Feature selection with fuzzy-rough minimum classification error criterion

Changzhong Wang, Yuhua Qian, Weiping Ding, Xiaodong Fan

***Abstract*—Classical fuzzy rough set often uses fuzzy rough dependency as an evaluation function of feature selection. However, this function only retains the maximum member- ship degree of a sample to one decision class, it can not describe the classification error. Therefore, in this work, a novel criterion function for feature selection is proposed to overcome this weakness. To characterize the classification error rate, we first introduce a class of irreflexive and symmetric fuzzy binary relations to redefine the concepts of fuzzy rough approximations. Then, we propose a novel concept of dependency: inner product dependency to describe the classification error, and construct a criterion function to evaluate the importance of candidate features. The proposed criterion function not only can maintain a maximum dependency function, but also guarantees the minimum classification error. The experimental analysis shows that the proposed criterion function is effective for data sets with a large overlap between different categories.**

***Index Terms*—Fuzzy rough set; Dependency function; Fuzzy inner product; Feature selection**

I. INTRODUCTION

owadays, data is growing in a large scale. It is always the case Nthat a data set has less samples and higher dimension. This brought a huge challenge to traditional classification learning algorithms. The existences of redundant features increase the negative effects on a machine learning task. It is a difficult problem for traditional learning algorithms to carry on the training of high dimensional data. How to eliminate redundant or irrelevant features from high dimensional data for avoiding the dimension disaster is an urgent problem to solve. Feature selection, or feature reduction is an effective technique to find a low-dimensional data set of interest from a high- dimensional one according to some evaluation criteria. Its main task is to acquire the characteristics of high-dimensional data through the analysis of low-dimensional data, and to simplify data analysis. At present, feature selection is one of the important research topics in pattern recognition and machine learning, and has been widely used in practice.

Selecting an appropriate evaluation criterion for feature selection is a critical step. It directly affects the performance of feature selection. There are two ways for construction of evaluation criterion functions: filter and wrapper. Wrapper approach uses a classifier to evaluate the selected feature subsets. However, not every classifier can be used as a criterion function to evaluate features. A classifier, which is suitable for wrapper algorithm, needs to be capable of dealing with features with high dimension, and can still get good classification result when the number of samples is limited. Filter approach selects features with an evaluation criterion that is independent of learning algorithms. To evaluate the merits of feature subsets, many feature evaluation criterions, such as consistency [1], correlation [2], mutual information [3], and Euclidean distance [4], have been developed for feature selection.

Rough set theory is a useful data preprocessing method that has been widely applied in reasoning with uncertainty [5]-[8], feature selection [9]-[14] and rule extraction [15]-[19]. It employs the dependency function for a feature evaluation function and can effectively reduce irrelevant or redundant features for classification tasks. However, the classical rough set models are only suitable for discrete features; they cannot be directly used to process real-valued data. Before feature selection, real-valued data must be discretized. This is a main limitation for the classical rough set models.

Fuzzy rough set model is one of the most important generalizations of classical model [20]-[23]. As it combines the advantages of rough and fuzzy sets, this model has been widely used to deal with the feature reduction of real-valued data [24]-[34]. Jensen and Shen first used fuzzy rough set theory to propose the concept of fuzzy rough dependency functions and designed a fast reduction algorithm to reduce redundant features [35]. Bhatt and Gopal presented a compact computing domain for the fast feature reduction algorithm, which effectively improved the computing efficiency of the algorithm [36]. Chen et al. defined the notion of fuzzy discernibility matrix using fuzzy rough lower approximations and employed it to calculate the feature reduction of a data set [37]. Dai et al. presented a

This work was supported by the National Natural Science Foundation of China under Grants 61976027, 61976120, 61572082, LiaoNing Revitalization Talents Program under Grant XLYC2008002, the Natural Science Foundation of Jiangsu Province under Grant BK20191445, Six Talent Peaks Project of Jiangsu Province under Grant XYDXXJS- 048, and sponsored by Qing Lan Project of Jiangsu Province.

C. Z. Wang, D.X. Fan are with the Department of Mathematics, Bohai University, Jinzhou, Liaoning 121000, P.R. China (e-mail: changzhong wang@126.com; bhdxfxd@163.com).

Y. H. Qian is with School of Computer and Information Technology, Shanxi University, Taiyuan 030006, P.R. China (e-mail: jinchengqyh@126.com).

W. P. Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

maximal discernibility pair-based approach to attribute reduction with fuzzy rough sets [38]. Mieszkowicz-rolka proposed a variable precision model of fuzzy rough sets for processing noise data [39]. Zhao et al. used the variable precision model to deal with misclassification and disturbance noise of data [40]. The common idea of these algorithms is to construct so-called fuzzy rough dependency for evaluating the classification ability of real valued feature subsets. A fuzzy rough dependency function was commonly formulated to the proportion of membership cardinality of fuzzy positive domain to the whole data set. However, the fuzzy positive domain can only retain the maximum membership degrees of samples to decision classes, but cannot keep the minimum classification error. Therefore, the criterion of fuzzy rough dependency determined by fuzzy positive domain cannot truly characterize the classification capability of candidate features. In this paper, we analyze the relationship between the classification error rate and membership functions of decision classes and point out that the overlap degree of membership functions of different classes is tightly linked to classification error. To characterize the classification error rate, we first introduce a class of irreflexive and symmetric fuzzy binary relations to redefine the rough approximations of a decision. Then, we propose a novel concept of dependency: inner product dependency to describe the classification error. Based on the concept of inner product dependency, we define a new feature selection criterion to determine the importance of feature subsets. The proposed feature selection criterion can keep the maximum dependency of each decision class and guarantees the minimum classification error rate.

The structure of this paper is as follows. Section 2 reviews classical fuzzy rough set model and analyze the weakness of the model. Section 3 reconstructs the fuzzy rough approximations of decision classes and then introduces the concept of inner product dependency. Section 4 designs a heuristic algorithm for feature selection using the proposed criterion. Section 5 verifies the effectiveness and feasibility of the proposed algorithm. Section 6 draws the conclusions and future works.

II. Classical Fuzzy Rough Set Model and Its Weakness

Let U be a given set and F be a mapping from U to the interval $[0,1]$, i.e., $F(\cdot):U\to[0,1]$, then F is referred to as a fuzzy set on U. For any $x\in U$, $F(x)$ is referred to as the membership degree of x to F.

Given a finite set of samples $U=\{x_1,x_2,\cdots,x_n\}$ and a set of real-valued features $A=\{a_1,a_2,\cdots,a_m\}$ describing the samples, these features can generate a fuzzy binary relation R_A on U [19]. R_A is then said to be a fuzzy similarity relation if it meets

(1) $R_A(x_i,x_i)=1$ for any $x_i\in U$;

(2) $R_A(x_i,x_j)=R_A(x_j,x_i)$ for any $x_i,x_j\in U$.

For any $x_i\in U$, the fuzzy similarity class associated with x and R_A is denoted by $[x_i]_A$ and defined as $[x_i]_A(x_j)=R_A(x_i,x_j)$, $x_j\in U$. Obviously, it is a fuzzy set on U. The similarity class $[x_i]_A$ is often called the fuzzy neighborhood of x_i. If a fuzzy similarity relation degenerates to a crisp one, the generated fuzzy neighborhoods degrade to crisp ones.

Suppose that D is a label feature that group the samples in U into r classes, that is, $U/D=\{D_1,D_2,\cdots D_r\}$. Then we call the triplet (U,A,D) a decision table.

Let $B\subseteq A$ and R_B be the fuzzy similarity relation on U related to B. For any $D_k\in U/D$, the fuzzy rough approximations of X_k are formulated as follows.

$$\underline{R}_B(D_k)(x_i)=\inf_{x_j\notin D_k}\{1-R_B(x_i,x_j)\} \tag{1}$$

$$\overline{R}_B(D_k)(x_i)=\sup_{x_j\in U}R_B(x_i,x_j) \tag{2}$$

for $x_i\in U$.

The fuzzy positive region of D upon B is given by

$$POS_B(D)(x_i)=\bigcup_{k=1}^{r}\underline{R}_B(D_k)(x_i),\ x_i\in U \tag{3}$$

It indicates that the sample x_i is assigned to a certain decision class by the degree of $POS_B(D)(x_i)$. Based on the concept of the fuzzy positive region, the fuzzy dependency function is expressed as

$$\gamma_B(D)=\frac{\sum_{x_i\in U}POS_B(D)(x_i)}{|U|} \tag{4}$$

Obviously, the dependency function can be explained as the proportion of the cardinal number of fuzzy positive region to the number of all the samples. It is commonly used for evaluating the importance of a feature subset in fuzzy rough set theory.

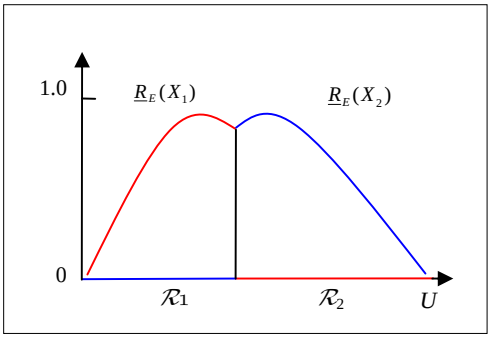

Fig. 1 The membership curves of two classes in a feature subspace E

However, the formulas (1) and (2) has the following weakness.

