ATTRIBUTE REDUCTION OF NUMERICAL DECISION TABLE BASED ON FUZZY DISTANCE USING HEURISTIC ALGORITHM

Lai Kien Cuong, Nguyen Trung Hieu, Quach Nhu The, Cao Chinh Nghia
People's Police Academy

Abstract: *Attribute reduction of the decision table is the process of selecting a subset of the conditional attribute set that preserves the degree of information measurement. Attribute reduction of numerical decision table according to fuzzy rough set approach has been studied extensively in recent years. Attribute reduction methods according to the fuzzy rough set approach inherit traditional rough set methods with fuzzy equivalent relations replacing the equivalent relations. The value of fuzzy similarity is in the range [0, 1] which shows the close or similar properties of two objects. The fuzzy equivalent relations process directly on numeric value domain without through steps of discrete data. The original method is a fuzzy positive region. Researchers have developed methods using fuzzy entropy and fuzzy distance to improve the quality of the classification accuracy and reduce the execution time of the algorithm. This paper proposed a fuzzy distance and constructed a heuristic algorithm to find one reduction set of numerical decision tables which is called the reduct. The proposed method preserves information measurement of the conditional attribute set. Experiments on datasets taken from the UCI repository show that the proposed method improves the quality of classification accuracy and execution time of the algorithm compared with the methods using fuzzy positive region and fuzzy entropy on most experimental datasets.*

Keywords: *Fuzzy distance, fuzzy rough set, heuristic, numerical decision table, reduce.*

Nhận bài ngày 25.8.2024; gửi phản biện, chỉnh sửa, duyệt đăng ngày 25.10.2024
Liên hệ tác giả: Cao Chính Nghĩa; email: ccnghia@gmail.com

1. INTRODUCTION

Attribute reduction is an important problem in the data preprocessing step which goal reduces the number of attributes by eliminating redundant data to improve the efficiency of data mining and machine learning algorithms. Attribute reduction of the decision table is the process of selecting a subset of the conditional attribute set that preserves the decision table's hierarchical information, called reduce. It's difficult to find the best reduction for the problem related to the class of NP-Hard problems. The results of attribute reduction directly affect the efficiency of implementing mining tasks: increased speed of excution and improved the quality of classification accuracy.

Dubois, D et al (1990) 1 proposed fuzzy rough set theory as a combination of rough set theory (Pawlak et al (1982) 2) and fuzzy set theory (Zadeh et al (1965) 3) to approximate fuzzy sets based on fuzzy equivalence relation which is determined by the attribute value domain.

Attribute reduction of the numerical decision table based on fuzzy rough set has good results in recently years.

Fuzzy rough set theory uses the fuzzy equivalence relation to replace the classical equivalence relation. The value similarity is in the range [0, 1] shows the close or similar properties of two objects. Attribute reduction methods based on fuzzy rough set approach have the potential to preserve the classification accuracy and not transform the value of attribute in the data table. The filter-type approach performs attribute selection independent of the mining algorithm that is used later. Attributes are selected based on their importance. The heuristic of the algorithm finds one reduct by following these steps: Define the reduct, define the importance of the attribute and then construct a heuristic algorithm to find one reduct. Common results of the filter-type approach are:

The method uses fuzzy positive region (Pawlak et al (1982) 2, Hu et al (2006) 4): considered the original method, derived from rough set theory to find one reduct of decision tables with discrete value domains. Later, Dubois, D. et al (1990) 1 developed the fuzzy rough set theory to reduce properties of the numeric value domain decision table.

The method uses fuzzy entropy (Dai et al (2013) 5, Xie et al (2023) 6, Li et al (2023) 7): entropy is an important measure in Shanon's theory of information. Based on entropy construct measurements that evaluate the dependency of the attribute reduction of the decision table.

The method uses fuzzy distance (Giang et al (2012) 8, Qian et al (2015) 9, Tan et al (2022) 10): has developed from the fuzzy entropy method which has similarities to the fuzzy entropy method. Although the technique distance was published later but it has played an important role in data mining. Especially with attribute reduction according to the fuzzy rough set theory approach has obtained many good results. In Vietnam, the distance method has been studied and achieved certain results by Giang N.L et al.

Attribute reduction of numerical decision table based on fuzzy rough set to preserve the measurement is diagram as follows by Xu et al (2007) 11 and Zhong Yuan et al (2021) 12.

