A New Method of Attribute Reduction Based On Information Quantity in An Incomplete System

Xu E
College of Information Technology, Bohai University, Jinzhou 121000, P.R.China
Email: EXU21@163.COM

Yuqiang Yang
College of Information Technology, Bohai University, Jinzhou 121000, P.R.China
Email: JZYANGYUQIANG@126.COM

Yongchang Ren
College of Information Technology, Bohai University, Jinzhou 121000, P.R.China
Email: BDRENYONGCHANG@YAHOO.COM.CN

***Abstract*—For an incomplete information system, attribute reduction is an important problem. To dealing with it, this paper proposed a new attribute reduction method based on information quantity. On one hand, this approach improved traditional tolerance relationship calculation methods using an extension of tolerance relationship in rough set theory. On the other hand, a new method was present for calculating the core attributes based on the extensive tolerance relationship, which can get core attribute set directly. And more, the method took attribute significance as the heuristic knowledge to calculate the candidate attribute expansion. Experiment results show that the method is simple and effective.**

***Index Terms*—attribute reduction, rough set, information quantity, an incomplete information system**

I. INTRODUCTION

Rough set theory was put forward by Prof Pawlak who was a Polish mathematician in 1980s, which is a tool to deal with uncertainty and vagueness of data [1-3]. It can effectively analyze inaccurate, inconsistent and incomplete information. Based on the perspective of knowledge classification, rough set theory processes in the approximation space and under the premise of maintaining the ability of classification. Rough set theory searches the implicit knowledge and reveals the potential rules through knowledge reduction, which does not require priori rules, avoiding the impact of personal preferences. Attributes reduction is the essence of rough set theory, which is an important research content and hot spot of rough set theory. Attributes reduction is a significant way to acquire simple expression of knowledge from the information systems by eliminating redundancy of attributes under the condition of unchanging the classification ability of original knowledge. Founding on classical rough set theory, a great deal of research has been done in complete information systems and a lot of effective attributes reduction methods have been put forward. However, in real life, due to errors of data measurement or misunderstanding or the restrictions or some other reasons, knowledge acquisition often faces with incomplete systems. There may be some objects with unknown attribute values, which greatly obstruct the development of rough set theory in practice. Thus, it is necessary to do some research for getting some methods to process the incomplete information systems via rough set theory.

There are two main approaches to deal with incomplete information systems using rough set theory [4-6]: One is the indirect approach, that is, incomplete information can be completed through some certain methods that are called data filling. The other is the direct method, that is, appropriate extensions are carried out in rough set theory to deal with the incomplete information systems. Indirect method is to deal with null values, where incomplete information system is firstly transformed into a complete information system via data filling method, and then is treated as complete information system[7-10]. There are some indirect methods such as using statistical analysis to fill the null values, using other condition attribute values and decision attribute values or relationship attributes to estimate the null values, asking experts to give the estimated value of the null values in accordance with some certain conditions, using Bayesian model and evidence theory to filll the missing data. But there are many drawbacks in indirect methods. For examples, Bayesian model needs to know the probability density; evidence theory requires evidence functions, which are often difficult to get, subjectivity and arbitrariness are also big concerns in some of these methods. As the computation complexity is too high, the efficiency is extremely low. Some of these methods can not deal with incomplete information system when there are a lot of null values in information systems, and at the same time the knowledge we get may not be reliable[11-14]. As the result in contrast with the indirect

doi:10.4304/jsw.7.8.1881-1888

method, the direct method maintains the original structure of information systems, and avoids human subjectivity. Direct methods are also more effective and reliable in dealing with the situation with many missing data[15,16].

In information systems with massive data sets, due to huge number of attributes and examples, the attribute reduction algorithm efficiency is particularly complex. So far there is no accepted and efficient algorithm in reduction algorithms based on rough set theory[18]. In practical applications, it is required to obtain as a relative attribute reduction.

To deal with the attribute reduction in an incomplete system, deficiency of attributes reduction algorithm in reference[12] was discussed and analyzed; Secondly, tolerance relationship calculation method was improved, and then a new method of seeking core attributes was given; Thirdly, a new attributes reduction algorithm based on information quantity in incomplete information system was designed.; Finally, the example was done and shows that the algorithm is effective.

II. ROUGH SET CONCEPTS AND THEOREMS

In order to describe attributes reduction, we define some conception and prove some thorems as below.

A. Rough Set Concepts

Information System: In rough set, an information system can be represented as

$$S=(U,A,V,f) \tag{1}$$

Where U is the universe, a finite set of N objects $U=\{x_1,x_2,...,x_n\}$, A is a finite set of attributes, which are divided into disjoint sets, i.e. $A=C\cup D$, where C is the set of condition attributes and D is the set of decision attribute. $V=\bigcup_{q\in A}V_q$ is the total decision function such that $f(x,q)\in V_q$ for every $q\in A, x\in U$.

Incomplete Information System: Suppose information system $S=(U,A,V,f)$, where U is universe; A is finite, nonempty set of attributes; C is condition attributes set and D is decision attribute set, $A=C\cup D$, $C\cap D=\varphi$; V is value domain of A; $f:A\to V$ is the mapping from attributes to domain; If there is at least an attribute $a\in C$ contains the null value, that is $f(x, a)=*$, then this information system is called incomplete information system, or it is called complete information system. Information systems often abbreviates to write as (U, A). Set B with missing attribute values, $B\in A$, missing values are remark as"*".

