Metric Based Attribute Reduction in Incomplete Decision Tables*

Long Giang Nguyen[1] and Hung Son Nguyen[2]

[1] Institute of Information Technology, VAST, Vietnam
nlgiang@ioit.ac.vn
[2] Institute of Mathematics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland
son@mimuw.edu.pl

Abstract. Metric technique has recently been applied to solve such data mining problems as classification, clustering, feature selection, decision tree construction. In this paper, we apply metric technique to solve a attribute reduction problem of incomplete decision tables in rough set theory. We generalize Liang entropy in incomplete information systems and investigate its properties. Based on the generalized Liang entropy, we establish a metric between coverings and study its properties for attribute reduction. Consequently, we propose a metric based attribute reduction method in incomplete decision tables and perform experiments on UCI data sets. The experimental results show that metric technique is an effective method for attribute reduction in incomplete decision tables.

Keywords: Rough sets, feature selection and extraction, Liang's entropy, metric based reducts.

1 Introduction

Classical rough set theory based on equivalent relation has been introduced by Pawlak [11] as one of the effective tools for rule induction, object classification in complete decision tables. Attribute reduction is one of the crucial problems in rough set theory. Recently, there have been many attribute reduction algorithms in complete decision tables based on the equivalent relation [17]. In fact, there are many cases that decision tables contain missing values for at least one conditional attribute in the value set of that attribute and these decision tables are called incomplete decision tables. To extract decision rules directly from incomplete decision tables, Kryszkiewicz [5] has extended the equivalent relation in classical rough set theory to tolerance relation and proposed tolerance rough set. Based on the tolerance relation, many uncertainty measures and attribute reduction algorithms for incomplete decision tables have been investigated [7],

* The authors are supported by grants 2011/01/B/ST6/03867 from the Polish National Science Centre (NCN), and the grant SP/I/1/77065/10 in frame of the strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information" founded by the Polish National Centre for Research and Development (NCBiR).

D. Ciucci et al. (Eds.): RSFDGrC 2013, LNAI 8170, pp. 99–110, 2013.

[8], [9], [12], [13]. Huang et al [4] proposed an attribute reduction algorithm based on information quantity. Zhou et al [22], Huang et al [3] proposed attribute reduction algorithms based on tolerance matrix. The time complexity of these algorithms is $O\left(|A|^3|U|^2\right)$, where $|A|$ is the number of conditional attributes and $|U|$ is the number of objects. Zhang et al [21] improved the algorithm from [4] and the time complexity is down to $O\left(|A|^2|U|^2\right)$. Dai et al [1] presented an attribute reduction algorithm based on the coverage of an attribute set.

Metric is a distance measure between two sets [2]. In recent researches, metric technique has been applied to solve problems in data mining and rough set theory. Mantaras [16], Simovici and Jaroszewicz [18], [19] used a metric as the attribute selection criterion in the process of decision tree construction. Nguyen [10] proposed a metric based attribute reduction method in complete decision tables. Qian et al [14], [15] proposed knowledge distances between coverings in incomplete information systems and investigate its properties.

In this paper, we propose a metric based attribute reduction method in incomplete decision tables. Firstly, we generalize Liang entropy [6] in incomplete information systems and investigate its properties. Secondly, we establish a metric between coverings based on the generalized Liang entropy and study its properties in incomplete decision tables for attribute reduction. Finally, we define a reduct based on the metric, significance of attribute based on the metric and propose an attribute reduction heuristic algorithm in incomplete decision tables. The time complexity of proposed algorithm is $O\left(|A|^2|U|^2\right)$.

The structure of this paper is as follows. Section 2 presents the concept of attribute reduction in rough set theory. Section 3 presents a generalized Liang entropy in incomplete information systems and investigate its properties. Section 4 establishes a metric between coverings based on the generalized Liang entropy and study its properties. Section 5 presents a metric based attribute reduction method in incomplete decision tables. In Section 6, we perform some experiments of the proposed algorithm. The conclusions are presented in the last section.

2 Basic Notions

In this section, we introduction some basic concepts in rough set theory related to attribute reduction.

An information system [11] is a pair $\mathbb{S} = (U, A)$, where U is a non-empty, finite collection of objects and A is a non-empty, finite set, of attributes. Each $a \in A$ corresponds to the function $a : U \to V_a$, where V_a is called the value set of a. Elements of U can be interpreted as, e.g., cases, patients, observations, etc. Without loss of generality, we will assume that $U = \{u_1, ..., u_{|U|}\}$.

For a given information system $\mathbb{S} = (U, A)$, the function $\mu_{\mathbb{S}} : \mathbb{P}(A) \longrightarrow \mathbb{R}^+$, where $\mathbb{P}(A)$ is the power set of A, is called *the monotone evaluation function* if:

1. $\mu_{\mathbb{S}}(B)$ can be computed using information from B and U for any $B \subset A$;
2. $\mu_{\mathbb{S}}(.)$ is monotone, i.e., for any $B, C \subset A$, if $B \subset C$, then $\mu_{\mathbb{S}}(B) \leq \mu_{\mathbb{S}}(C)$.

