Kernelized Fuzzy Rough Sets and Their Applications

Qinghua Hu, *Member*, *IEEE*, Daren Yu, Witold Pedrycz, *Fellow*, *IEEE*, and Degang Chen

Abstract—Kernel machines and rough sets are two classes of commonly exploited learning techniques. Kernel machines enhance traditional learning algorithms by bringing opportunities to deal with nonlinear classification problems, rough sets introduce a human-focused way to deal with uncertainty in learning problems. Granulation and approximation play a pivotal role in rough sets-based learning and reasoning. However, a way how to effectively generate fuzzy granules from data has not been fully studied so far. In this study, we integrate kernel functions with fuzzy rough set models and propose two types of kernelized fuzzy rough sets. Kernel functions are employed to compute the fuzzy T-equivalence relations between samples, thus generating fuzzy information granules in the approximation space. Subsequently fuzzy granules are used to approximate the classification based on the concepts of fuzzy lower and upper approximations. Based on the models of kernelized fuzzy rough sets, we extend the measures existing in classical rough sets to evaluate the approximation quality and approximation abilities of the attributes. We discuss the relationship between these measures and feature evaluation function ReliefF, and augment the ReliefF algorithm to enhance the robustness of these proposed measures. Finally, we apply these measures to evaluate and select features for classification problems. The experimental results help quantify the performance of the KFRS.

Index Terms—Rough set, fuzzy rough set, kernel, feature evaluation, feature selection.

1 Introduction

Rough set theory has received considerable attention in machine learning and pattern recognition in last decade. It has been found useful in dealing with imperfect and inconsistent information [32], which is quite often encountered in machine learning and data mining. The classical rough set model, proposed by Pawlak [32], dwells on Boolean equivalence relations, where objects taking the same values of feature are said to be indiscernible or equivalent. Pawlak's rough set model has been widely exploited in feature selection, attribute reduction, rule extraction, and reasoning in presence of uncertainty [15], [19], [40].

A certain disadvantage of the generic version of the Pawlak's rough sets stems from the fact that this model is concerned with categorical features assuming some discrete values [20], [21]. In fact, categorical, numerical, fuzzy and interval-valued features usually coexist in real-world databases [51], such as those existing in medical analysis and fault diagnosis [17], [18], [20], [21]. One of the feasible solutions to arrive at categorical variables is to divide the domain of the corresponding numerical feature into several intervals making use of a discretization algorithm. However, one has to be aware that discretization is typically an important source of information loss when dealing with complex data [20]. Another alternative to deal with numerical and fuzzy data is to develop a fuzzy rough set model [5], [6], [17], [18], [20], [27]. There are two important issues to be addressed when developing a fuzzy rough set model: generating fuzzy relations between the samples and inducing a set of fuzzy granules with the fuzzy relations; approximating fuzzy concepts with these induced fuzzy information granules.

Fuzzy rough sets have attracted attention in the last years. Dubois and Prade developed their first model [6], where fuzzy equivalence relations satisfying the properties of reflexivity, symmetry, and max-min transitivity form its cornerstone. In addition, in defining fuzzy lower and upper approximation operators used were t-norms and t-conorms [23]. Radzikowska and Kerre introduced a more general definition of fuzzy rough sets in [33]. They defined a broad family of fuzzy rough sets with respect to a fuzzy similarity relation where the fuzzy lower and upper approximations are determined by a border implicator and a t-norm, respectively. In [28], t-norms and T-residuated implications were introduced to define fuzzy rough sets. Mi and Zhang proposed a new definition of fuzzy rough sets based on the residual implication θ and its dual [27]. In [45], Yeung et al., reviewed the previous work and showed two approaches to define fuzzy rough sets which are based on arbitrary fuzzy relations.

Moris and Yakout replaced the min operator used in [6] to a class of t-norms and gave the axiomatics of upper and lower approximation operators. Wu and Zhang discussed the constructive and axiomatic approaches to fuzzy approximation operators based on general fuzzy relations and the min-max operators [43]. Yeung et al., discussed the characterization of different classes of generalized upper and lower approximation operators of fuzzy sets by forming different sets of axioms [45]. Fuzzy rough sets came with some successful applications such as feature

- Q.H. Hu and D.R. Yu are with the Harbin Institute of Technology, Harbin 150001, China. E-mail: {huqinghua, yudaren}@ hit.edu.cn.
- W. Pedrycz is with the Department of Electrical and Computer Engineering, University of Alberta, AB, Canada, and the Systems Science Institute, Polish Academy of Sciences, Warsaw, Poland. E-mail: pedrycz@ee.ualberta.ca.
- D. Chen is with the Department of Mathematics and Physics, North China Electric Power University, Beijing 102206, P.R. China.

Manuscript received 25 Jan. 2008; revised 31 Jan. 2009; accepted 14 Dec. 2009; published online 22 Dec. 2010.
Recommended for acceptance by D. Gunopulos.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-01-0054.
Digital Object Identifier no. 10.1109/TKDE.2010.260.

1041-4347/11/$26.00 © 2011 IEEE Published by the IEEE Computer Society

evaluation [18], attribute reduction [17], [20], [21], [49], [50], [52], [53], rule extraction [15], [39], [41], classification tree induction [1], medical analysis [14], stock prediction [40], and case-based reasoning [9] to point at the representative examples.

Although the models and applications of fuzzy rough sets have been discussed in a comprehensive manner, there is still an important problem to be addressed. As pointed out above, that granulation and approximation are two issues when applying fuzzy rough sets to real-world problems. The studies carried out so far are focused almost exclusively on defining fuzzy approximation operators and very little work was devoted to the problem of extracting fuzzy relations from data. Most of the fuzzy rough set models are constructed in the fuzzy granulated spaces induced by fuzzy T-similarity relations, but how to effectively generate fuzzy similarity relations from data has not been systematically discussed so far. In fact, it seems that fuzzy similarity relations are not easy to be computed in the setting of a given application. Nevertheless, the way to generate fuzzy relations from data has a direct impact on the performance of the resulting model. Therefore, it becomes important to develop a systematic and effective approach to determining fuzzy relations starting from available experimental data. Jensen and Shen claimed that fuzzy equivalence relations and fuzzy equivalent classes were used in their studies; yet no detailed algorithm has been offered in [20] and [21].

Moser showed that any kernel satisfying reflexivity and symmetry is at least $T_{\cos}$-transitive [29], [30]. Then, the relation computed with this kind of kernel functions is a fuzzy T-similarity relation. With the fuzzy relation generated by kernel functions, we can granulate the universe of discourse and form a family of fuzzy information granules. Next, such fuzzy granules can be used to approximate arbitrary subsets of the universe. We construct different kernel-based granulated spaces and form different fuzzy rough sets with various kernel functions. We will be referring to them as kernelized fuzzy rough sets (**KFRS**). KFRS constructs a bridge in-between kernel machines and rough sets. In this manner, some important theoretical findings in kernel methods can be introduced in rough set theory. The main reason of the use of kernels is that they support mapping the data into a high dimensional feature space in order to improve the classification power of linear machines without introducing significant computational overhead [36]. Support vector machines [4], multikernel regularized classifiers [42], kernel principal component analysis [35], and kernel canonical correlation [11] are just some of the algorithms that make use of kernels to deal with nonlinear problems. Kernelized fuzzy rough sets add a new member into the family of kernel machines. On the other hand, rough sets and their generalization such as fuzzy rough sets employ the idea of granular computing [47], [48] which is in line with a way people make rational decisions in complex environments. In essence, we granulate the complex universe into a collection of a limited number of information granules based on the relations between objects and roughly approximate the decision with lower and upper approximations. The main advantage of fuzzy rough sets is to achieve tractability, robustness, lower solution cost, and better rapport with reality. In this sense, kernelized fuzzy rough sets combine the advantages of kernel methods and rough sets. In this work, we show the combination of kernel and rough sets, discuss the model and properties of kernel fuzzy rough sets and elaborate on some potential applications of the proposed model. Moreover, we will reveal the relation between the kernel fuzzy rough sets and some other classical algorithms [32].

The contribution of the work is three folds. First, we integrate kernel functions with fuzzy rough sets and propose the model of kernelized fuzzy rough sets, which forms a bridge between kernel machines and rough set-based data analysis. Second, we discuss the relationship between the feature evaluation coefficients based on kernelized fuzzy rough sets and ReliefF. This analysis provides a new viewpoint to understand and extend rough sets. Finally, we show some generalized feature evaluation functions and attribute reduction algorithms based on the proposed model and validate the effectiveness of the proposed technique.

The paper is organized as follows: In Section 2, we present some pertinent preliminaries about rough sets, fuzzy operators, and fuzzy rough sets. Then, the model of kernelized fuzzy rough sets is introduced in Section 3. Some measures to evaluate the approximation quality are presented in Section 4. The relationship between kernel fuzzy rough sets and Relief algorithms is discussed in Section 5. We propose some feature evaluation functions and feature selection algorithms in Section 6. Experimental results are covered in Section 7. Finally, conclusions and the future work are included in Section 8.

2 PRELIMINARIES

Rough sets and fuzzy sets are two classes of powerful tools to deal with uncertainty and granularity of information. This section will review some definition and notations being used in the subsequent sections of this study [45].

2.1 Operations on Fuzzy Sets

Let U be a nonempty set, the class of all subsets and fuzzy subsets of U is denoted by $P(U)$ and $F(U)$, respectively. A binary operator T on the unit interval $I = [0, 1]$ is said be a triangular norm, if $\forall a, b, c \in I$, if the following conditions are satisfied

1. commutativity $T(a, b) = T(b, a)$;
2. associativity: $T(T(a, b), c) = T(a, T(b, c))$;
3. monotonicity: $a \leq c, b \leq d \Rightarrow T(a, b) \leq T(c, d)$; and
4. boundary condition: $T(a, 1) = a$, $T(1, a) = a$.

We say that a binary operator S on the unit interval I is a triangular-conorm (shortly, t-conorm or s-norm) if it is increasing, associative, and commutative and satisfies the boundary condition of the form $S(a, 0) = a$, $\forall a \in [0, 1]$.

A negator (complement) N is a decreasing mapping from $[0, 1] \rightarrow [0, 1]$ satisfying $N(0) = 1$ and $N(1) = 0$. The negator $N_s(a) = 1 - a$ is usually referred to as the standard negator. A negator N is said to be involutive if $N(N(a)) = a$, $\forall a \in [0, 1]$.

Given a t-norm T, a s-norm S, and a negator N, we say T and S are dual with respect to N if de Morgan laws are satisfied, namely, $S(a, b) = N(T(N(a), N(b)))$, $T(a, b) = N(S(N(a), N(b)))$, $\forall a, b \in [0, 1]$.

Let T, S, and N be a t-norm, a s-norm, and a negator, respectively. Given $A, B, C \in F(U)$,

$$\begin{tikzcd}
T \arrow[r, leftrightarrow, "N"] \arrow[d] & S \arrow[d] \\
\theta \arrow[r, leftrightarrow, "N"] & \sigma
\end{tikzcd}$$

Fig. 1. Relationship between operations on fuzzy sets.

1. if $\forall x \in U,\ C(x) = S(A(x), B(x))$, we denote $C = A \cup_S B$, and call C the S-norm union of A and B;
2. if $\forall x \in U,\ C(x) = T(A(x), B(x))$, we denote $C = A \cap_T B$, and call C the T-norm intersection of A and B.

Given a t-norm T, the binary operator θ: $\theta(a,b) = \sup\{c \in [0,1] : T(a,c) \le b\}$ is called an implication based on T. Moreover, θ is called a residual implication of T if T is lower semicontinuous. In this case θ is also called T-residuated implication.

Correspondingly, we introduce another fuzzy operator based on S: $\sigma(a,b) = \inf\{c \in [0,1] : S(a,c) \ge b\}$.

θ and σ are dual in terms of N if T and S are dual with respect to N, i.e., $\forall a,b \in [0,1]$: $\sigma(a,b) = N(\theta(N(a), N(b)))$ or $\theta(a,b) = N(\sigma(N(a), N(b)))$.

The relationships between fuzzy operations T, S, θ, and σ are shown as Fig. 1.

Tables 1 and 2 show some commonly encountered operators used in fuzzy reasoning, where T_M and S_M are called the standard min and max operators; T_P and S_P are algebraic product and probabilistic sum operators; T_L and S_L are Lukasiewicz norms; while $T_{\cos}$ and $S_{\cos}$ are named as cosine norms.

2.2 Rough Sets and Fuzzy Rough Sets

Given a nonempty and finite set U of objects, called universe, R is an equivalence relation on U if for $\forall x,y,z \in U$, we have 1) $R(x,x) = 1$, 2) $R(x,y) = R(x,y)$, and 3) $R(x,y) = R(y,z) \Rightarrow R(x,y) = R(x,z)$. The equivalence relation partitions the universe into a family of disjoint subsets, called equivalence classes. The equivalence class including x is denoted by $[x]_R$. We call $AS = <U,R>$ an approximation space. For arbitrary subset of objects $X \subseteq U$, the lower and upper approximations of X in $<U,R>$ are defined as [32]

$$\begin{cases} \underline{R}X = \{[x]_R \mid [x]_R \subseteq X\}, \\ \overline{R}X = \{[x]_R \mid [x]_R \cap X \neq \emptyset\}. \end{cases}$$

It is easy to show that $\underline{R}X \subseteq X \subseteq \overline{R}X$. We say X is a rough set in the approximation space if $\underline{R}X \neq \overline{R}X$; otherwise, we say X is definable. As to the rough set X, we use two unions of equivalent classes to approximate it. The lower approximation is the subset of objects whose equivalent class is completely contained by X, while the upper approximation is the subset of objects whose equivalent class at least has an object contained by X.

The above model employs an equivalence relation to granulate the universe and generate Boolean elemental granules. As pointed out by Zadeh in [47] and [48] that modes of information granulation in which the granules are Boolean play an important role in a wide variety of methods, approaches, and techniques. With the fuzzy information, fuzzy relations and information granules are more effective to capture the essence of the problems in which we encounter categorical and numeric data. Based on this observation, it is advantageous to replace the equivalence relations by fuzzy similarity relations.