Tolerance Relationship: In order to deal with incomplete information systems, the tolerance relationship is an extension of equivalence relationship in rough set. In incomplete information system $S=(U,A,V,f)$, the tolerance relationship T is defined as follow:

$$\forall_{x,y\in U}(T_B(x,y)\Leftrightarrow\forall_{c_j\in B}(c_j(x)=c_j(y)\vee c_j(x)=*\vee c_j(y)=*)) \tag{2}$$

T is reflexive and symmetric, but not necessarily transitive. Via T, the definition of tolerance class is:

$$T_B(x)=\{y|y\in U\wedge T_B(x,y)\} \tag{3}$$

Based on the tolerance relationship, Attribute set $X\subseteq U$, the X lower approximations can be defined as:

$$D_B(X)=\{x|x\in U\wedge T_B(x)\subseteq X\} \tag{4}$$

Information Quantity: Give incomplete information system $S=(U,A,V,f)$, $A=C\cup D$, information quantity of attribute set $B\subseteq C$ is defined as:

$$I(B)=1-\frac{1}{|U|^2}\sum_{i=1}^{|U|}|T(x_i)| \tag{5}$$

where $U=\{x_1,x_2,...,x_n\}$, $|X|$ express the cardinality of set X.

Condition Information Quantity: Give incomplete information system $S=(U,A,V,f)$, $A=C\cup D$, attribute set $B\subseteq C$, the condition information quantity with respect to D is defined as follow: $I(B|D)=I(B\cup D)-I(B)$.

Positive Region: In incomplete information system $S=(U,A,V,f)$, P and Q are two knowledge of universe, $POS_P(Q)$ denotes positive region of Q with respect to P.

$$POS_P(Q)=\bigcup_{X\in U/Q}D_P(X) \tag{6}$$

Attributes Significance: In incomplete information system $S=(U,A,V,f)$, significance of attribute $b\notin B\subseteq C$ is defined as:

$$Sig_B(b)=I(B|D)-I((B\cup\{b\})|D) \tag{7}$$

Attributes Reduction: In incomplete information system $S=(U,A,V,f)$, if and only if R satisfies both the condition (1) and (2), attributes set R（$R\subseteq C$）is an attributes reduction of D with respect to B.

Condition: (1) $I(R|D)=I(C|D)$

(2) $\forall b\in R\Rightarrow I((R-\{b\})|D)\neq I(C|D)$

A New Compatible Granular Information Ssystem: Assumed $U/C=\{[x_1']_c,[x_2']_c,...,[x_m']_c\}$ and $POS_C(D)=\{[G_1]_C\cup[G_2]_C\cup...\cup[G_t]_C\}$, $GR=\{x_1',x_2',...,x_m'\}$, $\{G_1,G_2,...,G_t\}\subseteq GR$, $|[G_s]_C/D|=1$, called $GR_{POS}=\{G_1,G_2,...,G_t\}$, $GR=GR_{POS}\cup GR_{NEG}$, a new compatible granular information system is defined as bellow.

$$GRS=(GR,C,D,V',f') \tag{8}$$

B. THROREMS AND PROOF

Theorem 1: Given $S=(U,A,V,f)$ is an incomplete information system, attributes set $B(B\subseteq C)$,

$c_{m+1} \in C - B$, $\forall x \in U$, under tolerance relationship in incomplete information system, $T_B(x) \supseteq T_{B\cup\{c_{m+1}\}}(x)$.

Proof: Take a random $y \in T_{B\cup\{c_{m+1}\}}(x)$, according to the definition of tolerance class, we can know that:

If $\forall c \in B\bigcup\{c_{m+1}\}(f(y,c)=f(x,c)\vee f(y,c)=*\vee f(x,c)=*)$, then $\forall c \in B\bigcup(f(y,c)=f(x,c)\vee f(y,c)=*\vee f(x,c)=*)$. So we can get $y \in T_B(x)$ as y is arbitrary. Proof finished.

Theorem 2: Given an incomplete informationsystem, $S=(U,A,V,f)$, $C=\{c_1,c_2,..,c_n\}$, $B=\{c_1,c_2,..,c_m\}$, $(1\le m\le n)$ $\forall x \in U$, if $T_B(x) \supset T_{B\cup\{c_{m+1}\}}(x)$, that is tolerance class of x changes after adding attribute c_{m+1}. If $\exists y \in T_B(x) - T_{B\cup\{c_{m+1}\}}(x)$ and y is satisfied with the following conditions, then c_{m+1} is a core attribute of incomplete information system.

