "A->V".If there exists a collection C which contains an attribute "a" at least and V contains null, namely f(x, a)=*, then the decision table is defined as incomplete decision table, else defined as complete decision table.

Definition 1: In the incomplete decision table S (S= (U,A,V,f)), $P \subseteq A$, $IND(P) = \{(x,y) \mid f(x,a) = f(y,a), \forall a \in P\}$ is called the indiscernibility relation of S. The relationship $IND(P)$ constitutes a partition of U, represented by $U/IND(P)$, denoting as $U/P$. Elements of $U/P$ are called as the equivalence class.

Definition 2: In the incomplete decision table S (S= (U, A, V, f)), $P \subseteq A$ and $X \subseteq U$, denoting $U/R = \{R_1, R_2 ...., R_n\}$, $\underline{R}(X)$ is called the under approximation set of X on R, where $\underline{R}(X) = \cup\{R_i' R_i \in U/R, R_i \subseteq X\}$. $\overline{R}(X)$ is called the upper approximation set of X on R, where $\overline{R}(X) = \cup\{R_i \mid R_i \in U/R, R \cap X \neq \phi\}$. $POS_R(X)$ is called the positive domain of R about X, where $POS_R(X) = \underline{R}(X)$.

Definition 3: In the incomplete decision table S(S= (U, A, V, f)), supposing C is condition attribute and D is decision attribute, U/D is called division of the domain U by the decision attributes D, where $U/D = \{D_1, D_2 ... D_n\}$ .U/C is known as division of the domain U by the condition attributes C, where $U/C = \{C_1, C_2 ... C_n\}$. $POS_C(D)$ is the positive domain of C about D, where $POS_C(D) = \underline{C}(D)$.

While $B \subseteq C$ , if $POS_B(D) = POS_{(B-\{R\})}(D)$ , r is the attribute of B relative to D which can be omitted. Otherwise r is the attribute of B relative to D which can not be omitted. If $POS_B(D) = POS_C(D)$ for the independent subset B of C, B is called the relative reduction of B.

Definition 4: In the incomplete decision table S(S= (U, A, V, f)), T(C) is defined as tolerance relations on U, which is expressed as {(x,y) $\in U \times U \mid \forall b \in C$ },f(x, b )=f(y ,b ) $\vee f(x,b) = * \vee f(y,b) = *\}$ . $T_C(x)$ is used for representing the set of objects in the form of $\{y \in U \mid (x,y) \in T(C)\}$.

## III. DISCERNIBILITY MATRIX

In the attribute reduction algorithm based on discernibility matrix, the first step is to calculate the discernibility matrix according to the decision table generally. While the scale of object in the decision table is large, there will be some shortcomings such as time-consuming, large storage space, which will be the bottleneck affecting the efficiency of the algorithm. Since the elements of discernibility matrixes produced from the any two objects x and y in the same equivalence class are the same as the elements of discernibility matrixes from object z which is not belong to the equivalence class. It only need put forward one representative tuple composing a simplified decision table in the same equivalence class, which will greatly reduce the dimensions of the discernibility matrixes, the computational complexity and storage space. So the method can improve the efficiency of the algorithm. Now do some modification of discernibility matrix's definition.

Definition 5: In the incomplete decision table S (S= (U, A, V, f)), $U/D = \{d_1, d_2 ... d_n\}$ , the dependence of decision attribute sets $d_i$ to condition attribute set C which denotes $\delta_C(x)$ is defined as following: $\delta_C(x) = |T_C(x) \cap d_i| / |T_C(x)|$。

Definition 6: In the incomplete decision table S (S= (U, A, V, f)), the elements ( $m_{ij}$ ) of binary discernibility matrix M are defined as following:

$$m_{ij} = \begin{cases} 1 & f(x_i,D) \neq f(x_j,D) \wedge \delta_C(x_i)=1 \wedge \delta_C(x_j)=1 \vee (\delta_C(x_i)=1 \vee \delta_C(x_j)=1) \\ 0 & others \end{cases}$$

