A Fitting Model for Feature Selection with Fuzzy Rough Sets

Changzhong Wang, Yali Qi, Mingwen Shao, Qinghua Hu, Degang Chen, Yuhua Qian, and Yaojin Lin

***Abstract*—Fuzzy rough set is an important rough set model used for feature selection. It uses the fuzzy rough dependency as a criterion for feature selection. However, this model can merely maintain a maximal dependency function. It does not fit a given data set well and cannot ideally describe the differences in sample classification. Therefore, in this study, we introduce a new model for handling this problem. First, we define the fuzzy decision of a sample using the concept of fuzzy neighborhood. Then, a parameterized fuzzy relation is introduced to characterize the fuzzy information granules, using which the fuzzy lower and upper approximations of a decision are reconstructed and a new fuzzy rough set model is introduced. This can guarantee that the membership degree of a sample to its own category reaches the maximal value. Furthermore, this approach can fit a given data set and effectively prevents samples from being misclassified. Finally, we define the significance measure of a candidate attribute and design a greedy forward algorithm for feature selection. Twelve data sets selected from public data sources are used to compare the proposed algorithm with certain existing algorithms, and the experimental results show that the proposed reduction algorithm is more effective than classical fuzzy rough sets, especially for those data sets for which different categories exhibit a large degree of overlap.**

***Index Terms*—Dependency function, Fuzzy rough set, Fuzzy similarity relation, Feature selection.**

I. Introduction

With the development of computer and database technology, a large number of attributes can be acquired and stored in databases for several real-world applications. Some of the attributes may be irrelevant or redundant for classification learning; they may greatly reduce the performance of classifiers and lead to a high degree of computational complexity. Therefore, before using a data set, it is necessary to preprocess the data to remove redundant features. Feature selection or attribute reduction, an important technique for reducing the number of redundant features, is used to find an optimal feature subset for performing classification under the premise of maintaining classification accuracy. In recent years, feature selection has been widely used in data processing, pattern recognition, and machine learning [12], [21], [27], [50]–[59].

There are two main problems related to feature selection that must be solved: one is the construction of a feature evaluation function, the other is the application of a strategy to the search for optimal features. A feature evaluation function is used to measure the quality of a candidate subset. This is related to the classification ability of a feature subset. In general, a good feature evaluation function can lead to high classification accuracy. The search strategy involves finding an optimal feature subset according to a certain evaluation function. This includes sequential forward search and backward selection algorithms. Greedy searching [20], [23], genetic algorithms [32], [34], and branch and bound [33], [47] are well-known search strategies. The feature evaluation plays a very important role in feature selection. To determine the relevance between a decision and features, many feature evaluation functions have been investigated in recent years. Distance [26], [45], correlation [11], consistency [6], and mutual information [25], [37], [48] are usually regarded as being feasible feature evaluation functions.

The classical rough set model, introduced by Pawlak [35], has been successfully used as a feature selection tool [22], [24], [28], [39], [46]. In this model, a crisp equivalence relation and crisp equivalence classes are used to characterize the dependency function between decision and condition attributes. The dependency function is used to determine the relevance between the decision and conditional attributes and to evaluate the classification ability of the attributes. For a given data set, it is possible to find a minimal subset of the conditional attributes that are the most informative using the dependency function. However, numerical data sets must be discretized before attribute reduction. Unfortunately, discretization greatly reduces the difference between the attribute values and leads to information loss [15], [17].

The combination of rough sets and fuzzy sets, as proposed by Dubois and Prade [7] gives rise to the notion of fuzzy rough sets. This provides an effective means of overcoming the problem of discretization and can be directly applied to the reduction of numerical or continuous attributes [5], [29], [36], [38], [57]. In the framework of fuzzy rough sets, a fuzzy similarity relation is defined by real-valued conditional attributes and is employed to

Manuscript received August 28, 2015; revised October 17, 2015, January 09, 2016 and March 22, 2016; accepted May 02, 2016. This work was supported by the National Natural Science Foundation of China under Grants 61572082, 61473111, 61363056, 61303131, the Natural Science Foundation of Liaoning Province (2014020142).

C. Z. Wang and Y. L. Qi are with the Department of Mathematics, Bohai University, Jinzhou 121000, China (e-mail: changzhongwang@ 126.com).

M. W. Shao is with the College of Computer and Communication Engineering, Chinese University of Petroleum, Qingdao 266580 (e-mail: smw278@126.com).

Q. H. Hu is with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: huqinghua@hit.edu.cn).

D. G. Chen is with the Department of Mathematics & Physics, North China Electric Power University, Beijing 102206 (e-mail: chengdegang@263.net).

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

Y. J. Lin is with the School of Computer Science, Minnan Normal University, Zhangzhou 363000, P.R. China (yjlin@mnnu.edu.cn).

measure the similarity between samples. The fuzzy upper and lower approximations of a decision are then defined based on a fuzzy similarity relation. Numerical attribute values are no longer needed for discretization before attribute reduction. Rather, they are converted to the corresponding membership degrees of samples to the upper and lower approximations. As fuzziness is introduced into the rough set theory, more classification information related to the continuous attributes is easily held. In recent years, fuzzy rough sets have attracted considerable attention [30], [31], [38], [40], [44], [54], [60]. Jensen and Shen first introduced the concept of a dependency function in a classical rough set model into the fuzzy case and proposed an attribute reduction algorithm based on fuzzy rough sets [16]. Some research efforts into attribute reduction with fuzzy rough sets mainly aimed at improving the pioneering algorithm. In [1], the authors presented the concept of a computational domain to improve the computational efficiency of the algorithm in [16]. In [17]–[19], the algorithm based on the fuzzy rough sets was further improved and developed with several heuristic algorithms to find optimal reducts. Meanwhile, the classical fuzzy rough set model was also improved to analyze the noisy data [3], [13], [41], [58]. In 2004, a model based on variable-precision fuzzy rough sets was introduced in [41], where the fuzzy memberships of a sample to the lower and upper approximations were computed with fuzzy inclusion. In 2009, Zhao et al. constructed a new model, called fuzzy variable-precision rough sets, to handle noise of misclassification and perturbation [58]. Another class of attribute reduction algorithm with fuzzy rough sets is based on a discernibility matrix. Skowron and Rauszer first presented an attribute reduction method based on a discernibility matrix in the context of Pawlak's rough sets [46]. Chen et al. extended the idea to fuzzy rough sets and proposed the concept of a fuzzy discernibility matrix for application to attribute reduction [3], [4], [49]. In addition, Hu et al. introduced information entropy to fuzzy rough sets to measure the dependency between conditional attributes and decisions [14], and applied the proposed measure to calculate the uncertainty in the fuzzy-rough approximation space [15].

The basic idea of the fuzzy rough model is that a fuzzy similarity relation is used to construct the fuzzy lower and upper approximations of a decision. The sizes of the lower and upper approximations reflect the discriminating capability of a feature subset. The union of fuzzy lower approximations forms the fuzzy positive region of decision. As the membership degree of a sample to the positive region increases, the possibility of it belonging to some category also increases. The fuzzy dependency is defined as the ratio of the sizes of the positive region over all the samples in the feature space. It is used to evaluate the significance of a subset of features. When a candidate feature is added to the existing feature pool and produces the greatest significance increment, the candidate feature will be regarded as being optimal and is therefore adopted into the feature pool. However, the definition of the classical fuzzy positive region cannot accurately reflect the classification ability of a subset of features because it only maintains the maximal dependency function and cannot guarantee the maximal membership degree of a sample to its own category. According to the definition of the fuzzy lower approximation, a training sample belonging to class *A* can be classified into class *B* by using the maximum membership principle. Thus, the classical fuzzy rough set model does not fit the training data set well and cannot ideally describe the differences in the sample classification. In particular, when different categories of a data set exhibit a large degree of overlap, it easily results in the samples being misclassified. This issue will be discussed in detail in Section 3.

In this study, we introduce a new fuzzy rough set model. It can fit a given data set and guarantee the maximal membership degree of a sample to its own category. It provides an effective means of preventing the misclassification of the training samples. First, we define the fuzzy decision of a sample by using the concept of fuzzy neighborhood. A parameterized fuzzy relation is introduced to characterize fuzzy information granules for the analysis of real-valued data sets, and a new fuzzy dependency function is proposed. Then, we define the significance measure of a candidate attribute, and present a greedy forward algorithm for attribute reduction. Finally, we compare the proposed algorithm with existing algorithms. The experimental results show that the proposed reduction algorithm is more feasible and effective, especially for those data sets for which different categories exhibit a large degree of overlap.

This paper is organized as follows. In Section 2, we review some basic concepts related to classical fuzzy rough sets and introduce the rough approximation of the fuzzy decision of samples. In Section 3, we develop a fitting fuzzy rough set model. In Section 4, we present a heuristic algorithm for feature selection. In Section 5, we verify the feasibility and stability of the proposed algorithm. Section 6 concludes the paper.

II. Rough Approximations Of fuzzy Decision

Suppose U is a universe and $\tilde{A}(\cdot): U \to [0,1]$ is a mapping function, $\tilde{A}$ is then called a fuzzy set on U; for any $x \in U$, $\tilde{A}(x)$ is called the membership degree of x to $\tilde{A}$. The class of all fuzzy sets on U is denoted as $F(U)$. Obviously, a crisp set is a special fuzzy set.

The domain of a membership function is $[0,1]$. As the value of $\tilde{A}(x)$ approaches 1, the degree of x belonging to $\tilde{A}$ increasees. As an upper bound, $\tilde{A}(x)=1$, it indicates that x completely belongs to $\tilde{A}$; otherwise, as the value of $\tilde{A}(x)$ approaches 0, the degree of x belonging to $\tilde{A}$ falls. As a lower bound, $\tilde{A}(x)=0$, it indicates that x does not belong to $\tilde{A}$ at all.

