Information Sciences 181 (2011) 5169–5179

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets

Degang Chen [a,*], Qinghua Hu [b], Yongping Yang [a]

[a] North China Electric Power University, Beijing 102206, PR China
[b] Harbin Institute of Technology, Harbin 150001, PR China

ARTICLE INFO

Article history:
Received 15 September 2008
Received in revised form 24 June 2011
Accepted 5 July 2011
Available online 23 July 2011

Keywords:
Gaussian kernels
Fuzzy rough sets
Feature selection
Discernibility matrix

ABSTRACT

Fuzzy rough sets are considered as an effective tool to deal with uncertainty in data analysis, and fuzzy similarity relations are used in fuzzy rough sets to calculate similarity between objects. On the other hand in kernel tricks, a kernel maps data into a higher dimensional feature space where the resulting structure of the learning task is linearly separable, while the kernel is the inner product of this feature space and can also be viewed as a similarity function. It has been reported there is an overlap between family of kernels and collection of fuzzy similarity relations. This fact motivates the idea in this paper to use some kernels as fuzzy similarity relations and develop kernel based fuzzy rough sets. First, we consider Gaussian kernel and propose Gaussian kernel based fuzzy rough sets. Second we introduce parameterized attribute reduction with the derived model of fuzzy rough sets. Structures of attribute reduction are investigated and an algorithm with discernibility matrix to find all reducts is developed. Finally, a heuristic algorithm is designed to compute reducts with Gaussian kernel fuzzy rough sets. Several experiments are provided to demonstrate the effectiveness of the idea.

1. Introduction

As a general pursuit in the domain of machine learning, kernel trick allows mapping data from input space into a higher dimensional feature space through kernel functions in order to simplify learning tasks and make them linear (viz. solvable by linear classifiers [37]). In this way, a number of linear learning algorithms can be extended to deal with nonlinear tasks, such as nonlinear SVMs [42], kernel perceptron [5], kernel discriminant analysis [37], nonlinear component analysis [37], kernel matching pursuit [43], etc. Most of them employ feature selection as a preprocessing step. According to [34], feature selection aims at picking out some of the original input features (i) for performance improvement by facilitating data collection and reducing storage space and classification time, (ii) to perform semantics analysis in helping understand the problem, and (iii) to improve prediction accuracy by avoiding "curse of dimensionality".

According to [2,12,13,24], feature selection approaches can be divided into filters [8,9,14,15,28], wrappers [24,44] and embedded approaches [1,4,51]. Acquiring no feedback from classifiers, the filter methods estimate the classification performance by some indirect assessments, such as distance measures which reflect how well the classes separate from each other. The wrapper methods, on the other hand, take the classification performance of a learning machine as a measure of goodness of a subset of features. Wrapper methods usually provide more accurate solutions than filter methods [26,27,44], but are more computationally expensive. Finally, embedded approaches simultaneously perform feature selection and classification modeling in the training process.

* Corresponding author.
E-mail address: chengdegang@263.net (D. Chen).

0020-0255/$ - see front matter © 2011 Published by Elsevier Inc.
doi:10.1016/j.ins.2011.07.025

Fuzzy rough sets are extended from Pawlak's rough sets [35] for dealing with decision tables with real-valued attributes rather than symbolic ones. In the existing framework of fuzzy rough sets a fuzzy similarity relation is employed to measure similarity between two objects. There are two topics related with fuzzy rough sets: developing different models of fuzzy rough sets and performing attribute reduction with fuzzy rough sets. Since the pioneering work in [6], many efforts [29–31,36,45,46] have been put on the first topic. Detailed summarizations on models of fuzzy rough sets can be found in [48]. On the other hand, attribute reduction with fuzzy rough sets was first proposed in [20], where fuzzy dependency function was employed to measure goodness of attributes by fuzzy rough sets proposed in [6] and an algorithm to compute a reduct was developed. Some researches on attribute reduction with fuzzy rough sets were mainly contributed to improve the method in [20], and some key concepts in the traditional rough sets, such as core of reducts, were also generalized to fuzzy rough sets [3,16,18–23,41,49,50].

As the role of fuzzy similarity relation in the framework of fuzzy rough sets, kernels also play the role as similarity measures in the framework of kernel tricks. It was pointed out in [32] that there is a closed relationship between kernels and fuzzy similarity relations, i.e., kernels mapping to the unit interval with 1 in their diagonal are a class of fuzzy similarity relations. This fact naturally motivates the idea in this paper to consider such kind of kernels as fuzzy similarity relations to develop kernel based fuzzy rough sets, and attribute reduction with kernel based fuzzy rough sets can be proposed as preprocessing step for kernel tricks.

In this paper we select the well-known Gaussian kernels as fuzzy similarity relations in the framework of fuzzy rough sets. We first develop Gaussian kernels based fuzzy rough sets and discuss their granular structures. Second we define parameterized attribute reduction with Gaussian kernel based fuzzy rough sets and develop an algorithm with discernibility matrix to compute all the reducts. Here we employ positive region in fuzzy rough sets to measure goodness of subsets of attributes and attributes are distinguished according to their importance related to the decision. Heuristic algorithm to find reducts is also proposed. At last, we employ the reduction algorithm for Gaussian kernel SVM as a preprocessing step in classification learning.

However, it is notable that every kernel mapping to the unit interval with 1 in its diagonal can also be considered as a fuzzy similarity function in fuzzy rough sets, and different kernels may have different techniques to perform attribute reduction. We discuss Gaussian kernels in this paper since they are widely used in the field of machine learning. The proposed idea can be employed as a preprocessing step of kernel trick related to Gaussian kernels.

The rest of this paper is organized as follows: Section 2 introduces kernel trick and fuzzy rough sets. Section 3 develops attribute reduction with Gaussian kernel based fuzzy rough sets; the structure of selected attribute set is characterized by approach of discernibility matrix in this section. Experimental results are described in Section 4. Conclusions are presented in Section 5.

2. Reviews on kernels and fuzzy rough sets

2.1. Positive definite kernels

Supposed $X \subset R^n$, H is a Hilbert space. $k(x,x')$ is a continuous and symmetric function on $X \times X$, if there exists a function $\Phi : X \to H$ satisfying that $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_H$, then $k(x,x')$ is called a positive definite kernel [37]. With a positive definite kernel $k(x,x')$, input vectors are mapped into a Hilbert space H, called feature space. According to [37], kernel trick means that given an algorithm which is formulated in terms of a positive definite kernel $k(x,x')$, one can construct an alternative algorithm by replacing $k(x,x')$ with another positive definite kernel $\tilde{k}(x,x')$. In view of definition of positive definite kernel, the justification for this procedure is that the original algorithm can be thought of as a dot product based algorithm operating on the data $\Phi(x_1),\ldots,\Phi(x_m)$. The algorithm obtained by replacing $k(x,x')$ with $\tilde{k}(x,x')$ is then exactly the same dot product based algorithm. The only difference comes from that they operate on $\tilde{\Phi}(x_1),\ldots,\tilde{\Phi}(x_m)$.

Generally speaking, there are mainly two kinds of kernels: translation invariant kernels and dot product kernels. The translation invariant kernels are independent of the absolute position of input x and x'. They only depend on the difference between x and x'. So we have $k(x,x') = k(x - x')$. Gaussian kernel $k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ is a well known translation invariant kernel. Some other translation invariant kernels include B_n-splines kernels, Dirichlet kernels and Periodic kernels. The second important family of kernels can be efficiently described in term of dot product, i.e., $k(x,x') = k(\langle x,x' \rangle)$, including homogeneous polynomial kernels $k(x,x') = \langle x,x' \rangle^p$ and inhomogeneous polynomial kernels $k(x,x') = (\langle x,x' \rangle + c)^p$ with $c \geqslant 0$.

