Neurocomputing 74 (2011) 1585–1594

Contents lists available at ScienceDirect

Neurocomputing

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

Neighborhood based sample and feature selection for SVM classification learning

Qiang He [a], Zongxia Xie [b], Qinghua Hu [c,*], Congxin Wu [a]

[a] Department of Mathematics, Harbin Institute of Technology, 150001 Harbin, PR China
[b] Harbin Institute of Technology, 150001 Harbin, PR China
[c] Power Engineering, Harbin Institute of Technology, 150001 Harbin, PR China

ARTICLE INFO

Article history:
Received 4 August 2010
Received in revised form
7 January 2011
Accepted 8 January 2011
Communicated by D. Wang
Available online 21 March 2011

Keywords:
Support vector machine
Rough set
Neighborhood relation
Sample selection
Feature selection

ABSTRACT

Support vector machines (SVMs) are a class of popular classification algorithms for their high generalization ability. However, it is time-consuming to train SVMs with a large set of learning samples. Improving learning efficiency is one of most important research tasks on SVMs. It is known that although there are many candidate training samples in some learning tasks, only the samples near decision boundary which are called support vectors have impact on the optimal classification hyper-planes. Finding these samples and training SVMs with them will greatly decrease training time and space complexity. Based on the observation, we introduce neighborhood based rough set model to search boundary samples. Using the model, we firstly divide sample spaces into three subsets: positive region, boundary and noise. Furthermore, we partition the input features into four subsets: strongly relevant features, weakly relevant and indispensable features, weakly relevant and superfluous features, and irrelevant features. Then we train SVMs only with the boundary samples in the relevant and indispensable feature subspaces, thus feature and sample selection is simultaneously conducted with the proposed model. A set of experimental results show the model can select very few features and samples for training; in the mean time the classification performances are preserved or even improved.

1. Introduction

In the last decade, we are witnessing great success of support vector machines (SVMs) in a lot of theoretical research [5,7,22,26,30,33] and practical applications [9,23,25,31]. However, SVM learning algorithms suffer from exceeding time and memory requirements if the training pattern set is very large because the algorithm requires solving a quadratic programming (QP) with time complexity of order $O(M^3)$ and space complexity $O(M^2)$, where M is the number of training samples [4,29]. In order to deal with the large scale quadratic programming, the decomposition methods were proposed to divide the large QP problem into a sequence of small problems so the memory difficulty is avoided [5,10,14,24]. However, for huge problems with many support vectors, the decomposition method still suffers from slow convergence [17,32].

In Ref. [1], Cortes and Vapnik showed that the weights of optimal classification hyper-plane in feature space can be represented as linear combination of support vectors, which shows optimal hyper-plane is independent of other training samples except support vectors. One can select only a part of the samples, so-called support vectors, to train SVM, rather than the whole training sets if they can be found in advance. By this way, the learning time and space complexity will be greatly reduced [21,27]. Based on this observation, some researches were reported to select patterns for SVM. Lee and Mangasarian [16] chose a random subset of the original samples and then learning classification plane with the subset. The power of the method strongly depends on the quality of resampling, and there still are some irrelevant samples in the random subset. Furthermore, it is not clear how many samples should be included in the random subset. Almeida et al. [1] grouped the training samples into some clusters with k-means clustering, and the clusters with homogeneous class are replaced with the centroid of the cluster. It is difficult to specify the number of clusters in a complex learning task in applications. Koggalage and Halgamuge [15] showed another clustering based sample selection algorithm for SVM, where they assumed the cluster centers were known in advance. In Ref. [11] Hoegaerts et al. introduced a subset based least squares subspace regression method, where some feature extraction algorithms were conducted to construct subspace regression algorithms, such as kernel partial least squares and kernel

* Corresponding author.
E-mail address: huqinghua@hit.edu.cn (Q. Hu).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.01.019

canonical correlation analysis. However, this work did not consider boundary sample extraction and classification problems. Shin and Cho proposed a neighborhood property based pattern selection algorithm, where neighborhood entropy was defined. They associated each samples with k nearest neighbors, then checked the entropy of the neighborhood. If the entropy is not zero, then the samples are considered as boundary set [28]. Furthermore, they gave the proof that neighborhood relation between training samples in input space is preserved in the feature space [29]. This lays a firm groundwork for boundary sample selection.