Given a nonempty and finite set U of objects, and R is a binary relation on U. R is said to be a fuzzy equivalence relation if for $\forall x,y,z \in U$, we have 1) reflexivity: $R(x,x) = 1$, 2) symmetry: $R(x,y) = R(x,y)$, and 3) min-max transitivity: $\min_y(R(x,y), R(y,z)) \le R(x,z)$. More generally, we say R is a fuzzy T-equivalence relation if for $\forall x,y,z \in U$, R satisfies reflexivity, symmetry, and T-transitivity, that is $T(R(x,y), R(y,z)) \le R(x,z)$.

Let R be a fuzzy equivalence relation on U. For $\forall x \in U$, we associate a fuzzy equivalence class $[x]_R$ with x. The membership function of y to $[x]_R$ is defined as $[x]_R(y) = R(x,y), \forall y \in U$. The family of fuzzy equivalence classes

TABLE 1
Selected t-Norms and Their Duals (S Conorms)

	Operators *T*	**Operators *S***
1	$T_M(a,b) = \min(a,b)$	$S_M(a,b) = \max(a,b)$
2	$T_p(a,b) = a \times b$	$S_p(a,b) = a + b - ab$
3	$T_L(a,b) = \max(a+b-1,\ 0)$	$S_L(a,b) = \min(a+b, 1)$
4	$T_{\cos}(a,b) = \max(ab - \sqrt{1-a^2}\sqrt{1-b^2},\ 0)$	$S_{\cos}(a,b) = \min(a+b-ab+\sqrt{2a-a^2}\sqrt{2b-b^2},\ 1)$

TABLE 2
Residual Implication Induced by the t-Norms and Their Duals

	Residual implication θ	**Operator σ**
1	$\theta_M(a,b) = \begin{cases} 1, & a \le b \\ b, & a > b \end{cases}$	$\sigma_M(a,b) = \begin{cases} 0, & a \ge b \\ b, & a < b \end{cases}$
2	$\theta_p(a,b) = \begin{cases} 1, & a = 0 \\ \min(1, b/a), & \text{otherwise} \end{cases}$	$\sigma_p(a,b) = \begin{cases} 1, & a = 0 \\ \max\left(0, \dfrac{b-a}{1-a}\right), & \text{otherwise} \end{cases}$
3	$\theta_L(a,b) = \min(b-a+1, 1)$	$\sigma_L(a,b) = \min(0, b-a)$
4	$\theta_{\cos}(a,b) = \begin{cases} 1, & a \le b \\ ab + \sqrt{1-a^2}\sqrt{1-b^2}, & a > b \end{cases}$	$\sigma_{\cos}(a,b) = \begin{cases} 0, & a > b \\ a+b-ab-\sqrt{2a-a^2}\sqrt{2b-b^2}, & a \le b \end{cases}$

forms a set of fuzzy elemental granules to approximate arbitrary subset of the universe. We call $FAS = <U, R>$ a fuzzy approximation space. Given $FAS = <U, R>$ and a fuzzy subset $X \in U$, the lower approximation and upper approximation of X in $<U, R>$ are defined as [6]

$$\begin{cases} \underline{R_{\max}}X(x) = \inf\limits_{y\in U} \max(1 - R(x,y), X(y)), \\ \overline{R_{\min}}X(x) = \sup\limits_{y\in U} \min(R(x,y), X(y)). \end{cases}$$

The T-equivalence relation is used to define fuzzy rough sets in [28]. Given a fuzzy T-equivalence relation on U and θ is a residual implication induced with T, the fuzzy lower and fuzzy upper approximations of fuzzy subset $X \in U$ are defined as

$$\begin{cases} \underline{R_{\theta}}X(x) = \inf\limits_{y\in U} \theta(R(x,y), X(y)), \\ \overline{R_T}X(x) = \sup\limits_{y\in U} T(R(x,y), X(y)). \end{cases}$$

Furthermore, based on T-equivalence relations, residual implication θ and its dual σ, Mi and Zhang gave another definition of fuzzy rough sets [27]

$$\begin{cases} \underline{R_{\theta}}X(x) = \inf\limits_{y\in U} \theta(R(x,y), X(y)), \\ \overline{R_{\sigma}}X(x) = \sup\limits_{y\in U} \sigma(N(R(x,y)), X(y)). \end{cases}$$

The above definitions of fuzzy rough sets were all constructed with fuzzy equivalence relations or fuzzy T-equivalence relations. They are the straighforward generalizations of the classical rough set model. These models will reduce to the classical one if the underlying relation is an equivalence relation. More generally, Yeung et al., proposed a model of fuzzy rough sets with a general fuzzy relation [45]

$$\begin{cases} \underline{R_S}X(x) = \inf\limits_{y\in U} S(N(R(x,y)), X(y)), \\ \overline{R_T}X(x) = \sup\limits_{y\in U} T(R(x,y), X(y)). \end{cases}$$

As a whole, there are three definitions of fuzzy lower approximation operators: $\underline{R_{\max}}$, $\underline{R_{\theta}}$, and $\underline{R_S}$ and three upper approximation operators: $\overline{R_{\min}}$, $\overline{R_T}$, and $\overline{R_{\sigma}}$. However, $\underline{R_{\max}}$ and $\overline{R_{\min}}$ are the special cases of $\underline{R_S}$ and $\overline{R_T}$, where $S = \max$ and $T = \min$. Therefore, we have two definitions of lower approximations and upper approximations, respectively.

It is easy to show that the model of fuzzy rough sets is the natural extension of Pawlak's rough sets and Neighborhood rough sets [19]. If the relation used in granulating the universe is neighborhood relation and the subset of the objects to approximate is crisp, these fuzzy rough set models degrade to neighborhood rough set model [19]. Furthermore, these fuzzy models degenerate to Pawlak's model if relations are equivalence relations.

3 Kernelized Fuzzy Rough Set Model

No matter what fuzzy rough set model is employed in applications, we have to develop an approach to determining fuzzy relations from data. Most of the definitions and properties of fuzzy rough sets are discussed in the context of fuzzy T-equivalence relations. Fuzzy T-equivalence relations are very useful in fuzzy rough set-based data analysis. In this section, we will introduce kernel functions to compute fuzzy T-equivalence relations and propose kernelized fuzzy rough sets.

3.1 Kernel Fuzzy Rough Sets

Definition 1. [4] *Give a nonempty and finite set U, a real-valued function $k : U \times U \rightarrow R$ is said to be a kernel if it is symmetric, that is, $k(x,y) = k(y,x)$ for all $\forall x, y \in U$, and positive-semidefinite.*

Theorem 1. [30] *Any kernel $k : U \times U \rightarrow [0,1]$ with $k(x,x) = 1$ is (at least) $T_{\cos}$-transitive, where*

$$T_{\cos}(a,b) = \max(ab - \sqrt{1-a^2}\sqrt{1-b^2}, 0).$$

As some of kernel functions are reflexive, $k(x,x) = 1$, symmetric $k(x,y) = k(y,x)$, and $T_{\cos}$-transitive, then the relations computed with these kernel functions are fuzzy T-equivalence relations. Hereafter, in this study T stands for $T_{\cos}$.

Some widely encountered kernel functions satisfying the above properties are [10]:

1. Gaussian kernel: $k_G(x,y) = \exp(-\frac{\|x-y\|^2}{\delta})$;
2. Exponential kernel: $k_E(x,y) = \exp(-\frac{\|x-y\|}{\delta})$;
3. Rational quadratic kernel: $k_R(x,y) = 1 - \frac{\|x-y\|^2}{\|x-y\|^2+\delta}$;
4. Circular kernel:

$$k_C(x,y) = \frac{2}{\pi}\arccos\left(\frac{\|x-y\|}{\delta}\right) - \frac{2}{\pi}\frac{\|x-y\|}{\delta}\sqrt{1-\left(\frac{\|x-y\|}{\delta}\right)^2}$$
$$\text{if } \|x-y\| < \delta;$$

5. Spherical kernel:

$$k_S(x,y) = 1 - \frac{3}{2}\frac{\|x-y\|}{\delta} + \frac{1}{2}\left(\frac{\|x-y\|}{\delta}\right)^3 \text{ if } \|x-y\| < \delta.$$

It is easy to show that the above kernel functions are reflexive $k(x,x) = 1$ and symmetric $k(x,y) = k(y,x)$. Moreover, they are $T_{\cos}$-transitive. Thus, the relations computed with these functions are fuzzy T-equivalence relations. With the kernel function we can substitute fuzzy relations in fuzzy rough sets.

Definition 2. *Given a nonempty universe U and a kernel function k being reflexive, symmetric, and $T_{\cos}$-transitive, for arbitrary fuzzy subset $X \in F(U)$, the fuzzy lower and upper approximation operators are defined as*

1. S-kernel fuzzy lower approximation operator: $\underline{k_S}X(x) = \inf_{y\in U} S(N(k(x,y)), X(y))$;
2. θ-kernel fuzzy lower approximation operator: $\underline{k_{\theta}}X(x) = \inf_{y\in U} \theta(k(x,y), X(y))$;
3. $\overline{T}$-kernel fuzzy upper approximation operator: $\overline{k_T}X(x) = \sup_{y\in U} T(k(x,y), X(y))$; and
4. σ-kernel fuzzy upper approximation operator: $\overline{k_{\sigma}}X(x) = \sup_{y\in U} \sigma(N(k(x,y)), X(y))$.

Theorem 2. *For any $\{A_i : i \in I\} \in F(U)$, we have the following properties:*

1. $\underline{k_S}(\cap_{i\in I} A_i) = \cap_{i\in I}\underline{k_S}A_i$, $\overline{k_T}(\cup_{i\in I} A_i) = \cup_{i\in I}\overline{k_T}A_i$;
2. $\underline{k_\theta}(\cap_{i\in I} A_i) = \cap_{i\in I}\underline{k_\theta}A_i$, $\overline{k_\sigma}(\cup_{i\in I} A_i) = \cup_{i\in I}\overline{k_\sigma}A_i$;

Theorem 3. *Suppose that k is a T-equivalence relation on U computed with kernel function $k(x,y)$. For $\forall X \in F(U)$ the following statements hold:*

1. $\underline{k_S}X \subseteq X$;
2. $\overline{k_T}X \supseteq X$;
3. $\underline{k_\theta}X \subseteq X$;
4. $\overline{k_\sigma}X \supseteq X$.
5. $\overline{k_T}x(y) = \overline{k_T}y(x)$
6. $(\underline{k_S}(U - \{y\}))(x) = (\underline{k_S}(U - \{x\}))(y)$
7.

$$\underline{k_S}(\underline{k_S}X) = \underline{k_S}X,\ \underline{k_\theta}(\underline{k_\theta}X) = \underline{k_\theta}X,\ \overline{k_T}(\overline{k_T}X) = \overline{k_T}X, \overline{k_\sigma}(\overline{k_\sigma}X) = \overline{k_\sigma}X.$$

The first four properties can be obtained from the reflexivity of kernel, properties 5 and 6 can be derived from the symmetry of kernel, and property 7 can be obtained from the T-transitivity of kernels.

Theorem 4. *Suppose that k is T-equivalence relation on U computed with kernel function $k(x,y)$, $\underline{k_S}$, $\overline{k_T}$, $\underline{k_\theta}$, and $\overline{k_\sigma}$ have the following properties:*

1. All of $\underline{k_S}$, $\overline{k_T}$, $\underline{k_\theta}$, and $\overline{k_\sigma}$ are monotone;
2.

$$\overline{k_T}(\underline{k_\theta}X) = \underline{k_\theta}X, \underline{k_\theta}(\overline{k_T}X) = \overline{k_T}X, \overline{k_\sigma}(\underline{k_S}X) = \underline{k_S}X, \underline{k_S}(\overline{k_\sigma}X) = \overline{k_\sigma}X;$$

3. $\overline{k_T}X = X \Leftrightarrow \underline{k_\theta}X = X, \overline{k_\sigma}X = X \Leftrightarrow \underline{k_S}X = X$.

3.2 Approximating Classification with Kernel

Classification is one of the most important problems in machine learning and pattern recognition. In this problem, the given learning samples are assigned to several decision labels (categories). Now we consider the fuzzy lower approximation of classification with kernel functions.

Typically, the classification can be formulated as $<U, A, D>$, where U is the nonempty and finite set of samples, A is the set of features characterizing the classification, D is the class attribute which divides the samples into subset $\{d_1, d_2, \ldots, d_K\}$. For $\forall x \in U$,

$$d_i(x) = \begin{cases} 0, x \notin d_i \\ 1, x \in d_i \end{cases}.$$

Assume that kernel function k is used to compute the fuzzy similarity relation between samples. Then, we approximate the decision subsets with the fuzzy granules induced by a kernel. Take the ith class as an example,

1.
$$\begin{aligned}\underline{k_S}d_i(x) &= \inf_{y\in U} S(N(k(x,y)), d_i(y)) \\ &= \inf_{y\in U} S((1-(k(x,y)), d_i(y)) \\ &= \inf_{y\in U} \min\Bigg(1, 1 - k(x,y) + k(x,y)d_i(y) \\ &\quad + \sqrt{1-k^2(x,y)}\sqrt{2d_i(y) - d_i^2(y)}\Bigg).\end{aligned}$$

If $d_i(y) = 1$, i.e.,

$$\begin{aligned}y \in d_i, \min(1, 1 - k(x,y) + k(x,y)d_i(y) \\ + \sqrt{1-k^2(x,y)}\sqrt{2d_i(y) - d_i^2(y)}) = 1.\end{aligned}$$

If $d_i(y) = 0$, i.e., $y \notin d_i$, we obtain $\underline{k_S}d_i(x) = \inf_{y\notin d_i}(1 - k(x,y))$.

2.
$$\begin{aligned}\underline{k_\theta}d_i(x) &= \inf_{y\in U} \theta(k(x,y), d_i(y)) \\ &= \inf_{y\in U}\left(\begin{cases} 1, & \text{k(x, y)} \le d_i(y), \\ k(x,y)d_i(y) \\ \quad + \sqrt{1-k^2(x,y)} \\ \quad \sqrt{1-d_i^2(y)}, & \text{k(x, y)} > d_i(y). \end{cases}\right)\end{aligned}$$

If $d_i(y) = 1$, i.e., $y \in d_i$, in this case $k(\text{x, y}) \le d_i(y)$, then we get $\underline{k_\theta}d_i(x) = \theta(k(x,y), d_i(y)) = 1$;

If $d_i(y) = 0$, i.e., $y \notin d_i$, in this case $k(\text{x, y}) > d_i(y)$, $\underline{k_\theta}d_i(x) = \theta(k(x,y), d_i(y)) = \sqrt{1-k^2(x,y)}$.