(1) $f(x,D) \neq f(y,D)$;

(2) $\min\{|\partial_C(x)|,|\partial_C(y)|\}=1$;

(3) $\forall c_i \in \{c_{m+2},...,c_n\}$ $f(x,c_i)=f(y,c_i)$;

Proof: According to the condition: $\exists y \in T_B(x) - T_{B\cup\{c_{m+1}\}}(x)$, then we can know that $f(x,c_{m+1}) \neq f(y,c_{m+1})$, and because $y \in T_B(x)$, then we can know that $\forall c_j \in \{c_1,...,c_m\}$, $f(x,c_j)=f(y,c_j)$, according to condition 3 ,we can get that $\forall c_i \in \{c_{m+2},...,c_n\}$, $f(x,c_i)=f(y,c_i)$, it shows that there is only one attribute's value between x and y is different, the attribute is attribute c_{m+1}, the others are all same. According to the definition of tolerance relationship, we can know that $y \notin T_C(x)$, but $y \in T_{C-\{c_{m+1}\}}(x)$, and according to condition 1, $f(x,D) \neq f(y,D)$, as for $U/ind(D)=\{Q_1,...Q_r\}$, x and y can not belong to any division subset of D relative to U. Suppose $x \in Q_s$, $y \in Q_t$, $1 \le s,t \le r$, $s \neq t$, according to the definition of lower approximation under the tolerance relationship, we can get $y \notin Q_S$, then $y \notin D_{C-\{c_{m+1}\}}Q_S$,because of $y \in T_{C-\{c_{m+1}\}}(x)$, $T_{C-\{c_{m+1}\}}(x) \subset D_{C-\{c_{m+1}\}}Q_s$, then $x \notin D_{C-\{c_{m+1}\}}Q_S$, in the same way, $y \notin D_{C-\{c_{m+1}\}}Q_t$, $x \notin D_{C-\{c_{m+1}\}}Q_t$, according to the definition of D positive region relative to P, we can know that $x \notin POS_{C-\{c_{m+1}\}}(D)$ $y \notin POS_{C-\{c_{m+1}\}}(D)$, according to condition 2 $\min\{|\partial_C(x)|,|\partial_C(y)|\}=1$: (1) when $|\partial_C(x)|=1$, according to the definition of the generalized decision function, we can know that $T_C(x) \subseteq Q_s$, then $x \in D_C Q_s \subseteq POS_C(D)$, because $x \notin POS_{C-\{c_{m+1}\}}(D)$, then $POS_{C-\{c_{m+1}\}}(D) \subset POS_C(D)$, as positive region is monotonically increasing when the attributes are added under tolerance relationship, we cannot get positive region of complete attribute set C when retaining other all attributes except attribute c_{m+1}, so c_{m+1} is the core attribute of incomplete information system; (2) when $|\partial_C(y)|=1$, according to the definition of the generalized decision function, we can know that $T_C(y) \subseteq Q_t$, then $y \in D_C Q_t \subseteq POS_C(D)$, as $y \notin POS_{C-\{c_{m+1}\}}(D)$, then $POS_{C-\{c_{m+1}\}}(D) \subset POS_C(D)$, c_{m+1} is the core attribute of incomplete information system. Proof finished.

Theorem 3: Let $S=(U,A,V,f)$ where $U=U^0\bigcup U'$, U^0 is the total sample set with complete attribute values, U' is the sample set that we only know the partial values. $A=C^0\bigcup C'\bigcup D$, C^0 is the significant attribute set, C' is the redundant attribute set, D is the decision attribute set. $If\ \forall a \in U', \forall b \in U^0, \forall c \in C'$, $c(a)=c(b)$, then conclude that the information system's certainty is stable.

Proof: Let any classification of the system is $E_i \in U \mid IND(C), (i=1,2,...,m)$, where m is the number of the classification divided by the condition attribute set C, $\{X_1,X_2,...,X_n\}=U \mid IND(D)$, then for some certain classification $E \in U \mid IND(C)$, it's certainty to the decision attribute class is as following:

$$\mu_{\max}(E)=\max\left(\left\{|E\cap X_i|/|E| : X_i \in U \mid IND(D)\right\}\right)$$

Then we can induce the information system certainty as below formula:

$$\mu_{\max}(S)=\sum_{i=1}^{m}\frac{|E_i|}{|U|}*\mu_{\max}(E_i) \tag{8}$$

Based on the above formula of the information system certainty, we can discuss the above theorem from two angles:

if C' has only one element c, then we can regard the formula $E_i' \in U \mid IND(C^0\bigcup(c))$, $(i=1,2,...,m')$ as the classes determined by the condition attribute set $C=C^0\bigcup\{c\}$. Since c is the redundant attribute, we consequently obtain $U \mid IND(C^0)=U \mid IND(C^0\bigcup\{c\})$, namely adding redundant attribute can't affect the classes in the information system S, i.e. $E=E'$, therefore , we can induce result that if $\forall E,E \in U \mid IND(C^0)$, then there must be the formula $\exists E \in U \mid IND(C^0\bigcup\{c\})$, we can obtain the formula, $\mu_{\max}(E)=\mu_{\max}(E')$; that is to say, $\mu_{\max}(S)$ in the information table is not changed.

In the same way, if $C'=\{C_1',C_2',...,C_m'\}$ is the redundant attribute set, then the $\mu_{\max}(S)$ in the information table will not changed.

Theorem 4: If $P \subseteq C$ and $\forall a \in (C-P)$ in new granular space $GRS=(GR,C,D,V',f')$, then $GR/(P\cup\{a\})=\bigcup_{X\in GR/P}(X/\{a\})$.

