

Fuzzy Distance Based Attribute Reduction in Decision Tables

Cao Chinh Nghia, Vu Duc Thi, Nguyen Long Giang, Tan Hanh

Abstract: In recent years, fuzzy rough set based attribute reduction has attracted the interest of many researchers. The attribute reduction methods can perform directly on the decision tables with numerical attribute value domain. In this paper, we propose a fuzzy distance based attribute reduction method on the decision table with numerical attribute value domain. Experiments on data sets show that the proposed method is more efficient than the ones based on Shannon's entropy on the executed time and the classification accuracy of reduct.

Keywords: Fuzzy rough set, fuzzy decision table, fuzzy equivalence relation, fuzzy distance, attribute reduction, reduct.

I. INTRODUCTION

Attribute reduction is an important issue in data preprocessing steps which aims at eliminating redundant attributes to enhance the effectiveness of data mining techniques. Rough set theory [12] is an effective approach to solve feature selection problems with discrete attribute value domain. Traditional rough set based attribute reduction techniques have many limitations when performing on tables with numerical attribute value domain. Data needs to be discretized before performing attribute reduction techniques. The major limitation of rough set theory based attribute reduction is losing information in the discrete processing, which will affect the quality of data classification. To solve the problem of attribute reduction directly on decision table with numerical data, fuzzy rough set based approach has recently been developed [3-6, 10, 16, 17].

Dubois D., and Prade H., proposed fuzzy rough set theory [3, 4] which is a combination of rough set theory [12] and fuzzy set theory [18] in order to

approximate fuzzy sets based on fuzzy equivalence relation. In rough set theory, two objects are called equivalent on R attribute set (the similarity is 1) if their attribute values are equal on all attributes of R . Conversely, they are not equal (the similarity is 0). Equivalence relation is the foundation to determine the partitions of the objects on a space object. The equal values on the same attribute set belong to the equivalence class. In the fuzzy rough set theory, in order to determine the equivalence of the two objects, the concept of equivalence relation is no longer valid and replaced by a fuzzy equivalence relation. The value equivalence in the range $[0, 1]$ shows the close or similar properties of two objects. The equivalence relation determines fuzzy partitions on a space object, the equivalence class of an object is the entire universal. Thus, if a data set has n objects, it would have n fuzzy equivalence classes.

Fuzzy rough set based attribute reduction methods focus on two directions: fuzzy partition and fuzzy equivalence relation. The first direction is to propose attribute reduction methods based on fuzzy partition. Jensen and Shen [9, 10] have proposed a heuristic algorithm to find one reduction of decision table. However, the biggest drawback of the algorithm is its computational complexity, the complexity in the worst case is exponentially increased [9, 10, 16] with respect to the conditional attribute set. Thus, this approach is only academic, not so feasible when applied in reality, and just few experts are interested in this research. The second direction is to propose attribute reduction methods based on fuzzy equivalence relation matrix. The fuzzy equivalence relation matrix is calculated based on a fuzzy equivalence relation defined on values of attribute sets. Then the general computational complexity is polynomial function [5, 6, 10, 16, 17]. According to this direction, Degang Chen et al. [1, 16] have proposed algorithm finding all

reducts by extending attribute reduction methods based on discernibility matrix in traditional rough set theory. Dai Jianhua et al. [5] have calculated fuzzy information gain of the Shannon's entropy based on fuzzy equivalence classes and they have proposed a heuristic algorithm to find a best reduct based on fuzzy information gain. From their experiments, they also demonstrated that their method is better than the traditional rough set methods on the classification accuracy of data. Though the time complexity of the algorithm is polynomial, the calculation time of this method is still long due to the usage of logarithm formulas, especially on large data sets.

In this paper, we have proposed a heuristic algorithm to find the best reduct of decision tables with numerical attribute value domain using fuzzy distance, called F_DBAR algorithm. By experiments on data sets from UCI [19], we will show that the execution time of F_DBAR is smaller than that of algorithm GAIN_RATIO_AS_FRS based on fuzzy information gain [5]. Furthermore, the classification accuracy of reduct generated by algorithm F_DBAR is higher than that of reduct generated by GAIN_RATIO_AS_FRS [5]. The structure of the paper is as follows. Section II presents some basic concepts of fuzzy rough set theory. Section III presents some concepts of fuzzy distances between two finite sets. Section IV presents an attribute reduction algorithm using fuzzy distance and an example of the algorithm. Section V presents some experiments on data sets from UCI [19]. Finally, Section VI gives a conclusion and future research.