Neighborhood is a kind of important topological property of sample spaces. In fact, neighborhood relations and neighborhood properties were used to extend Pawlak's rough set model [18–20,34–36], where each object in the universe is assigned with a subset of objects. The objects in the neighborhood are near the center object measured with some distance function. The subset is called a neighborhood information granule. The family of neighborhood granules forms a cover of the object space. Arbitrary subset of the universe, called a concept, can be approximated with part of the neighborhood granules. The lower approximation of the concept is the maximal union of objects whose neighborhoods completely belong to the concept, while upper approximation is the minimal union of objects whose neighborhoods completely contain the concept. The difference of the lower and upper approximations is called boundary. Connecting the definition of boundary in neighborhood model and that presented in Refs. [28,29], we can find that they refer to the same nature but present different forms. However, neighborhood rough set model presents a more soundly and systematically theoretical framework about this problem. Furthermore, the neighborhood model can conduct feature subset selection as well as sample selection in one framework [12]. Feature subset selection is an effective technique to improve generalization and reduce classification cost [8,13,23,37]. In this paper, we will introduce neighborhood rough set model and construct algorithms for simultaneous feature and sample selection for training support vector machine. The contributions of the work are twofold. On the one hand, we show a systematic solution to feature selection and boundary sample discovery for learning SVM, which shows the essential characteristic of boundary samples in classification problems. On the other hand, we present an application of neighborhood rough sets.

The rest of the paper is organized as follows. Neighborhood rough set model and its properties are shown in Section 2. Feature and sample selection algorithms are constructed in Section 3. Section 4 presents some experiments with artificial data and UCI data sets. Finally, the conclusion comes in Section 5.

2. Neighborhood based rough set model

In this section we mainly review the basic concepts and theoretical results of neighborhood based rough set model. More details can be found in Ref. [12]. Both rough sets and SVM deal with learning problems with structural data. Formally, the data can be written as a $IS=\langle U,C,D\rangle$, where U is the nonempty set of samples $\{x_1,x_2,\ldots,x_n\}$, called a universe or sample space, C is the set of input variables $\{c_1,c_2,\ldots c_m\}$, or condition attributes, D is the output, also called decision. We call $IS=\langle U,C,D\rangle$ a decision system.

Definition 1. Given arbitrary $x_i\in U$ and $B\subseteq C$, the neighborhood $\delta_B(x_i)$ of x_i in the subspace B is defined as

$$\delta_B(x_i)=\{x_j|x_j\in U,\Delta_B(x_i,x_j)\le\delta\},$$

where Δ is a metric function, $\forall x_1,x_2,x_3\in U$, which satisfies:

1) $\Delta(x_1,x_2)\ge 0$, $\Delta(x_1,x_2)=0$, if and only if $x_1=x_2$;
2) $\Delta(x_1,x_2)=\Delta(x_2,x_1)$;
3) $\Delta(x_1,x_3)\le\Delta(x_1,x_2)+\Delta(x_2,x_3)$.

Obviously, neighborhood relations are one class of similarity relations, which satisfy reflexivity and symmetry. Neighborhood relations draw the objects together for similarity or indistinguishability in terms of distances.

Definition 2. Consider a metric space $\langle U,\Delta\rangle$, N is a neighborhood relation on U, $\{\delta(x_i)|x_i\in U\}$is the family of neighborhood granules. Then we call $\langle U,\Delta,\mathrm{N}\rangle$ a neighborhood approximation space.

Definition 3. Given neighborhood approximation space $\langle U,\Delta,N\rangle$, $X\subseteq U$, two subsets of objects, called lower and upper approximations, are defined as

$$\underline{\mathrm{N}}X=\{x_i|\delta(x_i)\subseteq X,x_i\in U\},$$

$$\overline{\mathrm{N}}X=\{x_i|\delta(x_i)\cap X\neq\emptyset,x_i\in U\}.$$

Theorem 1. Given a neighborhood approximation space $\langle U,\Delta,\mathrm{N}\rangle$, we have

1) $\forall X\subseteq U:\ \underline{\mathrm{N}}X\subseteq X\subseteq\overline{\mathrm{N}}X$;
2) $\underline{\mathrm{N}}\emptyset=\overline{\mathrm{N}}\emptyset=\emptyset$, $\underline{\mathrm{N}}U=\overline{\mathrm{N}}U=U$;
3) $\forall X,Y\subseteq U:\ X\subseteq Y\Rightarrow\underline{\mathrm{N}}X\subseteq\underline{\mathrm{N}}Y;\ \overline{\mathrm{N}}X\subseteq\overline{\mathrm{N}}Y$.

Proof. Straightforward.

Theorem 2. (*Hu et al. [12]*). Given $\langle U,\Delta,\mathrm{N}\rangle$ and two non-negative δ_1 and δ_2, if $\delta_1\ge\delta_2$, we have

1) $\forall x_i\in U:\ \mathrm{N}_1\supseteq\mathrm{N}_2,\ \delta_1(x_i)\supseteq\delta_2(x_i)$;
2) $\forall X\subseteq U:\ \underline{\mathrm{N}_1}X\subseteq\underline{\mathrm{N}_2}X;\overline{\mathrm{N}_2}X\supseteq\overline{\mathrm{N}_1}X$,

where N_1 and N_2 are the neighborhood relations induced with δ_1 and δ_2, respectively.