For any sample $x_j\in U$, if its class label is assigned to the class D_k, then $\underline{R}(D_l)(x_j)=0$ for any other class D_l ($l\neq k$) no matter whether the label is right or not. For example, Fig.1 indicates a binary classification problem in a feature subspace E, where $U/D=\{D_1,D_2\}$ and $\mathcal{R}_1$ and $\mathcal{R}_2$ are two decision regions. From the classical model of fuzzy rough set theory,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

$\underline{R}_E(D_1)(x_j)=0$ when $x_j \in \mathcal{R}_2$ and $\underline{R}_E(D_2)(x_j)=0$ when $x_j \in \mathcal{R}_1$. Hence, the overlap degree of the membership function curves of different decisions is zeros as shown in Fig.1. As we know, an overlap degree of the membership curves reflects the extent in which the samples are misclassified. Therefore, the classical model of fuzzy rough sets has no ability to reflect the classification error rate in its current formalism. The criterion of fuzzy positive region just considers the fuzzy dependency between decisions and features and omits a minimum classification error rate. Therefore, the fuzzy dependency function does not accurately characterize the classification information of feature subsets and it is easy to cause the error of sample classification in the overlapping region of samples with different labels.

III. INNER PRODUCT DEPENDENCY

As we know, the membership degree of a sample to one class is mainly determined by the similarity between itself and its neighbors, and has nothing to do with reflexivity. Hence, an irreflexive and symmetric relation is first reviewed to describe fuzzy information of data. Then the fuzzy rough approximations of a decision are reconstructed. To overcome the weakness that the fuzzy positive region can't guarantee the minimal classification error, we finally introduce a new feature evaluation function in this section.

Let $U=\{x_1,x_2,\cdots,x_n\}$ be a set of samples and A be a set of real-valued features. Again let B be a subset of A and R_B be a fuzzy relation on U related to B, then R_B is an irreflexive and symmetric relation if it satisfies

(1) $R_B(x,x)=0$ for any $x \in U$;

(2) $R_B(x_i,x_j)=R_B(x_j,x_i)$ for any $x_i,x_j \in U$.

For the sake of simplicity, we still call R_B a fuzzy similarity relation. This means that the fuzzy similarity relations in the next discussion are referred as to irreflexive and symmetric fuzzy relations.

Let B be a subset of A, $a \in B$, and R_a be an irreflexive and symmetric fuzzy relation (i.e., fuzzy similarity relation) related to a. We stipulate that $R_B=\bigcap_{a\in B} R_a$. In order to obtain the classification information of feature subsets under different granularity, a parameterized fuzzy similarity relation is constructed as follows.

Definition 1. Let $U=\{x_1,x_2,\cdots,x_n\}$ be a sample set, A be a set of features and $B \subseteq A$, a parameterized fuzzy similarity relation on U is defined as

$$R_B^{\varepsilon}(x_i,x_j)=\begin{cases}0, & R_B(x_i,x_j)\le \varepsilon \\ R_B(x_i,x_j), & R_B(x_i,x_j)>\varepsilon\end{cases} \tag{5}$$

where ε is a parameter that controls the similarity of samples. Obviously, R_B^{ε} is also an irreflexive and symmetric fuzzy relation. For any $x_i \in U$, the corresponding fuzzy similarity class is denoted as $[x_i]_B^{\varepsilon}(x_j)=R_B^{\varepsilon}(x_i,x_j)$, $x_j \in U$.

Obviously, the fuzzy similarity relation R_B^{ε} is affected by ε and B. The membership degrees of R_B^{ε} decrease with the increase of features in B and grow with the decrease of the value of ε.

Proposition 1. Let B be a subset of A, then $R_A^{\varepsilon} \subseteq R_B^{\varepsilon}$.

Proposition 2. Let $\varepsilon_1 \le \varepsilon_2$, then $R_B^{\varepsilon_2} \subseteq R_B^{\varepsilon_1}$.

Let $U/D=\{D_1,D_2,\cdots D_r\}$, B be a subset of A, and R_B^{ε} be a fuzzy similarity relation related to B. Then the fuzzy rough approximations of D_k related to B are redefined as

$$\underline{R}_B^{\varepsilon}(D_k)(x_i)=\min_{x_j\notin D_k}\left\{1-R_B^{\varepsilon}(x_i,x_j)\right\} \tag{6}$$

$$\overline{R}_B^{\varepsilon}(D_k)(x_i)=\max_{x_j\in D_k} R_B^{\beta}(x_i,x_j) \tag{7}$$

for any $x_i \in U$. Because the similarity relation used here is irreflexive, the lower approximation of any decision class cannot be equal to zeros in a general case; it reflects the membership degree to a decision class. It is easily seen that the fuzzy rough approximations of a sample are just related to its class label and the similarity between itself and its neighbors and have nothing to do with reflexivity.

Example 1. Table 1 represents a decision table, where $U=\{x_1,x_2,\cdots,x_6\}$, $A=\{a_1,a_2,a_3\}$, and D is a label feature.

TABLE 1 DECISION TABLE

	a_1	a_2	a_3	D
x_1	1.0	5.0	1.3	1
x_2	1.4	4.5	1.5	1
x_3	0.8	4.4	1.9	2
x_4	0.5	4.0	2.1	2
x_5	2.1	3.6	1.7	3
x_6	2.8	4.1	2.0	3

First, the three features are standardized to the interval [0, 1]. Then the similarity degree r_{ij} ($i \ne j$) among samples are calculated according to the following formula.

$$r_{ij}=1-\frac{1}{m}\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})\wedge 2}$$

Thus, we obtain

$$R_A=(r_{ij})=\begin{pmatrix}0 & 0.84 & 0.71 & 0.58 & 0.59 & 0.55\\ 0.84 & 0 & 0.81 & 0.69 & 0.75 & 0.69\\ 0.71 & 0.81 & 0 & 0.87 & 0.72 & 0.70\\ 0.58 & 0.69 & 0.87 & 0 & 0.70 & 0.66\\ 0.59 & 0.75 & 0.72 & 0.70 & 0 & 0.80\\ 0.55 & 0.69 & 0.70 & 0.66 & 0.80 & 0\end{pmatrix}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

The label feature D partitions the sample set U into three decision classes $\{D_1, D_2, D_3\}$, where $D_1 = \{x_1, x_2\}$, $D_2 = \{x_3, x_4\}$, $D_3 = \{x_5, x_6\}$. According to Formula (6), the lower approximations of these classes are calculated as shown in Table 2.

TABLE2 ROUGH APPROXIMATIONS OF DECISION CLASSES

U	$\underline{R}(D_1)$	$\underline{R}(D_2)$	$\underline{R}(D_3)$
x_1	0.29	0.16	0.16
x_2	0.19	0.16	0.16
x_3	0.13	0.19	0.13
x_4	0.13	0.30	0.13
x_5	0.20	0.20	0.25
x_6	0.20	0.20	0.30

Because there is no the phenomenon that the approximation of a decision class equals to zeros, one can employ these membership functions to analyze misclassification case and improve the performance of feature selection. Next we give another pair of commonly used rough approximation decision operators.

Let $U/D = \{D_1, D_2, \cdots, D_r\}$. According to the literatures [41], [42], the corresponding fuzzy decisions of samples can be obtained. Assume that $\{\tilde{D}_1, \tilde{D}_2, \cdots \tilde{D}_r\}$ is the fuzzy decisions corresponding to $\{D_1, D_2, \cdots, D_r\}$. For any $D_k \in U/D$, the fuzzy rough approximations of D_i upon feature subset B can be then reformulated as follows (see Ref. [41] and [42]).

$$\underline{R}_B^{\varepsilon}(D_k)(x_i) = \inf_{x_j \in U} \max\left\{1 - R_B^{\varepsilon}(x_i, x_j), \tilde{D}_k(x_j)\right\} \tag{8}$$

$$\overline{R}_B^{\varepsilon}(D_k)(x_i) = \max_{x_j \in U} \inf\left\{R_B^{\varepsilon}(x_i, x_j), \tilde{D}_k(x_j)\right\} \tag{9}$$

for any $x_i \in U$.

According to Propositions 1 and 2, and the definitions of fuzzy rough approximations, the following properties can be easily derived.

Proposition 3. If $B_1 \subseteq B_2$, then $\underline{R}_{B_1}^{\varepsilon}(D_k) \subseteq \underline{R}_{B_2}^{\varepsilon}(D_k)$ and $\overline{R}_{B_2}^{\varepsilon}(D_k) \subseteq \overline{R}_{B_1}^{\varepsilon}(D_k)$

Proposition 4. If $0 < \varepsilon_1 \le \varepsilon_2$, then $\underline{R}_B^{\varepsilon_1}(D_k) \subseteq \underline{R}_B^{\varepsilon_2}(D_k)$ and $\overline{R}_B^{\varepsilon_2}(D_k) \subseteq \overline{R}_B^{\varepsilon_1}(D_k)$.