II. BASIC CONCEPTS IN FUZZY ROUGH SET

II.1. Fuzzy relation matrix

Definition 1 [7, 8, 15]. Let $U = \{x_1, \dots, x_n\}$ be a non-empty finite set and R be a relation on U . The relation matrix of R , denoted by $M(R)$, is defined as

$$M(R) = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{bmatrix}$$

where $r_{ij} = R(x_i, x_j)$ is the relation value of x_i and x_j , $r_{ij} \in [0, 1]$.

Definition 2 [7, 8, 15]. A relation R defined on U is called fuzzy equivalence relation if it satisfies the following conditions:

- 1) Reflectivity: $R(x, x) = 1, \forall x \in U$
- 2) Symmetry: $R(x, y) = R(y, x), \forall x, y \in U$
- 3) Transitivity: $R(x, z) \geq \min\{R(x, y), R(y, z)\} \forall x, y, z \in U$

Definition 3 [8]. Let U be a non-empty finite set and R be a fuzzy equivalence relation on U . Some operations of R are defined as

- 1) $R_1 = R_2 \Leftrightarrow R_1(x, y) = R_2(x, y), \forall x, y \in U$
- 2) $R = R_1 \cup R_2 \Leftrightarrow R(x, y) = \max\{R_1(x, y), R_2(x, y)\}$
- 3) $R = R_1 \cap R_2 \Leftrightarrow R(x, y) = \min\{R_1(x, y), R_2(x, y)\}$
- 4) $R_1 \subseteq R_2 \Leftrightarrow R_1(x, y) \leq R_2(x, y)$

II.2. Fuzzy partition

Definition 4 [8]. Let $U = \{x_1, \dots, x_n\}$ be a non-empty finite set and R be a fuzzy equivalence relation on U . Then, a fuzzy partition is defined as

$$U / R = \{[x_i]_R\}_{i=1}^n$$

where $[x_i]_R$ is a fuzzy set, $[x_i]_R$ is also called a fuzzy equivalence class.

$$[x_i]_R = \left\{ \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \dots + \frac{r_{in}}{x_n} \right\}$$

The cardinality of fuzzy set $[x_i]_R$ is calculated as

$$|[x_i]_R| = \sum_{j=1}^n r_{ij} \quad (1)$$

Let $DS = (U, C \cup D)$ be a decision table with numerical attribute value domain, $P, Q \subseteq C$ and $R(P), R(Q)$ are fuzzy equivalence relations R on P, Q

corresponding. Then we have $R(P \cup Q) = R(P) \cap R(Q)$ [8], it means that for any $x, y \in U$, $R(P \cup Q)(x, y) = \min\{R(P)(x, y), R(Q)(x, y)\}$. Suppose that $M(R(P)) = [r_{ij}^{R(P)}]_{n \times n}$, $M(R(Q)) = [r_{ij}^{R(Q)}]_{n \times n}$ are relation matrices of R on the attribute sets P, Q corresponding, then the relation matrix of R on the attribute sets $P \cup Q$ is defined as

$$M(R(P \cup Q)) = [r_{ij}^{R(P \cup Q)}]_{n \times n} \text{ where } r_{ij}^{R(P \cup Q)} = \min\{r_{ij}^{R(P)}, r_{ij}^{R(Q)}\} \quad (2)$$

Example 1. A decision table $DS = (U, C \cup \{d\})$ is shown in Table 1 where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, $C = \{c_1, c_2, c_3, c_4\}$.

Table 1. The decision table with numerical attribute value.