Let B be a subset of real-valued attributes describing an object set $U=\{x_1, x_2, \cdots, x_n\}$, on which these attributes can induce a fuzzy binary relation R_B. We say that R_B is a fuzzy similarity relation if it satisfies

(1) Reflexivity: $R_B(x,x)=1$, $\forall x \in U$;

(2) Symmetry: $R_B(x,y)=R_B(y,x)$, $\forall x, y \in U$;

The fuzzy similarity class $[x]_B$ associated with x and R_B is a fuzzy set on U. It is also called the fuzzy neighborhood of x, i.e., $[x]_B(y)=R_B(x,y)$, $y \in U$.

Assume that D is a decision attribute on U and partitions the sample set U into r crisp equivalence classes $U/D=\{D_1,D_2,\cdots D_r\}$. In the following, we introduce the concept of the fuzzy decision of a sample.

Definition 1. Let U be a universe and $\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_r\}$ be a family of fuzzy sets on U, then $\{\tilde{D}_1,\tilde{D}_2,\cdots,\tilde{D}_r\}$ is called a fuzzy partition if they satisfy $\sum_{i=1}^{r}\tilde{D}_i(x)=1,\forall x\in U$.

Definition 2. Assume that D is a decision attribute and $U/D=\{D_1,D_2,\cdots,D_r\}$, R_B is a fuzzy similarity relation on U induced by B. $\forall x\in U$, the fuzzy decision of x, is defined as follows.

$$\tilde{D}_i(x)=\frac{|[x]_B\cap D_i|}{|[x]_B|},\quad i=1,2,\ldots,r,\quad x\in U,$$

where $\tilde{D}_i$ is a fuzzy set and $\tilde{D}_i(x)$ indicates the membership degree of x to D_i. We call $\{\tilde{D}_1,\tilde{D}_2,\cdots\tilde{D}_r\}$ the fuzzy decision of samples induced by the decision D and the attribute subset B. Obviously, $\{\tilde{D}_1,\tilde{D}_2,\cdots\tilde{D}_r\}$ is a fuzzy partition on U.

Based on fuzzy similarity relations and the fuzzy decision of samples, fuzzy rough sets were introduced as follows.

Let $U=\{x_1,x_2,\cdots,x_n\}$ be a set of objects, AT be a set of real-valued attributes, D be a decision attribute defined on U, $B\subseteq AT$. Assume that a decision D partitions the objects into r crisp equivalence classes $U/D=\{D_1,D_2,\cdots,D_r\}$. R_B is a fuzzy similarity relation induced by B, $\{\tilde{D}_1,\tilde{D}_2,\cdots\tilde{D}_r\}$ is the fuzzy decision of samples induced by D and AT. Given a decision equivalence class $D_i\in\{D_1,D_2,\cdots,D_r\}$, the fuzzy lower and upper approximations are defined as follows, respectively.

$$\underline{B}D_i(x)=\inf_{y\in U}\max\{1-R_B(x,y),\tilde{D}_i(y)\},$$
$$\overline{B}D_i(x)=\max_{y\in U}\inf\{R_B(x,y),\tilde{D}_i(y)\}.$$

The membership of an object $x\in U$ to the fuzzy positive region is given by

$$POS_B(D)(x)=\bigcup_{i=1}^{r}\underline{B}D_i(x).$$

With the definition of fuzzy positive region, one can compute the fuzzy dependency function by using the following formula:

$$\gamma_B(D)=\frac{\sum_{x\in U}POS_B(D)(x)}{|U|}.$$

The fuzzy dependency is defined as the ratio of the sizes of the positive region over all the samples in the feature space. It is used to evaluate the significance of a subset of features. However, the definition of the fuzzy positive region can only maintain the maximal dependency function. It cannot guarantee the maximal membership degree of a sample to its own category. Therefore, such fuzzy dependency easily results in the training samples being misclassified and it cannot accurately reflect the classification ability of a subset of features.

III. Fitting Model Based On Fuzzy Rough Sets

The structure of a dataset used for classification learning can be written as a decision table and be denoted by $<U,A,D>$, where $U=\{x_1,x_2,\cdots,x_n\}$ is a nonempty set of samples, called a universe; $A=\{a_1,a_2,\cdots,a_m\}$ is a set of conditional attributes for characterizing the samples, and D is a decision attribute. In the following discussion, we assume that the universe is partitioned into r crisp equivalence classes by the decision D, denoted as $U/D=\{D_1,D_2,\cdots D_r\}$, and that the corresponding fuzzy decision of samples is induced by D and A and denoted as $\{\tilde{D}_1,\tilde{D}_2,\cdots\tilde{D}_r\}$. Let $B\subseteq A$, $a\in B$, and R_a be a fuzzy similarity relation induced by the attribute a, then $R_B=\bigcap_{a\in B}R_a$.

Fuzzy rough sets employ fuzzy similarity relations to construct a fuzzy rough dependency function at one level of granularity. In fact, more classification information can be acquired if a dependency function is constructed at different levels of granularity. We can study the influence of different granularities on the results of feature selection and select an optimal feature subset by adjusting the granularity [3], [13]. In addition, a multiple-granularity rough set model is also helpful for controlling the noise in data [41], [58]. The membership degree of one sample pair to a fuzzy similarity relation reflects the relationship tie of the sample pair. When the relationship tie of two samples is very weak, the membership degree of the sample pair can be regarded as being zero because this kind of weak relationship tie may be caused by data noise. To achieve this, a parameterized fuzzy information granule is constructed by introducing a parameter ε, as follows:

$$[x]_B^{\varepsilon}(y)=\begin{cases}0, & R_B(x,y)<\varepsilon,\\ R_B(x,y), & R_B(x,y)\geq\varepsilon,\end{cases}$$

where $\varepsilon\in[0,1)$. According to the above definition, we can easily see that the parameter ε influences the size of a fuzzy information granule. We refer to ε as the radius of the fuzzy neighborhood of a sample.

We can see that there are two factors impacting the membership degrees of a fuzzy similarity relation. One is parameter ε, the other is feature subset B. For a given parameter ε, the membership degrees become smaller as the number of features in B increases. Given a decision table $<U,A,D>$, $0<\varepsilon<1$, and $B\subseteq A$, R_B is the fuzzy similarity relation on U induced by ε and B. We denote it by R_B^{ε}. From the above discussion, we can derive the following properties.

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

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

Attribute reduction with the dependency function of classical fuzzy rough sets can only maintain the maximal fuzzy dependency. It does not fit a given data set well and cannot guarantee the maximal membership degree of a sample to its own category. An example is given as follows.

Example 1. A decision table $< U, A, D >$ is given in Table 1, where U is an object set and $U = \{x_1, x_2, x_3, x_4\}$, $A = \{a_1, a_2, a_3, a_4\}$ is a conditional attribute set, D is a decision attribute.

TABLE 1
DECISION TABLE

U	a_1	a_2	a_3	a_4	D
x_1	3	5	3	6	1
x_2	4	3	2	7	1
x_3	0	1	4	2	2
x_4	9	3	1	3	3

First, all the numerical attributes are normalized into the interval [0, 1]. Then, we use the following formula to compute the fuzzy similarity degree r_{ij} between objects x_i and x_j with respect to the attribute set A.

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

Thus, we obtain

$$R_A = \begin{bmatrix} 1 & 0.95 & 0.8 & 0.85 \\ 0.95 & 1 & 0.75 & 0.8 \\ 0.8 & 0.75 & 1 & 0.95 \\ 0.85 & 0.8 & 0.95 & 1 \end{bmatrix}.$$

The decision attribute D partitions objects U into three parts, that is, $U/D = \{D_1, D_2, D_3\}$, where $D_1 = \{x_1, x_2\}$, $D_1 = \{x_3\}$, $D_1 = \{x_4\}$.

$\forall x \in U$, we can compute the fuzzy decision of x using the following formula:

$$\tilde{D}_i(x) = \frac{|[x]_A \cap D_i|}{|[x]_A|},\ i = 1, 2, \ldots, r\ .$$

Therefore, we can obtain the fuzzy decision matrix of objects, as follows:

$$\tilde{D} = \left[\tilde{D}_1, \tilde{D}_2, \tilde{D}_3\right] = \begin{bmatrix} 0.54 & 0.22 & 0.24 \\ 0.56 & 0.21 & 0.23 \\ 0.44 & 0.29 & 0.27 \\ 0.46 & 0.26 & 0.28 \end{bmatrix},$$

where $\tilde{D}_1 = [0.54, 0.56, 0.44, 0.46]^T$, $\tilde{D}_2 = [0.22, 0.21, 0.29, 0.26]^T$, $\tilde{D}_3 = [0.24, 0.0.23, 0.0.27, 0.0.28]^T$. Here, T stands for the transpose operation of matrix. This gives:

$$\begin{aligned} \underline{R}(D_1)(x_1) &= \inf_{y \in U} \max\left\{1 - R_A(x_1, y), \tilde{D}_1(y)\right\} \\ &= \inf_{y \in U}(0.54, 0.56, 0.44, 0.46) \\ &= 0.44. \end{aligned}$$

Similarly, we have that $\underline{R}(D_1)(x_2) = 0.44$, $\underline{R}(D_1)(x_3) = 0.44$, $\underline{R}(D_1)(x_4) = 0.44$; $\underline{R}(D_2)(x_1) = 0.21$, $\underline{R}(D_2)(x_2) = 0.21$, $\underline{R}(D_2)(x_3) = 0.22$, $\underline{R}(D_2)(x_4) = 0.21$; $\underline{R}(D_3)(x_1) = 0.23$, $\underline{R}(D_3)(x_2) = 0.23$, $\underline{R}(D_3)(x_3) = 0.24$, $\underline{R}(D_3)(x_4) = 0.23$.

