Fuzzy rough attribute reduction for categorical data

Changzhong Wang, Yan Wang, Mingwen Shao, Yuhua Qian, Degang Chen

***Abstract*—Classical rough set theory is considered as a useful tool for dealing with the uncertainty of categorical data; the major deficiency of this model is that the classical rough set model is sensitive to noise in classification learning due to the stringent condition of equivalence relation. Thus, a class of fuzzy similarity relations was introduced to describe the similarity between samples with categorical attributes. However, these kinds of similarity relations also have deficiencies when they are used in fuzzy rough computation. In this paper, we propose a new fuzzy-rough-set model for categorical data by introducing a variable parameter to control the similarity of samples. This model employs the iterative computation strategy to define fuzzy rough approximations and dependency functions. It is proved that the proposed rough dependency function is monotonic. Finally, the proposed model is applied to the attribute reduction of categorical data. The experimental results indicate that the proposed model is more effective for categorical data than some existing algorithms.**

***Index Terms*—Fuzzy rough set; rough approximation; categorical data; attribute reduction**

I. INTRODUCTION

Rough set methodologies have become a popular class of approaches for dealing with data with uncertainty. One significant advantage of these approaches is that they use only internal information in the data and do not depend on any prior knowledge as fuzzy modes and probabilistic methods do. In recent years, rough set methodologies have received wider attention and have been extensively applied in the fields of feature selection, reasoning with uncertainty, and classification learning.

Rough set methodologies mainly deal with two types of data: categorical and numerical data. Classical rough set theory considers only categorical data [1]. In this framework, samples are described by a set of categorical attributes, which is viewed as a discrete feature space. Categorical attributes can induce equivalence relations and partition the feature space into mutually exclusive information granules (equivalence classes). Samples in the same information granule are indiscernible. A decision variable or target subset in the feature space can be approximated by information granules [2]. However, there is one main drawback of classical rough set theory. That is, the model is sensitive to noise in classification learning. In a given equivalence class, even if there is only one sample having different class label from the others, the equivalence class will be grouped into the boundary region.

For numerical data, several extensions of the classical rough set model have been developed, which include neighborhood rough sets [3]-[5], dominance rough sets [6],[7], fuzzy rough sets [8]-[10], and so on [11]-[17].

Fuzzy rough set is one of the most important generalization models for numerical data, and it has attracted considerable attention [18]-[28]. Dubois and Prade first introduced fuzzy rough sets based on fuzzy equivalence relations by combining fuzzy sets and rough sets [8]. Radzikowska and Kerre used fuzzy logic operators to generalize the model [10]. Mi and Zhang defined a class of new fuzzy rough sets using implication operators [27]. Wu and Zhang studied the axiomatic methods for fuzzy rough approximation under min-max operators [28]. In recent years, fuzzy rough sets have had successful applications in attribute reduction [29]-[36], approximate reasoning [26], [37], and classification learning [29], [38]. For example, Jensen proposed fuzzy rough dependency functions to evaluate the classification ability of attributes and developed an attribute-reduction algorithm [35]. Chen introduced the concept of a fuzzy discernible matrix and used it to reduce redundant attributes [31]. Hu et al. combined kernel functions and fuzzy rough sets to define fuzzy dependency functions [36]. Wang introduced a fitting fuzzy rough set model and used it to perform attribute reduction [25]. Zhao et al. proposed a novel fuzzy-rough-set method for constructing a robust fuzzy rough classifier [38].

Although fuzzy rough sets can successfully handle numerical data, they will be degraded to classical rough set models when facing categorical data. That is, fuzzy rough set models have the same drawback as classical rough sets do when they are used to deal with categorical data. The reason is that fuzzy equivalence relations will be degraded to crisp equivalence relations in this case. Similarly, other generalization models of rough sets also confront the same problem when they are used

This work was supported by the National Natural Science Foundation of China under Grants 61976027, 61572082, and 61773349; the Foundation of Educational Committee of Liaoning Province (LZ2016003); and the Natural Science Foundation of Liaoning Province (20170540012, 20170540004).

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

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

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

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

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.2019.2949765, IEEE Transactions on Fuzzy Systems

to manage categorical data.

Categorical data is an important class of data in machine learning. Several improved models have been proposed to overcome the weakness of classical rough sets in dealing with categorical data. Ziako introduced the model of variable-precision rough set (VPRS) [39]. Yao proposed a probabilistic rough-set model [40]. Duntch and Gediga developed some types of conditional entropies for attribute reduction [41]. Among these models, the VPRS model and information entropy were extensively discussed and used in coping with noisy data. In the VPRS model, the concept of inclusion was introduced to compute the lower and upper approximations of a target decision. The method of information entropy uses equivalence relations to define information entropy and conditional entropy of feature subsets and employs them to measure the uncertainty in Pawlak's approximation space [42]-[44]. The two methods are more effective than other existing methods in dealing with the uncertainty of categorical data. Besides, Mieszkowicz-Rolka and Rolka introduced a variable-precision fuzzy-rough-set model to deal with noisy data [45], in which the fuzzy memberships to rough approximations were defined by fuzzy inclusion. Zhao et al. proposed a fuzzy rough variable-precision model to overcome the noise of perturbation [46]. Nevertheless, these variable-precision models will be degraded to classical variable-precision models of rough sets when they are confronted with categorical data.

The disadvantage of classical rough sets lies in the stringent conditions of defining equivalence relations. Two samples are equivalent if and only if their attribute values are equal to each other in each dimension. If there is an attribute such that two samples have different attribute values, the two samples will be grouped into different equivalence classes. Such clustering is a major reason why classical rough sets are sensitive to noise. In fact, we can introduce a similarity measure to describe the similarity between categorical samples and then use it to granulate the feature space into elementary information granules. However, in this model there is a constant parameter that controls the similarity between samples. For a high-dimensional feature space, the membership degrees of a fuzzy similarity relation may get very small when just a few attributes are included in rough computation. To overcome this problem, in this paper, we propose a novel fuzzy rough computation model for categorical data. This model employs the iterative computation strategy to define fuzzy rough approximations and dependency functions.

As we know, attribute reduction or feature selection is one important application of rough set theory. In a classification task, some of attributes may be redundant and do not provide classification information. They can destroy the performance of algorithms and bring high computational complexity. The main task of attribute reduction is to find an optimal attribute subset while keeping or improving classification accuracy. Therefore, attribute reduction has a close relationship to classification learning [47]-[54]. In this paper, we apply the proposed model to the attribute reduction of categorical data and verify that the proposed model is feasible and effective.

Compared to other rough set models, the proposed model has the following advantages: (1) A fuzzy similarity relation is used to describe the similarity of categorical samples. It can better characterize the similarity of categorical data than crisp equivalence relations. Correspondingly, the fuzzy rough computation model can elaborate on the uncertainty of categorical data more subtly. (2) The computational complexity of the proposed model is lower than that of classical fuzzy rough sets.

This paper is organized as follows. In Section 2, we introduce a fuzzy similarity relation to depict the similarity of categorical data and review the corresponding fuzzy rough computations. In Section 3, a new fuzzy-rough-set model, called the fuzzy-rough iterative computation model, is proposed for categorical data. In Sections 4 and 5, we apply the proposed model to attribute reduction to verify its feasibility and stability. Section 6 concludes the paper.

II. FUZZY ROUGH SETS FOR CATEGORICAL DATA

Let $U=\{x_1,x_2,\cdots,x_n\}$ be a set of samples, called the universe of discourse, $A=\{a_1,a_2,\cdots,a_m\}$ be a set of categorical attributes (features) to describe the samples, and D be a decision attribute. Then, the triplet $<U,A,D>$ is called a discrete feature space. Without loss of generality, the universe U is assumed to be divided into r crisp decision classes by D and is expressed as $U/D=\{D_1,D_2,\cdots D_r\}$. In this section, we first introduce a similarity measure to depict the similarity between categorical samples. Let $B\subseteq A$; a similarity measure with B is defined as follows:

$$R_B\left(x_i,x_j\right)=\frac{1}{c}\left|\left\{a\in B:a\left(x_i\right)=a\left(x_j\right)\right\}\right| \tag{1}$$

where $|\cdot|$ denotes the cardinality of a set and c is a constant parameter that adjusts the $R_B(x_i,x_j)$ to be in the interval [0,1] for any sample pair, $a(x)$ is the attribute value of sample x on attribute a. Here, $c=|A|$. Obviously, R_B satisfies symmetry. It is a fuzzy similarity relation and can be used to characterize the similarity of categorical samples. In general, R_B can be represented by a matrix and expressed as $R_B=(r_{ij})_{n\times n}$, where $0\le r_{ij}\le 1$, $i=1,2,\cdots,n$, $j=1,2,\cdots,n$.

We can see that the attribute subset B influences the membership degrees of R_B. The increasing number of attributes in B means the greater the membership degrees of R_B. That is, R_B has the following property.

Property 1. Let $B_1\subseteq B_2$, then $R_{B_1}\subseteq R_{B_2}$.