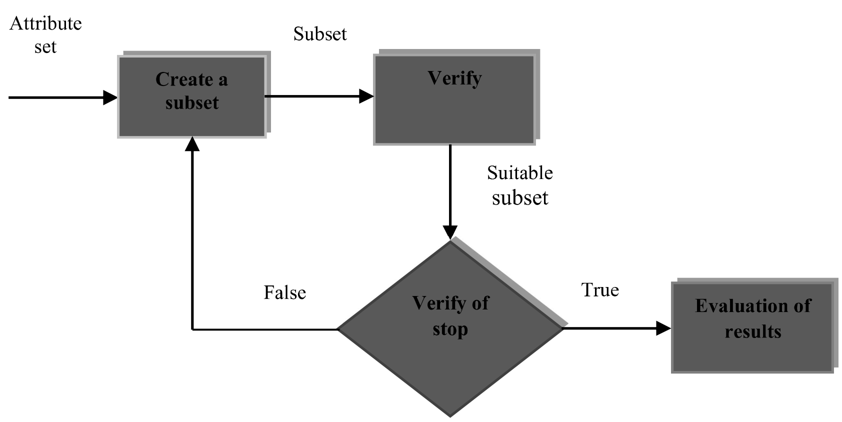

Figure 1: The diagram of the heuristic algorithm finds one reduction

The criteria of comparing and evaluating among methods is execution time of the algorithm and the classification accuracy of reduct.

The structure of this paper is as follows. Part 1 is the introduction. Part 2 gives some basic concepts of fuzzy rough set theory, presents a method of attribute reduction using fuzzy distance according to the heuristic algorithm and evaluates the experimental results. Part 3 is some conclusions and future development trends.

2. CONTENT

2.1. Definitions

Definition 2.1. The decision table with numerical attributes

Let decision table $DT = (U, C \cup D)$, if the value domain of all attributes $c \in C$ is numerical attributes then the decision table DT is a numerical decision table.

Example 2.1. Let a numerical decision table $DT = (U, C \cup D)$ *(Table 1)* with $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$,$C = \{c_1, c_2, c_3, c_4, c_5, c_6\}$.

Table 1: The decision table with numerical attributes

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

Definition 2.2. The fuzzy equivalence relation

Let decision table $DT = (U, C \cup D)$ with numerical attributes, the relation $\tilde{R}$ defined on value attribute domain is called fuzzy similarity relation if it satisfies the following conditions $x, y, z \in U$

1) Reflexive: $\tilde{R}(x, x) = 1$;

2) Symetric: $\tilde{R}(x, y) = \tilde{R}(y, x)$;

3) Max-min transitive: $\tilde{R}(x, z) \geq \min\{\tilde{R}(x, y), \tilde{R}(y, z)\}$;

Let two fuzzy equivalence relation $\tilde{R}_P$ and R_Q which define on attribute set P and Q, for any $x, y \in U$ we have (Qian et al (2015) 9):

1) $\tilde{R}_P = \tilde{R}_Q \Leftrightarrow \tilde{R}_P(x, y) = \tilde{R}_Q(x, y)$; (1)

2) $\tilde{R}_{P \cap Q} = \tilde{R}_P \cup \tilde{R}_Q \Leftrightarrow \tilde{R}(x, y) = \max\{\tilde{R}_P(x, y), \tilde{R}_Q(x, y)\}$; (2)

3) $\tilde{R}_{P \cup Q} = \tilde{R}_P \cap \tilde{R}_Q \Leftrightarrow \tilde{R}(x, y) = \min\{\tilde{R}_P(x, y), \tilde{R}_Q(x, y)\}$; (3)

4) $\tilde{R}_P \subseteq \tilde{R}_Q \Leftrightarrow \tilde{R}_P(x, y) \leq \tilde{R}_Q(x, y)$. (4)

Definition 2.3. The fuzzy equivalence matrix

Let decision table $DT = (U, C \cup D)$ with all $c \in C$ are numerical attributes and $U = \{x_1, x_2, \ldots, x_n\}$ and $\tilde{R}_P$ is the fuzzy relation which defined on $P \subseteq C$. The relation $\tilde{R}_P$ denoted by $M(\tilde{R}_P) = [p_{ij}]_{n\times n}$ is defined as Qian et al (2015) 9:

$$M(\tilde{R}_P) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \quad (5)$$

where $p_{ij} = \tilde{R}_P(x_i, x_j)$ is fuzzy relation value of x_i and x_j on P, $p_{ij} \in [0,1]$, $x_i, x_j \in U$, $1 \le i, j \le n$.