U	c_1	c_2	c_3	c_4	d
u_1	0.8	0.1	0.1	0.5	1
u_2	0.3	0.5	0.2	0.8	1
u_3	0.2	0.2	0.6	0.7	0
u_4	0.6	0.3	0.1	0.2	1
u_5	0.3	0.4	0.3	0.3	0
u_6	0.2	0.3	0.5	0.3	0

A fuzzy equivalence relation $R(\{c_k\})$ is defined on attribute $c_k \in C$ as follows

$$R(\{c_k\})(u_i, u_j) = \begin{cases} 1 - 4 * \frac{|u_i - u_j|}{|\max(c_k) - \min(c_k)|}, & \text{if} \\ \frac{|u_i - u_j|}{|\max(c_k) - \min(c_k)|} \leq 0.25 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Where: $\max(c_k), \min(c_k)$ are maximum value, minimum value of the attribute c_k , respectively.

Then the relation matrix on attribute c_1 is calculated as follows

$$M(R(\{c_1\})) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.33 & 0 & 1 & 0.33 \\ 0 & 0.33 & 1 & 0 & 0.33 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0.33 & 0 & 1 & 0.33 \\ 0 & 0.33 & 1 & 0 & 0.33 & 1 \end{bmatrix}$$

The fuzzy equivalence class of object u_1 is denoted by

$$[u_1]_{R(\{c_1\})} = \frac{1}{u_1} + \frac{0}{u_2} + \frac{0}{u_3} + \frac{0}{u_4} + \frac{0}{u_5} + \frac{0}{u_6}$$

Similarly, $M(R(\{c_2\})), M(R(\{c_3\})), M(R(\{c_4\}))$ are calculated and $M(R(C))$ is calculated.

II.3. Fuzzy rough set

Definition 5. Given a finite object set U , a fuzzy equivalence relation R and a fuzzy set F . Then, the fuzzy lower approximation set $\underline{R}(F)$ and the fuzzy upper approximation set $\bar{R}(F)$ of F are fuzzy sets, the membership function of objects $x_i \in U$ is defined as [3, 4]

$$\mu_{\underline{R}(F)}(x) = \inf_{y \in U} \max(1 - \mu_R(x, y), \mu_F(y)) \quad (4)$$

$$\mu_{\bar{R}(F)}(x) = \sup_{y \in U} \min(\mu_R(x, y), \mu_F(y)) \quad (5)$$

Where $\mu_{[x]_R}(y) = \mu_R(x, y)$, then the fuzzy lower approximation set $\underline{R}(F)$ and the fuzzy upper approximation set $\bar{R}(F)$ are rewritten as

$$\mu_{\underline{R}(F)}(x) = \inf_{y \in U} \max(1 - \mu_{[x]_R}(y), \mu_F(y)) \quad (6)$$

$$\mu_{\bar{R}(F)}(x) = \sup_{y \in U} \min(\mu_{[x]_R}(y), \mu_F(y)) \quad (7)$$

It is easy to see that the membership function of objects $u_j \in U$ in fuzzy equivalence class $[u_i]_R$ is $\mu_{[u_i]_R}(u_j) = R(u_i, u_j) = r_{ij}$. Then, $(\underline{R}(F), \bar{R}(F))$ is called the fuzzy rough set [3, 4]. It is obviously that the set $X \subseteq U$ can be seen as a fuzzy set where the membership function $\mu_X(y) = 1$ if $y \in X$ and

$\mu_X(y) = 0$ if $y \notin X$. The fuzzy rough set model can be considered as using of the fuzzy equivalence relation to approximate the fuzzy set (or crisp set) by the fuzzy lower approximation set and the fuzzy upper approximation.

III. FUZZY DISTANCE MEASURE BASED ON FUZZY RELATION MATRIX

III.1. Jaccard distance between two finite sets

Given a finite object set U and $X, Y \subseteq U$. Jaccard's distance measured the similarity between two sets X and Y is defined as [11]

$$D(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \quad (8)$$

Based on Jaccard's distance, the authors have proposed some attribute reduction methods in decision tables [11]. Given a decision table $DS = (U, C \cup D)$ where $U = \{x_1, \dots, x_n\}$ and $P \subseteq C$, suppose that $[x_i]_P$ is an equivalence class which contain x_i in partition U/P . Based on Jaccard's distance, the distance between two attribute sets C and $C \cup D$ is defines as [11]

$$d(C, C \cup D) = 1 - \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{|[x_i]_C \cap [x_i]_{C \cup D}|}{|[x_i]_C \cup [x_i]_{C \cup D}|} \quad (9)$$