In rough sets, reducts are the minimal subsets (with respect to the set inclusion) of attributes that contain a necessary portion of *information* about the objects, expressed by a *monotone evaluation function*.

Definition 1 (μ-reduct). *Any set $B \subseteq A$ is called* the reduct relative to a monotone evaluation function μ*, or briefly* μ-reduct*, if B is the smallest subset of attributes that $\mu(B) = \mu(A)$, i.e., $\mu(B') < \mu(B)$ for any proper subset $B' \subsetneq B$. We denote by $\mathcal{RED}(\mathbb{S}, \mu)$ the set of all μ-reducts, i.e.,*

$$\mathcal{RED}(\mathbb{S}, \mu) = \{R \subset A : R \textit{ is } \mu\textit{-reduct of } \mathbb{S}\} \tag{1}$$

The attribute $a \in A$ is called core attribute *if a presents in all reducts of A. The set of all core attributes is denoted by*

$$CORE(\mathbb{S}, \mu) = \bigcap_{\mathcal{RED}(\mathbb{S},\mu)} R \tag{2}$$

This definition is general for many existing definitions of reducts. Let us mention some well-known types of reducts used in rough set theory.

2.1 Decision Table and Decision Reducts

A decision table is a special information system $\mathbb{D} = (U, A \cup D)$, where attributes are of two types: conditional attributes (the attributes from A), and decision attributes (the attributes from D). The conditional attributes are also called *conditions*, while the decision attributes are briefly called *decisions*.

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

$$IND(P) = \{(x, y) \in U \times U : inf_P(x) = inf_P(y)\}\,. \tag{3}$$

It is obvious that $IND\,(P)$ is an equivalence relation, as it is reflexive, symmetric and transitive, over the set U. Any element $u \in U$ the set $[u]_P = \{v \in U\,|(u, v) \in IND\,(P)\}$ is called the equivalent class. The relation $IND\,(P)$ constitutes a partition of U, which is denoted by

$$U/_P = \{[u]_P : u \in U\} \tag{4}$$

Let $\mathbb{D} = (U, A \cup D)$ be a decision table. Any set $D_i \in U/_D$ is called the decision class of $\mathbb{D}$. For any $B \subset A$, the set

$$POS_B(D) = \{u \in U : [u]_B \subseteq D_i \text{ for some } D_i \in U/_D\} \tag{5}$$

is called the *B-positive region of D*. The decision table $\mathbb{D}$ is called consistent if and only if $POS_A(D) = U$. Otherwise, $\mathbb{D}$ is called the inconsistent decision table. Any minimal subset B of A such that $POS_B(D) = POS_A(D)$ is called the *decision reduct* (or reduct based on positive region) of $\mathbb{D}$. It has been shown in [9] that $\mu_{POS}(B) = |POS_B(D)|$ is a monotone evaluation function. Thus:

Proposition 1. *The set of attributes $R \subseteq A$ is* decision reduct *if and only if it is μ-reduct with respect to the measure $\mu_{POS}(B) = |POS_B(D)|$.*

2.2 Entropy Based Methods

Let $\mathbb{D} = (U, A \cup D)$ be a decision table and $C \subset A$ is an arbitrary set of attributes. Suppose that $U/_C = \{C_1, C_2, ..., C_m\}$and $U/_D = \{D_1, D_2, ..., D_n\}$, the conditional Shannon entropy of D with respect to $C \subset A$ is defined as

$$H\left(D\,|C\right) = -\sum_{i=1}^{m} \frac{|C_i|}{|U|} \sum_{j=1}^{n} \frac{|C_i \cap D_j|}{|C_i|} \log_2 \frac{|C_i \cap D_j|}{|C_i|} \tag{6}$$

Proposition 2 ([19]). *Let $\mathbb{D} = (U, A \cup D)$ be a decision table. If $Q \subseteq P \subseteq A$ then $H\left(D\,|Q\right) \geq H\left(D\,|P\right)$. The equality holds when $\forall X_u, X_v \in U/_P,\ X_u \neq X_v$, if $(X_u \cup X_v) \subseteq Y_k \in U/_Q$ then $\frac{|X_u \cap D_j|}{X_u} = \frac{|X_v \cap D_j|}{X_v}$ for $\forall j \in \{1, 2, ..., n\}$.*

Thus $H\left(D|C\right)$ is monotone function with respect to set inclusion. Any μ-reduct with respect to entropy measure $\mu_{Ent}(C) = M - H\left(D\,|C\right)$, where M is a constant, is called *a reduct of $\mathbb{D}$ based on conditional Shannon entropy.*

Let $\mathbb{S} = (U, A)$ be a complete information system, for any $P \subseteq A$ the value

$$E(P) = \sum_{i=1}^{m} \frac{|P_i|}{|U|} \left(1 - \frac{|P_i|}{|U|}\right) \tag{7}$$

where $U/_P = \{P_1, ..., P_m\}$, is called *the Liang entropy* [6].