In this paper, we use fuzzy equivalence according to formula (6) to construct fuzzy equivalent matrices directly from the attributes of numerical decision tables.

$$p_{ij} = \begin{cases} 1 - 4 * \dfrac{|p(x_i) - p(x_j)|}{|p_{\max} - p_{\min}|}, & \text{if } \dfrac{|p(x_i) - p(x_j)|}{|p_{\max} - p_{\min}|} \le 0.25 \\ 0, & otherwise \end{cases} \quad (6)$$

where $p(x_i)$ is the value of attribute p at object x_i; $p_{\max}$, $p_{\min}$ is maximum value and minimum value of attribute p. It is easy to see that the values of the elements of fuzzy equivalent matrices belong to the [0,1], if $p_{\max} = p_{\min}$ then $p_{ij} = 1$. Then the fuzzy equivalent relation in formulas (6) and (7) are the same.

$$p_{ij} = 1 \text{ if } x_j \in [x_i]_P \text{ and } p_{ij} = 0 \text{ if } x_j \notin [x_i]_P \quad (7)$$

In other word, the equivalent class $[x_i]_P$ can be considered a fuzzy equivalent class, denoted as $[x_i]_{\tilde{P}}$, with the function of $\mu_{[x_i]_{\tilde{P}}}(x_j) = 1$ if $x_j \in [x_i]_{\tilde{P}}$ and b $\mu_{[x_i]_{\tilde{P}}}(x_j) = 0$ if $x_j \notin [x_i]_{\tilde{P}}$.

Definition 2.4. The fuzzy partition and the fuzzy equivalent class

Let a decision table with numerical attribute $DT = (U, C \cup D)$ and $P, Q \subseteq C$. According to Qian et al (2015) 9 we have $R_P = \cap_{a\in P} R_a$ and $\tilde{R}_{P\cup Q} = \tilde{R}_P \cap \tilde{R}_Q$, which means that for any $x, y \in U, \tilde{R}_{P\cup Q}(x,y) = \min\{\tilde{R}_P(x,y), \tilde{R}_Q(x,y)\}$. Suppose that $M(\tilde{R}_P) = [p_{ij}]_{n\times n}$ and $M(\tilde{R}_Q) = [q_{ij}]_{n\times n}$ are fuzzy equivalent matrices of $\tilde{R}_P$, $\tilde{R}_Q$ corresponding, then the relational matrix on the attribute sets $S = P \cup Q$ is defined as:

$$M(\tilde{R}_S) = M(\tilde{R}_{P\cup Q}) = [s_{ij}]_{n\times n} \, with \, s_{ij} = \min\{p_{ij}, q_{ij}\} \quad (8)$$

For $P \subseteq C$, $U = \{x_1, x_2, \ldots, x_n\}$, fuzzy equivalent relation $\tilde{R}_P$ can be generated from the fuzzy equivalence relation $\pi(P) = U/\tilde{R}_P$ on U

$$\pi(\tilde{R}_P) = U/\tilde{R}_P = \{[x_i]_{\tilde{R}_P}\}_{i=1}^{n} = \{[x_1]_{\tilde{R}_P}, \ldots, [x_n]_{\tilde{R}_P}\} \quad (9)$$

Where $[x_i]_{\tilde{R}_P} = p_{i1}/x_1 + p_{i2}/x_2 + \ldots + p_{in}/x_n$ is a fuzzy set, is called a fuzzy equivalence of object x_i. The degree functions of the objects are determined by:

$$\mu_{[x_i]_{\tilde{R}_P}}(x_j) = \mu_{\tilde{R}_P}(x_i, x_j) = \tilde{R}_P(x_i, x_j) = p_{ij} \text{ for any } x_j \in U. \quad (10)$$

Then, the cardinality of fuzzy equivalence $[x_i]_{\tilde{R}_P}$ is calculated as Qian et al (2012) 9:

$$\left|[x_i]_{\tilde{R}_P}\right| = \sum_{j=1}^{n} p_{ij} \tag{11}$$

2.2. A fuzzy distance based on attribute reduction method in decision tables with numerical attributes using heuristic algorithm.