2.2. Fuzzy rough sets

In this subsection we first review fuzzy logic operators found in [29,31,36,48], then give a brief introduction of fuzzy rough sets.

Triangular norms (t-norms for short) have been originally studied within the framework of probabilistic metric spaces [38,39]. In this context, t-norms proved to be an appropriate concept when dealing with triangle inequalities. Latter on, t-norms and their dual version t-conorms have been used to model conjunction and disjunction for many-valued logic [7,11,25].

A t-norm is an increasing, associative and commutative mapping $T : [0,1] \times [0,1] \to [0,1]$ that satisfies the boundary condition ($\forall x \in [0,1]$, $T(x,1) = x$).

A triangular conorm (shortly t-conorm) is an increasing, associative and commutative mapping $S:[0,1]\times[0,1]\to[0,1]$ that satisfies the boundary condition ($\forall x\in[0,1]$, $S(x,0)=x$).

Given a t-norm T, the binary operation $\vartheta_T(\alpha,\gamma)=\sup\{\theta\in[0,1]:T(\alpha,\theta)\leqslant\gamma\}$ is called a R-implicator based on T. If T is lower semi-continuous, then ϑ_T is called a residuation implication of T, or a T-residuated implication. In [29] σ is defined by $\sigma(a,b)=\inf\{c\in[0,1]:S(a,c)\geqslant b\}$ as the residuated implication of a t-conorm S.

An information system is a pair $\mathbf{A}=(U,\mathbf{C})$, where $U=\{x_1,\ldots,x_n\}$ is a nonempty universe of discourse and $\mathbf{C}=\{a_1,a_2,\ldots,a_m\}$ is a nonempty finite set of attributes. With a subset of attributes $\mathbf{B}\subseteq\mathbf{C}$ we associate a binary relation $IND(\mathbf{B})$, called $\mathbf{B}$ – indiscernibility relation defined as $IND(\mathbf{B})=\{(x,y)\in U\times U:a(x)=a(y),\ \forall a\in\mathbf{B}\}$. Then $IND(\mathbf{B})$ is an equivalence relation and $IND(\mathbf{B})=\cap_{a\in\mathbf{B}}IND(\{a\})$. By $[x]_{\mathbf{B}}$ we denote the equivalence class of $IND(\mathbf{B})$ including x. For $X\subseteq U$ the sets $\cup\{[x]_{\mathbf{B}}:[x]_{\mathbf{B}}\subseteq X\}$ and $\cup\{[x]_{\mathbf{B}}:[x]_{\mathbf{B}}\cap X\neq\phi\}$ are called $\mathbf{B}$ – lower and $\mathbf{B}$ – upper approximations of X in $\mathbf{A}$, respectively, denoted by $\underline{\mathbf{B}}X$ and $\overline{\mathbf{B}}X$.

However, the above traditional rough set model can just deal with databases described with symbolic attributes. This limits the applications of rough sets. Several generalizations of the traditional rough sets were considered. Among these generalizations, the combination of rough sets and fuzzy sets develops a powerful tool, called fuzzy rough sets, to deal with real-valued datasets.

The definition of fuzzy rough sets was first proposed in [6]. Since then many efforts have been devoted to developing and characterizing models of fuzzy rough sets. Detailed summaries on this topic can be found in [41,45–48]. We here just offer the basic definitions of fuzzy rough sets.

Suppose U is a nonempty universe of discourses. As a similarity measure between two objects, a fuzzy T – similarity relation R is a fuzzy set on $U\times U$ which is reflexive, symmetric and T – transitive, namely $R(x,z)\geqslant T(R(x,y),R(y,z))$ holds. For $A\in F(U)$, the lower and upper approximations of A are defined as follows:

(1) T – upper approximation operator: $\overline{R_T}A(x)=\sup_{u\in U}T(R(x,u),A(u))$;
(2) S – lower approximation operator: $\underline{R_S}A(x)=\inf_{u\in U}S(N(R(x,u)),A(u))$;
(3) σ – upper approximation operator: $\overline{R_\sigma}A(x)=\sup_{u\in U}\sigma(N(R(x,u)),A(u))$;
(4) ϑ – lower approximation operator: $\underline{R_\vartheta}A(x)=\inf_{u\in U}\vartheta(R(x,u),A(u))$.

3. Approximations and attribute reduction with Gaussian kernels

Gaussian kernels are widely used in kernel tricks. In this section we consider Gaussian kernels as fuzzy T – similarity relations to develop Gaussian kernel based fuzzy rough sets and consider attribute reduction with Gaussian kernels.

3.1. Gaussian kernel based fuzzy rough sets

Suppose $U=\{x_1,x_2,\ldots,x_m\}$ is a finite universe of discourses, and every element $x_i\in U$ is described by a vector $(x_{i1},x_{i2},\ldots,x_{in})\in R^n$. Thus U is viewed as a subset of R^n. Since Gaussian kernel $k(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)$ takes values in $[0,1]$, it can be considered as a fuzzy relation. We denote this fuzzy relation by R_G^n, i.e., $R_G^n(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)$. Obviously R_G^n is reflexive and symmetric. In [32] it is pointed out that R_G^n is $T_{\cos}$ – transitive, where $T_{\cos}(a,b)=\max\left\{ab-\sqrt{1-a^2}\sqrt{1-b^2},0\right\}$ is a triangular norm. Thus R_G^n is a fuzzy $T_{\cos}$ – similarity relation. To obtain lower and upper approximations of fuzzy sets related to R_G^n, we first derive the residuated implication of $T_{\cos}$ by the following lemma.

Lemma 3.1.1 ([19,32]).

$$\vartheta_{T_{\cos}}(a,b)=\begin{cases}1, & a\leqslant b\\ ab+\sqrt{(1-a^2)(1-b^2)}, & a>b\end{cases}.$$

Proof. We have $\vartheta_{T_{\cos}}(a,b)=\sup\{\theta\in[0,1]:T_{\cos}(a,\theta)\leqslant b\}$, so if $a\leqslant b$, then $\vartheta_{T_{\cos}}(a,b)=1$. Suppose $a>b$. θ should satisfy $a\theta-\sqrt{(1-a^2)(1-\theta^2)}\leqslant b$. It means $a\theta-b\leqslant\sqrt{(1-a^2)(1-\theta^2)}$.

Let $f_1(\theta)=a\theta-b$ and $f_2(\theta)=\sqrt{(1-a^2)(1-\theta^2)}$. Then $f_1(\theta)$ strictly increases on $[0,1]$, and $f_2(\theta)$ strictly decreases on $[0,1]$. If $\theta=ab+\sqrt{(1-a^2)(1-b^2)}$, then $f_1(\theta)=f_2(\theta)$. So if $\theta\leqslant ab+\sqrt{(1-a^2)(1-b^2)}$, then $f_1(\theta)\leqslant f_2(\theta)$; if $\theta>ab+\sqrt{(1-a^2)(1-b^2)}$, then $f_1(\theta)>f_2(\theta)$. This implies that $\sup\{\theta\in[0,1]:T_{\cos}(a,\theta)\leqslant b\}=ab+\sqrt{(1-a^2)(1-b^2)}$.

It should be noted that result of Lemma 3.1.1 have been mentioned in [32] without proof, here we give a proof of it. □

In the rest of this paper we denote $\vartheta_{T_{\cos}}$ by $\vartheta_{\cos}$ for short. With $T_{\cos}$ and $\vartheta_{\cos}$ Gaussian kernel based fuzzy rough sets can be computed as: $\overline{R_G^n}A(x)=\sup_{u\in U}T_{\cos}(R_G^n(x,u),A(u))$, $\underline{R_G^n}A(x)=\inf_{u\in U}\vartheta_{\cos}(R_G^n(x,u),A(u))$.

The properties of $\overline{R_G^n}$ and $\underline{R_G^n}$ are discussed in [48] within the framework of general fuzzy rough sets. Here we only list the necessary ones for this paper.