Theorem 5: Assumed *GR* is a domain, *P* and Q are respectively the sets of equivalence relation in *GR*, if $P \subseteq C$, $Q \subseteq C$, then

$$GR / IND(P \cup Q) = GR / IND(P) \cap GR / IND(Q)$$

Proof:

Proof:

$$GR / IND(P \cup Q) = \bigcap_{a \in P \cup Q} IND(\{a\})$$

$$= (\bigcap_{a \in P - P \cap Q} IND(\{a\})) \cap (\bigcap_{a \in Q - P \cap Q} IND(\{a\})) \cap (\bigcap_{a \in P \cap Q} IND(\{a\}))$$

$$= ((\bigcap_{a \in P - P \cap Q} IND(\{a\})) \cap (\bigcap_{a \in P \cap Q} IND(\{a\}))) \cap$$

$$((\bigcap_{a \in Q - P \cap Q} IND(\{a\})) \cap (\bigcap_{a \in P \cap Q} IND(\{a\})))$$

$$= (\bigcap_{a \in P} IND(\{a\})) \cap (\bigcap_{a \in Q} IND(\{a\}))$$

$$= GR / IND(P) \cap GR / IND(Q)$$

From theorem 5, one conclusion was obtained as below

Assuming *P* and *Q* are equivalence relation of domain *GR*, when $1/|GR| < GD(R) < 1$, P and $Q \subseteq R$, $P = \{X_1, X_2, \cdots, X_n\}$, $Q = \{Y_1, Y_2, \cdots, Y_m\}$. In order to testify that granularity is drab diminishing with attribute increasing, it is right to testify

$$GD(P) - GD(P \cup Q) = \sum_{i=1}^{n} |IND(P)|^2 / |GR|^2 - \sum_{i=1}^{n} |IND(P \cup Q)|^2 / |GR|^2$$

$$= (\sum_{i=1}^{n} |IND(P)|^2 - \sum_{i=1}^{n} |IND(P \cup Q)|^2) / |GR|^2$$

$$= (\sum_{i=1}^{n} |IND(P)|^2 - \sum_{i=1}^{n} |IND(P) \cap IND(Q)|^2) / |GR|^2$$

III. THE NEW ATTRIBUTE REDUCTION ALGORITHM

A. Tolerance Class Algorithms

In reference [12], the basic idea of computing tolerance class is: when calculating each tolerance class $T_B(x)$ $(\forall x \in U)$, compare the other $|U| - 1$ objects' values of attributes set B in incomplete information system. So calculation of a tolerance class needs to calculate $|B|(|U| - 1)$ times and calculation of tolerance class $T_B(x)$ $(\forall x \in U)$ needs to calculate $|B||U|(|U| - 1)$ times. Reference [11] uses an important property of tolerance class to calculate tolerance class, that is $T_B(x) \supseteq T_{B \cup \{a\}}(x)$ $(a \in C - B, B \subseteq C)$. Hence, when calculating $T_{B \cup \{a\}}(x)$, compare x with other objects in tolerance class $T_B(x)$, do not compare x with the objects which are not in tolerance class $T_B(x)$. Thus it can reduce Computation greatly. However, tolerance relationship has another important property, that is T is reflexive and symmetric, $T_B(x, y) \Leftrightarrow T_B(y, x)$. Therefore, when comparing x and other objects each time whether they are tolerance relationship, f x and y is tolerance relation, then puts y into $T_B(x)$, at the same time, puts x into $T_B(y)$. So when calculating $T_B(y)$, x and y do not need to be compared repeatedly. It makes the calculation not more than $\frac{1}{2}|B||U|(|U| - 1)$. Therefore time complexity and calculation are greatly reduced.

Algorithm 1: Calculation of Tolerance Class

Calculate tolerance class $T_{B \cup \{c_{m+1}\}}(x_i)$ ($x_i \in U$, $C = \{c_1, c_2, .., c_n\}$, $B = \{c_1, c_2, .., c_m\}$, $1 \le m \le n$)

Input: $S = (U, A, V, f)$, $T_B(x_i)$.

Output: $T_{B \cup \{c_{m+1}\}}(x_i)$.

```
For i =1 to |U| do
{ T_{B∪{c_{m+1}}}(x_i) = T_{B∪{c_{m+1}}}(x_i) ∪ {x_i};
  If (f(x_i, c_{m+1}) == *)
  Then T_{B∪{c_{m+1}}}(x_i) = T_B(x_i);
  Else {
        For j=i+1 to |U| do
        {If (f(x_j, c_{m+1}) = *) ∨ f(x_j, c_{m+1}) = f(x_i, c_{m+1})
          T_{B∪{c_{m+1}}}(x_i) = T_{B∪{c_{m+1}}}(x_i) ∪ {x_j};
          T_{B∪{c_{m+1}}}(x_j) = T_{B∪{c_{m+1}}}(x_j) ∪ {x_i};
        }
      }
}
```

We use Table I to descript the algorithm.