Let $U/D = \{D_1, D_2, \cdots, D_r\}$ and $B \subseteq A$. From the classical fuzzy rough sets, the fuzzy positive region and dependency function of D upon B can be respectively redefined as

$$POS_B^{\varepsilon}(D) = \bigcup_{k=1}^{r} \underline{R}_B^{\varepsilon}(D_k) \tag{10}$$

$$\gamma_B^{\varepsilon}(D) = \frac{\sum_{x \in U} POS_B^{\varepsilon}(D)(x)}{|U|} \tag{11}$$

Proposition 5. If $B_1 \subseteq B_2$, then $POS_{B_1}^{\varepsilon}(D) \subseteq POS_{B_2}^{\varepsilon}(D)$.

Proposition 6. If $0 < \varepsilon_1 \le \varepsilon_2$, then $POS_B^{\varepsilon_1}(D) \subseteq POS_B^{\varepsilon_2}(D)$.

Proposition 5 shows that the positive region becomes greater with the increase in the size of a feature subset. Proposition 6 shows that the positive region is also affected by the similarity threshold.

The dependency function can be used as a criterion of feature selection, but it only considers the membership information of samples to the fuzzy positive region, and omits the information of minimum classification error rate. For example, Fig.2 and 3 show a binary classification problem in feature subspaces B and C, respectively. $\underline{R}_B^{\varepsilon}(D_1)$ ($\underline{R}_C^{\varepsilon}(D_1)$) and $\underline{R}_B^{\varepsilon}(D_2)$ ($\underline{R}_C^{\varepsilon}(D_2)$) indicate the membership function curves of the first and second classes in the subspace B (C), respectively. It is easy see that the fuzzy positive region in Fig.1 is greater than that in Fig.2. According to the classical fuzzy rough set theory, we consider that the feature subset B is more optimal than C when the fuzzy dependency function is used for the criteria of feature selection.

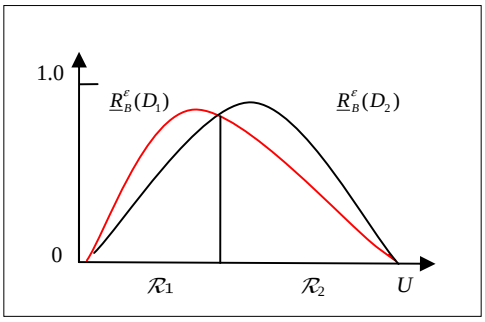

Fig. 2 The membership curves of two classes in feature subspace B

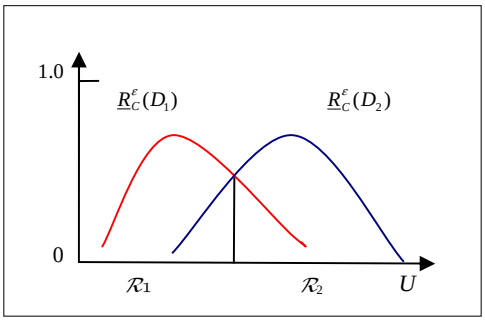

Fig. 3 The membership curves of two classes in feature subspace C

However, the difference between the two curves is smaller in the subspace B than that in C. According to the Bayes decision theory, the classification error rate gets smaller as the difference of the two membership function curves becomes larger. Therefore, the classification error rate in the feature subspace C is significantly less than that of B. This leads to a problem: how to employ the concept of fuzzy rough approximations to construct an effective evaluation function for feature selection? Intuitively, a subset of good features for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

classification learning should be having both the maximum fuzzy positive region and the minimum classification error rate. Based on the above analysis, it can be concluded that the fuzzy rough dependency functions are not good criterions for feature selection; they can just retain the maximal dependency between decision and features and cannot keep the minimum classification error. In the following, we introduce a new dependency function: inner product dependency, to describe the classification error of a classification problem.

Definition 2. Let $U/D=\{D_1,D_2,\cdots,D_r\}$ and $B\subseteq A$. Define

$$\omega_B^{\varepsilon}(D)=\frac{1}{\left|U^*\right|^2\left|r(r-1)\right|}\sum_{k\neq l}\sum_{x_i\in U^*}\underline{R}_B^{\varepsilon}(D_k)(x_i)\underline{R}_B^{\varepsilon}(D_l)(x_i) \quad (12)$$

where $U^*=\left\{x_i\in U\mid\sum_{k=1}^{r}\underline{R}_B^{\varepsilon}(D_k)(x_i)\neq 0\right\}$, $\omega_B^{\varepsilon}(D)$ is called the inner product dependency of D related to B. Obviously, $0\le\omega_B^{\varepsilon}(D)<1$.

The value of inner product dependency reflects the overlap degree of different classes. The smaller the degree of overlap, the smaller the inner product dependency. Because the overlap degree of decision classes is linked closely with the classification error, the concept of inner product dependency can describe the classification error.

The following properties can be derived directly from Proposition 3 and Definition 2.

Proposition 7. If $B_1\subseteq B_2$, then $\omega_{B_1}^{\varepsilon}(D)\subseteq\omega_{B_2}^{\varepsilon}(D)$.

Proposition 8. If $0<\varepsilon_1\le\varepsilon_2$, then $\omega_B^{\varepsilon_2}(D)\subseteq\omega_B^{\varepsilon_1}(D)$.

Propositions 7 shows that the inner product dependency function gets greater as the number of features increases. Propositions 8 shows the proposed dependency gets smaller with the increase of the value of the similarity threshold.

Theorem 1. If $\omega_B^{\varepsilon}(D)=\omega_A^{\varepsilon}(D)$, then $POS_B^{\varepsilon}(D)=POS_A^{\varepsilon}(D)$.

Proof. Since $B\subseteq A$, from Proposition 3 we have that $\underline{R}_B^{\varepsilon}(D_i)\subseteq\underline{R}_A^{\varepsilon}(D_i)$ for any $D_i\in U/D$. This means that $\underline{R}_B^{\varepsilon}(D_i)(x_j)\le\underline{R}_A^{\varepsilon}(D_i)(x_j)$ for any $D_i\in U/D$ and $x_j\in U$. Without loss of generality, we suppose that there exists a decision class $D_k\in U/D$ and a sample $x_l\in U$ such that $\underline{R}_B^{\varepsilon}(D_k)(x_l)<\underline{R}_A^{\varepsilon}(D_k)(x_l)$. By Definition 2, we have that $\omega_B^{\varepsilon}(D)\le\omega_A^{\varepsilon}(D)$, which is a contradict. Hence, $\underline{R}_B^{\varepsilon}(D_i)(x_j)=\underline{R}_A^{\varepsilon}(D_i)(x_j)$ holds for any $D_i\in U/D$ and $x_j\in U$, which implies that $\underline{R}_B^{\varepsilon}(D_i)=\underline{R}_A^{\varepsilon}(D_i)$ for any $D_i\in U/D$. It follows from Formula (10) that $POS_B^{\varepsilon}(D)=POS_A^{\varepsilon}(D)$.

It should be pointed out that the converse of the theorem is incorrect. The theorem shows that a feature subset can maintain the fuzzy dependency if it keeps the inner product dependency invariant. This means that the inner product dependency contain not only the classification information of the fuzzy positive region, but also additional information which the fuzzy positive region does not have. Let us analyze what is the additional information in the following.

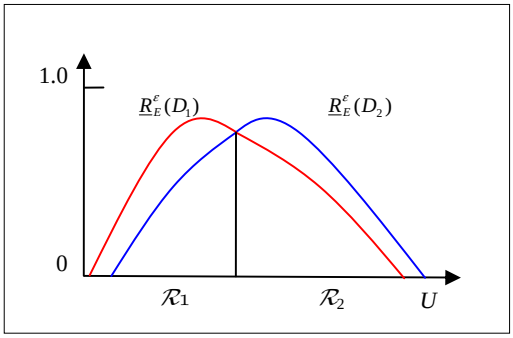

Fig. 4 The membership curves of two classes in the feature subspace E

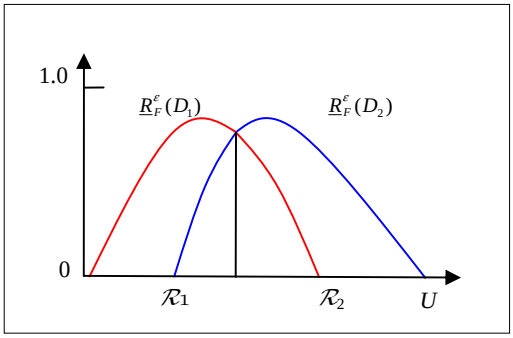

Fig. 5 The membership curves of two classes in the feature subspace F

As shown in Fig.4 and Fig.5, we temporarily assume that $\underline{R}_E^{\varepsilon}(D_1)\cup\underline{R}_E^{\varepsilon}(D_2)=\underline{R}_F^{\varepsilon}(D_1)\cup\underline{R}_F^{\varepsilon}(D_2)$, this means the fuzzy positive regions in the two subspaces are equal. According to the classical fuzzy rough set theory, the feature subsets E and F have the same classification abilities. The major disadvantage of the theory lies in that the fuzzy positive region omits the classification information of the lower parts of the membership function curves $\underline{R}_E^{\varepsilon}(D_1)$ and $\underline{R}_E^{\varepsilon}(D_2)$ or $\underline{R}_F^{\varepsilon}(D_1)$ and $\underline{R}_F^{\varepsilon}(D_2)$. In fact, the classification ability of a feature subset not only depends on the fuzzy positive region, is also related to the lower parts of the curves. The overlap degree of different membership function curves determines the classification error. The overlap degree in E is greater than that in F. Hence, The feature subset F should have the greater classification ability than E although they have the same fuzzy positive regions. We can easily see that the inner product dependency in the subspace F is less than that in E and that the concept of the inner product dependency can better reflect the classification error.

