International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

About a Distance Measure and Application for Finding Reduct in Incomplete Decision Tables

Nguyen Anh Tuan, Nguyen Long Giang

Abstract*: Tolerance rough set model is an effective tool to reduce attributes in incomplete decision tables. Over 40 years, several attribute reduction methods have been proposed to improve the efficiency of execution time and the number of attributes of the reduct. However, they are classical filter algorithms, in which the classification accuracy of decision tables is computed after obtaining the reducts. Therefore, the obtained reducts of these algorithms are not optimal in terms of reduct cardinality and classification accuracy. In this paper, we propose a filter-wrapper algorithm to find a reduct in incomplete decision tables. We then use this measure to determine the importance of the property and select the attribute based on the calculated importance (filter phase). In the next step, we find the reduct with the highest classification accuracy by iterating over elements of the set containing the sequence of attributes selected in the first step (wrapper phase). To verify the effectiveness of the method, we conduct experiments on 6 famous UCI data sets. Experimental results show that the proposed method increase classification accuracy as well as reduce the cardinality of reduct compared to Algorithm 1 [12].

Keywords*: Attribute reduction, Distance, Incomplete decision table, Reduct, Tolerance rough set.

I. INTRODUCTION

Rough set theory has long been conceived as a tool to conceptualize, organize, and analyze data types to deal with inaccurate, uncertain or ambiguous data in applications related to Machine Learning. In this paper, we present an approach to tolerance rough set for incomplete information systems, i.e. for systems in which the attribute values for objects may not be defined (missing, null). Our main concern is to find rules from such systems, which are used to calculate appropriate measurements, build efficient algorithms and extract important attributes without diminishing information. The tolerance rough set model we mentioned is inspired by from the tolerance relation extended by Kryszkiewicz[3] from the equivalence relation in the traditional rough set theory. Based on this model, many new methods have been published [1,4,8,11]. In this paper, we propose IDS_FW_DAR algorithm combining the two traditional methods, filter and wrapper. We call this the gap in the filter-wrapper hybrid approach. In the IDS_FW_DAR algorithm, the filter phase finds the most important attribute among candidates when adding to a reduct, while the wrapper phase finds the reduct with the highest classification accuracy. Test results on 6 sample data sets show that the proposed algorithm is better than some other algorithms in terms of classification accuracy and time for calculation. Moreover, the number of reduct attributes is less than some other filter algorithms.

This paper is organized as follows: Section 1 is an introduction. Section 2 introduces the basic concepts in tolerance rough set model. Section 3 presents the construction of distance measure in the incomplete decision table and presents the heuristic algorithm for finding a reduct using the proposed distance. Section 4 presents the experimental results for our algorithm. Section 5 presents our conclusions and suggests further research directions.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Nguyen Anh Tuan*, Vinh Phuc College; Phuc Yen City, Vinh Phuc Province, Vietnam. Email: tuanna573@gmail.com
Nguyen Long Giang, Institute of Information Technology, Vietnam Academy of Science and Technology; Hanoi city, Vietnam. Email: nlgiang@ioit.ac.vn

II. PRELIMINARIES

A. Incomplete Decision Table

The decision table is a pair $DS=(U,C\cup\{d\})$ where, U is an object set and $U\neq\varnothing$, C is a set of condition attributes and $C\neq\varnothing$, D is decision attribute set, and $C\cap D=\varnothing$. Each attribute $c\in C$ defines a mapping $c:U\to V_c$,where V_c is the value set of the attribute C.

If V_c contains missing values, the DS is called the incomplete decision table, otherwise it is complete decision table, the missing value is represented as '* '.

Then, incomplete decision tables is represented by $IDS=(U,C\cup\{d\})$ with $'*'\notin\{d\}$

B. Tolerance relation

Let $IDS=(U,C\cup\{d\})$ be an incomplete information system, for $B\subseteq C$, the subset B determines a binary relation, denoted by $SIM(B)$, which is defined as follows:

$$SIM(B)=\left\{(u,v)\in U\times U\,\middle|\,\forall b\in B, b(u)=b(v)\vee b(u)='*'\vee b(v)='*'\right\}$$

Then $SIM(B)$ is not an equivalence relation because it is symmetric and reflexive but not transitive. $SIM(B)$ is called the tolerance relation on U. Obviously $SIM(B)=\cap_{b\in B}SIM(\{b\})$.