TABLE I.
INCOMPLETE DECISION TABLE

U	a_1	a_2	a_3	a_4	D
1	1	1	2	1	1
2	2	*	2	1	1
3	*	*	1	2	2
4	1	*	2	2	1
5	*	*	2	2	3
6	2	1	2	*	1
7	1	1	2	1	1
8	2	*	2	1	1

According to the definition of tolerance class, we can calculate that:

$T_{a_1}(1) = \{1,3,4,5,7\}$,

$T_{a_1}(2) = \{2,3,5,6,8\}$,

$T_{a_1}(3) = \{1,2,3,4,5,6,7,8\}$,

$T_{a_1}(4) = \{1,3,4,5,7\}$,

$T_{a_1}(5) = \{1,2,3,4,5,6,7,8\}$,

$T_{a_1}(6)=\{2,3,5,6,8\}$,
$T_{a_1}(7)=\{1,3,4,5,7\}$,
$T_{a_1}(8)=\{2,3,5,6,8\}$,
$T_{\{a_1,a_2\}}(1)=\{1,3,4,5,7\}=T_{\{a_1,a_2\}}(4)=T_{\{a_1,a_2\}}(7)$,
$T_{\{a_1,a_2\}}(2)=\{2,3,5,6,8\}=T_{\{a_1,a_2\}}(6)=T_{\{a_1,a_2\}}(8)$,
$T_{\{a_1,a_2\}}(3)=\{1,2,3,4,5,6,7,8\}=T_{\{a_1,a_2\}}(5)$,
$T_{\{a_1,a_2,a_3\}}(1)=\{1,4,5,7\}$,
$T_{\{a_1,a_2,a_3\}}(2)=\{2,5,6,8\}$,
$T_{\{a_1,a_2,a_3\}}(3)=\{3\}$,
$T_{\{a_1,a_2,a_3\}}(4)=\{1,4,5,7\}$,
$T_{\{a_1,a_2,a_3\}}(5)=\{1,2,4,5,6,7,8\}$,
$T_{\{a_1,a_2,a_3\}}(6)=\{2,5,6,8\}$,
$T_{\{a_1,a_2,a_3\}}(7)=\{1,4,5,7\}$,
$T_{\{a_1,a_2,a_3\}}(8)=\{2,5,6,8\}$.

B. Calculation of Core Attributes

By means of calculating Algorithm 1, we can get all $T_B(x)$ $(\forall x\in U, B\subseteq C)$. If we can find attributes' core of incomplete information system by these $T_B(x)$, attributes reduction will become more accurate and efficient. In detail, reference [5] analyses the conditions which core attributes should satisfy in the incomplete information system and it also prove the correctness and necessity of adding generalized decision function to solve the problem of incompatible and inconsistencies.

Algorithm 2: Calculation of Core Attributes:

Input: all tolerance class $T_B(x_i)$ $(\forall x\in U, B\subseteq C)$

Output: core attribute set Core

For $i=|C|$ to 0 do

{ $B=\{c_1,c_2,...,c_i\}$; $A=\{c_1,c_2,...,c_{i-1}\}$;

For j = 1 to $|U|$ do

{if $T_A(x_j)\neq T_B(x_j)$

{if$\exists y\in T_B(x)-T_{B\cup\{c_{m+1}\}}(x)$ and y satisfy theorem 2;

Then Core = Core$\bigcup\{c_i\}$;

Break;

}

}

}

C. Description of Attributes Reduction Algorithm in Incomplete Information System

Algorithm 3: Attributes Reduction

Input: $S=(U,A,V,f)$, where $C=\{c_1,c_2,..,c_n\}$, $U=\{x_1,x_2,..,x_r\}$.

Output: attributes reduction R

(1) Suppose $\forall x_i\in U(I_\phi(x_i)=U)$, use algorithm 1 to calculate all $I_B(x_i)(B\subseteq C, x_i\in U)$, and calculate condition information quality $I(C|D)$ and information quality $I(D)$;

(2) Make $R=\phi$, use algorithm 2 to calculate core attributes set of incomplete decision table and $R=R\bigcup Core$;

(3) Calculate $I(C|D)$ and $I(R|D)$, if $I(R|D)=I(C|D)$, then output attribute reduction R, break up the algorithm, otherwise execute step (4);

(4) If $C-R=\phi$, then output attribute reduction R, break up the algorithm, otherwise, for each $c_i\in C-R$, use algorithm 1 to calculate $Sig_R(C_i)$, suppose $Sig_R(t)=\max_{c_i\in C-R} Sig_R(C_i)$, $R=R\bigcup\{t\}$, according to definition of attribute significance, the attribute with the biggest significance is added into R, go to step(3).

The time complexity of step (1) is $\frac{1}{2}|C\cup D||U|^2$; The time complexity of step (2) needs only to extend $|C||U|^2$ in the worst circumstance and the best time complexity is $|C|$; After entering the Step(3), the time complexity of step (3) is $\frac{1}{2}|R\cup D||U|^2=\frac{1}{2}|C||U|^2$; From step (4) to step (3), the time complexity is $|C-R|\times\frac{1}{2}|C||U|^2=\frac{1}{2}|C|^2|U|^2$. From the above, we can know that the time complexity of the attributes reduction algorithm 3 is $\frac{1}{2}|C|^2|U|^2$.