Finally, we arrive at $\underline{k_\theta}d_i(x) = \inf_{y\notin d_i}(\sqrt{1-k^2(x,y)})$.

3.
$$\begin{aligned}\overline{k_T}d_i(x) &= \sup_{y\in U} T(k(x,y), d_i(y)) \\ &= \sup_{y\in U} \max(0, k(x,y)d_i(y) \\ &\quad - \sqrt{1-k^2(x,y)}\sqrt{1-d_i^2(y)}).\end{aligned}$$

If $d_i(y) = 1$, i.e., $y \in d_i$, in this case, we get $\max(0, k(x,y)d_i(y) - \sqrt{1-k^2(x,y)}\sqrt{1-d_i^2(y)}) = k(x,y)$.

If $d_i(y) = 0$, i.e., $y \notin d_i$, in this case

$$\begin{aligned}&\max(0, k(x,y)d_i(y) - \sqrt{1-k^2(x,y)}\sqrt{1-d_i^2(y)}) \\ &\quad = \max(0, -\sqrt{1-k^2(x,y)}) = 0.\end{aligned}$$

We get $\overline{k_T}d_i(x) = \sup_{y\in d_i} k(x,y)$.

4.
$$\overline{k_\sigma}d_i(x) = \sup_{y\in U} \sigma(N(k(x,y)), d_i(y))$$

$$= \sup_{y\in U}\left(\begin{cases} 0, & \text{k(x, y)} \\ & > d_i(y), \\ (1-k(x,y)) + d_i(y) \\ \quad -(1-k(x,y))d_i(y) \\ \quad -\sqrt{2(1-k(x,y)) - (1-k^2(x,y))} \\ \quad \sqrt{2d_i(y) - d_i^2(y)}, & \text{k(x, y)} \\ & \le d_i(y). \end{cases}\right)$$

If $d_i(y)=1$, i.e., $y\in d_i$, in this case,

$$\sup_{y\in U}\left(\begin{cases} 0, & k(\mathrm{x,y}) > d_i(y), \\ (1-k(x,y))+d_i(y) \\ \quad -(1-k(x,y))d_i(y) \\ \quad -\sqrt{2(1-k(x,y))-(1-k(x,y))^2} \\ \quad \sqrt{2d_i(y)-d_i^2(y)}, & k(\mathrm{x,y}) \le d_i(y). \end{cases}\right)$$

$$\begin{aligned} &= (1-k(x,y))+d_i(y)-(1-k(x,y))d_i(y) \\ &\quad -\sqrt{2(1-k(x,y))-(1-k(x,y))^2}\sqrt{2d_i(y)-d_i^2(y)} \\ &= 1-\sqrt{2(1-k(x,y))-(1-k(x,y))^2}. \end{aligned}$$

If $d_i(y)=0$, i.e., $y\notin d_i$, in this case one has

$$\sup_{y\in U}\left(\begin{cases} 0, & k(\mathrm{x,y}) > d_i(y), \\ (1-k(x,y))+d_i(y) \\ \quad -(1-k(x,y))d_i(y) \\ \quad -\sqrt{2(1-k(x,y))-(1-k(x,y))^2} \\ \quad \sqrt{2d_i(y)-d_i^2(y)}, & k(\mathrm{x,y}) \le d_i(y). \end{cases}\right)$$
$$= 0.$$

Overall we have

$$\begin{aligned} \overline{k_\sigma}d_i(x) &= \sup_{y\in d_i}\left(1-\sqrt{2(1-k(x,y))-(1-k(x,y))^2}\right) \\ &= \sup_{y\in d_i}\left(1-\sqrt{1-k^2(x,y)}\right). \end{aligned}$$

Now, we construct the algorithms for computing the fuzzy lower and upper approximations for a given kernel function.

1. $\underline{k_S}d_i(x)=\inf_{y\notin d_i}(1-k(x,y))$;
2. $\underline{k_\theta}d_i(x)=\inf_{y\notin d_i}(\sqrt{1-k^2(x,y)})$;
3. $\overline{k_T}d_i(x)=\sup_{y\in d_i}k(x,y)$; and
4. $\overline{k_\sigma}d_i(x)=\sup_{y\in d_i}(1-\sqrt{1-k^2(x,y)})$.

We can see that we have the following properties:

Theorem 5. *Give a decision system* $<U,A,D>$, *where* d_i *is one of the decision classes,* $\forall x\in U, \underline{k_S}d_i(x)\le\underline{k_\theta}d_i(x)$, $\overline{k_T}d_i(x)\ge\overline{k_\sigma}d_i(x)$.

Proof. It is obvious that $\underline{k_S}d_i(x)\ge 0$ and $\underline{k_\theta}d_i(x)\ge 0$ □

$$\begin{aligned} \underline{k_S}d_i(x) &= \inf_{y\notin d_i}(1-k(x,y)), \\ \Rightarrow (\underline{k_S}d_i(x))^2 &= \inf_{y\notin d_i}(1-k^2(x,y)-2k(x,y)+2k^2(x,y)) \\ &= \inf_{y\notin d_i}(1-k^2(x,y)-2k(x,y)(1-k(x,y))) \\ &\le \inf_{y\notin d_i}(1-k^2(x,y)) = (\underline{k_\theta}d_i(x))^2. \end{aligned}$$

So $\underline{k_S}d_i(x)\le\underline{k_\theta}d_i(x)$.

Similarly, we can get $\overline{k_T}d_i(x)\ge\overline{k_\sigma}d_i(x)$.

As an example, let us consider the Gaussian kernel [55] to explain the essence of kernel-based fuzzy approximations.

1.
$$\begin{aligned} \underline{k_S}d_i(x) &= \inf_{y\notin d_i}(1-k(x,y)) \\ &= \inf_{y\notin d_i}\left(1-\exp\left(-\frac{\|x-y\|^2}{\delta}\right)\right). \end{aligned}$$
   If $x\in d_i$, we have to find a nearest neighbor of x from other classes to compute the lower approximation. Taking $1-\exp(-\frac{\|x-y\|^2}{\delta})$ as a generalized distance function, the membership of x to its class depends on the nearest sample in a distinct class. However, if $x\notin d_i$, the nearest sample of x out of d_i is x itself. In this case $\underline{k_S}d_i(x)=0$ because $k(x,x)=1$;
2. Analogically, if
$$x\in d_i, \underline{k_\theta}d_i(x)=\inf_{y\notin d_i}\left(\sqrt{1-\exp^2\left(-\frac{\|x-y\|^2}{\delta}\right)}\right);$$
   otherwise, $x\notin d_i$, $\underline{k_\theta}d_i(x)=0$;
3. If $x\in d_i$, $\overline{k_T}d_i(x)=\sup_{y\in d_i}k(x,y)$. Obviously, $\sup_{y\in d_i}k(x,y)=1$ because $k(x,x)=1$.
4. If $x\notin d_i$, we need to find a sample $y\in d_i$ such that $\overline{k_T}d_i(x)=\max_{y\in d_i}\exp(-\frac{\|x-y\|^2}{\delta})$. This means that y is the nearest sample from x in d_i.
5. If
$$\begin{aligned} & x\in d_i, \overline{k_\sigma}d_i(x) \\ &\quad = \sup_{y\in d_i}(1-\sqrt{2(1-k(x,y))-(1-k(x,y))^2}).k(x,y) \\ &\quad = 1 \end{aligned}$$
   if $x=y$. Here, $\overline{k_\sigma}d_i(x)=1$.
6. If $x\notin d_i$, $\overline{k_\sigma}d_i(x)=\sup_{y\in d_i}(1-\sqrt{1-k^2(x,y)})$. Let $k(x,y)=\exp(-\frac{\|x-y\|^2}{\delta})$, we get
$$\overline{k_\sigma}d_i(x)=\sup_{y\in d_i}\left(1-\sqrt{1-\left(\exp\left(-\frac{\|x-y\|^2}{\delta}\right)\right)^2}\right).$$
   Clearly, the fuzzy upper approximation depends on the nearest sample of x from d_i.

The above analysis shows that the membership of x to the lower approximation of x's decision is determined by the closest sample with different decision, while the membership of x to the lower approximation of other decision is zero. Correspondingly, the membership of x to the upper approximation of x's decision is always 1, while the membership of x to the upper approximation of another decision depends on the closest sample from this class. Furthermore, the distinct definition leads to different computing of the lower and upper approximations.

Example 1. Given is a classification task in which we encounter 12 patterns; those are shown in Table 3. Each sample is characterized by means of a certain condition attribute A. D is a two-valued decision variable assuming values d_1 and d_2.

We consider the Gaussian kernel function to compute the similarity relation between samples. The kernel parameter is equal to 0.2, $\delta = 0.2$:

$$\begin{aligned}
\underline{k_S}d_1(x_1) &= \inf_{y \notin d_1}(1 - k(x_1, y)) = \inf_{y \in d_2}(1 - k(x_1, y)) \\
&= \inf_{y \in d_2}\left(1 - \exp\left(-\frac{\|x_1 - y\|^2}{0.2}\right)\right) \\
&= 1 - \exp\left(-\frac{\|x_1 - x_7\|^2}{0.2}\right) = 0.7276, \\
\underline{k_S}d_2(x_1) &= \inf_{y \notin d_2}(1 - k(x_1, y)) = \inf_{y \in d_1}(1 - k(x_1, y)) \\
&= \inf_{y \in d_1}\left(1 - \exp\left(-\frac{\|x_1 - y\|^2}{0.2}\right)\right) \\
&= 1 - \exp\left(-\frac{\|x_1 - x_1\|^2}{0.2}\right) = 0,
\end{aligned}$$

$$\begin{aligned}
\underline{k_\theta}d_1(x_1) &= \inf_{y \notin d_1}\left(\sqrt{1 - k(x_1, y)^2}\right) = \inf_{y \in d_2}\left(\sqrt{1 - k(x_1, y)^2}\right) \\
&= \sqrt{1 - \left(\exp\left(-\frac{\|x_1 - x_7\|^2}{0.2}\right)\right)^2} = 0.9622, \\
\underline{k_\theta}d_2(x_1) &= \inf_{y \notin d_2}\left(\sqrt{1 - k^2(x_1, y)}\right) = \inf_{y \in d_1}\left(\sqrt{1 - k^2(x_1, y)}\right) \\
&= \sqrt{1 - \left(\exp\left(-\frac{\|x_1 - x_1\|^2}{0.2}\right)\right)^2} = 0,
\end{aligned}$$

$$\begin{aligned}
\overline{k_T}d_1(x_1) &= \sup_{y \in d_1} k(x, y) = \sup_{y \in d_1} \exp\left(-\frac{\|x_1 - y\|^2}{0.2}\right) \\
&= \exp\left(-\frac{\|x_1 - x_1\|^2}{0.2}\right) = 1, \\
\overline{k_T}d_2(x_1) &= \sup_{y \in d_2} k(x, y) = \sup_{y \in d_2} \exp\left(-\frac{\|x_1 - y\|^2}{0.2}\right) \\
&= \exp\left(-\frac{\|x_1 - x_7\|^2}{0.2}\right) = 0.2724,
\end{aligned}$$

$$\begin{aligned}
\overline{k_\sigma}d_1(x_1) &= \sup_{y \in d_1}\left(1 - \sqrt{1 - k^2(x_1, y)}\right) \\
&= \sup_{y \in d_1}\left(1 - \sqrt{1 - \left(\exp\left(-\frac{\|x_1 - y\|^2}{0.2}\right)\right)^2}\right) \\
&= 1 - \sqrt{1 - \left(\exp\left(-\frac{\|x_1 - x_1\|^2}{0.2}\right)\right)^2} = 1, \\
\overline{k_\sigma}d_2(x_1) &= \sup_{y \in d_2}\left(1 - \sqrt{1 - k^2(x_1, y)}\right) \\
&= \sup_{y \in d_2}\left(1 - \sqrt{1 - \left(\exp\left(-\frac{\|x_1 - y\|^2}{0.2}\right)\right)^2}\right) \\
&= 1 - \sqrt{1 - \left(\exp\left(-\frac{\|x_1 - x_7\|^2}{0.2}\right)\right)^2} = 0.0378.
\end{aligned}$$

$\underline{k_S}d_i(x)$ and $\underline{k_\theta}d_i(x)$ are the degrees the sample x certainly belongs to class d_i, while $\overline{k_T}d_i(x)$ and $\overline{k_\sigma}d_i(x)$ are the degrees this sample x possibly belongs to class d_i. In this example, x_1 certainly belongs to class 1 with degrees 0.7276 and 0.9622 computed with $\underline{k_S}d_i(x)$ and $\underline{k_\theta}d_i(x)$, respectively. At the same time, x_1 probably belongs to class 1 with degree 1 and probably belongs to class 2 with degree 0.2724 computed with the use of $\overline{k_T}d_i(x)$. Moreover, x_1 probably belongs to class 1 with degree 1 and probably belongs to class 2 with degree 0.0378 computed by means $\overline{k_\sigma}d_i(x)$. We have $\underline{k_S}d_1(x_1) + \overline{k_T}d_2(x_1) = 1$, $\underline{k_\theta}d_1(x_1) + \overline{k_\sigma}d_2(x_1) = 1$.

The membership values of the fuzzy lower approximations and upper approximations of classes 1 and 2 are shown in Fig. 2, where LM1 and LM2 are the membership values of the samples to fuzzy lower approximations of class 1 and class 2, respectively, while where UM1 and UM2 are the membership values of the samples to fuzzy upper approximations of class 1 and class 2, respectively.

4 Characterization of Approximation Quality

Some quantitative coefficients are defined to characterize the quality of approximation in classical rough set model. We here generalize these coefficients to handle the cases of fuzzy approximations and discuss some new measures.

Obviously, $\underline{k_S}d_i$, $\underline{k_\theta}d_i$, $\overline{k_T}d_i$, and $\overline{k_\sigma}d_i$ are all fuzzy sets, and we have $\underline{k_S}d_i \subseteq d_i \subseteq \overline{k_T}d_i$, and $\underline{k_\theta}d_i \subseteq d_i \subseteq \overline{k_\sigma}d_i$.