Theorem 3.1.2 [48]. *$\overline{R_G^n}$ and $\underline{R_G^n}$ satisfy the following properties:*

(1) $\underline{R_G^n}A \subseteq A \subseteq \overline{R_G^n}A,\ \overline{R_G^n}A = A \Longleftrightarrow \underline{R_G^n}A = A$;
(2) *$\underline{R_G^n}$ and $\overline{R_G^n}$ are monotone;*
(3) $\underline{R_G^n}(\underline{R_G^n}A) = \underline{R_G^n}A,\ \overline{R_G^n}(\overline{R_G^n}A) = \overline{R_G^n}A;\ \overline{R_G^n}(\underline{R_G^n}A) = \underline{R_G^n}A,\ \underline{R_G^n}(\overline{R_G^n}A) = \overline{R_G^n}A$;
(4) $\overline{R_G^n}(\cup_{t\in T}A_t) = \cup_{t\in T}\overline{R_G^n}A_t,\ \underline{R_G^n}(\cap_{t\in T}A_t) = \cap_{t\in T}\underline{R_G^n}A_t$;
(5) *If $R_G^m \subseteq R_G^n$, then $\underline{R_G^n}A \subseteq \underline{R_G^m}A \subseteq A \subseteq \overline{R_G^m}A \subseteq \overline{R_G^n}A$.*

By (1) we know $\overline{R_G^n}A$ and $\underline{R_G^n}A$ are a pair of fuzzy sets approximating A as upper and lower bounds, respectively, and by (5) we can get that a smaller fuzzy relation can offer more precise approximations. These properties are the theoretical foundation of attribute reduction described in Section 3.2.

The following theorem shows the granular structure of $\overline{R_G^n}A$ and $\underline{R_G^n}A$.

Theorem 3.1.3. *$\overline{R_G^n}A = \cup\{\overline{R_G^n}x_\lambda : x_\lambda \subseteq A\},\ \underline{R_G^n}A = \cup\left\{\overline{R_G^n}x_\lambda : \overline{R_G^n}x_\lambda \subseteq A\right\}$, here x_λ is a fuzzy set, called fuzzy point, defined as* $x_\lambda(y) = \begin{cases} \lambda, & y = x \\ 0, & y \neq x \end{cases}$.

Proof. Since $A = \cup\{x_\lambda : x_\lambda \subseteq A\}$, by (4) of Theorem 3.1.2 $\overline{R_G^n}A = \cup\{\overline{R_G^n}x_\lambda : x_\lambda \subseteq A\}$ is obviously true.

Suppose $\underline{\underline{R_G^n}}A = \cup\{\overline{R_G^n}z_\gamma : \overline{R_G^n}z_\gamma \subseteq A\}$. For every $x \in U$, suppose $\lambda = \underline{\underline{R_G^n}}A(x)$, then we have

$$\overline{R_G^n}x_\lambda(y) = T_{\cos}(R(x,y),\lambda) = T_{\cos}\left(R(x,y), \sup\left\{\overline{R_G^n}z_\gamma(x) : \overline{R_G^n}z_\gamma \subseteq A\right\}\right) = \sup\left\{T_{\cos}(R(x,y), T_{\cos}(R(z,x),\gamma)) : \overline{R_G^n}z_\gamma \subseteq A\right\}$$
$$\leqslant \sup\left\{T_{\cos}(R(z,y),\gamma) : \overline{R_G^n}z_\gamma \subseteq A\right\} = \underline{\underline{R_G^n}}A(y).$$

Thus $\overline{R_G^n}x_\lambda \subseteq \underline{\underline{R_G^n}}A$ holds. And for any $\lambda' > \lambda$, clearly $\overline{R_G^n}x_{\lambda'}$ cannot be included by A; otherwise $\underline{\underline{R_G^n}}A(x) = \lambda'$. It is a contradiction. So it implies that $\overline{R_G^n}x_\lambda$ is the maximal one in the collection $\{\overline{R_G^n}x_\eta : \eta \in (0,1]\}$ to be included by A.

On the other hand, for $u \in U$ we have $\overline{R_G^n}x_\lambda \subseteq \cup\left\{\overline{R_G^n}x_\beta : \overline{R_G^n}x_\beta(u) \leqslant A(u)\right\}$. Clearly $\overline{R_G^n}x_\lambda \subseteq \cap_{u\in U}\left(\cup\left\{\overline{R_G^n}x_\beta : \overline{R_G^n}x_\beta(u) \leqslant A(u)\right\}\right)$. Since $\cup\overline{R_G^n}x_\beta \in \left\{\overline{R_G^n}x_\eta : \eta \in (0,1]\right\}$, we have $\cap_{u\in U}\left(\cup\left\{\overline{R_G^n}x_\beta : \overline{R_G^n}x_\beta(u) \leqslant A(u)\right\}\right) \in \{\overline{R_G^n}x_\eta : \eta \in (0,1]\}$; thus $\overline{R_G^n}x_\lambda = \cap_{u\in U}\left(\cup\left\{\overline{R_G^n}x_\beta : \overline{R_G^n}x_\beta(u) \leqslant A(u)\right\}\right)$. Since $\overline{R_G^n}x_\lambda$ is the maximal one in the collection $\{\overline{R_G^n}x_\eta : \eta \in (0,1]\}$ included by A, it implies $\lambda = \overline{R_G^n}x_\lambda(x) = \inf_{u\in U}\sup\{\beta : T_{\cos}(R(x,u),\beta) \leqslant A(u)\} = \inf_{u\in U}\vartheta_{\cos}(R(x,u),A(u))$. □

According to Theorem 3.1.3, $M_G^n = \{\overline{R_G^n}x_\eta : x \in U, \eta \in (0,1]\}$ can be employed as the basic granular set to construct $\overline{R_G^n}$ and $\underline{R_G^n}$. This statement plays a key role in subsection 3.2 when we characterize the structure of reducts.

3.2. Parameterized attribute reduction related to Gaussian kernel based fuzzy rough sets

Suppose $U = \{x_1, x_2, \ldots, x_m\}$ is a finite universe. Each element $x_i \in U$ is described by a set $\mathbf{C}$ of n attributes with numerical values. The attribute value of x_i related to the jth attribute is x_{ij}. The pair $(U,\mathbf{C})$ is an information system. Suppose U is divided into several disjoint parts $D_1, D_2, \ldots, D_s$ with a decision attribute D. Then the triple $(U,\mathbf{C},D)$ is called a decision system.

A subset of $\mathbf{C}$ induces a fuzzy $T_{\cos}$ – similarity relation with the Gaussian kernel. We denote $R_G^{(j)}(x_i,x_k) = \exp\left(-\frac{\|x_{ij}-x_{kj}\|^2}{2\sigma^2}\right)$ as the one computed with the jth attribute in $\mathbf{C}$, then $\mathbf{C}$ can be equivalently written by $\mathbf{C} = \left\{R_G^{(1)}, R_G^{(2)}, \ldots, R_G^{(n)}\right\}$.

Firstly, we should give the aggregation operator of multiple elements in $\mathbf{C}$. In the existing fuzzy rough sets [3,16–23,41,50] t-norm *Min* is used as the aggregation operator of several fuzzy relations, and the fuzzy relation after aggregation is just the intersection of these fuzzy relations. However, if we select *Min* as the aggregation operator of elements in $\mathbf{C} = \left\{R_G^{(1)}, R_G^{(2)}, \ldots, R_G^{(n)}\right\}$, the resulting aggregation does not coincide with R_G^n due to the following property of Gaussian kernels: $R_G^n(x_i,x_k) = \exp\left(-\frac{\|x_i-x_k\|^2}{2\sigma^2}\right) = \prod_{s=1}^n R_G^{(s)}(x_i,x_k)$. Instead of the t-norm *Min*, we introduce the algebraic product $T_P(x,y) = x \cdot y$ as the aggregation operator. Clearly, in this case the resulting aggregation of elements in $\mathbf{C} = \left\{R_G^{(1)}, R_G^{(2)}, \ldots, R_G^{(n)}\right\}$ is equal to R_G^n. In the following we denote the fuzzy relation aggregated by $T_P(x,y) = x \cdot y$ with elements in $\mathbf{P} \subseteq \mathbf{C}$ by $R_G^{\mathbf{P}}$ and still denote $R_G^{\mathbf{C}}$ by R_G^n.

