Annales Univ. Sci. Budapest., Sect. Comp. 42 (2014) 157–172

METRIC BASED ATTRIBUTE REDUCTION IN DYNAMIC DECISION TABLES

János Demetrovics (Budapest, Hungary)

Vu Duc Thi (Ha Noi, Viet Nam)

Nguyen Long Giang (Ha Noi, Viet Nam)

Dedicated to András Benczúr on the occasion of his 70th birthday

Communicated by András A. Benczúr

(Received June 1, 2014; accepted July 1,2014)

Abstract. In the past two decades, several results appeared on feature reduction applying rough set theory. However, most of these methods are implemented on static decision tables. Using a distance measure, in this paper we propose algorithms to find the reducts of decision tables when adding or deleting objects. Since we can avoid re-running the original algorithms over the entire set of objects, our methods significantly reduce the running time for attribute reduction in dynamic data.

1. Introduction

Attribute or feature selection is one of the crucial problems of data mining and machine learning. Feature selection methods that apply rough set theory are also called attribute reduction. Attribute reduction in decision tables aims to find the minimal subset of conditional features that preserves the decision

Key words and phrases: Rough set, decision system, attribute reduction, reduct, metric.
2010 Mathematics Subject Classification: 68T20, 68U35
This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant 102.05-2013.37.

power of the original table. The feature subset is called a reduct. In the past two decades, feature reduction has attracted much attention from researchers of rough sets. However, most of the proposed algorithms can only be applied to static data sets. In the real world, decision tables are usually updated and changed with time. Modifications can vary from adding or deleting objects or features to updating existing objects. As a suboptimal solution for a changing table, we have to repeatedly run the existing algorithms to find reducts. In this case, the time spent for recomputation is quite large.

In the last few years, some researchers have developed incremental methods to find reducts on dynamic decision tables based on different measures. In [2, 3, 13], authors used positive region and discernibility matrix when adding new objects. In [7, 10, 11], authors constructed formulas for three entropies (Shannons entropy, Liang entropy and combination entropy) when adding or deleting objects. However, these formulas are quite complex. Moreover, the methods mentioned above have not completely dealt with dynamic decision tables.

In this paper we propose a distance measure between two attribute sets of a decision table. Using the measure, we give two algorithms for finding the reduct of a decision table after adding or deleting objects. Similar to other incremental algorithms, our ones save running time after addition or deletion by not recomputing the reducts on the entire object set but only update them. Even better, since our distance formula is less complicated than those of Shannons entropy, our algorithms run faster than those of [7, 10, 11].

This paper is organized as follows. In Section 2, we summarize some preliminary knowledge on rough set theory and related work on feature reduction in decision tables. Section 3 briefly presents the attribute reduction method based on distance measures. In Section 4, we construct two algorithms for finding reducts when one object is added or deleted. The conclusion of the paper and further research are presented in Section 5.

2. Basic concepts

In this section we summarize some basic concepts in rough set theory [9] and an overview of rough set based methods for attribute reduction in decision tables.

An information system is a couple $IS = (U, A)$ where U is a finite nonempty set of objects and A is a finite nonempty set of features. Each $a \in A$ determines a map $a : U \to V_a$ where V_a is the value set of a.

Given an information system $IS = (U, A)$, each $P \subseteq A$ determines an equivalence relation

$$IND(P) = \{(u, v) \in U \times U \,|\forall a \in P,\ a(u) = a(v)\}.$$

Let the partition of U by *IND(P)* be denoted by *U/P*, and the equivalence class containing u by $[u]_P$. Then let $[u]_P = \{v \in U \,|(u, v) \in IND(P)\}$.

Given an information system $IS = (U, A)$, $B \subseteq A$ and $X \subseteq U$, let $\underline{B}X = \{u \in U \,|[u]_B \subseteq X\}$ and $\overline{B}X = \{u \in U \,|[u]_B \cap X \neq \varnothing\}$, respectively, denote the lower and upper approximation of X with respect to B.

A decision table is a special form of an information system, where A includes two separate subsets, the condition attribute subset C and the decision attribute subset D. In other words, a decision table is $DS = (U, C \cup D)$ where $C \cap D = \varnothing$.

Let $DS = (U, C \cup D)$ be a decision table. Then $POS_C(D) = \bigcup_{D_i \in U/D} (\underline{C}D_i)$ is called the C-Positive region of D. One can easily obtain that $POS_C(D)$ is a set of objects belonging to U that can be partitioned by C into decision classes of D. A decision table DS is consistent if and only if $POS_C(D) = U$; otherwise, it is inconsistent.

Attribute reduction is the task to select the minimal subset of the condition attribute set that preserves the ability of the original decision table to partition the objects. In the past two decades, heuristic attribute reduction methods based on rough set theory have attracted attention of many researchers. These heuristic methods find the best reduct with respect to the classification quality of the features, also referred to as feature significance. In [5, 12], the authors summarize and categorize feature reduction methods in decision tables into three groups: (1) Positive region methods, including attribute reduction methods based on positive region; (2) Shannons entropy methods, including the method using Shannons entropy and method using relational algebra; (3) Liang entropy methods, including the method using Liang entropy, methods using information entropy and methods using discernibility matrix.