Consider a discrete feature space $<U,A,D>$, $x\in U$, $B\subseteq A$, and $U/D=\{D_1,D_2,\cdots D_r\}$. Let R_B be the fuzzy similarity relation on U introduced by formula (1). The fuzzy similarity class $[x]_B$ associated with x and R_B is a fuzzy set on U, defined as $[x]_B(y)=R_B(x,y)$, $y\in U$. All the fuzzy similarity classes $\{[x]_B\mid x\in U\}$ constitute the fuzzy

information granules of $<U, A, D>$. For any decision class $D_i \in U/D$, from the theory of fuzzy rough sets [8], the decision class D_i can be approximately characterized by fuzzy information granules as follows.

$$\underline{R_B}(D_i)(y) = \min_{x \notin D_i} \{1 - R_B(x, y)\}, \quad y \in U. \tag{2}$$

$$\overline{R}_B(D_i)(y) = \max_{x \in D_i} R_B(x, y), \quad y \in U. \tag{3}$$

$\underline{R_B}(D_i)$ and $\overline{R}_B(D_i)$ are called lower and upper approximations of D_i relative to B respectively.

$\underline{R_B}(D_i)(y)$ denotes the membership degree of y certainly being included in the equivalence class D_i. When sample y does not belong to class i, the value of $\underline{R_B}(D_i)(y)$ is the smallest. Otherwise, it is equal to the smallest value of dissimilar degrees between y and the samples not falling into class i. $\overline{R_B}(D_i)(y)$ represents the membership degree of sample y possibly belonging to equivalence class D_i. If sample y belongs to class i, the value of $\overline{R_B}(D_i)(y)$ is the largest. If not, it is equal to the max-value of the fuzzy similarities between y and all the samples in class i.

Fuzzy positive region and dependency function are two important concepts in fuzzy rough set theory [35]. As they are uncertain measures that represent the classification ability of attribute subsets, they are usually used as feature-evaluation functions in feature selection or attribute reduction [29]-[36]. The fuzzy positive region of decision D relative to B can be formulated as

$$POS_B(D) = \bigcup_{i=1}^{r} \underline{R_B}(D_i). \tag{4}$$

The dependency function of D relative to B can be formally described by

$$\partial_B(D) = \frac{\sum_{x_i \in U} POS_B(D)(x_i)}{|U|}. \tag{5}$$

Intuitively, the samples with greater memberships are more easily classified into one of the decision classes with lesser uncertainty. The dependency function is defined as the ratio of the size of the positive region over all samples in the feature space. Evidently, $0 \le \partial_B(D) \le 1$.

Property 2. Let $B_1 \subseteq B_2 \subseteq A$, $POS_{B_2}(D) \subseteq POS_{B_1}(D)$.

Proof. Because $B_1 \subseteq B_2$, it follows from Property 1 that $R_{B_1}(x, y) \le R_{B_2}(x, y)$ for any $x, y \in U$. By formula (2), we conclude that $\underline{R}_{B_2}(D_i)(y) \le \underline{R}_{B_1}(D_i)(y)$ for any $y \in U$, which means that $\underline{R}_{B_2}(D_i) \subseteq \underline{R}_{B_1}(D_i)$. The result is derived from formula (4).

Property 3. If $B_1 \subseteq B_2 \subseteq A$, then $\partial_{B_2}(D) \le \partial_{B_1}(D)$.

It is worth noting that Properties 2 and 3 differ from those in classical fuzzy rough set theory, where the monotonic relationship is reversed.

According to the rough set theory, the dependency function indicates the approximating ability of an attribute subset for a decision. It can be used as a measure for evaluating an attribute's significance [35].

Let $B \subseteq A$ and $a \in A - B$. From Property 3, the significance of a relative to B is formulated as

$$SIG(a, B, D) = \partial_B(D) - \partial_{B \cup \{a\}}(D). \tag{6}$$

The significance of attribute a is related to attribute subset B. An attribute is advantageous for classification if it has the greater significance. Formula (6) also differs from ones in classical fuzzy rough sets, where the subtracting relationship is reversed.

III. FUZZY ROUGH COMPUTATION MODEL

As discussed above, fuzzy relations introduced by formula (1) are crucial for defining a fuzzy rough computation model. However, there is a constant parameter c in the formula (1). For a dataset with a large number of attributes, the membership degrees of samples to a relation can get very small when a few of the attributes are included in rough computation. That is to say, the lower the number of the included features in rough computation, the smaller the discrimination of memberships.

To overcome this problem, in this section we propose a fuzzy rough iterative computation model for categorical data. We first deduce the iterative computation method under general conditions. Then, we present the iterative formulas in the case of natural sequences and show the monotonicity proof of attribute significance.

In the following, we use the strategy of an increasing sequence as the value of c in formula (1) to deduce fuzzy-rough computation. For instance, if a dataset contains 200 features, we select the sequence (20, 40, 60, 80, 100, 120, 140, 160, 180, 200) for c. First, let $c = 20$ when the number of the included features is less than 20. If the number of the included features is between 20 and 40, let $c = 40$, and so on. Then, the relationship between fuzzy lower approximations under different values of parameter c can be characterized as follows.

Theorem 1. Let $<U, A, D>$ be a discrete feature space, $\{c_i : i = 1, 2, \cdots, l\}$ be an increasing sequence of numbers as the value domain of c in formula (1), $B \subseteq A$, and $D_h \in U/D$. When $c = c_i$, the lower approximation of D_h is rewritten as $\underline{R}_B(D_h)_i$. When $c = c_{i+1}$, the lower approximation of D_h is rewritten as $\underline{R}_B(D_h)_{i+1}$, then

$$\underline{R}_B(D_h)_{i+1} = 1 - \frac{c_i}{c_{i+1}}\left(1 - \underline{R}^i_B(D_h)\right) = \frac{c_{i+1} - c_i}{c_{i+1}} + \frac{c_i}{c_{i+1}} \underline{R}_B(D_h)_i. \tag{7}$$

Proof. Denote formula (1) as

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.2019.2949765, IEEE Transactions on Fuzzy Systems

$$\left(R_B\right)_i\left(x_k,x_l\right)=\frac{1}{c_i}\left|\left\{a\in B:a\left(x_k\right)=a\left(x_l\right)\right\}\right|$$

when $c=c_i$. Then, we have

$$\left(R_B\right)_i\left(x_k,x_l\right)=\frac{1}{c_i}\left|\left\{a\in B:a\left(x_k\right)=a\left(x_l\right)\right\}\right| \text{ and}$$

$$\left(R_B\right)_{i+1}\left(x_k,x_l\right)=\frac{1}{c_{i+1}}\left|\left\{a\in B:a\left(x_k\right)=a\left(x_l\right)\right\}\right|.$$

It follows from Formula (2) that

$$\begin{aligned}\underline{R}_B(D_h)_i(x_k)&=\min_{x_l\notin D_h}\left\{1-(R_B)_i(x_k,x_l)\right\}\\&=\min_{x_l\notin D_h}\left\{1-\frac{1}{c_i}\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}\\&=1-\frac{1}{c_i}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}.\end{aligned}$$

Similarly,

$$\begin{aligned}\underline{R}_B(D_h)_{i+1}(x_k)&=\min_{x_l\notin D_h}\left\{1-(R_B)_{i+1}(x_k,x_l)\right\}\\&=1-\frac{1}{c_{i+1}}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}.\end{aligned}$$

Then we have

$$\frac{1-\underline{R}_B(D_h)_i(x_k)}{1-\underline{R}_B(D_h)_{i+1}(x_k)}=\frac{c_{i+1}}{c_i}.$$

Thus,

$$\underline{R}_B(D_h)_{i+1}(x_k)=1-\frac{c_i}{c_{i+1}}\left(1-\underline{R}_B(D_h)_i(x_k)\right),$$

So

$$\begin{aligned}\underline{R}_B(D_h)_{i+1}(x_k)&=1-\frac{c_i}{c_{i+1}}\left(1-\underline{R}_B(D_h)_i(x_k)\right)\\&=\frac{c_{i+1}-c_i}{c_{i+1}}+\frac{c_i}{c_{i+1}}\underline{R}_B(D_h)_i(x_k).\end{aligned}$$

This completes the proof of the theorem.

This theorem shows that the fuzzy lower approximation $\underline{R}_B(D_h)_{i+1}$ of D_h under parameter value c_{i+1} is equal to the linear combination of the lower approximation under parameter value c_i.