We can see that the calculation times of step (4) to step (3) is the most in all steps of attributes reduction algorithm 3, but calculation of core attributes can greatly reduce the calculation. Sometimes, there is no need to execute step (4) at all. So in practice, the time complexity of the algorithm will be much less than $\frac{1}{2}|C|^2|U|^2$.

D. Description of A Decision Information System GRS

Algorithm 4: *GRS Algorithm*

Input: inconsistent decision information system table $S=(U,A,V,f)$, $U=(x_1,x_2,\ldots\ldots x_m)$, $C=\{C_1,C_2,\ldots.C_n\}$

Output: consistent decision information system *GRS*,

Step 1 GR=null;

t=1;

Gt={x1};

Step 2 For (I =2; $i\le m$; I ++)

If f (xi, c j)= f（xi-1，cj）and

f(xi, D）=f(xi-1,D)

Then Gt=Gt∪{xi};

flag=1;

Else

if f(xi, c j）= f（xi-1，cj）and

f(xi, D ≠f(xi-1,D)

then $G_t=G_t\cup\{x_i\}$;

f(G_i,D)=max(f(x_i,D)+1
flag=0;
Gt.count++;
Until S=null;
step 3 GR_{pos}=null;GR_{neg}=null;
if flag=1;
GR_{pos}= GR_{pos} ∪ {G_t};
Else
GR_{neg} = GR_{neg} ∪ {G_t};
GR= GR_{pos} ∪ GR_{neg}.

IV. ILLUSTRATION AND ANALYSIS

A. Illustration 1 Description

As shown in Table 1, we can do as below.

Step1: By the caculation, we can get that :

$T_{\{a_1,a_2,a_3,a_4\}}(1)=\{1,7\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(2)=\{2,6,8\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(3)=\{3\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(4)=\{4,5\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(5)=\{4,5,6\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(6)=\{2,5,6,8\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(7)=\{1,7\}$,

$T_{\{a_1,a_2,a_3,a_4\}}(8)=\{2,6,8\}$,

$T_{C\cup D}(1)=\{1,7\}$,

$T_{C\cup D}(2)=\{2,6,8\}$,

$T_{C\cup D}(3)=\{3\}$,

$T_{C\cup D}(4)=\{4\}$,

$T_{C\cup D}(5)=\{5\}$,

$T_{C\cup D}(6)=\{2,6,8\}$

$T_{C\cup D}(7)=\{1,7\}$,

$T_{C\cup D}(8)=\{2,6,8\}$,

$T_D(1)=\{1,2,4,6,7,8\}=T_D(2)=T_D(4)=T_D(6)$,

$T_D(3)=\{3\}$,

$T_D(5)=\{5\}$,

$I(C|D)=I(C\cup D)-I(C)=\frac{1}{16}$, $I(D)=\frac{13}{32}$.

Step 2: According to definition of generalized decision function, we can get:

$\partial_C(1)=\partial_C(2)==\partial_C(7)=\partial_C(8)=\{1\}$,

$\partial_C(3)=\{2\}$,

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

Then we find that $T_{\{a_1,a_2,a_3,a_4\}}(1)\neq T_{\{a_1,a_2,a_3\}}(1)$, $T_{\{a_1,a_2,a_3\}}(1)-T_{\{a_1,a_2,a_3,a_4\}}(1)=\{4,5\}$, and 5 satisfies the conditions of Theorem 2, then a_4 is the core attribute of incomplete information system, quit from circulation; Then we find out $T_{\{a_1,a_2,a_3\}}(3)\neq T_{\{a_1,a_2\}}(3)$, $T_{\{a_1,a_2\}}(3)-T_{\{a_1,a_2,a_3\}}(3)=\{1,2,4,5,6,7,8\}$, After calculating, we find that 4 satisfies the conditions of Theorem 2, so a_3 is the core attribute of incomplete information system, quit from this circulation . Next we will compare $T_{\{a_1,a_2\}}(x_i)$ and $T_{\{a_1\}}(x_i)$, the results show that $T_{\{a_1,a_2\}}(x_i)=T_{\{a_1\}}(x_i)$, according to the conditions of Theorem 2, we judge that there is no y exist , therefore a_2 is not core attribute. Finally, we compare $T_{\{a_1\}}(x_i)$ and $T_\phi(x_i)=U$, but there are not items which satisfy the conditions of Theorem 2, so a_1 is not core attribute too. Go to step 3.

Step 3 : Now the attributes reducetion set $R=\{a_3,a_4\}$, calculate the tolerance class under attribute a_3 and a_4, we can get that :

$T_{\{a_3,a_4\}}(1)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}}(2)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}}(3)=\{3\}$,

$T_{\{a_3,a_4\}}(4)=\{4,5,6,\}$

$T_{\{a_3,a_4\}}(5)=\{4,5,6,\}$,

$T_{\{a_3,a_4\}}(6)=\{1,2,4,5,6,7,8\}$,

$T_{\{a_3,a_4\}}(7)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}}(8)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}\cup D}(1)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}\cup D}(2)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}\cup D}(3)=\{3\}$,

$T_{\{a_3,a_4\}\cup D}(4)=\{4,6\}$,

$T_{\{a_3,a_4\}\cup D}(5)=\{5\}$,

$T_{\{a_3,a_4\}\cup D}(6)=\{1,2,4,6,7,8\}$,

$T_{\{a_3,a_4\}\cup D}(7)=\{1,2,6,7,8\}$,

$T_{\{a_3,a_4\}\cup D}(8)=\{1,2,6,7,8\}$,

From the above, calculate the information quality of attribute set R, calculation process as follows:

$$I(R|D) = I(a_3, a_4|D) - I(a_3, a_4) = \frac{1}{16}$$

As a result of $I(R|D) = I(C|D)$, the process ends, output attribute set $R = \{a_3, a_4\}$, R is the relative attributes reduction which we want to require.

