Gaussian Kernel Based Fuzzy Rough Sets: Model, Uncertainty Measures and Applications

Qinghua Hu[a, b], Lei Zhang[b], Degang Chen[c], Witold Pedrycz[d], Daren Yu[a]

a) Harbin Institute of Technology, Harbin 150001, China, b) Department of computing, The Hong Kong Polytechnic University c) North China Electric Power University, Beijing 102206, P. R. China, d) Department of Electrical and Computer Engineering, University of Alberta, Canada

Abstract: Kernel methods and rough sets are two general pursuits in the domain of machine learning and intelligent systems. Kernel methods map data into a higher dimensional feature space, where the resulting structure of the classification task is linearly separable; while rough sets granulate the universe with the use of relations and employ the induced knowledge granules to approximate arbitrary concepts existing in the problem at hand. Although it seems there is no connection between these two methodologies, both kernel methods and rough sets explicitly or implicitly dwell on relation matrices to represent the structure of sample information. Based on this observation, we combine these methodologies by incorporating Gaussian kernel with fuzzy rough sets and propose a Gaussian kernel approximation based fuzzy rough set model. Fuzzy T-equivalence relations constitute the fundamentals of most fuzzy rough set models. It is proven that fuzzy relations with Gaussian kernel are reflexive, symmetric and transitive. Gaussian kernels are introduced to acquire fuzzy relations between samples described by fuzzy or numeric attributes in order to carry out fuzzy rough data analysis. Moreover, we discuss information entropy to evaluate the kernel matrix and calculate the uncertainty of the approximation. Several functions are constructed for evaluating the significance of features based on kernel approximation and fuzzy entropy. Algorithms for feature ranking and reduction based on the proposed functions are designed. Results of experimental analysis are included to quantify the effectiveness of the proposed methods

Keyword: fuzzy set; rough set; Gaussian kernel; uncertainty measure; feature selection

1. Introduction

In the recent years, we have witnessed two types of methodologies which are widely discussed in pattern recognition and machine learning domains: kernel methods and rough sets. The first one allows mapping data into a higher dimensional feature space in order to simplify classification tasks and made them linear (viz. solvable by linear classifiers [1]). In this way, a number of linear learning algorithms can be used to deal with nonlinear tasks, such as nonlinear SVM [2, 3], kernel perceptron [4], kernel discriminant analysis [5], nonlinear component analysis [6], kernel matching pursuit [7], etc. Rough sets, forming an important conceptual tool for granular computing [10, 50, 51] [8], offer a uniform induction framework in machine learning [9]. Using this methodology, we granulate the universe of discourse into a family of elemental concepts to describe the objects, and then use these elemental concepts to approximate arbitrary subsets of the universe. Feature evaluating, variable selection, attribute reduction [10, 56], rule extraction [11, 54, 60], ensemble learning [12] and uncertainty reasoning [13, 58, 59, 61, 62] are the main developments encountered in rough sets.

Although these two learning methodologies have been widely studied, relatively little attention has been paid to explore relationships between them. In some literature, we can some hybrid structure which combine the advantages of these techniques where one reduces data with rough sets and then carry out the development of classifiers which operate on such reduced data [14, 53]. However, the linkages between these two methodologies are not discussed explicitly and rough sets based feature selection method can be replaced with any other feature selection algorithm. Asharaf et al. [15] defined a rough sphere having an inner radius R defining its lower approximation and an outer radius T>R defining its upper approximation. With these definitions a rough support vector clustering algorithm was developed. Following a similar idea, Lingras and Butz embedded rough set methodology into support vector machines for multi-class tasks [16]. Rough sets used here are employed to represent the lower and upper boundary of patterns. These studies attempted to incorporate the idea of rough sets into support vector based learning,. It seems that there are very limited developments in a hybrid, combined use of rough sets and kernel machines. However, if we observe these two learning schemes, we can note that kernel methods and rough sets based data analysis share some interesting commonalities. Let us recall the basic procedures of the two methods in pattern analysis. On one hand, a typical kernel learning algorithm consists of two functional modules: nonlinear mapping realized by kernel functions and pattern classification being completed with kernel machines [1]. Nonlinear mapping transfers the original data matrix into a kernel matrix (also called Gram matrix) which presents the structure and describes relationships between samples. Kernel matrix plays an important role in kernel learning algorithms as it contains all the information available in order to perform further learning. The learning algorithm relies on information about the training data available through the kernel matrix. On the other hand, there are also two modules in the rough set methodology: (a) granulation of data (samples) into a set of information granules according to the relation of objects and (b) approximate classification realized in the presence of such induced information granules. The rough set methodology helps extract a relation (relation matrix) dealing with samples and subsequently granulates the set of objects into a set of information granules according to the relation between objects. The objects in the granule are indistinguishable in terms of this relation. Then the information granules induced by the relation are used to approximate the classification of the universe. Obviously, relation and relation matrix form the fundamentals of rough set models. They play the same conceptual role in rough sets as kernel matrix in kernel machines. The types of rough set models are determined by the algorithms being used to extract the relationship between samples. For example, the generic rough set model considers into account an equivalence relation to partition the samples into disjoint equivalence classes [17]; neighborhood rough sets group the samples into different neighborhood information granules [18], fuzzy rough sets segment the universe with a fuzzy relation into a set of fuzzy granules and approximate fuzzy sets with these fuzzy granules [19-23, 55, 57, 58]. We can find a high level of similarity between kernel methods and rough set algorithms if we take the kernel matrix as a relation matrix or consider the relation matrix as a kernel one. In fact, one can show that the most relation matrices used in the existing rough set models satisfy the conditions of kernel functions. They are positive-semidefinite and symmetric. At the same time, kernel matrices are symmetric and some of them are reflective [24, 25]. This means that some of kernel matrices could be used as fuzzy relation matrices in fuzzy rough sets. Taking this into account, we can form a bridge between rough sets and kernel methods with the relation matrices.

We can make use of kernel functions to extract fuzzy relations for rough sets based data analysis. Although different models of fuzzy rough sets were proposed and properties of these models were discussed in literatures [19-22], little attention was paid to extract fuzzy relations from data and integrate these relations into fuzzy rough sets. The models and theories about fuzzy rough sets available in the existing literature just give a one-sided view at fuzzy rough computation as most of the existing fuzzy rough set models are constructed based on the fuzzy granulated spaces induced by fuzzy *T*-equivalence relations. Nevertheless the issue of how to generate an effective fuzzy *T*-equivalence relation from data has not been systematically discussed so far. Subsequently, the effective solutions are not present in applications. Obviously the way to generate fuzzy relations from data substantially influences the performance of rough set-based intelligent data analysis. The absence of effective techniques in this regard constitutes an obstacle for pursuing applications of fuzzy rough sets. In this study, we will introduce Gaussian kernel functions to extract fuzzy similarity relations between samples for fuzzy rough set based data analysis. Then we construct fuzzy rough models based on Gaussian kernel induced by fuzzy relations. In this way, we effectively combine fuzzy rough sets with kernel methods.

As most of the existing fuzzy rough set models are constructed based on the fuzzy granulated spaces induced by fuzzy *T*- equivalence relations, it is desirable that the extracted fuzzy relations are fuzzy *T*- equivalence relations. In this context, Moser showed that the kernel matrix computed with a reflexive kernel taking values from the unit interval is a fuzzy *T*- equivalence relation [24, 25]. Therefore such kernel functions can be considered to directly induce fuzzy *T*-equivalence relations from data. In [26], Hu et al. introduced Gaussian kernels to compute similarity between samples in fuzzy rough set based attribute reduction. The fact that Gaussian kernel matrix is a fuzzy *T*-equivalence relation was not emphasized and fully discussed at that time. Gaussian functions are reflexive and symmetric taking values in the unit interval. This emphasizes that Gaussian functions can be integrated with fuzzy rough sets to support extraction of fuzzy *T*-equivalence relations. However, to our best knowledge, no detailed analysis with this regard has been reported yet. In this study, we will construct a novel fuzzy rough set model with Gaussian kernel approximation, where sample spaces are granulated into fuzzy information granules in terms of fuzzy *T*-equivalence relations computed with Gaussian kernel. We discuss the uncertainty measures of Gaussian kernel approximation and adapt the proposed measures to evaluate the quality of the features. Some attribute ranking and reduction algorithms are proposed. The experimental analysis is covered to quantify the performance of the method.

The paper is organized as follows. In Section 2, some basic notations about fuzzy rough sets are briefly reviewed. In Section 3, the fuzzy rough set model based on Gaussian kernel approximation is proposed. Uncertainty measures of Gaussian kernel approximation are discussed in Section 4. Section 5 shows the applications of Gaussian kernel approximation to feature evaluating and feature reduction. Numeric experiments are reported in Section 6.

2. Rough sets and fuzzy rough sets: some preliminary knowledge

Given an information system $IS=(U,A)$, where $U=\{x_1,...,x_m\}$ is a nonempty finite set of objects and $A=\{a_1,a_2,...,a_n\}$ is a nonempty finite set of attributes to characterize the objects, we associate a binary relation $IND(B)$ with a subset of attributes $B\subseteq A$, called $B-$ indiscernibility relation, defined as $IND(B)=\{(x,y)\in U\times U: a(x)=a(y),\forall a\in B\}$. $IND(B)$ is an equivalence relation and $IND(B)=\bigcap_{a\in B}IND(\{a\})$. The equivalence relation partitions the objects into a family of disjoint subsets, called elemental concepts. By $[x]_B$ we denote the equivalence class induced by $IND(B)$ including x. $U/IND(B)=\{[x]_B \mid x\in U\}$. For arbitrary subset $X\subseteq U$, two sets of equivalence classes, called $B-$lower and $B-$upper approximations, are defined as $\underline{B}X=\bigcup\{[x]_B:[x]_B\subseteq X\}$ and $\overline{B}X=\bigcup\{[x]_B:[x]_B\cap X\neq\varnothing\}$, respectively. X is said definable if $\underline{B}X=\overline{B}X$; otherwise X is a rough set. As to rough set X, we call $BN(X)=\overline{B}X-\underline{B}X$ the boundary of X in (U,B).

Although a lot of applications of the classical rough set model are found, there it is a certain point that deserves more attention. That is, given the equivalence relations the above model is able to deal with symbolic-valued databases. This somewhat limits the applications of rough sets. Several generalizations of this model were proposed in [19, 22, 23]. Among these generalizations, the combination of rough sets and fuzzy sets called fuzzy rough sets offers a useful opportunity to deal with real-valued datasets where fuzzy similarity relations between samples are determined.

The concept of fuzzy rough sets was first proposed by Dubois and Prade [19]. Given a fuzzy relation R 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: $R(x,y)=R(y,x)$ and 3) transitivity: $\min_y\left(R(x,y),R(y,z)\right)\leq R(x,z)$. 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: $T\big(R(x,y), R(y,z)\big) \le R(x,z)$, where T is some triangular norm.

Let R be a fuzzy equivalence relation on *U* and X be a fuzzy subset of *U*. Then the lower and upper approximations of *X* were defined as [19]

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

The pair of Min and max is the two aggregation operations used in these calculations. In fact, there are a number of t-norms and s-norms for fuzzy aggregation. To generalize the above definition of fuzzy rough sets, *T*-equivalence relations were introduced in [20]. Given a fuzzy *T*-equivalence relation on *U* where θ is a residual implication induced with *T*, the fuzzy lower and fuzzy upper approximations of fuzzy subset X were 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 as [28]

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

More generally, Yeung, Chen, et al. proposed a model of fuzzy rough sets with a pair of t-norm T and t-conorms S in [45].

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

Overall, there are three definitions of fuzzy lower approximation operators: $\underline{R_{\max}}$, $\underline{R_{\theta}}$, $\underline{R_S}$ 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 arrive at two definitions of lower approximations and upper approximations, respectively.

The above definitions of fuzzy rough sets were constructed making use of fuzzy equivalence relations or fuzzy *T*-equivalence relations. They are straightforward generalizations of the classical rough set model. All of them reduce to the original concept of rough sets when the underlying relation is a Boolean one and *X* is a subset of *U*.

There are three essential problems to be addressed when employing a fuzzy rough set model to real-world applications: computing the fuzzy relation from data, aggregating multiple relations extracted from a set of features and defining the lower and upper approximations. The above work mainly focused on the definitions of lower and upper approximations. We will introduce Gaussian kernel function to compute the fuzzy equivalence relations between samples and discuss the aggregation of features.