Theorem 2. Let $<U,A,D>$ be a discrete feature space, $\{c_i:i=1,2,\cdots,l\}$ be an increasing sequence of numbers as the value domain of c in formula (1), and $B\subseteq A$. When $c=c_i$, the fuzzy positive region of decision D is rewritten as $POS_B(D)_i$. When $c=c_{i+1}$, the fuzzy positive region is rewritten as $POS_B(D)_{i+1}$, then

$$POS_B(D)_{i+1}=1-\frac{c_i}{c_{i+1}}\left(1-POS_B(D)_i\right)=\frac{c_{i+1}-c_i}{c_{i+1}}+\frac{c_i}{c_{i+1}}POS_B(D)_i. \quad (8)$$

Proof. Let $U/D=\{D_1,D_2,\cdots D_r\}$. For $D_h\in U/D$ and any $x_k\in U$, similar to the proof of Theorem 1, we have

$$\underline{R}_B(D_h)_i(x_k)=1-\frac{1}{c_i}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\},$$

$$\underline{R}_B(D_h)_{i+1}(x_k)=1-\frac{1}{c_{i+1}}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}.$$

Then,

$$\begin{aligned}POS_B(D)_i(x_k)&=\max_{D_h\in U/D}\left\{\underline{R}_B(D_h)_i(x_k)\right\}\\&=\max_{D_h\in U/D}\left\{1-\frac{1}{c_i}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}\right\}\\&=1-\frac{1}{c_i}\min_{D_h\in U/D}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}\end{aligned}$$

and

$$POS_B(D)_{i+1}(x_k)=1-\frac{1}{c_{i+1}}\min_{D_h\in U/D}\max_{x_l\notin D_h}\left\{\left|\left\{a\in B:a(x_k)=a(x_l)\right\}\right|\right\}.$$

Thus,

$$\frac{1-POS_B(D)_i(x_l)}{1-POS_B(D)_{i+1}(x_l)}=\frac{c_{i+1}}{c_i}.$$

Therefore,

$$POS_B(D)_{i+1}(x_l)=1-\frac{c_i}{c_{i+1}}\left(1-POS_B(D)_i(x_l)\right).$$

So

$$POS_B(D)_{i+1}=1-\frac{c_i}{c_{i+1}}\left(1-POS_B(D)_i\right)=\frac{c_{i+1}-c_i}{c_{i+1}}+\frac{c_i}{c_{i+1}}POS_B(D)_i.$$

Theorem 3. Let $<U,A,D>$ be a discrete feature space, $\{c_i:i=1,2,\cdots,l\}$ be an increasing sequence of numbers as the value domain of c in formula (1), and $B\subseteq A$. When $c=c_i$, the dependency function is rewritten as $\partial_B(D)_i$. When $c=c_{i+1}$, the dependency function is rewritten as $\partial_B(D)_{i+1}$, then

$$\partial_B(D)_{i+1}=1-\frac{c_i}{c_{i+1}}\left(1-\partial_B(D)_i\right)=\frac{c_{i+1}-c_i}{c_{i+1}}+\frac{c_i}{c_{i+1}}\partial_B(D)_i. \quad (9)$$

Proof. For any $x_l\in U$, it follows from Theorem 2 that

$$POS_B(D)_{i+1}(x_l)=1-\frac{c_i}{c_{i+1}}\left(1-POS_B(D)_i(x_l)\right), \text{ and}$$

$$\sum_{x_l\in U}POS_B(D)_{i+1}(x_l)=|U|-\frac{c_i}{c_{i+1}}\left(|U|-\sum_{x_l\in U}POS_B(D)_i(x_l)\right).$$

So

$$\partial_B(D)_{i+1}=1-\frac{c_i}{c_{i+1}}\left(1-\partial_B(D)_i\right)=\frac{c_{i+1}-c_i}{c_{i+1}}+\frac{c_i}{c_{i+1}}\partial_B(D)_i.$$

For a given subset of attributes B, one can obtain different fuzzy rough approximations, fuzzy positive regions, and dependency functions when different parameters $(c_i:i=1,2,\cdots,l)$

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.2019.2949765, IEEE Transactions on Fuzzy Systems

are used. Theorems 1-3 present the iterative computation approaches of rough approximations by using incremental sequences.

Theorem 4. let $<U, A, D>$ be a discrete feature space, $\{c_i : i = 1, 2, \cdots, l\}$ be an increasing sequence of numbers as the value domain of c in the formula (1), $B \subseteq A$, and $a \in A - B$. When $c = c_i$, the dependency function is rewritten as $\partial_B(D)_i$. When $c = c_{i+1}$, the dependency function is rewritten as $\partial_B(D)_{i+1}$. The significance of a with respect to B is denoted as $SIG_i(a, B, D)$, then

(A) $SIG_i(a,B,D) = \partial_B(D)_{i+1} - \partial_{B\cup\{a\}}(D)_{i+1} = \frac{c_i}{c_{i+1}}\left(\partial_B(D)_i - \partial_{B\cup\{a\}}(D)_i\right)$,

(B) $SIG_i(a,B,D) = \frac{c_{i+1}-c_i}{c_{i+1}} + \frac{c_i}{c_{i+1}}\partial_B(D)_i - \partial_{B\cup\{a\}}(D)_{i+1}$. (10)

Proof. (A) From Definition 4, we have that $SIG_i(a, B, D) = \partial_B(D)_{i+1} - \partial_{B\cup\{a\}}(D)_{i+1}$. it follows from Theorem 3 that

$$\partial_B(D)_{i+1} = \frac{c_{i+1}-c_i}{c_{i+1}} + \frac{c_i}{c_{i+1}}\partial_B(D)_i \text{ and}$$

$$\partial_{B\cup\{a\}}(D)_{i+1} = \frac{c_{i+1}-c_i}{c_{i+1}} + \frac{c_i}{c_{i+1}}\partial_{B\cup\{a\}}(D)_i .$$

Thus, $SIG_i(a,B,D) = \frac{c_i}{c_{i+1}}\left(\partial_B(D)_i - \partial_{B\cup\{a\}}(D)_i\right)$.

(B) Similarly, we can get the result.

Theorem 4 presents a formula for computing the significance of an attribute. It shows that the significance numerically equals the linear difference between the fuzzy dependencies of the adjacent feature subsets.

From Theorem 1 to Theorem 4, we see that an artificially increasing sequence is used to derive the main formulas of fuzzy rough approximations. The purpose of such a design is just to illustrate that the proposed fuzzy rough computation is established under general conditions, not limited to a specific circumstance. If the increasing sequence is set to a natural sequence, there are no longer any parameters in formulas (7) - (10). In the following theorem, we present the formula for computing the significance of an attribute in the case of a natural sequence.

Theorem 5. Let $<U, A, D>$ be a discrete feature space, $\mathbb{N} = \{1, 2, 3, \cdots\}$ be the sequence of natural numbers as the value domain of c in formula (1), $B \subseteq A$, and $a \in A - B$. Let $c = i$ when $|B| = i \in \mathbb{N}$, then

$$SIG_i(a,B,D) = \frac{1}{|B|+1} + \frac{|B|}{|B|+1}\partial_B(D)_i - \partial_{B\cup\{a\}}(D)_{i+1}. \quad (11)$$

Proof. It follows immediately from Theorem 4.

Theorem 6. Let $<U, A, D>$ be a discrete feature space, $\mathbb{N} = \{1, 2, 3, \cdots\}$ be the sequence of natural numbers as the value domain of c in formula (1), and $B \subseteq A$. Let $c = i$ when $|B| = i \in \mathbb{N}$, then $SIG_i(a, B, D) > 0$ for any $a \in A - B$.

Proof. According to Theorem 4, we have

$$SIG_i(a,B,D) = \partial_B(D)_{i+1} - \partial_{B\cup\{a\}}(D)_{i+1}.$$

It follows from Property 3 that $\partial_B(D)_{i+1} \geq \partial_{B\cup\{a\}}(D)_{i+1}$, which implies that $SIG_i(a, B, D) > 0$ for any $a \in A - B$.

Theorem 7. Let $<U, A, D>$ be a discrete feature space, $\mathbb{N} = \{1, 2, 3, \cdots\}$ be the sequence of natural numbers as the value domain of c in formula (1), and $B \subseteq A$. Let $c = i$ when $|B| = i \in \mathbb{N}$, then $SIG_i(a, B, D)$ is monotonic as $i \to |A|$.