Secondly, we should develop a method to measure goodness of subsets of conditional attributes. For each D_t, $t = 1,2,\ldots,s$, if $x \notin D_t$, then we have $\underline{R_G^n}D_t(x) = 0$. If $x \in D_t$, $\underline{R_G^n}D_t(x) = \inf_{u\in U}\vartheta_{\cos}(R_G^n(x,u), D_t(u)) = \inf_{u\notin D_t}\vartheta_{\cos}(R_G^n(x,u), D_t(u)) =$

$\inf_{u\notin D_t}\sqrt{1-(R_G^n(x,u))^2}$. $\underline{R_G^n}D_t(x)$ can be understood as the certainty degree to that x belongs to D_t according to the attributes in $\mathbf{C}$. One obvious observation is that $\underline{R_G^n}D_t(x)$ is determined by the smallest one (the worst case) of $\sqrt{1-(R_G^n(x,u))^2},\ u\notin D_t$. Thus if there is $u_0\notin D_t$ such that $R_G^n(x,u_0)$ is great enough, i.e., x is quite similar to an object in other classes, then $\underline{R_G^n}D_t(x)$ should be very small. Another observation is that $\underline{R_G^n}D_t(x)=1$ is never true because $R_G^n(x,u)\neq 0$ always holds, i.e., every pair of objects are similar in a certain degree with respect to Gaussian kernels. $Pos_{\mathbf{C}}(D)=\cup_{t=1}^{s}\underline{R_G^n}D_t$ is called the positive region of decision attribute D related to the conditional attribute set $\mathbf{C}$, we will employ positive region as a measure of goodness of attributes.

However, if we use Gaussian function as the similarity function, deleting any attribute C from $\mathbf{C}$ will result in $Pos_{\mathbf{C}}D(x_i) > Pos_{\mathbf{C}-\{C\}}D(x_i)$ for every $i=1,2,\ldots,m$. So we cannot employ the idea in the traditional rough sets [35,40] and existing fuzzy rough sets [41] to define attribute reduct as the minimal subset of $\mathbf{C}$ to keep the positive region invariant. This issue is also not mentioned in [19]. We overcome this problem by considering a threshold ε of the positive region, and we can define a parameterized attribute reduct with Gaussian kernel based fuzzy rough sets by limiting the change of positive region within the given threshold ε. The idea can be formulated as follows.

Definition 3.2.1. Suppose $(U,\mathbf{C},D)$ is a decision system, $\varepsilon\in[0,1]$. For $C\in\mathbf{C}$, if $Pos_{\mathbf{C}}D(x_i)-Pos_{\mathbf{C}\ -\{C\}}D(x_i)\leqslant\varepsilon$ for every $i=1,2,\ldots,m$, then C is called ε – superfluous in $\mathbf{C}$ relative to D; otherwise C is called ε – indispensable in $\mathbf{C}$ relative to D. For every $\mathbf{P}\subseteq\mathbf{C}$, if $Pos_{\mathbf{C}}D(x_i)-Pos_{\mathbf{P}}D(x_i)\leqslant\varepsilon$ for every $i=1,2,\ldots,m$, and every element in $\mathbf{P}$ is indispensable, then $\mathbf{P}$ is called a ε – reduct of $\mathbf{C}$ relative to D. The collection of all the ε – indispensable elements in $\mathbf{C}$ is called the ε – core of $\mathbf{C}$ relative to D, denoted by $Core_D(\mathbf{C})$, and we have the following theorem for the core.

Theorem 3.2.1. *$Core_D(\mathbf{C})=\cap Red_D(\mathbf{C})$, where $Red_D(\mathbf{C})$ is the collection of all the ε – reduct of $\mathbf{C}$ relative to D.*

Proof. If C is ε – indispensable in $\mathbf{C}$ relative to D, then C should be included in every ε – reduct of $\mathbf{C}$. Hence $Core_D(\mathbf{C})\subseteq\cap Red_D(\mathbf{C})$. On the other hand, if C is ε – superfluous in $\mathbf{C}$ relative to D, then $\mathbf{C}-\{C\}$contains a ε – reduct of $\mathbf{C}$, thus there is a ε – reduct of $\mathbf{C}$ that does not include C, hence $C\notin\cap Red_D(\mathbf{C})$. It implies $Core_D(\mathbf{C})\supseteq\cap Red_D(\mathbf{C})$. □

For $C\in\mathbf{C}$, clearly C is ε – superfluous in $\mathbf{C}$ relative to D if and only if $\underline{R_G^n}D_t(x)-\varepsilon\leqslant\underline{R_G^{\mathbf{C}-\{C\}}}D_t(x)$ for every D_t, $t=1,2,\ldots,l$ and $x\in D_t$, and if and only if $x_{\lambda(x)}\subseteq\underline{R_G^{\mathbf{C}-\{C\}}}D_t$ for $x\in D_t$ and $\lambda(x)=\underline{R_G^n}D_t(x)-\varepsilon$, and if and only if $\overline{R_G^{\mathbf{C}-\{C\}}}x_{\lambda(x)}\subseteq\underline{R_G^{\mathbf{C}-\{C\}}}D_t$ for $x\in D_t$ by Theorem 3.1.3, here $x_{\lambda(\mathrm{x})}$ is a fuzzy point. Thus we have the following theorem.

Theorem 3.2.2. *Suppose $\mathbf{P}\subseteq\mathbf{C}$. $\mathbf{P}$ contains a ε – reduct of $\mathbf{C}$ if and only if $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}(z)=0$ for $x\in D_t$, $z\notin D_t$, $t=1,2,\ldots,l$.*

Proof. $\mathbf{P}$ contains a ε – reduct of $\mathbf{C}$ if and only if $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}\subseteq\underline{R_G^{\mathbf{P}}}D_t$ for $x\in D_t$. If $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}\subseteq\underline{R_G^{\mathbf{P}}}D_t$, then clearly $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}(z)=0$ for $z\notin D_t$. □

Conversely, if $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}(z)=0$ for $z\notin D_t$, then $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}\subseteq D_t$ which implies $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}\subseteq\underline{R_G^{\mathbf{P}}}D_t$ by Theorem 3.1.3.

Theorem 3.2.3. *Suppose $\mathbf{P}\subseteq\mathbf{C}$. $\mathbf{P}$ contains a ε – reduct of $\mathbf{C}$ if and only if there is $\mathbf{Q}\subseteq\mathbf{P}$ such that $R_G^{\mathbf{Q}}(x,z)\leqslant\sqrt{1-\lambda^2(x)}$ for $x\in D_t$, $z\notin D_t$, $t=1,2,\ldots,l$.*

Proof. $\overline{R_G^{\mathbf{P}}}x_{\lambda(x)}(z)=0$ for $x\in D_t, z\notin D_t,\ t=1,2,\ldots,l\iff\sup_{u\in U}T_{\cos}(R_G^{\mathbf{P}}(z,u),x_{\lambda(x)}(u))=0\iff T_{\cos}(R_G^{\mathbf{P}}(x,z),\lambda(x))=0\iff R_G^{\mathbf{P}}(x,z)\cdot\lambda(x)-\sqrt{(1-(R_G^{\mathbf{P}}(x,z))^2)(1-\lambda^2(x))}\leqslant 0\iff R_G^{\mathbf{P}}(x,z)\leqslant\sqrt{1-\lambda^2(x)}\iff$ there is $\mathbf{Q}\subseteq\mathbf{P}$ such that $R_G^{\mathbf{Q}}(x,z)\leqslant\sqrt{1-\lambda^2(x)}$.