Definition 4. Given a neighborhood decision table $NDT=\langle U,A\cup D,V,f\rangle$, $X_1,X_2,\cdots,X_N$ are the object subsets with decisions 1 to N, $\delta_B(x_i)$ is the neighborhood information granules including x_i and generated by attributes $B\subseteq A$, Then the lower and upper approximations of the decision D with respect to attributes B are defined as

$$\underline{\mathrm{N}_B}D=\bigcup_{i=1}^{N}\underline{\mathrm{N}_B}X_i,$$

$$\overline{\mathrm{N}_B}D=\bigcup_{i=1}^{N}\overline{\mathrm{N}_B}X_i,$$

where

$$\underline{\mathrm{N}_B}X=\{x_i|\delta_B(x_i)\subseteq X,x_i\in U\},$$

$$\overline{\mathrm{N}_B}X=\{x_i|\delta_B(x_i)\cap X\neq\emptyset,x_i\in U\}.$$

The decision boundary region of D with respect to attributes B is defined as

$$BN(D)=\overline{\mathrm{N}_B}D-\underline{\mathrm{N}_B}D.$$

Decision boundary is the object subset whose neighborhoods come from more than one decision class. On the other hand, the lower approximation of decision, also called *positive region* of

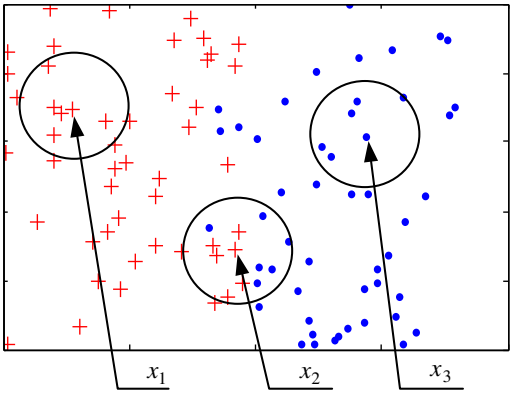


Fig. 1. An example with two classes.

decision, denoted by $POS_B(D)$, is the subset of objects whose neighborhoods consistently belong to one of the decision classes.

Example 1. Fig. 1 shows an example of binary classification in 2-D space, where d_1 is labeled with "plus" and d_2 is labeled with "point". Consider samples x_1, x_2, x_3, we assign circle neighborhoods to these samples. We can find $\delta(x_1)\subseteq d_1$ and $\delta(x_3)\subseteq d_2$, while $\delta(x_2)\cap d_1\neq\emptyset$, $\delta(x_2)\cap d_2\neq\emptyset$. According to the above definitions: $x_1\in \underline{N}d_1$, $x_3\in \underline{N}d_2$ and $x_2\in BN(D)$. Support vectors probably lie in the decision boundary region. Therefore, we can select the boundary samples defined in neighborhood rough set model to train SVM.

The samples in different feature subspaces will have different boundary regions. The size of boundary region reflects the discriminability of the classification problem in the corresponding subspaces. It also reflects the recognition power or characterizing power of the condition attributes. More specifically, the greater the boundary region is, the weaker the characterizing power of the condition attributes is. It can be formulated as follows.

Definition 5. The dependency degree of D to B is defined as the ratio of consistent objects

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

$\gamma_B(D)$ reflects the ability of B to approximate D. Obviously, $0\leq\gamma_B(D)\leq 1$. We say D completely depends on B if $\gamma_B(D)=1$, denoted by $B\Rightarrow D$; otherwise we say $D\gamma$ *depends* on B, denoted by $B\Rightarrow_r D$.

Theorem 3. (*Hu et al. [12]*). $\langle U, A, D\rangle$ is a neighborhood decision system; $B_1,B_2\subseteq A$, $B_1\subseteq B_2$, then we have

1) $N_{B1}\supseteq N_{B2}$;
2) $\forall X\subseteq U$, $N_{B_1}X\subseteq N_{B_2}X$, $\overline{N_{B_1}}X\supseteq\overline{N_{B_2}}X$;
3) $POS_{B_1}(D)\leq POS_{B_2}(D)$, $\gamma_{B_1}(D)\leq\gamma_{B_2}(D)$.

Theorem 3 shows dependence monotonously increases with attributes, which means that adding a new attribute in the attribute subset at least does not decrease the dependence. The property is very important for constructing feature selection algorithms. Generally speaking, we hope to find a minimal feature subset which has the same characterizing power as the whole set of original features.

Definition 6. Given a neighborhood decision table $NDT=\langle U,A,D\rangle$, $B\subseteq A$, we say attribute subset B is a relative reduct if

1) $\gamma_B(D)=\gamma_A(D)$;
2) $\forall a\in B,\gamma_B(D)>\gamma_{B-a}(D)$.