TABLE 2
LOWER APPROXIMATION OF THE DECISION

U	$\underline{R}(D_1)$	$\underline{R}(D_2)$	$\underline{R}(D_3)$
$x1$	0.44	0.21	0.23
$x2$	0.44	0.21	0.23
$x3$	0.44	0.22	0.24
$x4$	0.44	0.21	0.23

From Table 1, we can see that samples x_3 and x_4 belong to the second and third categories, respectively. However, both are grouped into the first category according to the lower approximations of decision and the maximum membership decision principle. This shows that the classical fuzzy rough set model does not fit a given data set well and cannot ideally describe differences in the sample classifications. To overcome this problem, the following definition is given. This definition can guarantee the maximal membership degree of a sample to its own category and fit a given data set well.

Definition 1. Given a decision table $< U, A, D >$, $U/D = \{D_1, D_2, \cdots D_r\}$, $\{\tilde{D}_1, \tilde{D}_2, \cdots \tilde{D}_r\}$ is the fuzzy decision of samples induced by A and D, $0 < \varepsilon < 1$, and $B \subseteq A$. R_B^ε is the fuzzy similarity relation on U induced by ε and B, the lower and upper approximations of decision D with respect to B are defined as:

$$\underline{R_B^\varepsilon}(D_i)(y) = \begin{cases} \min\limits_{x \in U} \max\left\{1 - R_B^\varepsilon(x, y), \tilde{D}_i(x)\right\}, & y \in D_i; \\ 0, & y \notin D_i. \end{cases}$$

$$\overline{R_B^\varepsilon}(D_i)(y) = \begin{cases} \max\limits_{x \in U} \min\left\{R_B^\varepsilon(x, y), \tilde{D}_i(x)\right\}, & y \in D_i; \\ 0, & y \notin D_i. \end{cases}$$

The fuzzy lower approximation of decision equivalence class D_i is also called the fuzzy positive region of D_i.

In the same way as in the case of classical rough sets, fuzzy rough sets employ the fuzzy neighborhood and fuzzy decision of an object to determine its membership degree to any one decision class. $\underline{R_B}(D_i)(y)$ indicates the membership degree of an object y certainly belonging to class i. $\overline{R}_B(D_i)(y)$ represents the membership degree of an object y possibly belonging to class i.

There are two main differences between the proposed model and classical fuzzy rough sets.

(1) For a given decision equivalence class D_i and a sample $y \in U$, according to the classical fuzzy rough sets, the membership degree of y to the fuzzy lower approximation of D_i is computed by $\underline{R_B^\varepsilon}(D_i)(y) = \min\limits_{x \in U} \max\left\{1 - R_B^\varepsilon(x, y), \tilde{D}_i(x)\right\}$ regardless of the category to which the sample belongs. This easily results in the samples being misclassified when using the maximum membership principle because the model cannot ensure the maximal membership degree of a sample to its own category. Thus, the classical model cannot fit the given data set or satisfactorily describe the differences in sample classification. In contrast, the proposed model considers two different cases when it computes the membership degree. This overcomes the

drawbacks with the classical model and results in the data fitting well.

(2) The proposed model is somewhat less complex than the classical model. For any sample, the classical model computes the membership degree of the sample for each decision equivalence class, while the proposed model only considers the decision equivalence class to which the sample belongs.

Definition 2. Given a decision table $<U, A, D>$, $0<\varepsilon<1$, $B \subseteq A$, and $U/D=\{D_1, D_2, \cdots D_r\}$, $\{\tilde{D}_1, \tilde{D}_2, \cdots \tilde{D}_r\}$ is the fuzzy decision for the samples induced by A and D. The fuzzy positive region of decision D with respect to B is defined as $POS_B^{\varepsilon}(D)=\sum_{i=1}^{r}\underline{R_B^{\varepsilon}}(D_i)$.

The fuzzy positive region is defined as the sum of the lower approximations of the decision equivalence classes. The advantage of the new fuzzy rough set model lies in the fact that, as the membership degree of a sample to the positive region increases, so too does the possibility of it belonging to its own category. Intuitively, those samples with large membership degrees are easily classified into their respective decision equivalence class. The size of the positive region reflects the classification ability of the condition attributes. Classification tasks in different feature subspaces have different fuzzy positive regions. One usually attempts to find a feature subset in which a classification task has a great positive region. The significance of a feature subset can be described by the fuzzy positive region and is formally defined as follows.

Definition 3. Given a decision table $<U, A, D>$, $0<\varepsilon<1$, $B \subseteq A$, and $U/D=\{D_1, D_2, \cdots D_r\}$, the dependency degree of decision D to B is defined as

$$\partial_B^{\varepsilon}(D)=\frac{\sum_{x_i \in U} POS_B^{\varepsilon}(D)(x_i)}{|U|}=\frac{\sum_{x_i \in U}\sum_{i=1}^{r}\underline{R_B^{\varepsilon}}(D_i)(x_i)}{|U|}.$$

The dependency degree is also called the fuzzy dependency function. Obviously, $0 \le \partial_B^{\varepsilon}(D) \le 1$. This reflects the classification power of a conditional attribute subset. We can say that decision D is completely dependent on B if $\partial_B^{\varepsilon}(D)=1$; otherwise, D depends on B in the degree of $\partial_B^{\varepsilon}(D)$. Each sample in B belongs entirely to its own decision equivalence class when $\partial_B^{\varepsilon}(D)=1$. Intuitively, we hope to find a feature subspace in which the value of the fuzzy dependency function is a maximum because the error rate of classification is smaller in such a case.

Theorem 1. Given a decision table $<U, A, D>$ and $0<\varepsilon<1$, if $B_1 \subseteq B_2 \subseteq A$, then $POS_{B_1}^{\varepsilon}(D) \subseteq POS_{B_2}^{\varepsilon}(D)$.

Proof. As $B_1 \subseteq B_2$, we have $R_{B_2}^{\varepsilon} \subseteq R_{B_1}^{\varepsilon}$ according to Proposition 1. For any $y \in D_i$ and $x \in U$, we have $1-R_{B_2}^{\varepsilon}(x,y) \ge 1-R_{B_1}^{\varepsilon}(x,y)$. From the definition of lower approximation, it follows that $\underline{R_{B_1}^{\varepsilon}}(D_i)(y) \le \underline{R_{B_2}^{\varepsilon}}(D_i)(y)$. Hence, $POS_{B_1}^{\varepsilon}(D) \subseteq POS_{B_2}^{\varepsilon}(D)$.

Theorem 2. Given a decision table $<U, A, D>$ and $0<\varepsilon<1$, if $\varepsilon_1 < \varepsilon_2$, then $POS_B^{\varepsilon_1}(D) \subseteq POS_B^{\varepsilon_2}(D)$.

Proof. As $\varepsilon_1 \le \varepsilon_2$, we have $R_B^{\varepsilon_2} \subseteq R_B^{\varepsilon_1}$ according to Proposition 2. For any $y \in D_i$ and $x \in U$, we have $1-R_B^{\varepsilon_2}(x,y) \ge 1-R_B^{\varepsilon_1}(x,y)$. Given the definition of lower approximation, it follows that $\underline{R_B^{\varepsilon_1}}(D_i)(y) \le \underline{R_B^{\varepsilon_2}}(D_i)(y)$. Hence, $POS_B^{\varepsilon_1}(D) \subseteq POS_B^{\varepsilon_2}(D)$.

Obviously, the dependency function has the following properties.

Theorem 3. Given a decision table $<U, A, D>$ and $0<\varepsilon<1$, if $B_1 \subseteq B_2 \subseteq \cdots \subseteq B_m \subseteq A$, then $\partial_{B_1}^{\varepsilon}(D) \le \partial_{B_2}^{\varepsilon}(D) \le \cdots \le \partial_{B_m}^{\varepsilon}(D)$.

Theorem 4. Given a decision table $<U, A, D>$ and $0<\varepsilon_i<1$, if $\varepsilon_1 \le \varepsilon_2 \le \cdots \le \varepsilon_m$, then $\partial_B^{\varepsilon_1}(D) \le \partial_B^{\varepsilon_2}(D) \le \cdots \le \partial_B^{\varepsilon_m}(D)$.

The above theorems show that the dependency function monotonically increases with the size of the attribute subset and the parameter ε. Such property is very important for designing a forward algorithm because it guarantees that adding a candidate feature to the existing feature subset will not decrease the dependency of the new subset. When such a dependency function is employed as a criterion for feature selection, the step for search stop should be easily implemented.

Definition 4. Given a decision table $<U, A, D>$, $0<\varepsilon<1$, and $B \subseteq A$, for any $a \in B$, if $\partial_{B-a}^{\varepsilon}(D) \ne \partial_B^{\varepsilon}(D)$, we say attribute a is indispensable in B. Otherwise, we say a is redundant or superfluous in B. If any attribute a in B is indispensable, we say B is independent.

If an attribute is redundant, it can be removed because the dependency does not change. A redundant attribute does not provide more classification information. Rather, it confuses the learning algorithm during training. Therefore, it must be deleted from the condition attribute set before classification learning.

Definition 5. Given a decision table $<U, A, D>$, $0<\varepsilon<1$, and $B \subseteq A$, we say B is a reduct of A if it satisfies

(1) $\forall a \in B$, $\partial_{B-a}^{\varepsilon}(D) < \partial_B^{\varepsilon}(D)$; (2) $\partial_A^{\varepsilon}(D) = \partial_B^{\varepsilon}(D)$.

According to this definition, a reduct of A is a minimal subset of those attributes that have the same classification ability as the whole set of attributes

IV. Attribute Reduction Algorithm Based On Fitting Fuzzy Rough Set Model

In this section, we first present the definition of significance measure of a candidate feature, then propose a greedy forward algorithm for attribute reduction and discuss its time complexity.

Definition 6. Given a decision table $<U, A, D>$, $B \subseteq A$, and $a \in A-B$, the significance of a with respect to B is defined as $SIG^{\varepsilon}(a, B, D)=\partial_{B\cup\{a\}}^{\varepsilon}(D)-\partial_B^{\varepsilon}(D)$.