3. Gaussian kernel based fuzzy rough set model

3.1 Approximating fuzzy sets with Gaussian kernel

Gaussian functions constitute a widely used category of kernels in SVM and other fields such as RBF neural networks [52]. Good performance and computational effectiveness is usually obtained with Gaussian kernel to embed nonlinear problems in higher dimensional feature spaces. In this section we introduce Gaussian kernel for computing fuzzy $T-$equivalence relations in fuzzy rough sets and thus approximate arbitrary fuzzy subsets with kernel induced fuzzy granules.

Suppose U is a finite set of samples. $x_i \in U$ is described by a vector $< x_{i1}, x_{i2}, ..., x_{in} > \in R^n$, thus U can be viewed as a subset of R^n.

The similarity between two samples is computed with Gaussian kernel function $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\delta^2)$, where $\|x_i - x_j\|$ is the Euclidean distance between samples x_i and x_j. we have

1) $k(x_i, x_j) \in [0,\ 1]$;

2) $k(x_i, x_j) = k(x_j, x_i)$;

3) $k(x_i, x_i) = 1$.

Therefore Gaussian kernel induces a fuzzy relation satisfying the properties of reflexivity and symmetry. We denote this fuzzy relation by R_G^n. In [24, 25], it was shown that R_G^n also satisfies $T_{\cos}-$transitive where $T_{\cos}(a,b) = \max\{ab - \sqrt{1-a^2}\sqrt{1-b^2}, 0\}$ is a triangular norm.

Theorem 1. [25] Any kernel $k: U \times U \to [0,1]$ with $k(x,x) = 1$ is (at least) $T_{\cos}-$ transitive where $T_{\cos}(a,b) = \max(ab - \sqrt{1-a^2}\sqrt{1-b^2},\ 0)$.

Corollary 1. Fuzzy relation R_G computed with Gaussian kernel is a $T_{\cos}-$equivalence relation.

According to the definitions of $\underline{R_\theta}$ and $\overline{R_\sigma}$, we should obtain the residual implication of $T_{\cos}$ and its dual for computing lower and upper approximations of fuzzy sets related to R_G^n. We derive the residual implication of $T_{\cos}$ by the following lemma.

Lemma 1. $\theta_{T_{\cos}}(x,y)=\begin{cases}1, & a\le b\\ ab+\sqrt{(1-a^2)(1-b^2)}, & a>b\end{cases}$.

Proof. We have $\theta_{T_{\cos}}(a,b)=\sup\{\vartheta\in[0,1]:T_{\cos}(a,\vartheta)\le b\}$, so if $a\le b$, then $\theta_{T_{\cos}}(a,b)=1$.

Suppose $a>b$. We have ϑ should satisfies $a\vartheta-\sqrt{(1-a^2)(1-\vartheta^2)}\le b$ which implies $a\vartheta-b\le\sqrt{(1-a^2)(1-\vartheta^2)}$. Let $f_1(\vartheta)=a\vartheta-b$, $f_2(\vartheta)=\sqrt{(1-a^2)(1-\vartheta^2)}$, then $f_1(\vartheta)$ strictly increases on $[0,1]$, and $f_2(\vartheta)$ strictly decreases in interval $[0,1]$. If $\vartheta=ab+\sqrt{(1-a^2)(1-b^2)}$, then $f_1(\vartheta)=f_2(\vartheta)$. So if $\vartheta\le ab+\sqrt{(1-a^2)(1-b^2)}$, then $f_1(\vartheta)\le f_2(\vartheta)$; if $\vartheta>ab+\sqrt{(1-a^2)(1-b^2)}$, then $f_1(\vartheta)>f_2(\vartheta)$, this implies $\sup\{\vartheta\in[0,1]:T_{\cos}(a,\vartheta)\le b\}=ab+\sqrt{(1-a^2)(1-b^2)}$, which completes the proof.

In what follows, we use a compact notation by denoting $\theta_{T_{\cos}}$ by θ.

Definition 1. Given an information system $IS=(U,A)$ and $X\in F(U)$, where $F(U)$ is the power set of fuzzy sets, the fuzzy lower and upper approximations of X related to R_G^n are defined as

1) S-Gaussian fuzzy lower approximation operator: $\underline{R_{G_S}^n}X(x)=\inf_{y\in U}S(N(R_G^n(x,y)),X(y))$;

2) θ- Gaussian fuzzy lower approximation operator: $\underline{R_{G\theta}^n}X(x)=\inf_{y\in U}\theta(R_G^n(x,y),X(y))$;

3) T - Gaussian fuzzy upper approximation operator: $\overline{R_{GT}^n}X(x)=\sup_{y\in U}T(R_G^n(x,y),X(y))$;

4) σ - Gaussian fuzzy upper approximation operator: $\overline{R_{G\sigma}^n}X(x)=\sup_{y\in U}\sigma(N(R_G^n(x,y)),X(y))$.

In this work, we will focus on the definitions of $\underline{R_{G\theta}^n}X$ and $\overline{R_{GT}^n}X$. In a similar way, we can establish properties of the remaining operators. If no confusion occurs, we use a shorthand notation for $\underline{R_{G\theta}^n}$ and $\overline{R_{GT}^n}$ in the form $\underline{R_G^n}$ and $\overline{R_G^n}$, respectively.

Since R_G^n is a class of fuzzy *T*-equivalence relations, the properties of $\underline{R_G^n}$ and $\overline{R_G^n}$ are the same as those shown in [22]. Here we list some of them.

Theorem 2 [22] Given an information system $IS=(U,A)$ and $X\in F(U)$, $\underline{R_G^n}$ and $\overline{R_G^n}$ satisfy the following properties:

1) $\underline{R_G^n}X \subseteq X \subseteq \overline{R_G^n}X$, $\overline{R_G^n}X = X \Leftrightarrow \underline{R_G^n}X = X$;

2) $\underline{R_G^n}(\underline{R_G^n}X) = \underline{R_G^n}X$, $\overline{R_G^n}(\overline{R_G^n}X) = \overline{R_G^n}X$; $\overline{R_G^n}(\underline{R_G^n}X) = \underline{R_G^n}X$, $\underline{R_G^n}(\overline{R_G^n}X) = \overline{R_G^n}X$;

3) $\overline{R_G^n}(\bigcup_{t\in T} X_t) = \bigcup_{t\in T} \overline{R_G^n}X_t$, $\underline{R_G^n}(\bigcap_{t\in T} X_t) = \bigcap_{t\in T} \underline{R_G^n}X_t$;

4) If $R_G^m \subseteq R_G^n$, then $\underline{R_G^n}X \subseteq \underline{R_G^m}X \subseteq X \subseteq \overline{R_G^m}X \subseteq \overline{R_G^n}X$.

By 1) we know $\overline{R_G^n}X$ and $\underline{R_G^n}X$ are a pair of fuzzy sets approximating X as upper and lower bounds, respectively, and 4) indicates that a finer fuzzy relation can offer more precise approximations than the coarser one. These properties give a foundation for defining and developing attributes reduction in the following subsection.

3.2 Approximating decision regions with Gaussian Kernel

If the set of samples is assigned with a decision attribute D , we call the triple $< U, C, D >$ a decision system, where C is the set of condition attributes and D is the decision.

Each subset of C can be used to induce a fuzzy $T_{\cos}$ – equivalence relation over U by computing similarity with Gaussian kernel. We compute $R_G^{(j)}(x_i, x_k) = \exp(-\frac{\|x_{ij} - x_{kj}\|^2}{2\delta^2})$ as the similarity of samples x_i and x_k with respect to attribute j . Then the information hidden in C can be equivalently expressed as $R_G = \{R_G^{(1)}, R_G^{(2)}, \cdots, R_G^{(j)}, \cdots, R_G^{(n)},\}$, where n is the number of condition attributes.

After computing the relation making use of a single attribute, we require aggregating them for providing information for decision. If there are multiple fuzzy relations, the aggregation operator of fuzzy relations usually employs the $t-$norm treated as the min operator in the existing fuzzy rough sets. For example, given attribute a and b , the relations between samples x_i and x_k are $R^a(x_i, x_k)$ and $R^b(x_i, x_k)$, respectively. Then $R^{\{a\}\cup\{b\}}(x_i, x_k) = \min(R^a(x_i, x_k), R^b(x_i, x_k))$. In this work, we use the algebraic product, $T_P(x, y) = x \cdot y$, to carry out aggregation. This implies that the Gaussian kernel induced the individual fuzzy relations comes in the form

$$R_G^n(x_i, x_k) = \exp(-\frac{\|x_i - x_k\|^2}{2\delta^2}) = \prod_{s=1}^n R_G^{(s)}(x_i, x_k) .$$

In what follows, we denote the fuzzy relation induced by attribute subset $P \subseteq C$ by R_G^P .

Assume decision D divides the samples into subsets $\{d_1, d_2, \cdots, d_I\}$. Here we encounter the following relationship $\forall x \in U$, $d_i(x) = 1$ if $x \in d_i$; otherwise, $d_i(x) = 0$.

Now we approximate the decision regions with the fuzzy granules induced by Gaussian function. Take the i*th* class as an example,

1) $\underline{R_G^n} d_i(x) = \inf_{y\in U} \theta(R_G^n(x,y), d_i(y)) = \inf_{y\in U}\left(\begin{cases}1, & R_G^n(\mathrm{x,y}) \le d_i(y)\\ R_G^n(x,y)d_i(y)+\sqrt{1-R_G^{n\,2}(x,y)}\sqrt{1-d_i{}^2(y)}, & R_G^n(\mathrm{x,y}) > d_i(y)\end{cases}\right)$

If $d_i(y)=1$, i.e. $y\in d_i$, in this case $R_G^n(\mathrm{x,y}) \le d_i(y)$, then we get $\theta\left(R_G^n(x,y), d_i(y)\right)=1$;

if $d_i(y)=0$, i.e. $y\notin d_i$, in this case $R_G^n(\mathrm{x,y}) > d_i(y)$, $\theta\left(R_G^n(x,y), d_i(y)\right)=\sqrt{1-R_G^{n\,2}(x,y)}$.

Finally, we obtain $\underline{R_G^n} d_i(x) = \inf_{y\notin d_i}\left(\sqrt{1-R_G^{n\,2}(x,y)}\right)$.

2) $\overline{R_G} d_i(x) = \sup_{y\in U} T(R_G(x,y), d_i(y)) = \sup_{y\in U}\max\left(0,\ R_G(x,y)d_i(y)-\sqrt{1-R_G{}^2(x,y)}\sqrt{1-d_i{}^2(y)}\right)$

If $d_i(y)=1$, i.e. $y\in d_i$, we get $\max\left(0,\ R_G(x,y)d_i(y)-\sqrt{1-R_G{}^2(x,y)}\sqrt{1-d_i{}^2(y)}\right)=R_G(x,y)$;

if $d_i(y)=0$, $\max\left(0,\ R_G(x,y)d_i(y)-\sqrt{1-R_G{}^2(x,y)}\sqrt{1-d_i{}^2(y)}\right)=\max\left(0,\ -\sqrt{1-R_G{}^2(x,y)}\right)=0$.

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

The fuzzy lower and upper approximations of a decision in terms of a Gaussian kernel based fuzzy relation are computed as

1) $\underline{R_G^n} d_i(x) = \inf_{y\notin d_i}\left(\sqrt{1-R_G^{n\,2}(x,y)}\right)$;

2) $\overline{R_G} d_i(x) = \sup_{y\in d_i} R_G(x,y)$.

To be more specific, we have $\underline{R_G} d_i(x) = \inf_{y\notin d_i}\left(\sqrt{1-\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)^2}\right)$. If $x\in d_i$, we find a nearest neighbor of *x* from other classes to compute the lower approximation. However, if $x\notin d_i$, the nearest sample of *x* out of d_i is *x* itself. In this case $\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)=1$, so $\underline{k_\theta} d_i(x)=0$. As the upper approximation is concerned, if $x\in d_i$, $\overline{R_G} d_i(x)=\sup_{y\in d_i}R_G(x,y)$. Obviously, $\sup_{y\in d_i}k(x,y)=1$ as $k(x,x)=1$. If $x\notin d_i$, we find a nearest sample y of *x* in class d_i and $\overline{R_G} d_i(x)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$.

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 decisions, while the membership of *x* to the lower approximation of other decisions is zero. However, 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.