Definition 7: The elements ( $\beta_l (l = 1,2...,C)$ ) of $m_{ij}$ are defined as following:

$$\beta_S = \begin{cases} 1 & f(x_i,a) \neq f(x_j,a) \wedge f(x_i,a) \neq * \wedge f(x_j,a) \neq * \\ 0 & f(x_i,a) = f(x_j,a) \vee f(x_i,a) = * \vee f(x_j,a) \neq * \end{cases}$$

T heorem 1: Supposed c $\in$ C, if $\forall m_{ij} \in M = 1$ , $v_i^c \neq v_j^c$ .

Proof: Supposed the equation "$v_i^c = v_j^c$" is true, the conclusion that $m_{ij}$ =0 can be drawn by definition 6 and definition 7, which is contradict with the known conditions $\forall m_{ij} = 1$ .Therefore we can prove the equation $v_i^c \neq v_j^c$ established. If $v_i^c = v_j^c$ , $\exists a \in C$ , there will be $v_i^{c+a} \neq v_j^{c+a}$ .

If there is a line whose elements in the discernibility matrix are all zero, indicates that any attribute of the element can not be used for distinguishing two objects according to theorem 1, which means that all elements with value 0 is meaningless in the discernibility matrix. If there is a line whose elements in the discernibility matrix are all one, indicates that any attribute of the element can be used for distinguishing objects, which means that all elements with value one is meaningless in the discernibility matrix. In the attribute reduction algorithms based on discernibility matrix, need to delete the elements of nuclear-valued attribute where the column value is 1 after obtaining attribute of nuclear value from paper [8]. If there be relationship" a=a + b", the element "a " will be deleted, which will not affect the efficiency of attribute reduction.The elements of new discernibility matrix will be further reduced to obtain the smaller size of new discernibility matrix according to the above discussion.Delete the rows where the value of all elements is 1 or 0 in the discernibility matrix. If two lines exists the relationship "a=a+b", delete "a". The reduced discernibility matrix denotes $M'$ and $M' = m_{ij}'$ .

Theorem 2[5]: In the incomplete decision table S (S= (U, A, V, f)), if the condition which is $\exists B \subseteq C$ , $\forall m_{ij}' \in M \neq \phi$ is true, the conclusion of $B \cap m_{ij}' \neq \phi \Leftrightarrow POS_C(D) = POS_B(D)$ will be proved correct.

Proof: Supposing the representation of $POS_C(D) \neq POS_B(D)$ is tenable.

For $\exists x_i \in U \Rightarrow T_C(x_i) \subseteq POS_C(D) \wedge T_B(x_i) \not\subset POS_C(D)$, there exists $x_j$ which will meet the condition of $x_j \in U$ .So from f ( $x_i$,C ) =f( $x_j$,C ) $\wedge$ f( $x_i$,C )= $* \wedge$ f( $x_j$,C) = $* \wedge$ f( $x_i$,C) = f($x_j$,D) , the conclusion of $T_C(x_j) \neq T_C(x_i)$ will be drawn. And there must be $\exists a \in C - B \Rightarrow f(x_i,a) \neq f(x_j,a)$. Because of $x \in U$ and $\delta_C(x_i) = 1$, $a \in m_{ij}'$ will be true. While $m_{ij}' \neq \phi$ and $x_j \in U$, we can be able to reach the conclusion of $B \cap m_{ij}' = \phi$ , which is contradiction with the known condition. So the formula $B \cap m_{ij}' = \phi \Rightarrow POS_C(D) = POS_B(D)$ can be proved to be true.

Supposing $\exists \phi \neq m_{ij}' \in M$, the formula of $R \cap m_{ij}' = \phi$ is true. There must be an attribute "a" while $a \in C \wedge a \notin B$ meeting the condition of $a \in m_{ij}'$ . Then $f(x_i,a) \neq f(x_j,a) \wedge f(x_i,a) \neq * \wedge f(x_j,a) \neq *$ , as to $\delta_C(x_i) = 1$ and $\delta_C(x_j) = 1$ ,at least one is correct.