Definition 3. *Let X be a fuzzy subset, the cardinality of X is defined as $|X| = \sum_{x \in U} X(x)$, where $X(x)$ is the membership of x to X.*

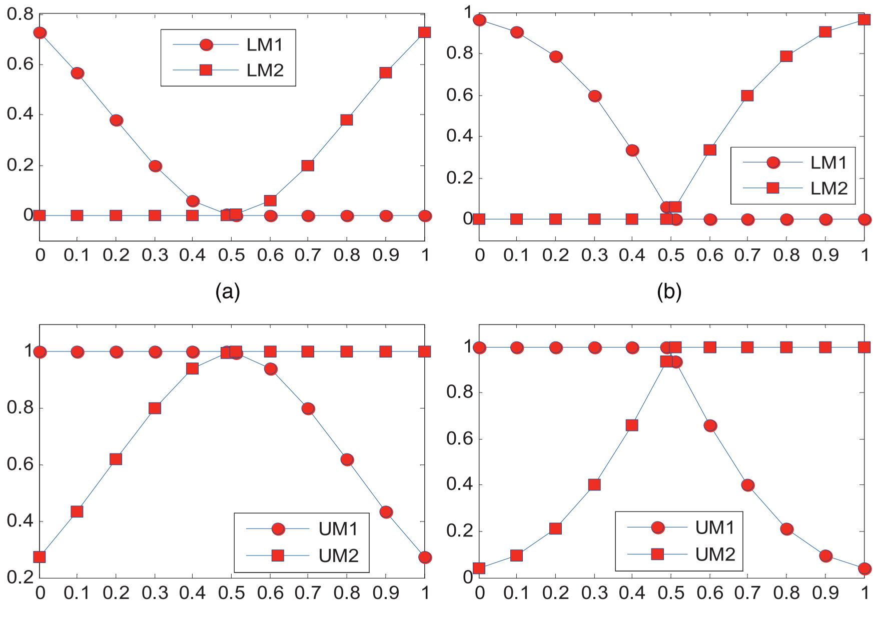
Definition 4. *Let X be a fuzzy subset, the approximation accuracy of set X in the approximation space is defined as*

$$\mu_A^{S-T}(X) = \frac{|\underline{k_S}X|}{|\overline{k_T}X|}, \; \mu_A^{\theta-\sigma}(X) = \frac{|\underline{k_\theta}X|}{|\overline{k_\sigma}X|}.$$

Definition 5. *Let X be a fuzzy set, the approximation quality of set X in the approximation space is defined as*

TABLE 3
Example Data

Sample	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
A	0	0.1	0.2	0.3	0.4	0.49	0.51	0.6	0.7	0.8	0.9	1
D	d_1	d_1	d_1	d_1	d_1	d_1	d_2	d_2	d_2	d_2	d_2	d_2



(a) (b) (c) (d)

Fig. 2. Membership values of fuzzy lower and upper approximations computed with Gaussian kernel (a) Membership function computed with $\underline{k_S}d_i(x)$. (b) Membership function computed with $\underline{k_\theta}d_i(x)$. (c) Membership function computed with $\overline{k_T}d_i(x)$. (d) Membership function computed with $\overline{k_\sigma}d_i(x)$.

$$\eta_A^{S-T}(X) = \frac{|\underline{k_S}X|}{|X|}, \quad \eta_A^{\theta-\sigma}(X) = \frac{|\underline{k_\theta}X|}{|X|}.$$

Definition 6. *Given a classification problem* $<U, A, D>$, *k is T-equivalence relation on Ucomputed with kernel function* $k(x, y)$ *in the feature space* $B \subseteq A$. *U is divided into* $\{d_1, d_2, \dots, d_I\}$ *with the decision attribute. The fuzzy positive regions of D in term of B are defined as*

$$POS_B^S(D) = \bigcup_{i=1}^{I} \underline{k_S}d_i, \quad POS_B^\theta(D) = \bigcup_{i=1}^{I} \underline{k_\theta}d_i,$$

where I is the number of classes.

Definition 7. *Given a classification problem* $<U, A, D>$, *k is T-equivalence relation on U computed with kernel function* $k(x, y)$ *in the feature space* $B \subseteq A$. *U is divided into* $\{d_1, d_2, \dots, d_I\}$ *with the decision attribute. The quality of classification approximation is defined as*

$$\gamma_B^S(D) = \frac{|\cup_{i=1}^{I} \underline{k_S}d_i|}{|U|} \text{ or } \gamma_B^\theta(D) = \frac{|\cup_{i=1}^{I} \underline{k_\theta}d_i|}{|U|}.$$

The coefficients of classification quality reflect the approximation ability of the approximation space or the ability of the granulated space induced by attribute subset B to characterize the decision. These coefficients are also called the dependency between the decision and condition attributes. We say that decision D is dependent on B with degree $\gamma_B^S(D)$ or $\gamma_B^\theta(D)$, denoting by $B \Rightarrow_\gamma D$.

Theorem 6. *Given a classification problem* $<U, A, D>$, $B_1 \subseteq B_2 \subseteq A$, k_1 *and* k_2 *are two T-equivalence relations on U computed with kernel function* $k(x, y)$ *in the feature space* B_1 *and* B_2, *respectively. Then, we have*

1. $k_1 \supseteq k_2$;
2. $\underline{k_{1S}}d_i \subseteq \underline{k_{2S}}d_i, \underline{k_{1\theta}}d_i \subseteq \underline{k_{2\theta}}d_i$;
3. $\overline{k_{1T}}d_i \supseteq \overline{k_{2T}}d_i, \overline{k_{1\sigma}}d_i \supseteq \overline{k_{2\sigma}}d_i$;
4. $POS_{B_1}^S(D) \subseteq POS_{B_2}^S(D), POS_{B_1}^\theta(D) \subseteq POS_{B_2}^\theta(D)$;
5. $\mu_{B_1}^{S-T}(d_i) \le \mu_{B_2}^{S-T}(d_i), \mu_{B_1}^{\theta-\sigma}(d_i) \le \mu_{B_2}^{\theta-\sigma}(d_i)$;
6. $\eta_{B_1}^{S-T}(d_i) \le \eta_{B_2}^{S-T}(d_i), \eta_{B_1}^{\theta-\sigma}(d_i) \le \eta_{B_2}^{\theta-\sigma}(d_i)$;
7. $\gamma_{B_1}^S(D) \le \gamma_{B_2}^S(D)$ and $\gamma_{B_1}^\theta(D) \le \gamma_{B_2}^\theta(D)$.

Proof. Properties 2-6 can be derived from the monotonicity of the lower and upper approximations [45]. Now, we just require showing the proof of property 1. Take Gaussian kernel as an example: $k_G(x, y) = \exp(-\frac{\|x-y\|^2}{\delta})$. □

Assume that $|B_1| = N_1$, $|B_2| = N_2$. As $B_1 \subseteq B_2$, we have $N_1 \le N_2$. Without loss of generality, we take two arbitrary samples to compute the fuzzy relations with Gaussian kernel function.

In the feature space B_1, $\|x - y\|_{B_1}^2 = \sum_{i=1}^{N_1} (f(x, a_i) - f(y, a_i))^2$, where $f(x, a_i)$ is the value of sample x in feature a_i. In the feature space B_2, $\|x - y\|_{B_2}^2 = \sum_{i=1}^{N_1} (f(x, a_i) - f(y, a_i))^2 + \sum_{i=N_1}^{N_21} (f(x, a_i) - f(y, a_i))^2$. Therefore, $\|x - y\|_{B_2}^2 \ge \|x - y\|_{B_1}^2$ and $k_{B_1}(x, y) \ge k_{B_2}(x, y)$. Then $k_1 \supseteq k_2$.

Theorem 6 shows that as the features increase, the approximation quality, the approximation accuracy, and the classification quality monotonously increase. These properties are consistent with our intuition. New features maybe bring new information about granulation and classification. Correspondingly, the induced approximation space with more features becomes finer and more precise approximations are generated in the finer approximation space. As a result, the approximation quality, accuracy, and classification quality become improved.

The coefficients introduced above reflect the quality of approximation of a set in the granulated or approximation space induced by a kernel function. Moreover, we can also show some measures to characterize the contribution of the samples to classification learning. As we pointed before, $\underline{k_S}d_i(x)$ and $\underline{k_\theta}d_i(x)$ are the degrees the sample x certainly belongs to class d_i, while $\overline{k_T}d_i(x)$ and $\overline{k_\sigma}d_i(x)$ are the degrees quantifying that sample x possibly belongs to class d_i. Generally speaking, one hopes that the training samples certainly belong to one of the decision class. In this case, $\underline{k_S}d_i(x)$ and $\underline{k_\theta}d_i(x)$ are great while the values of $\overline{k_T}d_j(x)$ and $\overline{k_\sigma}d_j(x)$ are small. Based on this observation we can define a measure of classification certainty of samples as follows:

Definition 8. *Given $<U, A, D>$, k is T-equivalence relation on U computed with kernel function $k(x, y)$ in the feature space $B \subseteq A$. U is divided into $\{d_1, d_2, \dots, d_I\}$ with the decision attribute. If $x \in d_l$, we define the classification certainty of x in feature space B as*

$$\omega_B^{S-T}(x) = \underline{k_S}d_l(x) - \sum_{d_i \in D - d_l} \overline{k_T}d_i(x) \text{ or } \omega_B^{\theta-\sigma}(x) = \underline{k_\theta}d_l(x) - \sum_{d_i \in D - d_l} \overline{k_\sigma}d_i(x).$$

The sum of classification certainty of samples reveals the overall classification certainty in the corresponding feature space.

Definition 9. *Given $<U, A, D>$, k is T-equivalence relation on U computed with kernel function $k(x, y)$ in the feature space $B \subseteq A$. The classification certainty of B is defined as*

$$\omega_B^{S-T}(D) = \frac{1}{|U|}\sum_{x \in U} \omega_B^{S-T}(x) \text{ or } \omega_B^{\theta-\sigma}(D) = \frac{1}{|U|}\sum_{x \in U} \omega_B^{\theta-\sigma}(x).$$

In this section, we introduce some coefficients to characterize the approximation quality of a set in a granulated space and approximation quality of a classification. Moreover, we also discuss the classification certainty of samples and classification certainty of the granulated space. In the next section, we will discuss the relationship kernel fuzzy rough set-based measures and the famous Relief algorithm.

5 Relationship between Kernelized Fuzzy Rough Sets and Relief Series

Relief and its variants ReliefF, RReliefF, and I-relief are a family of well-known feature evaluation and attribute weighting techniques for classification and regression [32], [34], [38]. In this section, we will show that kernel fuzzy rough set-based measures share the similar idea with Relief and introduce the ideas in Relief and its variants to extend the measures of dependency and classification certainty.

The key idea in the original Relief, which was proposed by Kira and Rendell in 1992 [22], is to estimate the significance of attributes according to how well their values distinguish between instances that are near to each other. As to arbitrary sample x with decision d_i, Relief searches for its two nearest neighbors: one from d_i, called nearest hit H, and the other from the different classes, called nearest miss M. Then the distance difference is used to estimate the classification certainty of sample:$\|M - x\| - \|H - x\|$, where $\|M - x\|$ is a general distance function. As a whole, $\sum_{x \in U} \|M - x\| - \|H - x\|$ is used to evaluate the quality of features. If there are too many samples available, a subset of samples can randomly be extracted from the original data and the quality of features are estimated with the generated subset of samples.

Algorithm Relief
Input: $<U, A, D>$
Output: $W = <w_1, w_2, \dots, w_N>$ of estimations of the qualities of attributes

```
set all weights $W \leftarrow 0$;
for $i = 1$ to $m$ do begin
    randomly select an instance $x_i$;
    find nearest hit $H_i$ and nearest miss $M_i$;
    for $A = 1$ to $N$ do
        $W \leftarrow W + \|M_i - x_i\|/m - \|H_i - x_i\|/m$;
    end
end
```

Relief algorithm searches two nearest samples of a sample within its same class and out of its class to estimate the classification certainty of the sample. Intuitively, we hope the sample is far from the samples with different classes and is close to the sample of the same decision. In fact, the fuzzy lower and upper approximations share the similar idea in kernelized fuzzy rough sets.

Take Gaussian kernel as an example. Given $<U, A, D>$, assume the samples are divided into two classes $\{d_1, d_2\}$, $x \in d_1$, and the nearest sample from class d_2 is M, then the membership of x to the fuzzy lower approximation is computed as

$$\underline{k_S}d_1(x) = \inf_{y \notin d_1}(1 - k(x, y)) = 1 - k(x, M).$$

Let ϕ be the nonlinear map from the input space A to the feature selection F, and $k(x, y) = <\phi(x), \phi(y)>$. Then

$$\begin{aligned}\|\phi(x) - \phi(M)\|^2 &= \phi(x)\phi(x) + \phi(M)\phi(M) - 2\phi(x)\phi(M) \\ &= 2 - 2\phi(x)\phi(M) = 2 - 2k(x, M),\end{aligned}$$

where $\| \bullet \|$ is the 2-norm of vectors. We conclude that $\underline{k_S}d_1(x) = \|\phi(x) - \phi(M)\|^2/2$. In other words, the membership of x to the fuzzy lower approximation of its class is the distance between this sample and its nearest sample with different classes in the kernel space.

Similarly, we can also consider $\sqrt{1 - \exp^2(-\frac{\|x-M\|^2}{\delta})}$ as a distance in Gaussian function induced kernel space. With different kernel functions and fuzzy operators, we can obtain various distance measures in the corresponding kernel spaces.

As the analysis in Section 3.2, the membership of x to the upper approximation of its class $\overline{k_T}d_i(x) = 1$, we make a slight modification of the definition of the upper approximation as

$$\overline{k'_T}d_i(x) = \sup_{\substack{y\in d_i\\ y\neq x}} k(x,y).$$

Let H be the nearest sample of x from the same class and $H \neq x$, then $\overline{k'_T}d_i(x) = k(x,H) = 1 - \|\phi(x) - \phi(H)\|^2/2$.

We compute the Relief coefficient in the kernel space:

$$\begin{aligned}&\|\phi(M) - \phi(x)\|^2 - \|\phi(H) - \phi(x)\|^2\\ &\quad = 2\underline{k_S}d_1(x) - (2 - 2\overline{k'_T}d_1(x)) = 2(\underline{k_S}d_1(x) + \overline{k'_T}d_1(x) - 1).\end{aligned}$$

This shows that Relief is the trade-off of the lower approximation and the upper approximation in the kernel space. As we pointed out before, $\underline{k_S}d_1(x)$ is the degree of x certainly belonging to d_1, while $\overline{k'_T}d_1(x)$ is the degree of x possibly belonging to d_1, $\|\phi(M) - \phi(x)\|^2 - \|\phi(H) - \phi(x)\|^2$ is the trade-off of these two degrees. In this context, kernelized fuzzy rough sets produce a solution that extends Relief to the kernelized Relief.