The first condition guarantees that $POS_B(D)=POS_A(D)$. The second condition shows there is no superfluous attribute in the reduct. Therefore, a reduct is the minimal subset of attributes which has the same approximating power as the whole attribute set. This definition will guide us to find optimal feature subsets.

There are usually multiple reducts in an information system. In other words, we can find more than one subset of features which has the same prediction capability as the whole features, and each reduct presents a point of view to understand the classification problem.

Let $\langle U,A,D\rangle$ be a decision table and $\{B_j|j\leq r\}$ is the set of reducts, we denote the following attribute subsets:

$$Core=\bigcap_{j\leq r}B_j,\quad K=\bigcup_{j\leq r}B_j-Core,\quad K_j=B_j-Core\quad I=A-\bigcup_{j\leq r}B_j.$$

Definition 7. *Core* is the attribute subset of *strong relevance*, which cannot be deleted from any reduct, otherwise the prediction power of the system will decrease. Namely, $\forall a\in Core$, $\gamma_{A-a}(D)<\gamma_A(D)$. Therefore the core attributes will be in all of the reducts. I is the *completely irrelevant* attribute set. The attribute in I will not be included in any reduct, which means I is completely useless in the system. K_j is a *weak relevant attribute set*. The union of *Core* and K_j forms a reduct of the information system. Given a feature subset $B=core\cup k_i$, then $\forall a\in k_j, j\neq i$, is said to be redundant.

Neighborhood rough set model discover the structures of sample spaces and feature spaces. It segments the sample set into two classes: positive region and boundary. Generally speaking, samples in boundary region provide more information for classification learning. Support vector machines use the samples near the decision boundary, namely, support vectors, to support classification hyper-plane. On the other hand, neighborhood rough set model divides the feature set into four classes: strongly relevant, weakly relevant and indispensable, weakly relevant and redundant, and irrelevant features. This will put the two tasks of sample selection and feature selection in one framework and they can be implemented in parallel.

Training SVM just with the boundary samples in the reduced attribute subspace will speed up the learning process, improve generalization power of trained classifiers and reduce the cost in measuring and storing data. The following section will present the algorithms to search reducts and discover boundary samples.

3. Algorithm design

In this section we will construct two algorithms for feature selection and boundary sample discovery, respectively. First we find a feature subset based on the neighborhood rough set model with the proposed algorithm. Then we search boundary samples in the reduced subspaces.

3.1. Feature selection based on neighborhood model

As mentioned previously, the motivation of rough set based feature selection is to select a minimal attribute subset, which has the same characterizing power as the whole attribute set and without any redundant attribute. In other word, the dependency of the selected attributes is the same as the original attributes, and the dependence will decrease if any selected attribute is deleted.

Definition 8. Given a decision system $\langle U,A,D\rangle$, $B\subseteq A$, $a\in B$, we define the significance of an attribute as

$$SIG(a,B,D)=\gamma_B(D)-\gamma_{B-a}(D).$$

The attribute's significance is the function of three variables: a, B and D. An attribute a may be of great significance in B_1but of little significance in B_2. Furthermore, the attribute's significance is

different for each decision if there are multiple decision attributes in a decision table.

Definition 9. We say attribute a is *superfluous* in B with respect to D if $SIG(a,B,D)=0$, otherwise a is indispensable. We say B is dependent if $\forall a \in B$, a is indispensable.

From another standpoint, we also can define the significance of an attribute as follows.

Definition 10. Given a decision system $\langle U,A,D\rangle$, $B\subseteq A$, $a\notin B$, the significance of an attribute is

$$SIG(a,B,D)=\gamma_{B\cup a}(D)-\gamma_B(D).$$

It is a combinational optimization problem to find all of the reducts. There are $2^{|A|}$ combinations of attribute subsets. It is not practical to search all of the reducts in $2^{|A|}$ combinations. Fortunately, in practice, we usually just require one of the reducts to train a classifier, and we do not care much whether the reduct is really the minimal one or not. Then a tradeoff solution can be constructed, such as greedy forward search algorithm.

Algorithm. Forward attribute reduction based on neighborhood model (FARNeM)

Input: $\langle U,A,d\rangle$ and δ //δ is the threshold to control the size of neighborhood
Output: reduct *red*
Step 1: $\emptyset \rightarrow red$; // *red* is the pool to contain the selected attributes
Step 2: For each $a_i \in A-red$
compute $SIG(a_i,B,D)=\gamma_{red\cup ai}(D)-\gamma_{red}(D)$, // Here we define $\gamma_\emptyset(D)=0$
end
Step 3: select the attribute a_k which satisfies:
$SIG(a_k,B,D)=\max_i(SIG(a_i,red,B))$
Step 4: if $Sig(a_k,B,D)>0$,
$red\cup a_k \rightarrow red$
go to step2
else
return *red*
Step 5: end

Here the FARNeM algorithm adds an attribute with the great increment of dependence into the reduct in each circle until the dependence does not increase, namely, adding any new attribute will not increase the dependence in this case. The time complexity of the algorithm is $O(N\times N)$, where N is the number of candidate attributes.