In this part, we introduce a fuzzy distance-based attribute reduction method that performs directly on the decision tables with numerical attributes that preserve the information measurement (Cao Chinh Nghia et al (2016) 13).

Definition 2.5. Let a decision table with numerical attributes $DT = (U, C \cup D)$ with $\pi(\tilde{R}_P)$, $\pi(\tilde{R}_Q)$ are two fuzzy partitions which construct by fuzzy equivalent relation $\tilde{R}_P$, $\tilde{R}_Q$ on P, $Q \subseteq C$. Then, the fuzzy distance between two attribute sets P and Q, denoted as $d_{NF}(P,Q)$, is defined between two fuzzy partitions $\pi(\tilde{R}_P)$ and $\pi(\tilde{R}_Q)$, it means that $d_{NF}(P,Q) = D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right)$.

Theorem 2.1. Let $DT = (U, C \cup D)$ is a numerical decision table where $U = \{x_1, x_2, \ldots, x_n\}$ and $\pi(\tilde{R}_P)$, $\pi(\tilde{R}_Q)$ are two fuzzy partitions which generated two fuzzy equivalent relations $\tilde{R}_P$, $\tilde{R}_Q$ on $P, Q \subseteq C$. Then:

$$D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{\tilde{R}_P}\right| + \left|[x_i]_{\tilde{R}_Q}\right| - 2\left|[x_i]_{\tilde{R}_P} \cap [x_i]_{\tilde{R}_Q}\right|}{n}\right) \tag{12}$$

is fuzzy distance between $\pi(\tilde{R}_P)$ and $\pi(\tilde{R}_Q)$.

Proof: We have $D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right) \geq 0$ and $D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right) = D_{NF}\left(\pi(\tilde{R}_Q), \pi(\tilde{R}_P)\right)$. We need to prove triangular inequality. No loss of generality, with all $\pi(\tilde{R}_P), \pi(\tilde{R}_Q), \pi(\tilde{R}_S) \in \mathcal{P}$ we have proved $D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right) + D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_S)\right) \geq D_{NF}\left(\pi(\tilde{R}_Q), \pi(\tilde{R}_S)\right)$. For any $x_i \in U$ we have: $d_{NF}\left([x_i]_{\tilde{R}_P}, [x_i]_{\tilde{R}_Q}\right) + d_{NF}\left([x_i]_{\tilde{R}_P}, [x_i]_{\tilde{R}_S}\right) \geq d_{NF}\left([x_i]_{\tilde{R}_Q}, [x_i]_{\tilde{R}_S}\right)$. Therefore:

$$D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right) + D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_S)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d_{NF}\left([x_i]_{\tilde{R}_P}, [x_i]_{\tilde{R}_Q}\right)}{n} + \frac{1}{n}\sum_{i=1}^{n}\frac{d_{NF}\left([x_i]_{\tilde{R}_P}, [x_i]_{\tilde{R}_S}\right)}{n} \geq \frac{1}{n}\sum_{i=1}^{n}\frac{d_{NF}\left([x_i]_{\tilde{R}_Q}, [x_i]_{\tilde{R}_S}\right)}{n}$$

$$= D_{NF}\left(\pi(R_Q), \pi(R_S)\right)$$

It easy to see that $0 \leq D_{NF}\left(\pi(\tilde{R}_P), \pi(\tilde{R}_Q)\right) \leq 1$.

Proposition 2.1. Let $DT = (U, C \cup D)$ is a numerical decision table $U = \{x_1, x_2, \ldots, x_n\}$ and $\tilde{R}$ is a fuzzy equivalence relation determined on conditional attributes. Then, the fuzzy distance between two attribute sets C *and* $C \cup D$ which is determined as:

$$\boldsymbol{d_{NF}(C, C \cup D) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{\tilde{R}_C}\right| - \left|[x_i]_{\tilde{R}_C} \cap [x_i]_{\tilde{D}}\right|}{n}\right)} \quad (13)$$