ReliefF [24], a variant of Relief, is proposed to deal with multiple-class problems. ReliefF is more robust to noisy samples and can deal with the data with missed values. Similarly to Relief, ReliefF randomly selects an instance x_i , but then searches for k of its nearest neighbors from the same class, called nearest hits H_i and also k nearest neighbors from each of the different classes, called nearest misses M_j, The update formula is similar to that of Relief, except that we average the contribution of all the hits and all the misses. The contribution for each class of the misses is weighted with the prior probability of that class, which is estimated from the training set.

Rough set-based algorithms were reported to be sensitive to noise in feature evaluation, attribute reduction, and rule induction [54]. Here, we can also introduce the ideas used in ReliefF to increase the robustness of fuzzy rough set-based coefficients.

As the dependency is usually used to evaluate the quality of features in attribute reduction, we extend the dependency as follows.

Definition 10. *Given a classification problem $<U, A, D>$, k is T-equivalence relation on U computed with kernel function $k(x,y)$ in the feature space $B \subseteq A$. U is divided into $\{d_1, d_2, \ldots, d_I\}$ with the decision attribute. For $\forall x_i \in U$, we search for the nearest k samples of x_i from each class except d_i (d_i is the decision of x_i) , denoted by $H^i_{d_i} = \{H^i_1, H^i_2, \ldots, H^i_k\}$ The generalized measures of the quality of classification approximation are defined as*

$$g\gamma^S_B(D) = \frac{1}{(I-1)k|U|}\sum_{x_i\in U}\sum_{d\in D, d\neq d_i}\sum_{y\in H^i_d}[1 - k(x,y)] \text{ and}$$

$$g\gamma^\theta_B(D) = \frac{1}{(I-1)k|U|}\sum_{x_i\in U}\sum_{d\in D, d\neq d_i}\sum_{y\in H^i_d}\sqrt{1 - k(x,y)^2}.$$

Along this direction, we generalize the measures of classification certainty.

Definition 11. *Given $<U, A, D>$, k is T-equivalence relation on U computed with kernel function $k(x,y)$ in the feature space $B \subseteq A$. U is divided into $D = \{d_1, d_2, \ldots, d_I\}$ with the decision attribute. For $\forall x_i \in U$, we search for the nearest k samples of x_i from the same class denoted by $H^i = \{H^i_1, H^i_2, \ldots, H^i_k\}$ and the nearest k samples of x_i from each different classes, denoted by*

$$M^i = \begin{vmatrix} M^i_{11} & M^i_{12} & \cdots & M^i_{1k}\\ M^i_{21} & M^i_{22} & \cdots & M^i_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ M^i_{(I-1)1} & M^i_{(I-1)2} & \cdots & M^i_{(I-1)k}\end{vmatrix},$$

where M^i_{lm} is the m*th* nearest sample of x_i from class d_l. The generalized measures of the classification certainty are defined as

$$\begin{aligned} g\omega^{S-T}_B(D) = {}& \frac{1}{(I-1)k|U|}\sum_{x_i}\left\{\sum_{H^i}[1 - k(x_i, H^i_j)]\right.\\ &\left. - \sum_l\sum_m k(x_i, M^i_{lm})\right\} \text{ or}\\ g\omega^{\theta-\sigma}_B(D) = {}& \frac{1}{(I-1)k|U|}\sum_{x_i}\left\{\sum_{H^i}\sqrt{1 - k(x_i, H^i_j)^2}\right.\\ &\left. - \sum_l\sum_m\left[1 - \sqrt{1 - k(x_i, M^i_{lm})^2}\right]\right\}. \end{aligned}$$

6 Feature Selection with Kernelized Fuzzy Rough Sets

The above analysis shows that the model of kernel-based fuzzy rough sets can be used to compute the membership degrees of a sample to the lower approximation and upper approximation of its decision and calculate the memberships of a sample to the upper approximations of other classes. Moreover, we define the approximation quality and approximation accuracy of a set or classification. The measures of dependency and classification certainty are also introduced. This section presents some potential applications of these coefficients.

6.1 Feature Evaluation

Feature evaluation and attribute reduction are among the most important applications of fuzzy rough set model. In this section, we discuss feature evaluation with kernelized fuzzy rough sets.

In Sections 3 and 4, we generalize the dependency function to the fuzzy case and give two kernelized fuzzy dependency functions $0 \le \gamma^S_B(D) \le 1$ and $0 \le \gamma^\theta_B(D) \le 1$. These functions reflect the overall degree at which the samples certainly belong to one of the decision classes; therefore dependency characterizes the power of features to predict the decision. Boolean dependency function, which is the percentage of positive region over the set of samples, was widely used in heuristic attribute reduction [19], [21], where a sample either belongs to the positive region, or belongs to classification boundary. While fuzzy

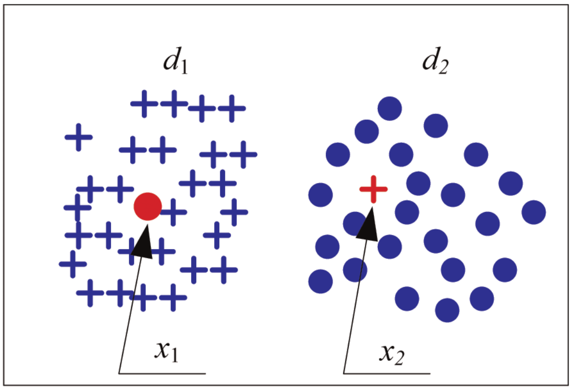


Fig. 3. Learning samples with label noise.

dependency function reflects the overall membership of samples belonging to the positive region. Generally speaking, one select the features which produce a high value of dependency. Moreover, there is another advantage that we take dependency to evaluate the features. That is, the function of dependency is monotonous, $\gamma_B^S(D) \le \gamma_C^S(D)$ and $\gamma_B^\theta(D) \le \gamma_C^\theta(D)$ if $B \subseteq C$. Monotonicity is important for some search strategies and the implementation of the stopping criterions [5], [37].

However, there are still two problems in using fuzzy dependency functions to evaluate the quality of features. First, dependency just considers the membership of a sample to the lower approximation of its class, and neglects the membership of the sample to the upper approximation of other classes. Overall one hopes to find a feature subspace where samples not only certainly belong to their classes, but also their membership degrees to the upper approximations of other classes are as low as possible. Therefore, a sound evaluating function should take the memberships of upper approximations into account. Second, dependency function is not robust enough for dealing with noisy information. As we know, dependency is computed from the distance between a sample and its nearest neighbor with different class. Given $x \in d_i$, it is mislabeled with d_j. Then, the value of the memberships of sample $y \in d_i$ is wrong because the samples from decision d_i may take x as the nearest sample with different class. As shown in Fig. 3, the samples are divided into two classes and two samples x_1 and x_2 are mislabeled. Then, we will get that x_1 is the nearest miss of most samples in d_1 and x_2 is the nearest miss of most samples in d_2. Obviously, mislabeling has great influence on the value of dependency.

The first problem is addressed by introducing the coefficient of classification certainty in Section 4, where the classification certainty of a sample is defined as the difference of the membership of the lower approximation of its class and the sum of the membership of the upper approximations of other classes:

$$\begin{aligned}\omega_B^{S-T}(x) &= \underline{k_S} d_l(x) - \sum_{d_i \subseteq D - d_l} \overline{k_T} d_i(x) \text{ or } \omega_B^{\theta-\sigma}(x) \\ &= \underline{k_\theta} d_l(x) - \sum_{d_i \subseteq D - d_l} \overline{k_\sigma} d_i(x).\end{aligned}$$

The idea behind the classification certainty is that a good feature subspace should make samples get great memberships to the lower approximations of their classes, in the same time; it should make samples far from other classes. This idea is similar to that conveyed by the Fisher's criterion [7]. Then, the quality of features is characterized as the classification certainty:

$$\omega_B^{S-T}(D) = \frac{1}{|U|} \sum_{x \in U} \omega_B^{S-T}(x) \text{ or } \omega_B^{\theta-\sigma}(D) = \frac{1}{|U|} \sum_{x \in U} \omega_B^{\theta-\sigma}(x).$$

The second problem is alleviated by introducing the idea present in ReliefF. We search k nearest neighbors from each class for computing fuzzy dependency and classification certainty, and use their average to estimate the quality of features. We call them generalized dependency functions or generalized classification certainty functions,

$$\begin{aligned}g\gamma_B^S(D) &= \frac{1}{(I-1)k|U|} \sum_{x_i \in U} \sum_{d \in D, d \neq d_i} \sum_{y \in H_d^i} [1 - k(x,y)], \\ g\gamma_B^\theta(D) &= \frac{1}{(I-1)k|U|} \sum_{x_i \in U} \sum_{d \in D, d \neq d_i} \sum_{y \in H_d^i} \sqrt{1 - k(x,y)^2}, \\ g\omega_B^{S-T}(D) &= \frac{1}{(I-1)k|U|} \sum_{x_i} \left\{ \sum_{H^i} \left[1 - k\left(x_i, H_j^i\right)\right] \right. \\ &\quad \left. - \sum_l \sum_m k\left(x_i, M_{lm}^i\right) \right\}, \quad \text{and} \\ g\omega_B^{\theta-\sigma}(D) &= \frac{1}{(I-1)k|U|} \sum_{x_i} \left\{ \sum_{H^i} \sqrt{1 - k\left(x_i, H_j^i\right)^2} \right. \\ &\quad \left. - \sum_l \sum_m \left[1 - \sqrt{1 - k\left(x_i, M_{lm}^i\right)^2}\right] \right\}.\end{aligned}$$

Although $\omega_B^{S-T}(D)$, $g\gamma_B^S(D)$, $g\gamma_B^\theta(D)$, $g\omega_B^{S-T}(D)$, and $g\omega_B^{\theta-\sigma}(D)$ reflect more upon information of the data, and may be more robust to noisy samples than dependency, they are not monotonous with respect to the features, just as Relief and ReliefF coefficient. Namely, if $B \subseteq C$, we cannot guarantee that

$$\begin{aligned}\omega_B^{S-T}(D) \le \omega_C^{S-T}(D), \quad g\gamma_B^S(D) \le g\gamma_C^S(D), \quad g\gamma_B^\theta(D) \\ \le g\gamma_C^\theta(D), \quad g\omega_B^{S-T}(D) \le g\omega_C^{S-T}(D),\end{aligned}$$

and $g\omega_B^{\theta-\sigma}(D) \le g\omega_C^{\theta-\sigma}(D)$.

6.2 Search Strategies

There are two key problems in constructing an algorithm for feature selection: feature evaluation and search strategies. The first one is to tell us how good it is if we are given a set of features or a feature and the second one is how to search the optimal features with respect to the given evaluation function. Given N features to be selected, there are $2^N - 1$ subsets of features. It is infeasible to evaluate the subsets one by one even if we are confronted with a data set of moderate size. Therefore, suboptimal solutions should be developed [26].

There are several methods to find suboptimal subsets of features making use of the proposed measures. First, we can evaluate each feature with these coefficients and rank them in terms of the feature quality. Then m best features are selected, where $m < N$ may be prespecified according to the application context. We call it ranking technique [26], [16]. Since we do not repeatedly compute the quality of feature subsets in ranking, this technique is usually efficient. However, ranking can just reflect the dependency between input features and decision, it is not able to reduce the redundant features if there are some redundant features which produce a great value of dependency and they are also high correlative with each other. Generally speaking, the objective of feature selection is to find a minimal subset of features which are sufficient and indispensable [12], [46]. Therefore, these redundant features should be deleted.

In order to reduce the redundancy of the selected features, forward or backward greedy algorithms based on heuristic knowledge can be constructed. In the forward search, we start with an empty set of features, and select one or several best features in terms of the evaluating coefficient in each round until adding any rest feature does not bring significantly improvement of classification quality. On the other hand, backward search starts with the whole set of features, and we delete one or several worst features in each round until the classification quality decreases. The feature selection algorithms with these strategies are usually $O(N^2)$ time complexity, where N is the number of candidate features. Sometimes, the computational complexity of $O(N^2)$ is also too high when we are dealing with data of great volume.

In [46], Yu and Liu proposed a more efficient search technique for feature selection. The method involves two related steps: first, we compute the dependency between the input features and decision, where dependency is estimated with symmetrical uncertainty, and selects part of the relevant features based on a predefined threshold. Second, the method ranks these features in the descending order based on the values of dependency. The so-called predominant features are selected again and those features whose correlation with these predominant features is higher than a given threshold are deleted. The first step guarantees the selected are relevant to the decision as the features with greater relevance are selected, while the second step reduce the redundant information present in the original data. This technique was shown to be very efficient in most applications. How to specify these two thresholds used in the algorithm becomes the main problem encountered in applications.

How to design the stopping criterion is another important issue in constructing an effective algorithm for feature selection and classification learning. Stopping the search too early leads to insufficient features used for learning; otherwise overfitting may occur. In decision tree learning, one usually introduces a postpruning strategy to prevent overfitting, where the decision tree fully grows in the first step, and then it is pruned with a cross validation technique [2], [8]. In this work, we will bring this idea to the problem of feature selection. In the first step, we select features by running a forward greedy search algorithm and record the order the features have been selected. In the second step, we evaluate the selected features one by one and keep the best subset. Namely, in k*th* time, we evaluate the first k features which are selected in the first step with a classification algorithm based on cross validation. Generally speaking, the classification performance improves in the beginning as we add the features one by one, and reaches a peak, and afterward decreases or is kept at the same value. We select the subset of the features which generates the highest classification accuracy and is of the lowest cardinality. This technique does not require prespecifying the threshold to terminate the search.

In evaluating features, we encounter the same level of computational complexity as in ReliefF, where k nearest samples from the same class and k nearest samples from each distinct class are to be found. The time complexity in evaluating N features is $O(N \cdot n \cdot \log n)$ if there are n samples [34]. The total complexity of selecting features with a forward greedy search strategy is $O(N^2 \cdot n \cdot \log n)$.