Generally, $TC_B(u)$ denotes the maximal set of objects which are possibly indistinguishable by B with object u. Equivalently,

$TC_B(u)=\{v\in U\,|(u,v)\in SIM(B)\}$. Let $U/SIM(B)$ denote the family set $\{TC_B(u)|u\in U\}$, which is the classification induced by B. Any element from $U/SIM(B)$ will be called a tolerance class. Tolerance classes in $U/SIM(B)$ do not constitute a partition of U in general. They form a covering of U, i.e., $TC_B(u)=\varnothing$ for every $u\in U$, and $\bigcup_{u\in U}TC_B(u)=U$

C.Covering

For $B\subseteq C$, the tolerance relation $SIM(B)$ determines a covering on U, denoted by

$K(B)=U/SIM(B)=\{TC_B(u)|u\in U\}$.

Denote $COVER(U)=\{K(B)|B\subseteq C\}$ is the set of all the covering of U generated by subsets of attribute $B\subseteq C$ On $COVER(U)$, the smallest element $K(\omega)=\{TC_B(u)|\,TC_B(u)=\{u\},u\in U\}$ is called a discrete covering, the largest element $K(\delta)=\{TC_B(u)|\,TC_B(u)=U,\,u\in U\}$ is called a block covering

A partial relation $\preceq$ on $COVER(U)$ is defined as follows

$K(P)\preceq K(Q)\Leftrightarrow TC_P(u)\subseteq TC_Q(u),\forall u\in U$

Equation sign $K(P)=K(Q)\Leftrightarrow TC_P(u)=TC_Q(u),\forall u\in U$

$K(P)\prec K(Q)\Leftrightarrow K(P)\preceq K(Q)$. and $K(P)\neq K(Q)$.

D. Definition 2.1. Let $IDS=(U,C\cup D)$ be an incomplete decision system, for $U=\{u_1,u_2,...,u_n\}$ and $B\subseteq C$. Then, the tolerance matrix of tolerance relation $SIM(P)$, denoted by $M(B)=[b_{ij}]_{n\times n}$ is defined as:

$$M(B)=\begin{bmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & b_{nn} \end{bmatrix}$$

where $b_{ij}\in\{0,1\}$. $b_{ij}=1$ if $u_j\in TC_B(u_i)$ and $b_{ij}=0$ if $u_j\notin TC_B(u_i)$ with $i,j=1..n$

For the representation of tolerance $SIM(P)$. with the tolerance matrix $M(B)$, we have $\forall u_i\in U$, $TC_B(u_i)=\{u_j\in U|b_{ij}=1\}$ and $|TC_B(u_i)|=\sum_{j=1}^{n}b_{ij}$. With $P,Q\subseteq C,u\in U$, we have $TB_{P\cup Q}(u)=TB_P(u)\cap TB_Q(u)$.

Suppose $M(P)=[p_{ij}]_{n\times n}$, $M(Q)=[q_{ij}]_{n\times n}$ are two tolerance matrices of $SIM(P)$, $SIM(Q)$, then the tolerance matrix on the set of attribute $S=P\cup Q$ is: $M(S)=M(P\cup Q)=[s_{ij}]_{n\times n}$ with $s_{ij}=p_{ij}.q_{ij}$.

III. BUILDING A DISTANCE MEASURE IN INCOMPLETE DECISION SYSTEMS

In this section, we present the attribute reduction method in the incomplete decision table based on the distance built in section III.B. The proposed method includes the following steps:

- Define of reduct based on distance
- Define the importance of an attribute based on distance.
- Building a heuristic algorithm to find reduct in the direction of the filter and wrapper

A. Building a distance measure between two sets

Let U is a objects set, $U\neq\varnothing$. A distance on U is a mapping $d:U\times U\to[0,\infty)$ that satisfies the following three conditions:

(1) $d(u,v)\geq 0$ with $\forall u,v\in U$. $d(u,v)=0\Leftrightarrow u=v$;

(2) $d(u,v)=d(v,u)$ with $\forall u,v\in U$;

(3) $d(u,v)+d(v,t)\geq d(u,t)$ with $\forall u,v,t\in U$.