This definition is used to compute the increment of the classification ability that is introduced by an attribute. It can be used as the significance measure of a candidate feature. With the proposed measure, we construct a heuristic algorithm for attribute reduction as follows. The algorithm starts with an empty set, and adds one attribute with the greatest significance into a feature pool at each iteration until the value of the dependency does not increase further.

To make full use of the classification information contained in the data and demonstrate the performance of the proposed attribute reduction method, we introduce another parameter $\lambda \in [0,1)$ to control the fuzzy decision of the samples. Given a decision table $< U, A, D >$, assume that decision D partitions the objects into r crisp equivalence classes $U/D = \{D_1, D_2, \cdots, D_r\}$, R_A is a fuzzy similarity relation on U induced by A. $\forall x \in U$, the fuzzy decision $\tilde{D}_i(x)$ of x is computed by

$$\tilde{D}_i(x) = \frac{\left|[x]_A^\lambda \cap D_i\right|}{\left|[x]_A^\lambda\right|},\ i = 1, 2 \cdots r,$$

where $[x]_A^\lambda$ is a parameterized fuzzy set and is defined as:

$$[x]_A^\lambda(y) = \begin{cases} 0, & R_A(x,y) < \lambda, \\ R_A(x,y), & R_A(x,y) \geq \lambda. \end{cases}$$

We call λ the radius of the fuzzy neighborhood of D.

Algorithm: Heuristic algorithm based on fitting fuzzy rough sets (NFRS)

Input: Decision table $< U, A, D >$, thresholds ε and λ // ε is the threshold for the fuzzy neighborhood of a sample. similarity. λ is the threshold for the fuzzy neighborood of decision D.

Output: One reduct *red*.

1: $\forall a \in A$: compute the relation matrix R_a;
2: Compute the fuzzy decision $\tilde{D} = \{\tilde{D}_1, \tilde{D}_2, \cdots \tilde{D}_r\}$;
3: Initialize: $red = \varnothing$, $B = A - red$, start = 1; // *red* is the pool containing the selected attributes and B is for the left attributes.
4: while start
5: $T \leftarrow \varnothing$
6: for each $a_i \in B$
7: $T \leftarrow red \cup \{a_i\}$;
8: Compute fuzzy similarity relation R_T^ε.
9: for each $x_j \in U$, suppose $x_j \in D_i$;
10: Compute fuzzy lower approximation $\underline{R_T^\varepsilon}(D_i)(x_j)$.
11: end for
12: $\partial_{red \cup a_i}^\varepsilon(D)$ =sum(($\max_{D_i \in U/D} \underline{R_T^\varepsilon}(D_i)$)/n;
15: end for
16: Find attribute a_k with maximum value $\partial_{red \cup a_k}^\varepsilon(D)$.
17: Compute $SIG^\varepsilon(a_k, red, D) = \partial_{red \cup a_k}^\varepsilon(D) - \partial_{red}^\varepsilon(D)$.
18: if $SIG^\varepsilon(a_k, red, D) > 0$
19: $red \leftarrow red \cup a_k$;
20: $B \leftarrow B - red$;
21: else
22: start=0;
23: end if
24: end while
25: return *red*

As described above, this algorithm terminates when the addition of any remaining attribute does not increase the dependency $\partial_B^\varepsilon(D)$. If there are N condition attributes, the time complexity for computing the fuzzy similarity relations is N, and the worst search time for a reduct will result in $N \times N$ evaluations of the dependency function. The overall time complexity of the algorithm is $O(N^2)$.

V. Experimental Analysis

In this section, we compare the proposed algorithm (NFRS) with existing attribute reduction algorithms. The existing algorithms are the classical rough set based algorithm (RS)[35], fuzzy information entropy based algorithm (FISEN)[14], [15] and algorithm of fuzzy rough dependency constructed from the intersection operations of fuzzy similarity relations (FRSA) [16]. These algorithms employ a sequential forward-search strategy to identify the optimal features. The experimental comparison is conducted based on a 10-fold cross-validation. That is to say, the original data set is randomly divided into ten subsets, of which nine are used as the training data and the remaining one is used for testing. Feature selection is performed on the training set; the reduced training set and test set are then sent to a classifier to attain the classification accuracy. After 10 rounds, the average value and variation of the classification accuracies are computed as the final performance. In the experiments, three indices, including the number of selected features, classification performance, and running time, are used in the comparison. All the algorithms are performed in Matlab 2013b and run in a hardware environment with a Intel (R) Core (TM) i7-4790 CPU @ 3.60 GHz, with 16.0 GB RAM.

TABLE I
DESCRIPTION OF DATA SETS

No	Data sets	Sample	Attributes	Classes
1	Glass	214	10	6
2	Wdbc	569	30	2
3	Ionos	351	33	2
4	Sonar	208	60	2
5	Diabetes	768	8	2
6	Gearbox	1603	72	4
7	Segment	2310	19	7
8	Brain	90	5920	5
9	Breast	84	9217	5
10	AMLALL	72	7129	2
11	Prostate1	136	12600	2
12	Prostate2	102	10509	2

Two classical classifiers are employed to evaluate the classification accuracies of the original and reduced data. They are the support vector machine (RBF-SVM) and k-nearest neighbor rule (K－ NN, K = 3). Twelve data sets are used in the experimental analysis, some are selected from the UCI Machine Learning Repository [2], the others are downloaded at Keng Ridge Bio-medical (KRBM) Data Set Repository [61]. The information contained in these data sets is outlined in Table I. All of the numerical attributes are first normalized into the interval [0, 1]. The fuzzy similarity degree r_{ij} between objects x_i and x_j with respect to an attribute is computed by

$$r_{ij} = \begin{cases} 1-|x_i - x_j|, & |x_i - x_j| \le 1-\varepsilon; \\ 0, & |x_i - x_j| > 1-\varepsilon. \end{cases}$$

As the classical rough set considers only categorical data, it is necessary to preprocess numerical data by a discretization algorithm such as equal scale, equal frequency, maximum entropy, and so forth [15], [42], [43]. In general, different discretization methods may result in selecting different feature subsets. The problem has been intensively studied [42], [43]. The experimental results described in ref. [15] show that fuzzy C-means discretization offers a better level of performance than either equal scale or equal frequency. In the following numerical experiments, we employ a fuzzy C-means clustering (FCM) technique to discretize numerical data before attribute reduction with classical rough sets. The numeric attributes are discretized into four intervals. To compare the NFRS and FRSA algorithms, two parameters ε and λ are also introduced into the FRSA algorithm in the same way as the NFRS algorithm. They are used to control the fuzzy neighborhood and fuzzy decision of a sample, respectively. They have a great impact on different data sets. We set ε to a value between 0.1 and 0.5 in steps of 0.05 and λ to a value between 0.1 and 0.6 in steps of 0.1. As different learning algorithms may require different feature subsets to produce the best classification accuracy, all of the experimental results in the following tables are presented with the highest classification accuracy based on 10-fold cross-validation.

Table II presents the four algorithms for reducing the features of data sets. The average sizes of the feature subsets selected with FCM + RS are smaller than those selected with the other algorithms in most cases. The reason for this result may be due to the information loss caused by data discretization. The average subset sizes with FISEN, NFRS, and FRSA are roughly the same except for the Sonar and Brain data sets. However, the FCM + RS algorithm yields an empty set when it is applied to the Diabetes data set. This is because the dependency of each single feature is zero in the first loop of each training fold and the algorithm stops at the last fold. To obtain a feature subset for classification learning, we randomly select a feature in this case. Thus, the FCM + RS algorithm continues to run rather than stopping. Finally, a subset of the features is selected and the corresponding average subset size is marked by an asterisk, as shown in Table II.

TEBLE II
AVERAGE SIZES OF FEATURE SUBSETS

Data sets	Raw data	FCM+RS	FISEN	NFRS	FRSA
Glass	10	5.9	4.9	3.6	4.4
Wdbc	30	8.2	12.2	9.5	11.3
Ionos	33	9.0	7.8	10.5	10.8
Sonar	60	6.1	28.7	19.7	19.4
Diabetes	8	7.0*	5.1	6.1	5.2
Gearbox	72	7.7	9.1	9.2	9.2
Segment	19	12.4	9.7	8.9	8.8
Brain	5920	4.8	10.6	15.6	15.2
Breast	9217	4.2	12.6	10.0	10.3
AMLALL	7129	2.5	7.8	5.1	4.9
Prostate1	12600	6.6	9.5	8.2	9.3
Prostate2	10509	3.7	8.6	6.7	8.0
Average	3800.58	6.46	10.55	9.43	9.73

Tables III and IV present the comparative performance of the four reduction algorithms, where the underlined symbols indicate the highest classification accuracies among the reduced data sets. From the results listed in Tables III and IV, it is easy to see that the classification accuracies based on the NFRS and FISEN methods are comparable with but obviously higher than the other two methods in most cases. Out of the 24 cases, the NFRS and FISEN methods achieve the highest classification accuracies in 13 and 9 cases, respectively. The FRSA method obtains it in 3 cases, and FCM + RS attains it for 2 case. Most cases of the NFRS algorithm are higher than the FRSA algorithm. In particular, the accuracies of the Sonar, Diabetes and Prostate1 data sets have been greatly improved. This shows that different categories of these data sets exhibit a large degree of overlap. Because the FRSA algorithm cannot guarantee that the membership degree of a sample to its own category is maximal, the resulting classification accuracies of these data sets are relatively low. Thus, the NFRS algorithm is more effective than the FRSA algorithm. It should be pointed out that the accuracies of the FCM + RS algorithm are zeros when it is applied to the Diabetes data set. As mentioned above, regarding the data set, no feature is selected in the first loop. Therefore, their classification accuracies are zeros. As a feature is randomly selected in this case, we obtain the classification accuracies of the data set as shown in Tables III and IV, where the corresponding accuracy is also marked by an asterisk.