According to the results in [7], the formula (9) can be rewritten as follows

$$\begin{aligned} d(C, C \cup D) &= 1 - \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{|[x_i]_C \cap [x_i]_C \cap [x_i]_D|}{|[x_i]_C \cup ([x_i]_C \cap [x_i]_D)|} \\ &= 1 - \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{|[x_i]_C \cap [x_i]_D|}{|[x_i]_C|} \end{aligned} \quad (10)$$

The measure distance in the formula (10) characterizes the similarity between the conditional attribute set C and the decisional attribute set D . Based on the measure distance, authors [11] proposed an attribute reduction method in the decision tables, including: defined reduct based on the distance, defined the importance of the attribute based on the distance, designed a heuristic algorithm to find one reduct based

on the distance. Authors [11] also have proved by theoretical and experimental that the distance method is more effective than some other methods using Shannon entropy.

III.2. Fuzzy Jaccard distance measure between two finite sets

Using the distance measure in the formula (10), we have designed the fuzzy distance measure based on the fuzzy relational matrix according to fuzzy rough set approach.

Definition 6. Given a decision table with numerical attribute value $DS = (U, C \cup D)$, suppose that two fuzzy equivalence relations R_C and R_D are defined on two attribute sets C and D corresponding. Let r_{ij}^C be the elements of the fuzzy relation matrix $M(R_C)$, r_{ij}^D be the elements of the fuzzy relation matrix $M(R_D)$ where $1 \leq i, j \leq n$. Based on the formula (10), Definition 3 and Definition 4, fuzzy distance measure between two attribute sets C and $C \cup D$ is defined as

$$d_F(C, C \cup D) = 1 - \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{\sum_{j=1}^n \min\{r_{ij}^C, r_{ij}^D\}}{\sum_{j=1}^n r_{ij}^C} \quad (11)$$

Proposition 1. Given a decision table with numerical attribute value $DS = (U, C \cup D)$ and R_C, R_D are two fuzzy equivalence relations defined on C, D . Then, we have:

- 1) $0 \leq d_F(C, C \cup D) \leq 1$
- 2) $d_F(C, C \cup D) = 0$ when $R_C \subseteq R_D$

Proof:

1) According to formula (11), it is easy to see $0 \leq d_F(C, C \cup D) \leq 1$.

2) According to definition 3 and [7], we have $R_C \subseteq R_D \Leftrightarrow R_C(x, y) \leq R_D(x, y) \Leftrightarrow r_{ij}^C \leq r_{ij}^D, \forall i, j \in [1, n]$. By using formula (11) we have $d_F(C, C \cup D) = 0$.

Proposition 2. Given a decision table with numerical attribute value $DS = (U, C \cup D)$ and $B \subseteq C$, then we have $d_F(B, B \cup D) \geq d_F(C, C \cup D)$.

Proof: According to [7] we have $B \subseteq C \Leftrightarrow U/C \prec U/B$ (the partition U/C is much finer than the partition U/B) if and only if $[u]_C \subseteq [u]_B$. According to Definition 3 and [7] we have

$$[u]_C \subseteq [u]_B \Leftrightarrow [u]_{R(C)} \leq [u]_{R(B)} \Leftrightarrow \left| \sum_{i,j=1}^n r_{ij}^C \right| \leq \left| \sum_{i,j=1}^n r_{ij}^B \right|$$

$$\Leftrightarrow \sum_{i,j=1}^n r_{ij}^C \leq \sum_{i,j=1}^n r_{ij}^B. \text{ By } r_{ij}^C, r_{ij}^B \in [0,1] \text{ we have}$$

$$\frac{r_{ij}^D}{r_{ij}^C} \geq \frac{r_{ij}^D}{r_{ij}^B} \Leftrightarrow (1 - \frac{r_{ij}^D}{r_{ij}^C}) \leq (1 - \frac{r_{ij}^D}{r_{ij}^B}).$$

Instead formula (11) we have $d_F(B, B \cup D) \geq d_F(C, C \cup D)$.