Let $P, Q \subseteq A$ be arbitrary sets of attributes and let $U/_P = \{P_1, ..., P_m\}$, $U/_Q = \{Q_1, ..., Q_n\}$. The *conditional Liang entropy* is defined as follows:

$$E(Q|P) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|Q_i \cap P_j|}{|U|} \frac{|Q_i^c - P_j^c|}{|U|} \tag{8}$$

where $Q_i^c = U - Q_i,\ P_j^c = U - P_j$ (see [6]).

It has been shown in [6] that both Liang entropy and conditional Liang entropy measures are monotone with respect to set inclusion. Thus the μ-reducts with respect to either $\mu_1(P) = E(P)$ or $\mu_2(P) = E(D|P)$ are called the Liang entropy based reducts.

3 Reducts for Incomplete Information Systems

An information system $\mathbb{S} = (U, A)$ is called *incomplete*, or IIS for short, if the value $a(u)$ is not always determined for $a \in A$ and $u \in U$. Furthermore, we will denote the missing value by * [5]. Analogically, incomplete decision table, briefly IDT, is an incomplete information system $\mathbb{D} = (U, A \cup \{d\})$ where $d \notin A$ and $* \notin V_d$. Let $\mathbb{S} = (U, A)$ be an IIS, for any $P \subseteq A$ we define a binary relation on U as follows:

$$SIM\left(P\right) = \left\{(u, v) \in U^2 : \forall a \in P, a(u) = a(v) \vee a(u) = * \vee a(v) = *\right\} \tag{9}$$

Let us notice that $SIM\left(P\right)$ is a tolerance relation (as it is reflexive and symmetric) on U and that $SIM\left(P\right) = \bigcap_{a \in P} SIM\left(\{a\}\right)$. For any object $u \in U$ and set of attributes $P \subset A$, the set $S_P(u) = \{v \in U : (u,v) \in SIM(P)\}$ is called *the tolerance class of* u, or granule of information. Let $K(P)$ denote the family of tolerance classes of all objects from U, called *the knowledge base of* P, i.e.

$$K(P) = U/_{SIM(P)} = \{S_P(u) : u \in U\} = \left\{S_P(u_1), S_P(u_2), ..., S_P\left(u_{|U|}\right)\right\}.$$

It is clear that the tolerance classes in $K(P)$ do not constitute a partition of U in general. They constitute a covering of U, i.e., $S_P(u) \neq \emptyset$ for every $u \in U$, and $\bigcup_{u \in U} S_P(u) = U$. We will denote by $COVER(U) = \{K(P) : P \subset A\}$ the set of all possible coverings on U defined by attributes from A. A partial ordered relation $(COVER(U), \preceq)$ can be defined on $COVER(U)$ as follows

1. $K(P)$ is the same as $K(Q)$, denoted by $K(P) = K(Q)$, if and only if $\forall u \in U,\ S_P(u) = S_Q(u)$.
2. $K(P)$ is finer than $K(Q)$, denoted by $K(P) \preceq K(Q)$, if and only if $\forall u \in U,\ S_P(u) \subseteq S_Q(u)$.

Let $\mathbb{S} = (U, A)$ be an IIS. The family $\omega = \{S_A(u) = \{u\} \,|u \in U\}$ is called *the discrete covering* and $\delta = \{S_A(u) = U \,|u \in U\}$ is called *the complete covering.*

Definition 2 (generalized Liang entropy). *Let $\mathbb{S} = (U, A)$ be an IIS and $P \subseteq A$. The* **generalized Liang entropy** *of P is defined by*

$$IE(P) = \sum_{i=1}^{|U|} \frac{1}{|U|}\left(1 - \frac{|S_P(u_i)|}{|U|}\right) = 1 - \frac{1}{|U|^2}\sum_{i=1}^{n} |S_P(u_i)| \tag{10}$$

where $|S_P(u)|$ denotes the cardinality of $S_P(u)$.

Obviously, we have $0 \le IE(P) \le 1 - \frac{1}{|U|}$. Function $IE(P)$ achieves the maximum value $1 - \frac{1}{|U|}$ if $K(P) = \omega$, and the minimum value 0 when $K(P) = \delta$.