Proof: We have:

$$d_{NF}(C, C \cup D) = D_{NF}\left(\pi(\tilde{R}_C), \pi(\tilde{R}_{C\cup D})\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{\tilde{R}_C}\right| + \left|[x_i]_{\tilde{R}_{C\cup D}}\right| - 2\left|[x_i]_{\tilde{R}_C} \cap [x_i]_{\tilde{R}_{C\cup D}}\right|}{n}\right) =$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{\tilde{R}_C}\right| + \left|[x_i]_{\tilde{R}_C} \cap [x_i]_{\tilde{R}_D}\right| - 2\left|[x_i]_{\tilde{R}_C} \cap [x_i]_{\tilde{R}_D}\right|}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{\tilde{R}_C}\right| - \left|[x_i]_{\tilde{R}_C} \cap [x_i]_{\tilde{R}_D}\right|}{n}\right) =$$
$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{\tilde{R}_C}\right| - \left|[x_i]_{\tilde{R}_C} \cap [x_i]_{\tilde{D}}\right|}{n}\right)$$

It is easy to see that $0 \le d_{NF}(C, C \cup D) \le 1 - \frac{1}{n}$.

In the next part, we present a heuristic method to find one reduct of the decision table with numerical attributes using the fuzzy distance. Our method includes: defining the reduct based on fuzzy distance, defining the importance of the attribute and designing a heuristic algorithm to find one reduct based on the importance of the attribute.

Definition 2.6. Let $DT = (U, C \cup D)$ is a numerical decision table where $B \subseteq C$ and $\tilde{R}$ is a fuzzy equivalent relation which determined by the value of the conditional set. If:

1) $\boldsymbol{d_{NF}(B, B \cup D) = d_{NF}(C, C \cup D)}$ (14)

2) $\boldsymbol{\forall b \in B,\ d_{NF}(\{B - \{b\}\}, \{B - \{b\}\} \cup D)) \ne d_{NF}(C, C \cup D)}$ (15)

Then B is a reduct of C based on fuzzy partition. Formula (14) ensures that the preservation of the measurement of the information of the reduced attribute set and the condition attribute set compared to the decisive attribute set is the same.

Definition 2.7. Let $DT = (U, C \cup D)$ is a numerical decision table with $B \subset C$ and $b \in C - B$. The importance of attribute b with respect to B is defined as

$$\boldsymbol{SIG_B(b) = d_{NF}(B, B \cup D) - d_{NF}(B \cup \{b\}, B \cup \{b\} \cup D)} \quad (16)$$

The importance of $SIG_B(b)$ characterizes the classification accuracy of attribute *b* with respect to decision attribute *D*. It is used as the attribute selection criterial for heuristic algorithms to find the one reduct.

The diagraming of the algorithm constructing steps is as follows

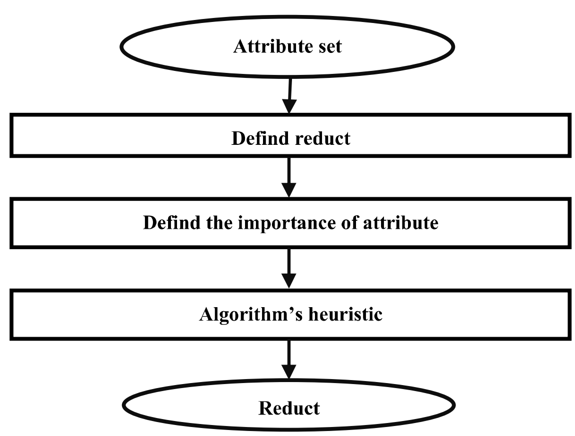

Figure 2: Steps of heurictic's method finds one reduct

The algorithm's heurictis finds one reduct of the numerical decision table as follows

Algorithm NF_DBAR: A heuristic algorithm finds one reduct by using fuzzy distance.
Input: Decision table with numerical attributes $DT = (U, C \cup D)$, fuzzy equivalence relation $\tilde{R}$
Output: The Reduct B
1. $B \leftarrow \emptyset; M(\tilde{R}_B) = [1]_{n\times n};$
2. Calculate the relation matrix $M(\tilde{R}_C)$, calculate equivalence matrix $M(IND(D))$, calculate fuzzy distance $d_{NF}(C, C \cup D)$;
3. While $d_{NF}(B, B \cup D) \neq d_{NF}(C, C \cup D)$ Do
4. Begin
5. For each $a \in C - B$ Do
$SIG_B(a) = d_{NF}(B, B \cup D) - d_{NF}(B \cup \{a\}, B \cup \{a\} \cup D)$;
6. Select $a_m \in C - B$ That $SIG_B(a_m) = \underset{a \in C-B}{Max}\{SIG_B(a)\}$;
7. $B = B \cup \{a_m\}$
8. End;
For each $a \in B$
9. Begin
10. Calculate $d_{NF}(B - \{a\}, B - \{a\} \cup D)$;
11. If $d_{NF}(B - \{a\}, B - \{a\} \cup D) = d_{NF}(C, C \cup D)$ then $B = B - \{a\}$;
11. End;
12. Return B;

By using NF_DBAR algorithm to find one reduct of table 1, we have reduct $B = \{c_4, c_1\}$.