Based on the above observation, the inner product dependency not only contains the classification information of the fuzzy positive region, but also considers the error

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

information of classification.

Definition 3. Let (U, A, D) be a decision table and $B \subseteq A$. Then we call B a reduct of (U, A, D) if it has the minimum classification error and satisfies the following conditions:

(1) $\gamma_B^{\varepsilon}(D) = \gamma_A^{\varepsilon}(D)$;

(2) $\gamma_{B-\{a\}}^{\varepsilon}(D) < \gamma_B^{\varepsilon}(D)$ for $\forall a \in B$.

It is easy to see that a reduct is a minimal feature subset that has the minimum classification error and has the same classification ability as the whole set of features.

IV. Fuzzy rough minimum misclassification criterion

In practice, the incremental strategy is usually used to search for the optimal subset of features. The search begins with a nonempty set, finds one feature with great significances according to the feature evaluation criterion, and puts it into the selected feature subset each time. In views of the results in the previous section, we know that an optimal feature subset should be with the greatest fuzzy positive region and with the minimum classification error. Because the inner product dependency considers both of the classification error and the classification information in the fuzzy positive region, we introduce a novel criterion for feature selection as follows.

Definition 4. Let $B \subseteq A$ and $a \in A - B$. Then the importance of a related to B is defined by

$$\phi(a; B, D) = \frac{\gamma_{B\cup\{a\}}^{\varepsilon}(D) - \gamma_B^{\varepsilon}(D)}{\sqrt{\omega_{B\cup\{a\}}^{\varepsilon}(D)}} \tag{13}$$

The numerator $\gamma_{B\cup\{a\}}^{\varepsilon}(D) - \gamma_B^{\varepsilon}(D)$ denotes the increment of the fuzzy dependency caused by the feature a, and the denominator is defined as the square root of $\omega_{B\cup\{a\}}^{\varepsilon}(D)$ by considering dimensional scaling. We can easily see that the more significant the feature a is, the greater the value of $\phi(a; B, D)$ is. The proposed significance measure considers both the increment of the fuzzy dependency and the minimum misclassification rate. Therefore, it can better reflect the classification capability of a candidate feature.

On the basis of the above observations, we can design a heuristic algorithm for feature selection as follows.

Algorithm: Fuzzy rough algorithm with minimum misclassification rate (FRMR):

Input: A decision table $DS = (U, A, D)$ and parameters ε and δ

Output: one reduct red

1: Compute the fuzzy similarity relation R_a for any $a \in A$.
2: Let start=1, $red = \varnothing$, $B = A - red$.
3: while start
4: for each $a_i \in B$
5: Compute fuzzy relation $R_{red\cup\{a\}}^{\varepsilon}$
6: for each $x_j \in U$
7: for each $D_m \in U/D$
8: Compute the lower membership function $\underline{R_{red\cup\{a\}}}(D_m)(x_j)$
9: end for
10: end for
11: Compute $\gamma_{red\cup\{a\}}^{\varepsilon}$, $\omega_{red\cup\{a\}}^{\varepsilon}(D)$ and $\phi(a; red, D)$
12: end for
13: Find the feature a_k with maximum value $\phi(a_k; red, D)$
14: if $\gamma_{red\cup a_k}^{\varepsilon}(D) - \gamma_{red}^{\varepsilon}(D) > \delta$
15: $red = red \cup \{a_k\}$
16: $B = B - red$
17: else
18: start=0
19: end if
20: end while
21: return red

In the proposed algorithm, the parameter ε is used for controlling the similarity of samples and δ is used to terminate the main loop. In fact, the optimal values of parameters ε and δ are different for different data sets. In Section 5, we will discuss the search method for the optimal values of the two parameters by using the ten-fold cross validation technique.

Suppose that the numbers of training samples and features are n and m, respectively. It is easy to know that the computational complexity for the fuzzy similarity relations is $n^2 m$. The procedure from step 4 to 12 is used for computing the importance of each candidate feature. Step 13 is used to find the feature with maximum significance. The part of procedure from step 14 to 19 is used for terminating the main loop. The maximum search time for one optimal feature subset will lead to $(m^2 + m)/2$ evaluations of the criterion function. Therefore, the total time complexity of the proposed algorithm is in $O(n^2 m + (m^2 + m)/2)$.

V. Experimental analysis

In this section, we compare some existing algorithms with the proposed algorithm. Three representative feature selection algorithms in fuzzy rough set theory are selected. These are fuzzy information entropy algorithm (FRSE) [25], fitting fuzzy rough algorithm (FITF) [42], and fuzzy variable precision rough set algorithm (FPRS) [40]. Two classical classifiers including SVM classifier and KNN classifier (K=3) are used to evaluate the performance of these feature selection algorithms. The parameters in SVM are set to the default values. All the algorithms are run in Matlab 2013b. Fourteen data sets were downloaded in the UCI and KRB database for experimentation. The basic description related to these data sets is shown in Table 3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

TABLE 3 DESCRIPTION OF EXPERIMENTAL DATA

No	Data sets	Sample	Features	Classes
1	gamma	19020	10	2
2	glass	214	10	7
3	horse	368	22	2
4	iris	150	4	3
5	mushroom	8124	22	2
6	segmentation	2310	18	7
7	sonar	208	60	2
8	wdbc	569	30	2
9	wine	178	13	3
10	hill	1212	101	2
11	colon	62	2000	2
12	Breast	84	9216	5
13	prostate	102	10509	2
14	MLL	72	12582	3

A. Training Parameters

As shown in Section 4, FRMR algorithm has two parameters ε and δ which are used for the thresholds of fuzzy similarity and algorithm termination, respectively. The effective selection of the two parameters is the necessary guarantee for the algorithm to output the optimal feature subset. Theoretically, the optimal solution of the parameters should be searched from the entire range space for each data set. Fortunately, as discussed in the literature [9]-[14], [24]-[29], [41],[42] for such an algorithm with two parameters in rough set models, if one parameter is controlled at a certain value and the optimal value of another parameter is searched in its entire range space, then the optimal performance of the algorithm can be approximately obtained. According to this idea, the value of parameter ε will be set to a constant 0 in the following experiments and then the optimal value of algorithm termination δ is set to be searched in the interval [0,0.05] with step 0.001.

TABLE 4 TRAINING RESULTS FOR OPTIMAL SHUTDOWN THRESHOLDS

Dataset	FRMR	FRSE	FITF	FPRS
gamma	0.006	0.000	0.0001	0.0001
glass	0.026	0.011	0.026	0.021
horse	0.025	0.013	0.026	0.025
iris	0.031	0.014	0.017	0.019
mushroom	0.005	0.006	0.022	0.021
segmentation	0.008	0.009	0.004	0.005
sonar	0.010	0.001	0.002	0.002
wdbc	0.028	0.001	0.001	0.003
wine	0.023	0.008	0.011	0.003
hill	0.001	0.000	0.0001	0.0002
colon	0.032	0.013	0.012	0.010
Breast	0.017	0.018	0.015	0.017
prostate	0.016	0.011	0.020	0.022
MLL	0.023	0.015	0.022	0.023

We use 10-fold cross validation to perform feature selection on these data sets. That is to say, for a value of termination parameter δ and a data set, we divide the data set into ten parts, of which nine parts are used as the training set and one part is for the testing set. In the training stage, feature selection is conducted on the training set and an optimal feature subset is selected. In the testing stage, a subdata is extracted from the testing data by using the optimal feature subset. The extracted subdata is then sent to SVM and 3NN classifiers for computing the classification accuracy of the subdata. After ten cycles, the average value of the results of ten cycles is taken as the final result of feature selection. In experiments, we try δ from 0 to 0.05 with step 0.001 and find the optimal value of δ for each data set. Similarly, we have done the experiments with the same parameter search way for the other three algorithms that need to be compared. The optimal values of the parameter δ in the experimental results are listed in Table 4. Thereafter, these values of the parameter δ in these algorithms are then all used in the next series of experiments.

B. Correlation Analysis of Different Classification Indexes

In order to analyze the relationship among the inner product dependency (abbreviated as IPD), the classical fuzzy rough set dependency (abbreviated as FRD) and classification accuracy. We select the wine dataset; randomly generate 10 subsets with 5 features. Then, we calculate the fuzzy rough set dependencies, the inner product dependencies and the 3NN classifier accuracies and SVM classifier accuracies for each subset. The results are listed in Table 5.