Using distance measures, the authors of [8] proposed a method for attribute reduction based on the Jaccard distance between two infinite sets and proved this method belongs to the group Shannons entropy methods. In the next section, we construct a new metric between two infinite sets with a corresponding method for attribute reduction.

3. Metric based attribute reduction

3.1. Metric between two knowledges and properties

A metric on the set U is a map $d : U \times U \to [0, \infty)$ that satisfies the following conditions for any $x, y, z \in U$ [1]

P(1) $d(x,y) \geq 0$, $d(x,y) = 0$ if and only if $x = y$,

P(2) $d(x,y) = d(y,x)$,

P(3) $d(x,y) + d(y,z) \geq d(x,z)$.

Theorem 3.1. [4] *Given an infinite set of objects U and the family subsets* $\mathcal{P}(U)$ *of U, for any* $X, Y \in \mathcal{P}(U)$, $d(X,Y) = |X \cup Y| - |X \cap Y|$ *is a metric between X and Y.*

From the metric between two infinite sets as Theorem 3.1, we construct the metric between two knowledges as defined next, generated by two attribute sets on a decision table.

Given a decision table $DS = (U, C \cup D)$, for each $P \subseteq C$, $K(P) = \{[u_i]_P \,|u_i \in U\}$ is called a knowledge of P on U [9]. *K(P)* includes $|U|$ elements where each one is a partition in U/P , also referred as a knowledge granule. Let the family of all knowledges on U be denoted by $\mathcal{K}(U)$.

Theorem 3.2. *The map* $d : \mathcal{K}(U) \times \mathcal{K}(U) \to [0, \infty)$ *defined by*

$$d(K(P), K(Q)) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left(\left| [u_i]_P \cup [u_i]_Q \right| - \left| [u_i]_P \cap [u_i]_Q \right| \right)$$

is a metric between K(P) and K(Q).

Proof. *P(1)* Applying Theorem 3.1 for two sets $[u_i]_P$ and $[u_i]_Q$ with $u_i \in U$, one can obtain $\left| [u_i]_P \cup [u_i]_Q \right| - \left| [u_i]_P \cap [u_i]_Q \right| \geq 0$, as a result $d(K(P), K(Q)) \geq 0$; $d(K(P), K(Q)) = 0$ if and only if $\left| [u_i]_P \cap [u_i]_Q \right| = \left| [u_i]_P \cup [u_i]_Q \right| \Leftrightarrow [u_i]_P \cap [u_i]_Q = [u_i]_P \cup [u_i]_Q \Leftrightarrow [u_i]_P = [u_i]_Q$ for any $u_i \in U$, i.e. *K(P)=K(Q).*

P(2) According to the definition, $d(K(P), K(Q)) = d(K(Q), K(P))$ for any $K(P), K(Q) \in \mathcal{K}(U)$.

P(3) According to the definition, one can obtain

$d(K(P), K(Q)) + d(K(Q), K(R)) =$

$$= \tfrac{1}{|U|^2} \sum_{i=1}^{|U|} \left(\left| [u_i]_P \cup [u_i]_Q \right| - \left| [u_i]_P \cap [u_i]_Q \right| \right) +$$

$$+\frac{1}{|U|^2}\sum_{i=1}^{|U|}\left(\left|[u_i]_Q \cup [u_i]_R\right| - \left|[u_i]_Q \cap [u_i]_R\right|\right)$$
$$= \frac{1}{|U|^2}\sum_{i=1}^{|U|} d\left([u_i]_P, [u_i]_Q\right) + \frac{1}{|U|^2}\sum_{i=1}^{|U|} d\left([u_i]_Q, [u_i]_R\right)$$
$$= \frac{1}{|U|^2}\sum_{i=1}^{|U|} d\left([u_i]_P, [u_i]_Q\right) + d\left([u_i]_Q, [u_i]_R\right)$$
$$\geq \frac{1}{|U|^2}\sum_{i=1}^{|U|} d\left([u_i]_P, [u_i]_R\right) = d\left(K(P), K(R)\right).$$

From *(P1), (P2), (P3)* we get that $d(K(P), K(Q))$ is a metric on $\mathcal{K}(U)$. ■

Proposition 3.1. *Given a decision table* $DS = (U, C \cup D)$ *and* $P, Q \subseteq C$, *we have*

(i) $d(K(P), K(Q))$ *reaches the minimum value 0 if and only if* $K(P) = K(Q)$,

(ii) $d(K(P), K(Q))$ *reaches the maximum value* $1 - \frac{1}{|U|}$ *if and only if* $K(P) = \{[u_i]_P = U \,|u_i \in U\}$, $K(Q) = \left\{[u_i]_Q = \{u_i\} \,|u_i \in U\right\}$ *or*

$K(P) = \{[u_i]_P = \{u_i\} \,|u_i \in U\}$, $K(Q) = \left\{[u_i]_Q = U \,|u_i \in U\right\}$.