In reference [12], it's attributes reduction algorithm can not calculate core attributes, but calculates every attribute's significance, and then computes the reduction. So it's calculation and time complexity is higher than the new one. This new method can calculate the tolerance class effectively and can acquire the core attribute directly, so there are less attributes which need to calculate the attribute significanc. In algorithm, bacause the calculation of attribute significanc is the largest, this method can greatly reduce time complexity. In this example, there is no need to calculate any attribute's significanc to acquire the attributes reduction set.

B. Illustration 2 Description

As shown in Table II, we can do as below.

In Table Ⅱ, an attribute a_5 is added into the

TABLE II
ONE INCONSISTENT DECISION INFORMATION SYSTEM S

U	a	b	c	d	D
X1	1	*	0	1	1
X2	1	2	0	1	1
X3	*	0	0	1	0
X4	0	0	1	2	1
X5	2	1	*	2	1
X6	0	0	1	2	2
X7	2	0	0	1	0
X8	0	*	2	2	1
X9	2	1	0	2	2
X10	*	0	0	1	0

incomplete information system. The calculation in step 1 of example 1 is still useful. We do not need to calculate again in step 1 of example 2.

Firstly， change the Table S into GRS as Table III.

Calculate IND(C) by the granularity formula.

$$IND(a) = \{\{G_1\}, \{G_2, G_4\}, \{G_3, G_5\}\}$$
$$IND(b) = \{\{G_1,\}\{G_2, G_3\}, \{G_4, G_5\}\}$$
$$IND(c) = \{\{G_1, G_2, G_4\}, \{G_3\}, \{G_5\}\}$$
$$IND(d) = \{\{G_1, G_4\}, \{G_2, G_3, G_5, G_6, G_7, G_8\}\}$$
$$IND(D) = \{\{G_1, G_2\}, \{G_3, G_4G_5\}\}$$

According to granularity fineness formula, obtain the fellowing result.

$$GD(a) = \frac{1^2 + 2^2 + 2^2}{5^2} = \frac{9}{25}$$
$$GD(b) = \frac{1^2 + 2^2 + 2^2}{5^2} = \frac{9}{25}$$
$$GD(c) = \frac{3^2 + 1^2 + 1^2}{5^2} = \frac{11}{25}$$
$$GD(d) = \frac{2^2 + 3^2}{5^2} = \frac{13}{25}$$

According to algorithm three, the thinner the granularity, the higher the distinguish rate, so it is the more important to decide the thinnest granularity. According to algorithm, if the importance is same, it is important to decide the first attribute. So get Attribute *a*.

Similarly knowable:

$$IND(a \cup b) = \{\{G_1\}, \{G_2\}, \{G_3\}, \{G_4\}, \{G_5\}\},$$

so the attribute granularity is

$$GD(a \cup b) = \frac{1^2 + 1^2 + 1^2 + 1^2 + 1^2}{5^2} = \frac{1}{5}$$

According to the examples, it reduces space waste to reduce incompatible division table into compatible division table, and shorts search time. And it reduces a lot of unnecessary operations with incremental method to calculate the size.

The algorithm introduced fineness concept of knowledge granular based on knowledge granular definition, and redefine a simplified granular space to overcome the error in uncertainty system reducing, so we can reduce the uncertainty information system into compatible information system. The method not only applicable to incompatible information system, and experiments show that it can eliminate the repeat factors in original information table, and make the reduction in new simplified system. We design a new reasonable measurement granularity fineness calculation formula of attribute importance for the purpose to rapid reduce search space, and gives the recursive formula. Using this formula as heuristic information, we designed attribute reduction of which the time complexity is $\max(O(|C||GR|), O(|C|^2|U/C|)$ based on granularity fineness. The theoretical analysis and practical simulation results show that: the method greatly reducing waste of space; time complexity is relatively low; reduce the computation time in a certain extent, thus provide effective methods for calculating the minimum reduction. The advantage of using granularity to reduce is: making the meticulous division to the information system; working out the relatively accurate reduction. However, it increase time and space complexity undoubtedly in large list. It will be our further work to fuse the granularity fineness importance and discernibility matrix.

C. Experiment

To test the method better, some practical data was extracted and an experiment was done, the result was as Figure 1.