By Theorem 3.2.2 we finish the proof. □

Theorem 3.2.3 will be applied to study the structure of ε – reduction of $\mathbf{C}$ and design algorithms to compute all the ε – reducts of $\mathbf{C}$ in the following subsection.

3.3. Discernibility matrix based attribute reduction

Discernibility matrix is a key concept to investigate attribute reduction in the rough set framework [40]. A reasonable definition of discernibility matrix can reveal the structure of attribute reduction, furthermore, it is the theoretical foundation to design algorithms to compute reducts. In this subsection we develop an approach to find the ε – reducts based on discernibility matrix.

Definition 3.3.1. Suppose $(U,\mathbf{C},D)$ is a decision system. By $M(U,\mathbf{C},D)$ we denote a $m\times m$ matrix $(c_{ij})_{m\times m}$, called the discernibility matrix of $(U,\mathbf{C},D)$, defined as

(1) if x_i and x_j belong to different decision classes, $c_{ij} = \{\wedge\mathbf{P} : \mathbf{P} \subseteq \mathbf{C}\}$, here $\wedge\mathbf{P}$ is the conjunction of elements in $\mathbf{P}$, and $\mathbf{P}$ satisfies $R_G^{\mathbf{P}}(x_i,x_j) \leqslant \sqrt{1-\lambda^2(x_i)}$ and for $\mathbf{Q} \subseteq \mathbf{P}$ such that $R_G^{\mathbf{Q}}(x_i,x_j) \leqslant \sqrt{1-\lambda^2(x_i)}$, then $\mathbf{Q} = \mathbf{P}$.
(2) $c_{ij} = \phi$, otherwise.

Clearly c_{ij} is the collection of all the conjunctions of elements in $\mathbf{P} \subseteq \mathbf{C}$ thus that $\mathbf{P}$ is a minimal one satisfying $R_G^{\mathbf{P}}(x_i,x_j) \leqslant \sqrt{1-\lambda^2(x_i)}$. It is remarkable that $M(U,\mathbf{C},D)$ may not be symmetric in this case.

Theorem 3.3.1. *Suppose $(U,\mathbf{C},D)$ is a decision system, $\mathbf{P} \subseteq \mathbf{C}$. We have the following two statements:*

(1) *$\mathbf{P}$ contains a ε – reduct of $\mathbf{C}$ if and only if $\mathbf{P} \cap c_{ij} \neq \phi$ for $c_{ij} \neq \phi$, here $\mathbf{P} \cap c_{ij}$ defined as $\cup\{\mathbf{Q} \subseteq \mathbf{P} : \wedge\mathbf{Q} \in c_{ij}\}$.*
(2) *$Core_D(\mathbf{C}) = \cup\{\mathbf{Q}_{ij} \subseteq \mathbf{C} : \mathbf{Q}_{ij} = \cap\{\mathbf{P} : \wedge\mathbf{P} \in c_{ij}\},\ i,j = 1,2,\ldots,m\}$.*

Proof

(1) If $\mathbf{P}$ contains a ε – reduct of $\mathbf{C}$, then for x_i and x_j belong to different decision classes, there exists a minimal $\mathbf{Q} \subseteq \mathbf{P}$ such that $R_G^{\mathbf{Q}}(x_i,x_j) \leqslant \sqrt{1-\lambda^2(x_i)}$, thus $\wedge\mathbf{Q} \in c_{ij}$ and $\mathbf{P} \cap c_{ij} \neq \phi$.

Conversely, if $\mathbf{P} \cap c_{ij} \neq \phi$ for $c_{ij} \neq \phi$, then there exists a minimal $\mathbf{Q} \subseteq \mathbf{P}$ such that $R_G^{\mathbf{Q}}(x_i,x_j) \leqslant \sqrt{1-\lambda^2(x_i)}$, thus $\mathbf{P}$ contains a ε – reduct of $\mathbf{C}$.
(2) If $C \in Core_D(\mathbf{C})$, then there exist x_i and x_j belonging to different decision classes such that $R_G^{\mathbf{C}-\{C\}}(x_i,x_j) > \sqrt{1-\lambda^2(x_i)}$. So if $\mathbf{P} \subseteq \mathbf{C}$ such that $R_G^{\mathbf{P}}(x_i,x_j) \leqslant \sqrt{1-\lambda^2(x_i)}$, then $C \in \mathbf{P}$ must hold. This implies

$$C \in \cap\{\mathbf{P} : \wedge\mathbf{P} \in c_{ij}\} \text{ and } Core_D(\mathbf{C}) \subseteq \cup\left\{\mathbf{Q}_{ij} \subseteq \mathbf{C} : \mathbf{Q}_{ij} = \cap\{\mathbf{P} : \wedge\mathbf{P} \in c_{ij}\},\ i,j = 1,2,\ldots,m\right\}.$$

Conversely, if $C \in \cup\{\mathbf{Q}_{ij} \subseteq \mathbf{C}:\mathbf{Q}_{ij} = \cap\{\mathbf{P} : \wedge\mathbf{P} \in c_{ij}\}, i,j = 1,2,\ldots,m\}$, then there exist x_i and x_j belonging to different decision classes such that $C \in \cap\{\mathbf{P} : \wedge\mathbf{P} \in c_{ij}\}$, so $R_G^{\mathbf{C}-\{C\}}(x_i,x_j) > \sqrt{1-\lambda^2(x_i)}$ holds, which implies $C \in Core_D(\mathbf{C})$. Thus we finish the proof. □

(2) of Theorem 3.3.1 proposes a formula to compute the relative core by discernibility matrix, this formula will play a key role when design algorithm to compute one reduct in Section 4.1.

Corollary 3.3.2. *Suppose $(U,\mathbf{C},D)$ is a decision system, $\mathbf{P} \subseteq \mathbf{C}$. $\mathbf{P}$ is a ε – reduct of $\mathbf{C}$ if and only if $\mathbf{P}$ is the minimal subset of $\mathbf{C}$ satisfying $\mathbf{P} \cap c_{ij} \neq \phi$ for $c_{ij} \neq \phi$.*

A discernibility function $f(U,\mathbf{C},D)$ for $(U,\mathbf{C},D)$ is a Boolean function of n Boolean variables $\overline{C_1},\overline{C_2},\ldots,\overline{C_n}$ corresponding to the attributes $C_1,C_2,\ldots,C_n$ in $\mathbf{C}$, respectively, and defined as $f(U,\mathbf{C},D)(\overline{C_1},\overline{C_2},\ldots,\overline{C_n}) = \wedge\{\vee(c_{ij}) : c_{ij} \neq \phi,\ 1 \leqslant i,\ j \leqslant m\}$, where $\vee(c_{ij})$ is the disjunction of all elements in c_{ij} as $\wedge\mathbf{P}$. By using of the discernibility function, we have the following theorem to compute all the ε – reducts of $\mathbf{C}$.