TABLE III
COMPARISION OF CLASSIFICATION ACCURACIES OF REDUCED DATA WITH SVM

Data sets	Raw data	FCM + RS	FISEN	NFRS	FRSA
Glass	93.22 ± 5.74	92.93 ± 6.86	94.84 ± 4.75	93.36 ± 5.88	92.38 ± 5.12
Wdbc	96.52 ± 2.05	96.13 ± 2.59	96.66 ± 2.38	97.19 ± 2.22	96.90 ± 1.45
Ionos	90.52 ± 4.65	93.16 ± 4.31	94.29 ± 3.57	94.22 ± 3.14	94.29 ± 3.31
Sonar	85.44 ± 7.88	75.09 ± 7.15	84.61 ± 7.71	87.62 ± 8.16	84.24 ± 7.72
Diabetes	76.16 ± 4.12	75.47 ± 4.48*	77.22 ± 4.79	76.43 ± 6.09	68.35 ± 3.98
Gearbox	98.88 ± 1.26	99.50 ± 0.39	99.50 ± 0.71	99.19 ± 1.51	98.63 ± 1.62
Segment	95.10 ± 1.17	95.84 ± 1.28	95.41 ± 2.08	95.19 ± 1.57	95.11 ± 1.54
Brain	66.67 ± 11.71	76.67 ± 11.77	82.22 ± 11.94	84.44 ± 10.54	81.11 ± 11.05
Breast	39.17 ± 12.67	73.33 ± 13.37	92.25 ± 5.49	91.37 ± 8.44	91.37 ± 9.22
AMLALL	65.24 ± 13.66	93.49 ± 8.71	97.14 ± 6.03	98.57 ± 6.03	97.46 ± 4.52
Prostate1	72.29 ± 11.97	87.86 ± 10.45	92.57 ± 4.85	93.57 ± 6.26	93.29 ± 5.39
Prostate2	61.17 ± 16.45	91.00 ± 7.38	92.33 ± 8.76	94.33 ± 7.87	94.17 ± 6.89
Average	78.36 ± 7.78	87.54 ± 6.75	91.59 ± 5.26	92.01 ± 5.64	90.61 ± 5.15

TABLE IV
COMPARISION OF CLASSIFICATION ACCURACIES OF REDUCED DATA WITH 3NN

Data sets	Raw data	FCM + RS	FISEN	NFRS	FRSA
Glass	90.80 ± 7.35	89.28 ± 8.04	94.44 ± 6.58	95.31 ± 2.26	93.16 ± 4.94
Wdbc	97.00 ± 2.24	94.20 ± 2.03	96.13 ± 3.27	97.36 ± 1.91	97.01 ± 1.87
Ionos	85.96 ± 5.38	88.88 ± 4.16	90.89 ± 3.98	92.02 ± 4.63	91.95 ± 5.41
Sonar	83.33 ± 8.06	73.71 ± 9.66	85.94 ± 7.99	84.29 ± 8.99	83.28 ± 6.35
Diabetes	74.00 ± 4.83	73.33 ± 4.56*	72.14 ± 5.38	73.56 ± 8.55	67.71 ± 5.13
Gearbox	99.69 ± 1.44	99.63 ± 0.32	99.69 ± 0.33	99.50 ± 1.07	98.81 ± 1.72
Segment	96.01 ± 1.16	96.23 ± 1.16	96.45 ± 1.18	95.76 ± 1.38	95.71 ± 1.89
Brain	86.67 ± 10.03	72.22 ± 10.21	82.03 ± 12.83	82.82 ± 10.73	81.91 ± 12.36
Breast	71.25 ± 13.24	70.50 ± 14.20	90.42 ± 8.84	92.67 ± 12.91	92.67 ± 12.91
AMLALL	85.71 ± 9.52	94.02 ± 8.46	94.60 ± 7.03	97.89 ± 6.45	98.76 ± 6.03
Prostate1	77.14 ± 15.36	85.43 ± 7.69	89.71 ± 7.61	89.00 ± 9.60	85.14 ± 8.44
Prostate2	84.33 ± 11.17	84.33 ± 12.78	93.33 ± 9.03	94.07 ± 6.69	93.33 ± 6.44
Average	85.99 ± 7.48	85.14 ± 7.16	90.48 ± 6.17	91.19 ± 6.26	89.95 ± 6.12

For SVM, NFRS outperforms all the raw data sets for classification tasks. At the same time, NFRS outperforms the raw data 8 times with respect to 3NN. Moreover, the average accuracies of NFRS are also comparable to any other feature selection algorithm in terms of SVM and 3NN learning algorithms.

In addition, we apply the Friedman test [10] and the Bonferroni–Dunn test [8] to show the statistical significance of the results. The Friedman statistic is defined as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} R_i^2 - \frac{k(k+1)^2}{4}\right) \text{ and } F_F = \frac{(\text{N-1})\chi_F^2}{N(k-1)-\chi_F^2},$$

where N is the number of data sets, k is the number of algorithms, and R_i is the average rank of algorithm i among all the data sets. F_F follows a Fisher distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. If the null hypothesis is rejected under the Friedman test statistic, a post-hoc test such as the Bonferroni–Dunn test can be used to further explore which algorithms are different in statistical terms. According to the results of this test, the performance of two algorithms is regarded as being significantly different if the distance of the average ranks exceeds the critical distance

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6N}},$$

where q_α is the critical tabulated value for this test [9].

To explore whether the classification performances of each classifier with the four feature selection algorithms are significantly different, we performed two Friedman tests. The null hypothesis of the Friedman test is that all of the algorithms are equivalent in terms of the classification performance. Tables V and VI indicate the rankings of the four feature selection algorithms under different classifiers. The values of F_F for different evaluation measures are listed in Table Ⅶ. The critical value of $F(3,33)$ for $\alpha = 0.1$ is 2.23.

From Table Ⅶ, we can reject the null hypothesis at $\alpha = 0.1$ and accept the alternative hypothesis that the four algorithms are different under different classifiers. Therefore, two Bonferroni–Dunn tests were conducted. In [9], we can find that the critical value $q_{0.10} = 2.128$, such that $CD_{0.10} = 1.122$, $(k = 4,\ N = 12)$.

For both of SVM and 3NN, the Bonferroni–Dunn tests demonstrate that NFRS is statistically better than FCM + RS and FRSA with α = 0.1, respectively. There is no consistent evidence, however, to indicate the statistical differences from FISEN.

TABLE V
RANK OF THE FEATURE SELECTION ALGORITHMS WITH SVM

Data sets	FCM + RS	FISEN	NFRS	FRSA
Glass	3	1	2	4
Wdbc	4	3	1	2
Ionos	4	2	2	2
Sonar	4	2	1	3
Diabetes	3	1	2	4
Gearbox	1.5	1.5	3	4
Segment	1	2	3	4
Brain	4	2	1	3
Breast	4	1	2.5	2.5
AMLALL	4	3	1	2
Prostate1	4	3	1	2
Prostate2	4	3	1	2
Average	3.38	2.04	1.71	2.88

TABLE VI
RANK OF THE FEATURE SELECTION ALGORITHMS WITH SVM

Data sets	FCM + RS	FISEN	NFRS	FRSA
Glass	4	2	1	3
Wdbc	4	3	1	2
Ionos	4	3	1.5	1.5
Sonar	4	1	2	3
Diabetes	2	3	1	4
Gearbox	2	1	3	4
Segment	2	1	3	4
Brain	4	2	1	3
Breast	4	3	1.5	1.5
AMLALL	4	3	2	1
Prostate1	3	1	2	4
Prostate2	4	2.5	1	2.5
Average	3.42	2.13	1.66	2.79

TABLE Ⅶ
The VALUE OF F_F FOR DIFFERENT CLASSIERS

	SVM	3NN
F_F	6.21	6.04

Table Ⅷ lists the running times of the four reduction algorithms. We can find that the FCM + RS algorithm spend more running time in most cases than that with the other algorithms. This is because the algorithm not only need some time to reduce attributes, but also spend additional time for data discretization. Of the four reduction algorithms, FISEN runs the fastest. This is because the FISEN algorithm does not need to compute the lower approximation of each sample. Therefore, it effectively saves time. The running time of NFRS is shorter than that of the FRSA algorithm. From Tables Ⅲ and Ⅳ, we know that most of the classification accuracies of the NFRS algorithm are higher than that of the FRSA. This shows that the use of the NFRS algorithm not only increases the classification accuracy, but also does not increase the time complexity. Thus, it can be seen that the NFRS method is both feasible and effective.

TABLE Ⅷ
RUNNING TIME OF REDUCTION WITH DIFFERENT ALGORITHMS

Data sets	FCM + RS	FISEN	NFRS	FRSA
Glass	0.65 ± 0.09	0.37 ± 0.09	0.40 ± 0.13	0.52 ± 0.19
Wdbc	3.23 ± 0.30	2.65 ± 0.23	2.82 ± 0.36	2.88 ± 0.21
Ionos	1.95 ± 0.25	1.32 ± 0.16	1.62 ± 0.33	1.66 ± 0.26
Sonar	2.80 ± 0.26	1.89 ± 0.16	1.76 ± 0.20	1.62 ± 0.23
Diabetes	1.79 ± 0.33*	1.13 ± 0.13	1.15 ± 0.17	1.13 ± 0.19
Gearbox	17.45 ± 2.12	78.58 ± 6.22	98.34 ± 10.17	107.82 ± 10.23
Segment	5.94 ± 0.42	34.58 ± 0.99	35.95 ± 3.36	78.65 ± 9.87
Brain	76.96 ± 11.84	25.74 ± 2.89	44.97 ± 13.54	50.70 ± 12.34
Breast	108.17 ± 18.05	48.69 ± 6.73	33.54 ± 6.10	50.22 ± 5.56
AMLALL	55.22 ± 6.01	17.83 ± 4.55	18.89 ± 4.19	18.59 ± 4.17
Prostate1	383.82 ± 15.57	139.42 ± 16.83	131.11 ± 12.48	127.89 ± 10.69
Prostate2	127.33 ± 15.22	44.34 ± 5.07	54.93 ± 11.83	53.12 ± 2.81
Average	65.44 ± 5.87	33.05 ± 3.67	35.46 ± 5.24	41.23 ± 4.73