B. Building a distance measure between two coverings

Proposition 1

Let X, Y are two finite sets. The distance between X and Y is defined as $d(X,Y)=|X\cup Y|-|X\cap Y|$

Proof

Obviously $|X\cup Y|\geq|X\cap Y|$ should $d(X,Y)\geq 0$.

Moreover, $d(X,Y)=d(Y,X)$. Without loss of generality, we prove:

$d(X,Y)+d(X,Z)\geq d(Y,Z)$

Suppose $U=\{u_1,u_2,...,u_n\}$. we perform subset $X\subseteq U$ by an n-dimensional vector $V^X=(v_1^X,v_2^X,...,v_n^X)$ with $v_k^X=1$ *if* $u_k\in X$ and $v_k^X=0$ *if* $u_k\notin X$.

Set $V^{XY}=V^XV^Y=v_1^X.v_1^Y+v_2^X.v_2^Y+...+v_n^X.v_n^Y$, then $|X\cap Y|=V^{XY}$ and $|X\cup Y|=V^{XX}+V^{YY}-V^{XY}$, $d(X,Y)=V^{XX}+V^{YY}-2V^{XY}$.

From there we have:

$d(X,Y)+d(X,Z)-d(Y,Z)=V^{XX}+V^{YY}-2V^{XY}+V^{XX}+V^{ZZ}-2V^{XZ}-V^{YY}-V^{ZZ}+2V^{YZ}$

$=2V^{XX}-2V^{XY}-2V^{XZ}+2V^{YZ}$

Other way, $(V^X-V^Y)(V^X-V^Z)\geq 0$ or $V^{XX}-V^{XY}-V^{XZ}+V^{YZ}\geq 0$ satisfies because the k-th element of $(V^X-V^Y)(V^X-V^Z)$ is 0 and 1.

Therefore, $d(X,Y)+d(X,Z)-d(Y,Z)\geq 0$ or $d(X,Y)+d(X,Z)\geq d(Y,Z)$.

Proposition 2

Let $IDS=(U,C\cup\{d\})$ with $U=\{u_1,u_2,...,u_n\}$ and $U/SIM(X)$, $U/SIM(Y)$ are two covering generator by $X,Y\subseteq C$. Then the distance between two covering $U/SIM(X)$ and $U/SIM(Y)$ referred to as the distance between two attribute set X and Y is determined as follows:

$$D(X,Y)=\frac{1}{n^2}\sum_{i=1}^{n}\left(|TC_X(u_i)\cup TC_Y(u_i)|-|TC_X(u_i)\cap TC_Y(u_i)|\right)$$

Proof

Obvious $D(X,Y)\geq 0$ and $D(X,Y)=D(Y,X)$. We need prove triangle inequality. Without loss of generality, with $\forall X,Y,Z\subseteq C$. We prove: $D(X,Y)+D(X,Z)\geq D(Y,Z)$

From Proposition 3.1, with $\forall u_i\in U$, we have: $d(TC_X(u_i),TC_Y(u_i))+d(TC_X(u_i),TC_Z(u_i))\geq d(TC_Y(u_i),TC_Z(u_i))$

therefrom: $D(X,Y)+D(X,Z)=$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\left(|TC_X(u_i)\cup TC_Y(u_i)|-|TC_X(u_i)\cap TC_Y(u_i)|\right)$$

$$+\frac{1}{n^2}\sum_{i=1}^{n}\left(|TC_X(u_i)\cup TC_Z(u_i)|-|TC_X(u_i)\cap TC_Z(u_i)|\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} d\left(TC_X(u_i), TC_Y(u_i)\right) + \frac{1}{n^2}\sum_{i=1}^{n} d\left(TC_X(u_i), TC_Z(u_i)\right)$$

$$\geq \frac{1}{n^2}\sum_{i=1}^{n} d\left(TC_Y(u_i), TC_Z(u_i)\right) = D(Y,Z)$$

Obviously, $D(X,Y)$ reaches a minimal is 0 $\Leftrightarrow K(X)=K(Y)$ or $TC_X(u_i)=TC_Y(u_i), \forall u_i \in U$ and $D(X,Y)$

$D(X,Y)$ reaches a minimax is $1-\frac{1}{n}$ $\Leftrightarrow K(X)=K(\omega)$ and $K(Y)=K(\delta)$ (or $K(X)=K(\delta)$ and $K(Y)=K(\omega)$).