Theorem 3.3.3. *Suppose $(U,\mathbf{C},D)$ is a decision system; $M(U,\mathbf{C},D) = (c_{ij} : i,j \leqslant n)$ is the discernibility matrix of $(U,\mathbf{C},D)$ and $f(U,\mathbf{C},D)$ is the discernibility function of $(U,\mathbf{C},D)$. If $f(U,\mathbf{C},D) = \vee_{k=1}^{l}(\wedge\mathbf{\Delta}_k)(\mathbf{\Delta}_k \subseteq \mathbf{C})$ is computed from $f(U,\mathbf{C},D)$ by applying the multiplication and absorption laws as many times as possible such that every element in $\mathbf{\Delta}_i$ only appears one time, then the set $\{\mathbf{\Delta}_k : k \leqslant l\}$ is the collection of all the ε – reducts of $\mathbf{C}$, i.e., $Red_D(\mathbf{C}) = \{\mathbf{\Delta}_1,\ldots,\mathbf{\Delta}_l\}$.*

Proof. For every $k = 1,\ldots,l$, we have $\mathbf{\Delta}_k \cap c_{ij} \neq \phi$. Since $f(U,\mathbf{C},D) = \vee_{k=1}^{l}(\wedge\mathbf{\Delta}_k)$, for every $\mathbf{\Delta}_k$, if we reduce an element C in $\mathbf{\Delta}_k$ ($\mathbf{\Delta}'_k = \mathbf{\Delta}_k - \{C\}$), then $f(U,\mathbf{C},D) \neq \vee_{r=1}^{k-1}(\wedge\mathbf{\Delta}_r) \vee (\wedge\mathbf{\Delta}'_k) \vee (\vee_{r=k+1}^{l}\mathbf{\Delta}_r)$ and $f(U,\mathbf{C},D) < \vee_{r=1}^{k-1}(\wedge\mathbf{\Delta}_r) \vee (\wedge\mathbf{\Delta}'_k) \vee (\vee_{r=k+1}^{l}\mathbf{\Delta}_r)$. If $\forall c_{ij}$, we have $\mathbf{\Delta}'_k \cap c_{ij} \neq \phi$, then $\wedge\mathbf{\Delta}'_k \leqslant \vee c_{ij}$, which implies

$$f(U,\mathbf{C},D) \geqslant \bigvee_{r=1}^{k-1}(\wedge\mathbf{\Delta}_r) \vee (\wedge\mathbf{\Delta}'_k) \vee \left(\bigvee_{r=k+1}^{l}\mathbf{\Delta}_r\right) \quad \text{and} \quad f(U,\mathbf{C},D) = \bigvee_{r=1}^{k-1}(\wedge\mathbf{\Delta}_r) \vee (\wedge\mathbf{\Delta}'_k) \vee \left(\bigvee_{r=k+1}^{l}\mathbf{\Delta}_r\right)$$

it is a contradiction. Hence there exists $c_{i_0j_0}$ such that $\mathbf{\Delta}'_k \cap c_{i_0j_0} = \phi$, which implies $\mathbf{\Delta}_k$ is a reduction of $(U,\mathbf{C},D)$.

For every $\mathbf{X} \in Red_D(\mathbf{C})$, we have $\mathbf{X} \cap c_{ij} \neq \phi$ for every $c_{ij} \neq \phi$. So we have $f(U,\mathbf{C},D) \wedge (\wedge\mathbf{X}) = \wedge(\vee c_{ij}) \wedge (\wedge\mathbf{X}) = \wedge\mathbf{X}$. This implies $\wedge\mathbf{X} \leqslant f(U,\mathbf{\Delta},D)$. Suppose that for every k we have $\mathbf{\Delta}_k - \mathbf{X} \neq \phi$. Then for every k one can find $C_k \in \mathbf{\Delta}_k - \mathbf{X}$. By rewriting $f(U,\mathbf{C},D) = (\vee_{k=1}^{l}C_k) \wedge \Phi$, we have $\wedge\mathbf{X} \leqslant \vee_{k=1}^{l}C_k$. So there is C_{k_0}, such that $\wedge\mathbf{X} \leqslant C_{k_0}$. This implies $C_{k_0} \in \mathbf{X}$, which is a contradiction. So $\mathbf{\Delta}_{k_0} \subseteq \mathbf{X}$ for some k_0, since both $\mathbf{X}$ and $\mathbf{\Delta}_{k_0}$ are reducts. We have $\mathbf{X} = \mathbf{\Delta}_{k_0}$. Hence $Red_D(\mathbf{C}) = \{\mathbf{\Delta}_1,\ldots,\mathbf{\Delta}_l\}$. □

Now we can conclude that $\mathbf{C}$ can be categorized into three parts according to their importance related to the classification: (1) elements in the core of reducts which should be included in every reduct; (2) elements cannot be included in any reduct; (3) elements belong to some but not all reducts. This partition also seems reasonable in the practical viewpoint.

It is worth pointing out that the proposed idea in this paper is not only limited to Gaussian kernels, but also applicable to all kernels mapping to the unit interval with 1 in its diagonal. We can develop fuzzy rough sets and consider attribute reduction with this kind of kernels. However, different kernels may have different techniques to perform attribute reduction. For example, we employ the algebraic product $T_P(x,y) = x \cdot y$ as the aggregation operator for Gaussian kernels, and in the existing attribute reducts [3,10,16–23,41,50] t-norm *Min* is employed as aggregation operator for fuzzy *Min* – similarity relations, this difference may lead to different formulation of discernibility matrixes. In addition, since a kernel plays the same role as a similarity measure in both attribute reduction and kernel trick, we suggest to use the same kernel when attribute reduction is employed as a preprocessing step of kernel trick.

4. Experiments and comparisons

In this section we will design an algorithm to compute reducts; we will also perform attribute reduction as a preprocessing step for Gaussian kernel support vector machines in order to test the effectiveness of the proposed work.

Table 1
Description of experimental data.

	Data	Samples	Features	Classes
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	Wine	178	13	3
8	Wpbc	198	33	2

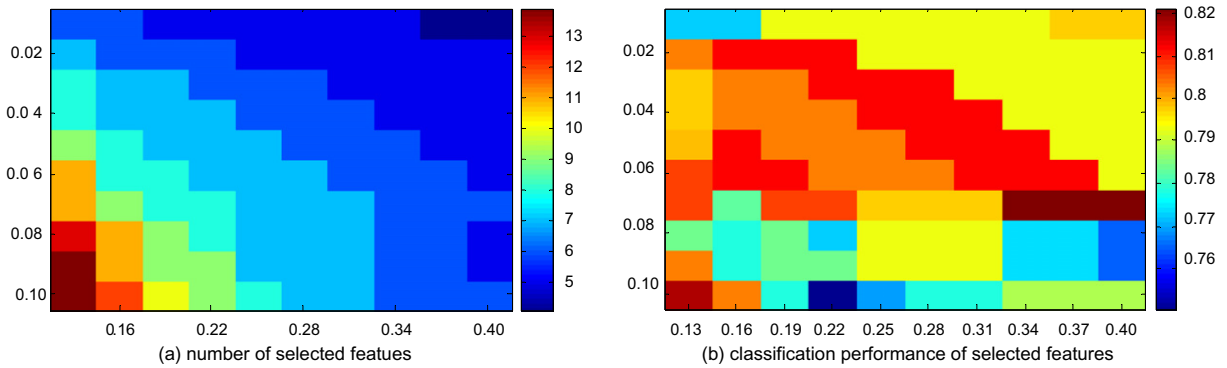

Fig. 1. Variation of size of selected features and corresponding classification performance (sonar).

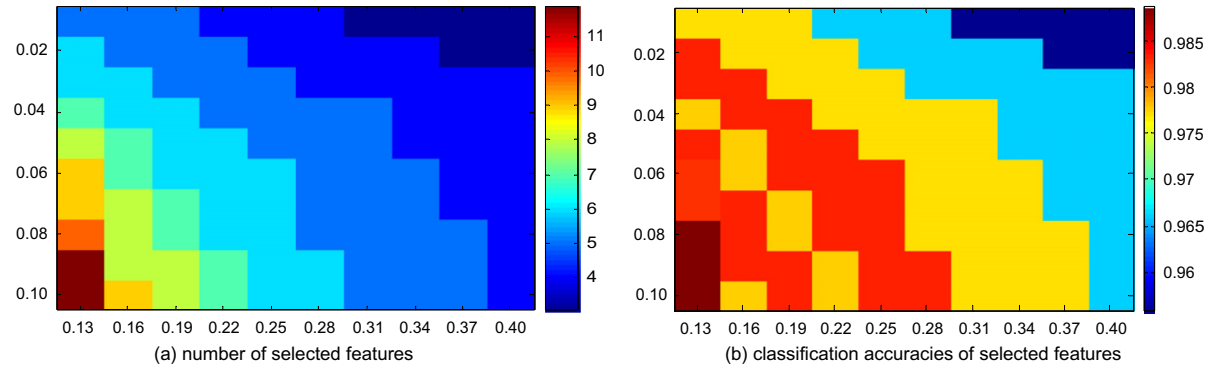

Fig. 2. Variation of size of selected features and corresponding classification performance (wine).