IV. ATTRIBUTE REDUCTION BASED ON FUZZY DISTANCE MEASURE

In this section, we present an attribute reduction method of the decision table with numerical attribute value using the fuzzy distance measure. Similar to attribute reduction methods in traditional rough set theory, our method includes: defining the reduct based on fuzzy distance, defining the importance of the attribute and designing a heuristic algorithm to find the best reduct based on the importance of the attribute.

Definition 7. Given a decision table $DS = (U, C \cup D)$ with numerical attribute value and attribute set $R \subseteq C$. If

$$1) d_F(R, R \cup D) = d_F(C, C \cup D)$$

$$2) \forall r \in R, d_F(R - \{r\}, (R - \{r\}) \cup D) \neq d_F(C, C \cup D)$$

then R is a reduct of C based on fuzzy distance.

Definition 8. Given a decision table $DS = (U, C \cup D)$, $B \subseteq C$ and $b \in C - B$. The importance of attribute b to B is defined as

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

The importance of the attribute characterizes the classification quality of conditional attributes which respect to the decision attribute. It is used as the attribute selection criterion for heuristic algorithm to find the reduct.

F_DBAR Algorithm (Fuzzy Distance based Attribute Reduction): a heuristic algorithm to find the best reduct by using fuzzy distance.

Input: The decision table with numerical attribute value $DS = (U, C \cup D)$, the fuzzy relation equivalence R .

Output: The best reduct P

1. $P \leftarrow \emptyset; M(R_P) = 0;$
2. Calculate the relation matrix $M(R_C), M(R_D);$
3. Calculate the fuzzy distance $d_F(C, C \cup D);$
// Adding gradually to P an attribute having the greatest importance
4. For $d_F(P, P \cup D) \neq d_F(C, C \cup D)$ Do
5. Begin
6. For each $a \in C - R$
7. Begin
8. Calculate $d_F(P \cup \{a\}, P \cup \{a\} \cup D);$
9. Calculate
 $SIG_P(a) = d_F(P, P \cup D) - d_F(P \cup \{a\}, P \cup \{a\} \cup D);$
10. End;
11. Select $a_m \in C - P$ so that
 $SIG_P(a_m) = \max_{a \in C - P} \{SIG_P(a)\};$
12. $P = P \cup \{a_m\};$
13. Calculate $d_F(P, P \cup D);$
14. End;
- //Remove redundant attribute in P*
15. For each $a \in P$
16. Begin

17. Calculate $d_F(P - \{a\}, (P - \{a\}) \cup D)$;

18. If $d_F(P - \{a\}, (P - \{a\}) \cup D) = d_F(C, C \cup D)$

then $P = P - \{a\}$;

19. End;

20. Return P ;

The computational complexity of fuzzy equivalence relation matrix is $O(|C||U|^2)$ with $|C|$, the number of attribute of the data set, $|U|$ the number of element of the data set. Hence, the complexity of F_DBAR algorithm is $O(|C|^3|U|^2)$.

Example 2. Given a decision table with numerical attribute value $DS = (U, C \cup D)$ (Table 2) where

$$U = \{u_1, u_2, u_3, u_4, u_5, u_6\}, C = \{c_1, c_2, c_3, c_4, c_5, c_6\}.$$

Table 2. The decision table in the Example 2.

U	c_1	c_2	c_3	c_4	c_5	c_6	D
u_1	0.8	0.2	0.6	0.4	1	0	0
u_2	0.8	0.2	0	0.6	0.2	0.8	1
u_3	0.6	0.4	0.8	0.2	0.6	0.4	0
u_4	0	0.4	0.6	0.4	0	1	1
u_5	0	0.6	0.6	0.4	0	1	1
u_6	0	0.6	0	1	0	1	0

By using steps of F_DBAR algorithm, firstly we use the fuzzy similarity measure in formula (3) to calculate some relation matrices.