Therefore, $0 \leq D(X,Y) \leq 1-\frac{1}{n}$.

Proposition 3

Let $IDS=(U,C\cup\{d\})$ *with* $U=\{u_1,u_2,...,u_n\}$ *and* $M(C)=\left[c_{ij}\right]_{n\times n}$, $M(\{d\})=\left[d_{ij}\right]_{n\times n}$ are the tolerance matrices on C and d, respectively. Then, the distance between two sets of attributes C and $C\cup\{d\}$ is determined as follows:

$$D(C,C\cup\{d\}) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\left|TC_C(u_i)\right| - \left|TC_C(u_i)\cap TC_{\{d\}}(u_i)\right|\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(c_{ij} - c_{ij}.d_{ij}\right)$$

Proof

From proposition 2, we have:

$$D(C,C\cup\{d\}) = \frac{1}{n^2}\sum_{i=1}^{n}\left(\left|TC_C(u_i)\cup TC_{C\cup\{d\}}(u_i)\right| - \left|TC_C(u_i)\cap TC_{C\cup\{d\}}(u_i)\right|\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\left(\left|TC_C(u_i)\cup\left(TC_C(u_i)\cap S_{\{d\}}(u_i)\right)\right| - \left|TC_C(u_i)\cap TC_{\{d\}}(u_i)\right|\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\left(\left|TC_C(u_i)\right| - \left|TC_C(u_i)\cap TC_{\{d\}}(u_i)\right|\right)$$

Obviously, $D(C,C\cup\{d\})=0$ then $K(C)\preceq K(\{d\})$ and $D(C,C\cup\{d\})=1-\frac{1}{n}$ then $K(C)=K(\delta)$ and $K(\{d\})=K(\omega)$.

Proposition 4

Let $IDS=(U,C\cup\{d\})$ *with* $U=\{u_1,u_2,...,u_n\}$. *If* $P\subseteq C$ *then* $D(P,P\cup\{d\}) \geq D(C,C\cup\{d\})$

Proof

With $\forall u_i \in U, i=1..n$ we have $TC_C(u_i)\subseteq TC_P(u_i)$, therefore:

$$\left(TC_P(u_i)-TC_C(u_i)\right)\cap TC_{\{d\}}(u_i) \subseteq TC_P(u_i)-TC_C(u_i)$$

$$\Leftrightarrow \left(TC_P(u_i)\cap TC_{\{d\}}(u_i)\right) - \left(TC_C(u_i)\cap TC_{\{d\}}(u_i)\right) \subseteq TC_P(u_i)-TC_C(u_i)$$

$$\Leftrightarrow \left|\left(TC_P(u_i)\cap TC_{\{d\}}(u_i)\right) - \left(TC_C(u_i)\cap TC_{\{d\}}(u_i)\right)\right| \leq \left|TC_P(u_i)-TC_C(u_i)\right| \quad (1)$$

By $TC_C(u_i)\subseteq TC_P(u_i)$ then $TC_C(u_i)\cap TC_{\{d\}}(u_i)\subseteq TC_P(u_i)\cap TC_{\{d\}}(u_i)$ and (1) $\Leftrightarrow$

$$\left|TC_P(u_i)\cap TC_{\{d\}}(u_i)\right| - \left|TC_C(u_i)\cap TC_{\{d\}}(u_i)\right| \leq \left|TC_P(u_i)\right| - \left|TC_C(u_i)\right|$$

$$\Leftrightarrow \left|TC_P(u_i)\right| - \left|TC_P(u_i)\cap TC_{\{d\}}(u_i)\right| \geq \left|TC_C(u_i)\right| - \left|TC_C(u_i)\cap TC_{\{d\}}(u_i)\right| \quad (2)$$

By $TC_P(u_i)\cap TC_{\{d\}}(u_i)\subseteq TC_P(u_i), TC_C(u_i)\cap TC_{\{d\}}(u_i)\subseteq TC_C(u_i)$ then (2) $\Leftrightarrow$

$$\left|TC_P(u_i)\cup\left(TC_P(u_i)\cap TC_{\{d\}}(u_i)\right)\right| - \left|TC_P(u_i)\cap\left(TC_P(u_i)\cap TC_{\{d\}}(u_i)\right)\right|$$

$$\geq \left|TC_C(u_i)\cup\left(TC_C(u_i)\cap TC_{\{d\}}(u_i)\right)\right| - \left|TC_C(u_i)\cap\left(TC_C(u_i)\cap TC_{\{d\}}(u_i)\right)\right| \quad (3)$$