Definition 3 (Conditional generalized Liang entropy). *Let $\mathbb{S} = (U, A)$ be an IIS and $P, Q \subseteq A$. The generalized Liang entropy of Q conditioned on P is defined by*

$$IE(Q\,|P) = \frac{1}{|U|}\sum_{i=1}^{|U|}\left(\frac{|S_P(u_i)| - |S_Q(u_i) \cap S_P(u_i)|}{|U|}\right) \tag{11}$$

It has been shown that Liang entropy $E(P)$ presented in [6] is a particular case of the generalized Liang entropy, and the conditional Liang entropy $E(Q\,|P)$ is a particular case of the conditional generalized Liang entropy $IE(Q\,|P)$. Moreover, let $\mathbb{S} = (U, A)$ be an IIS and $P, Q, R \subseteq A$, the following properties hold:

$P1$) If $K(P) \preceq K(Q)$ then $IE(P) \ge IE(Q)$ and $IE(P) = IE(Q)$ if and only if $K(P) = K(Q)$.
$P2$) If $K(P) \preceq K(Q)$ then $IE(P \cup Q) = IE(P)$.
$P3$) $IE(P \cup Q) \ge IE(P)$ and $IE(P \cup Q) \ge IE(Q)$.

$P4)$ $IE(P \cup Q) = IE(P) + IE(Q|P) = IE(P) + IE(P|Q)$.
$P5)$ $0 \le IE(Q|P) \le 1 - \frac{1}{|U|}$; the equality $IE(Q|P) = 0$ holds iff $K(P) \preceq K(Q)$ and the equality $IE(Q|P) = 1 - \frac{1}{|U|}$ holds iff $K(P) = \delta$ and $K(Q) = \omega$.
$P6)$ If $U/_{SIM(P)} \preceq U/_{SIM(Q)}$ then $IE(R|Q) \ge IE(R|P)$.
$P7)$ If $U/_{SIM(P)} \preceq U/_{SIM(Q)}$ then $IE(P|R) \ge IE(Q|R)$.
$P8)$ $IE(Q|P) + IE(P|R) \ge IE(Q|R)$.

Let $\mathbb{D} = (U, A \cup \{d\})$ be an IDT, Huang Bing et al [4] defined the reducts based on information quantity as the minimal subsets of attributes B such that $IE(B|\{d\}) = IE(A|\{d\})$ They are, in fact, the μ-reducts with respect to the *conditional generalize Liang entropy* measure, defined by

$$\mu_{IE}(B) = IE(B|\{d\}) = IE(B \cup \{d\}) - IE(B) \tag{12}$$

4 Metric between Coverings and Properties

Recall that any map $d : X \times X \to [0, \infty)$ that satisfies the following conditions:

$M1)$ $d(x, y) \ge 0$, $d(x, y) = 0$ if and only if $x = y$.
$M2)$ $d(x, y) = d(y, x)$.
$M3)$ $d(x, y) + d(y, z) \ge d(x, z)$.

for any $x, y, z \in X$ is called *a metric on* X [2].

The condition $M3)$ is called the triangular inequality. The pair (X, d) is called a metric space. Based on the generalized Liang entropy, in this Section we establish a metric between coverings and study some properties of the proposed metric for attribute reduction in incomplete decision tables.

Theorem 1 (Metric). *For any incomplete information system* $\mathbb{S} = (U, A)$*, the map* $d_E : COVER(U) \times COVER(U) \to [0, \infty)$*, defined by*

$$d_E(K(P), K(Q)) = IE(P|Q) + IE(Q|P) \tag{13}$$

where $P, Q \subset A$*, is a metric on* $COVER(U)$*.*

Proof. We will show that d_E satisfies three properties of metric functions:

(M1) From Property $P5)$ we have $d_E(K(P), K(Q)) \ge 0$ for any $P, Q \subset A$ and the equality holds if and only if $(IE(Q|P) = 0)$ and $(IE(P|Q) = 0)$, i.e.,

$$\left(U/_{SIM(P)} \preceq U/_{SIM(Q)}\right) \wedge \left(U/_{SIM(Q)} \preceq U/_{SIM(P)}\right) \Leftrightarrow K(P) = K(Q)$$

(M2) From the definition of d_E, it is easy to see that

$$d_E(K(P), K(Q)) = d_E(K(Q), K(P))$$

for any $K(P), K(Q) \in COVER(U)$.

(M3) For any $P, Q, R \subset A$, from Property $P5)$ we have

$$IE(Q|P) + IE(P|R) \ge IE(Q|R) \text{ and } IE(R|P) + IE(P|Q) \ge IE(R|Q)$$

Thus we have $d_E(K(Q), K(P)) + d_E(K(P), K(R)) \ge d_E(K(Q), K(R))$

Therefore all conditions $(M1), (M2), (M3)$ are satisfied, we can conclude that d_E is a metric on $COVER(U)$

The following propositions present some properties of the metric d_E. The proofs of those facts are omitted due to lack of space.