Some conclusions can be drawn clearly from Table 5. It is easy to see that IPD is relevant to the classification capability of feature subsets. The experimental results show that the IPD for the fifth feature subset is the smallest, and the corresponding classification accuracy of 3NN and SVM classifiers with ten-fold cross-validation is the highest. For the third feature subset, the IPD is relatively large, and the classification accuracy of the two classifiers is also relatively low. In addition, the correlation analysis was used to calculate the correlation coefficient between IPD and 3NN classifier accuracy as -0.7908, and the correlation between IPD and SVM classifier accuracy is -0.8221. Obviously this correlation is strongly related.

TABLE 5 FUZZY ROUGH SET DEPENDENCY, INNER PRODUCT DEPENDENCY AND CLASSIFICATION ACCURACY

ID	Features	FRD	IPD	3NN	SVM
1	2,3,1,10,4	0.2039	0.1382	89.38	92.71
2	13,4,10,2,11	0.2199	0.1286	93.89	92.78
3	6,7,2,4,12	0.1686	0.1476	84.24	86.04
4	9,11,3,6,4	0.1701	0.1404	83.26	89.93
5	13,6,10,5,7	0.2247	0.1237	96.04	97.15
6	4,2,7,11,5	0.1873	0.1361	91.60	93.89
7	5,1,9,11,7	0.2124	0.1379	93.13	94.38
8	3,7,5,13,8	0.2079	0.1351	92.78	95.00
9	7,2,10,4,3	0.1976	0.1345	93.82	93.26
10	1,12,3,13,6	0.2343	0.1368	94.38	90.97

At the same time, it can be clearly seen from Table 5 that the IPD is also related to the FRD. That is, the smaller the IPD, the larger the FRD of the feature subset. Of course, this conclusion is not always established. For example, in the calculation of No. 2, the value of IPD is relatively small, and the accuracy performance of the two classifiers is not very satisfactory. This is because the IPD just has the mathematical meaning of the misclassification rate. It can only reflect the misclassified information of the feature subset. This information reflects one side of the classification ability, and does not fully embody the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

classification ability of feature subset. Furthermore, we did the same experiment in other datasets. The experimental results show that the correlation between IPD and classification accuracies is similar to the wine data set.

For the four data sets of glass, iris, soybean, and mushroom, we employ the FRMR algorithm to output a sequence of selected feature subsets , At this point, we set the shutdown condition as δ =0. Along with the original data set, we separately plot the variation of the IPD, 3NN accuracy and SVM accuracy as the number of features increases in Fig. 6 to Fig. 8.

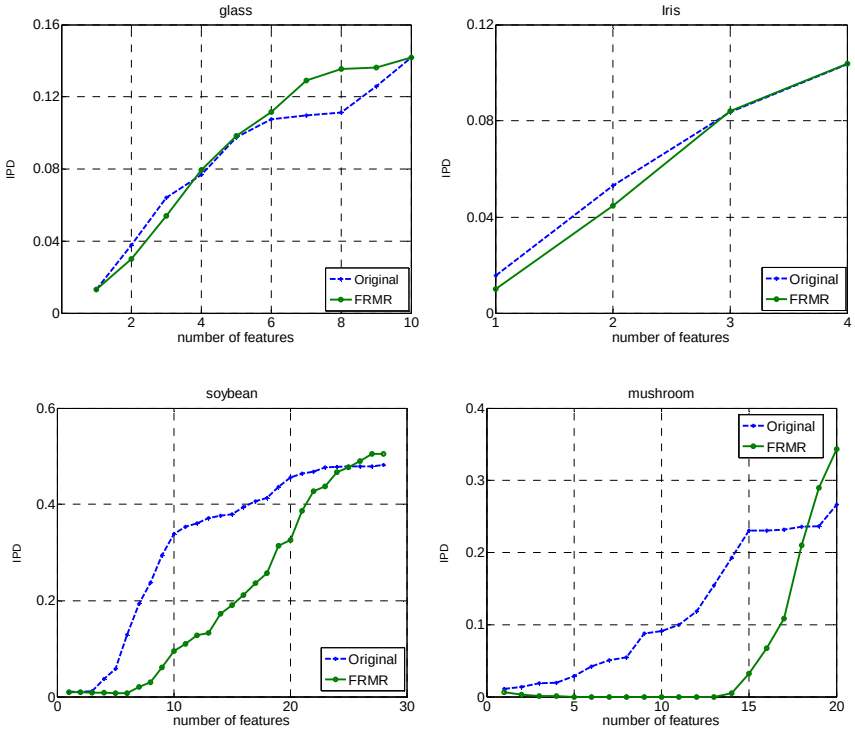

Fig. 6 The inner product dependence curve with the features increases

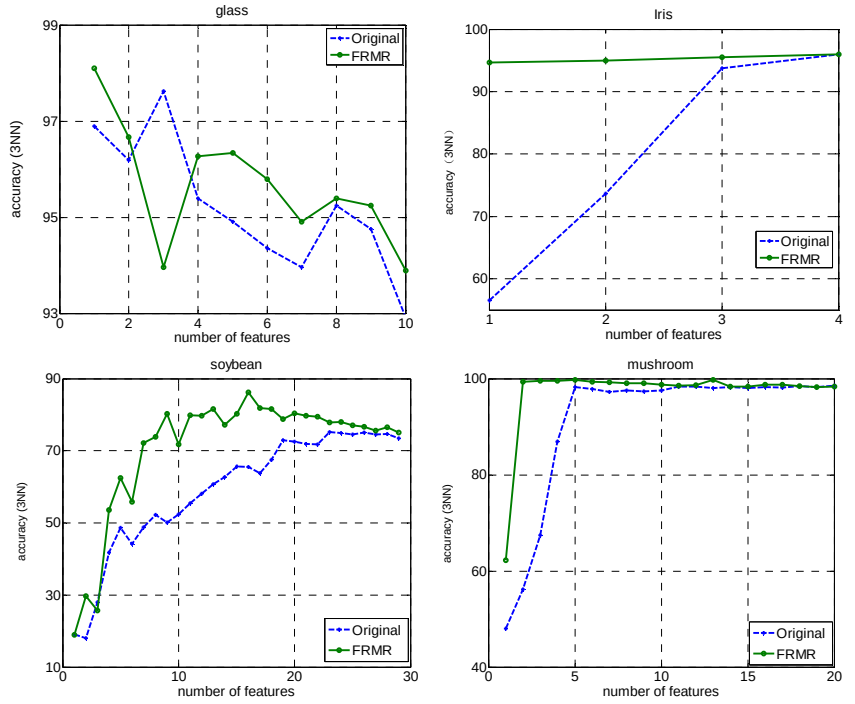

Fig. 7 3NN classifier accuracy curve with the features increases

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

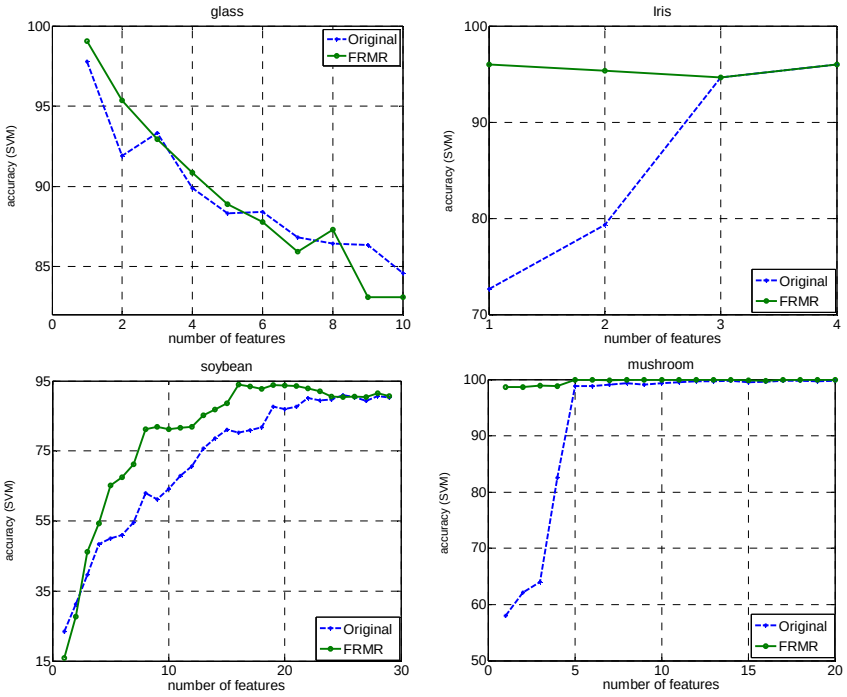

Fig. 8 SVM classifier accuracy curve with the features increases

We first analyzed the graph of the IPD curves as Fig 6. Through observation, we can find that the same situation is reflected in the four sets of data: in the initial stage of feature sequence, the IPD computed by the proposed algorithm is less than that computed by the original feature sequence. As the number of features increases, the IPD of the FRMR algorithm will eventually intersect with the IPD of the original data, even exceeding it. This shows that the feature subset output by the FRMR algorithm in the initial stage can effectively reduce the possibility of misclassification.

We again compared the 3NN, SVM classifier accuracy curve. It is easy to see that the FRMR algorithm improves the classification accuracies over the original data sets to a large extent. Moreover, when the classification accuracy calculated by FRMR algorithm reaches the highest, the corresponding the IPD is smaller than the IPD of the original feature sequence. For example, for glass data, for the SVM classifier and 3NN classifier, the FRMR algorithm selects the first feature with the highest precision output, while the IPD is the lowest at this time. Therefore, it can be concluded that the optimal feature subset appears when the IPD of the FRMR algorithm is less than the IPD of the original data.