Supposing $\delta_C(x_i) = 1$ is correct, there will be $D_i \in U/D \wedge T_C(x_i) \subseteq D_i$ to meet the formula $x_i \in POS_C(D)$ .Because of $B \cap m_{ij}' \neq \phi$ , there will be

$T_C(x_i) \cup T_C(x_j) \subseteq T_B(x_i)$ and $x_i \notin POS_B(D)$. So we come to this conclusion: $POS_C(D) \neq POS_B(D)$, which is contradiction with the known condition. Therefore the formula of $POS_C(D) = POS_B(D) \Leftrightarrow B \bigcap m_{ij}^{'} \neq \phi$ can be proved to be correct.

Attribute reduction algorithm based on reduced discernibility matrix is equivalent to the attribute reduction algorithm based on positive region from theorem 2. So the attribute reduction of decision table can be obtained based on the method of reduced discernibility matrix.

## IV. ATTRIBUTE REDUCTION ALGORITHM BASED ON REDUCED DISCERNIBILITY MATRIX

In order to improve the computational speed, the discernibility matrix $M^{'}$ is calculated from information system U firstly. Then the sum of element attribute and the sum of element of discernibility matrix are treated as heuristic information to guide implementation of the algorithm. So the algorithm is efficient and does not require lots of storage space.

The attribute reduction algorithm based on reduced discernibility matrix is described as following:

Input: an incomplete decision table S (S= (U, A, V, f)), A= $C \cup D$, Reduce(C) = $\phi$

Output: the attribute reduction of incomplete decision table

Step1: Calculate the dependence of decision attribute set on condition attribute set, denoting $\delta_C(x)$.

Step2: Construct discernibility matrix according to definition 6.If there exist such relationship as $m_{ij} \subseteq m_{ij}^{'}$, then $M^{'} = M^{'} \cup m_{ij}$.

Step3: For any $m_{ij}^{'} \in M^{'}$, count the sum of $m_{ij}^{'}$, denoting A ($m_{ij}^{'}$).

Step4: Choosing A ($m_{ij}^{'}$) =1, R=R $\cup\{a\}$, C=C-{a} and removing the elements of including elements with value of "a" in $M^{'}$, if $M^{'} \neq \phi$, the algorithm terminates and outputs the reduced attribute R.

Step5: As to any a belonging to C, count the number of attribute "a"'s emerging time, denoting Sum (a). Choosing $\max_{a \in C}(sum(a))$, the attributes which meet the condition are not unique, select any one. R=R $\cup\{a\}$, C=C-{a} and removing the elements of including elements with value of "a" in $M^{'}$, if $M^{'} \neq \phi$, go to step4.

In the algorithm, the time complexity of step1 is O(|C||U|).The time complexity of step2 is O(|C| $\sum_{1 \le i < j \le |U|} |T_C(x_i)||T_C(x_j)|$). The time complexity of step3 is the same as the step2. The time complexity of step4 and step5 are O (|C|$^2$ $\sum_{1 \le i < j \le |U|} |T_C(x_i)||T_C(x_j)|$) respectively. So the time complexity of the algorithm is O (|C|$^2$ $\sum_{1 \le i < j \le |U|} |T_C(x_i)||T_C(x_j)|$).

For $\sum_{1 \le i \le |U|} |T_C(x_i)| << |U|$,

O (|C|$^2$ $\sum_{1 \le i < j \le |U|} |T_C(x_i)||T_C(x_j)|$) <O (|C|$^2$|U|$^2$).

The key point is to store the discernibility matrix coming from incomplete decision table while carries out attribute reduction based on discernibility matrix. Because of referring the additional information ( $\delta_C(x)$ ) while construct the discernibility matrix, the space complexity of the algorithm is O(|C| $\sum_{1 \le i < j \le |U|} |T_C(x_i)||T_C(x_j)|$ ). In addition, the redundant elements in the discernibility matrix are deleted in the algorithm .So the space complexity is less than O (|C| $\sum_{1 \le i < j \le |U|} |T_C(x_i)||T_C(x_j)|$ ) generally. Through the above discussion, the time and space complexity are better than the algorithm of paper [5].