Certainly, there are still other search techniques to be used in feature selection, such as Genetic Algorithms, Particle Swarm Optimization, Branch and Bound, etc., [12], [21], [26], [27]. While those are promising optimization vehicles, they are not within the scope of this study.

7 Experimental Studies

Feature evaluation and selection are the main applications of kernelized fuzzy rough sets which will be discussed in detail. For comparative reasons, we consider the use of some classical techniques. First, we introduce Relief and its variant ReliefF [34] to compare the performance in feature evaluation and their ranking. As we design different coefficients to evaluate the features based on k nearest neighbors, we compare Relief with fuzzy dependency and classification certainty in $S-T$ fuzzy rough sets and $\theta-\sigma$ fuzzy rough sets. Second, we use CFS [46] and FCBF [13] to compare the performance of feature selection. Moreover, CART [3], NEC [19], linear SVM, and RBF-SVM [4] are employed to validate the selected subsets of features, where all the parameters used in these algorithms are specified as the default values and OSU-SVM3.00 software package is used (http://www.ece.osu.edu/~maj/osu_svm/). As we know, CART, NEC, and RBF-SVM are nonlinear classifiers, while linear SVM is a linear learner, we intend to compare the performance of these classifiers.

The data used in the experiments, described in Table 4, come from the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/), where N1 and N2 stand for the numbers of numerical and nominal features, respectively, For CART, NEC, linear-SVM, and RBF-SVM we use the average classification accuracies and standard deviation of the raw data sets based on 10-fold cross validation.

We first evaluate the quality of single features and rank the candidate features. In order to compare the effectiveness of different feature ranking techniques, we generate a series of subsets of the k best features with respect to the corresponding evaluation methods, where $\mathrm{k} = 1, 2, 3, \ldots$ We determine the classification performance of these feature subsets based on 10-fold cross validation. Data hepatitis, sonar, and wine are used in these experiments. The results are presented in Figs. 4, 5, 6, 7, 8, and 9, where GDS and GD-theta stand for the classification accuracies of generalized dependency based on $S-T$ fuzzy rough sets and $\theta-\sigma$ fuzzy rough sets, respectively, while GWS and GW-theta denote the classification accuracies of generalized classification certainty based on $S-T$ fuzzy rough sets and $\theta-\sigma$ fuzzy rough sets, respectively. Moreover, $\mathrm{K} = 1$ states that we compute dependency and classification certainty with a single nearest neighbor, while $\mathrm{K} = 5$ states that five nearest neighbors are used.

The experimental results demonstrate that GD-theta and GW-theta are better than GDS and GWS in most cases. As we presented in theorem 4 that $\underline{k_S} d_i(x) \leq \underline{k_\theta} d_i(x)$ and

TABLE 4
Data Description and Classification Performance Obtained on Raw Data

Data	Samples	N1	N2	Class	CART	NEC	Linear-SVM	RBF-SVM
anneal	798	6	32	5	99.89±0.35	89.01±0.46	100.00±0.0	99.89±0.35
credit	690	6	9	2	82.73±14.86	80.72±15.45	85.48±18.51	81.44±7.18
german	1000	3	17	2	69.90±3.54	73.50±3.44	73.70±4.72	70.40±0.52
heart	270	7	6	2	74.07±6.30	80.00±5.57	83.33±5.31	81.11±7.50
hepatitis	155	6	13	2	91.00±5.45	85.33±5.25	86.17±7.70	83.50±5.35
horse	368	7	15	2	95.92±2.30	90.78±4.02	92.96±4.43	72.30±3.63
iono	351	34	0	2	87.55±6.93	64.12±1.08	87.57±6.45	93.79±5.08
sick	2800	6	23	2	98.46±1.18	93.79±0.25	93.89±0.11	93.82±0.24
sonar	208	60	0	2	72.07±13.94	84.69±8.85	77.86±7.05	85.10±9.49
wdbc	569	31	0	2	90.50±4.55	96.14±2.31	97.73±2.48	98.08±2.25
wine	178	13	0	3	89.86±6.35	97.15±3.01	98.89±2.34	98.89±2.34
wpbc	198	33	0	2	70.63±7.54	79.26±6.20	77.37±7.73	77.37±7.73

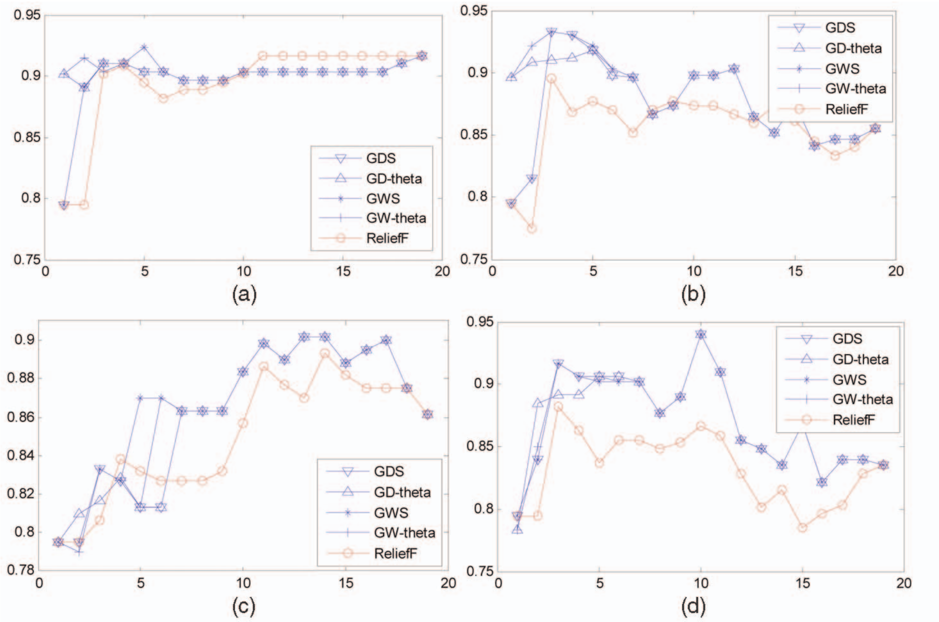


Fig. 4. Average classification accuracies based on ranking; Hepatitis data ($K = 1$). (a) CART. (b) NEC. (c) Linear SVM. (d) RBF-SVM.

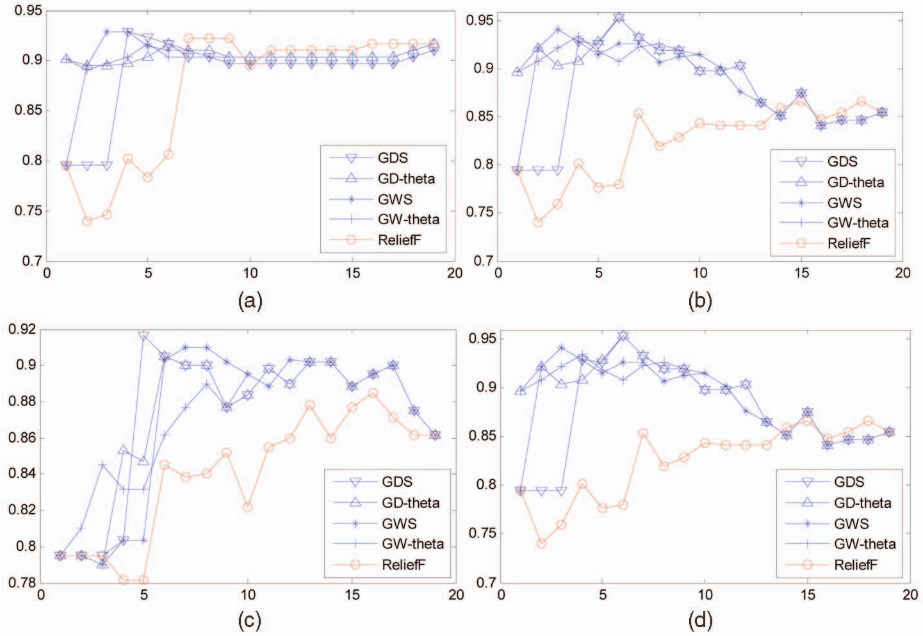


Fig. 5. Average classification accuracies based on ranking; Hepatitis data ($K = 5$). (a) CART. (b) NEC. (c) Linear SVM. (d) RBF-SVM.

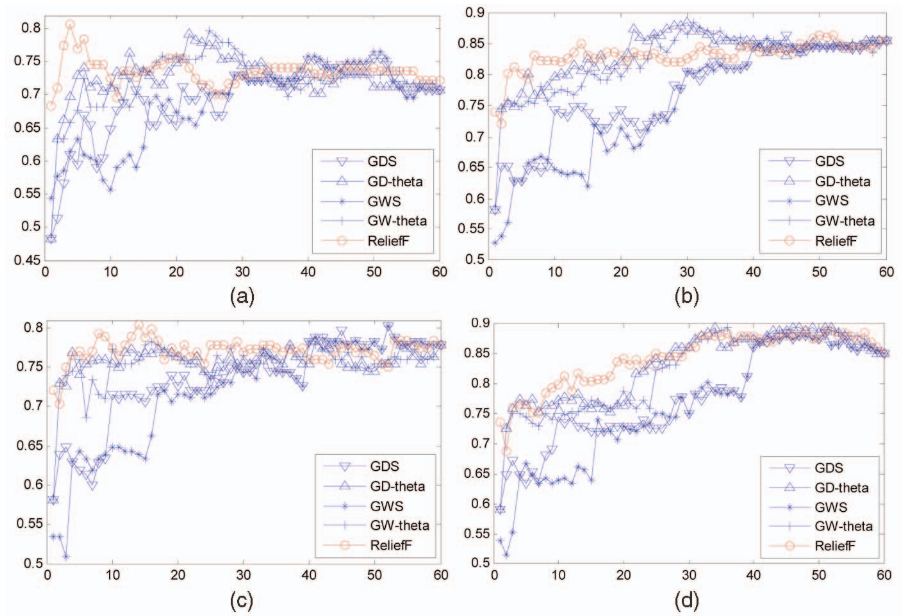


Fig. 6. Average classification accuracies based on ranking (Sonar data, $\mathrm{k} = 1$). (a) CART. (b) NEC. (c) Linear SVM. (d) RBF-SVM.

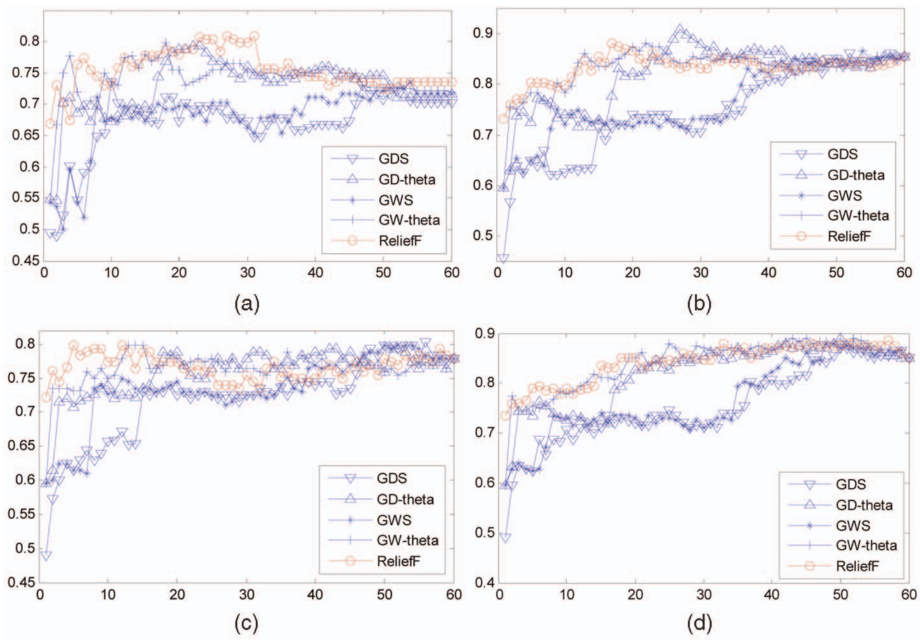


Fig. 7. Average classification accuracies based on ranking (Sonar data, $\mathrm{k} = 5$). (a) CART. (b) NEC. (c) Linear SVM. (d) RBF-SVM.

$\overline{k_T} d_i(x) \geq \overline{k_\sigma} d_i(x)$, then it is easy to show that $\gamma_B^S(D) \leq \gamma_B^\theta(D)$, $g\gamma_B^S(D) \leq g\gamma_B^\theta(D)$, $\omega_B^{S-T}(D) \leq \omega_B^{\theta-\sigma}(D)$, and $g\omega_B^{S-T}(D) \leq g\omega_B^{\theta-\sigma}(D)$. It means that with the same kernel function and the same features, we can more precisely approximate the decision with $\theta - \sigma$ fuzzy rough sets. In essence, this property results from the local approximation of these two kinds of fuzzy operators because the function $\sqrt{1 - k^2(x, y)}$ converges to zero faster than $1 - k(x, y)$. Moreover, generalized dependency and classification certainty are better than or equivalent to Relief and ReliefF algorithms in both cases of $\mathrm{k} = 1$ and $\mathrm{k} = 5$.

Ranking-based feature selection cannot delete the relevant but redundant features from the original data. Now, we introduce a two-step strategy to select features. First, we reduce features with a forward greedy search strategy directed by the feature evaluation functions GDS, GD-theta, GWS, and GW-theta, and then we use cross validation to further eliminate useless features from the selected subsets. At the same time, CFS and FCBF algorithms are introduced for comparative analysis. The classification results of the selected features are given in Tables 5, 6, 7, and 8, where in boldface we show the cases where the average accuracies of the algorithms are higher than both CFS and FCBF. The numbers of the corresponding features are also shown (Tables 9, 10, 11, and 12).