Proof. From Theorem 3.2 we have $d(K(P), K(Q))$ reaches the minimum value 0 if and only if $K(P) = K(Q)$. $d(K(P), K(Q))$ reaches the maximum value when $\left|[u_i]_P \cup [u_i]_Q\right|$ reaches the maximum value $|U|$ and $\left|[u_i]_P \cap [u_i]_Q\right|$ reaches the minimum value 1, i.e. $[u_i]_P = U$, $[u_i]_Q = \{u_i\}$ or $[u_i]_P = \{u_i\}$, $[u_i]_Q = U$. The maximum value is $\frac{1}{|U|^2}\sum_{i=1}^{|U|}(|U| - 1) = 1 - \frac{1}{|U|}$. ■

Proposition 3.2. *Given a decision table* $DS = (U, C \cup D)$ *and two partitions* $U/C = \{C_1, C_2, ..., C_m\}$, $U/D = \{D_1, D_2, ..., D_n\}$, *we have*

$$d(K(C), K(C \cup D)) = \frac{1}{|U|^2}\sum_{i=1}^{n}\sum_{j=1}^{m}|D_i \cap C_j|\,|C_j - D_i|.$$

Proof. Let $D_i \cap C_j = \left\{u_{i1}, u_{i2}, ..., u_{is_j}\right\}$ for $|D_i \cap C_j| = s_j$ and $|D_i| = t_i$. Then $\sum_{j=1}^{m} s_j = t_i$ and $\sum_{i=1}^{n} t_i = |U|$. We have

$D_i \cap C_j = [u_{i1}]_D \cap [u_{i1}]_C = [u_{i2}]_D \cap [u_{i2}]_C = ... = \left[u_{is_j}\right]_D \cap \left[u_{is_j}\right]_C$,

$|D_i \cap C_j| = |[u_{i1}]_D \cap [u_{i1}]_C| = |[u_{i2}]_D \cap [u_{i2}]_C| = ... = \left|\left[u_{is_j}\right]_D \cap \left[u_{is_j}\right]_C\right| = s_j$,

$|D_i \cap C_j|\,|C_j - D_i| = |D_i \cap C_j|\,|C_j - (D_i \cap C_j)| =$

$= |[u_{i1}]_C - ([u_{i1}]_D \cap [u_{i1}]_C)| + ... + \left|\left[u_{is_j}\right]_C - \left(\left[u_{is_j}\right]_D \cap \left[u_{is_j}\right]_C\right)\right|$

$= \sum_{k=1}^{s_i} |[u_{ik}]_C - ([u_{ik}]_D \cap [u_{ik}]_C)|,$

$\sum_{j=1}^{m} |D_i \cap C_j|\,|C_j - D_i| = \sum_{j=1}^{m}\sum_{k=1}^{s_j} |[u_{ik}]_C - ([u_{ik}]_D \cap [u_{ik}]_C)|$

$= \sum_{k=1}^{t_i} |[u_{ik}]_C - ([u_{ik}]_D \cap [u_{ik}]_C)|.$

So that,

$\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\,|C_j - D_i| = \sum_{i=1}^{n}\sum_{k=1}^{t_i} |[u_{ik}]_C - ([u_{ik}]_D \cap [u_{ik}]_C)|$

$= \sum_{i=1}^{|U|} |[u_i]_C - ([u_i]_D \cap [u_i]_C)|,$

$\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\,|C_j - D_i| = \sum_{i=1}^{|U|} |[u_i]_C - [u_i]_{C\cup D}| = \sum_{i=1}^{|U|} (|[u_i]_C| - |[u_i]_{C\cup D}|)$

$= \sum_{i=1}^{|U|} (|[u_i]_C \cup [u_i]_{C\cup D}| - |[u_i]_C \cap [u_i]_{C\cup D}|).$

Consequently, $d(K(C), K(C\cup D)) = \frac{1}{|U|^2}\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\,|C_j - D_i|$. ■

Proposition 3.3. *Given a decision table $DS = (U, C\cup D)$, one can obtain $d(K(C), K(C\cup D)) = E(D|C)$, where $E(D|C)$ is the Liang conditional entropy defined in [6].*

Proof. Let $U/C = \{C_1, C_2, ..., C_m\}$ and $U/D = \{D_1, D_2, ..., D_n\}$. According to the definition of the Liang entropy in [5], we have

$E(D|C) = \frac{1}{|U|^2}\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\left|D_i^c - C_j^c\right| = \frac{1}{|U|^2}\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\,|D_i^c - C_j|$

$= \frac{1}{|U|^2}\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\,|C_j - (D_i \cap C_j)| = \frac{1}{|U|^2}\sum_{i=1}^{n}\sum_{j=1}^{m} |D_i \cap C_j|\,|C_j - D_i|$

$= d(K(C), K(C\cup D))$. ■

Definition 3.1. *Given a decision table $DS = (U, C\cup D)$, $c \in C$ is dispensable in DS if $d(K(C-\{c\}), K(C-\{c\}\cup D)) = d(K(C), K(C\cup D))$; otherwise, c is indispensable. The set of all indispensable attributes in DS is called the core and denoted by CORE(C).*

Definition 3.2. *Given a decision table $DS = (U, C\cup D)$, $R \subseteq C$. If*

(i) $d(K(R), K(R\cup D)) = d(K(C), K(C\cup D))$ *and,*

(ii) $\forall r \in R, \quad d(K(R-\{r\}), K(R-\{r\} \cup D)) \neq d(K(C), K(C \cup D))$

then R is a reduct of C based on metric.