By $\begin{cases} TC_{C\cup\{d\}}(u_i) = TC_C(u_i)\cap TC_{\{d\}}(u_i) \\ TC_{P\cup\{d\}}(u_i) = TC_P(u_i)\cap TC_{\{d\}}(u_i) \end{cases}$ then (3)

$$\Leftrightarrow \frac{1}{n^2}\sum_{i=1}^{n}\left|TC_P(u_i)\right| - \left|TC_{P\cup\{d\}}(u_i)\right| \geq \frac{1}{n^2}\sum_{i=1}^{n}\left|TC_C(u_i)\right| - \left|TC_{C\cup\{d\}}(u_i)\right| \quad (4)$$

Form proposition 3 and (4) $\Leftrightarrow D(P,P\cup\{d\}) \geq D(C,C\cup\{d\})$.

C. A distance based attribute reduction in incomplete decision tables

Define 1

Let $IDS=(U,C\cup\{d\})$ with $R\subseteq C$,

If $\begin{cases} D(R,R\cup\{d\}) = D(C,C\cup\{d\}) \\ D\left(R-\{c\},\{R-\{c\}\}\cup\{d\}\right) \neq D(C\cup\{d\}) \end{cases}$ $\forall c\in R$

then R is reduct of C based on distance.

Define 2

Let $IDS=(U,C\cup\{d\})$ with $R\subset C$ and $c\in C-R$. The significance measure of feature *c* for attribute set *R* is defined by: $SIG_R(c)=D(R,R\cup\{d\})-D(R\cup\{c\},R\cup\{c\}\cup\{d\})$

Form proposition 4, we have $SIG_R(c)\geq 0$

$SIG_R(c)$ is classification quality of attribute *c* for decision attribute *d* and is used as the attribute selection criterion for heuristic algorithm to find reduct.

Next, we built a heuristic algorithm to find reduct by the approach of filter and wrapper.

The idea of the algorithm is to start from the empty set, $P=\varnothing$, in turn adding to the set *P* the attribute of greatest importance until the reduct is found.

Algorithm IDS_FW_DAR

(Filter- Wrapper Distance based Attribute Reduction in Incomplete Decision Tables)

Input: An incomplete decision tables $IDS=(U,C\cup\{d\})$

Output: An Attribute Reduct R_{best}

1. Let $P\leftarrow\varnothing$ and $W\leftarrow\varnothing$;
2. compute $M(P)$, $M(C)$, $M(\{d\})$, $D(P,P\cup\{d\})$, $D(C,C\cup\{d\})$

// filter phase

3.While $D(P,P\cup\{d\})\neq D(C,C\cup\{d\})$ do

4.Begin

5. for each $p\in C-P$ compute

$SIG_P(p)=D(P,P\cup\{d\})-D(P\cup\{p\},P\cup\{p\}\cup\{d\})$

6. Chooice $p_m\in C-P$ so that $SIG_P(p_m)=\underset{p\in C-P}{Max}\{SIG_P(p)\}$;

7. $P:=P\cup\{p_m\}$;

8. $T:=T\cup B$;

9. Compute tolerance matrix $M(P)$, distance $D(P,P\cup\{d\})$;

10.End;

// wrapper phase

11. Let $w=|W|$ // $W=\left\{\{c_{i_1}\},\{c_{i_1},c_{i_2}\},...,\{c_{i_1},c_{i_2},...,c_{i_w}\}\right\}$;w is elements of W

12. Let $W_1=\{c_{i_1}\},W_2=\{c_{i_1},c_{i_2}\},...,W_w=\{c_{i_1},c_{i_2},...,c_{i_w}\}$

13. For *j = 1* to *w*

14. Use the 10fold method to calculate classification accuracy on W_j with a classifier

15. $R_{best}=W_{jo}$.

16.Return R_{best};

Time complexity analysis of algorithm IDS_FW_DAR

Symbols $|C|,|U|$ are the number of conditional attributes and object numbers of DS.