Definition 2. Given a decision table $< U, C, D >$, R_G is *T*-equivalence relation on *U* computed with Gaussian kernel in feature space $B \subseteq C$. *U* is divided into $\{d_1, d_2, \cdots, d_I\}$ with the decision attribute. The fuzzy positive regions of *D* in term of *B* are defined as

$$POS_B(D) = \bigcup_{i=1}^{I} \underline{R_G} d_i .$$

Positive region of *D* is a fuzzy set, the membership of a sample to the positive regions of decision reflects the degree of the sample necessarily belong to its decision class. The higher the membership is, the more certain the classification outcome is.

3.3 Approximating quality and reducts

Definition 3. Given a decision table $< U, C, D >$, R_G is *T*-equivalence relation on *U* computed with Gaussian kernel in feature space $B \subseteq C$. *U* is divided into $\{d_1, d_2, \cdots, d_I\}$ with the decision attribute. The fuzzy positive regions of *D* in term of *B* are given by as $\bigcup_{i=1}^{I} \underline{R_G} d_i$. The quality of approximating classification is defined as

$$\gamma_B(D) = \frac{| \bigcup_{i=1}^{I} \underline{R_G} d_i |}{|U|},$$

where $| \bigcup_{i=1}^{I} \underline{R_G} d_i | = \sum_i \sum_{x \in d_i} \underline{R_G} d_i(x)$.

The coefficient of approximating quality reflects the approximation abilities of the granulated space induced by attribute subset *B* to characterize the decision. This coefficient is also called the dependency between the decision and condition attributes. We say that decision *D* is dependent on B with degree $\gamma_B(D)$, denoting by $B \Rightarrow_\gamma D$. We say that the decision system is consistent if $\gamma_B(D) = 1$.

Theorem 3. Given a decision system $< U, C, D >$, $B_1 \subseteq B_2 \subseteq C$, R_1 and R_2 are two *T*-equivalence relations on *U* computed with Gaussian function $G(x, y)$ in B_1 and B_2, respectively. Then we have

1) $R_1 \supseteq R_2$;

2) $\underline{R_1} d_i \subseteq \underline{R_2} d_i$;

3) $\overline{R_1} d_i \supseteq \overline{R_2} d_i$;

4) $POS_{B_1}(D) \subseteq POS_{B_2}(D)$;

5) $\gamma_{B_1}(D) \le \gamma_{B_2}(D)$.

Proof. Properties 4 (2)-(4) can be derived from the monotonicity of the lower and upper approximations [22]. Here we just show the proof of the first property. Assuming 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, we obtain $\| x-y \|_{B_1}{}^2 = \sum_{i=1}^{N_1} \left(a_i(x)-a_i(y)\right)^2$, where $a_i(x)$ is the value of sample x in feature a_i. In feature space B_2, $\| x-y \|_{B_2}{}^2 \quad = \sum_{i=1}^{N_1} (a_i(x)-a_i(y))^2 + \sum_{i=N_1}^{N_{21}} (a_i(x)-a_i(y))^2$. So $\| x-y \|_{B_2}{}^2 \ge \| x-y \|_{B_1}{}^2$ and $R_1(x,y) \ge R_2(x,y)$. Then $R_1 \supseteq R_2$.

Theorem 4. Given a decision system $<U,C,D>$, R_G is *T*-equivalence relation on *U* computed with Gaussian function in $B \subseteq C$ and R_D is the equivalence relation induced by *D*. The decision system is consistent if and only if $R_G \subseteq R_D$, or for $\forall x, y \in U$, $R_G(x,y)=0$ if $x \in d_i$ and $y \notin d_i$.

Proof. Without loss of generality, we discuss the proof using two arbitrary samples. Assume the decision system is consistent. $\forall x, y \in U$, there are two cases, i.e.1) x and y belong to the same class; 2) x and y belong to different classes. As to case 1, $R_D(x,y)=1$, obviously, $R_G(x,y) \le R_D(x,y)$. As to case 2, $R_D(x,y)=0$. Since the system is consistent, we have $\gamma(D)=1$. We know $0 \le \underline{R_G} d_i(x) \le 1$. So $\forall x \in U$, $\underline{R_G} d_i(x) = 1$ if $x \in d_i$. Assume that $R_G(x,y)>0$, then we have $\underline{R_G^n} d_i(x) = \inf_{u \notin d_i} \left(\sqrt{1-R_G{}^2(u,y)} \right) \le \sqrt{1-R_G{}^2(x,y)} < 1$. This is in conflict with the fact that the system is consistent. Thus $R_G(x,y)=0$, and $R_G(x,y) \le R_D(x,y)$. Now we assume that $R_G \subseteq R_D$, then $\forall x, y \in U$, $R_G(x,y)=0$ if $x \in d_i$ and $y \notin d_i$. $\underline{R_G^n} d_i(x) = \inf_{u \notin d_i} \left(\sqrt{1-R_G{}^2(u,y)} \right) = \sqrt{1-R_G{}^2(x,y)} = 1$, so $\gamma(D)=1$. The system is consistent.

Theorem 4 shows that as the number of features increases, the approximation quality, the classification quality increases as well. These properties are consistent with our intuition that new features bring new information about granulation and classification. Correspondingly, the induced approximation space with more features becomes finer and can generate more precise approximations of decisions. As a result, the quality of approximating classification increases.

The quality of approximating classification, also called dependency between the decision and condition attributes, reflects the average degree of the fact that the samples certainly belong to their classes. Ideally, all the samples should be classified without error. Namely, for $\forall x \in d_i$, $\underline{R} d_i(x) = 1$. Accordingly, $\gamma_B(D)=1$. However, as there is some level of uncertainty in real-world decision systems caused by noise or insufficient features to distinguish all the objects, the dependency level between condition and decision is usually less than 1.

Like in the classical rough set model [8], we can define concepts such as redundancy, indispensability and reducts of decision systems based on fuzzy dependency.

Definition 4. Given $<U,C,D>$, $a \in B \subseteq C$. If $\gamma_{B_1}(D) = \gamma_{B-a}(D)$, we say *a* is redundant in *B* with respect to *D*; otherwise, we say *a* is indispensable in *B* to *D*.

Definition 5. Given $<U,C,D>$, $B \subseteq C$. We say B is a relative reduct to *D* if *B* satisfies the following conditions:

1) sufficient condition: $\gamma_B(D) = \gamma_C(D)$;

2) necessary condition: for $\forall a \in B$, *a* is indispensable in *B* to *D.*

Definition 5 shows a reduct is a subset of attributes which not only produces the same dependency as the whole attributes but also has no superfluous one. Obviously such attribute subsets are desirable in feature selection.

It is notable that given a data set there is usually more than one relative reduct. We can find a set of feature subsets { $B_1, B_2, \cdots, B_k$ } which all preserve the dependency of decision on condition attributes, which shows we can get multiple viewpoints to consider classification tasks. We name the intersection of all reducts $Core = \bigcap_i B_i$ as the core features of the decision table in discourse.

3.4 Connections between the proposed model and other models

In this section, we discuss the connections of fuzzy rough sets, rough set, neighborhood rough sets and the famous Relief algorithm [36, 37, 38 39].

The definition of the lower approximation in Gaussian kernel based fuzzy rough sets is a direct and intuitively appealing generalization of rough set and the neighborhood rough set [18]. As to rough sets themselves, only discrete variables can be analyzed. Given x, $y \in U$, we define a distance function in a discrete space:

$$\| x - y \| = \begin{cases} \infty, \ x \neq y \\ 0, x = y \end{cases}.$$

According to rough sets, if $\| x - y \| = 0$, then $y \in [x]$. If $y \in d_i$ and $x \notin d_i$. In this case, $\underline{R_G} d_i(x) = \inf_{u \notin d_i} \left(\sqrt{1 - \exp\left(-\frac{\| x-u \|^2}{2\delta^2} \right)^2} \right) = \sqrt{1 - \exp\left(-\frac{\| x-y \|^2}{2\delta^2} \right)^2} = 0$. In fact, in rough sets, we also know that

$x \notin \underline{R_G} d_i$. For the case where $x_i \in d_i$ if for $x_i \in U$, $\| x_i - x \| = 0$, we have

$$\underline{R_G} d_i(x) = \inf_{u \notin d_i}\left(\sqrt{1-\exp\left(-\frac{\| x-u \|^2}{2\delta^2}\right)^2}\right) = \sqrt{1-\exp\left(-\frac{\infty^2}{2\delta^2}\right)^2} = 1 .$$

Certainly, we also have $x \in \underline{R_G} d_i$. The above analysis shows Gaussian kernel based fuzzy rough sets can degrade to Pawlak rough sets.

Neighborhood rough sets realize an idea similar to the one captured by rough sets. Being different from rough sets, neighborhood rough sets use a general distance function, rather than discrete distance used in the previous construct [18]. In neighborhood rough sets, we consider that *x* belongs to the lower approximation of its class if the distance between *x* and its nearest sample with a different class is greater than δ, which is a certain threshold specified in advance. Here neighborhood rough sets extend rough sets by generalizing the distance function, while Gaussian kernel based fuzzy rough sets generalize neighborhood rough sets through extending the binary membership {0, 1} to a fuzzy membership function $\underline{R_G} d_i(x) = \inf_{u \notin d_i} \sqrt{1-(R_G^n(x,u))^2}$. Assuming that $y \in d_i$ if $\| x - y \| \le \delta$ for $\forall y \in U$, then $\underline{R_G} d_i(x) \ge \sqrt{1-\exp\left(-\frac{\delta^2}{2\delta^2}\right)^2}$. We introduce a cut operator and say that $\underline{R_G} d_i(x) = 1$ if $\underline{R_G} d_i(x) \ge \sqrt{1-\exp\left(-\frac{\delta^2}{2\delta^2}\right)^2}$; otherwise, $\underline{R_G} d_i(x) = 0$. Then fuzzy rough sets degenerate to neighborhood rough sets.

Furthermore, we have defined $\gamma_B(D) = \frac{|\bigcup_{i=1}^{I} \underline{R_G} d_i|}{|U|}$, where $|\bigcup_{i=1}^{I} \underline{R_G} d_i| = \sum_i \sum_{x \in d_i} \underline{R_G} d_i(x)$. We also get $\underline{R_G} d_i(x) = \inf_{u \notin d_i} \sqrt{1-(R_G(x,u))^2}$. As we know, $R_G(x,u)$ reflects the similarity degree, thus $\sqrt{1-(R_G(x,u))^2}$ can be considered as a general distance function. Then dependency $\gamma_B(D)$ is the sum of distances between each sample and its nearest sample with different classes.

It is an interesting conclusion stating that dependency of *D* to *B* is the sum of distances between each sample and its nearest sample with a different class in feature space *B*. Let's review the well-known feature evaluation algorithm called Relief [36, 37, 38 39]. In Relief, one finds x_i's nearest sample from the same class, called the nearest hit H_i, and the nearest sample from other classes, called the nearest miss M_i, then computes the distances $\| x_i - H_i \|$ and $\| x_i - M_i \|$ and afterwards uses $\sum_{i=1}^{m} \| x_i - M_i \| - \| x_i - H_i \|$ to evaluate the quality of a feature, where *m* is the number of the samples in training set or a subset of samples randomly drawn from the training set.

We see fuzzy rough sets and algorithm Relief share a common idea that the feature space where samples are far from other classes should produce a great weight. The greater the inter-class distance is, the greater the weight should be.

4. Uncertainty measures of fuzzy rough sets

Uncertainty measures in approximation space are important in rough approximation. They can be used to evaluate the quality of a set of condition attributes, and then be incorporated with a feature selection algorithm [26, 29, 30]. Moreover, these measures can also be used in inducing a fuzzy decision tree [31]. Duntsch and Gediga [32] systematically discussed the measurement of uncertainty in predicting based on rough sets. Qian et al. [33] pointed out that the measure of approximation accuracy cannot produce an elaborate characterization of the uncertainty of approximation and introduced three new measures. Here we will introduce and adapt the fuzzy entropy discussed in [23, 26, 34, 35] to compute the uncertainty present in the Gaussian kernel approximation.

Given a decision system $<U,C,D>$, the fuzzy $T_{\cos}-$equivalence relation matrix induced by $a_j \in C$ is denoted by

$$R_G^{(j)} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{pmatrix}.$$

The element of the relation $r_{ik} = R_G^{(j)}(x_i, x_k) = \exp(-\frac{\left\|x_{ij} - x_{kj}\right\|^2}{2\delta^2})$ quantifies the similarity degree between samples x_i and x_j when being considered in terms of attribute j. Then with each sample $x_i \in U$, we associate a fuzzy information granule of the following form $FIG(x_i) = \frac{r_{1i}}{x_1} + \frac{r_{2i}}{x_2} + \cdots + \frac{r_{mi}}{x_m}$. The family of fuzzy information granules, called fuzzy elemental concepts, form a fuzzy covering of the universe, denoted by $U/R = \{FIG(x), x \in U\}$. The fuzzy cardinality of $FIG(x_i)$ is computed in the form $|FIG(x_i)| = \sum_{l=1}^{m} r_{li}$.