In fact, neighborhood rough set model just presents some ways to measure significance of attributes. They are independent of attribute search strategies. Other search strategies, such as GA, Branch & Bound, also can be introduced. Detailed analysis on search strategies can be found in Ref. [6]. In this work, our focus is not to compare these search strategies.

3.2. Boundary sample selection

In fact, FARNeM has to find the positive-region samples for evaluating the significance of attributes in step 2. According to the property showed in Section 2, we know $BN(D)=U-POS_B(D)$. However, the aim of FARNeM is to find feature subset which can distinguish the samples. It is different from discovering boundary samples. In order to well support classification hyper-plane, one requires a set of boundary samples with an appropriate size. Too few boundary samples are not enough to support the optimal hyper-plane. Therefore, on one hand, we should delete most of the samples in the positive region; on the other hand, we should keep enough samples near the decision boundary to support the optimal hyper-plane. Intuitively, if we increase the threshold δ, then $\delta(x_i)$ will include more samples. Then $\delta(x_i)$ will more probably belong to the decision boundary. Hence, the decision boundary will expand with threshold δ. The boundary set will be empty if $\delta=0$ and the learning task is consistent. In the other end, all of the training samples will be in boundary if $\delta=\infty$. So it is critical to select a proper threshold δ.

The second problem is to recognize and reduce noise samples in training set.

Example 2. As shown in Fig. 2(1), there are some noise samples in the data, such as x_1 and x_2, we can find all of the samples near x_1 belong to class one except itself, while the neighborhood of x_2 belong to class 2 except itself, which shows that x_1 and x_2 are probably mislabeled. Figs. 2(2), (3), (4) present the boundary samples discovered with neighborhood rough set model, where $\delta=0.1$, $\delta=0.2$, $\delta=0.4$, respectively. Obviously the boundary region increases with threshold δ. And we also find some of the samples far away from the decision boundary are also included in the boundary set due to the noise samples. These samples will have negative effect on the classification hyper-plane. They should be deleted from the boundary set.

Here we introduce the idea of variable precision rough sets to deal with this problem [38].

Definition 11. Given two sets X and Y, we define the inclusion degree of X in Y as

$$I(X,Y)=|X\cap Y|/|X|$$

Definition 12. Given a family of neighborhood information granules $\delta(x_i)$, $i=1,2,\ldots,n$, $X\subseteq U$, the variable precision lower approximation and upper approximation is defined as

$$\underline{N^\beta}X=\{x_i|I(\delta_B(x_i),X)\geq\beta,x_i\in U\},$$

$$\overline{N^\beta}X=\{x_i|I(\delta_B(x_i),X)>1-\beta,x_i\in U\},$$

where $\beta>0.5$.

Definition 12 relaxes the condition of strict inclusion or strict exclusion in Definition 3, replacing them with majority inclusion and majority exclusion. Based on the variable precision neighborhood model, x_1,x_2 and their neighborhood will be classified into the positive regions of class 1 and class 2, respectively.