Proof. According to the significance of attributes, we assume that the order of attributes is determined as follows:

$$a_{l_1} \succ a_{l_2} \succ a_{l_3} \succ \cdots \succ a_{l_{|A|-1}} \succ a_{l_{|A|}},$$

where the term $a_{l_k} \succ a_{l_{k+1}}$ means that attribute a_{l_k} is more important than $a_{l_{k+1}}$. For any $B \subseteq A$, without loss of generality, let $a_{l_k}, a_{l_{k+1}} \notin B$. It follows from Property 3 that

$$\partial_B^{c_t}(D) \geq \partial_{B\cup\{a_{l_k}\}}^{c_t}(D) \geq \partial_{B\cup\{a_{l_k}\}\cup\{a_{l_{k+1}}\}}^{c_t}(D),$$

and

$$\partial_B^{c_t}(D) - \partial_{B\cup\{a_{l_k}\}}^{c_t}(D) \geq \partial_{B\cup\{a_{l_k}\}}^{c_t}(D) - \partial_{B\cup\{a_{l_k}\}\cup\{a_{l_{k+1}}\}}^{c_t}(D),$$

where $\partial_B^{c_t}(D)$ denotes the fuzzy dependency when $c = c_t$ and c_t is a distance parameter. By Theorem 3, when $c = |B| = i$, we have

$$\partial_B(D)_i = 1 - \frac{c_t}{i}\left(1 - \partial_B^{c_t}(D)\right),$$

$$\partial_{B\cup\{a_{l_k}\}}(D)_{i+1} = 1 - \frac{c_t}{i+1}\left(1 - \partial_B^{c_t}(D)\right),$$

$$\partial_{B\cup\{a_{l_k}\}\cup\{a_{l_{k+1}}\}}(D)_{i+2} = 1 - \frac{c_t}{i+2}\left(1 - \partial_{B\cup\{a_{l_k}\}}^{c_t}(D)\right).$$

From Theorem 5, it follows that

$$\begin{aligned} SIG_i(a_{l_k}, B, D) &= \frac{1}{i+1} + \frac{i}{i+1}\partial_B(D)_i - \partial_{B\cup\{a_{l_k}\}}(D)_{i+1} \\ &= \frac{1}{i+1} + \frac{i}{i+1}\left(1 - \frac{c_t}{i}\left(1 - \partial_B^{c_t}(D)\right)\right) - \left(1 - \frac{c_t}{i+1}\left(1 - \partial_{B\cup\{a_{l_k}\}}^{c_t}(D)\right)\right) \\ &= 1 - \frac{c_t}{i+1}\left(1 - \partial_B^{c_t}(D)\right) - 1 + \frac{c_t}{i+1}\left(1 - \partial_{B\cup\{a_{l_k}\}}^{c_t}(D)\right) \\ &= \frac{c_t}{i+1}\left(\partial_B^{c_t}(D) - \partial_{B\cup\{a_{l_k}\}}^{c_t}(D)\right). \end{aligned}$$

Similarly,

$$SIG_{i+1}(a_{l_{k+1}}, \{a_{l_k}\} \cup B, D) = \frac{c_t}{i+2}\left(\partial_{B\cup\{a_{l_k}\}}^{c_t}(D) - \partial_{B\cup\{a_{l_k}\}\cup\{a_{l_{k+1}}\}}^{c_t}(D)\right).$$

Thus, $SIG_i(a_{l_k}, B, D) \geq SIG_{i+1}(a_{l_{k+1}}, \{a_{l_k}\} \cup B, D)$, namely, $SIG_i(a, B, D)$ is monotonically decreasing with the increase of i.

Along with the fact that $SIG_i(a,B,D) > 0$ for any $a \in A-B$, we know that $SIG_i(a,B,D)$ is bounded and monotonic. It can be concluded that $SIG_i(a,B,D)$ converges as $i \to |A|$.

The iterative computation of attribute significance takes full advantage of the monotonic property of rough approximations. Theorem 5 shows that the significance of an attribute can be obtained by adjacent dependency functions. Theorems 6 and 7 prove that attribute significance by iterative computation still converges as the number of features grows to the cardinality of attribute set A.

In the case of natural sequence, the proposed fuzzy rough computation essentially uses the following measure to calculate the similarity of two samples with categorical attributes.

$$\left(R_B\right)_i(x_k,x_l) = \frac{1}{|B|}\left|\{a \in B : a(x_k) = a(x_l)\}\right|, \tag{12}$$

where $i = |B|$.

It is easy to see that formula (12) is different from formula (1), in which the parameter c is a constant number. The membership degrees of a fuzzy similarity relation computed by formula (12) cannot get small even when only a few of the features are included in rough computation. In addition, it has more advantages than equivalence relations in characterizing the similarity of categorical data. For example, there are two samples with 100 categorical attributes. For classical rough sets, the two samples are equivalent (similar) if and only if the two samples take the same value on each attribute. If there is only one attribute such that the two samples have different attribute values on it, then the two samples are not equivalent (similar). If we use formula (12) to characterize the similarity of the two samples, then the similarity degree of the two samples is 0.99. The two samples are almost equivalent. Hence, formula (12) can take some important characteristics of categorical data into consideration in describing the similarity of samples. Such representation was recalled by Zhao and Yao in 2007 [55]. The difference is that they used a threshold to control the similarity. For the full content, the reader may refer to the literature.

From formula (12), the definitions of fuzzy lower approximation, fuzzy positive region, and dependency function can be rewritten as follows.

$$\underline{R}_B(D_h)_i(x_k) = \min_{x_l \notin D_h}\{1-(R_B)_i(x_k,x_l)\}, \tag{13}$$

$$POS_B\left(D\right)_i = \bigcup_{h=1}^{r}\underline{R}_B\left(D_h\right)_i, \tag{14}$$

$$\partial_B(D)_i = \frac{\sum_{x_i \in U} POS_B(D)_i(x_l)}{|U|}. \tag{15}$$

where $i = |B|$.

According to formulas (11) - (15), the fuzzy rough computation model can be conducted. It overcomes the shortcoming that is caused by considering c as a constant number in rough computation. This is because the proposed model can elaborate the uncertainty of rough computation more subtly for symbolic data than classical fuzzy rough sets or Pawlak's rough sets.

In the following, we apply the proposed model to the attribute reduction of categorical data and verify its feasibility and effectiveness.

IV. Attribute-reduction algorithm for categorical data

In this section, we employ formulas (11) - (15) to compute a fuzzy dependency function and use it as a measure for evaluating an attribute subset. In addition, a heuristic algorithm for attribute reduction is designed, and its computational complexity is analyzed.

Algorithm: Fuzzy rough computation algorithm (FRC)
Input: Data table $<U, A, D>$, threshold δ // δ is set to stop the algorithm.
Output: one optimal attribute subset red.

1: Initialize: $red = \varnothing$, $B = A - red$, start=1; // red is the variable container to store the selected attributes and B is for the left attributes.
2: Compute $U/D = \{D_1, D_2, \cdots D_r\}$. Define a $1\times|U-D_k|$ null vector for each $x_j \in D_k$, i.e., let $sv(j) = (0)_{1\times|U-D_k|}$ for each $x_j \in D_k$, $k = 1,2,\cdots,r$.
3: while start
4: Define an $n \times r$ null matrix $lower\_appr = zeros(n,r)$, where $n = |U|$.
5: for each $a_l \in B$
6: for each $x_i \in D_k$, $k = 1,2,\cdots,r$
7: Compute vector $(r_{ij})_{1\times|U-D_k|}$, where $x_j \in U - D_k$, $r_{ij} = 1$ if $a_l(x_i) = a_l(x_j)$, otherwise, $r_{ij} = 0$.
8: Let $temp = sv(i) + (r_{ij})_{1\times|U-D_k|}$.
9: Let $lower\_appr(i,k) = \min\left\{1 - \frac{1}{|red|+1}temp\right\}$.
10: end for
11: Compute

$$\partial_{red\cup\{a_l\}}\left(D\right)_{|red|+1} = sum\left\{\max_{k=1}^{r}\left\{lower\_appr(i,k)\right\}\right\}\Big/|U|.$$

12: end for
13: Find a_k with maximum value $\partial_{red\cup\{a_k\}}\left(D\right)_{|red|+1}$.
14: Compute

$$SIG_{|red|}(a,red,D) = \frac{1}{|red|+1} + \frac{|red|}{|red|+1}\partial_B(D)_{|red|} - \partial_{red\cup\{a_k\}}(D)_{|red|+1}.$$

15: if $SIG_{|red|}(a,red,D) > \delta$
16: $red = red \cup a_k$.
17: for any $x_i \in D_k$, $k = 1,2,\cdots,r$
18: compute $(r_{ij})_{1\times|U-D_k|}$, where $x_j \in U - D_k$, $r_{ij} = 1$ if $a_k(x_i) = a_k(x_j)$, otherwise, $r_{ij} = 0$.
19: Let $sv(i) = sv(i) + (r_{ij})_{1\times|U-D_k|}$.

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.2019.2949765, IEEE Transactions on Fuzzy Systems

20: end for
21: $B = B - red$
22: else
23: start = 0;
24: end if
25: end while
26: return *red* .

In this algorithm, a pre-set null vector is designed for each sample in Step 2. It is first stored, then called by the algorithm. The complexity of Step 2 is in $O(n|U-D_p|)$, where $|D_p| = \min_{D_i \in U/D}\{|D_i|\}$. For the main block of the algorithm, the worst search time for an optimal attribute subset will result in $m(m+1)/2$ loops to evaluate the dependency function. In each loop, for each sample, the sample needs to be compared $n|U-D_p|$ times with other samples with different class labels. Thus, the time complexity from Step 3 to Step 26 is $O((m^2+m)n|U-D_p|/2)$, and the overall complexity of Algorithm 2 is in $O(n|U-D_p|+m(m+1)n|U-D_p|/2)$. It is just in $O(m(m+1)n|U-D_p|/2)$. Although classical fuzzy-rough-set-based algorithms [29] - [36] can also deal with categorical data sets, the computational complexity for most of these algorithms is $O(m(m+1)n^2)$. From the perspective of space efficiency, the proposed algorithm assigns each sample with a vector whose length equals the number of samples with different decision labels from itself, while the algorithms based on matrix calculation assign each sample with a vector whose length equals the number of all samples. Furthermore, each vector variable in the FRC algorithm can release the computational space after the calculation of the vector, while a matrix variable in matrix-based algorithms can only release the computational space after the calculation of all the samples. These algorithms often overflow memory for larger-scale data sets and make computation fail. Obviously, the proposed algorithm is more efficient compared to those classical algorithms.