Definition 6. Given a decision system $<U,C,D>$, R_G is a fuzzy relation induced with Gaussian kernel and attribute $B \subseteq C$. We call $<U,R_G>$ a fuzzy approximation space. The uncertainty of the approximation space is expressed in the form

$$H(B) = H(R_G) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|FIG(x_i)|}{|U|}.$$

It is easy to note that $\log|U| \geq H(B) \geq 0$ Furthermore $H(B)=0$ if and only if $\forall x, y \in U$, $R(x,y)=1$. $H(B)=0$ which means that each pair of samples is not distinguishable and the granularity of the system is the greatest and the system is the coarsest in this case. $H(B)=\log|U|$ if and only if $\forall x \neq y$, $R(x,y)=0$. In this case, all the samples are distinguishable and the fuzzy approximation space is the finest one.

$\frac{|FIG(x_i)|}{|U|}$ can be considered as the local probability estimated with samples around x_i. In this case, Gaussian functions are viewed as a window function. So $H(B)$ can also be understood as differential entropy in the viewpoint of probability estimation with window functions. If we interpret Gaussian functions as fuzzy neighborhoods of samples, then the measure is fuzzy information entropy.

Theorem 5. Given a decision system $<U,C,D>$, $a_i, a_j \in C$, $R_G^{(i)}$ and $R_G^{(j)}$ are two fuzzy similarity relation matrices induced by a_i and a_j with Gaussian kernel. We have $H(R_G^{(i)}) \geq H(R_G^{(j)})$ if $R_G^{(i)} \subseteq R_G^{(j)}$, where $R_G^{(i)} \subseteq R_G^{(j)}$ means that $\forall r_{kl}^{(i)}, r_{kl}^{(j)}:\ r_{kl}^{(i)} \leq r_{kl}^{(j)}$.

Proof. $FIG_i(x)$ and $FIG_j(x)$ stand for the fuzzy information granules generated with $R_G^{(i)}$ and $R_G^{(j)}$, respectively. If $R_G^{(i)} \subseteq R_G^{(j)}$, for $\forall x \in U$, we have $|FIG_i(x)| \leq |FIG_j(x)|$ because $\forall r_{kl}^{(i)}, r_{kl}^{(j)}:\ r_{kl}^{(i)} \leq r_{kl}^{(j)}$. Therefore, $\forall x \in U$, $-\frac{1}{m}\log\frac{|FIG_i(x)|}{m} \geq -\frac{1}{m}\log\frac{|FIG_j(x)|}{m}$. We arrive at the conclusion we have $-\frac{1}{m}\sum_{x\in U}\log\frac{|FIG_i(x)|}{m} \geq -\frac{1}{m}\sum_{x\in U}\log\frac{|FIG_j(x)|}{m}$ and $H(R_G^{(i)}) \geq H(R_G^{(j)})$.

Corollary 2. (Type-1 monotonicity, parameter monotonicity) Given a decision system $<U,C,D>$, $a \in C$, the similarity relation matrices $R_G^{(1)}$ and $R_G^{(2)}$ between samples are computed as $\exp(-\frac{\|x-y\|^2}{2\delta_1^2})$ and $\exp(-\frac{\|x-y\|^2}{2\delta_2^2})$ in terms of attribute *a*. $H(R_G^{(1)}) \geq H(R_G^{(2)})$ if $\delta_1 \leq \delta_2$ and $H(R_G)=0$ when $\delta \to \infty$.

Proof. Given arbitrary samples *x* and *y*, it is clear that $\exp(-\frac{\|x-y\|^2}{2\delta_1^2}) \leq \exp(-\frac{\|x-y\|^2}{2\delta_2^2})$ if $\delta_1 \leq \delta_2$. Therefore $R_G^{(1)} \subseteq R_G^{(2)}$ and $H(R_G^{(1)}) \geq H(R_G^{(2)})$.

Kernel parameter δ plays a role of controlling the granularity of approximation. The fuzzy information granules induced at a greater value of δ are greater than those induced with lower value δ. If $\delta \to \infty$, $\forall x, y \in U$, the similarity between them is intended to be 1. This means that all objects are indistinguishable in the context of infinitely high granularity. From this viewpoint, fuzzy entropy reflects the refinement or granularity of fuzzy sets induced by the corresponding kernel matrix.

Corollary 3. (Type-2 monotonicity, attribute monotonicity) Given a decision system $<U, C, D>$, $B, B' \in C$, $B \supseteq B'$, R_G and $R_G^{'}$ are two fuzzy similarity relation matrices induced by B and B' with Gaussian kernel. We have $H(R_G) \geq H(R_G^{'})$.

Proof. Assume that $B' \bigcup B_1 = B$. $R_G^{(1)}$ and $R_G^{(2)}$ are the kernel matrices induced by B' and B_1, respectively. Then the kernel matrix induced by B is computed in the form $R_G^{(1)} \otimes R_G^{(2)}$, where the element in $R_G^{(1)} \otimes R_G^{(2)}$ is $R_G^{(1)}(x_i, x_k) \cdot R_G^{(2)}(x_i, x_k)$. Since $R_G^{(1)}(x_i, x_k) \leq 1$ and $R_G^{(2)}(x_i, x_k) \leq 1$, thus we have $R_G^{(1)}(x_i, x_k) \cdot R_G^{(2)}(x_i, x_k)$ $\leq R_G^{(1)}(x_i, x_k)$. In the sequel $R_G \subseteq R_G^{'}$ and $H(R_G) \geq H(R_G^{'})$.

Corollary 3 states that the kernel matrix and the corresponding information granules induced by the relation matrix could be further refined once new features have been added.

Definition 7. Given a decision system $<U, C, D>$, $B_1, B_2 \subseteq C$, kernel matrices $R_G^{(1)}$ and $R_G^{(2)}$ are induced by B_1 and B_2. The joint entropy of attributes B_1 and B_2 is expressed as

$$H(B_1 \bigcup B_2) = H(R_G^{(1)} \otimes R_G^{(2)}).$$

It is worth noting that the definition of joint entropy is different with that presented in [23, 35]. In [23, 35], the operator of composition of fuzzy relations is realized using the “min”operation ; in our study we have confined to the algebraic product.

Given Corollary 3 and Definition 7, we have $H(B_1 \bigcup B_2) \leq H(B_1)$ and $H(B_1 \bigcup B_2) \leq H(B_2)$.

Definition 8. Given a decision system $<U, C, D>$, $B \subseteq C$, B generates a fuzzy similarity relation computed with Gaussian kernel, while D induces a Boolean equivalence relation on U. Then the conditional entropy of D to B is defined as

$$H(D \mid B) = H(B \bigcup D) - H(B).$$

Conditional entropy $H(D|B)$ reflects the uncertainty of D if B is given. In virtue of Corollary 3 and Definition 7, one can show $H(D|B) \geq 0$.

Theorem 6. Given $<U,C,D>$, $R_G^{(C)}$ and $R^{(D)}$ are induced by C and D. $H(D|C)=0$ if $<U,C,D>$ is consistent.

Proof. If $<U,C,D>$ is consistent, we have $R_G^{(C)} \subseteq R_G^{(D)}$. For $\forall x_i, x_j$, $r_{ij}^{(C)} \leq r_{ij}^{(D)}$. Assumed that $r_{ij}^{(D)} = 0$, we have $r_{ij}^{(C)} = 0$, so $r_{ij}^{(C)} \cdot r_{ij}^{(D)} = r_{ij}^{(C)} = 0$; otherwise, $r_{ij}^{(D)} = 1$, $r_{ij}^{(C)} \cdot r_{ij}^{(D)} = r_{ij}^{(C)}$. Therefore, if $R_G^{(C)} \subseteq R_G^{(D)}$, we have $R_G^{(C)} \otimes R_G^{(D)} = R_G^{(C)}$, $H(D|C) = H(R_G^{(C)} \otimes R_G^{(D)}) - H(R_G^{(C)}) = 0$.

Theorem 7. Given a decision system $<U,C,D>$, $B_1, B_2 \subseteq C$, kernel matrices $R_G^{(1)}$ and $R_G^{(2)}$ are induced by B_1 and B_2. If $R_G^{(1)} \subseteq R_G^{(2)}$, we have $H(D|B_1) \leq H(D|B_2)$.

Proof. It is straightforward.

Corollary 4. If $B_1 \supseteq B_2$, we have $H(D|B_1) \leq H(D|B_2)$.

Corollary 4 shows that addition of any new attribute will not lead to the increase of conditional entropy. Here we see again that new attributes introduce additional information supporting classification. The proposed entropy measures determine the uncertainty degree in relations and granulation induced by the relations. There are two factors influencing the granularity of the collection of fuzzy granules induced by attributes and Gaussian kernel. With the same attributes, a greater value of kernel parameter δ induced a coarser granulation. This conclusion is consistent with our previous observations. If δ takes greater values, the similarity degrees between any pair of samples become larger. In this case, an arbitrary sample can be difficult to distinguish from others. As a result, lower values of entropy are obtained. Furthermore, given the value of δ (viz. the level of granularity of analyzing the classification problem), the entropy gets larger when more attributes become available. The increase in the values of entropy can be used to evaluate the usefulness of attributes in the classification problem.

5. Attribute evaluation and reduction with Gaussian kernel rough sets

One of the most important applications of rough set theory is to evaluate the classification power of attributes in a decision system by computing the dependency between condition attributes and the resulting decision. The dependency function is used as a sort of heuristics in constructing efficient greedy attribute reduction algorithms [10, 11, 12]. In [29], Shen et al. generalized the function of dependency to the case of fuzzy sets and proposed a fuzzy dependency function.

In the generic model of rough sets, dependency is defined as $\gamma_B(D) = |POS_B(D)| / |U|$, where $POS_B(D) = \bigcup_{i=1}^{I} \underline{B} d_i$, $d_i \in U / D$. Dependency is the percentage of samples in the positive region, which is defined as the set of samples unquestionably belonging to one of the decision classes.

As to the Gaussian kernel based rough sets, it has been mentioned in Subsection 3.2 that for $\forall d_i, i = 1,2,..., I$, if $x \notin d_i$, $\underline{R_G^n} d_i(x) = 0$ and if $x \in d_i$, $\underline{R_G^n} d_i(x) \ = \inf_{y \notin d_t} \sqrt{1-(R_G^n(x,y))^2}$. This facts indicates that the value of $\underline{R_G^n} d_i(x)$ is determined by the minimal value of $\sqrt{1-(R_G^n(x,y))^2}$, $y \notin d_i$. A sample's membership belonging to its class's lower approximation depends on its nearest sample with distinct classes according to $\underline{R_G} d_i(x) = \inf_{y \notin d_t} \sqrt{1-(R_G^n(x,y))^2}$.

Given $x \in d_i$, $\underline{R_G} d_i(x)$, the membership of sample x to the fuzzy lower approximation of its class d_i reflects the degree at which x certainly belongs to its decision, while $\overline{R_G} d_i(x)$ is the degree at which x possibly belongs to its decision. In feature selection, we naturally wish that we can find a feature subspace $B \subseteq C$ where each sample belongs to its decision with the greatest certainty; meanwhile, there is not a redundant attribute in *B*. The total certainty of samples belonging to its decision can be measured with the fuzzy dependency $\gamma_B(D)$. Formally, the computation of fuzzy dependency is described as follows.

Algorithm 1. Dependency with Gaussian kernel approximation (**DGKA**)

Input: sample set $U = \{x_1, x_2, \cdots, x_m\}$, feature set B, decision *D* and parameter δ

Output: dependency γ of *D* to *B*

```
$\gamma_B(D) \leftarrow 0$
    for $i$=1 to $m$
        find the nearest sample $M_i$ of $x_i$ with a different class
        $\gamma_B(D) \leftarrow \gamma_B(D) + \sqrt{1-\left[\exp\left(-\frac{\| x_i - M_i \|^2}{\delta}\right)\right]^2}$
    end
output $\gamma_B(D)$.
```

This algorithm is easy to implement and its time complexity is the same as of Relief. To evaluate *n* features with *m* samples, the time complexity is $O(nm\log m)$ [37]. Similar to Relief, algorithm 1 can just be used to evaluate the significance of features and rank them. Irrelevant features will receive low dependency values and could be removed from the data. However, it was pointed out that features ranking can not remove the redundant features because two features producing great dependency values may be redundant. Redundant features exist in a lot of databases [40, 41]. Attribute reduction need eliminate not only the irrelevant, but also the redundant variables from the data.