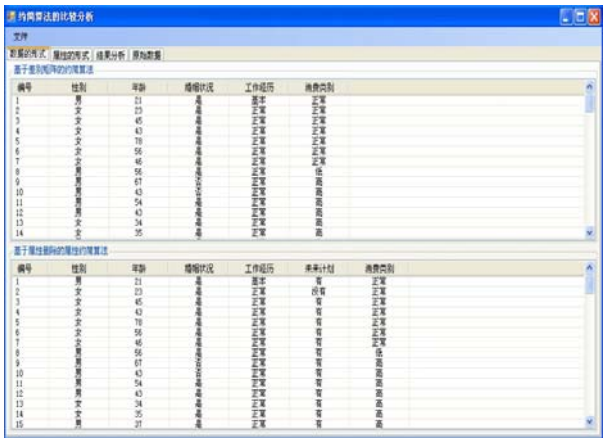

Figure 1. Note how the caption is centered in the column.

V. CONCLUSION

In this paper, at first discusses the shortcomings of ordinary attribute reduction based on information quantity in incomplete information systems; And via using an important property of tolerance class, presents an improved calculating algorithm of tolerance class. This method greatly reduces the calculating complexity of tolerance class; Secondly, by analyzing the computed results of tolerance class, presents a new method of calculating core attributes based on information quantity in incomplete information system and proves that it is correct; Finally, using information quantity as heuristic information and as the condition to determine whether it is an attributes reduction, designs a new attributes reduction algorithm under tolerance relation in incomplete information system. The analysis of the realistic example shows that the algorithm is accurate and effective. This algorithm of attributes reduction can acquire core attributes directly based on the tolerance class' computation and greatly reduce the calculating complexity of attributes reduction. The algorithm also is a basis for the research of attributes reduction when there are one or more attributes adding into the attributes set of incomplete information system.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 60674056, 70771007,70971059; Liaoning doctoral funds under Grant No.20091034, Liaoning higher education funds under Grant No. 2008T090 and Chinese postdoctoral funds under Grant No.20100471475.

REFERENCES

[1] Pawak Z, Grzymala-Busse J , Slowinski R, "Rough sets", *Communications of the ACM*, vol.8, pp.89-95,1995.

[2] Pawak Z, "Rough set theory and its application to data analysis". *Cybernetics and Systems*, vol.9, no.4, pp. 661-668, 1998.

[3] Pawlak Z, "Rough sets and intelligent data analysis", *Information Sciences*, vol.147, no.124, pp.1212-1218, 2002.

[4] Krysikiewicz M, "Rough set approach to incomplete information system", *Information Sciences*, Vol.112, pp.39-49, 1998.

[5] E Xu, Shao Liangshan, Ye Baiqing, Li Sheng, "Algorithm for rule extraction based on rough set", *Journal of Harbin Institute of Technology*, Vol 14, pp. 34-37, 2007.

[6] Wang guoyin, "Incomplete information system rough set extension", *Computer Research and Development*, vol.39, no.8, pp. 1238-1243, 2002.

[7] Liang J Y, Xu Z B, "An algorithm for knowledge in incomplete in－formation systems", *International Journal of Uncertainty, Fuzzinessand Knowledge-based Systems*, vol.10, no.1, pp. 95-103, 2002.

[8] Hexiangang, Huangbing, Wenpingchuan, "A heuristic algorithm for reduction of knowledge under incomplete information systems", *Piezoelectrics&Acoustooptics*, vol.26, no.2, pp. 158-160, 2004.

[9] Li Xiu-Hong , SHI Kai-Quan, "A knowledge granulation-based algorithm for attribute reduction under incomplete information systems", *Computer Science*, vol.33, no.11, pp.169-171, 2006.

[10] Guoyin Wang, "Calculation methods for core attributes of decision table", *Chinese Journal Of Computers*, Vol.26, No.6, pp.622-615, 2003.

[11] Huang B, He X, Zhou X Z, "Rough computational methods based on tolerance matrix" , *Acta Automatica Sinica*, vol.30, no.2, pp. 363-370, 2004.

[12] Huang Bing,Zhou Xian-zhong, Zhang Rong-rong, "Attribute Reduction Based on Information Quantity under Incomplete Information Systems" , *System Engineering-Theory&Practice*, vol.25, no.4, pp. 55-60, 2005.

[13] Zhang Qing-guo,Zhang Xue-feng,Zhang Ming-de,YU Yi-ke, "New attribute reduction algorithm of incomplete decision table of information quantity", *Conputer Engineering AndApplications*, vol.46, no.2, pp.19-21, 2010.

[14] Duoqian Miao, Guirong Hu, "A heuristic algorithm for reduction of knowledge", *Journal Of Computer Research And Development*, Vol.36, No.6, pp.681-684, 1999.

Xu E was born in 1971. He received his Ph.D. degree from University of Science and Technology of Beijing in 2006. He is now a professor in the College of Information Technology, Bohai University. His recent research interests include data mining, knowledge discovery in database and artificial intelligence.

Yuqiang Yang was born in 1965. He received his master degree in the School of Electronic and Informational Engineering, Beihang University in 1998. His recent research interests include information system, knowledge discovery in database and artificial intelligence.

Yongchang Ren was born in 1969. received his Ph.D. degree from Liaoning Technical University in 2008. He is now a professor in the College of Information Technology, Bohai University. His recent research interests include software management, knowledge discovery in database and artificial intelligence.