Proposition 3. *Let $\mathbb{S} = (U, A)$ be an incomplete information system. For any subsets $B, C \subset A$:*

$$a) \quad d_E\left(K\left(B\right), K\left(C\right)\right) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_B(u_i)| - |S_C(u_i)|}{|U|} \tag{14}$$

$$b) \quad \text{if } B \subseteq C \text{ then } d_E\left(K\left(B\right), K\left(B \cup \{d\}\right)\right) \geq d_E\left(K\left(C\right), K\left(C \cup \{d\}\right)\right) \tag{15}$$

Proposition 3 b) states that the bigger the attribute set B is, the smaller the metric $d_E\left(K\left(B\right), K\left(B \cup \{d\}\right)\right)$ is, and vice versa. In other words, the metric decreases as tolerance classes become smaller through finer classification.

5 Metric Based Reducts in Incomplete Decision Tables

In next content, we define the reduct based on the proposed metric and prove that this reduct is the same as the reduct based on information quantity.

Definition 4. *If the set of attributes $R \subseteq A$ satisfies the following conditions:*

(1) $d_E\left(K\left(R\right), K\left(R \cup \{d\}\right)\right) = d_E\left(K\left(A\right), K\left(A \cup \{d\}\right)\right)$
(2) $\forall r \in R,\ d_E\left(K\left(R - \{r\}\right), K\left(\left(R - \{r\}\right) \cup \{d\}\right)\right) \neq d_E\left(K\left(A\right), K\left(A \cup \{d\}\right)\right)$

then R is called a reduct of A based on metric.

Proposition 4. *Let $\mathbb{D} = (U, A \cup \{d\})$ be an incomplete decision table and $B \subseteq A$. Then $d_E(K(B), K(B \cup \{d\})) = d_E(K(A), K(A \cup \{d\}))$ if and only if*

$$IE\left(B\left|\{d\}\right.\right) = IE\left(A\left|\{d\}\right.\right).$$

Proof. Let us consider $U = \{u_1, ..., u_n\}$ and $B \subseteq A$. Since $B \subset B \cup \{d\}, A \subset A \cup \{d\}$, and $d_E\left(K\left(B\right), K\left(B \cup \{d\}\right)\right) = d_E\left(K\left(A\right), K\left(A \cup \{d\}\right)\right)$, it follows from Proposition 3 that

$$\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_B(u_i)| - \left|S_{B\cup\{d\}}(u_i)\right|}{|U|} = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_A(u_i)| - |S_{A\cup\{d\}}(u_i)|}{|U|} \Leftrightarrow$$

$$\Leftrightarrow \left(1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|S_{B\cup\{d\}}(u_i)\right|\right) - \left(1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}|S_B(u_i)|\right)$$

$$= \left(1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left|S_{A\cup\{d\}}(u_i)\right|\right) - \left(1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}|S_A(u_i)|\right)$$

According to Equation 12, the last equation is equivalent to

$$IE\left(B\cup\{d\}\right)-IE\left(B\right)=IE\left(A\cup\{d\}\right)-IE\left(A\right)$$

which is equivalent to $IE\left(B\left|\{d\}\right.\right)=IE\left(A\left|\{d\}\right.\right)$. This completes the proof.

Therefore, we can conclude from Proposition 4 that the reduct based on proposed metric is the same as that based on information quantity in incomplete decision tables.

Definition 5. *Let $\mathbb{D}=(U,A\cup\{d\})$ be an incomplete decision table and $B\subseteq A$. The significance of attribute $b\in A-B$ is defined as*

$$SIG_B(b)=d_E(K(B),K(B\cup\{d\}))-d_E(K(B\cup\{b\}),K(B\cup\{b\}\cup\{d\})),$$

where $S_{\emptyset}(u_i)=U$ for any $u_i\in U,i=1,...,|U|$.

Definition 5 implies that the significance of attribute $b\in A-B$ is measured by the changes of the metric $d_E\left(K\left(B\right),K\left(B\cup\{d\}\right)\right)$ when b is added to B, the bigger the value of $SIG_B(b)$, the more important the attribute b. This significance of attribute will be treated as the attribute selection criterion in our heuristic algorithm for attribute reduction

The heuristic search for short metric based reducts in incomplete decision tables is presented in Algorithm 1 (Algorithm **MBR**). In order to find the best reduct, the algorithm begins with $R=\emptyset$, then the most important attribute is chosen from searching space and added into R. The above processes are done until we get the best reduct.