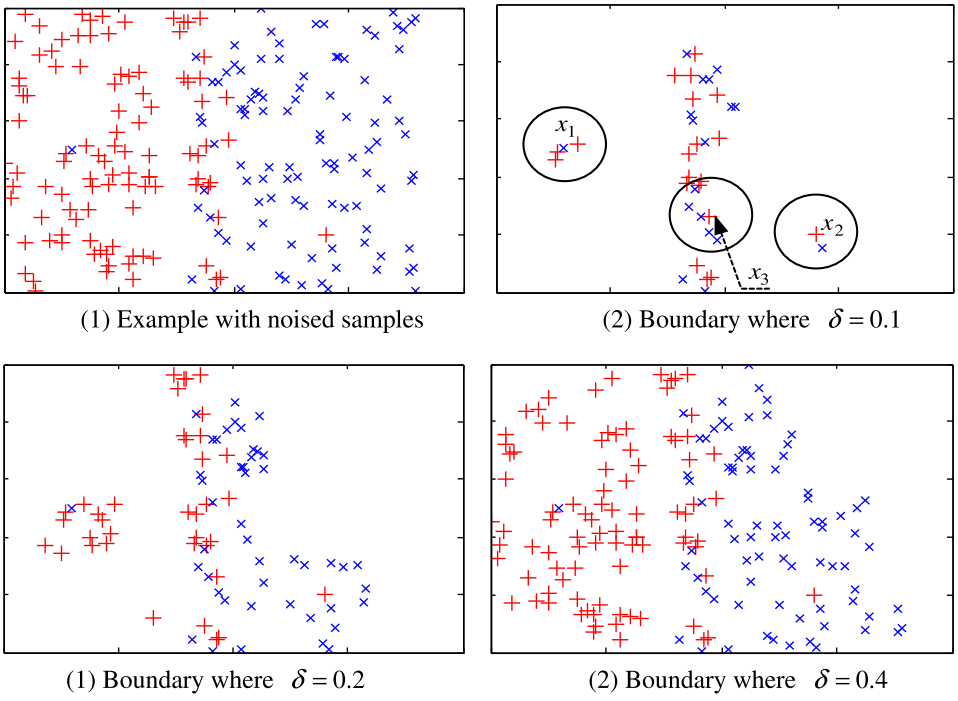


Fig. 2. An example data with noised samples.

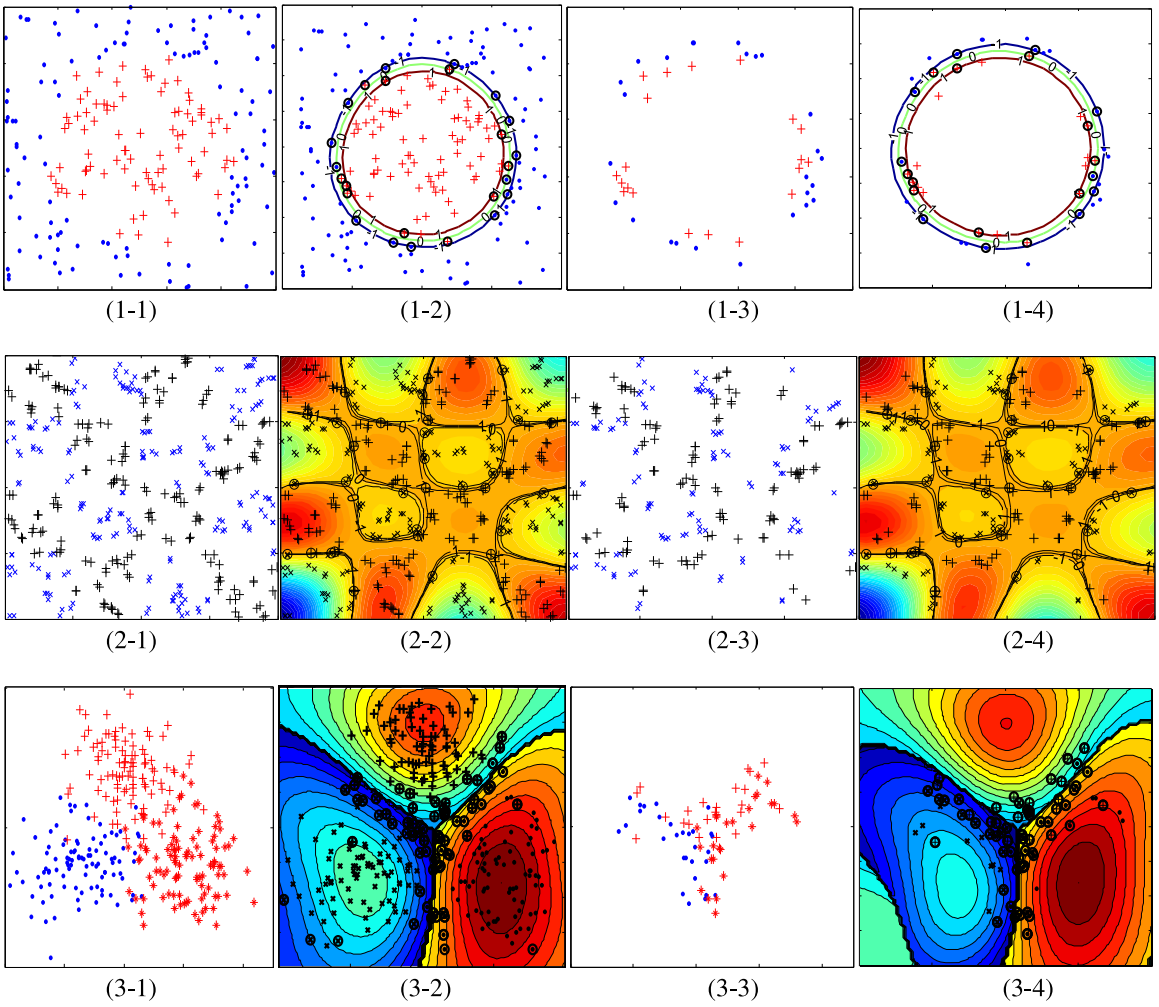


Fig. 3. Three toy examples.

While samples like x_3 are still grouped into the decision boundary.