4.1. Algorithm design and complexity analysis

The algorithm by discernibility matrix is helpful to find all the reducts of the dataset, but the time complexity to find all the reducts increases exponentially with the number of attributes $O(|U|^2 \times 2^{|\mathbf{C}|})$ [40], where $|U|$ is the size of the universe, $|\mathbf{C}|$ is the number of conditional attributes.

In real applications, it is not necessary to find all the reducts. It is enough to address the real problem by using one of the reducts. In the following we provide a heuristic algorithm to find a reduct.

Input: $(U, \mathbf{C}, D)$, Reduct $\leftarrow \{\}$
Step 1: Compute the similarity relation of the set of all condition attributes: R_G^n;
Step 2: Compute $Pos_{\mathbf{C}}(D) = \cup_{t=1}^{s} \underline{R_G^n} D_t$;
Step 3: Compute c_{ij} by its definition in Section 3;
Step 4: Compute $Core_D(\mathbf{C}) = \cup \{\mathbf{Q}_{ij} \subseteq \mathbf{C} : \mathbf{Q}_{ij} = \cap\{\mathbf{P} : \wedge\mathbf{P} \in c_{ij}\}, i,j = 1,2,\ldots,m\}$; Delete those c_{ij} with nonempty overlap with$Core_D(\mathbf{C})$;
Step 5: Let Reduct = $Core_D(\mathbf{C})$;
Step 6: Add the element a whose frequency of occurrence is maximum in all c_{ij} into Reduct; and delete those c_{ij} with nonempty overlap with Reduct;
Step 7: If there still exist some $c_{ij} \neq \phi$, go to Step 6; Otherwise, go to Step 8;
Step 8: If Reduct is not independent, delete the redundant elements in Reduct;
Step 9: Output Reduct.

The computational complexity of this algorithm is $O(|U|^2 \times |\mathbf{C}|)$.

4.2. Experimental analysis

In this subsection, we will perform experiments to examine effectiveness of our idea. We select Gaussian kernel SVM as a classifier to validate the quality of the features selected by our technique.

Eight datasets are downloaded from UCI machine learning repository [33], described in Table 1.

First, we consider the impact of parameters on feature selection. We set σ from 0.1 to 0.4 with step 0.03. In the meanwhile, ε is set as 0.01 to 0.1 with step 0.01. With these parameters, we can get 100 subsets of attributes and the corresponding classification performance. We perform experiments on data sets *sonar* and *wine*. The results are shown in Figs. 1 and 2, where the x-axis is σ and y-axis isε.

As the objective of feature selection is to find a minimal subspace which has good classification performance, so it is expected that the size of the selected feature is relatively small and the corresponding classification performance is good enough. We can see from the above results that [0.1,0.2] and [0.01,0.02] are proper value domains for σ and ε, respectively.

Table 2
Numbers of selected features.

Data	Raw data	KFRS	CFS	NRS	RS
Credit	15	12	8	12	11
Heart	13	11	10	12	0
Hepatitis	19	12	6	11	12
Horse	22	8	7	8	4
Iono	34	20	4	18	8
Sonar	60	9	9	16	0
Wine	13	8	5	9	4
Wpbc	33	13	3	10	7

Table 3
Classification accuracies based on Gaussian kernel SVM (%).

Data	Raw data	KFRS	CFS	NRS	RS
Credit	81.44 ± 7.18	85.63 ± 18.5	85.48 ± 18.51	85.48 ± 18.5	85.48 ± 18.5
Heart	81.11 ± 7.50	85.93 ± 6.25	84.44 ± 6.00	83.33 ± 6.59	–
Hepatitis	83.50 ± 5.35	90.83 ± 6.54	91.50 ± 6.40	89.00 ± 4.46	85.00 ± 7.24
Horse	72.30 ± 3.63	91.84 ± 4.05	90.76 ± 4.82	87.24 ± 3.61	89.11 ± 4.45
Iono	93.79 ± 5.08	95.19 ± 4.03	87.84 ± 5.39	87.26 ± 6.06	83.30 ± 5.97
Sonar	85.10 ± 9.49	87.50 ± 6.89	76.52 ± 7.10	74.05 ± 7.60	–
Wine	98.89 ± 2.34	98.89 ± 2.34	95.49 ± 3.54	97.22 ± 2.93	95.00 ± 4.10
Wpbc	77.37 ± 7.73	81.89 ± 5.71	76.32 ± 3.04	80.37 ± 5.33	78.37 ± 5.06

We can see that the classification accuracies of the reduced data are relatively high and the sizes of the reduced data are small if we let σ and ε take values in these domains, respectively.

Now we compare the number of the selected features and classification performances of the reducts, shown in Tables 2 and 3, where reduct is computed by the proposed algorithm and the classification performances of reducts are attained with Gaussian kernel SVM based on the 10-fold cross validation technique and Gaussian kernel SVM is implemented with osu_svm3.00 toolbox. δ and ε are specified as 0.1 and 0.02, respectively.

Now we analyze experimental results in Tables 2 and 3. Compared with the raw data, we see that (i) among the 8 data sets, our proposed attribute reduction method performs well on six data sets: hepatitis, horse, wpbc, anneal, iono, sonar and wine. For these six data sets, numbers of attributes greatly decrease after reduction compared with the raw data, and performances of classifier (SVM) are improved distinctly. This implies our proposed attribute reduction method can really delete redundant attributes from these data sets; (ii) for data sets credit and heart, few attributes are deleted, and improvements of performances of classifier are not significant. However, this may due to that there are less redundant attributes in these two data sets since the original numbers of attributes in these two data sets are few.

In order to compare the proposed techniques with the existing one, we use neighborhood rough set approach (NRS) [18] and correlation based feature selection (CFS) [14] on these data sets. These techniques can deal with numerical features directly. From Tables 2 and 3, we can also see that fuzzy rough sets are better than other algorithms inmost cases. In addition, we also introduce the classical rough set technique to select features with a forward greedy search strategy, denoted by RS. As to datasets heart and sonar, no feature is returned. This phenomenon has been mentioned in the previous work as any single feature produces the dependency of zero. So the algorithm stops here.

In addition, we gather six cancer recognition tasks outlined in Table 4. The numbers of features are much more than the numbers of samples in these tasks. The detailed description about these tasks can be gotten from the webpage (http://www.gems-system.org/). Overfitting is the most important challenge in gene classification. Attribute reduction may help overcome this problem. We perform attribute reduction based on techniques of neighborhood rough sets and fuzzy rough sets, respectively. The results are presented in Tables 5 and 6. We see that most of candidate genes are removed from classification learning and only a few genes are selected. Moreover, the genes selected by FRS are a little more than that by NRS; however; the classification performance is greatly improved by FRS compared with the raw data and those selected by NRS. These results show fuzzy rough sets are useful in gene selection for cancer recognition.

Table 4
Gene expression data sets.

Data	Genes	Class	Samples
Leuk1	7129	3	72
Leuk2	12,582	3	72
SRBCT	2308	4	83
Breast	9216	5	84
Lung2	12,600	5	203
DLBCL	4026	6	88

Table 5
Number of the features selected.

Data	Raw	NRS	FRS
Breast	9216	5	13
DLBCL	4026	4	15
Leukemial1	7129	2	4
Leukemial2	12582	2	9
Lung2	12600	5	14
SRBCT	2308	3	12

Table 6
Accuracy of the selected genes.