The proposed 4 algorithms outperform CFS and FCBF as to classification algorithms CART, NEC, and RBF-SVM in most cases. GDS, GWS, GD-theta, and GW-theta produce higher classification accuracies with fewer features. This

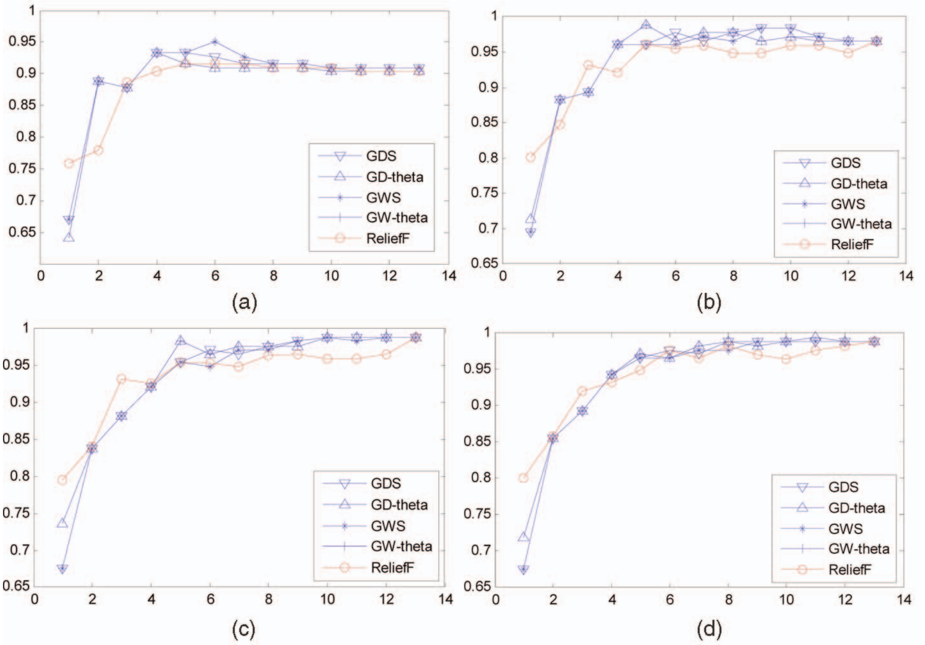


Fig. 8. Average classification accuracies based on ranking (Wine data, $k = 1$). (a) CART. (b) NEC. (c) Linear SVM. (d) RBF-SVM.

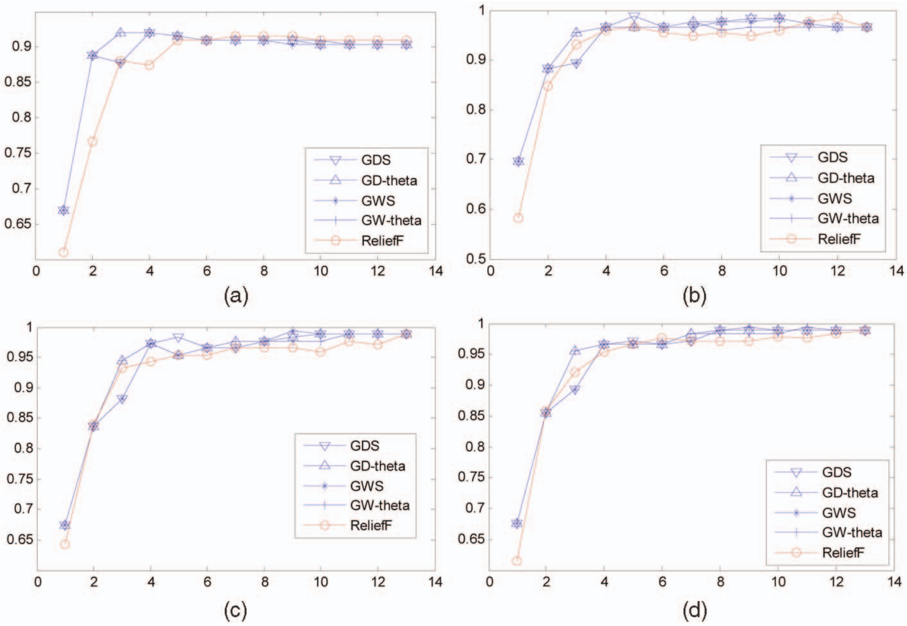


Fig. 9. Average classification accuracies based on ranking (Wine data, $k = 5$). (a) CART. (b) NEC. (c) Linear SVM. (d) RBF-SVM.

shows that these four evaluation functions are able to find the useful and irredundant subsets of features for CART, NEC, and RBF-SVM, which are nonlinear learning machines.

From Tables 6 and 10, we can get that GDS, GWS, GD-theta, and GW-theta are equivalent to or even worse than CFS and FCBF with respect to linear SVM. As we know, linear SVM is a linear classification algorithm and we also know that not only GDS, GWS, GD-theta, and GW-theta, but also CFS and FCBF cannot reflect the difference between linear and nonlinear classification problems because the feature evaluation functions in these algorithms are not sensitive to nonlinearity of data. In this sense, these feature evaluation functions are not applicable to select features for linear learning algorithms.

8 Conclusion and Future Work

There are two important aspects to be discussed in fuzzy rough set-based data analysis. One is how to generate fuzzy relations and fuzzy information granules from the data; the other is how to approximate arbitrary fuzzy subset with the fuzzy information granules. The existing research is mainly focused on the second problem, and little work has been done to develop a technique for generating effective fuzzy granulation of the universe. In this study, we introduce a class of kernel functions to extract fuzzy T-equivalence relations and fuzzy T-equivalent information granules from the given data, and then use the fuzzy granules induced by the kernel to approximate the decision. It is interesting to

TABLE 5
CART (Percent)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**100±0.0**	**100±0.0**	**100±0.0**	**100±0.0**	96.33±1.81	88.08±0.53
credit	**82.28±14.79**	**83.32±13.29**	79.68±14.55	80.86±13.43	81.43±12.50	80.12±13.89
german	**71.00±5.50**	67.70 ±3.65	**70.20±2.44**	69.70±4.57	69.80±4.87	69.70±4.72
heart	77.41±8.81	75.19±6.99	**80.00±7.65**	77.41±8.81	74.81±7.37	77.41±7.08
hepatitis	92.33±6.68	90.33±3.31	92.33±6.68	91.00±4.46	91.00±5.45	93.00±7.11
horse	**96.47±1.30**	**96.20±1.88**	91.30±4.21	**96.47±1.30**	93.48±4.06	95.93±1.90
iono	**90.68±5.87**	**90.08±5.24**	89.79±3.70	88.13±6.62	90.01±5.41	87.80±6.96
sick	**98.36±0.78**	**98.32±0.67**	**98.54±1.08**	**98.36±0.78**	95.18±1.30	95.21±1.28
sonar	**71.62±5.76**	**71.62±5.76**	**75.05±11.06**	**74.48±7.43**	70.17±12.36	70.62±12.11
wdbc	94.20±1.66	94.03±2.74	94.20±1.66	94.20±1.66	94.56±3.45	94.02±4.58
wine	**92.08±4.81**	**92.08±4.81**	**92.08±4.81**	**92.08±4.81**	91.46±6.87	90.42±6.50
wpbc	**73.16±13.05**	71.61±13.23	**76.26±7.17**	71.051±2.19	71.58±12.87	72.66±10.62

TABLE 6
NEC (Percent)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**100±0.0**	**100±0.0**	**100±0.0**	**100±0.0**	96.33±1.81	96.33±1.81
credit	85.48±18.50	**85.63±18.54**	85.48±18.51	85.20±18.23	85.48±18.45	85.48±18.51
german	71.00±3.37	73.10±4.91	70.50±2.37	73.70±4.67	74.70±4.47	74.70±4.47
heart	80.00±6.58	79.26±10.36	**81.48±7.41**	**81.85±6.86**	80.74±6.00	80.37±4.95
hepatitis	**93.67±5.08**	**91.50±8.97**	**92.17±6.94**	**91.50±8.97**	90.83±9.07	90.17±8.55
horse	**95.90±3.53**	**95.90±3.53**	89.95±4.24	**95.90±3.53**	91.58±5.29	91.87±3.94
iono	**92.92±5.56**	**92.10±4.98**	**90.66±5.74**	88.14±5.62	88.39±5.17	88.07±3.92
sick	**94.04±0.17**	**94.00±0.15**	**94.04 ±0.17**	**94.04±0.17**	93.89±0.11	93.89±0.11
sonar	**82.19±6.85**	**82.19±6.85**	**86.02±7.55**	**84.67±7.30**	78.00±8.34	80.31±4.65
wdbc	95.61±2.22	95.61±2.22	**96.32±1.92**	**96.14±1.98**	95.61±1.48	94.74±2.15
wine	98.26±2.80	97.22±3.93	98.33±2.68	98.26±2.80	93.19±4.56	98.33±2.68
wpbc	**78.71±7.77**	**77.82±5.74**	**80.39±9.18**	**78.71±7.77**	77.39±7.72	77.34±5.16

TABLE 7
Linear SVM (Percent, $k = 1$)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**99.89±0.3**	**99.89±0.3**	**96.33±1.05**	**96.33±1.05**	93.88±0.78	87.20±0.91
credit	85.48±18.51	85.48±18.51	56.23±1.25	85.48±18.51	85.48±18.51	85.48±18.51
german	70.30±1.83	70.50±3.50	70.00±0.00	**73.70±4.57**	72.90±5.26	72.90±5.26
heart	84.07±6.06	82.22±7.96	82.96±5.30	83.33±5.59	84.44±6.00	82.22±5.47
hepatitis	90.33±4.57	85.00±6.53	91.17±7.20	88.83±5.67	91.50±6.40	90.17±6.59
horse	90.22±4.13	**91.57±5.18**	90.21±3.91	90.22±4.13	90.76±4.82	91.03±4.96
iono	**88.96±6.39**	84.43±5.58	**90.33±3.81**	85.28±6.62	87.84±5.39	83.22±6.35
sick	93.89±0.11	93.89±0.11	93.89±0.11	93.89±0.11	93.89±0.11	93.89±0.11
sonar	76.40±8.54	76.40±8.54	**78.81±9.91**	76.93±5.47	76.52±7.10	77.93±7.12
wdbc	**96.84±1.59**	95.96±2.03	**97.55±2.06**	**97.20±2.36**	96.32±2.26	95.80±2.84
wine	98.33±2.68	95.56±4.38	98.33±2.34	98.33±2.68	95.49±3.54	98.89±2.34
wpbc	76.32±3.04	76.32±3.04	76.32±3.04	76.32±3.04	76.32±3.04	76.32±3.04

TABLE 8
RBF-SVM (Percent)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**99.89±0.3**	**99.89±0.3**	**99.89±0.50**	**99.89±0.3**	93.66±1.15	87.20±0.91
credit	**85.92±18.39**	**85.63±18.48**	**56.37±1.25**	**85.63±18.48**	84.61±17.95	85.05±1.779
german	**71.80±2.44**	**72.70±4.74**	70.00±0.00	**71.90±2.69**	71.10±3.96	71.10±3.96
heart	**85.93±6.25**	**81.11±7.90**	**85.93±6.94**	**85.93±6.25**	80.74±6.94	80.74±5.47
hepatitis	**91.67±6.89**	85.67±6.30	**92.17±1.58**	**90.83±6.54**	89.67±7.11	89.67±5.54
horse	91.05±3.96	92.13±4.81	91.82±1.26	91.05±3.96	88.59±5.20	91.59±5.13
iono	**93.22±4.82**	**92.08±5.76**	**95.16±8.66**	**92.90±5.22**	90.93±5.77	89.51±3.89
sick	93.93±0.17	93.93±0.17	93.89±0.11	93.93±0.17	93.89±0.11	93.89±0.11
sonar	79.76±8.30	79.76±8.30	**82.26±2.46**	**81.26±5.66**	76.05±7.62	80.29±8.35
wdbc	96.67±2.38	96.32±1.73	**97.90±3.52**	**97.90±2.31**	96.49±2.02	96.50±2.71
wine	98.33±2.68	96.67±3.88	98.89±2.68	98.33±2.68	95.49±3.54	98.89±2.34
wpbc	**77.34±4.66**	76.84±4.61	**79.37±3.04**	76.84±4.03	76.32±3.04	76.32±3.04

TABLE 9
Number of Selected Features (CART)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**3**	**3**	**3**	**3**	5	6
credit	**6**	**6**	**5**	**6**	8	7
german	6	9	1	10	5	5
heart	**5**	**6**	**4**	**5**	10	6
hepatitis	**5**	**5**	**6**	8	6	7
horse	**4**	3	**5**	**4**	7	8
iono	6	6	13	8	4	4
sick	10	8	15	10	6	6
sonar	**6**	**6**	**8**	**6**	9	10
wdbc	**4**	**4**	**4**	**4**	6	7
wine	**3**	**3**	**3**	**3**	5	10
wpbc	5	5	7	6	3	2

TABLE 10
Number of Selected Features (NEC)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**3**	**3**	**3**	**3**	5	6
credit	**6**	**6**	**5**	**6**	8	7
german	6	**5**	**1**	**4**	5	5
heart	7	7	8	9	10	6
hepatitis	7	**1**	9	**1**	6	7
horse	**1**	**1**	**5**	**1**	7	8
iono	6	5	6	5	4	4
sick	8	9	9	8	6	6
sonar	**6**	**6**	10	**6**	9	10
wdbc	**3**	**3**	11	12	6	7
wine	**5**	**4**	8	**5**	5	10
wpbc	4	6	13	4	3	2

TABLE 11
Number of Selected Features (Linear SVM)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**3**	**3**	**3**	**3**	5	6
credit	**4**	**5**	**1**	**4**	8	7
german	9	9	1	10	5	5
heart	9	6	10	7	10	6
hepatitis	**5**	**5**	**6**	7	6	7
horse	**4**	**7**	**4**	**4**	7	8
iono	**2**	**5**	11	9	4	4
sick	**1**	**1**	**1**	**1**	6	6
sonar	**4**	**4**	**7**	**6**	9	10
wdbc	7	**4**	13	11	6	7
wine	6	**4**	6	6	5	10
wpbc	**1**	**1**	**1**	**1**	3	2

TABLE 12
Number of Selected Features (RBF-SVM)

Data	GDS	GD-THETA	GWS	GW-THETA	CFS	FCBF
anneal	**3**	**3**	**3**	**3**	5	6
credit	**5**	**5**	**1**	**5**	8	7
german	7	6	1	7	5	5
heart	**6**	**6**	**4**	**6**	10	6
hepatitis	**3**	**5**	**5**	**6**	6	7
horse	**5**	**6**	**3**	**5**	7	8
iono	8	5	9	9	4	4
sick	8	9	*1*	8	6	6
sonar	**4**	**4**	14	**6**	9	10
wdbc	7	4	13	11	6	7
wine	6	**4**	**4**	6	5	10
wpbc	7	6	6	10	3	2

note that the fuzzy relations computed with the reflexive and symmetric kernel functions are fuzzy T-equivalence relations [29], [30], which is the footstone for most of fuzzy rough set models. By introducing this model, we develop an important bridge between rough sets and kernel techniques.