Here we redefine the positive region and decision boundary as follows:

$$POS_B(D) = \{x_i | \exists d_j, I(\delta(x_i), d_j) \geq \beta\},$$

$$BN(D) = \{x_i | \forall d_j, I(\delta(x_i), d_j) < \beta\}.$$

We also have

$$U = POS_B(D) \cup BN(D).$$

In this case, the noise in positive region can be recognized as

$$N(D) = \{x_i | x_i \in d_k, \exists d_j, I(\delta(x_i), d_j) \geq \beta, k \neq j\}.$$

By specifying proper thresholds δ and β, we can find a boundary region with appropriate size and delete the noised samples from the boundary set. Several literatures discussed the problem of specifying the value of parameterβ [2,3]. At large, β can take values in arrange [0.7, 1]. The value of δ depends on applications. Generally speaking, if the inter-class distance of a learning sample set is large, we should assign δ with a large value to get enough boundary samples to support the optimal hyperplane, and vice versa. Largely, δ can take value in arrange [0.1, 0.5] if numerical attributes are normalized to the unit interval [0,1].

Table 1
Data description.

	Data set	Abbreviation	Samples	Numeric features	Classes
1	Ionosphere	Iono	351	34	2
2	Sonar, Mines vs. Rocks	Sonar	208	60	2
3	Small Soybean	Soy	47	35	4
4	Wisconsin Diagnostic Breast Cancer	WDBC	569	31	2
5	Wisconsin Prognostic Breast Cancer	WPBC	198	33	2
6	Wine recognition	Wine	178	13	3
7	Spambase	Spambase	4601	58	2
8	Waveform	Waveform	5000	22	3

Table 2
Comparison of feature numbers with three definitions of neighborhoods, $\delta = 0.125$.

Data	1-norm		2-norm		Inf-norm	
	N	Accuracy	*N*	Accuracy	*N*	Accuracy
Iono	6	0.9122 ± 0.0501	9	0.9264 ± 0.0517	12	0.9293 ± 0.0627
Sonar	5	0.7783 ± 0.1100	6	0.7543 ± 0.1309	7	0.8364 ± 0.0837
Soy	2	1.0000 ± 0.0000	2	1.0000 ± 0.0000	2	1.0000 ± 0.0000
WDBC	6	0.9614 ± 0.0259	8	0.9667 ± 0.0207	21	0.9790 ± 0.0161
WPBC	5	0.7632 ± 0.0304	6	0.7632 ± 0.0304	11	0.7842 ± 0.0769
Wine	4	0.9660 ± 0.0294	5	0.9493 ± 0.0412	6	0.9833 ± 0.0268
Aver.	4.67	0.8969	6	0.8933	9.83	0.9187

4. Experimental analysis

First, we give some toy examples in Fig. 3. There are three typical classification problems. The first one is a binary classification problem with circle classification plane. The second one is 4×4 checkerboard problem, and the third one is a three-class problem. 1-1, 2-1 and 3-1 show the raw sample set. 2-1, 2-2 and 3-2 show the optimal classification planes trained with the raw data. 1-3, 2-3 and 3-3 show the boundary samples found with 1-norm neighborhood rough set model. Finally, 1-4, 2-4 and 3-4 present the optimal hyperplanes trained with the boundary samples. We can see that the two classes of classification plane are quite similar although most of the learning samples do not take part in the training process.

In order to test the proposed algorithms, some data sets are downloaded from the machine learning data repository, University of California at Irvine. The data sets are outlined in Table 1.

First, we compare the feature selection algorithms based on neighborhood model with other existing methods reported in literatures [6,13,35]. Table 2 shows the numbers of selected features and classification accuracies based on neighborhood rough set model with different distance metrics. Before conducting the reduction, all the numerical attributes are normalized into interval [0,1]. From Table 2, we can find 1-norm neighborhood rough set model requires the least features, while infinite norm model requires the most features. We use the selected features to train RBF-SVM, and find that average classification accuracy of infinite norm neighborhood model is the highest one, and 1-norm neighborhood model is better than the 2-norm model. Notice that the number of features based on 1-norm are almost half of the features selected with ∞-norm. If we consider the cost of decision in measuring and storing the features, sometimes we may prefer the solution found with 1-norm model. We can find that the average number of features in the raw data is 34.17, while there are just 4.67 features in the reduced data with 1-norm model; in the mean time, the classification accuracies do not decrease remarkably. It should be noted that here $\delta = 0.125$ because we hope that the neighborhood based algorithms are consistent with the experiments in Table 3. There, we discretized the numerical attributes into four intervals for classical rough set based reduction and consistency based algorithm as they can only deal with discrete data. If $\delta = 0.125$, the diameter of neighborhoods is 0.25. Then we can consider the neighborhood model which can also divide a unit interval into four parts.

Table 3 shows the comparison of numbers of selected features and accuracies with the reduced data, where, the first two columns

Table 3
Comparison of numbers of features and accuracies based on different feature selection algorithms.