C. Performance Comparison of Different Algorithms

To demonstrate the anti-noise effect of our proposed algorithm, 5% label noises were generated in data sets: gamma, mushroom, segmentation, sonar and wine, that is, the label orders of 5% samples in these datasets were randomly shuffled. Then, feature reductions are performed on these data sets. The experimental results are shown in Table 6. It is easy to see that the proposed method has better anti-noise ability than other methods, which is due to the introduction of the misclassification rate index in our algorithm.

Based on the results of the parameter training in Tables 4, we used the 3NN classifier and the SVM classifier to perform data experiments in 14 selected data sets. The numbers of selected optimal features and the running time are listed in Table 7. Obviously, all four algorithms achieve the goal of feature selection. Especially for the FRMR algorithm, the number of selected features is relatively small in most of data sets. However, for the sonar data set, FRMR algorithm selects more features than the other three algorithms. If we compare the results in Tables 8 and 9, it is not difficult to find out that for the sonar data set, the 3NN and SVM accuracy of the FRMR algorithm are obviously superior to the accuracy of other algorithms. Moreover, for the original 60 features, the FRMR algorithm selected 12 features, that is a large degree of reduction. For data sets with fewer categories, the running time of the FRMR algorithm is about the same as that of the other algorithms. In data sets with a large number of categories, the FRMR algorithm takes more time than other algorithms. This is because the FRMR algorithm not only needs to compute the lower approximations of categories like other algorithms, but also needs to compute the inner product of the lower approximations, while other algorithms only need to compute the lower approximation of categories.

The test accuracy of the two classifiers for these data sets is shown in Tables 8 and 9, respectively, where the underlined portion is the maximum test accuracy of the current data set. For most data sets, the classification accuracy guided by our proposed algorithm is higher than the accuracy of the original data set. Moreover, compared with the other three algorithms, the highest precision of most data sets is also the most frequent occurrence of the FRMR algorithm. This shows that the selected feature subset with FRMR has stronger classification ability. For the glass and colon datasets, the FRMR algorithm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

selected the least subset of features; however, it did show better classification accuracy. For breast, gamma datasets, the classification accuracies of feature subsets obtained by FRMR algorithm are much higher than those obtained by other algorithms. The emergence of the manifestations indicates that the inner product dependency can reach a higher level in eliminating redundant features.

TABLE 6 CLASSIFICATION ACCURACIES OF NOISED DATA AT 5% NOISE LEVEL

Data sets	Classifier	Raw data	Noised data	FRMR	FRSE	FITF	FPRS
gamma	SVM	68.78	65.67	70.44	66.53	64.33	65.84
	3NN	80.71	77.56	78.35	76.20	76.47	76.28
mushroom	SVM	99.94	96.00	96.31	92.58	95.76	93.86
	3NN	99.57	97.29	97.43	91.29	96.35	91.45
segmentation	SVM	97.14	91.52	91.83	88.81	91.53	90.14
	3NN	96.12	90.57	90.35	89.44	90.41	89.89
sonar	SVM	88.11	86.04	84.42	72.45	72.82	72.14
	3NN	84.34	84.14	78.35	72.85	77.73	71.89
wine	SVM	96.67	90.35	92.24	89.95	91.34	89.98
	3NN	96.04	93.26	92.76	89.72	90.45	90.61
average	—	90.74	87.24	87.25	82.98	84.72	83.21

TABLE 7 THE SELECTED FEATURE NUMBERS AND THE RUNNING TIME OF THE FOUR ALGORITHMS (NUMBER/TIME)

Data sets	Raw data	FRMR	FRSE	FITF	FPRS
gamma	10	5.8/989.29	2.0/788.56	6.1/941.39	6.2/950.48
glass	10	1.0/0.19	2.1/0.11	2.3/0.09	2.1/0.16
horse	22	8.8/1.31	6.7/0.85	7.2/0.98	9.2/1.01
iris	4	1.2/0.05	1.6/0.05	3.1/0.02	1.7/0.02
mushroom	22	5.6/ 292.88	7.2/ 286.20	6.4/ 256.12	6.8/ 269.70
segmentation	18	5.6/74.14	5.8/34.53	8.2/35.13	7.3/29.53
sonar	60	12.5/2.18	8.6/1.36	8.6/1.90	8.7/1.87
wdbc	30	3.1/2.54	7.2/3.31	6.7/3.25	6.5/3.07
wine	13	4.9/0.23	5.2/0.09	4.6/0.20	7.9/0.58
hill	100	6.0/66.66	6.4/48.17	4.4/42.53	5.5/45.88
colon	2000	4.9/8.58	6.8/5.78	8.7/6.91	5.2/5.38
breast	9216	8.8/181.19	7.3/35.55	8.9/51.84	8.6/73.22
prostate	10509	9.7/143.26	7.5/96.61	6.6/70.36	6.4/77.41
MLL	12582	6.7/114.71	11.5/43.12	9.4/48.02	8.1/45.45
average	2471.14	6.06/134.08	6.12/96.02	6.51/104.20	6.44 /107.41

TABLE 8 CLASSIFICATION ACCURACIES OF REDUCED DATA WITH 3NN

Data sets	Raw data	FRMR	FRSE	FITF	FPRS
gamma	80.68 ± 1.83	83.12 ± 1.16	78.21 ± 1.81	79.88 ± 2.16	79.32 ± 2.54
glass	91.29 ± 5.04	99.07 ± 1.28	96.82 ± 2.55	96.48 ± 2.54	97.07 ± 2.48
horse	90.76 ± 3.79	92.61 ± 3.63	90.53 ± 3.27	91.19 ± 3.99	90.38 ± 4.11
Iris	95.33 ± 1.83	96.67 ± 2.98	96.00 ± 2.79	96.67 ± 2.98	95.33 ± 1.83
mushroom	99.57 ± 0.41	99.75 ± 0.26	99.51 ± 0.31	99.88 ± 0.28	99.57 ± 0.32
segmentation	96.16 ± 0.73	96.36 ± 0.83	95.37 ± 0.75	95.28 ± 0.78	95.19 ± 1.02
sonar	84.19 ± 8.95	85.12 ± 5.73	80.03 ± 6.12	81.64 ± 6.13	80.98 ± 6.35
wdbc	96.68 ± 2.25	96.54 ± 1.78	96.29 ± 1.46	96.48 ± 2.42	97.20 ± 1.77
wine	95.95 ± 3.25	97.75 ± 3.34	96.63 ± 3.62	96.63 ± 3.56	97.21 ± 3.44
hill	50.89 ± 3.73	52.80 ± 4.24	52.39 ± 3.31	48.51 ± 3.53	47.87 ± 4.15
colon	71.19 ± 10.69	93.57 ± 10.89	88.81 ± 11.06	86.28 ± 12.33	85.66 ± 10.26
breast	69.19 ± 15.74	100.00 ± 0.00	94.04 ± 5.26	90.29 ± 6.17	90.62 ± 4.88
prostate	83.50 ± 8.26	96.00 ± 6.80	93.17 ± 8.45	96.00 ± 10.07	95.83 ± 9.89
MLL	83.78 ± 10.27	97.32 ± 3.68	95.89 ± 8.19	97.14 ± 3.91	96.77 ± 4.87
average	84.94 ± 5.48	91.91 ± 3.47	89.55 ± 4.21	89.45 ± 4.35	89.21 ± 4.14

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

TABLE 9 CLASSIFICATION ACCURACIES OF REDUCED DATA WITH SVM

Data sets	Raw data	FRMR	FRSE	FITF	FPRS
gamma	68.89 ± 2.56	73.11 ± 1.66	68.02 ± 2.13	70.58 ± 1.35	70.88 ± 2.23
glass	93.21 ± 4.52	99.06 ± 1.28	97.67 ± 1.28	98.14 ± 1.04	97.12 ± 1.23
horse	90.75 ± 3.64	91.31 ± 3.36	91.04 ± 3.84	91.04 ± 3.86	89.87 ± 4.44
iris	94.67 ± 2.14	96.00 ± 2.79	95.33 ± 1.83	95.33 ± 1.83	95.33 ± 1.83
mushroom	99.94 ± 0.14	100.00 ± 0.00	99.88 ± 0.28	99.75 ± 0.28	100.00 ± 0.00
segmentation	97.27 ± 0.64	96.82 ± 0.43	95.58 ± 0.58	95.19 ± 0.70	96.72 ± 1.13
sonar	88.02 ± 9.15	88.96 ± 6.12	81.88 ± 5.80	85.50 ± 5.65	85.08 ± 6.21
wdbc	97.19 ± 2.08	97.49 ± 1.33	97.88 ± 1.89	97.11 ± 1.99	96.31 ± 2.04
wine	96.62 ± 3.17	98.86 ± 3.33	97.22 ± 3.79	97.19 ± 3.40	97.13 ± 3.75
hill	55.26 ± 5.33	56.36 ± 5.12	55.74 ± 4.78	51.22 ± 3.50	51.48 ± 3.96
colon	64.76 ± 11.06	90.48 ± 9.87	86.37 ± 11.28	85.48 ± 11.57	85.71 ± 12.84
breast	38.16 ± 13.48	98.82 ± 2.63	93.90 ± 6.43	94.12 ± 4.56	95.05 ± 4.94
prostate	56.59 ± 11.35	100 ± 0.00	96.17 ± 7.87	96.02 ± 9.91	96.55 ± 6.13
MLL	38.93 ± 12.56	98.57 ± 3.19	96.50 ± 5.36	98.57 ± 3.19	97.29 ± 5.66
average	77.16 ± 5.84	91.84 ± 2.94	89.51± 4.08	89.76 ± 3.78	89.55 ± 4.03

Next, we analyzed the significant differences in the experimental results of these algorithms. We selected statistics Friedman test [43] and Bonferroni–Dunn test [44] to evaluate these experimental results.