Data	Raw (%)	NRS (%)	FRS (%)
Breast	44.05	72.08	100.0
DLBCL	87.50	76.99	99.00
Leukemial1	54.17	88.87	97.32
Leukemial2	39.62	91.71	94.28
Lung2	69.83	80.31	90.56
SRBCT	73.86	76.23	82.05

5. Conclusion and future work

Fuzzy rough sets are a hot topic in granular computing. In this paper we introduce Gaussian kernel into fuzzy rough sets for computing fuzzy similarity relation and develop a novel method of attribute reduction with parameter based on the proposed model.

We discuss the structure of subsets of selected attributes with fuzzy discernibility matrix. Attributes can be grouped as three collections according to their importance related to the decision. The main purpose of this paper is to develop attribute reduction with kernel tricks. We use the UCI machine learning data sets and cancer classification tasks to test the proposed technique. The experimental results shows Gaussian kernel based fuzzy rough sets can find good subsets of attributes for classification learning.

Although Gaussian kernel is frequently used, there are also some other kernel functions can be introduced into fuzzy rough sets. We will work on other kernels and develop a set of attribute reduction techniques based on fuzzy rough sets and kernels.

Acknowledgements

This paper is partly supported by National Natural Science Foundation under Grants 70871036, 60703013, and 10978011 and a grant of National Basic Research Program of China (2009CB219801-3).

References

[1] M.F. Balcan, A. Blum, S. Vempala, Kernels as features: on kernels, margins, and low-dimensional mappings, Machine Learning 65 (2006) 79–94.
[2] O. Barzilay, V.L. Brailovsky, On domain knowledge and feature selection using a support vector machine, Pattern Recognition Letters 20 (1999) 475–484.
[3] R.B. Bhatt, M. Gopal, On fuzzy rough sets approach to feature selection, Pattern recognition Letters 26 (2005) 965–975.
[4] P.S. Bradley, O.L. Mangasarian, Feature selection via concave minimization and support vector machine, in: Proceedings of the 15th International Conference on Machine Learning, San Francisco, CA, USA, 1998, pp. 82–90.
[5] J.H. Chen, C.S. Chen, Fuzzy kernel perceptron, IEEE Transactions on Neural Networks 13 (2002) 1364–1373.
[6] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (1990) 191–209.
[7] D. Dubois, H. Prade, A review of fuzzy set aggregation connectives, Information Sciences 36 (1985) 85–121.
[8] R. Duda, P. Hart, D. Stork, Pattern Classification, second ed., John Wiley & Sons, New York, NY, USA, 2000.
[9] T. Evgeniou, M. Pontil, C. Papageorgiou, T. Poggio, Image representations and feature selection for multimedia database search, IEEE Transactions on Knowledge and Data Engineering 15 (2003) 911–920.
[10] S. Fernandez, J.M. Murakami, Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations, Fuzzy Sets and Systems 139 (2003) 635–660.
[11] S. Gottwald, Fuzzy Sets and Fuzzy Logic, Vieweg, Braunschweig, 1993.
[12] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[13] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.
[14] M. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the 17th ICML, CA, 2000, pp. 359–366.
[15] C.L. Huang, C.J. Wang, A GA-based feature selection and parameters optimization for support vector machines, Expert Systems with Applications 31 (2006) 231–240.
[16] 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.
[17] 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.
[18] Q.H. Hu, D.R. Yu, J. F Liu, C. X Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.
[19] Q.H. Hu, L. Zhang, D.G. Chen, W. Pedrycz, D. Yu, Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications, International Journal of Approximating Reasoning 51 (2010) 453–471.
[20] R. Jensen, Q. Shen, Fuzzy-rough attributes reduction with application to web categorization, Fuzzy Sets and Systems 141 (2004) 469–485.
[21] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15 (2007) 73–89.
[22] 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.
[23] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Transactions on Fuzzy Systems 17 (2009) 824–838.
[24] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Proceedings of the 11th International Conference on Machine Learning, 1994, pp. 121–129.
[25] E.P. Klement, R. Mesiar, E. Pap, Triangular norms, Trends in Logic, vol. 8, Kluwer Academic Publishers, Dordrecht, 2000.
[26] J. Kohavi, Wrappers for feature subset selection, AIJ issue on relevance, 1995
[27] Y. Liu, Y.F. Zheng, FS-SFS: a novel feature selection method for support vector machines, Pattern Recognition 39 (2006) 1333–1345.
[28] K.Z. Mao, Feature subset selection for support vector machines through discriminative function pruning analysis, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 34 (2004) 60–67.
[29] J.S. Mi, W.X. Zhang, An axiomatic characterization of a fuzzy generalization of rough sets, Information Sciences 160 (2004) 235–249.
[30] J.S. Mi, Y. Leung, H.Y. Zhao, Generalized fuzzy rough sets determined by a triangular norm, Information Sciences 178 (2008) 3203–3213.
[31] N.N. Morsi, M.M. Yakout, Axiomatics for fuzzy rough sets, Fuzzy Sets and Systems 100 (1998) 327–342.
[32] B. Moser, On representing and generating kernels by fuzzy equivalence relations, Journal of Machine Learning Research 7 (2006) 2603–2620.
[33] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI Repository of machine learning databases, University of California, Department of Information and Computer Science, Irvine, CA, 1998. <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
[34] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Machine Learning 61 (2005) 129–150.
[35] Z. Pawlak, Rough sets, International Journal of Computer Information Science 11 (1982) 341–356.
[36] A.M. Radzikowska, E.E. Kerre, A comparative study of fuzzy rough sets, Fuzzy Sets and Systems 126 (2002) 137–155.
[37] B. Scholkopf, A.J. Smola, Learning with Kernels, The MIT Press, 2002.
[38] B. Schweizer, A. Sklar, Associative functions and statistical triangle inequalities, Publicationes Mathematicae – Debrecen 8 (1961) 169–186.
[39] B. Schweizer, A. Sklar, Probabilistic Metric Spaces, North-Holland, Amsterdam, 1983.

[40] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision support, Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, 1992.
[41] C.C.E. Tsang, D.G. Chen, S.D. Yueng, W.T.J. Lee, X.Z. Wang, Attribute reduction using fuzzy rough sets, IEEE Transactions on Fuzzy Systems 16 (2008) 1130–1142.
[42] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[43] P. Vincent, Y. Bengio, Kernel matching pursuit, Machine Learning 48 (2002) 165–187.
[44] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, vol. 13, MIT Press, Cambridge, MA, USA, 2001, pp. 668–674.
[45] W.Z. Wu, W.X. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Information Sciences 159 (2004) 233–254.
[46] W.Z. Wu, J.S. Mi, W.X. Zhang, Generalized fuzzy rough sets, Information Sciences 151 (2003) 263–282.
[47] W.Z. Wu, Attribute reduction based on evidence theory in incomplete decision systems, Information Sciences 178 (2008) 1355–1371.
[48] S.D. Yeung, D.G. Chen, C.C.E. Tsang, W.T.J. Lee, X.Z. Wang, On the generalization of fuzzy rough sets, IEEE Transactions on Fuzzy Systems 13 (2005) 343–361.
[49] D.R. Yu, Q.H. Hu, C.X. Wu, Uncertainty measures on fuzzy relations and their applications, Applied Soft Computing 7 (2007) 1135–1143.
[50] S.Y. Zhao, E.C.C. Tsang, On fuzzy approximation operators in attribute reduction with fuzzy rough sets, Information Sciences 178 (2008) 3163–3176.
[51] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani, 1-Norm support vector machines, in: S. Thrun, L. Saul, B. Scholkopf (Eds.), Advances in Neural Information Processing Systems, vol. 13, MIT Press, Cambridge, MA, USA, 2004.