TABLE Ⅸ
THE FEATURE SUBSETS NFRS AND FRSA ALGORITHM

Data sets	NFRS	FRSA
Glass	1, 4, 9, 7	1, 4, 9, 6, 5
Wdbc	28, 23, 16, 7, 22, 8, 25, 18, 26, 21	28, 23, 7, 22, 16, 8, 25, 6, 21, 26, 2
Ionos	1, 4, 26, 23, 27, 9, 20, 7, 19, 2, 5	1, 4, 26, 23, 27, 9, 20, 7, 19, 2, 5
Sonar	44, 21, 35, 12, 27, 29, 54, 24, 31, 16, 8, 37, 59, 49, 32, 53, 23, 26, 20, 55, 42	21, 25, 36, 30, 16, 12, 54, 23, 32, 27, 10, 45, 35, 41, 53, 49, 37, 59
Diabetes	2, 7, 5, 1, 8, 4	5, 7, 1, 8, 4
Gearbox	34, 29, 65, 20, 62, 8, 25, 56, 2, 35, 26	65, 20, 17, 35, 62, 26, 11, 7, 8
Segment	18, 16, 2, 1, 17, 4, 5, 3, 13	18, 11, 2, 1,17, 4, 5, 13, 3
Brain	3962, 4269, 2532, 2703, 5846, 903, 5062, 2739, 868, 1294, 497, 5011, 1116, 1498, 2762, 2756	3962, 4269, 2532, 5899, 2703, 52, 4993, 4705, 5062, 3094, 5182, 868, 4001, 2357, 1498
Breast	4085, 7797, 7966, 1315, 6383, 6436, 8506, 7489, 4478, 5477	4085, 7797, 7966, 6383, 6436, 8506, 1315, 4478, 5477, 7489
AMLALL	758, 1882, 4050, 2642, 3252	758, 1882, 4050, 4680, 1779
Prostate1	8850, 12067, 9850, 11125, 2802, 6195, 6367, 231	8850, 12067, 9850, 11125, 2802, 6367, 3480, 11478
Prostate2	4823, 7372, 1899, 7310, 10417, 6880, 5820	4823, 5820, 7310, 8545, 7372, 6640, 6880, 5227

All of the experimental results reported above are given out by a 10-fold cross validation. To show the selected feature subset of a data set, in the following we employ the FRSA and FNRS algorithms to reduce the entire data set based on parameters where the classification accuracies are obtained in the above experiments. The selected feature subsets are listed in Table 9, which shows that most of the features selected for the NFRS and FRSA algorithms are the same. For the Ionos and Breast data sets, in particular, the selected feature subsets are identical and the classification accuracies for the NFRS and FRSA algorithms are almost the same. The slight differences for Ionos may be due to the fact that the selected feature subsets are given out by reducing the entire data set, while the classification accuracies are based on a 10-fold cross-validation. In this case, we specify the same ranking for the NFRS and FRSA algorithms as shown in Tables Ⅴ and Ⅵ. This result illustrates that the NFRS algorithm constitutes an improvement over the FRSA algorithm.

Finally, a series of experiments was conducted by 10-fold cross validation to demonstrate the variation in the classify-cation accuracy with ε and λ. We set the value of ε to vary from 0.1 to 0.5 in steps of 0.05, and λ to vary from 0.1 to 0.6 in steps of 0.1. Figures 1–12 show only the accuracy curves for some datasets with SVM. The experimental results obtained using 3NN are roughly consistent with SVM.

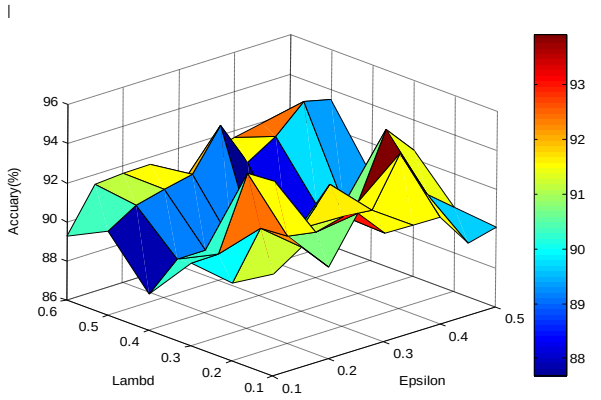


Fig. 1 Accuracy varying with thresholds ε and λ (Glass)

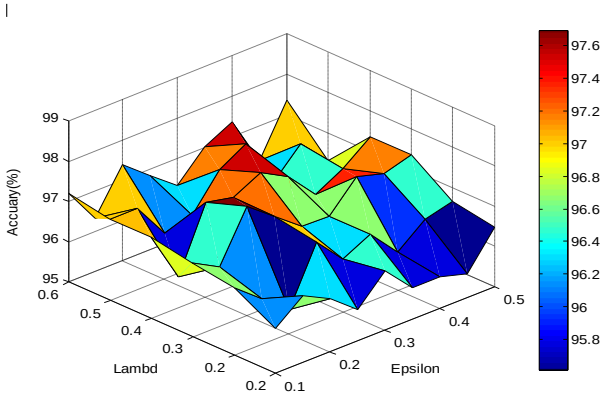


Fig. 2 Accuracy varying with thresholds ε and λ (Wdbc)

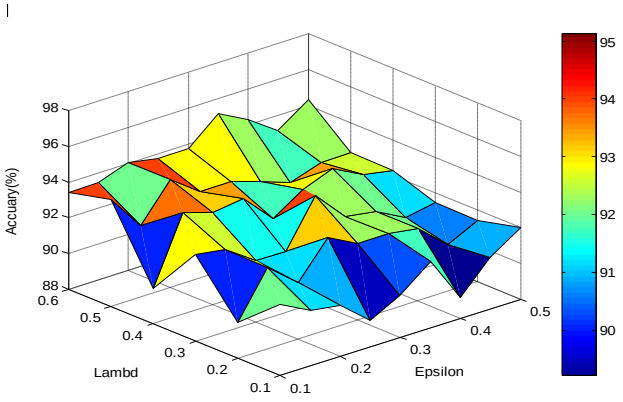


Fig. 3 Accuracy varying with thresholds ε and λ (Ionos)

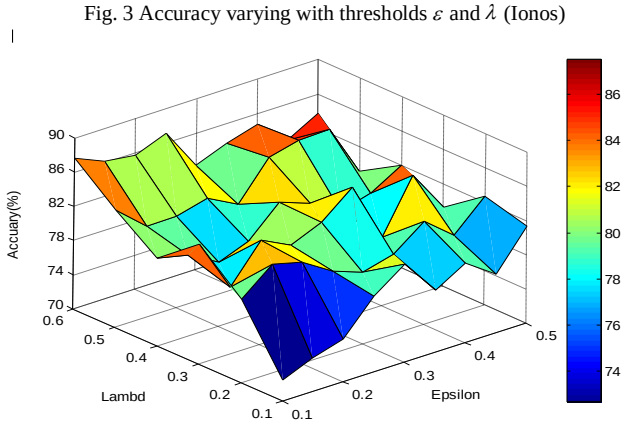


Fig. 4 Accuracy varying with thresholds ε and λ (Sonar)

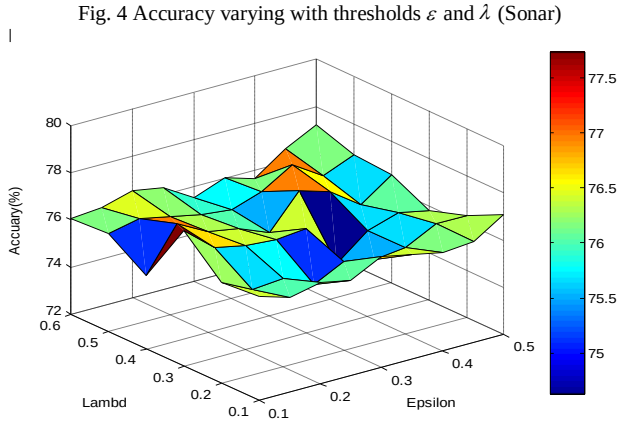


Fig. 5 Accuracy varying with thresholds ε and λ (Diabetes)

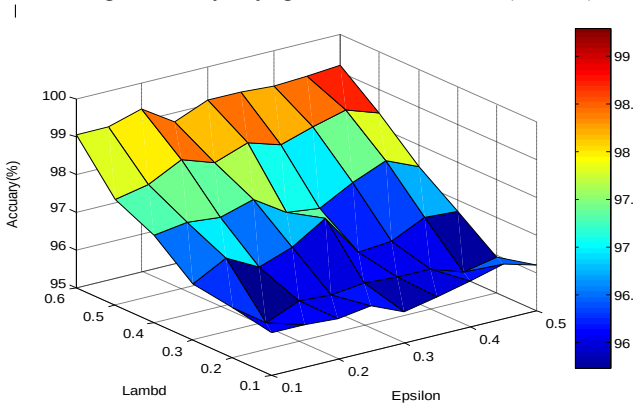


Fig. 6 Accuracy varying with thresholds ε and λ (Gearbox)