The Friedman statistic is formulated as:

$$\chi_F^2 = \frac{12n}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right) \text{ and } F = \frac{(n-1)\chi_F^2}{n(k-1)-\chi_F^2} \qquad (14)$$

where k is the number of algorithms, r_i is the average rank of algorithm i and n is the number of data sets, F follows the Fisher distribution with $k-1$ and $(k-1)(n-1)$ degrees of freedom. As shown in Ref. [43], the critical value $F(3,39) = 2.555$ when the significance level $\alpha = 0.1$.

The null hypothesis tested by the Friedman was that all algorithms were the same in classification performance. If the null hypothesis is rejected, the Bonferroni–Dunn test is performed to further examine which algorithms are different. According this test, the performance of two algorithms is thought to be significantly different when their average rank distance exceeds the critical distance

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6n}} \qquad (15)$$

where $q_{0.10} = 2.291$ as shown in [45].

Tables 10 and 11 show the rankings of these algorithms under the two classifiers. It follows from Formula (14) that $F = 6.748$ for 3NN classifier and $F = 9.332$ for SVM classifier. Both were greater than the threshold of significance $F(3,39)$ at the level α (= 0.1). Thus, we can consider these algorithms to be significantly different. According to Formula (15), it follows that $CD_{0.10} = 1.118 (k = 4, N = 14)$.

One can observe from Table 10 that the average rank of FRMR is greater than 1.118 plus that of any of other algorithms for 3NN. Thus, the tests demonstrate that FRMR is statistically better than other three algorithms for 3NN classifier. From Table 11, the tests demonstrate that FRMR is also statistically better than other algorithms for SVM classifier.

TABLE 10 RANK OF THE FOUR ALGORITHMS WITH 3NN

Data sets	FRMR	FRSE	FITF	FPRS
gamma	1	4	2	3
glass	1	2	3	4
horse	1	3	2	4
iris	1.5	3	1.5	4
mushroom	2	4	1	3
segmentation	1	2	3	4
sonar	1	4	2	3
wdbc	2	4	3	1
wine	1	3.5	3.5	2
hill	1	2	3	4
colon	1	2	3	4
breast	1	2	4	3
prostate	1.5	4	1.5	3
MLL	1	4	2	3
average	1.25	3.17	2.54	3.08

TABLE 11 RANK OF THE FOUR ALGORITHMS WITH SVM

Data sets	FRMR	FRSE	FITF	FPRS
gamma	1	4	3	2
glass	1	3	2	4
horse	1	2.5	2.5	4
Iris	1	3	3	3
mushroom	1.5	3	4	1.5
segmentation	1	3	4	2
sonar	1	4	2	3
wdbc	2	1	3	4
wine	1	2	3	4
hill	1	2	4	3
colon	1	2	4	3
breast	1	4	3	2
prostate	1	3	4	2
MLL	1.5	4	1.5	3
average	1.21	2.89	3.07	2.82

D. Fixed Number of Selected Features

In general, for a given data set and a classifier, an optimal subset of features exists and is not unique. Most feature selection algorithms have a termination condition that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2021.3097811, IEEE Transactions on Fuzzy Systems

determines the number of selected features. Different values of the termination condition will cause an algorithm to select different number of features. However, in some practical problems, due to the restriction of objective conditions, the optimal feature subset is not needed. Instead, the required number of features is given based on a particular problem, and then feature selection is carried out according to the fixed number. For example, the approximately optimal numbers of features were given for datasets: breast, colon and prostate in literature [46], [47]. Next, we set the number of selected features as D = 6 for high-dimensional datasets like the way in [47], and compare the performance of the four algorithms. The classification results for the reduced data sets are shown in Tables 12 and 13, respectively.

TABLE 12 CLASSIFICATION ACCURACIES OF REDUCED DATA WITH 3NN (D = 6)

Data sets	Raw data	FRMR	FRSE	FITF	FPRS
horse	90.76	91.61	90.51	90.51	90.05
mushroom	99.57	99.62	99.36	99.71	99.26
segmentation	96.16	96.08	95.21	94.94	94.56
sonar	84.19	83.18	75.36	77.94	76.86
wdbc	96.68	96.31	93.83	95.95	95.55
hill	50.89	51.14	52.11	48.19	47.59
colon	71.19	92.15	86.25	85.42	85.42
breast	69.19	95.00	92.75	89.83	90.27
prostate	83.50	95.00	92.00	94.17	92.10
MLL	83.78	95.71	91.43	95.71	95.71
average	82.59	89.58	86.88	87.24	86.74

TABLE 13 CLASSIFICATION ACCURACIES OF REDUCED DATA WITH SVM (D = 6)

Data sets	Raw data	FRMR	FRSE	FITF	FPRS
horse	90.75	91.01	89.39	89.39	89.56
mushroom	99.94	99.88	99.76	99.76	99.82
segmentation	97.27	96.03	95.21	94.97	94.31
sonar	88.02	78.38	75.94	73.98	75.89
wdbc	97.19	96.67	94.03	97.01	95.01
hill	55.26	51.74	51.24	50.91	50.71
colon	64.76	85.00	82.08	82.50	83.15
breast	38.16	96.25	90.00	90.42	92.12
prostate	56.59	95.00	91.00	92.17	91.00
MLL	38.93	95.71	90.32	94.29	93.41
average	72.69	88.57	85.90	86.54	86.50

It is easy to see that our method has a great advantage over other methods when the number of selected features is fixed at D = 6. The statistical significance test was also carried out for this experiment, and it can be also verified that FRMR is significantly better than the other three methods according to the Friedman and Bonferroni–Dunn statistics.

This experiment indicates that we do not need to calculate the parameters in these algorithms for some hybrid systems where the required number of selected features is given, so can avoid a lot of repetitive computations.

VI. CONCLUSION

Fuzzy dependency function is often used as a feature selection criterion in fuzzy rough set theory. This function only considers the classification information in the upper branches of the membership function curves of decision classes, and ignores the information in the lower branches of these curves. In this paper, we introduced concept of inner product dependency to characterize the classification error and propose a novel criterion for feature selection. This criterion makes full use of the classification information provided by the membership function curves of decision classes and overcomes the shortcoming of fuzzy rough dependency function. Using ten UCI and KRB datasets, we conducted a series of numerical experiments to evaluate the proposed approach. The experimental results indicate that the proposed approach can select fewer features and maintain higher classification accuracy in most of data sets. It is also found in experiments that the thresholds in the compared algorithms have impacts on the performance of feature selection. The optimal values of thresholds should be trained before feature selection for each data set.

In future work, we will introduce Bayesian minimum error rate into other rough set models and apply these models to pattern recognition problems such as classification and rule extraction.

REFERENCES

[1] M. Dash and H. Liu, “Consistency-based search in feature selection,” Artificial Intelligence, vol. 151, no. 1/2, pp. 155–176, 2003.
[2] M. A. Hall, “Correlation-based feature selection for discrete and numeric class machine learning,” In Proc. 17th Int. Conf. Machine Learning, pp. 359–366, 2000.
[3] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.
[4] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502, 2005.
[5] A. Tan, W. Wu, S. Shi, S. Zhao, "Granulation selection and decision making with making with multi-granulation rough set over two universes ," International Journal of Machine Learning & Cybernetics, vol.10, no.9, pp. 2501- 2513, 2019.
[6] X. He, L. Wei, Y. She, "L-fuzzy concept analysis for three-way decisions: basic definitions and fuzzy inference mechanisms," International Journal of Machine Learning and Cybernetics, vol. 9, no.11, pp. 1857–186, 2018.
[7] W. Xu, J. Yu, "A novel approach to information fusion in multi-source datasets: A granular computing viewpoint," Information Sciences, vol. 378, pp. 410 – 423, 2017.
[8] J. Zhan, H. Malik, M. Akram, "Novel decision-making algorithms based on intuitionistic fuzzy rough environment," International Journal of Machine Learning & Cybernetics, vol. 8, no.6, pp.1459-1485, 2018.
[9] J. Dai, Q. Hu, J. Zhang, H. Hu, N. Zheng, "Attribute selection for partially labeled categorical data by rough set approach," IEEE Trans. Cybernetics, vol. 47, no.9, pp. 2460-2471, 2017.
[10] W. Ding, C. Lin, M. Prasad, Z. Cao, and J. Wang, "A layered-coevolution-based attribute-boosted reduction using adaptive quantum behavior PSO and its consistent segmentation for neonates brain tissue," IEEE Transactions on Fuzzy Systems, vol. 26, no. 3, pp.1177-1191, 2018.
[11] G. Lang, Q. Li, M. Cai, T. Yang, Q. Xiao, "Incremental approaches to knowledge reduction based on characteristic matrices," International Journal of Machine Learning and Cybernetics, vol. 8, no.1, pp. 203-222, 2017.
[12] J. Y. Liang, F. Wang, C. Y. Dang, “A group incremental approach to feature selection applying rough set technique,” IEEE Transaction on Knowledge and Data Engineering, vol. 26, no. 2, pp. 294–304, 2014.
[13] L. Sun, J. Xu, Y. Tian, "Feature selection using rough entropy-based uncertainty measures in incomplete decision systems," Knowledge-Based Systems, vol.36, pp.206-216, 2012.
[14] L. Sun, X. Zhang, Y. Qian, J. Xu, S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification," Information Sciences, vol. 502, pp. 18-41, 2019.