2.3. Experiments

We have selected six sample datasets taken from the UCI data repository (2024) 14 with numerical value domains to conduct the test. The test environment is a PC with a Pentium core i5 2.4 GHz CPU, 8 GB of RAM, using the Windows 10 operating system.

We have chosen algorithms FA_FPR (Hu et al (2007) 15) (find one reduct based on the fuzzy positive domain) and GAIN_RATIO_AS_FRS (Dai et al (2013) 5) (called GRAF) (find one reduct based on the fuzzy entropy) to compare with the algorithm proposed algorithm on the classification accuracy of reduct and execution time. For testing, we perform the following tasks:

1) Code FA_FPR, GRAF and NF_DBAR algorithms by program C#, Algorithms used the fuzzy equivalence relation defined by the formula (6).

2) Execute three algorithms on six datasets by environment testing.

3) Use C4.5 algorithm in WEKA (Eibe Frank et al (2016) 16) tool to evaluate the classification accuracy of three algorithms by selecting 2/3 first objects as the training set and the remainder of the objects as the testing set.

Table 2 shows the testing results of six datasets where |U| is the number of objects, $|C|$ is the number of the conditional attribute, $|R|$ is the number of attributes of the reduct for each algorithm, t is the average time after three performance of the attribute reduction of algorithms (in seconds).

Table 2: The experimental result of three algorithms FA_FPR, GRAF and NF_DBAR

Id	Datasets	$\lvert C\rvert$	FA_FPR		GRAF		NF_DBAR	
			$\lvert R\rvert$	t	$\lvert R\rvert$	t	$\lvert R\rvert$	t
1	Fisher_Order	35	21	0.193	21	0.107	18	**0.079**

2	Iris	4	2	0.003	2	0.003	1	**0.002**
3	Glass	10	7	0.036	8	0.028	7	**0.024**
4	Sonar	60	12	2.889	23	**1.980**	13	2.433
5	Sensor_Readings_ 24	24	15	2.465	12	**1.986**	14	2.005
6	EEG_Eye_State	14	7	4.069	7	3.790	7	**3.046**

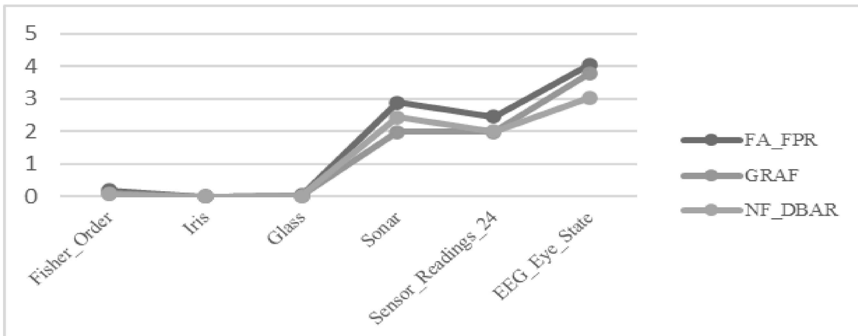

Figure 3: The executed time of FA_FPR, GRAf and NF_DBAR

Table 3 represents the number of attributes of reduct and the classification accuracy of three algorithms GRAF, FA_FPR and NF_DBAR on six experimental datasets.

Table 3: The classification accuracy C4.5 of FA_FPR, GRAF and NF_DBAR

Id	Datasets	$\|U\|$	$\|C\|$	FA_FPR		GRAF		NF_DBAR	
				$\|R\|$	Classification accuracy of ***C4.5*** *(%)*	$\|R\|$	Classification accuracy of ***C4.5*** *(%)*	$\|R\|$	Classification accuracy of ***C4.5*** *(%)*
1	Fisher_Order	47	35	21	76.60	21	76.60	18	**78.72**
2	Iris	150	4	2	94.00	2	94.00	1	**94.67**
3	Glass	214	10	7	81.56	8	**81.70**	7	81.56
4	Sonar	208	60	12	70.60	23	71.67	13	**76.25**
5	Sensor_Readings_24	5456	24	15	**95.12**	12	92.25	14	94.84
6	EEG_Eye_State	14980	14	7	81.25	7	81.25	7	81.25

Figure 4 is a chart comparing the classification accuracy by the C4.5 algorithm in WEKA's J48 tool (Eibe Frank et al (2016) 16) of three algorithms on six datasets.