Algorithm 1. MBR: metric-based reduct for incomplete decision table

Data: An incomplete decision table $\mathbb{D}=(U,A\cup\{d\})$;
Output: The short metric-based reduct R of $\mathbb{D}$;
$R=\emptyset$;
Calculate $d_E\left(K\left(R\right),K\left(R\cup\{d\}\right)\right)$ and $T=d_E\left(K\left(A\right),K\left(A\cup\{d\}\right)\right)$;
`// Iterative insertion of the most important attribute to` R
while $d_E(K(R),K(R\cup\{d\}))\neq T$ **do**
    for *each* $a\in A-R$ **do**
        Calculate $S=d_E(K(R\cup\{a\}),K(R\cup\{a\}\cup\{d\}))$;
        $SIG_R(a)=d_E(K(R),K(R\cup\{d\}))-S$;
    $R=R\cup\left\{\underset{a\in A-R}{ArgMax}\left\{SIG_R\left(a\right)\right\}\right\}$;
    Calculate $d_E\left(K\left(R\right),K\left(R\cup\{d\}\right)\right)$;
`// Deleting redundant attributes in` R
for *each* $a\in R$ **do**
    Calculate $d_E\left(K\left(R-\{a\}\right),K\left(R-\{a\}\cup\{d\}\right)\right)$;
    if $d_E\left(K\left(R-\{a\}\right),K\left(R-\{a\}\cup\{d\}\right)\right)=T$ **then** $R=R-\{a\}$
return R;

Let us consider While loop from command line 3 to 8. To calculate $SIG_R(a)$, we need to calculate $S_{R\cup\{a\}}(u_i)$, $S_{R\cup\{a\}\cup\{d\}}(u_i)$ because $S_R(u_i), S_{R\cup\{d\}}(u_i)$ have already calculated in the previous step. According to Zhang et al [21], the time complexity to calculate $S_{R\cup\{a\}}(u_i)$ for $\forall u_i \in U$ when $S_R(u_i)$ calculated is $O\left(|U|^2\right)$. So the time complexity to calculate all $SIG_E(a)$ is

$$(|A| + (|A| - 1) + ... + 1) * |U|^2 = (|A| * (|A| - 1)/2) * |U|^2 = O\left(|A|^2|U|^2\right),$$

where $|A|$ is the number of conditional attributes and $|U|$ is the number of objects. The time complexity to choose the attribute with maximum significance is $|A| + (|A| - 1) + ... + 1 = |A| * (|A| - 1)/2 = O\left(|A|^2\right)$. Hence, the time complexity of While loop is $O\left(|A|^2|U|^2\right)$. Similarly, the time complexity of For loop from command line 10 to 12 is $O\left(|A|^2|U|^2\right)$. Consequently, the time complexity of Algorithm 1 is $O\left(|A|^2|U|^2\right)$, which is less than that of [3], [4], [22]. However, the time complexity of Algorithm 1 is the same as that of [21].

5.1 Example

Table 1. Car descriptions

Car	*Price*	*Mileage*	*Size*	*Max-speed*	*d*
u_1	High	High	Full	Low	Good
u_2	Low	*	Full	Low	Good
u_3	*	*	Compact	High	Poor
u_4	High	*	Full	High	Good
u_5	*	*	Full	High	Excellent
u_6	Low	High	Full	*	Good

In this Section we consider the descriptions of cars as in Table 1 [4]. This is an incomplete decision table $\mathbb{D} = (U, A \cup \{d\})$, where

$$U = \{u_1, u_2, u_3, u_4, u_5, u_6\} \text{ and } A = \{Car, Price, Mileage, Size, Max\text{-}speed\}.$$

For simplification we will denote the attributes by a_1, a_2, a_3, a_4 respectively. Firstly, let us calculate the knowledge bases of the following sets of attributes:

$$\begin{aligned}
K(\{a_1\}) =& \{\{u_1, u_3, u_4, u_5\}, \{u_2, u_3, u_5, u_6\}, U, \{u_1, u_3, u_4, u_5\}, U,\\
& \{u_2, u_3, u_5, u_6\}\}\\
K(\{a_2\}) =& \{U, U, U, U, U, U\}\\
K(\{a_3\}) =& \{\{u_1, u_2, u_4, u_5, u_6\}, \{u_1, u_2, u_4, u_5, u_6\}, \{u_3\}, \{u_1, u_2, u_4, u_5, u_6\},\\
& \{u_1, u_2, u_4, u_5, u_6\}, \{u_1, u_2, u_4, u_5, u_6\}\}\\
K(\{a_4\}) =& \{\{u_1, u_2, u_6\}, \{u_1, u_2, u_6\}, \{u_3, u_4, u_5, u_6\}, \{u_3, u_4, u_5, u_6\},\\
& \{u_3, u_4, u_5, u_6\}, U\}
\end{aligned}$$

$$\begin{aligned} K(A) =& \{\{u_1\},\{u_2,u_6\},\{u_3\},\{u_4,u_5\},\{u_4,u_5,u_6\},\{u_2,u_5,u_6\}\} \\ K(\{d\}) =& \{\{u_1,u_2,u_4,u_6\},\{u_1,u_2,u_4,u_6\},\{u_3\},\{u_1,u_2,u_4,u_6\},\{u_5\}, \\ & \{u_1,u_2,u_4,u_6\}\} \end{aligned}$$