The complexity of the tolerance matrix, the distance in statement 2 is $O\left(|C|*|U|^2\right)$. Consider the While loop from statements 3 through 9, to calculate $SIG_P(p)$ we have to calculate: $D\left(P\cup\{p\},P\cup\{p\}\cup\{d\}\right)$ because $D\left(P,P\cup\{d\}\right)$ was calculated in the previous step. The complexity to calculate $D\left(P\cup\{p\},P\cup\{p\}\cup\{d\}\right)$ when knowing $D\left(P,P\cup\{d\}\right)$ is $O\left(|U|^2\right)$. Because there are two nested loops in $|C|$, the complexity of the While loop is $O\left(|C|^2*|U|^2\right)$. Therefore the complexity of the filter stage is $O\left(|C|^2*|U|^2\right)$. The complexity of the wrapper phase depends on the complexity of the classifier used. Assuming the complexity of the classifier is $O(T)$, then the complexity of the wrapper stage is $O\left(|C|*T\right)$.

Therefore, the complexity of IDS_FW_DAR algorithm is $O\left(|C|^2*|U|^2\right)+O\left(|C|*T\right)$

Example 1. Consider an incomplete decision tables shown in Table 1,

$U=\{u_1,u_2,u_3,u_4,u_5,u_6\}$ is the object set, $C=\{c_1,c_2,c_3,c_4\}$ is the condition attribute set, and $D=\{d\}$ is the decision attribute set. For convenience, in the sequel, c_1, c_2, c_3 *and* c_4 will stand for Price, Mileage, Size, Max-speed, respectively. Find attribute reduct $P\subseteq C$.

Table 1: An incomplete decision tables.

Car	Price	Mileage	Size	Speed	Acceleration
u_1	Medium	Medium	Medium	Medium	Excel
u_2	Low	*	Medium	Medium	Excel
u_3	*	*	Full	Medium	Poor
u_4	Medium	*	Medium	High	Excel
u_5	*	*	Medium	High	Good
u_6	Low	Medium	Medium	*	Excel

Init: Let $P:=\varnothing$; $TC_P(u)=U$ for $\forall u\in U$;

- Compute $M(P)$, $M(C)$, $M(\{d\})$

$$M(C)=\begin{bmatrix}1&0&0&0&0&0\\0&1&0&0&0&1\\0&0&1&0&0&0\\0&0&0&1&1&0\\0&0&0&1&1&1\\0&1&0&0&1&1\end{bmatrix},M(\{d\})=\begin{bmatrix}1&1&0&1&0&1\\1&1&0&1&0&1\\0&0&1&0&0&0\\1&1&0&1&0&1\\0&0&0&0&1&0\\1&1&0&1&0&1\end{bmatrix}$$

- Comput the distances $D\left(P,P\cup\{d\}\right)$, $D\left(C,C\cup\{d\}\right)$

$$D\left(P,P\cup\{d\}\right)=D\left(\varnothing,\{d\}\right)=\frac{18}{36}$$

$$D\left(C,C\cup\{d\}\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(c_{ij}-c_{ij}.d_{ij}\right)=\frac{1}{6^2}\sum_{i=1}^{6}\sum_{j=1}^{6}\left(c_{ij}-c_{ij}.d_{ij}\right)=\frac{4}{36}$$

Because of $D\left(P,P\cup\{d\}\right)=\frac{18}{36}\neq\frac{4}{36}=D\left(C,C\cup\{d\}\right)$ then doing the loop, we have: $p\in C-P=\{c_1,c_2,c_3,c_4\}$ *and we get*

$D\left(\{c_1\},\{c_1\}\cup\{d\}\right)=\frac{18}{36}\rightarrow$

$SIG_P(c_1)=D\left(P,P\cup\{d\}\right)-D\left(P\cup\{c_1\},P\cup\{c_1\}\cup\{d\}\right)=\frac{18}{36}-\frac{18}{36}=0\Leftrightarrow$

$SIG_P(c_1)=0$

$D\left(\{c_2\},\{c_2\}\cup\{d\}\right)=\frac{18}{36}\rightarrow SIG_P(c_2)=0$;

$D\left(\{c_3\},\{c_3\}\cup\{d\}\right)=\frac{8}{36}\rightarrow SIG_P(c_3)=\frac{10}{36}$;