From Proposition 3.3, one can see that the reduct based on the metric is equivalent to the reduct based on the Liang entropy. So, metric based attribute reduction belongs to the group of Liang entropy based methods.

Definition 3.3. *Given a decision table $DS = (U, C \cup D)$, $B \subset C$ and $b \in C - B$, the significance of is defined by*

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

where $U/\{\varnothing\} = U$.

According to [6], $E(D|B \cup \{b\}) \leq E(D|B)$. So

$$d(K(B \cup \{b\}),\ K(B \cup \{b\} \cup D)) \leq d(K(B),\ K(B \cup D))$$

and $SIG_B(b) \geq 0$. Hence, $SIG_B(b)$ is caculated by the amount of change in the distance between B and $B \cup D$ when adding b to B. The greater is $SIG_B(b)$, the greater is the amount of change, or the more significant b is and vice versa. This significance is the attribute selection criteria of the heuristic algorithm for finding reducts of decision tables.

Algorithm 3.1. Heuristic algorithm for finding the best reduct based on the metric.

Input: Decision table $DS = (U, C \cup D)$.

Output: The best reduct R.

//Finding the core set CORE(C);

1. $CORE(C) = \varnothing$;
2. For $c \in C$
3. If $d(K(C-\{c\}), K(C-\{c\} \cup D)) \neq d(K(C), K(C \cup D))$ then
$CORE(C) := CORE(C) \cup \{c\}$;

//Finding the reduct based on metric

4. $R = CORE(C)$
5. While $d(K(R), K(R \cup D)) \neq d(K(C), K(C \cup D))$ do
6. Begin
7. For $a \in C - R$ calculate $SIG_R(a)$;
8. Select $a_m \in C - R$ such that $SIG_R(a_m) = \underset{a \in C-R}{Max} \{SIG_R(a)\}$;
9. $R = R \cup \{a_m\}$;

10. End;
11. Return R;

Given a decision table $DS = (U, C \cup \{d\})$ by supposing that decision set D includes only one element $D = \{d\}$, according to [14], the time complexity (hereinafter referred to as complexity) for getting the conditional partition U/C is $O(|U||C|)$, hence the complexity for computing the metric

$$d(K(C),\ K(C \cup \{d\}))$$

is

$$O\left(|U||C| + |U| + \sum_{i=1}^{n} D_i \sum_{j=1}^{m} C_j\right) = O\left(|U||C| + |U|^2\right),$$

the complexity for computing the core set *CORE(C)* from steps 1 to 3 is $O\left(|C|\left(|U||C| + |U|^2\right)\right) = O\left(|C|^2|U| + |C||U|^2\right)$, and the complexity for computing the reduct from steps 4 to 9 is $O\left(|C|^2|U| + |C||U|^2\right)$. Hence, the complexity of algorithm 3.1 is $O\left(|C|^2|U| + |C||U|^2\right)$.

4. Algorithms for finding the reduct based on metric when adding or deleting one object

4.1. Formula for calculating the metric when adding one object

Given a decision table $DS = (U, C \cup D)$ and $B \subseteq C$, let

$$U/B = \{X_1, X_2, ...X_m\} \quad \text{and} \quad U/D = \{Y_1, Y_2, ..., Y_n\}.$$

The metric between two knowledges $K(B)$ and $K(B \cup D)$ on U is

$$d_U(K(B), K(B \cup D)).$$

Proposition 4.1. *Suppose that object x is added to U, then one can obtain:*
1) If $x \notin X_j$ *for any* $j = 1..m$ *and* $x \notin Y_i$ *for any* $i = 1..n$ *, then*
$d_{U\cup\{x\}}(K(B), K(B \cup D)) = \frac{|U|^2}{|U+1|^2} d_U(K(B), K(B \cup D))$.
2) If $x \notin X_j$ *for any* $j = 1..m$ *and* $x \in Y_q$ *for* $q \leq n$, *then*
$d_{U\cup\{x\}}(K(B), K(B \cup D)) = \frac{|U|^2}{|U+1|^2} d_U(K(B), K(B \cup D))$.

3) If $x \in X_p$ *for* $p \leq m$ *and* $x \notin Y_i$ *for any* $i = 1..n$, *then*

$d_{U\cup\{x\}}(K(B), K(B\cup D)) = \frac{1}{|U+1|^2}\left(|U|^2 d_U(K(B), K(B\cup D)) + 2|X_p|\right)$.

4) If $x \in X_p$ *for* $p \leq m$ *and* $x \in Y_q$ *for* $q \leq n$, *then*

$d_{U\cup\{x\}}(K(B), K(B\cup D)) =$

$= \frac{1}{|U+1|^2}\left(|U|^2 d_U(K(B), K(B\cup D)) + 2|X_p - Y_q|\right)$.