We extend the measures of approximation accuracy, approximation quality, and dependency to the fuzzy cases for evaluating quality of kernel-based approximation; we also introduce a new measure, called classification certainty, to reflect a sample's certainty belonging to one of the decisions. Based on these functions, we introduced several coefficients to evaluate approximation quality of attributes and discuss the relationship between the feature evaluation function ReliefF and fuzzy dependency and classification certainty. We have noted that fuzzy rough set-based dependency function can be understood as a kernelized version of ReliefF. We introduce the idea in ReliefF to enhance the robustness of rough set-based measures.

These generalized measures are used to feature evaluation and feature selection. We compare the proposed measures with ReliefF in feature ranking and the results show that some of the measures outperform ReliefF in most cases. Moreover, we also establish a two-step strategy to select relevant and necessary features for classification. The experiments show that the proposed dependency and classification certainty functions based on $S-T$ fuzzy rough sets and $\theta-\sigma$ fuzzy rough sets perform better than CFS and FCBF with respect to the quality of the nonlinear classification algorithms.

Kernelized fuzzy rough sets are not only used to evaluate the features, but also considered to compute the memberships of samples to the lower approximation and upper approximation of decision classes. These memberships are considered as the degree that a sample certainly belongs to the class and the degree that a sample possibly belongs to the classes, respectively. These memberships are useful for instance-based learning [31], [44] and fuzzy support vector machines [25], which require the membership values to weight the samples in classification. In the future, we intend to investigate the applications of kernelized fuzzy rough sets to these domains.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants 60703013, 10978011 and 70871036. Professor Daren Yu is supported by National Science Fund for Distinguished Young Scholars under Grant 50925625.

References

[1] R.B. Bhatt and M. Gopal, "FRCT: Fuzzy-Rough Classification Trees," *Pattern Analysis and Applications,* vol. 11, no. 1, pp. 73-88, 2008.

[2] M. Bohanec and I. Bratko, "Trading Accuracy For Simplicity in Decision Trees," *Machine Learning,* vol. 15, pp. 223-250, 1994.

[3] L. Breiman, *Classification and Regression Trees.* Chapman & Hall, 1993.

[4] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[5] M. Dash and H. Liu, "Consistency-Based Search in Feature Selection," *Artificial Intelligence,* vol. 151, nos. 1/2, pp. 155-176, 2003.

[6] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets," *Int'l J. General Systems,* vol. 17, nos. 2/3, pp. 191-209, 1990.

[7] R. Duda, P. Hart, and D.G. Stork, *Pattern Classification,* second ed. John Wiley and Sons, Inc., 2001.

[8] F. Esposito, D. Malerba, and G. Semeraro, "A Comparative Analysis of Methods for Pruning Decision Trees," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 476-491, May 1997.

[9] 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 Man and Cybernetics Part C-Applications and Rev.,* vol. 37, no. 1, pp. 138-146, Jan. 2007.

[10] M. Genton, "Classes of Kernels for Machine Learning: A Statistics Perspective," *J. Machine Learning Research,* vol. 2, pp. 299-312, 2001.

[11] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel Methods for Measuring Independence," *J. Machine Learning Research,* vol. 6, pp. 2075-2129, 2005.

[12] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[13] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," Hamilton, New Zealand, 1998.

[14] A. Hassanien, "Fuzzy Rough Sets Hybrid Scheme for Breast Cancer Detection," *Image and Vision Computing,* vol. 25, no. 2, pp. 172-183, 2007.

[15] T.P. Hong et al., "Learning a Coverage Set of Maximally General Fuzzy Rules by Rough Sets," *Expert Systems with Applications,* vol. 19, no. 2, pp. 97-103, 2000.

[16] S.J. Hong, "Use of Contextual Information for Feature Ranking and Discretization," *IEEE Trans. Knowledge and Data Eng.,* vol. 9, no. 5, pp. 718-730, Sep./Oct. 1997.

[17] Q.H. Hu, D.R. Yu, and Z.X. Xie, "Information-Preserving Hybrid Data Reduction Based on Fuzzy-Rough Techniques," *Pattern Recognition Letters,* vol. 27, no. 5, pp. 414-423, 2006.

[18] Q.H. Hu, Z.X. Xie, and D.R. Yu, "Hybrid Attribute Reduction Based on a Novel Fuzzy-Rough Model and Information Granulation," *Pattern Recognition,* vol. 40, no. 12, pp. 3509-3521, 2007.

[19] Q.H. Hu, D.R. Yu, and Z.X. Xie, "Neighborhood Classifiers," *Expert Systems with Applications,* vol. 34, pp. 866-876, 2008.

[20] R. Jensen and Q. Shen, "Fuzzy-Rough Sets Assisted Attribute Selection," *IEEE Trans. Fuzzy Systems,* vol. 15, no. 1, pp. 73-89, Feb. 2007.

[21] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches," *IEEE Trans. Knowledge and Data Eng.,* vol. 16, no. 12, pp. 1457-1471, Dec. 2004.

[22] K. Kira and L.A. Rendell, "A Practical Approach to Feature Selection," *Proc. Ninth Int'l Conf. Machine Learning (ICML '92),* D. Sleeman and P. dwards, eds., pp. 249-256. 1992.

[23] E.P. klement, R. Mesiar, and E. Pap, *Triangular Norms.* Kluwer Academic Publishers, 2001.

[24] I. Kononenko, "Estimating Attributes: Analysis and Extensions of Relief," *Machine Learning: ECML-94,* L. De Raedt and F. Bergadano, eds., pp. 171-182, Springer Verlag, 1994.

[25] C.F. Lin and S.D. Wang, "Fuzzy Support Vector Machines," *IEEE Trans. Neural Networks,* vol. 13, no. 2, pp. 464-471, Mar. 2002.

[26] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Trans. Knowledge and Data Eng.,* vol. 17, no. 4, pp. 491-502, Apr. 2005.

[27] J. Mi and W. Zhang, "An Axiomatic Characterization of a Fuzzy Generalization of Rough Sets," *Information Sciences,* vol. 160, pp. 235-249, 2004.

[28] N.N. Morsi and M.M. Yakout, "Axiomatics for Fuzzy Rough Set," *Fuzzy Sets System,* vol. 100, pp. 327-342, 1998.

[29] B. Moser, "On the T-Transitivity of Kernels," *Fuzzy Sets and Systems,* vol. 157, pp. 1787-1796, 2006.

[30] B. Moser, "On Representing and Generating Kernels by Fuzzy Equivalence Relations," *J. Machine Learning Research,* vol. 7, pp. 2603-2620, 2006.

[31] R. Paredes and E. Vidal, "Learning Weighted Metrics to Minimize Nearest-Neighbor Classification Error," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 7, pp. 1100-1110, July 2006.

[32] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data.* Kluwer Academic Publishers, 1991.

[33] A.M. Radzikowska and E.E. Kerre, "A Comparative Study of Fuzzy Rough Sets," *Fuzzy Sets and Systems,* vol. 126, pp. 137-155, 2002.

[34] M. Robnik-sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning,* vol. 53, pp. 23-69, 2003.

[35] B. Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. 10, pp. 1299-1319, 1998.

[36] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* Cambridge Univ. Press, 2004.

[37] P. Somol, P. Pudil, and J. Kittler, "Fast Branch & Bound Algorithms for Optimal Feature Selection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 7, pp. 900-912, July 2004.

[38] Y.J. Sun, "Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Application," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 6, pp. 1035-1051, June 2007.

[39] Y.C. Tsai, C.H. Cheng, and J.R. Chang, "Entropy-Based Fuzzy Rough Classification Approach for Extracting Classification Rules," *Expert Systems with Applications,* vol. 31, no. 2, pp. 436-443, 2006.

[40] Y.F. Wang, "Mining Stock Price Using Fuzzy Rough Set System," *Expert Systems with Applications,* vol. 24, no. 1, pp. 13-23, 2003.

[41] X.Z. Wang et al., "Learning Fuzzy Rules from Fuzzy Samples Based on Rough Set Technique," *Information Sciences,* vol. 177, no. 20, pp. 4493-4514, 2007.

[42] Q. Wu, Y. Ying, and D.-X. Zhou, "Multi-Kernel Regularized Classifiers," *J. Complexity,* vol. 23, pp. 108 -134, 2007.

[43] W. Wu and W. Zhang, "Constructive and Axiomatic Approaches of Fuzzy Approximation Operators," *Information Sciences,* vol. 159, pp. 233-254, 2004.

[44] R.R. Yager, "Using Fuzzy Methods to Model Nearest Neighbor Rules," *IEEE Trans. Systems, Man, and Cybernetics—Part B: Cybernetics,* vol. 32, no. 4, pp. 512-525, Aug. 2002.

[45] D.S. Yeung, D.-G. Chen, E.C.C. Tsang, J.W.T. Lee, and X.-Z Wang, "On the Generalization of Fuzzy Rough Sets," *IEEE Trans. Fuzzy Systems,* vol. 13, no. 3, pp. 343-361, June 2005.

[46] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *J. Machine Learning Research,* vol. 5, pp. 1205-1224, 2004.

[47] L.A. Zadeh, "Fuzzy $\text{Logic} = \text{Computing}$ with Words," *IEEE Trans. Fuzzy Systems,* vol. 4, no. 2, pp. 103-111, May 1996.

[48] L.A. Zadeh, "Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic," *Fuzzy Sets and Systems,* vol. 90, no. 2, pp. 111-127, 1997.

[49] W. Zhu and F.Y. Wang, "On Three Types of Covering-Based Rough Sets," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 8, pp. 1131-1144, Aug. 2007.

[50] W.-Z. Wu., "Attribute Reduction Based on Evidence Theory in Incomplete Decision Systems," *Information Sciences,* vol. 178, no. 5, pp. 1355-1371, 2008.

[51] Q.H. Hu, D.R. Yu, J. Liu, and C. Wu., "Neighborhood Rough Set Based Heterogeneous Feature Subset Selection," *Information Sciences,* vol. 178, no. 18, pp. 3577-3594, 2008.

[52] X. Liu, W. Pedrycz, and M. Song, "The Development of Fuzzy Rough Sets with the Use of Structures and Algebras of Axiomatic Fuzzy Sets," *IEEE Trans. Knowledge and Data Eng.,* vol. 23, no. 3, pp. 443-462, Mar. 2009.

[53] P. Maji and S.K. Pal, "Rough-Fuzzy C-Medoids Algorithm and Selection of Bio-Basis for Amino Acid Sequence Analysis," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 6, pp. 859-872, June 2007.

[54] Q. Hu, J. Liu, and D. Yu, "Stability Analysis on Rough Set Based Feature Evaluation," *Proc. Third Int'l Conf. Rough Sets and Knowledge Technology (RSKT '08),* pp. 88-96, 2008.

[55] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, "Gaussian Kernel Based Fuzzy Rough Sets: Model, Uncertainty Measures and Applications," *Int'l J. Approximate Reasoning,* vol. 51, no. 4, pp. 453-471, 2010.

Qinghua Hu received the BEng and ME degrees from Department of Power Engineering from Harbin Institute of Technology, Harbin, China in 1999 and 2002, respectively, and the PhD degree from Department of Control Science and Engineering, Harbin Institute of Technology in 2008. He is currently an associate professor with Harbin Institute of Technology, and he is also working as a postdoctoral fellow with Department of Computing, The Hong Kong Polytechnic University, Hong Kong. His research interests are focused on data mining, knowledge discovery with fuzzy, and rough techniques. He has authored or coauthored more than 60 journal and conference papers in the areas of machine learning, data mining, and rough set theory. He received the best student paper award from PRICAI2006 and Chinese Conference on Rough Sets and Soft Computing, 2007. He is now acting as PC chair for The Seventh International Conference on Rough Sets and Current Trends in Computing (RSCTC 2010). He is a member of the IEEE. For more details, please visit the URL: http://www.turbo.hit.edu.cn/members/huqinghua.

Daren Yu received the ME and PhD degrees from Harbin Institute of Technology, Harbin, China, in 1988 and 1996, respectively. Since 1988, he has been working at the School of Energy Science and Engineering, Harbin Institute of Technology. He is currently Cheung Kong professor with Harbin Institute of Technology and director of Institute of Advanced Power. He has published more than 200 conference and journal papers on power control and fault diagnosis. His main research interests are in modeling, simulation, and control of power systems.

Witold Pedrycz (M'88-SM'94-F'99) received the MSc, PhD, and DSci degrees from the Silesian University of Technology, Gliwice, Poland. He is currently a professor and a Canada Research chair (CRC) in the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. He is also with the Systems Research Institute of the Polish Academy of Sciences. He is actively pursuing research in computational intelligence, fuzzy modeling, knowledge discovery and data mining, fuzzy control including fuzzy controllers, pattern recognition, knowledge-based neural networks, relational computation, bioinformatics, and software engineering. He has published numerous papers in this area. He is also an author of nine research monographs covering various aspects of computational intelligence and software engineering. He has been a member of numerous program committees of conferences in the area of fuzzy sets and neurocomputing. He currently serves as an associate editor of the *IEEE Transactions on Systems, Man, and Cybernetics, the IEEE Transactions on Neural Networks*, and the *IEEE Transactions on Fuzzy Systems*. He is the editor-in-chief of Information Sciences and president of the International Fuzzy Systems Association (IFSA) and the North American Fuzzy Information Processing Society (NAFIPS). He is a fellow of the IEEE.

Degang Chen received the MS degree from Northeast Normal University, Changchun, Jilin, China, in 1994, and the PhD degree from Harbin Institute of Technology, Harbin, China, in 2000. He has worked as a postdoctoral fellow with Xi'an Jiaotong University, Xi'an, China, from 2000 to 2002 and with Tsinghua University, Tsinghua, China, from 2002 to 2004. Since 1994, he has been with Bohai University, Jinzhou, Liaoning, China. He is now working with the Department of Mathematics and Physics, North China Electric Power University, Beijing 102206, P.R. China. His research interests include fuzzy group, fuzzy analysis, rough sets, and SVM.