It is impractical to find the optimal subset of features from $2^n - 1$ candidates through exhaustive search, where n is the number of features. Greedy search guided by some heuristics is usually more efficient than the plain brute-force exhaustive search. In a forward greedy search, one starts with an empty set of attributes, and keeps adding features to the subset of selected attributes one by one. Each selected attribute maximizes the increment of dependence of the current subset; this implies the relevant but redundant attributes will not be included because it can not bring much new information about classification if the attribute is redundant. Formally, a forward search algorithm for feature selection based on Gaussian kernel approximation is written as follows.

Algorithm 2: Feature selection based on Gaussian kernel approximations (FS-GKA)

```
Input：  sample set $U = \{x_1, x_2, \cdots, x_m\}$, feature set $C$, decision $D$ and stopping threshold $\varepsilon$
Output:  reduct $red$
$red \leftarrow \varnothing$ ,  $\gamma \leftarrow 0$ ;
while $red \neq C$
    for each $a_i \in (C - red)$
        compute $\gamma_i = \gamma_{\{a_i\} \cup red}$
    end
    find the maximal $\gamma_i$ and the corresponding attribute $a_i$ ;
    if $\gamma_i - \gamma_{red}(D) > \varepsilon$
        $red \leftarrow red \bigcup a_i$ , $\gamma_{red} \leftarrow \gamma_i$ ;
    else
        exist while；
    end if
end while
return $red$
```

The time complexity of algorithm 2 is $O(n^2 m \log m)$, where n and m are the numbers of features and samples, respectively.

Besides dependency, conditional information entropy introduced in Section 4 can also be used to evaluate features. As we explain in Definition 8 and Theorem 8 that conditional entropy $H(D \mid B)$ is the uncertainty of D if condition attributes B are given, conditional entropy reflects the relevance between condition attributes and decision. We thus define the significance of attribute subset B in the following form

$$SIG(B, D) = H(D) - H(D \mid B) = H(D) + H(B) - H(BD) .$$

It is easy to observe that $SIG(B, D)$ becomes a symmetric uncertainty measure. In fact this is mutual information of B and D defined in Shannon's information theory if B and D generate Boolean equivalence relations

[23]. As it is well-known, mutual information is widely applied in evaluating features and constructing decision trees [40, 42, 43], but the classical definition of mutual information can just be used to deal with discrete features. But $SIG(B, D)$ defined here can be used to deal with numerical and fuzzy information. If we substitute mutual information for dependency in algorithm 2, a new feature selection algorithm based on fuzzy mutual information is derived.

Besides, it is worth noting that the proposed measures of dependency and mutual information can be incorporated with other search strategies used in other feature selection algorithms, such as ABB (Automatic Branch and Bound), SetCover, probabilistic search [44], and GP (Genetic programming) [45]. In this study, we are not going to compare and discuss the influence of search strategies on the results of feature selection. Here we focus on the comparison of the proposed method when dealing with different evaluation measures.

6. Experimental Studies

There are two objectives when carrying out a series of numerical experiments. First, when using Gaussian kernel to compute similarity relations between samples, we specify the parameter δ in Gaussian kernel. This parameter controls the granularity of the granulation space induced by the Gaussian functions. Considering its functionality, this parameter exhibits a significant impact on the effectiveness associated with the corresponding fuzzy rough sets. However, just like in Gaussian kernel support vector machines [46], no theoretical results have been obtained for specifying kernel parameters. The optimal value of the kernel parameter is dependent on the nature of the specific application. In this section, we report on a suite of experiments which helped us determine a range of “optimal” values of the kernel parameter. Second, as the main application of rough sets and fuzzy rough sets comes with attribute evaluation and reduction, we offer a comprehensive experimental evidence with this regard.

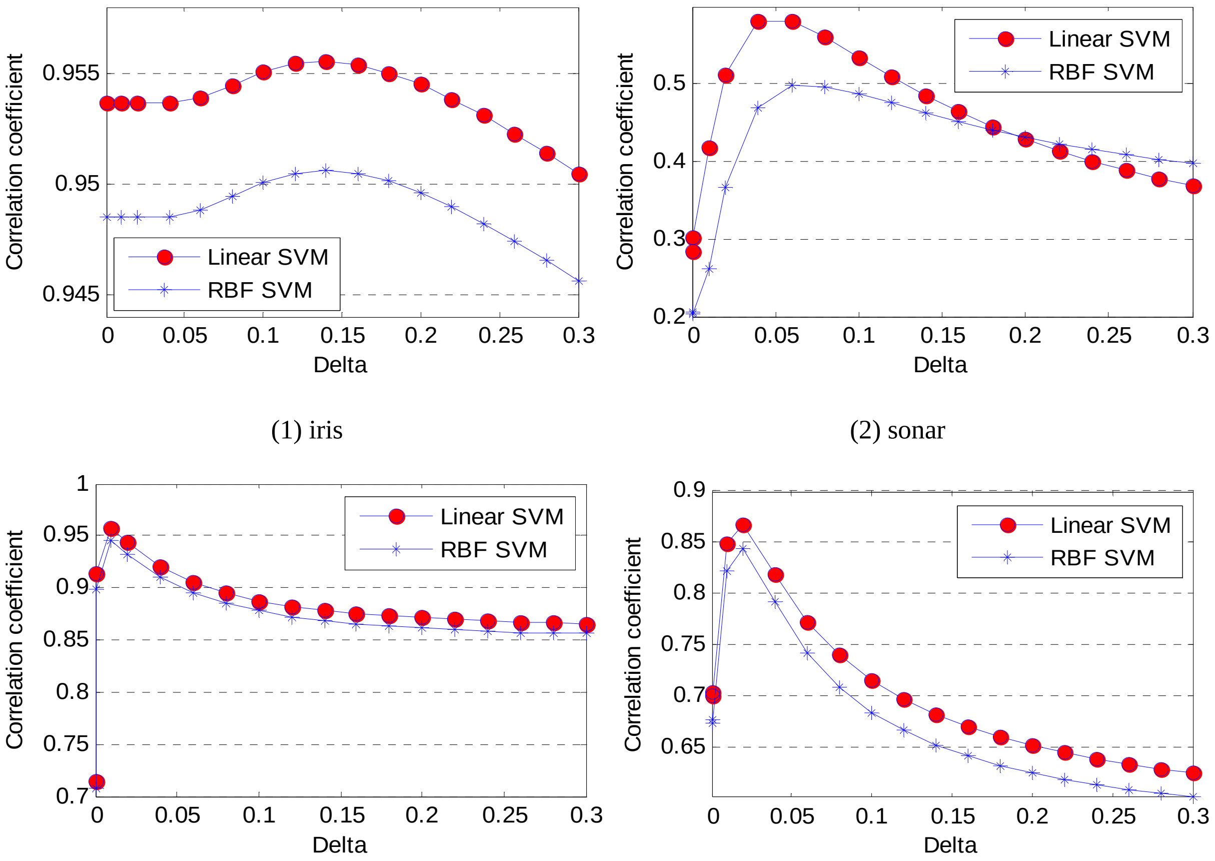
Some datasets used here come from the UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/); refer to Table 1. Numerical attributes are linearly normalized as follows $(x - x_{\min})/(x_{\max} - x_{\min})$ (with $x_{\min}$ and $x_{\max}$ being the bounds of the given attribute) before reduction and classification. In experiments, learning algorithms such as CART, linear SVM and RBF SVM are used. The experiments were run in a 10-fold cross validation mode. The parameters of the linear SVM and RBF SVM are taken as the default values (the use of the Matlab toolkit osu_svm3.00).

Table 1 Data description

ID	Data	samples	features	class
1	credit	690	15	2
2	heart	270	13	2
3	hepatitis	155	19	2
4	horse	368	22	2
5	iono	351	34	2
6	sonar	208	60	2
7	wdbc	569	31	2
8	wpbc	198	33	2
9	wine	178	13	3
10	iris	150	4	3

In computing the membership grades of samples belonging to the low approximation of decision with Gaussian kernel, one should specify kernel parameter δ. We experiment with a number of values of δ over different datasets, and compute the dependency of decision to each single feature. At the same time, we compute classification accuracies obtained for single features with linear SVM and RBF SVM. Finally, we determine the correlation coefficients between classification accuracies and dependencies. High values of correlation coefficient are reflective of the associated classification capabilities of the corresponding features. So the value domain of δ generating a great correlation coefficient is used in computing similarity. The underlying reason is that we hope this dependency becomes a sound estimate of classification abilities of the respective attributes. The values of δ were taken from the set {0.001, 0.005, 0.01, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.20, 0.22, 0.24, 0.26, 0.28, 0.30} when dealing with iris, sonar, wdbc and wine datasets, respectively.

The obtained values of the correlation coefficients vs. kernel parameters are presented in Figure 1. There is a uniform trend of variation of correlation coefficients vs. parameters, namely, the correlation goes up firstly achieves some peak, and afterwards decreases. This behaviour points at the optimal interval for the valurs of the kernel parameter. In what follows, we consider the range $\delta = [0.04,\ 0.06]$ in the evaluation of a single feature. We specify $\delta = [0.1,\ 0.12]$ when selecting features making use of algorithm 2. We learn that 0.04-0.06 is a sound interval for evaluating single features (refer to figure 1), however, we have found that the algorithm converges too early to find enough features in experiments if δ takes value in interval [0.04, 0.06]. Given this we have extended the range of the valurs and considered δ to be in the range of 0.1 and 0.15, $\delta = [0.1,\ 0.15]$.



(1) iris

(2) sonar

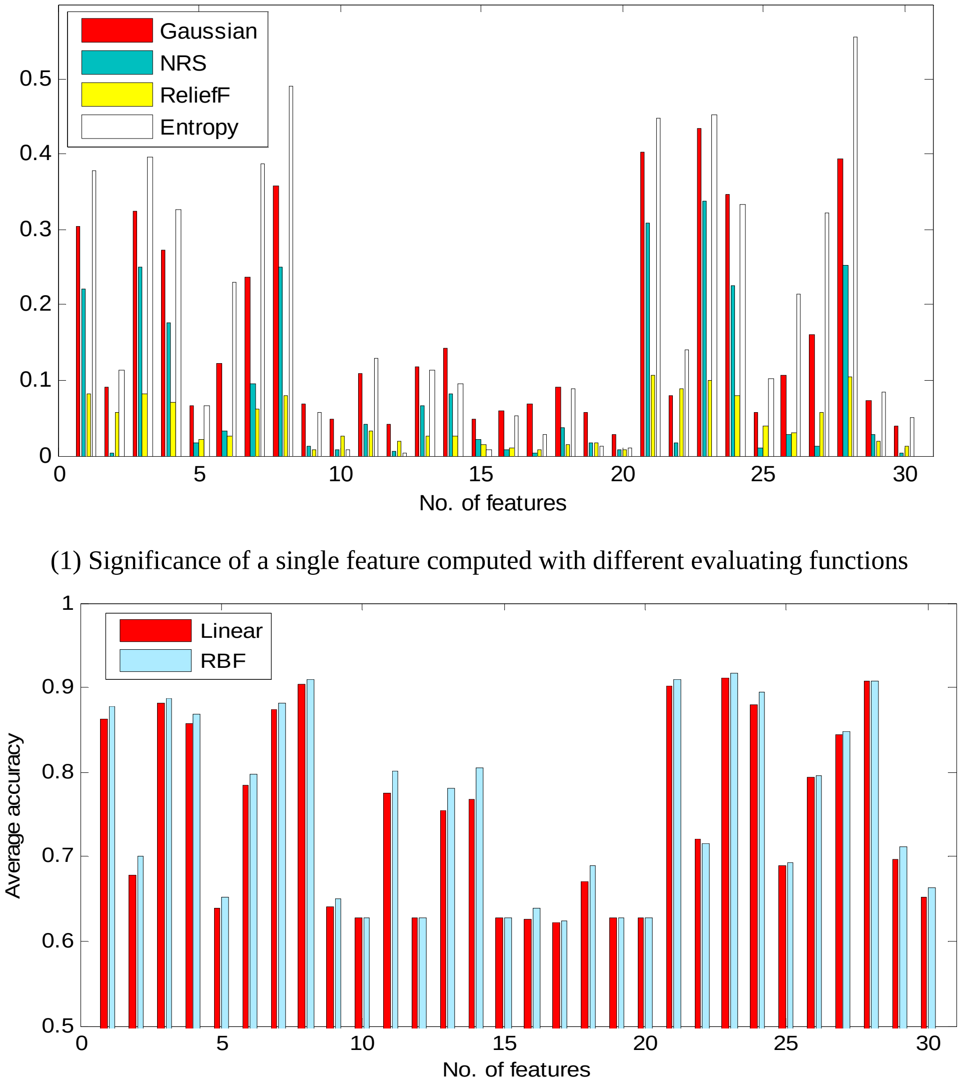
(3) wdbc

(4) wine

Fig. 1. Variation of correlation coefficients vs. kernel parameter.