Proof. 1) Suppose that $X_{m+1} = \{x\}$ and $Y_{n+1} = \{x\}$. We have

$$d_{U\cup\{x\}}(K(B), K(B\cup D)) = \frac{1}{|U+1|^2}\sum_{i=1}^{n+1}\sum_{j=1}^{m+1}|Y_i\cap X_j||X_j - Y_i|$$

$$= \frac{1}{|U+1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j||X_j - Y_i| + \sum_{j=1}^{m+1}|Y_{n+1}\cap X_j||X_j - Y_{n+1}|\right.$$

$$\left. + \sum_{i=1}^{n}|Y_i\cap X_{m+1}||X_{m+1} - Y_i|\right) = \frac{|U|^2}{|U+1|^2}d_U(K(B), K(B\cup D)).$$

2) Suppose that $X_{m+1} = \{x\}$ and $x \in Y_q$ for $q \leq n$. We have:

$$d_{U\cup\{x\}}(K(B), K(B\cup D)) = \frac{1}{|U+1|^2}\left(\sum_{i=1, i\neq q}^{n}\sum_{j=1}^{m+1}|Y_i\cap X_j||X_j - Y_i| + \right.$$

$$\left. + \sum_{j=1}^{m+1}|(Y_q\cup\{x\})\cap X_j||X_j - (Y_q\cup\{x\})|\right)$$

$$= \frac{1}{|U+1|^2}\left(\sum_{i=1, i\neq q}^{n}\sum_{j=1}^{m}|Y_i\cap X_j||X_j - Y_i| + \sum_{j=1}^{m}|Y_q\cap X_j||X_j - Y_q|\right)$$

$$= \frac{1}{|U+1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j||X_j - Y_i|\right) = \frac{|U|^2}{|U+1|^2}d_U(K(B), K(B\cup D)).$$

3) Suppose that $x \in X_p$ for $p \leq m$ and $Y_{n+1} = \{x\}$. We have:

$$d_{U\cup\{x\}}(K(B), K(B\cup D)) = \frac{1}{|U+1|^2}\left(\sum_{i=1}^{n+1}\sum_{j=1, j\neq p}^{m}|Y_i\cap X_j||X_j - Y_i| + \right.$$

$$\left.\sum_{i=1}^{n+1}|Y_i\cap(X_p\cup\{x\})||(X_p\cup\{x\}) - Y_i|\right)$$

$$= \frac{1}{|U+1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j||X_j - Y_i| + \sum_{i=1}^{n}|Y_i\cap X_p||\{x\}| + |X_p||\{x\}|\right)$$

$$= \frac{1}{|U+1|^2}\left(|U|^2 d_U(K(B), K(B\cup D)) + 2|X_p|\right).$$

4) Suppose that $x \in X_p$ for $p \leq m$ and $x \in Y_q$ for $q \leq n$. We have:

$$d_{U\cup\{x\}}(K(B), K(B\cup D)) = \frac{1}{|U+1|^2}\sum_{i=1, i\neq q}^{n}\sum_{j=1, j\neq p}^{m}|Y_i\cap X_j||X_j - Y_i| +$$

$$+\frac{1}{|U+1|^2}\sum_{j=1,j\neq p}^{m}|(Y_q\cup\{x\})\cap X_j|\,|X_j-(Y_q\cup\{x\})|+$$
$$+\frac{1}{|U+1|^2}\,|(Y_q\cup\{x\})\cap(X_p\cup\{x\})|\,|(X_p\cup\{x\})-(Y_q\cup\{x\})|+$$
$$+\frac{1}{|U+1|^2}\sum_{i=1,i\neq q}^{n}|Y_i\cap(X_p\cup\{x\})|\,|(X_p\cup\{x\})-Y_i|$$
$$=\frac{1}{|U+1|^2}\left(|U|^2 de_U\left(K\left(B\right),K\left(B\cup D\right)\right)+|X_p-Y_q|+|(U-Y_q)\cap X_p|\right)$$
$$=\frac{1}{|U+1|^2}\left(|U|^2 d_U\left(K\left(B\right),K\left(B\cup D\right)\right)+2\,|X_p-Y_q|\right).$$ ■

4.2. Formula for calculating the metric when deleting one object

Proposition 4.2. *Let $x\in U$ be the element to be deleted from U. Then*

1) If $\{x\}=X_p$ for $p\leq m$ and $\{x\}=Y_q$ for $q\leq n$, then

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=\frac{|U|^2}{|U-1|^2}d_U\left(K\left(B\right),K\left(B\cup D\right)\right).$$

2) If $\{x\}=X_p$ for $p\leq m$ and $x\in Y_q$ for $q\leq n$, then

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=\frac{|U|^2}{|U-1|^2}d_U\left(K\left(B\right),K\left(B\cup D\right)\right).$$

3) If $x\in X_p$ for $p\leq m$ and $\{x\}=Y_q$ for $q\leq n$, then

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=$$
$$=\frac{1}{|U-1|^2}\left(|U|^2 d_U\left(K\left(B\right),K\left(B\cup D\right)\right)-2\,|X_p|+2\right).$$

4) If $x\in X_p$ for $p\leq m$ and $x\in Y_q$ for $q\leq n$, then

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=$$
$$=\frac{1}{|U-1|^2}\left(|U|^2 d_U\left(K\left(B\right),K\left(B\cup D\right)\right)+|X_p\cap Y_q|-|X_p-Y_q|-|X_p|\right).$$

Proof. 1) Suppose that $X_m=\{x\}$ and $Y_n=\{x\}$. We have:

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=\frac{1}{|U-1|^2}\sum_{i=1}^{n-1}\sum_{j=1}^{m-1}|Y_i\cap X_j|\,|X_j-Y_i|$$
$$=\frac{1}{|U-1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j|\,|X_j-Y_i|-\sum_{j=1}^{m-1}|Y_n\cap X_j|\,|X_j-Y_n|-\right.$$
$$\left.-\sum_{i=1}^{n}|Y_i\cap X_m|\,|X_m-Y_i|\right)=\frac{|U|^2}{|U-1|^2}d_U\left(K\left(B\right),K\left(B\cup D\right)\right).$$

2) Suppose that $X_m=\{x\}$ and $x\in Y_q$ for $q\leq n$. We have:

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=\frac{1}{|U-1|^2}\left(\sum_{i=1,i\neq q}^{n}\sum_{j=1}^{m-1}|Y_i\cap X_j|\,|X_j-Y_i|+\right.$$

$$+\sum_{j=1}^{m-1}|(Y_q-\{x\})\cap X_j|\,|X_j-(Y_q-\{x\})|\Bigg)$$

$$=\tfrac{1}{|U-1|^2}\left(\sum_{i=1,i\neq q}^{n}\sum_{j=1}^{m}|Y_i\cap X_j|\,|X_j-Y_i|+\sum_{j=1}^{m}|Y_q\cap X_j|\,|X_j-Y_q|\right)$$

$$=\tfrac{1}{|U-1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j|\,|X_j-Y_i|\right)=\tfrac{|U|^2}{|U-1|^2}d_U\left(K\left(B\right),K\left(B\cup D\right)\right).$$

3) Suppose that $x\in X_p$ for $p\leq m$ and $Y_n=\{x\}$. We have:

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=\tfrac{1}{|U-1|^2}\left(\sum_{i=1}^{n-1}\sum_{j=1,j\neq p}^{m}|Y_i\cap X_j|\,|X_j-Y_i|+\right.$$

$$\left.+\sum_{i=1}^{n-1}|Y_i\cap(X_p-\{x\})|\,|(X_p-\{x\})-Y_i|\right)$$

$$=\tfrac{1}{|U-1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j|\,|X_j-Y_i|-\sum_{i=1}^{n-1}|Y_i\cap X_p|\,|X_p-Y_i|-\right.$$

$$\left.-|Y_n\cap X_p|\,|X_p-Y_n|+\sum_{i=1}^{n-1}|Y_i\cap X_p|\,(|X_p-Y_i|-1)\right)$$

$$=\tfrac{1}{|U-1|^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_i\cap X_j|\,|X_j-Y_i|-2\,|X_p|+2\right)$$

$$=\tfrac{1}{|U-1|^2}\Big(|U|^2d_U\left(K\left(B\right),K\left(B\cup D\right)\right)-2\,|X_p|+2\Big).$$

4) Suppose that $x\in X_p$ for $p\leq m$ and $x\in Y_q$ for $q\leq n$. We have:

$$d_{U-\{x\}}\left(K\left(B\right),K\left(B\cup D\right)\right)=\tfrac{1}{|U-1|^2}\sum_{i=1,i\neq q}^{n}\sum_{j=1,j\neq p}^{m}|Y_i\cap X_j|\,|X_j-Y_i|+$$

$$+\tfrac{1}{|U-1|^2}\sum_{j=1,j\neq p}^{m}|(Y_q-\{x\})\cap X_j|\,|X_j-(Y_q-\{x\})|+$$

$$+\tfrac{1}{|U-1|^2}\,|(Y_q-\{x\})\cap(X_p-\{x\})|\,|(X_p-\{x\})-(Y_q-\{x\})|+$$

$$+\tfrac{1}{|U-1|^2}\sum_{i=1,i\neq q}^{n}|Y_i\cap(X_p-\{x\})|\,|(X_p-\{x\})-Y_i|$$

$$=\tfrac{1}{|U+1|^2}\sum_{i=1,i\neq q}^{n}\sum_{j=1,j\neq p}^{m}|Y_i\cap X_j|\,|X_j-Y_i|+\tfrac{1}{|U+1|^2}\sum_{j=1,j\neq p}^{m}|Y_q\cap X_j|\,|X_j-Y_q|+$$

$$+\tfrac{1}{|U+1|^2}\,(|Y_q\cap X_p|-1)\,|X_p-Y_q|+\tfrac{1}{|U+1|^2}\sum_{i=1,i\neq q}^{n}|Y_i\cap X_p|\,(|X_p-Y_i|-1)$$

$$=\tfrac{1}{|U-1|^2}\Big(|U|^2d_U\left(K\left(B\right),K\left(B\cup D\right)\right)+|X_p\cap Y_q|-|X_p-Y_q|-|X_p|\Big).\quad\blacksquare$$

4.3. Algorithms for finding the reduct when adding or deleting one object

At first, we construct an incremental algorithm for finding the reduct after adding a new object. Here Proposition 4.3 is used for constructing the algorithm.

Proposition 4.3. *Given a decision table* $DS = (U, C \cup D)$, $B \subseteq C$ *is a reduct of DS based on the metric and* x *is the new element added to* U. *Let* $U/B = \{X_1, X_2, ...X_m\}$, $U/C = \{Y_1, Y_2, ..., Y_n\}$. *Then one can obtain:*

1) If $x \notin X_j$ *for any* $j = 1..m$, *then*

$d_{U\cup\{x\}}(K(B), K(B \cup D)) = d_{U\cup\{x\}}(K(C), K(C \cup D))$.