V. Experimental analysis

As analyzed in the above sections, although fuzzy rough sets can deal with data with both real-valued and normal attributes, they will be degraded to classical rough set models when facing categorical data. This means that fuzzy rough set models have the same drawback as classical rough sets do when they are used to handle symbolic data. The proposed method uses fuzzy similarity relations to characterize the similarity between samples with categorical attributes and presents a new fuzzy rough set model to approximately describe a decision variable. The shortcomings of classical rough set models are effectively overcome. In this section, we verify the viability and effectiveness of our method by comparing some existing algorithms. We first consider comparing our algorithm with fuzzy rough sets (FFRS) [25], as well as other representative attribute-reduction algorithms that are considered as to be better for categorical data, namely, classical rough sets (RS) [1], variable precision rough sets (VPRS) [39] and consistency algorithms (CONSIS) [56]. Three indexes are compared: the number of selected attributes, classification accuracy of the reduced data sets, and reduction time. All the algorithms are run in Matlab 2013b and a hardware environment with an Intel (R) Core (TM) i7-4790 CPU @ 3.60 GHz &16.0 GB RAM.

We employ 10-fold cross-validation to conduct these experiments. The original data set is equally divided into ten parts; one of them is used for testing; the remaining nine parts are used as the training set for attribute reduction. A classifier is then learned with the reduced training set, and the classification accuracy is calculated on the reduced testing data. After 10 loops, the average value of the classification accuracies are calculated and used as the final performance value. Two selected classifiers in WEKA [57], i.e., trees.J48 (C4.5) and bayes.NaiveBayes, are used to evaluate these attribute reduction algorithms. All the parameters in C4.5 are set to the default values. We download ten public data sets from the UCI Machine Learning Repository [58]. There are eight data sets that include only categorical attributes, and two mixed data sets with categorical and numerical attributes. All the numerical attributes are preprocessed by discretization using the fuzzy C-means clustering (FCM) technique, and each numerical attribute is discretized into four intervals. These data sets are described in Table 1.

Table 1 Description of data sets

NO	Data set	Samples	Attributes	Classes
1	Car	1728	6	4
2	Kr-vs-kp	3196	36	2
3	Lphography	148	18	4
4	Monk	432	6	2
5	Mushroom	8124	22	2
6	Spect	267	22	2
7	Soybean	683	35	19
8	Tic-tac-toe	958	9	2
9	Horse	369	23	2
10	Heart	270	13	2

In the VPRS algorithm, a parameter λ is introduced to control the variable precision. The value of λ is set to vary from 0.5 to 1 with a step of 0.015. To make a fair comparison, in the FRC algorithm, the value of δ is also set from 0.05 to 0.35 with a step of 0.01. Because different data sets use different attribute subsets to yield the best classification accuracy, in the following experiments, we only compare the experimental results corresponding to the highest classification accuracies.

Table 2 presents the average sizes of selected attribute sets with these algorithms. It is easily seen that these algorithms can effectively reduce the number of attributes. The average sizes of the selected attribute subsets with these methods are approximately the same except data sets Car, Kr-vs-kp, Tic-tac-toe, and Horse. For the Tic-tac-toe, Kr-vs-kp, and Car data sets, the numbers of selected attributes with the RS and

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.2019.2949765, IEEE Transactions on Fuzzy Systems

FFRS algorithms are roughly the same but less than the other three algorithms. Furthermore, the resulting classification performances are relatively low, as shown in Tables 3 and 4. This shows that the RS and FFRS algorithms remove too many attributes, and even some attributes that are beneficial to the classification are deleted. This implies that the FFRS algorithm has the same drawback as the RS algorithm when it is used to handle symbolic data. Compared to the results of VPRS, CONSIS, and algorithms, the average sizes of selected attribute subsets with FRC are relatively small, but the corresponding classification accuracies are the highest in most cases, as shown in Tables 3 and 4. These results show that the FRC algorithm not only can delete redundant attributes but also retains attributes with more classification information.

Table 2 Numbers of selected features

Data set	Raw data	RS	FFRS	VPRS	CONSIS	FRC
Car	6	2.0	2.8	5.2	5.0	5.0
Kr-vs-kp	36	4.8	4.2	8.6	7.0	8.8
Lphography	18	5.8	5.8	6.2	5.8	5.6
Monk	6	1.0	1.0	2.2	2.0	2.0
Mushroom	22	5.0	4.4	5.0	4.8	4.2
Spect	22	2.0	4.8	2.6	3.2	3.2
Soybean	35	14.0	14.0	14.0	13.0	11.4
Tic-tac-toe	9	1.0	1.5	6.8	8.0	8.0
Horse	22	5.2	4.6	4.8	5.2	2.2
Heart	13	5.8	7.2	7.8	7.2	7.8
Average	18.9	4.66	5.03	6.32	6.22	5.82

Table 3 Comparison of classification accuracies of reduced data with C4.5

Data set	Raw data	RS	FFRS	VPRS	CONSIS	FRC
Car	96.25 ± 1.40	77.78 ± 3.81	78.78 ± 2.09	86.70 ± 9.45	89.31 ± 10.71	96.31 ± 0.74
Kr-vs-kp	96.34 ± 0.45	76.81 ± 1.85	77.38 ± 1.86	94.09 ± 1.46	94.09 ± 1.21	96.72 ± 1.87
Lphography	80.09 ± 8.41	76.86 ± 3.86	82.43 ± 6.99	83.14 ± 7.57	78.19 ± 2.66	84.47 ± 6.67
Monk	81.36 ± 6.63	74.98 ± 2.44	75.02 ± 6.07	75.63 ± 6.67	74.98 ± 2.74	83.63 ± 6.07
Mushroom	99.93 ± 0.14	100 ± 0.00	99.43 ± 0.23	98.52 ± 0.18	100 ± 0.00	100 ± 0.00
Spect	80.14 ± 7.51	79.42 ± 4.45	80.93 ± 4.75	80.47 ± 5.85	77.16 ± 2.94	80.51 ± 6.27
Soybean	88.93 ± 5.72	87.69 ± 4.26	90.78 ± 3.03	91.67 ± 2.07	87.25 ± 1.92	90.48 ± 5.32
Tic-tac-toe	87.99 ± 3.39	65.36 ± 3.61	65.36 ± 4.98	82.13 ± 7.88	86.25 ± 1.82	89.54 ± 2.13
Horse	93.47 ± 2.06	93.11 ± 2.04	92.91 ± 2.93	92.11 ± 2.26	93.01 ± 2.85	95.13 ± 2.59
Heart	76.67 ± 2.11	74.44 ± 15.35	80.37 ± 4.66	80.15 ± 7.05	82.96 ± 5.77	83.85 ± 6.18
Average	88.12 ± 3.78	80.64 ± 4.17	82.34 ± 3.76	86.36 ± 5.04	86.32 ± 3.26	90.69 ± 3.78

Table 4 Comparison of classification accuracies of reduced data with NaiveBayes

Data set	Raw data	RS	FFRS	VPRS	CONSIS	FRC
Car	86.28 ± 1.25	77.78 ± 3.81	78.78 ± 2.09	80.91 ± 8.46	82.99 ± 6.37	86.29 ± 1.53
Kr-vs-kp	85.98 ± 1.26	77.22 ± 1.99	77.22 ± 1.77	91.11 ± 1.48	93.84 ± 1.18	93.84 ± 1.11
Lphography	83.19 ± 5.50	80.28 ± 3.10	83.81 ± 5.53	84.19 ± 7.65	80.95 ± 4.21	85.03 ± 4.83
Monk	66.42 ± 3.99	66.65 ± 5.65	66.69 ± 6.44	66.68 ± 6.44	66.65 ± 5.65	70.65 ± 5.01
Mushroom	88.95 ± 1.52	93.84 ± 0.52	93.84 ± 1.57	94.32 ± 1.66	91.79 ± 2.06	97.03 ± 1.33
Spect	77.70 ± 5.78	79.42 ± 4.45	80.28 ± 4.84	80.92 ± 6.10	76.06 ± 3.87	80.51 ± 6.06
Soybean	89.07 ± 4.44	86.22 ± 4.90	89.47 ± 3.98	90.65 ± 3.00	87.39 ± 2.80	91.07 ± 4.86
Tic-tac-toe	70.98 ± 2.25	65.36 ± 3.61	65.36 ± 4.98	67.94 ± 3.72	69.73 ± 3.05	80.56 ± 1.16
Horse	93.77 ± 2.92	91.85 ± 1.88	92.95 ± 3.36	94.29 ± 3.48	91.39 ± 1.18	94.13 ± 2.20
Heart	84.81 ± 4.22	73.73 ± 14.61	81.48 ± 7.01	80.25 ± 6.49	82.96 ± 5.77	85.93 ± 5.36
Average	82.72 ± 3.31	79.24 ± 4.45	80.98 ± 4.16	83.13 ± 4.85	82.38 ± 3.61	86.51 ± 3.35

Tables 3 - 4 show the classification results of the reduced data sets based on these algorithms. The highest accuracies are underlined.