Now we compare the effectiveness of fuzzy rough sets in evaluating feature quality. Sometimes one needs to compute the dependency of decision *D* for a single feature and find the relevance between input and output. One may anticipate that the evaluating function can reflect the classification performance in feature selection and feature ranking. We compute the significance of single features with four evaluation functions: dependency in Gaussian kernel approximation (**Gaussian**); fuzzy entropy in Gaussian kernel approximation (**Entropy**); dependency in neighborhood rough sets (**NRS**) [18] and **ReliefF** [37]. At the same time, we reported the classification accuracies of the corresponding features based on the use of the linear SVM and RBF SVM.

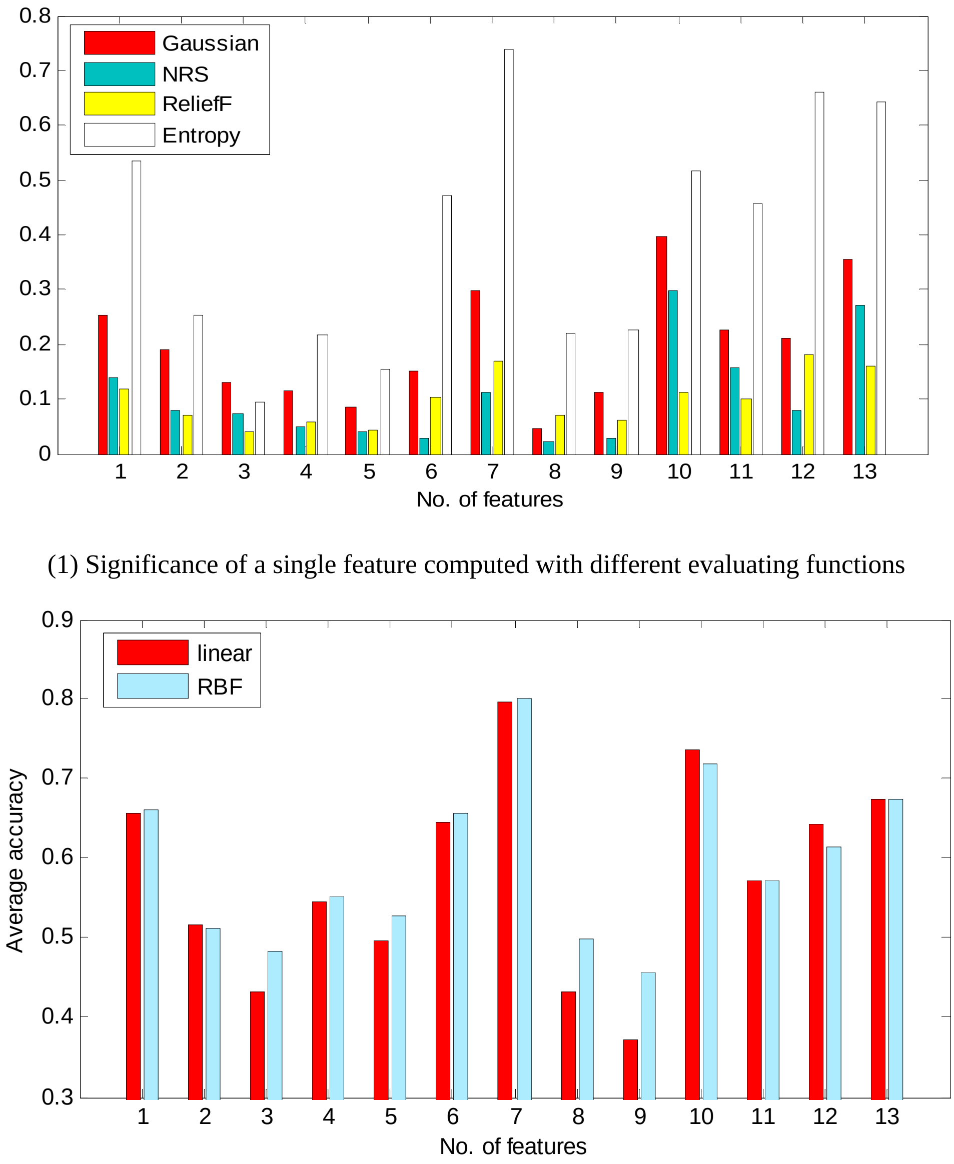
Two data sets wdbc and wine are used in experiments. There are 30 numerical features in wdbc and 13 numerical features in the wine dataset. The results are given in Figures 3 and 4, respectively. As to the wdbc data, features 1, 3, 4, 7, 8, 21, 23, 24, 28 produce higher values of all evaluating functions, as shown in figure 2 (1); at the same time, we can also find that these features produce higher values of classification accuracy (again shown in figure 2 (2)). As to the wine data, features 1, 6, 7, 10, 11, 12, 13 are better than others in terms of the four evaluating functions, corresponding the classification accuracies of features 1, 6, 7, 10, 11, 12, 13 are also higher than for other features. These results show that all the four evaluating functions can produce good estimates of classification ability of the features. There exist some differences between the evaluating functions in the two experimental results. We can find that the ordering of the feature is different if we rank the features considering individual evaluating functions. For example, for the wine data, the descending order of features induced by the fuzzy information entropy is 7, 12, 13, 1, 10, 6, 11, 2, 4, 9, 8, 5, 3; while the order induced by the ReliefF evaluation function is 7, 12, 13, 6, 10, 1, 9, 11, 8, 5, 4, 3, 2. Features 7, 12 and 13 are the three best features with respect to entropy and ReliefF. However, feature 1 ranks the fourth with respect to entropy, while it ranks the sixth as to ReliefF. Greedy search algorithms are sensitive to this little difference. Finally the little difference may leads to completely different feature subsets in Greedy algorithms.



(1) Significance of a single feature computed with different evaluating functions

(2) Classification accuracies obtained for single features when using linear SVM and RBF SVM

Fig.2 Significance and accuracy of single feature (**wdbc**)

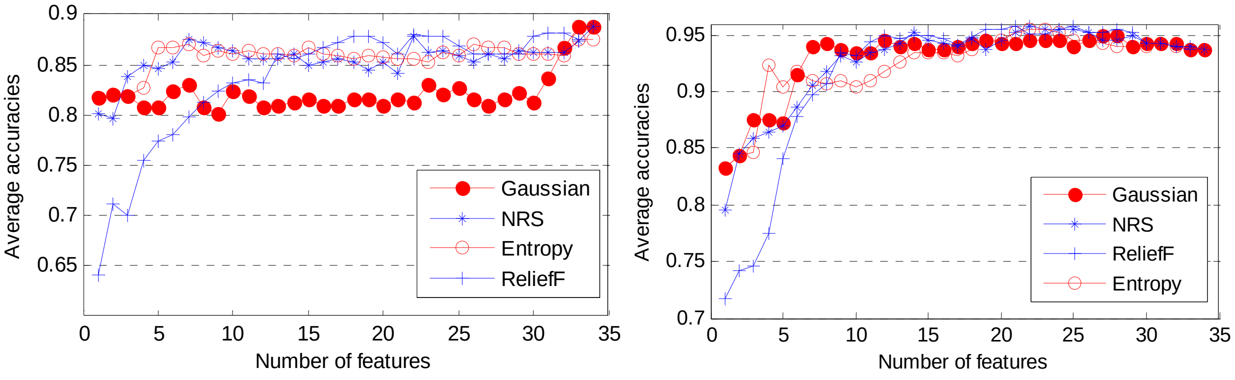


(1) Significance of a single feature computed with different evaluating functions

(2) Classification accuracies of single feature computed with linear SVM and RBF SVM

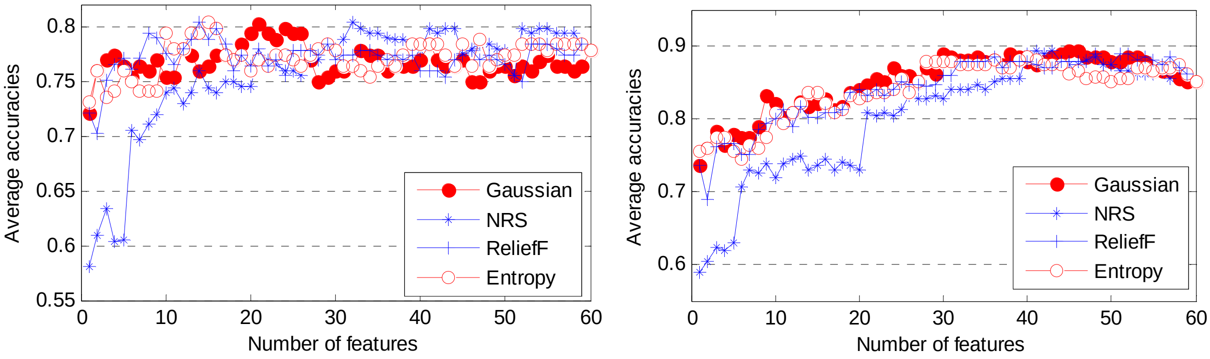
Fig.3 Significance and accuracy of single features (**wine**)

In ranking based feature selection, the first "k" best features are selected, where k is specified based on available domain knowledge. One can also add the best features one by one, and determine the classification performance of the current features in each round until the classification performance does not improve significantly when adding more features. Here we compare the four evaluation measures when working with the second strategy. Datasets of iono, sonar, wdbc and wine are used in experiments. We employ linear SVM and RBF SVM to validate the selected features. Figure 4-Figure 7 present the variation of classification performance over the number of selected features. The results show that classification accuracy increases with the number of selected features. The improvement is significant at the beginning of the selection process. Afterwards, the classification accuracy does not improve significantly once a certain number of features have been selected. Considering the cost of classification, we can delete the features which do not exhibit any significant influence on the quality of classification. Still we can find that fuzzy entropy and dependency in Gaussian kernel approximation are competent with neighborhood rough sets and ReliefF. Entropy and dependency sometimes are better than the other two algorithms.