2) If $x \in X_p$ *for* $p \leq m$ *and* $x \notin Y_i$ *for any* $i = 1..n$, *then*

$d_{U\cup\{x\}}(K(B), K(B \cup D)) \neq d_{U\cup\{x\}}(K(C), K(C \cup D))$.

3) If $x \in X_p$ *for* $p \leq m$ *and* $x \in Y_q$ *for* $q \leq n$, *then*

$d_{U\cup\{x\}}(K(B), K(B \cup D)) = d_{U\cup\{x\}}(K(C), K(C \cup D))$.

Proof. 1) and 2) can be directly drawn from Proposition 4.1 and Definition 3.2 of a reduct based on the metric. We will prove 3). According to Proposition 4.1, one can obtain

$$d_{U\cup\{x\}}(K(B), K(B \cup D)) =$$

$$= \frac{1}{|U+1|^2}\left(|U|^2 d_U(K(B), K(B \cup D)) + 2|X_p - D_r|\right)$$

for $D_r \in U/D$ and $x \in X_p, x \in D_r$.

$$d_{U\cup\{x\}}(K(C), K(C \cup D)) =$$

$$= \frac{1}{|U+1|^2}\left(|U|^2 d_U(K(C), K(C \cup D)) + 2|Y_q - D_r|\right)$$

for $x \in Y_q, x \in D_r$.

Since B is the reduct of the DS, then

$d_U(K(B), K(B \cup D)) = d_U(K(C), K(C \cup D))$.

According to Proposition 3.3 we have $E(D|B) = E(D|C)$. Since $B \subseteq C$, then $Y_q \subseteq X_p$. If $Y_q = X_p$, obviously we gain the proof. If $Y_q \subset X_p$, we may assume that $Y_q = X_p \cup X_k$. From $E(D|B) = E(D|C)$ and [6] we have $X_p \subseteq D_r, X_k \subseteq D_r$ or $X_p \subseteq D_r, Y_q \subseteq D_r$, so $X_p - D_r = \varnothing$ and $Y_q - D_r = \varnothing$. As a result, $d_{U\cup\{x\}}(K(B), K(B \cup D)) = d_{U\cup\{x\}}(K(C), K(C \cup D))$. ■

Proposition 4.3 shows that if x does not belong to any equivalence class in U/B and U/C, or x simultaneously belongs to one equivalence class in U/B and U/C, then the metric $d_U(K(B), K(B \cup D))$, $d_U(K(C), K(C \cup D))$ is

preserved, i.e the reduct is unchanged. From that, we construct the incremental algorithm for finding reducts as follow:

Algorithm 4.1. The incremental algorithm for finding reducts based on metric when adding a new object.

Input: Decision table $DS = (U, C \cup D)$, reduct R_U on U and new object x.

Output: Reduct $R_{U\cup\{x\}}$ on $U \cup \{x\}$.

1. Assign $R = R_U$, calculate $U/R = \{X_1, X_2, ...X_m\}$;
2. If $x \in X_p,\ X_p \in U/R$ then
3. Begin
4. Calculate $U/C = \{Y_1, Y_2, ..., Y_n\}$;
5. If $x \notin Y_q$, $\forall q = 1..n$ then
6. Begin
7. While $d_{U\cup\{x\}}(K(R), K(R \cup D)) \neq d_{U\cup\{x\}}(K(C), K(C \cup D))$ do
8. Begin
9. For $a \in C - R$ calculate $SIG_R(a)$;
10. Select $a_m \in C - R$ such that $SIG_R(a_m) = \underset{a\in C-R}{Max}\{SIG_R(a)\}$
11. $R = R \cup \{a_m\}$;
12. End;
13. End;
14. End;
15. For $a \in R$ do
16. Begin
17. Calculate $d_{U\cup\{x\}}(K(R - \{a\}), K((R - \{a\}) \cup D))$;
18. If $d_{U\cup\{x\}}(K(R - \{a\}), K((R - \{a\}) \cup D)) =$
$= d_{U\cup\{x\}}(K(R),\ K(R \cup D))$ then $R = R - \{a\}$;
19. End;
20. Return R.

According to [14], the complexity for calculating partition U/C is $O(|C||U|)$, hence the complexity of the incremental formula for calculating metric in Proposition 4.1 is $O(|C||U| + m|C| + |U| + |X_p||Y_q|) = O(|C||U| + |X_p||Y_q|)$. The complexity of the while loop between lines 7 to 12 is $O(|C|(|C||U| + |X_p||Y_q|))$. The complexity of the for loop between lines 15 to 19 is $O(|C|(|C||U| + |X_p||Y_q|))$. Hence the complexity of Algorithm 4.1 is $O\left(|C|^2|U| + |C||X_p||Y_q|\right)$. Obviously, $|X_p||Y_q|$ is much less than $|U|^2$, so we can say the complexity of Algorithm 4.1 is much less than that of the original Algorithm 3.1.

Similar to the case of adding one new object, the algorithm for finding reducts when deleting one object is based on Proposition 4.4.