[15] Y. Lin, H. Chen, G. Lin, J. Chen, Z. Ma, J. Li, "Synthesizing decision rules from multiple information sources: a neighborhood granulation viewpoint," International Journal of Machine Learning & Cybernetics, vol. 9, no.11, pp. 1919-1928, 2018.
[16] B. Sang, Y. Guo, D. Shi, W. Xu, "Decision-theoretic rough set model of multi-source decision systems," International Journal of Machine Learning & Cybernetics, vol. 9, no. 11, pp. 1941-1954, 2018.
[17] B. Sun, W. Ma, X. Chen, "Variable precision multigranulation rough fuzzy set approach to multiple attribute group decision-making based on similarity relation," Computers and Industrial Engineering, DOI:10. 1016 / j.cie.2018.10.009.
[18] D. Liang, D. Liu, W. Pedrycz, P. Hu, "Triangular fuzzy decision- theoretic rough sets," International Journal of Approximate Reasoning, vol.54, pp. 1087-1106, 2013.
[19] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences, vol. 11, no.5, pp. 341-356, 1982.
[20] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," International Journal of General Systems, vol. 17, pp. 191–208, 1990.
[21] N. Morsi, M. M. Yankout, "Axiomatic for fuzzy rough sets," Fuzzy sets and systems, vol.100, no.1-3, pp.327-342,1998.
[22] A.M. Radzikowska and E.E. Kerre, "A comparative study of fuzzy rough sets," Fuzzy Sets and Systems, vol. 126, no.22, pp.137–155, 2002.
[23] W. Wu and W. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," Information Sciences, vol. 159, pp. 233-254, 2004.
[24] Q. Hu, D. Yu, W. Pedrycz, D. Chen, "Kernelized fuzzy rough sets and their applications," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 11, pp. 1649 – 1667, 2011.
[25] Q. Hu, D. Yu, Z. Xie, J. Liu, "Fuzzy probabilistic approximation spaces and their information measures," IEEE Transactions on Fuzzy Systems, vol. 14, no. 2, pp. 191-201, 2006.
[26] Y. Yang, D. Chen, H. Wang, Eric C. C. Tsang, D. Zhang, "Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving," Fuzzy Sets and Systems, vol. 312, pp. 66-86, 2017.
[27] Y. Lin, Y. Li, C. Wang, J. Chen, "Attribute reduction for multi-label learning with fuzzy rough set," Knowledge-Based Systems, vol. 152 pp. 51-61, 2018.
[28] S. An, Q. Hu, W. Pedrycz, P. Zhu, Eric C. C. Tsang, "Data-distribution aware fuzzy rough set model and its application to robust classification," IEEE Transactions on Cybernetics, vol. 46, no.12, pp. 3073- 3085, 2016.
[29] P. Maji and P. Garai, "Fuzzy-rough simultaneous feature selection and feature extraction algorithm," IEEE Transactions on System, Man and Cybernetics, Part B, Cybernetics, vol. 99, pp.1-12, 2012.
[30] D. C. Martine, C. Cornelis and E. E. Kerre, "Fuzzy rough sets: The forgotten Step," IEEE Transaction on Fuzzy Systems, vol.15, no. 1, pp. 121-130. 2007.
[31] A. Tan , S. Shi, W. Wu, J. Li, and W. Pedrycz, "Granularity and Entropy of Intuitionistic Fuzzy Information and Their Applications," IEEE Trans. Cybernetics, Digital Object Identifier 10.1109/TCYB.2020.2973379.
[32] A. H. Tan, W.-Z. Wu, Y. H. Qian, J. Y. Liang, J. K. Chen, J. J. Li, "Intuitionistic fuzzy rough set-based granular structures and attribute subset selection," IEEE Transactions on Fuzzy Systems, vol. 27, no.3, pp.527-539, 2019.
[33] S. Zhao, H. Chen, C. Li, X. Du, H. Sun, "A novel approach to building a robust fuzzy rough classifier," IEEE Transactions on Fuzzy Systems, vol. 23, no.4, pp. 769–786, 2015.
[34] Eric C. C. Tsang, J. Song, D. Chen, X. Yang, "Order based hierarchies on hesitant fuzzy approximation space," International Journal of Machine Learning & Cybernetics, vol. 10, no. 6, pp. 1407-1422, 2019
[35] R. Jensen, Q. Shen, "Fuzzy-rough attributes reduction with application to web categorization," Fuzzy Sets and systems, vol. 141, pp. 469–485, 2004.
[36] R. B. Bhatt, M. Gopal, "On fuzzy-rough sets approach to feature selection," Pattern Recognition Letters, vol. 26, no.7, pp. 965-975, 2005.
[37] D. Chen, L. Zhang, S. Zhao, Q. Hu, P. Zhu, "A novel algorithm for finding reducts with fuzzy rough sets," IEEE Transaction on Fuzzy Systems, vol. 20, no.2, pp. 385–389, 2012.
[38] J. H. Dai, H. Hu, W. -Z. Wu, Y. H. Qian, D. B. Huang, "Maximal discernibility pair-based approach to attribute reduction in fuzzy rough sets," IEEE Transactions on Fuzzy Systems, vol. 26, no.4, pp. 2174-2187, 2018.
[39] A. Mieszkowicz-Rolka, L. Rolka, "Variable precision fuzzy rough sets," in: Transactions on Rough sets 1, LNCS-3100, Springer, Berlin, Cermany, 2004, pp. 144–160.
[40] S. Zhao, E. C. C. Tsang, and D. Chen, "The model of fuzzy variable precision rough sets," IEEE Transactions on Fuzzy Systems, vol. 17, no.2, pp. 451–467, 2009.
[41] C. Wang, Y. Huang, M. Shao, X. Fan, "Fuzzy rough set-based attribute reduction using distance measures," Knowledge-Based Systems, vol. 164, pp. 205–212, 2019.
[42] C. Wang, Y. Qi, M. Shao, Q. Hu, D. Chen, Y. Qian, Y. Lin, "A fitting model for feature selection with fuzzy rough sets," IEEE Transaction on Fuzzy Systems, vol. 25, no.4, pp.741–753, 2016.
[43] M. Friedman, "A comparison of alternative tests of significance for the problem of m ranking," Ann. Math. Statist., vol. 11, pp. 86–92, 1940.
[44] J. Dunn, "Multiple comparisons among means," Journal of the American Statistical Association, vol. 56, pp. 52-64, 1961.
[45] J. Demsar, "Statistical comparison of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.
[46] P. Maji, P. Garai, "On fuzzy-rough attribute selection: Criteria of Max-Dependency, Max-Relevance, Min-Redundancy, and Max-Significance," Applied Soft Computing, vol.13, pp.3968-3980, 2013.
[47] P. Maji, S. Paul, "Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data," International Journal of Approximate Reasoning, vol.52, pp.408-426, 2011.

Changzhong Wang received the M.S. degree from Bohai University, Jinzhou, China, the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2005, and 2008 respectively. He is currently a Professor with Bohai University.

His research interests are focused on fuzzy sets, rough sets, data mining, pattern recognition and statistical analysis.

He has authored or coauthored more than 50 journal and conference papers in the areas of machine learning, data mining, and rough set theory.

Yuhua Qian is a Professor of Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He received the M.S. degree and the PhD degree in Computers with applications at Shanxi University in 2005 and 2011, respectively.

He is actively pursuing research in pattern recognition, feature selection, rough set theory, granular computing and artificial intelligence. He has published more than 50 articles on these topics in international journals.

Weiping Ding received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. He was a Postdoctoral Fellow in the Brain Research Center, National Chiao Tung University, Hsinchu, Taiwan, in 2014.

He is a member of ACM and CCF. He has authored more than 60 papers in journals and conference proceedings. His current research interests include machine learning, data mining and their applications in big data. He serves as an Associate Editor of IEEE Transaction on fuzzy systems and Information Sciences.

Xiaodong Fan received the M.S. degree from Harbin University of Commerce in 2000, the M.S. degree in Mathematics from Guizhou University in 2007 and the Ph.D. degrees in Mathematics from Beijing University of Technology in 2013, respectively. He is currently an associate Professor with Bohai University. His research interests are focused on machine learning and optimization.