(1)Linear-SVM (2)RBF-SVM

Fig. 4. Variation of accuracies vs. number of selected features (**iono**)



(1) Linear SVM (2) RBF SVM

Fig. 5. Variation of accuracies vs. number of selected features (**sonar**)

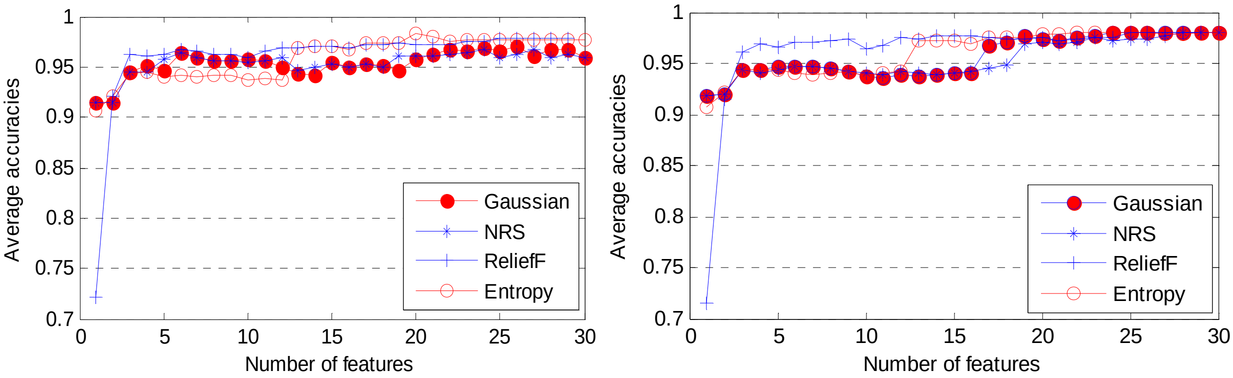


Fig. 6. Variation of accuracies vs. number of selected features (**wdbc**)

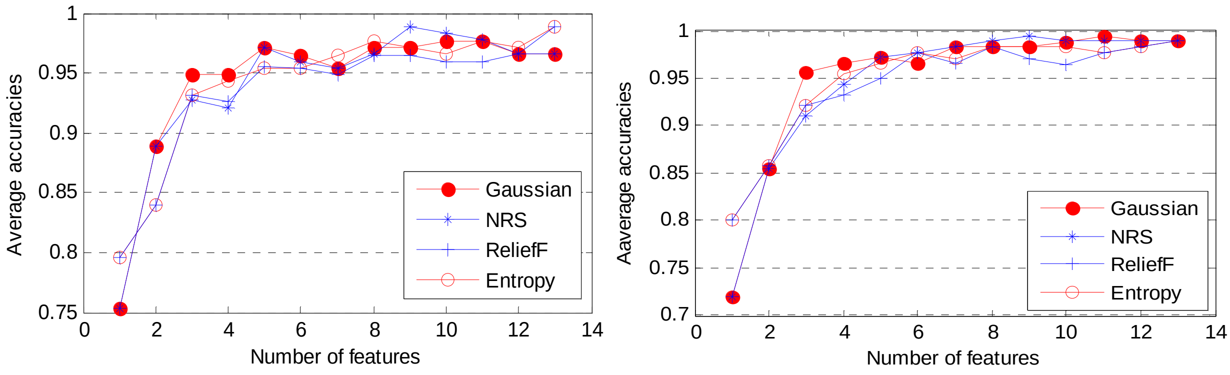


Fig. 7. Variation of accuracies vs. number of selected features (**wine**)

The above results show the proposed fuzzy rough sets and fuzzy entropy can be used to evaluate single attributes. Now we show the effectiveness in attribute reduction. As mentioned above that feature ranking can not delete the redundant information from data, while ReliefF was design to compute the weights of features and ranking them with the weights. We here compare Gaussian kernel based fuzzy rough sets and fuzzy entropy with Pawlak rough sets [8], where the entropy based discretization algorithm is introduced for transform the numerical features into discrete ones [63], neighborhood rough sets [18], Triangle similarity based fuzzy rough sets (shortly triangle) [30] and correlation based algorithm [43] in feature selection or attribute reduction.

The selected features with different algorithms are presented in Tables 2, 3 and 4, respectively. Regarding Gaussian kernel approximation, entropy, rough set and neighborhood rough sets, the orders of the features presented in the tables are the orders that the features are kept being added to the feature space. These orders reflect the relative significance of features in terms of the corresponding measures.

Table 2. Subsets of features selected with Gaussian kernel approximation and fuzzy entropy

Data	Gaussian kernel approximation	fuzzy entropy
credit	2,6,3,9,14	9,10
heart	8,1,4,10,3,12,13,7,2,5	13,12,3,11
hepatitis	18,14,15,1,11,17,9	18,17,15,11
horse	15,5,17,20	17,20,8,10,13,6
iono	3,31,24,16,5,9,34	5,6,8,25,28,24,34,7,3
sonar	44,11,27,21	11,17,37,48,27,22,29,12,33,36
wdbc	23,28,22,12,25,19,10,9,7,2,26,21,8,29	28,21,22,25,29,2,8,10,12
wpbc	1,12,7,23,32,22,6	13,32,33,24,6,23,20,21,26,12,1,2,28,10
wine	13,10,7,1,5,2	7,1,10,13,5,2

Table 3. Subsets of features selected with Pawlak rough sets and neighborhood rough sets

Data	**Pawlak rough set**	**NRS**
credit	4,7,9,15,1,3,11,6,14,8,2	15,8,6,9,2,3
heart	--	10,12,13,3,1,4,5,8,7
hepatitis	2,18,8,10,4,5,17,19,13,15,3,12	2,17,1,18,14,15,11
horse	15,3	5,20,17,10,8,13,1,11
iono	5,3,6,34,17,14,22,4	1,5,19,32,24,20,7,8,3
sonar	--	1,45,39,36,28,21,7
wdbc	24,8,22,26,13,5,14	23,28,2,29,5,16,25,9,22,10,12,11
wpbc	23,29,24,1,8,6,20,11	1,19,6,23,24,30,13
wine	10,13,7,2	13,10,7,5,11,1

Table 4. Subsets of features selected with triangle similarity based fuzzy rough sets and CFS

Data	**Triangle**	**CFS**
credit	5,7, 6, 9,10,12,13, 4	5,6,8,9,11,14,15
heart	10, 8, 1, 3,13,12, 7,11	3,7,8,9,10,12,13
hepatitis	2	1,2,6,11,14,17,18
horse	5, 4,18,19	1,3,5,7,15,17,20,21
iono	1	1,3,4,5,6,7,8,14,18,21,27,28,29,34
sonar	44,35,20,29,25,54,12	4,5,9,10,11,12,13,21,28,36,44,45,46,47,48,49,51,52,54
wdbc	23,28,22,12	2,7,8,14,19,21,23,24,25,27,28
wpbc	1, 7,12,23	1,33
wine	13,10, 7, 1, 5	1,2,3,4,5,6,7,10,11,12,13

Table 5. Number of features

Data	**Raw data**	**Gaussian**	**entropy**	**RS**	**NRS**	**Triangle**	**CFS**
credit	15	5	2	11	6	8	7
heart	13	10	4	0	9	8	7
hepatitis	19	7	4	12	7	1	7
horse	22	4	6	2	8	4	8
iono	34	9	9	8	9	1	14
sonar	60	4	10	0	7	7	19
wdbc	31	14	9	7	12	4	11
wpbc	33	7	14	8	7	4	2
wine	13	6	6	5	6	5	11

Some interesting results can be derived from the selected attributes. First, whatever attribute selection techniques have been used, most of the attributes in all datasets can be deleted. The reduction rate is high to 90% for some datasets, such as sonar and wpbc. Second, different algorithms produce distinct subsets of attributes. It is interesting that no two algorithms get the same subset of features for any database in the experiments except Gaussian fuzzy rough sets and fuzzy entropy for data wine. Even though, the orders of the selected features are different for this database. The best single feature is 13 in terms of Gaussian fuzzy rough sets, while feature 7 is the best one with respect to fuzzy entropy. This difference comes from the definitions of feature significance. The feature which is the best with respect to fuzzy information entropy is not necessarily good in terms of fuzzy dependency. The difference in the feature subsets also shows there are multiple subsets of features which have good classification power for a given classification task. Third, it is remarkable that Pawlak rough sets do not obtain any feature for data heart and sonar. As to forward greedy search algorithms, the algorithm will stop at the first round and output nothing if the significance of any single feature is zero. Sometimes no classification sample is consistent with respect to a single feature, thus the dependency defined in Pawlak rough sets is zero. This problem usually occurs in practice when conducting attribute reduction with Pawlak rough sets.

The great difference between these selected features may result from two factors. One is the difference between the qualities of features computed with different evaluation functions. As we know, we consider the ranking of features in feature selection, sometimes, a little difference in feature qualities may lead to completely different ranking. The other is the search strategy we used in these algorithms. We use greedy search procedure to find optimal features in terms of these evaluation functions. However, we know greedy search usually cannot get the optimal solutions to tasks. Furthermore, we may get completely different solutions if the first features selected with different algorithms are different. Although the selected features are different, they may all be effective for classification learning.

Another question is whether these selected features are effective for classification learning. Although we evaluate the features with different functions and the selected features get high scores in terms of these functions, the classification performance of the selected features have to be tested. We build classification models with the selected features and test their classification performance based on 10-fold cross validation. The average value and standard deviation are used to measure the classification performance.

We compare the raw data, Gaussian kernel based fuzzy rough sets, fuzzy information, Pawlak rough sets and neighborhood rough sets , triangle similarity based fuzzy rough sets and CFS in Tables 6, 7 and 8, where learning algorithms CART, linear SVM and RBF SVM are introduced to evaluate the selected features.

Table 6 Classification accuracies based on CART (%)

Data	Raw data	Gaussian	entropy	RS	NRS	Triangle	CFS
credit	82.73±14.86	82.28±14.79	85.48±18.5	82.88±14.34	82.28±14.79	83.90±16.90	80.12±14.08
heart	74.07±6.30	75.93±6.36	82.59±5.53	--	75.93±7.66	75.93±7.86	77.04±6.94
hepatitis	91.00±5.45	90.33±3.31	91.00±3.16	91.00±4.46	90.33±4.57	79.50±1.58	93.00±7.11
horse	95.92±2.30	96.47±1.30	89.92±4.53	93.49±5.12	88.87±5.57	71.15±6.83	95.93±1.90
iono	87.55±6.93	96.00±5.19	89.87±7.48	93.18±3.61	90.06±5.19	74.99±8.66	88.66±7.10
sonar	72.07±13.94	69.17±6.49	71.60±8.38	--	69.67±13.23	70.19±11.41	70.69±14.09
wdbc	90.50±4.55	91.93±4.31	91.58±3.62	94.20±3.43	94.02±4.19	94.20±16.6	92.79±4.81
wpbc	70.63±7.54	67.00±12.36	72.24±6.25	70.47±13.65	70.71±8.41	69.63±3.60	72.66±10.62
wine	89.86±6.35	92.08±4.81	91.53±4.83	92.08±4.81	91.53±6.09	92.08±4.81	89.86±6.35

Table 7 Classification accuracies based on linear SVM (%)

Data	Raw data	Gaussian	entropy	RS	NRS	Triangle	CFS
credit	85.48±18.5	85.48±18.5	85.48±18.51	85.48±18.51	85.48±18.5	85.48±18.51	85.48±18.51
heart	83.33±5.31	82.60±8.20	83.33±6.36	--	83.33±6.59	82.59±5.53	84.81±5.91
hepatitis	86.17±7.70	88.83±5.67	88.83±5.67	85.00±7.24	90.33±6.37	79.50±1.58	90.17±6.59
horse	92.96±4.43	89.68±4.78	90.22±4.13	63.04±1.26	90.49±4.98	63.04±1.26	91.03±4.96
iono	87.57±6.45	88.3191 %	85.26±6.10	83.30±5.97	87.26±6.06	74.99±8.66	86.38±5.35
sonar	77.86±7.05	76.41±8.54	77.90±7.13	--	70.21±7.68	71.19±7.76	78.38±5.58
wdbc	97.73±2.43	97.55±2.05	97.02±2.03	95.09±2.83	96.67±2.39	95.96±2.02	96.32±1.92
wpbc	77.37±7.73	76.32±3.04	76.84±4.61	76.32±3.04	76.32±3.04	76.32±3.04	76.32±3.04
wine	98.89±2.34	98.33±2.68	98.33±2.68	95.00±4.10	97.78±3.88	96.67±3.88	98.89±2.34

Table 8 Classification accuracies based on RBF SVM (%)

Data	Raw data	Gaussian	entropy	RS	NRS	Triangle	CFS
credit	81.44±7.18	**85.63±18.5**	85.48±18.51	81.00±16.25	**85.63±18.48**	82.88±9.73	85.05±17.79
heart	81.11±7.50	**85.93±6.25**	85.56±6.16	--	80.74±4.88	78.89±6.06	80.74±6.72
hepatitis	83.50±5.35	**90.83±6.54**	88.67±7.06	84.17±8.21	90.83±7.25	90.33±5.54	89.67±5.54
horse	72.30±3.63	**91.82±3.63**	**91.82±3.93**	63.04±1.26	88.86±2.99	82.59±5.40	91.59±5.13
iono	93.79±5.08	93.50±4.59	94.88±4.47	91.54±5.53	93.76±5.00	92.62±374	**95.19±4.43**
sonar	85.10±9.49	79.76±8.30	**83.71±8.10**	--	79.33±6.33	82.29±7.03	79.81±6.01
wdbc	98.08±2.25	**97.73±2.03**	**97.37±2.37**	95.61±2.37	96.67±2.09	96.49±2.61	96.84±1.80
wpbc	80.37±5.33	77.34±4.66	**80.37±5.83**	77.37±5.14	78.37±5.06	78.87±4.94	76.32±3.04
wine	98.89±2.34	98.33±2.68	98.33±2.68	97.22±2.93	**98.89±2.34**	97.15±3.99	**98.89±2.34**

Comparing the performance of raw data and fuzzy rough set based reducts, we can find although most of features have been removed in the reduct, most of the classification accuracies derived from the reduced data sets do not decrease, but increase. It shows there are redundant and irrelevant attributes in the raw data.