According to lines 1 and 2 of Algorithm 1, we set $R = \emptyset$ and calculate

$$T = d_E\left(K\left(A\right), K\left(A \cup \{d\}\right)\right) = \frac{1}{|U|^2} \sum_{i=1}^{6} \left(\left|S_A(u_i) - \left(S_A(u_i) \cap S_{\{d\}}(u_i)\right)\right|\right) = \frac{4}{36}.$$

Now, we start the first iteration of the While loop by the calculation of attribute significance:

$$SIG_{\emptyset}\left(a_1\right) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left(\left|S_{\emptyset}(u_i) - S_{\{d\}}(u_i)\right| - \left|S_{\{a_1\}}(u_i) - S_{\{a_1,d\}}(u_i)\right|\right) = 0.$$

Similarly, $SIG_{\emptyset}\left(a_2\right) = 0$, $SIG_{\emptyset}\left(a_3\right) = \frac{10}{36}$, $SIG_{\emptyset}\left(a_4\right) = \frac{8}{36}$. Choose a_3 which haves the most significance and $R = \{a_3\}$. After calculation of

$$d_E\left(K\left(\{a_3\}\right), K\left(\{a_3, d\}\right)\right) = \frac{8}{36},$$

we can see that $d_E\left(K\left(\{a_3\}\right), K\left(\{a_3, d\}\right)\right) \neq d_E\left(K\left(A\right), K\left(A \cup \{d\}\right)\right)$. Thus we have to perform the second loop.

$$SIG_{\{a_3\}}(a_1) = \frac{2}{36}, SIG_{\{a_3\}}(a_2) = 0, SIG_{\{a_3\}}(a_4) = \frac{4}{36}.$$

Choose a_4 which haves the most significance and $R = \{a_3, a_4\}$. Calculate

$$d_E\left(K\left(\{a_3, a_4\}\right), K\left(\{a_3, a_4, d\}\right)\right) = \frac{4}{36} = d_E\left(K\left(A\right), K\left(A \cup \{d\}\right)\right).$$

Hence, go to For loop. We can see that

$$d_E\left(K\left(\{a_3\}\right), K\left(\{a_3, d\}\right)\right) = \frac{8}{36} \neq T; \quad d_E\left(K\left(\{a_4\}\right), K\left(\{a_4, d\}\right)\right) = \frac{10}{36} \neq T.$$

As a consequence, the algorithm finishes and returns $R = \{a_3, a_4\}$ as the best reduct of A. This result is the same as the result in the example in reference [4].

6 Experiments

The experiments on PC (Pentium Dual Core 2.13 GHz, 1GB RAM, WINXP) are performed on 6 data sets obtained from UCI Machine Learning Repository [20]. We choose information quantity based attribute reduction algorithm [4] (IQBAR for short) to compare with the proposed algorithm. The results of experiments are showm in Table 2 and Table 3, where $|U|$, $|A|$, $|R|$ are the numbers of objects, primal condition attributes, and after reduction respectively, and t is the time of operation (calculated by second). Condition attributes will be denoted by $1, 2, ..., |A|$. The results show that the reduct of the proposed algorithm is the same as that of the IQBAR algorithm. However, the time of operation in the proposed algorithm is less than that in the IQBAR algorithm.

Table 2. The results of the proposed algorithm and IQBAR algorithm

Seq.	Data sets	$\|U\|$	$\|A\|$	Algorithm *IQBAR*		Algorithm **MBR**	
				$\|R\|$	Comp. time	$\|R\|$	Comp. time
1	Hepatitis	155	19	4	1.296	4	0.89
2	Lung-cancer	32	56	4	0.187	4	0.171
3	Automobile	205	25	5	3	5	1.687
4	Anneal	798	38	9	179	9	86.921
5	Voting Records	435	16	15	25.562	15	16.734
6	Credit Approval	690	15	7	29.703	7	15.687

Table 3. The reducts of the proposed algorithm and IQBAR algorithm

Seq	Data sets	The reducts of Alg. *IQBAR*	The reducts of Alg. **MBR**
1	Hepatitis	$\{1, 2, 4, 17\}$	$\{1, 2, 4, 17\}$
2	Lung-cancer	$\{3, 4, 9, 43\}$	$\{3, 4, 9, 43\}$
3	Automobile	$\{1, 13, 14, 20, 21\}$	$\{1, 13, 14, 20, 21\}$
4	Anneal	$\{1, 3, 4, 5, 8, 9, 33, 34, 35\}$	$\{1, 3, 4, 5, 8, 9, 33, 34, 35\}$
5	Voting Records	$\{1, 2, 3, 4, 5, 7, 8, 9, 10, 11,$ $12, 13, 14, 15, 16\}$	$\{1, 2, 3, 4, 5, 7, 8, 9, 10, 11,$ $12, 13, 14, 15, 16\}$
6	Credit Approval	$\{1, 2, 3, 4, 5, 6, 8\}$	$\{1, 2, 3, 4, 5, 6, 8\}$