## V. EXPERIMENTS

We take the car information table discussing in paper [9] as test information table in the algorithm, which is described in table 1 as following, where $U = \{x_1, x_2 ... x_6\}$ ,condition attribute set and decision attribute set are expressed as C={price, mileage ,size ,max-speed} and D={d} respectively.

TABLE Ⅰ. INCOMPLETE CAR TABLE

| *Car* | *Price* | *Mileage* | *Size* | *Max-speed* | *d* |
|---|---|---|---|---|---|
| 1 | High | High | Full | Low | Good |
| 2 | Low | * | Full | Low | Good |
| 3 | * | * | Compact | High | Poor |
| 4 | High | * | Full | High | Good |
| 5 | * | * | Full | High | Excellent |
| 6 | Low | High | Full | * | Good |

For the six objects in the incomplete decision table, calculate the tolerance relation separately on U according definition 4.

$T_C(x_1) = \{x_1\}$ ; $T_C(x_2) = \{x_2, x_6\}$ ; $T_C(x_3) = \{x_3\}$ ;

$T_C(x_4) = \{x_4, x_5\}$ ; $T_C(x_5) = \{x_4, x_5, x_6\}$ ; $T_C(x_6) = \{x_2, x_5, x_6\}$.

From definition 5, we can obtain the following results further: $\delta_C(x_1)$ =1, $\delta_C(x_2)$ =1, $\delta_C(x_3)$ =1, $\delta_C(x_4)$ =1/2, $\delta_C(x_5) = 1/3$, $\delta_C(x_6) = 2/3$. And the discernibility matrix is expressed as $M^{'}$, where $M^{'}$ = {0001, 1000, 0010}.Then the reduced attribute described as reduce(R) is {price, size, max-speed}. At this time $M^{'} = \phi$ and the algorithm terminates.

## VI. CONCLUSION

Attribute reduction is an important application on data analysis in rough set theory. The paper studies the attribute reduction based on discernibility matrix combining the dependence of decision attribute on condition attribute in incomplete table and presents a new attribute reduction algorithm based on incomplete decision table, which improves the two aspects of time and space complexity. Finally, the algorithm is proved to be effective by experiments.

## REFERENCES


[1] Pawlak Z.Rough sets [J]. International Journal of Computer and Information Science, 1982, 11(5):341-356.

[2] Pawlak Z,Skowron A. Rough set: Some ectensions[J], Information Science. 2007, 177 :41-73.

[3] Zhang weixu, Wu zhiwei, Liang jiye. Theories and Methods of Rough Set [M], Science Press,2001.

[4] Wang guoyin, Zhang qinghua. Uncertainty of Rough Sets in Different Knowledge Granularities [J], Chinese Journal of Computers, 2008,31(9):1588-1598.

[5] Shu wenhao,Xu zhangyan,Qian wenbin,Yang bingru. Quick Attribute Reduction Algorithm Based Imcomplete Decision Table [J], Chinese Journal of Computers, 2011,32(9),1867-1871.

[6] Teng shuhua ,Zhou shilin,Sun jixiang, Li zhiyong, Attribute Reduction Algorithm Based on Conditional Entropy under Incomplete Information System[J].Journal of National University of Defense Technology, 2010, 32(1),90-94.

[7] Li xiukai,Shi kaiquan, A Knowledge Granulation-based Algorithm for Attribute Reduction under Incomplete Information Systems[J], Compute Sciences,2006,33(11),169-170.

[8] Ge hao,Li longshu,Yang chuanjian, Discernibility Matrix Based on Credibility and Attribute Reduction Method[J],Journal of Sichuan University(Engineering Science Edition),2011,43(5),146-152.

[9] Kryszkiewicz M, Rough set approach to incomplete information systems [J]. Information Sciences,1998,112(1):39-49.