$P \leftarrow \emptyset$, $M(R_P) = 0$, $d_F(\emptyset, \emptyset \cup \{d\}) = 1$, calculate some fuzzy relation matrices $M(R\{c_1\})$, $M(R\{c_2\})$, $M(R\{c_3\})$, $M(R\{c_4\})$, $M(R\{c_5\})$, $M(R\{c_6\})$, $M(R\{C\})$, $M(\{D\})$

$$M(R\{c_1\}) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, M(R\{c_2\}) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$M(R\{c_3\}) = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, M(R\{c_4\}) = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$M(R\{c_5\}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.2 & 0.2 & 0.2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 1 & 1 & 1 \\ 0 & 0.2 & 0 & 1 & 1 & 1 \\ 0 & 0.2 & 0 & 1 & 1 & 1 \end{bmatrix}, M(R\{c_6\}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.2 & 0.2 & 0.2 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 1 & 1 & 1 \\ 0 & 0.2 & 0 & 1 & 1 & 1 \\ 0 & 0.2 & 0 & 1 & 1 & 1 \end{bmatrix}$$

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

Calculate:

$$d_F(C, C \cup D) = 0, d_F(\{c_1\}, \{c_1\} \cup \{D\}) = 0.38889$$

$$d_F(\{c_2\}, \{c_2\} \cup \{D\}) = 0.5, d_F(\{c_3\}, \{c_3\} \cup \{D\}) = 0.389$$

$$d_F(\{c_4\}, \{c_4\} \cup \{D\}) = 0.222, d_F(\{c_5\}, \{c_5\} \cup \{D\}) = 0.23958$$

$$d_F(\{c_6\}, \{c_6\} \cup \{D\}) = 0.23958, SIG_P(c_1) = 0.61111$$

$$SIG_P(c_2) = 0.5, SIG_P(c_3) = 0.611, SIG_P(c_4) = 0.778,$$

$$SIG_P(c_5) = 0.76042, SIG_P(c_6) = 0.76042. \text{ So attribute } \{c_4\} \text{ is selected.}$$

Similarity, $d_F(\{c_4, c_1\}, \{c_4, c_1\} \cup \{D\}) = 0$, checked $d_F(\{c_4, c_1\}, \{c_4, c_1\} \cup \{D\}) = d_F(C, C \cup D) = 0$, algorithm finished and $P = \{c_4, c_1\}$. Consequently, $P = \{c_4, c_1\}$ is the best reduct of DS .

V. EXPERIMENTS

We select the heuristic algorithm GAIN_RATIO_AS_FRS [5] (Called GRAF) to compare with algorithm F_DBAR on execution time, reduct and the classification accuracy of reduct generated two algorithms. We perform the following tasks:

1) Coding algorithm GRAF [5] and algorithm F_DBAR by C# language program. Both algorithms used the fuzzy equivalence relation defined by the formula (3).

2) On a PC with Pentium Core i3, 2.4 GHz CPU, 2 GB of RAM, using Windows 10 operating system, test two algorithms on 6 data sets from the UCI repository [19]. For each data set, assume that $|U|$ is the number of objects, $|R|$ is the number of attributes of the reduct, $|C|$ is the number of the conditional attributes, t is the time of operation (calculated by second), condition attributes will be denoted by 1, 2, ..., $|C|$.

The execution time and reduct of two algorithms are described in Table 3 and Table 4.

Table 3. The execution time of F_DBAR and GRAF [5]

No	Data set	$ U $	$ C $	F_DBAR		GRAF[5]	
				$ R $	t	$ R $	t
1	Ecoli	336	7	6	0.036	6	0.124
2	Fertility	100	9	8	0.017	7	0.021
3	Wdbc	569	30	15	9.624	17	12.146
4	Wpbc	198	33	16	5.016	17	6.725
5	Soybean (small)	47	35	19	0.079	21	0.105
6	Ionosphere	351	34	11	6.022	12	8.142

Table 4. Reducts of F_DBAR and GRAF[5]