$D\left(\{c_4\},\{c_4\}\cup\{d\}\right)=\frac{10}{36}\rightarrow SIG_P(c_4)=\frac{8}{36}$

Thus $SIG_P(c_3)=\frac{10}{36}>SIG_P(c_4)=\frac{8}{36}>SIG_P(c_1)=SIG_P(c_2)=0$ then chosing the attribute of c_3. So that $P:=P\cup\{c_3\}=\{c_3\}$

- Compute $D\left(P,P\cup\{d\}\right)=D\left(\{c_3\},\{c_3\}\cup\{d\}\right)=\frac{8}{36}$

Thus $D\left(P,P\cup\{d\}\right)=\frac{8}{36}\neq\frac{4}{36}=D\left(C,C\cup\{d\}\right)$ then move on the second loop.

With $p\in C-P=\{c_1,c_2,c_4\}$ *we get:* $D\left(P\cup\{c_1\},P\cup\{c_1\}\cup\{d\}\right)=\frac{8}{36}\rightarrow$

$SIG_P(c_1)=0$; $D\left(P\cup\{c_2\},P\cup\{c_2\}\cup\{d\}\right)=\frac{8}{36}\rightarrow SIG_P(c_2)=0$;

$D\left(P\cup\{c_4\},P\cup\{c_4\}\cup\{d\}\right)=\frac{4}{36}\rightarrow SIG_P(c_4)=\frac{4}{36}$

Thus $SIG_P(c_4)=\frac{4}{36}>SIG_P(c_1)=SIG_P(c_2)=0$ then chosing the attribute of c_4 and $P=\{c_3,c_4\}$

- Comput the distances $D\left(P,P\cup\{d\}\right)=D\left(\{c_3,c_4\},\{c_3,c_4\}\cup\{d\}\right)=\frac{4}{36}$

Thus $D\left(P,P\cup\{d\}\right)=\frac{4}{36}=D\left(C,C\cup\{d\}\right)$ so stop. Therefore, the subset $P=\{c_3,c_4\}$ is an attribute reduct of the system.

IV. EXPERIMENTAL ANALYSIS

We use algorithm 1 [12] to compare. Data taken from UCI[10] data store. Let ***N*** be the number of objects, ***C*** be the number of conditional attributes, ***R*** be the attribute number of the reduct, ***t*** is the time (in seconds) to perform the algorithm; Conditional attributes are numbered as 1,2, etc. The results are described in table 2 and table 3.

Table 2
Comparison of computation time between Algorithms 1[12] and IDS_FW_DAR algorithms

ID	Data sets	|N|	|C|	Algorithms [1]		IDS_FW_DAR algorithms	
				|R|	*t*	|R|	*t*
1	Hepa titis.data	155	19	4	1.29	4	0.33
2	Lung cancer.data	32	56	4	0.17	4	0.07
3	Auto_mobile.data	205	25	5	1.68	5	0.78
4	Anneal.data	798	38	9	178.0	9	48.17
5	Congressional_Voting_Records	435	16	13	16.73	15	8.14
6	Credit_Approval	690	15	7	15.68	4	5.70

Table 3
Comparison of the sizes between Algorithms 1[12] and IDS_FW_DAR algorithms

ID	Data sets	|N|	|C|	Algorithms [1]	IDS_FW_DAR algorithms
1	Hepa titis.data	155	19	{1,2,4,17}	{1,2,4,17}
2	Lung cancer.data	32	56	{3,4,9,43}	{3,4,9,43}
3	Auto_mobile.data	205	25	{1,13,14,20, 21}	{1,13,14,20, 21}
4	Anneal.data	798	38	{1,3,4,5,8,9, 33,34,35}	{1,3,4,5,8,9,3 3,34,35}
5	Congressional_ Voting_Records	435	16	{1,2,3,4,5,8, 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	690	15	{1,2,3,4,5,6, 8}	{2,3,6,8}

Test results show that:

- The execution time of IDS_FW_DAR Algorithm is faster.
- The reduct collected by the IDS_FW_DAR Algorithm is more optimal in the 6th data.

V. CONCLUSIONS

In recent years, there are a large number of research on the attribute reduction in the incomplete decision table based on the approach to the tolerance rough set. The incremental algorithm has been proposed to find the reduction in the incomplete decision table by the traditional filtering approach. However, the reduction collected is not optimal in terms of number of properties and classification accuracy.