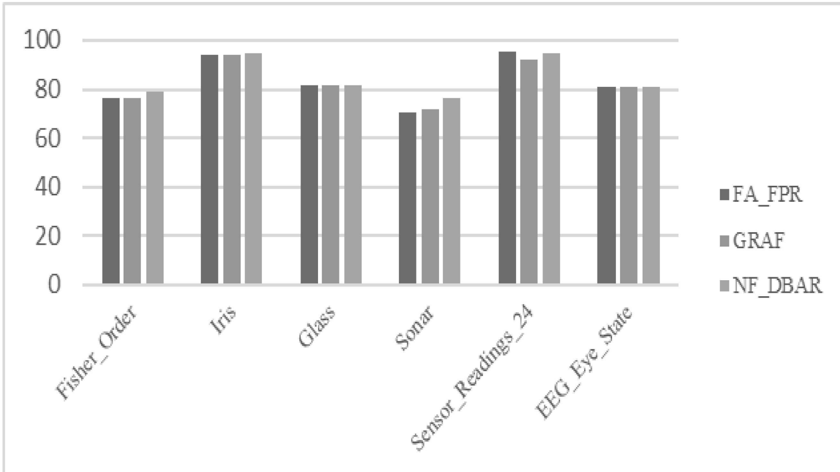

Figure 4: The classification accuracy C4.5 of FA_FPR, GRAF and NF_DBAR

The experimental results show that the method using fuzzy distance gives a better rate of data classification accuracy at 3/6 datasets, the method using fuzzy entropy and the method using fuzzy positive region have the same classification accuracy at 1/6 datasets and the three methods have the same classification accuracy at 1/6 datasets. The classification accuracy depends on special reduct according to each method. The execution time of three methods was averaged after three test runs with each dataset. Specifically, the method of using fuzzy distance is the fastest at 4/6 datasets, followed by the method of using fuzzy entropy is at 2/6 datasets and the method of using fuzzy positive region is at 0/6 datasets. If the methods for the same reduct with a particular dataset then the method using fuzzy distance has the shortest execution time, followed by the method using fuzzy entropy and finally the method using fuzzy positive region. This is in line with the theory because the computational complexity of methods in the general case are the same but the fuzzy positive region has the most complex calculation formula by the combined operations of the lower approximations. Methods of finding reduct using fuzzy entropy and fuzzy distance have simpler calculation formulas. However, the method of using fuzzy distance does not use logarithm functions such as the fuzzy entropy method, so the calculation time is minimized. Experimental results on six datasets show that the proposed method is more effective than methods using fuzzy entropy and fuzzy positive region according to the evaluation criteria of *execution time* and *classification accuracy* on some experimental datasets.

3. CONCLUSION

This paper proposed a distance measurement between two fuzzy partitions and constructed a heuristic algorithm to find one reduct of the decision table with numerical attributes. Experiments on six datasets taken from the UCI repository show that the proposed method improves the quality of classification accuracy and the execution time of algorithm compared with methods using fuzzy positive region and fuzzy entropy on most experimental datasets. The effectiveness of methods on each dataset also depends on the data distribution of each experimental dataset. The scientific contribution of this paper provides an effective attribute reduction method of numerical decision table according to the fuzzy rough set approach based on the fuzzy distance. In the future, we will focus on researching methods of attribute reduction of the numeric value domain decision table in the trend of filter/wrapper hybrid to improve the efficiency of data mining