No	Data set	F_DBAR	GRAF[5]
1	Ecoli	{1, 2, 3, 4, 6, 7}	{1, 2, 3, 4, 6, 7}
2	Fertility	{1, 2, 3, 5, 6, 7, 8, 9}	{1, 2, 3, 5, 6, 7, 8}
3	Wdbc	{1, 3, 4, 7, 8, 9, 12, 14, 16, 18, 19, 22, 24, 25, 30}	{1, 2, 4, 5, 7, 8, 9, 10, 12, 14, 16, 18, 19, 22, 23, 24, 30}
4	Wpbc	{1, 2, 5, 8, 9, 10, 13, 14, 15, 18, 19, 22, 23, 25, 28, 32}	{1, 3, 5, 7, 8, 9, 10, 13, 14, 15, 18, 19, 22, 23, 25, 28, 32}
5	Soybean (small)	{1, 2, 5, 7, 9, 10, 11, 13, 15, 16, 18, 19, 22, 25, 29, 30, 31, 32, 34}	{1, 3, 5, 7, 9, 10, 11, 13, 14, 15, 16, 18, 19, 20, 22, 25, 29, 30, 31, 32, 34}
6	Ionosphere	{1, 2, 8, 10, 12, 15, 18, 22, 28, 32, 34}	{1, 2, 4, 8, 9, 12, 15, 18, 22, 23, 28, 32}

The results of Table 3 and Table 4 show that the number of attributes of the reduct obtained by F_DBAR are smaller than that of the reduct obtained by GRAF (except Fertility). Furthermore, the executed

time of F_DBAR is less than that of GRAF. So F_DBAR is more effectively than GRAF in term of the executed time.

Next, we carry out some experiments to compare classification accuracy of the reduct obtained by F_DBAR and GRAF. The classification accuracy is conducted on two reducts of two algorithms with algorithm C4.5 in Weka [20] and 10-fold cross-validation. Specifically, given data set is randomly divided into ten parts of equal size. The nine parts of these ten parts are used to conduct as the training set and the rest part was taken as the testing set. Experimental results are shown in Table 5.

Table 5. A comparison of F_DBAR and GRAF[5] on classification accuracy

No	Data set	$ U $	$ C $	F_DBAR		GRAF[5]	
				$ R $	Accuracy	$ R $	Accuracy
1	Ecoli	336	7	6	0.802	6	0.802
2	Fertility	100	9	8	0.817	7	0.752
3	Wdbc	569	30	15	0.984	17	0.917
4	Wpbc	198	33	16	0.902	17	0.804
5	Soybean (small)	47	35	19	0.802	21	0.705
6	Ionosphere	351	34	11	0.942	12	0.904
Average					0.875		0.814

The results of Table 5 show that the average accuracy of F_DBAR is higher than that of GRAF on 6 data sets. That is F_DBAR is more effectively than GRAF on classification accuracy.

Consequently, experimental results on 6 data sets show that F_DBAR is more effectively than GRAF on the executed time and classification accuracy. That is the main result of this paper.

VI. CONCLUSION

Fuzzy rough set model proposed by Dubois D., and Prade H., [3, 4] is an effective approach to solve the issue of the attribute reduction on the decision table with numerical attribute value. In this paper, based on fuzzy distance we proposed an attribute reduction method on the decision table with numerical attribute value. The fuzzy distance measure is determined based on the equivalence relation matrix of attributes. The fuzzy equivalence relation matrix on

the value of attributes is determined by formula (3), the fuzzy equivalence matrix of attribute set is determined by formula (2). The experimental results on 6 data sets from UCI [19] show that the executed time of proposed algorithm F_DBAR is less than that of algorithm GRAF [5] and the classification accuracy of the reduct obtained by F_DBAR is higher than that of the reduct obtained by GRAF [5]. Our further research is to find the relation between reducts obtained by different methods according to fuzzy rough set approach.

ACKNOWLEDGEMENTS

This research has been funded by the Research Project, VAST 01.08/16-17. Vietnam Academy of Science and Technology.