7 Conclusion

Attribute reduction is one of the crucial problems in both rough set theory for complete information systems and tolerance rough set for incomple information systems. In this paper, a generalized Liang entropy is proposed based on Liang entropy [6] and some of its properties are considered in incomplete information systems. Based on the generalized Liang entropy, a metric is established between coverings and a metric based attribute reduction method in incomplete decision tables is proposed. To construct the metric based attribute reduction method, we define the reduct based on metric, the significance of an attribute based on metric. We use the significance of an attribute as heuristic information to design and implemement an efficient attribute reduction algorithm in incomplete decision tables. We also prove theoretically and experimentally that the reduct based on metric is the same as that base on information quantity [4] and the time complexity of the proposed algorithm is less than that of the information quantity based algorithm [4].

References

1. Dai, X.P., Xiong, D.H.: Research on Heuristic Knowledge Reduction Algorithm for Incomplete Decision Table. In: 2010 International Conference on Internet Technology and Applications, pp. 1–3. IEEE (2010)
2. Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer (2009)
3. Huang, B., He, X., Zhou, X.Z.: Rough Computational methods based on tolerance matrix. Acta Automatica Sinica 30, 363–370 (2004)

4. Huang, B., Li, H.X., Zhou, X.Z.: Attribute Reduction Based on Information Quantity under Incomplete Information Systems. Systems Application Theory and Practice 34, 55–60 (2005)
5. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Information Science 112, 39–49 (1998)
6. Liang, J.Y., Chin, K.S., Dang, C.Y., Richard, C.M.Y.: New method for measuring uncertainty and fuzziness in rough set theory. International Journal of General Systems 31, 331–342
7. Liang, J.Y., Qian, Y.H.: Axiomatic approach of knowledge granulation in information system. In: Sattar, A., Kang, B.-H. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 1074–1078. Springer, Heidelberg (2006)
8. Liang, J.Y., Qian, Y.H.: Information granules and entropy theory in information systems. Information Sciences 51, 1–18 (2008)
9. Liang, J.Y., Shi, Z.Z., Li, D.Y., Wierman, M.J.: The information entropy, rough entropy and knowledge granulation in incomplete information system. International Journal of General Systems 35(6), 641–654 (2006)
10. Nguyen, L.G.: Metric Based Attribute Reduction in Decision Tables. In: The 2012 International Workshop on Rough Sets Applications (RSA 2012), FedCSIS Proceedings, pp. 333–338 (2012), `http://fedcsis.org/proceedings/fedcsis2012/`
11. Pawlak, Z.: Rough sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers (1991)
12. Qian, Y.H., Liang, J.Y.: Combination Entropy and Combination Granulation in Incomplete Information System. In: Wang, G.-Y., Peters, J.F., Skowron, A., Yao, Y. (eds.) RSKT 2006. LNCS (LNAI), vol. 4062, pp. 184–190. Springer, Heidelberg (2006)
13. Qian, Y.H., Liang, J.Y.: New method for measuring uncertainty in incomplete information systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (2008)
14. Qian, Y.H., Liang, J.Y., Dang, C.Y.: Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. International Journal of Approximate Reasoning 50, 174–188 (2009)
15. Qian, Y.H., Liang, J.Y., Dang, C.Y., Wang, F., Xu, W.: Knowledge distance in information systems. Journal of Systems Science and Systems Engineering 16, 434–449 (2007)
16. Mantaras, R.L.: A Distance-Based Attribute Selection Measure for Decision Tree Induction. Machine Learning 6(1), 81–92 (1991)
17. Shifei, D., Hao, D.: Research and Development of Attribute Reduction Algorithm Based on Rough Set. In: IEEE, CCDC 2010, pp. 648–653 (2010)
18. Simovici, D.A., Jaroszewicz, S.: Generalized conditional entropy and decision trees. In: Proceeding of EGC, Lyon, France, pp. 369–380 (2003)
19. Simovici, D.A., Jaroszewicz, S.: A new metric splitting criterion for decision trees. International Journal of Parallel Emergent and Distributed Systems 21(4), 239–256 (2006)
20. The UCI machine learning repository, `http://archive.ics.uci.edu/ml/datasets.html`
21. Zhang, Q.G., Zheng, X.F., Xu, Z.Y.: Efficient Attribute Reduction Algorithm Based on Incomplete Decision Table. In: 2009 Second International Conference on Intelligent Computation Technology and Automation, pp. 192–195. IEEE (2009)
22. Zhou, X.Z., Huang, B.: Rough set-based attribute reduction under incomplete Information Systems. Journal of Nanjing University of Science and Technology 27, 630–636 (2003)