REFERENCES

1. Dubois, D., and Prade, H. (1990), Rough fuzzy sets and fuzzy rough sets, *International Journal of General System*, *17*(2-3), pp. 191-209.
2. Pawlak, Z. (1982), Rough sets, *International Journal of Computer & Information Sciences*, *11*(5), pp. 341-356.
3. Zadeh, L. A. (1965), Fuzzy sets, *Information and control*, *8*(3), pp. 338-353.
4. Hu, Q., Yu, D., and Xie, Z. (2006), Information-preserving hybrid data reduction based on fuzzy-rough techniques, *Pattern recognition letters*, *27*(5), pp. 414-423.
5. Dai, J., and Xu, Q. (2013), Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, *Applied Soft Computing*, *13*(1), pp. 211-221.
6. L. Xie, G. Lin, J. Li, and Y. Lin, (2023), "A novel fuzzy-rough attribute reduction approach via local information entropy," *Fuzzy Sets Syst.,* Vol. 473.
7. Z. Li and Y. Chen, (2023), "Feature selection in a discrete feature space based on fuzzy conditional information entropy iterative model and matrix operation," *Int. J. Gen. Syst*., Vol. 52, No. 5.
8. Nguyen, L. G. (2012, September), Metric based attribute reduction in decision tables, In *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on* (pp. 311-316), IEEE.
9. Qian, Y., Wang, Q., Cheng, H., Liang, J., and Dang, C. (2015), Fuzzy-rough feature selection accelerator, *Fuzzy Sets and Systems*, *258*, pp. 61-78.
10. Tan, S. Shi, W. Z. Wu, J. Li, and W. Pedrycz, (2022), "Granularity and Entropy of Intuitionistic Fuzzy Information and Their Applications", *IEEE Trans. Cybern*., Vol. 52, no. 1, pp. 192-20.
11. Xu, F., Miao, D., and Wei, L. (2007, August), An Approach for Fuzzy-Rough Sets Attributes Reduction via Mutual Information, *In FSKD (3)*, pp. 107-112.
12. Zhong Yuan, Hongmei Chen, Peng Xie, Pengfei Zhang, Jia Liu, Tianrui Li (2021), Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions, *Applied Soft Computing*, Vol 107.
13. Cao Chinh Nghia, Vu Duc Thi, Nguyen Long Giang, Tan Hanh (2016), "Fuzzy distance based attribute reduction in decision tables", *Research, Development and Application on Information and Communication Technology*, Vol 16 (36), Vietnam Ministry of Information and Communications, pp. 104-112
14. Datasets - UCI Machine Learning Repository, http://archive.ics.uci.edu/datasets.html
15. Hu, Q., Xie, Z., and Yu, D. (2007), Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, *Pattern recognition*, *40*(12), pp. 3509-3521.
16. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016), The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016, https://sourceforge.net/projects/weka

PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH CỦA BẢNG QUYẾT ĐỊNH MIỀN GIÁ TRỊ SỐ DỰA TRÊN KHOẢNG CÁCH MỜ SỬ DỤNG THUẬT TOÁN HEURISTIC

Tóm tắt: *Rút gọn thuộc tính của bảng quyết định là quá trình lựa chọn một tập con của thuộc tính điều kiện mà bảo toàn được thông tin phân lớp của bảng quyết định. Rút gọn thuộc tính của bảng quyết định miền giá trị số theo tiếp cận tập thô mờ thu hút được nhiều nhà nghiên cứu quan tâm trong giai đoạn hiện nay. Các phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ là sự kế thừa của tập thô truyền thống với quan hệ tương đương mờ thay thế cho quan hệ tương đương. Độ tương đương mờ của hai đối tượng là một giá trị nằm trong đoạn [0,1] cho thấy tính gần nhau, hay khả năng phân biệt giữa hai đối tượng. Quan hệ tương đương mờ giúp xử lý trực tiếp trên miền dữ liệu giá trị số mà không cần thông qua bước rời rạc hóa dữ liệu. Xuất phát từ phương pháp sử dụng miền dương mờ, nhiều phương pháp mới ra đời nhằm cải thiện chất lượng phân lớp dữ liệu và giảm* *thời gian thực hiện thuật toán. Bài báo này đề xuất một độ đo khoảng cách mờ và xây dựng phương pháp heuristic rút gọn thuộc tính của bảng quyết định miền giá trị số theo tiếp cận tập thô mờ bảo toàn được độ đo thông tin của tập thuộc tính điều kiện. Thực nghiệm trên một số bộ dữ liệu từ kho dữ liệu UCI cho thấy, phương pháp đề xuất hiệu quả hơn phương pháp sử dụng entropy mờ và phương pháp sử dụng miền dương mờ theo tiêu chí chất lượng phân lớp dữ liệu và thời gian thực hiện thuật toán.*

Từ khóa*: Bảng quyết định miền giá trị số, heuristic, khoảng cách mờ, tập thô mờ, tập rút gọn.*