In this paper, we propose a heuristic method combined two methods filter and wrapper to find the reduct in incomplete decision tables based on distance which include the following steps: building a distance measurement between two sets of attributes and Building a distance measure between two coverings; defining a reduct based on the distance; defining the significance measure of feature based on the distance and building algorithm to find a best reduct based on the distance.

We demonstrate the reduct based on the distance the optimal resulting in classification quality, efficiency of the number of the attribute comparision with some other filter algorithms.

The results of this paper supplement the attribute reduction methods in incomplete decision tables

The next development direction of the authors is to study the attribute reduction methods in incomplete decision tables with change attribute set in the case of adding and delete attribute sets. Building incremental filter-wrapper algorithm for distance based attribute reduction when adding and delete objects to find the reduct.

ACKNOWLEDGEMENTS

This research has been funded by the Research Project, VAST 01.10/20-21. Vietnam Academy of Science and Technology.

REFERENCES

1. Vu Van Dinh, Vu Duc Thi, Ngo Quoc Tạo, Nguyen Long Giang, “Partition Distance Based Attribute Reduction in Incomplete Decision Tables”, Journal of Information & Communications, Vol. V-2, No. 14(34), 2015.
2. Vu Van Dinh, Nguyen Long Giang, Duc Thi Vu, "Generalized Discernibility Function based Attribute Reduction in Incomplete Decision Systems", Serdica Journal of Computing 7 (2013), Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, No 4, 2013, pp. 375-388
3. The UCI machine learning repository, <http://archive.ics.uci.edu/ml/datasets.html>
4. Pawlak Z. (1982), “Rough sets”, International Journal of Computer and Information Sciences, 11(5): 341-356.
5. Nguyen Long Giang, Vu Van Dinh, Relationships Among the Concepts of Reduct in Incomplete Decision Tables, Frontiers in Artificial Intelligence and Applications (FAIA), Volume 252: Advanced Methods and Technologies for Agent and Multi-Agent Systems, IOS Press, 2013, pp. 417-426
6. Nguyen Long Giang, Vu Duc Thi, “Algorithm for finding all reducts of a decision”, Journal of Computer Science and Cybernetics, Vol 27, No 3 (2011), pp. 199-205.
7. Nguyen Long Giang, Nguyen Thanh Tung, “A new method for attribute reduction in decision tables based on metric”, Proceedings of 14th National Conference on Information Technology and Telecommunication, Can Tho, 10/2011, pp. 249-266.
8. Nguyen Long Giang, “Study on some data mining methods based on rough set theory”, Doctoral Dissertation, Institute of Information Technology, Vietnam Academy of Science and Technology (2012).
9. Nguyen Long Giang, Nguyen Hung Son, “Metric Based Attribute Reduction in Incomplete Decision Tables”, Proceedings of 14th International Conference, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2013, Halifax, NS, Canada, Lecture Notes in Computer Science, SpingerLink, Vol. 8170, 2013, pp. 99-110.
10. Kryszkiewicz M. (1998), “Rough set approach to incomplete information systems”, Information Science, Vol. 112, pp. 39-49.
11. Huang B., Li H. X. and Zhou X. Z., “Attribute Reduction Based on Information Quantity under Incomplete Information Systems”, Systems Application Theory & Practice, Vol. 34, 2005, pp. 55-60
12. Demetrovics Janos, Vu Duc Thi, Nguyen Long Giang, A Distance-based Method for Attribute Reduction in Incomplete Decision Systems, Serdica Journal of Computing 7 (2013), No 4, pp. 355-374

AUTHORS PROFILE

Nguyen Anh Tuan obtained the Master's degree on information technology at the University of Technology - Vietnam National University, Hanoi, 2000. The main areas include Artificial Intelligence, Data Mining, Rough Sets, Fuzzy Rough Sets.

Nguyen Long Giang obtained the PhD degree on Math Fundamentals for Informatics at Institute of Information Technology (IoIT), Vietnam Academy of Science and Technology (VAST) in 2012. His major fields include Artificial Intelligence, Data Mining, Soft Computing, Fuzzy Computing, Rough Sets, Fuzzy Rough Sets.