Proposition 4.4. *Given a decision table* $DS = (U, C \cup D)$*, a reduct* $B \subseteq C$ *of DS based on the metric and* $x \in U$*. Let* $U/B = \{X_1, X_2, ...X_m\}$*,* $U/C = \{Y_1, Y_2, ..., Y_n\}$*. Then one can obtain:*

1) If $x \notin X_j$ *for any* $j = 1..m$*, then*

$d_{U-\{x\}}(K(B), K(B \cup D)) = d_{U-\{x\}}(K(C), K(C \cup D))$.

2) If $x \in X_p$ *for* $p \leq m$ *and* $x \notin Y_i$ *for any* $i = 1..n$*, then*

$d_{U-\{x\}}(K(B), K(B \cup D)) \neq d_{U-\{x\}}(K(C), K(C \cup D))$.

3) If $x \in X_p$ *for* $p \leq m$ *and* $x \in Y_q$ *for* $q \leq n$*, then*

$d_{U-\{x\}}(K(B), K(B \cup D)) = d_{U-\{x\}}(K(C), K(C \cup D))$.

Applying the formula for calculating the metric when deleting one object in Proposition 4.2, Proposition 4.4 is similarly proved as the proof of Proposition 4.3. The algorithm for finding reducts in this case is worked out in the same way as in Algorithm 4.1.

5. Conclusions

We proposed effective methods to optimize the running time for finding reducts in databases that gradually get increased, changed and updated. Based on an incremetal calculation, in this paper we use a distance measure to construct two algorithms for finding reducts in the cases of adding or deleting one object. Our algorithms for finding reducts can easily be extended to the case when adding or deleting more than one objects. Also in this paper, we prove that the time complexity of our algorithms is less than that one of original algorithms. As further research, we could use the metric to constructing algorithms for finding reducts in case of updating objects.

References

[1] **Deza, M.M. and E. Deza,** *Encyclopedia of Distances*, Springer, 2009.

[2] **Demetrovics, J., V.D. Thi and N.L. Giang,** An effective algorithm for determining the set of all reductive attributes in incomplete decision

tables, *Cybernetics and Information Technologies CIT*, Sofia, Bulgarian Academy of Sciences, **13** (4) (2013), 118–126.

[3] **Guan, L. H.,** An incremental updating algorithm of attribute reduction set in decision tables, *FSKD'09 Proc. 6th Int. Conf. on Fuzzy Systems and Knowledge Discovery*, **2** (2009), 421–425.

[4] **Halmos, P.R.,** *Naive set theory*, The University Series in Undergraduate Mathematics, van Nostrand Company, 1960.

[5] **Hu, F., G.Y. Wang, H. Huang H. and Y. Wu,** Incremental attribute reduction based on elementary sets, *Proc. 10th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Regina, Canada, 2005,* 185–193.

[6] **Liang, J.Y, K.S. Chin, C.Y. Dang and C.M. R.C.M. Yam,** New method for measuring uncertainty and fuzziness in rough set theory, *International Journal of General Systems*, **31** (4) (2002), 331–342.

[7] **Liang, J.Y, F. Wang, C.Y. Dang and Y.H. Qian,** A group incremental approach to feature selection applying rough set technique, *IEEE Transactions on Knowledge and Data Engineering*, **26** (2) (2014), 294–308.

[8] **Long Giang Nguyen,** Metric based attribute reduction in decision tables, *The 2012 Int. Workshop on Rough Sets Applications (RSA2012), FedCSIS Proceedings*, IEEE, 2012, 333–338.

[9] **Pawlak, Z.,** *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, 1991.

[10] **Wang, F., J.Y. Liang and Y.H. Qian,** Attribute reduction: A dimension incremental strategy, *Knowledge-Based Systems*, **39** (2013), 95–108.

[11] **Feng Wang, Jiye Liang and Chuangyin Dang,** Attribute reduction for dynamic data sets, *Applied Soft Computing*, **13** (1) (2013), 676–689.

[12] **Wei, W., J.Y. Liang, Y.H. Qian, F. Wang and C.Y. Dang,** Comparative study of decision performance of decision tables induced by attribute reductions, *International Journal of General Systems*, **Vol. 39** (8) (2010), 813–838.

[13] **Zhang, C.S, J. Jing Ruan and Y.H. Tan,** An improved incremental updating algorithm for core based on positive region, *Journal of Computational Information Systems*, **7** (9) (2011), 3127–3133.

[14] **Xu, Z.Y., Z.P. Liu, B.R. Yang and W. Song,** A quick attribute reduction algorithm with complexity of max $\left\{O\left(|C| * |U|\right), O\left(|C|^2 * |U/C|\right)\right\}$, *Journal of Computer*, **29** (3) (2006), 391–398.

János Demetrovics
Institute for Computer Science and Control (MTA SZTAKI)
Hungarian Academy of Sciences
Budapest, Hungary
demetrovics@sztaki.mta.hu

Vu Duc Thi
Information Technology Institute
Vietnam National University (VNU)
Ha Noi, Viet Nam
vdthi@vnu.edu.vn

Nguyen Long Giang
Institute of Information Technology
Vietnam Academy of Science and Technology (VAST)
Ha Noi, Viet Nam
nlgiang@ioit.ac.vn