It can be seen from these results that the classification performance with the RS algorithm is obviously lower than the other four algorithms. For the FRC algorithm, not only are the number of selected features fewer, but also the classification accuracies are the highest. The FRC method achieves the highest accuracy 16 times in 20 cases. The VPRS, CONSIS, FFRS, and RS methods only obtain the highest accuracies in 3, 2, 1, and 1 cases, respectively. Furthermore, it is easily seen from the average accuracies that FRC outperforms any other reduction algorithm in terms of C4.5 and NaiveBayes classifiers. The reason why the FRC algorithm achieves such good performance may be that the FRC method fully considers the fuzzy information in categorical data and better elaborates the classification information of samples. Because the RS method is sensitive to noisy data, the corresponding performance is the worst.

Next, we analyze the influence of the stopping threshold δ on our proposed algorithm. The threshold δ is considered as a parameter to control the process of feature search; in other words, it is a threshold of stopping feature search. In the following, we demonstrate the effect of the stopping threshold

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.2019.2949765, IEEE Transactions on Fuzzy Systems

on the performance of the algorithm. Here, we randomly select six data sets and show the experimental results of these data sets on the C4.5 classifier. It is easily observed from Fig. 1 - 6 that the threshold δ has a great effect on the FRC algorithm. The number of selected attributes decreases monotonically with an increase in the value of δ , and the classification accuracies also have different changes. Fortunately, most of the data sets can achieve high accuracy in a wider area. For instance, the Car, Mushroom, and Horse data sets achieve good accuracies in terms of a few attributes selected in attribute reduction and exhibit a wide range of accuracy stability in their respective parameter domains. These results show that the proposed algorithm is stable and can provide an effective way to select an optimal subset of attributes for classification. The optimal location of the parameter δ is different among these datasets. Therefore, we should train parameters before reducing a data set in experiments. As described above, we trained the optimal values of δ by choosing its value from between 0.05 and 0.35 with a step of 0.01. It is easily observed from Fig. 1- 6 that the optimal values of the stopping threshold δ are taken in the interval [0.05, 0.15] in most cases.

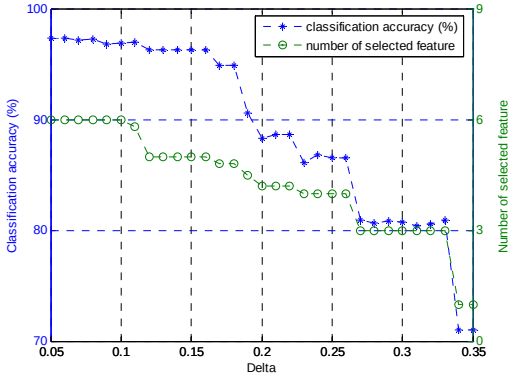

Fig. 1 Effect of δ on attribute reduction (Car)

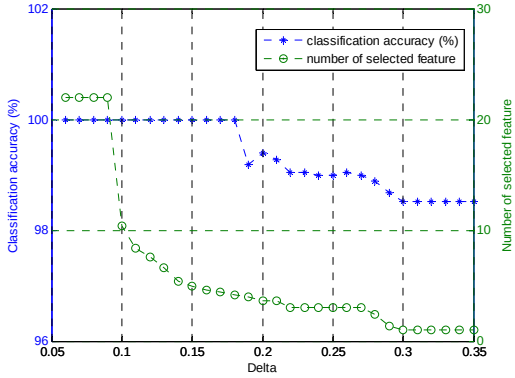

Fig. 2 Effect of δ on attribute reduction (Mushroom)

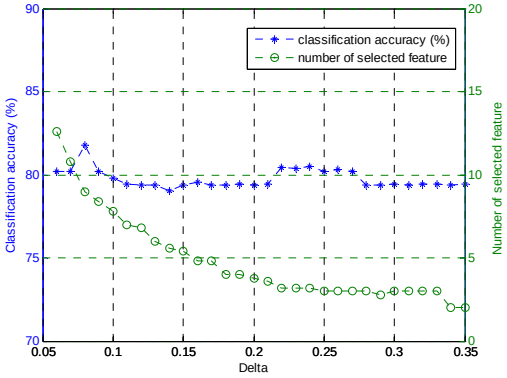

Fig. 3 Effect of δ on attribute reduction (Spect)

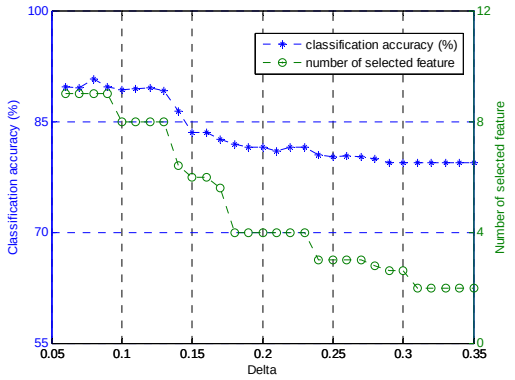

Fig. 4 Effect of δ on attribute reduction (Tic-tac-toe)

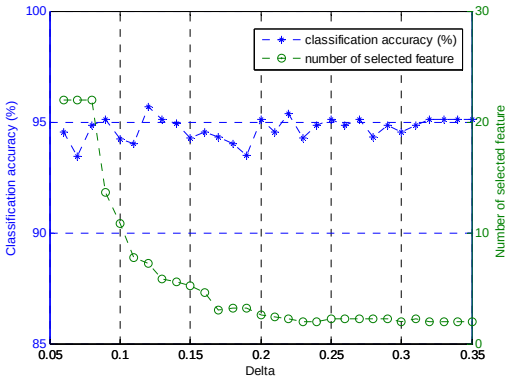

Fig. 5 Effect of δ on attribute reduction (Horse)

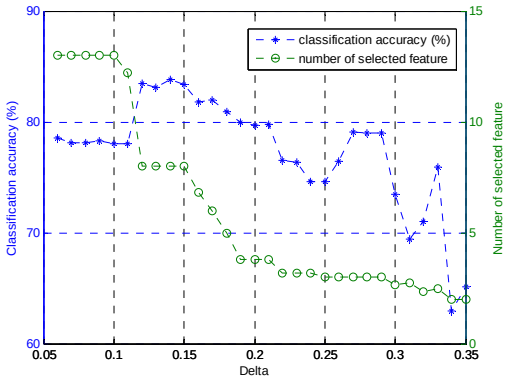

Fig. 6 Effect of δ on attribute reduction (Heart)

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.2019.2949765, IEEE Transactions on Fuzzy Systems

To demonstrate the selected attribute subset of a data set, we use the FFRS, VPRS, and FRC algorithms to reduce the eight categorical data sets. We reduce each data set based on the parameters where the classification accuracies were obtained in the above experiments. The selected attribute subsets are shown in Table 5. For the Car, Kr-vs-kp, Tic-tac-toe, and Monk datasets, the optimal attributes selected by FFRS are virtually subsets of the optimal attributes selected by VPRS or FRC. The corresponding classification accuracies of the reduced data sets are also lower than those of the VPRS and FRC algorithms, as shown in Tables 3 and 4. Most of the attributes selected by VPRS and FRC are the same, but the classification accuracies of FRC are higher than those of VPRS. This result shows that FRC can select the optimal attributes for classification. For the Car dataset, four of the five attributes selected by the two algorithms are identical; namely, they are 2 (maintenance), 4 (passenger capacity), 5 (luggage boot), 6 (safety). One attribute is different. The VPRS algorithm chooses 3 (number of doors), and the FRC algorithm chooses 1 (buying price). This indicates that *buying price* has more classification information than *number of doors*. The Kr-vs-kp, Tic-tac-toe, and Monk datasets also have a similar situation.

For the Soybean, Lphography, and Mushroom datasets, most of the attributes selected by the three algorithms are the same. The difference in the attribute subsets indicates that there are multiple attribute subsets that have acceptable classification power for a given classification task. For the Spect dataset, the selected attribute subsets were identical, and the classification accuracies were almost the same for the VPRS and FRC algorithms. The marginal differences could be because the selected attribute subsets were presented by reducing the entire dataset, whereas the classification accuracies were given based on ten-fold cross-validation.