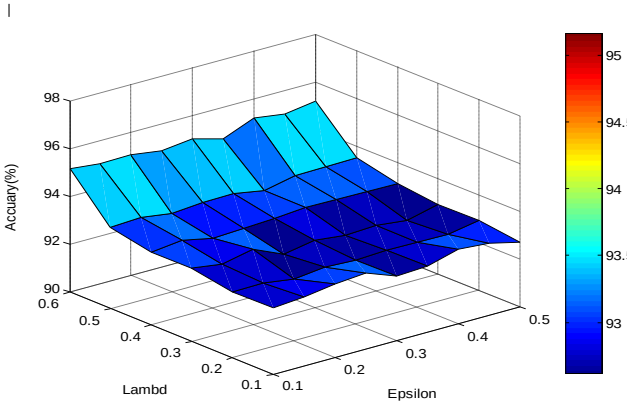


Fig. 7 Accuracy varying with thresholds ε and λ (Segment**)**

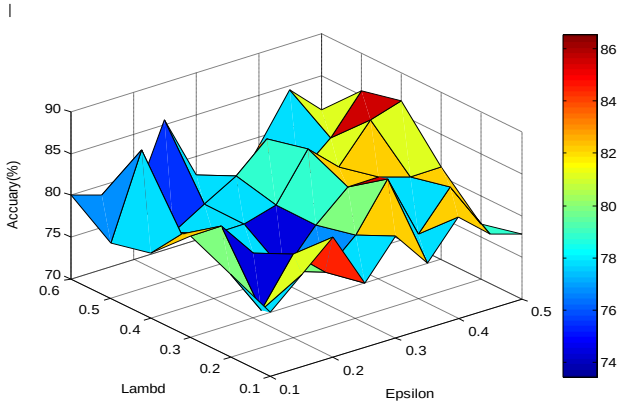


Fig. 8 Accuracy varying with thresholds ε and λ (Brain)

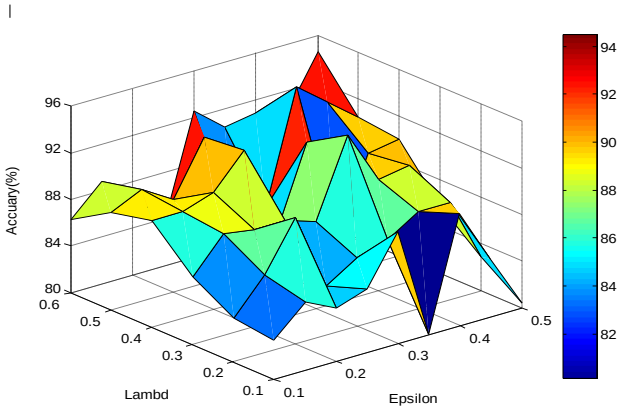


Fig. 9 Accuracy varying with thresholds ε and λ (Breast)

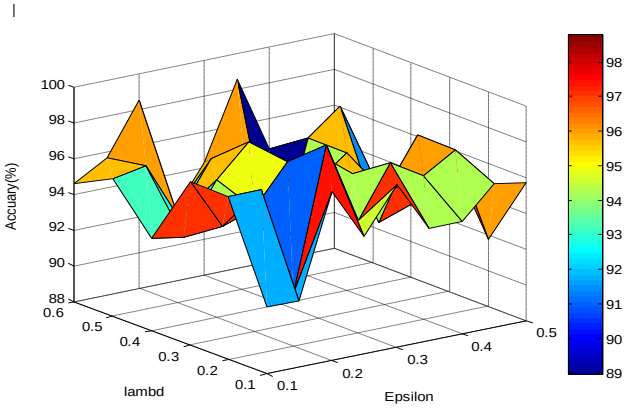


Fig. 10 Accuracy varying with thresholds ε and λ (AMLALL)

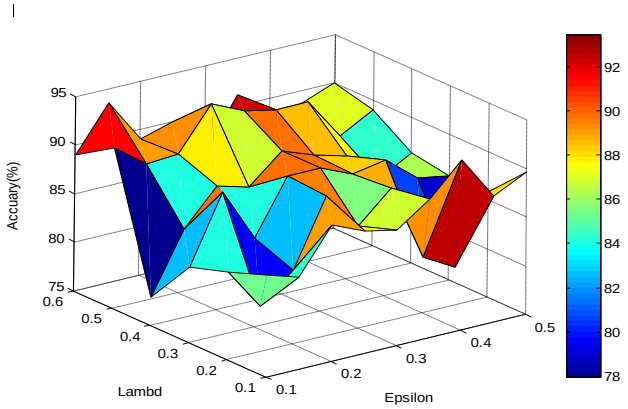


Fig. 11 Accuracy varying with thresholds ε and λ (Prostate1)

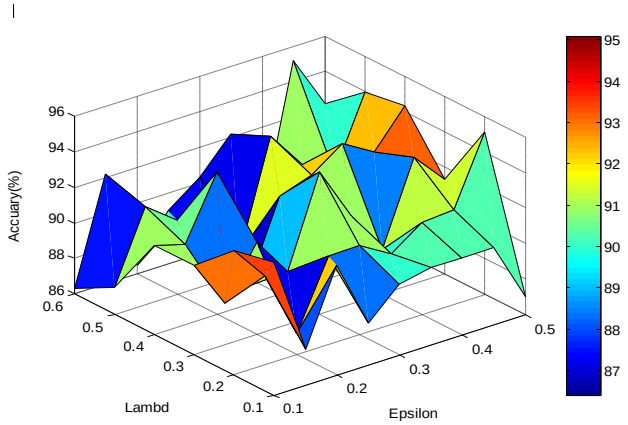


Fig. 12 Accuracy varying with thresholds ε and λ (Prostate2)

Figures 1–12 show the classification accuracies varying with ε and λ. We can clearly see that most of data sets exhibit higher classification accuracy over a greater area. In particular, Ionos, Wdbc, Gearbox, Segment and Prostate1 exhibit stability in their respective regions. Thus, it can be seen that the NFRS algorithm is both feasible and stable.

VI. Conclusion And Future Work

Feature selection is one of the important steps in classification learning. Reducing the number of redundant or irrelevant features can improve the classification performance in most cases. The fuzzy rough set model is one of the most important rough set methods used in attribute reduction. However, classical fuzzy-rough dependency cannot better reflect the classification ability of a subset of features because it merely keeps the fuzzy positive region maximal and cannot fit data well. In this study, we introduced a fitting fuzzy rough set model to overcome this problem. We defined the fuzzy decision of samples by introducing the concept of fuzzy neighborhood. To better determine the relevance between the decision and condition attributes, a parameterized fuzzy relation was introduced to construct a new fuzzy dependency function. The proposed method can fit a given data well and guarantee the maximal membership degree of a sample to its own category. The advantage of the proposed model lies in the fact that samples with great membership degrees can be easily classified into their decision equivalence class with low uncertainty. Twelve data sets, selected from UCI and KRBM are used to compare the performance of the proposed algorithm with that of existing algorithms. The experimental results show that the proposed reduction algorithm is more effective than classical fuzzy rough sets, especially for those data sets in which different categories have a large degree of overlap. Furthermore, most data sets can achieve a high degree of precision over a wide region.

However, there are still some problems to be considered, and further discussion on the proposed fuzzy rough set model is required. For example, 1) Is there an over-fitting phenomenon for some data sets with respect to the proposed model? If so, how can we characterize it and identify an effective means of avoiding it? 2) Similarly for classical fuzzy rough sets, how do we describe the proposed model by using the concept of the discernibility matrix? 3) We also need to investigate how the proposed model can be applied to the fields of classification learning. Further research into these problems will aid in the development of a systematic theory for the analysis of real-valued data sets using fuzzy rough sets.

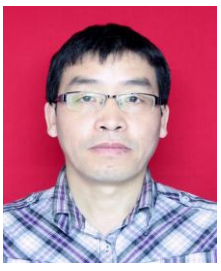
Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful and insightful comments and suggestions.

References

[1] R. B. Bhatt, M. Gopal, “On the compact computational domain of fuzzy rough sets,” Pattern Recognition Letter, vol. 26, pp. 1632–1640, 2005.
[2] C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Databases, 1998. [Online]. Available: http://www. ics.uci.edu/~mlearn/ MLRepository.html.
[3] D. G. Chen, Q. H. Hu and Y. P. Yang, “Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets,” Information Sciences, vol. 181, no. 23, pp. 5169–5179, 2011.
[4] D. G. Chen, L. Zhang, S. Y. Zhao, Q. H. Hu, P. F. Zhu, “A novel algorithm for finding reducts with fuzzy rough sets,” IEEE Transaction on Fuzzy Systems, vol. 20, no.2, pp. 385–389, 2012.
[5] C. Cornelis, R. Jensen, G. Hurtado, et al, “Attribute select with fuzzy decision reducts,” Information Sciences, vol. 177, pp. 3–20, 2007.
[6] M. Dash and H. Liu, “Consistency-based search in feature selection,” Artificial Intelligence, vol. 151, no. 1/2, pp. 155–176, 2003.
[7] D. Dubois and H. Prade, “Rough fuzzy sets and fuzzy rough sets,” International Journal of General Systems, vol. 17, pp. 191–208, 1990.
[8] O. J. Dunn, “Multiple comparisons among means,” Journal of the American Statistical Association, vol. 56, pp. 52-64, 1961.
[9] J. Demsar, “Statistical comparison of classifiers over multiple data sets,” J. Mach. Learn. Res., vol. 7, pp.1–30, 2006.
[10] M. Friedman, “A comparison of alternative tests of significance for the problem of m ranking,” Ann. Math. Statist., 11,vol. 86–92, 1940.
[11] 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.
[12] T. K. Ho and M. Basu, “Complexity measures of supervised classification problems,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 289–300, 2002.
[13] Q. H. Hu, Z. X. Xie, and D. R. Yu, “Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation,” Pattern Recognit., vol. 40, no. 12, pp. 3509–3521, 2007
[14] Q. H. Hu, D. R. Yu, Z. X. Xie, and J. F. Liu, “Fuzzy probabilistic approximation spaces and their information measures,” IEEE Transaction on Fuzzy Systems, vol. 14, no. 2, pp. 191–201, 2006.
[15] Q. H. Hu, D. R. Yu, Z. X. Xie, “Information-preserving hybrid data reduction based on fuzzy-rough techniques,” Pattern Recognition Letters, vol. 27, no. 5, pp. 414–423, 2006.
[16] R. Jensen, Q. Shen, “Fuzzy-rough attributes reduction with application to web categorization,” Fuzzy Sets and systems, vol. 141, pp. 469–485, 2004.
[17] R. Jensen, Q. Shen, “New approaches to fuzzy-rough feature selection,” IEEE Transactions on Fuzzy Systems, vol. 17, pp. 824–838, 2009.