REFERENCES

- [1] CHEN D. G., LEI Z., SUYUN Z., QING H. H. and PENG F. Z., *A Novel Algorithm for Finding Reducts With Fuzzy Rough Sets*, IEEE Transaction on Fuzzy Systems, Vol. 20, No. 2, 2012, pp. 385-389.
- [2] CHENG Y., *Forward approximation and backward approximation in fuzzy rough sets*, Neurocomputing, Volume 148, 2015, pp. 340-353.
- [3] DUBOIS D., PRADE H., *Putting rough sets and fuzzy sets together*, Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, 1992.
- [4] DUBOIS D., PRADE H., *Rough fuzzy sets and fuzzy rough sets*, International Journal of General Systems, 17, 1990, pp. 191-209.
- [5] DAI J. H., XU Q., *Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification*, Applied Soft Computing 13, 2013, pp. 211-221.
- [6] HE Q., WU C. X., CHEN D. G., ZHAO S. Y., *Fuzzy rough set based attribute reduction for information systems with fuzzy decisions*, Knowledge-Based Systems 24, 2011, pp. 689-696.
- [7] HU Q. H., YU D. R., XIE Z. X., *Information-preserving hybrid data reduction based on fuzzy-rough techniques*, Pattern Recognition Letters 27, 2006, pp. 414-423.
- [8] HU Q. H., YU D. R., *Fuzzy Probability Approximation Space and Its Information Measures*, IEEE Transaction on Fuzzy Systems, Vol 14, 2006.
- [9] JENSEN R., SHEN Q., *Fuzzy-Rough Sets for Descriptive Dimensionality Reduction*, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE'02, 2002, pp. 29-34.
- [10] JENSEN R., SHEN Q., *Fuzzy-rough attribute reduction with application to web categorization*, Fuzzy Sets and Systems, Volume 141, Issue 3, 2004, pp. 469-485.
- [11] NGUYEN LONG GIANG, *Rough Set Based Data Mining Methods*, Doctor of Thesis, Institute of Information Technology, 2012.
- [12] PAWLAK Z., *Rough sets*, International Journal of Computer and Information Sciences, 11(5), 1982, pp. 341-356.
- [13] 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, 2009, pp. 174-188.
- [14] QIAN Y. H., LIANG J. Y., WEI Z., WU Z., DANG C. Y., *Information Granularity in Fuzzy Binary GrC Model*, IEEE Transaction on Fuzzy Systems, Vol. 19, No. 2, 2011.
- [15] QIAN Y. H., LI Y. B., LIANG J. Y., LIN G. P., DANG C. Y., *Fuzzy granular structure distance*, IEEE Transactions on Fuzzy Systems, 23(6), 2015, pp. 2245-2259.
- [16] TSANG E.C.C., CHEN D. G., YEUNG D.S., XI Z. W., JOHN W. T. LEE, *Attributes Reduction Using Fuzzy Rough Sets*, IEEE Transactions on Fuzzy Systems, Volume 16, Issue 5, 2008, pp. 1130-1141.
- [17] XU F. F., MIAO D. Q., WEI L., *An Approach for Fuzzy-Rough Sets Attributes Reduction via Mutual Information*, Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD, 2007, Volume 3, pp. 107-112.
- [18] ZADEH L. A., *Fuzzy sets*, Information and Control, 8, 1965, pp. 338-353.
- [19] The UCI machine learning repository, <http://archive.ics.uci.edu/ml/datasets.html>
- [20] <https://sourceforge.net/projects/weka/>

AUTHOR'S BIOGRAPHIES

CAO CHINH NGHIA



He was born on 26/10/1977 in Ha Noi. Graduated from VNU University of Science in 1999. Received Master degree from VNU University of Engineering and Technology in 2006. Research interests include database, data mining and machine learning.

VU DUC THI



He was born on 07/04/1949 in Hai Duong. Graduated from VNU University of Science in 1971. Received the Ph.D degree from Hungary Academy of Sciences in 1987, specialized databases, Information Technology. Received the title of associate professor in 1991, received the title professor in 2009. Research interests include database, data mining and machine learning.

NGUYEN LONG GIANG



He was born on 05/06/1975 in Ha Tay. Graduated from Ha Noi University of Science and Technology in 1997. Received Master degree from VNU University of Engineering and Technology in 2003. Received the Ph.D degree in 2012 from Institute of Information Technology - Vietnamese Academy of Science and Technology (VAST). Research interests include database, data mining and machine learning.

TAN HANH



He was born on 10/01/1964 in Phnom Penh, Cambodia. Graduated from Ho Chi Minh City Pedagogical University in 1987. Received Master degree from VNU University of Science, Vietnam National University Ho Chi Minh City in 2002. Received the Ph.D degree from Grenoble Institute of Technology, France, in 2009, specialized distributed systems, Information Technology. Research interests include databases, Information retrieval, and distributed systems.