Table 5. Optimal attributes selected by FFRS, VPRS, and FRC algorithms

Data set	FFRS	VPRS	FRC
Car	4, 6, 1	4, 6, 5, 3, 2	4, 6 ,1, 2, 5
Kr-vs-kp	21, 10, 29, 14	21, 33, 10, 35, 11, 1, 15, 6, 14	21, 10, 29, 14, 28, 1, 15, 16, 6
Lphography	18, 2, 13, 14, 15, 5	13, 2, 15, 18, 14, 16	18, 2, 10, 13, 14, 15
Monk	5	5, 2	5, 1
Mushroom	5, 20, 8, 12	5, 8, 20, 12, 3	5, 9, 14, 20
Spect	17, 18, 1, 4, 16	1, 17, 18	17, 18, 1
Soybean	18, 26, 11, 12, 35, 29, 22, 1, 3, 6, 7, 10, 4 5	21, 29, 35, 26, 18, 1, 22, 3, 15, 14, 8, 9, 17, 11	18, 26, 11, 28, 22, 15, 29, 4, 1, 3, 24, 7, 10
Tic-tac-toe	2, 3	5, 1, 2, 3, 4, 7, 6, 8	1, 2, 3, 5, 8, 9, 7, 4

Most attribute reduction algorithms with fuzzy rough sets are based on matrix computation, but the algorithm proposed in this paper is based on vector computation. Thus, the computational complexity of the proposed algorithm is lower than that of classical fuzzy rough sets. The comparison of the running time of reduction in the FRC and FFRS algorithms is shown in Fig. 7 - 10, where the vertical axis represents reduction time (in seconds), and the horizontal axis represents the number of selected attributes. From Fig. 7 – 10, it is easily observed that the FRC algorithm executed much faster than the FFRS algorithm in the four data sets presented. The FRC algorithm is more than three times faster than the FFRS algorithm when selecting the same number of attributes.

Finally, numerical experiments on the monotonicity of the attribute-significance measure are demonstrated on some data sets. From Fig. 11 – 14, it is seen that the measure of attribute significance decreases monotonically with the increase of the number of selected attributes. However, a closer look shows that for the Mushroom dataset, when the number of attributes is 17, the corresponding attribute significance is slightly lower than it is when the number of attributes is 18. This may be caused by the noise of the data itself, but the overall trend is for the attribute significance to decrease gradually with an increasing number of attributes.

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.2019.2949765, IEEE Transactions on Fuzzy Systems

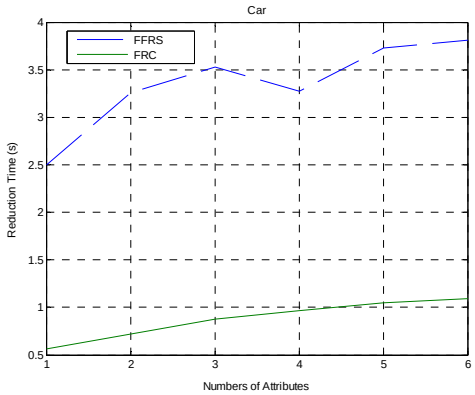

Fig. 7 Running time of reduction on Car

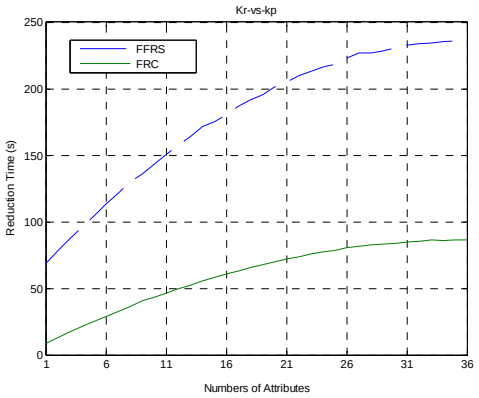

Fig. 8 Running time of reduction on Kr-vs-kp

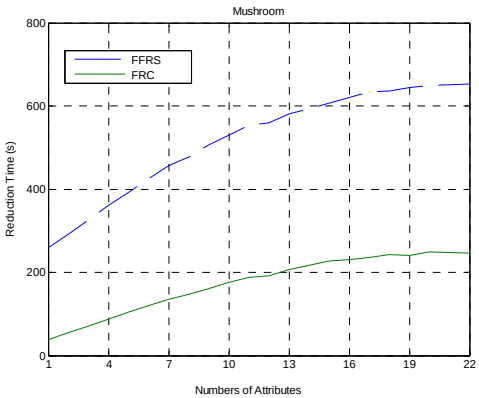

Fig. 9 Running time of reduction on Mushroom

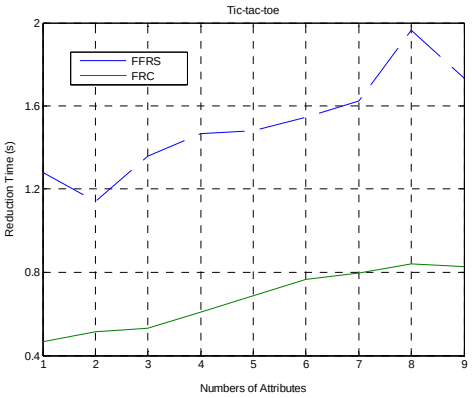

Fig. 10 Running time of reduction on Tic-tac-toe

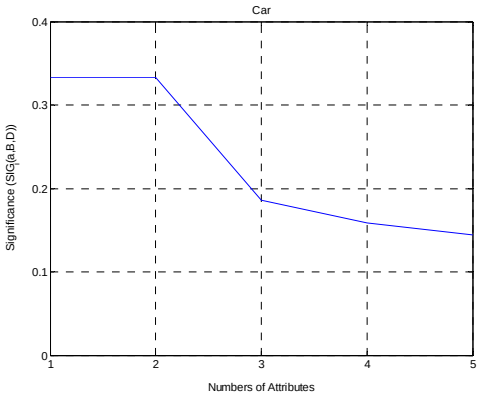

Fig. 11 Monotonicity of the attribute-importance measure with the selected number of attributes in Car

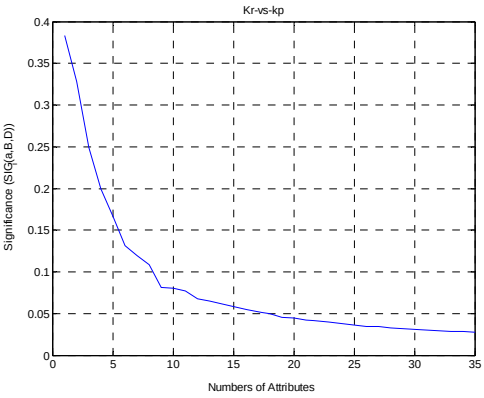

Fig. 12 Monotonicity of the attribute-importance measure with the selected number of attributes in Kr-vs-kp

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.2019.2949765, IEEE Transactions on Fuzzy Systems

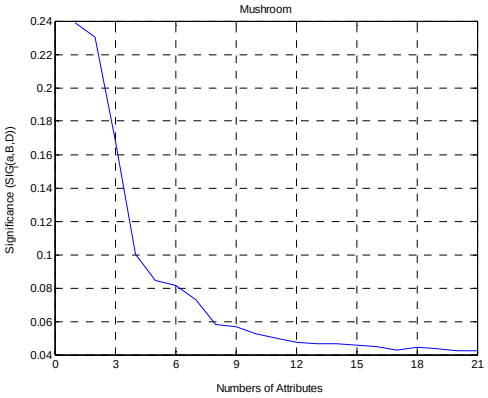

Fig. 13 Monotonicity of the attribute-importance measure with the selected number of attributes in Mushroom

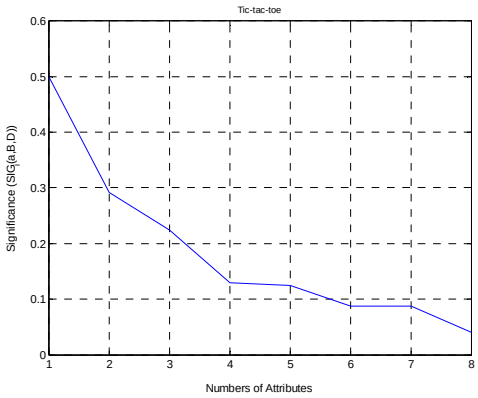

Fig. 14 Monotonicity of the attribute-importance measure with the selected number of attributes in Tic-tac-toe

VI. CONCLUSIONS

There are two important types of data in rough-set-based data analysis: categorical and numerical data. Fuzzy rough sets are used mainly to deal with numerical data, and classical rough sets are used mainly for categorical data. However, classical rough sets are sensitive to noise because this model is based on the stringent conditions of definitions of equivalence relations. In this paper, we propose a novel fuzzy-rough-set model to perform rough computations for categorical data. This model employs a fuzzy similarity relation to define rough approximations of a target decision and overcome the deficiency of classical rough sets. In the proposed model, there is a parameter that controls the fuzzy similarity between samples in a discrete feature space. If the parameter is viewed as a constant number, the membership degrees of samples to decision classes, computed by fuzzy rough approximation, may become very small for high-dimensional data when only a few of the features are included in rough computation. To solve the problem, we adopt an iterative method for rough computations and develop a rough-fuzzy iterative-computational model for categorical data. With ten data sets from the UCI data source, a series of experiments of attribute reduction are done for evaluating the proposed method. The experimental analysis indicates that the proposed method is effective for reducing redundant attributes and can keep high classification accuracy.