[18] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough-and fuzzy-rough-based approaches," IEEE Transaction on Knowledge and Data Engineering, vol. 16, no. 12, pp. 1457–1471, 2004.
[19] R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," IEEE Transactions on Fuzzy Systems, vol. 15, no. 1, pp.73–89, 2007.
[20] N. Kwak and C. H. Choi, "Input feature selection for classification problems," IEEE Transaction on Neural Network, vol. 13, no. 1, pp. 143–159, 2002.
[21] J. Li, H. Zhao, W. Zhu. "Fast randomized algorithm with restart strategy for minimal test cost feature selection," International Journal of Machine Learning and Cybernetics, vol. 6, no.3, pp. 435–442, 2015.
[22] G. Lang, Q. Li, T. Yang, "An incremental approach to attribute reduction of dynamic set-valued information systems," International Journal of Machine Learning and Cybernetics, vol. 5, no.5, pp. 775–788, 2014.
[23] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining," Boston, MA: Kluwer, 1998.
[24] J. Y. Liang, F. Wang, C. Y. Dang, Y. H. Qian, "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.
[25] J. Y. Liang, C. Y. Dang, K. Chin, C. Yam Richard, "A new method for measuring uncertainty and fuzziness in rough set theory," International Journal of General Systems, vol. 31, no. 4, pp.331–342, 2002.
[26] 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.
[27] J. Ma, Y. Leung, W. Zhang, "Attribute reductions in object-oriented concept lattice," International Journal of Machine Learning and Cybernetics, vol. 5, no.5, pp. 789–813, 2014.
[28] P. Maji, "A rough hypercuboid approach for feature selection in approximation spaces," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no.1, pp. 16–29, 2014.
[29] P. Maji and P. Garai, "On fuzzy-rough attribute selection: Criteria of max-Dependency, max-relevance, min-redundancy, and max-significance," Applied Soft Computing, vol. 13, no.9, pp. 3968–3980, 2013.
[30] J. S. Mi, Y. Leung, H. Y. Zhao, T. Feng, "Generalized fuzzy rough sets determined by a triangular norm," Information Sciences, vol. 178, no.16, pp. 3203–3213, 2008.
[31] J. S. Mi and W. X. Zhang, "An axiomatic characterization of a fuzzy generalization of rough sets," Information Sciences, vol. 160, no.1-4, pp. 235–249, 2004.
[32] D. Muni, N. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," IEEE Transactions on Systems, Man, and Cybernetics. Part B, vol. 36, no. 1, pp. 106–117, 2006.
[33] P. M. Narendra and K. Fukunaga, "A branch-and-bound algorithm for feature subset selection," IEEE Transaction on Computers, vol. C-26, no. 9, pp. 917–922, 1977.
[34] I. S. Oh, J. S. Lee, and B. R. Moon, "Hybrid genetic algorithms for feature selection," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, pp. 1424–1437, 2004.
[35] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences, vol. 11, no. 5, pp. 341–356, 1982.
[36] N. M. Parthaláin, R. Jensen, "Unsupervised fuzzy-rough set-based dimension-ality reduction," Information Sciences, vol. 229, pp. 106–121, 2013.
[37] 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.
[38] Y. Qian, J. Liang, W. Wu, C Dang, "Information granularity in fuzzy binary GrC model," IEEE Transactions on Fuzzy Systems, vol.19, no.2, pp. 253–264, 2011.
[39] Y. Qian J. Liang, W. Pedrycz, C. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," Artificial Intelligence, vol.174, pp. 597–618, 2010.
[40] 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.
[41] A. Mieszkowicz-Rolka and L. Rolka, "Variable precision fuzzy rough sets", Transactions on Rough sets 1,vol. LNCS-3100, Berlin, Germany: Springer, 2004,pp.144–160.
[42] S. Robert, "Analyzing discretizations of continuous attributes given a monotonic discrimination function", Intelligent Data Analysis, vol. 1, No.1-4, pp. 157–179, 1997
[43] C. T. Su, J. H. Hsu, "An extended Chi2 algorithm for discretization of real value attributes", IEEE Transactions on Knowledge and Data Engineering, vol. 17, No. 3, pp.437–441, 2005.
[44] Y. She, X. He, "Uncertainty measures in rough algebra with applications to rough logic," International Journal of Machine Learning and Cybernetics, vol. 5, no.5, pp. 671–681, 2014.
[45] M. Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," Machine Learning, vol. 53, no. 1/2, pp. 23–69, 2003.
[46] A. Skowron, C. Rauszer, "The discernibility matrices and functions in information systems," In: R. Slowinski (Ed.), Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht, pp. 331–362, 1992.
[47] P. Somol, P. Pudil, and J. Kittler, "Fast branch-and-bound algorithms for optimal feature selection," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 26, no. 7, pp. 900–912, 2004.
[48] K. Torkkola, "Feature extraction by nonparametric mutual information maximization," Journal of Machine Learning Reseach, vol. 3, no. 7/8, pp. 1415–1438, 2003.
[49] E. C. C. Tsang, D. G. Chen, D. S. Yeung, X. Z. Wang, "Attribute reduction using fuzzy rough sets," IEEE Transactions on Fuzzy Systems, vol. 16, no. 5, pp. 1130–1141, 2008.
[50] E. C. C. Tsang, Q. Hu, D. Chen, "Feature and instance reduction for PNN classifiers based on fuzzy rough sets," International Journal of Machine Learning and Cybernetics, DOI: 10.1007/s13042-014-0232-6, 2014.
[51] C. Wang, Q. He, D. Chen, Q. Hu, "A novel method for attribute reduction of covering decision systems," Information Sciences, vol. 254, pp.181–196, 2014.
[52] X. Z. Wang, L. C. Dong, J. H. Yan, "Maximum ambiguity based sample selection in fuzzy decision tree induction," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no.8, pp. 1491–1505, 2012.
[53] B. Wang, J. Liang, Y. Qian, "Determining decision makers' weights in group ranking: a granular computing method," International Journal of Machine Learning and Cybernetics, vol. 6, no.3, pp. 511–521, 2015.
[54] W. Z. Wu, Y. Leung, M. W. Shao, "Generalized fuzzy rough approximation operators determined by fuzzy implicators," International Journal of Approximate Reasoning, vol. 54, no.9, pp. 1388–1409, 2013.
[55] Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," Information Sciences, vol. 178, no. 17, pp. 3356–3373, 2008.
[56] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," Journal of Machine Learning Research, vol. 5, pp. 1205–1224, 2004.
[57] 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.
[58] 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.
[59] N. Zhong, J. Dong, S. Ohsuga, "Using Rough sets with heuristics for feature selection," Journal of Intelligent Information Systems, vol. 16, no. 3, pp. 199–214, 2001.
[60] X. Zhang, J. Dai, Y. Yu, "On the union and intersection operations of rough sets based on various approximation spaces," Information Sciences, vo. 292, pp. 214–229, 2015.
[61] Kent Ridge Bio-medical Dataset, http://datam.i2r.a-tar.edu.sg/datasets/krbd/index.html.



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.

His research interests are focused on fuzzy sets, rough sets, data mining and knowledge discovery. He has authored or coauthored more than 50 journal and conference papers in the areas of machine learning, data mining, and rough set theory.

Yali Qi received her B.Sc. degrees in mathematics from Bohai University in 2012. Now, she is a graduate student for a Master degree.

Her main research interests include fuzzy sets, rough sets, pattern recognition and knowledge discovery.

Mingwen Shao received the M.S. degree in mathematics from Guangxi University, China, in 2002, and the PhD degree in applied mathematics from Xi'an Jiaotong University, China, in 2005. He has publicshed more than 40 papers in international journals and international conferences. His current research interests include rough sets, fuzzy sets, formal concept analysis, and granular computing.

Qinghua Hu received the B.S., M.S., and Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008,respectively. He was a Post-Doctoral Fellow with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, from 2009 to 2011. He is currently a Full Professor and the Vice Dean of the School of Computer Science and Technology with Tianjin University, Tianjin, China. He has authored over 100 journal and conference papers in the areas of granular computing-based machine learning, reasoning with uncertainty, pattern recognition, and fault diagnosis. His research interests include rough sets, granular computing, and data mining for classification and regression.

Prof. Hu was acted as the Program Committee Co-Chair of the International Conference on Rough Sets and Current Trends in Computing in 2010, the Chinese Rough Set and Soft Computing Society in 2012 and 2014, and the International Conference on Rough Sets and Knowledge Technology and the International Conference on Machine Learning and Cybernetics in 2014，general Co-Chair of IJCRS 2015, and serves as a Referee for a great number of journals and conferences.

Degang Chen received the M.S. degree from Northeast North University, Changchun, Jilin, China, in 1994 and the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2000.

He was a Postdoctoral Fellow with Xi'an Jiaotong University, Xi'an, China, from 2000 to 2002, and with Tsinghua University, Beijing, China, from 2002 to 2004. Since 2006, he has been a Professor with North China Electric Power University, Beijing, China. He has authored or coauthored more than 140 research publications. His research interests include fuzzy groups, fuzzy algebra, fuzzy analysis, rough sets, machine learning.

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 best known for multi-granulation rough sets in learning from categorical data and granular computing. 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. On professional services, Qian has served as program chairs or special issue chairs of RSKT, JRS, and ICIC, and PC members of many machine learning, data mining, and granular computing. He also served on the Editorial Board of International Journal of Knowledge-Based Organizations and the Editorial Board of Artificial Intelligence Research.

Yaojin Lin received the Ph.D. degree in School of Computer and Information from Hefei University of Technology. He currently is an associate professor with Minnan Normal University and a postdoctoral fellow with Tianjin University.

His research interests include data mining, and granular computing. He has published more than 60 papers in international journals and international conferences.