Comparing fuzzy rough sets, fuzzy entropy with rough sets, no matter which classification algorithms are used, fuzzy rough sets and fuzzy entropy are almost consistently better than Pawlak rough sets. Pawlak Rough sets finds nothing for data sets heart and sonar, however, both fuzzy rough sets and fuzzy entropy output subsets of features of moderate size. At the same time, in the data sets of horse, iono, wdbc, wine, etc, fuzzy rough sets or fuzzy entropy are much better than rough sets.

As a whole, neighborhood rough sets outperform Pawlak rough sets with respect to linear SVM and RBF SVM, however, are worse than fuzzy rough sets or fuzzy entropy. As to CART and linear SVM learning algorithms, fuzzy rough sets or fuzzy entropy are better than or equivalent to neighborhood rough sets for eight of the nine databases, while as to RBF SVM, fuzzy rough sets or fuzzy entropy are better than neighborhood rough sets for all the databases.

Triangle functions are used to compute the fuzzy similarity between samples in [30]. Based on this function, Jensen and Shen proposed a number of measures to compute the importance of attributes without discretization. From Tables 5, 6, 7 and 8, we can see that their algorithm return two few attributes to keep the classification performance. The yielded features produce worse performance than the original data sets and other subsets. The reduction of performance results from the computation of similarity, which leads to early stopping of the algorithms. The features derived by Gaussian based fuzzy rough sets and fuzzy entropy get the higher classification accuracies in most of the datasets. Especially, for the linear SVM and RBF SVM, the proposed algorithm performs much better than the triangle similarity based technique.

7. Conclusion

Kernel methods and rough sets are two classes of commonly encountered learning methodologies in machine learning and pattern recognition. They have different application domains and it seems that there are no tangible links between these two methodologies. We stressed that there are some commonalities as these two approaches rely on the same format of representation of samples and relationships between them: exhibiting the same format of data, that is kernel matrices used in kernel methods and relation matrices considered in rough sets.

Here we incorporate Gaussian kernel with fuzzy rough sets and construct a Gaussian kernel approximation based fuzzy rough set model. In this model, we introduce Gaussian function to compute the similarities between samples and generate fuzzy information granules for each sample. Afterwards, these fuzzy granules are used to approximate the decision classes. Besides we introduce fuzzy entropy to measure the uncertainty in kernel approximation. Some theorems about granularity, approximation quality, kernel parameter and features have been provided. Based on the dependency and mutual information defined in Gaussian kernel approximation, we proposed two feature evaluation indexes and selection algorithms. When compared with rough sets, neighborhood rough sets, Relief and CFS, we showed that the proposed methods come with a better performance.

It is interesting that we find that the dependency function in Gaussian kernel approximation shares the similar idea with the Relief algorithm. It gives a new viewpoint for understanding and extending the existing rough set techniques. The future work could move along two directions. First, we will continue to construct different rough set models with various kernel functions and discuss the common properties of this kind of kernel based rough set models. Second, the existing feature selection algorithms based on rough sets sometimes might not be robust enough for real-world applications; we may contemplate introducing improvements similar to those discussed in the ReliefF series [37, 38, 39, 49].

Acknowledgement

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

References

[1] J. Shawe-Taylor, N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.

[2] C. Cortes, V. Vapnik. Support-vector networks. Machine learning, 20 (1995) 273-297.

[3] C. J. C. Burges. A tutorial on Support Vector Machines for pattern recognition. Data mining and knowledge discovery. 2(1998)121-167

[4] Chen, Jiun-Hung, Chen, Chu-Song. Fuzzy kernel perceptron. IEEE Transactions on Neural Networks, 13(2002) 1364-1373

[5] Baudat G, Anouar FE. Generalized discriminant analysis using a kernel approach. Neural Computation, 12 (2000) 2385-2404

[6] B. Scholkopf, A. Smola, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(1998)1299-1319

[7] Vincent, Pascal, Bengio, Yoshua. Kernel matching pursuit. Machine Learning, 48(2002)65-187

[8] Z. Pawlak, Rough Sets—Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.

[9] J. Wang, Q. Tao. Theory of rough sets and statistical learning. In knowledge science and computational science (R.Q.Lu eds.), Tsinghua Publishing House, 2003

[10] Q. H. Hu, Z. X. Xie, D. R. Yu. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recognition, 40(2007)3509-3521

[11] X. Hu, N. Cercone. Learning in relational databases: a rough set approach. Computational Intelligence, 11(1995) 323–338

[12] Q. H. Hu, D. R. Yu, Z. X. Xie, X. D. Li. EROS: ensemble rough subspaces. Pattern recognition, 40(2007) 3728–3739

[13] Z. Pawlak. Rough set theory and its applications to data analysis. Cybernetics & Systems, 29 (1998) 661-688

[14] R. M. Fang, H. Z. Ma. Hybrid rough set and support vector machine for faults diagnosis of power transformer. dynamics of continuous discrete and impulsive systems-series b-applications & algorithms, vol.13, pp. 1209-1213 Part 3 Suppl., 2006

[15] S. Asharaf, S.K. Shevade,M. Narasimha Murty. Rough support vector clustering. Pattern Recognition, 38 (2005) 1779–1783

[16] P. Lingras, C. Butz. Rough set based 1-v-1 and 1-v-r approaches to support vector machine multi-classification. Information Sciences, 177(2007)3782–3798

[17] Pawlak Z. Rough sets. International Journal of Information and Computer Science, 11(1982)314-356

[18] Q. H. Hu, D. R. Yu, Z. X. Xie. Neighborhood classifiers. Expert Systems with Applications, 34(2008)866-876

[19] D. Dubois and H. Prade, Rough fuzzy sets and fuzzy rough sets. Int. J. Gen. Syst., 17(1990)191-209.

[20] N. Morsi Nehad and M. M. Yakout, “Axiomatics for fuzzy rough sets,” Fuzzy Sets Syst., 100 (1998) 327-342

[21] W. -Z Wu, W. -X Zhang. Constructive and axiomatic approaches of fuzzy approximation operators. Inform. Sci., 159 (2004) 233-254.

[22] D. S. Yeung, D. G. Chen, E. C. C. Tsang, J. W. T. Lee, X. Z Wang. On the generalization of fuzzy rough sets. IEEE Transactions on fuzzy systems, 13 (2005) 343-361

[23] Q. H. Hu, D. R. Yu, Z. X. Xie, J. F. Liu. Fuzzy probabilistic approximation spaces and their information measures. IEEE Transactions on fuzzy systems, 14 (2006)191-201

[24] B. Moser. On the t-transitivity of kernels. Fuzzy Sets and Systems, 157(2006) 787–1796

[25] B. Moser. On Representing and Generating Kernels by Fuzzy Equivalence Relations. Journal of Machine Learning Research, 7 (2006)2603-2620

[26] Q. H. Hu, D. R. Yu, Z. X. Xie. Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognition Letters, 27(2006)414–423

[27] M. G. Genton. Classes of kernels for machine learning: a statistics perspective. Journal of machine learning research, 2 (2001) 299-312

[28] J. -S. Mi, , W. -X. Zhang., An axiomatic characterization of a fuzzy generalization of rough sets, Information Sciences, 160 (2004) 235-249

[29] R. Jensen, Q. Shen. Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. IEEE Transactions on knowledge and data engineering, 16 (2004) 1457-1471.

[30] R. Jensen, Q. Shen. New Approaches to Fuzzy-Rough Feature Selection. IEEE Transactions on fuzzy systems, 17 (2009) 824 -828

[31] E. Hernandez, J. Recasens. A reformulation of entropy in the presence of indistinguishability operators. Fuzzy sets and systems, 128 (2002) 185-196

[32] I. Duntsch, G. Gediga. Uncertainty measures of rough set prediction. Artificial intelligence, 106 (1998)109-137

[33] Y. H. Qian, J. Y. Liang, D. Y. Li, H. Y. Zhang, C. Dang. Measures for evaluating the decision performance of a decision table in rough set theory. Information Sciences. 178 (2008) 181-202

[34] Q. H. Hu, D. Yu. Entropies of fuzzy indiscernibility relation and its operations. Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems, 12 (2004)575–589

[35] D. Yu, Q. H. Hu, C. Wu. Uncertainty measures for fuzzy relations and their applications. Applied soft computing, 7 (2007) 1135-1143

[36] K. Kira, L. A. Rendell. A practical approach to feature selection. In D. Sleeman, & P. Edwards (Eds.), Machine Learning: Proceedings of International Conference (ICML'92) pp. 249–256. Morgan Kaufmann.

[37] M. Robnik-Sikonja, I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53 (2003) 23-69

[38] Y. Sun and J. Li, "Iterative RELIEF for Feature Weighting," Proc. 23rd Int'l Conf. Machine Learning, pp. 913-920, 2006.

[39] Y. J. Sun. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. IEEE Transactions on pattern analysis and machine intelligence, 29 (2007)1035-1051.

[40] L. Yu, H. Liu. Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5 (2004) 1205-1224.

[41] H. Liu, L. Yu. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering, 17(2005)491-502.

[42] R. Battiti. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5 (1994) 537-550

[43] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pp. 359-366, 2000.

[44] M. Dash, H. Liu. Consistency-based search in feature selection. Artificial Intelligence, 151(2003)155–176

[45] D. P. Muni, N. R. Pal, J. Das. Genetic programming for simultaneous feature selection and classifier design. IEEE Transactions on systems, man, and cybernetics—Part B: cybernetics, 36(2006)106-117.

[46] H. L. Xiong, M. N. S. Swamy, M. O. Ahmad. Optimizing the kernel in the empirical feature space. IEEE Transactions on neural networks, 16(2005) 460-474

[47] S. Yang, S. Yan, C. Zhang, et al. Bilinear analysis for Kernel selection and nonlinear feature extraction. IEEE Transactions on neural networks, 18 (2007)1442-1452.

[48] K. R. Muller, S. Mika, G. Ratsch, et al. An introduction to kernel-based learning algorithms. IEEE Transactions on neural networks, 12(2001)181-201

[49] R. Gilad-Bachrach, A., Navot, N. Tishby. Margin based feature selection- theory and algorithms. In proceeding of ICML 2004, pp. 43-50

[50] L. A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 90 (1997) 111-127

[51] W. Pedrycz, G. Vukovich. Feature analysis through information granulation and fuzzy sets. Pattern Recognition, 35 (2002) 825-834.

[52] S. Chen, C. F. N. Cowan. Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks, 2 (1991) 302-309.

[53] T.-L. Tseng, C.-C. Huang. Rough set-based approach to feature selection in customer relationship management.

Omega, 35 (2007) 365-383

[54] Y. Leung, M. M. Fischer, W.-Z. Wu, J.-S. Mi. A rough set approach for the discovery of classification rules in interval-valued information systems. International Journal of Approximate Reasoning, 47(2008) 233-246

[55] T. -J. Li, Y. Leung, W.-X. Zhang. Generalized fuzzy rough approximation operators based on fuzzy coverings. International Journal of Approximate Reasoning, 48 (2008) 836-856.

[56] W. -Z. Wu. Attribute reduction based on evidence theory in incomplete decision systems. Information Sciences, 178(2008) 1355-1371

[57] A. Mieszkowicz-Rolka, L. Rolka. Fuzzy rough approximations of process data. International Journal of Approximate Reasoning, 49 (2008) 301-315

[58] Y. Y. Yao. Two views of the theory of rough sets in finite universes. International Journal of Approximate Reasoning, 15(1996) 291-317

[59] Y. Y. Yao, Y. Zhao. Attribute reduction in decision-theoretic rough set models. Information Sciences, 178(2008) 3356-3373

[60] S. Zhao, E. C.C. Tsang. On fuzzy approximation operators in attribute reduction with fuzzy rough sets. Information Sciences, 178(2008) 3163-3176

[61] Y. H. Qian, J. Y. Liang, C. Y. Dang. Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. International Journal of Approximate Reasoning, 50 (2009) 174-188

[62] Y. J. Yang, R. I. John. Generalisation of roughness bounds in rough set operations. International Journal of Approximate Reasoning, 48 (2008) 868-878

[63] U. Fayyad, K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Proc. Thirteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 1993,1022–1027.