In the future, we will further study the applications of the proposed model in approximate reasoning and classification learning, and develop fuzzy-rough iterative computation theory for continuous data.

REFERENCES

[1] Z. Pawlak, “Rough sets,” International Journal of Computer and Information Sciences, vol. 11, no. 5, pp. 341–356, 1982.

[2] D. Miao, Y. Zhao, Y. Yao, H. X. Li, F. Xu, “Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model”, Information Sciences, vol.179, no. 24, pp. 4140-4150, 2009.

[3] T. Y. Lin, "Neighborhood systems – application to qualitative fuzzy and rough sets," In P. P. Wang, Advances in machine intelligence and soft computing, Department of Electrical Engineering, Duke University Durham, North Carolina, USA, 1997, pp. 132–155

[4] Q. H. Hu, D. Yu, J. F. Liu, and C. Wu, “Neighborhood-rough-set based heterogeneous feature subset selection,” Information Sciences, vol. 178, no. 18, pp. 3577–3594, 2008.

[5] Y. Yao, "Relational interpretations of neighborhood operators and rough set approximation operators, Information Sciences, vol. 101, pp. 239-259, 1998.

[6] S. Greco, B. Matarazzo, and R. "Slowinski, Rough approximation by dominance relations," International Journal of Intelligent Systems, vol. 17, pp. 153-171, 2002.

[7] M. Inuiguchi, Y. Yoshioka, Y.Kusunoki, "Variable-precision dominance based rough set approach and attribute reduction," International Journal of Approximate Reasoning, vol. 50, pp. 1199–1214, 2009.

[8] D. Dubois and H. Prade, “Rough fuzzy sets and fuzzy rough sets,” International Journal of General Systems, vol. 17, pp. 191–208, 1990.

[9] N. N. Moresi, M. M. Yankout, "Axiomatic for fuzzy rough sets," Fuzzy sets and systems, vol.100, no.1-3, pp.327-342,1998.

[10] 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.

[11] D. Kim, "Data classification based on tolerant rough set," Pattern Recognition, vol. 34, no. 8, pp. 1613–1624, 2001.

[12] J. Zhan, Q. Wang, "Certain types of soft coverings based rough sets with applications," International Journal of Machine Learning & Cybernetics, vol. 3, no. 5, pp.1065-1076, 2019.

[13] G. Lang, M., H. Fujita, Q. Xiao, "Related families-based attribute reduction of dynamic covering decision information systems," Knowledge -Based Systems, https://doi.org/10.1016/ j.knosys.2018.05.01 9.

[14] R. Slowinski, D. Vanderpooten, "A generalized definition of rough approximations based on similarity," IEEE Transactions on Knowledge and Data Engineering, vol. 2, pp. 331-336, 2000.

[15] A. Tan, W. Wu, S. Shi, S. Zhao, "Granulation selection and decision making with multigranulation rough set over two universes," Internatio-nal Journal of Machine Learning & Cybernetics, vol.10, no.9, pp. 2501-2513, 2019.

[16] B. Sun, W. Ma, X. Chen, "Variable precision multigranulation rough fuzzy set approach to multiple attribute group decision-making based on λ – similarity relation," Computers and Industrial Engineering, DOI:10. 1016 / j.cie.2018.10.009.

[17] W. Zhu and F. Wang, “On three types of covering rough sets,” IEEE Transactions on Knowledge and Data Engineering, vol. 19, pp. 1131–1144, 2007.

[18] Y. Qian, Q. Wang, H. Cheng, J. Liang, C. Dang, "Fuzzy-rough feature selection accelerator," Fuzzy Sets and Systems, vol., 258, pp. 61-78, 2016.

[19] C. Wang, M. Shao, Q. He, Y. Qian, Y. Qi, "Feature subset selection based on fuzzy neighborhood rough sets," Knowledge-Based Systems, vol.111, no.1, pp.173-179, 2016.

[20] 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.

[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] W. Wu, M. Shao, X. Wang, "Using single axioms to characterize (S, T)-intuitionistic fuzzy rough approximation operators," International Journal of Machine Learning & Cybernetics, vol. 10, no. 1, pp. 27-42, 2019.
[27] 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
[28] W. Wu and W. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," Information Sciences, vol. 159, pp. 233-254, 2004.
[29] 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.
[30] R. B. Bhatt, M. Gopal, "On the compact computational domain of fuzzy rough sets," Pattern Recognition Letter, vol. 26, pp. 1632–1640, 2005.
[31] 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.
[32] W. Ding, C. Lin, and Z. Cao, "Deep neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping PSO with nearest-neighbor memeplexes,"IEEE Transactions on Cybernetics, vol.49, no. 7, pp. 2744-2757, 2019.
[33] J. Dai, H. Hu, W. Wu, Y. Qian, D. Huang, Maximal discernibility pairs based approach to attribute reduction in fuzzy rough sets, IEEE Transactions on Fuzzy Systems, vol. 26, no.4, pp. 2174-2187.
[34] 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.
[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] 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.
[37] F. Fernandez-Riverola, F. Diaz, and J. M. Corchado, "Reducing the memory size of a fuzzy case-based reasoning system applying rough set techniques," IEEE Trans. Systems and Cybernetics Part C-Applications and Rev., vol.37, no.1, pp. 138-146, 2007.
[38] 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.
[39] W. Ziarko, "Variable precision rough set model," Journal of Computer and System Sciences, vol. 46, pp. 39–59, 1993
[40] Y. Yao, "Probabilistic rough set approximations," International Journal of Approximate Reasoning, vol. 49, no. 2, pp. 255–271, 2008
[41] I. Duntsch, G. Gediga, "Uncetainty measures of rough set prediction," Artificial Intelligence, vol. 106, pp.109-137, 1998
[42] 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.
[43] J.Y. Liang, J. Mi, W. Wei, F. Wang, "An accelerator for attribute reduction based on perspective of objects and attributes," Knowledge-Based Systems, vol.44, pp. 90-100, 2013.
[44] D. Slezak, "Approximate entropy reducts," Fundam. Inf., vol. 53, no.3-4, pp. 365-390, 2002
[45] 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.
[46] 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.
[47] 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.
[48] F. Min, Z. Zhang, W. Zhai, and R. Shen, Frequent pattern discovery with tripartition alphabets, Information Sciences doi:10.1016/j.ins.2018.04.013.
[49] W. Ding, C. Lin, and Z. Cao, "Shared nearest neighbor quantum game-based attribute reduction with hierarchical co-evolutionary Spark and its consistent segmentation application in neonatal cerebral cortical surfaces," IEEE Transactions on Neural Network and Learning System, vol. 30, no. 7, pp. 2013-2027, 2019.
[50] L. Sun, Y. Qian, J. Xu, S. Zhang, Y. Tian, "Joint neihborhood entropy-based gene selection method with fisher score for tumor classification," Applied Intelligence, vol. 49, no.4, pp. 1245-1259, 2019.
[51] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, Feature selection based on neighborhood discrimination index, IEEE Transactions on Neural Networks and Learning Systems, vol.29, no.7, pp. 2986–2999, 2018.
[52] 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.
[53] X. Che, J. Mi, "Attributes set reduction in multigranulation approximation space of a multi-source decision information system, " International Journal of Machine Learning & Cybernetics, vol. 10, no. 9, pp. 2297-2311, 2019.
[54] 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.
[55] Y. Zhao, Y. Yao, F. Luo, "Data analysis based on discernibility and indiscernibility," Information Sciences, vol. 177, pp. 4959–4976, 2007
[56] M. Dash and H. Liu, "Consistency-based search in feature selection," Artificial Intelligence, vol. 151, no. 1/2, pp. 155–176, 2003.
[57] Weka 3: Data Mining Software in Java, https://www.cs.waikato.ac.nz/ml/weka/
[58] C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Databases, 1998. [Online]. Available: http://www. ics.uci.edu/~mlearn/MLRepository.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. 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.

Yan Wang received his B.Sc. degrees in mathematics from Bohai University in 2016. Now, she is a graduate student for a Master's 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 public- shed more than 50 papers in international journals and international conferences. His current research interests include rough sets, fuzzy sets, formal concept analysis, and granular computing.

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.

Degang Chen received his M.S. degree from Northeast Normal University, Chan- gchun, China, in 1994 and his Ph.D. degree from the Harbin Institute of Tech- nology, 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.

His research interests include fuzzy groups, fuzzy analysis, rough sets, and support vector machines.