Data	Raw data		Classical rough sets		Consistency		Fuzzy entropy	
	N	Accuracy	N	Accuracy	N	Accuracy	N	Accuracy
Iono	34	0.9379 ± 0.0507	10	0.9348 ± 0.0479	9	0.9519 ± 0.0423	13	0.9462 ± 0.0365
Sonar	60	0.8510 ± 0.0948	6	0.7074 ± 0.1004	6	0.7843 ± 0.0742	12	0.8271 ± 0.0902
Soy	35	0.9300 ± 0.1135	2	1.0000 ± 0.0000	2	1.0000 ± 0.0000	2	1.0000 ± 0.0000
Wdbc	30	0.9808 ± 0.0225	8	0.9649 ± 0.0183	11	0.9579 ± 0.0238	17	0.9702 ± 0.0248
Wpbc	33	0.7779 ± 0.0420	7	0.7837 ± 0.0506	7	0.7632 ± 0.0304	17	0.8087 ± 0.0601
Wine	13	0.9889 ± 0.0234	4	0.9486 ± 0.0507	4	0.9486 ± 0.0507	9	0.9833 ± 0.0268
Aver.	**34.17**	**0.9111**	**6.17**	**0.8899**	**6.5**	**0.9010**	**11.67**	**0.9226**

present the numbers of features in the raw data and accuracies; then the second two columns are the numbers of selected features with classical rough set algorithm proposed in Ref. [2] and the corresponding classification accuracies with the reduced data; consistency based algorithm was proposed in Ref. [6]; while fuzzy entropy based method was introduced in Ref. [13]. Comparing Tables 2 and 3, we can see that the performance of all the feature subset selection algorithms is comparable. Although fuzzy entropy based method get the best classification accuracy, it requires the most features in the six algorithms.

Tables 4 and 5 show the size of positive region (**P**), boundary (**B**) and noises (**N**) found with 1-norm and 2-norm neighborhood rough sets in 1-norm neighborhood reduct subspaces, namely, we compute positive region, boundary and noise in the feature subspace found with 1-norm neighborhood. Then we delete the noise samples and train SVM with denoised data. The classification accuracies are shown in Tables 4 and 5. We can find all the classifications are improved with the denoised data.

In order to show the role of variable precision neighborhood rough set model, two particular experiments are conducted. We add some Gaussian white noise into wine and wdbc data, and let the amplitude of noise gradually increase. Then we train SVM used the data denoised by neighborhood rough sets. Fig. 4 shows the classification accuracies varying with the amplitude of noise. We can find that the accuracies denoised by neighborhood rough sets are more robust than the noised one. It also shows that variable precision neighborhood rough sets can find the noise samples.

Tables 6–9 show the classification results with the boundary samples, where SV means the numbers of support vectors in SVM. Positive and boundary mean the numbers of positive-region and boundary samples discovered from the samples with neighborhood rough sets, respectively, while SV1 means the number of the support vector found in training SVM just with the boundary samples, and "rate" is the classification accuracy of positive samples classified with SVM which is trained with the boundary samples.

In Table 6, with 1-norm neighborhood model, we search boundary samples in the feature subspace selected with 1-norm neighborhood model. Similarly, Table 7 is the result with 2-norm neighborhood based boundary in 1-norm feature subspace. Tables 8 and 9 are the results in 2-norm neighborhood based feature subspace. In these tables, we can find that just minority of samples was selected as boundary samples based on neighborhood model. Sometimes, boundary samples are even less than support vectors found in training with the whole sample set. In the mean time, most of the boundary samples selected with neighborhood model becomes support vectors. SVMs trained with the boundary samples are able to classify the samples in positive

Table 4
1-norm positive, noise and boundary with 1-norm attributes.

	P	N	B	Noised	Denoised
Iono	336	2	13	0.9122 ± 0.0501	0.9175 ± 0.0404
Sonar	172	0	36	0.7783 ± 0.1100	0.7783 ± 0.1100
Soy	47	0	0	1.0000 ± 0.0000	1.0000 ± 0.0000
Wdbc	541	1	27	0.9614 ± 0.0259	0.9631 ± 0.0254
Wpbc	153	0	45	0.7632 ± 0.0304	0.7632 ± 0.0304
Wine	164	1	13	0.9660 ± 0.0294	0.9778 ± 0.0287

Table 5
2-norm positive, noise and boundary with 1-norm attributes.

	P	N	B	Noised	Denoised
Iono	307	7	37	0.9122 ± 0.0501	0.9303 ± 0.0457
Sonar	97	1	110	0.7783 ± 0.1100	0.7869 ± 0.1054
Soy	47	0	0	1.0000 ± 0.0000	1.0000 ± 0.0000
WDBC	502	7	60	0.9614 ± 0.0259	0.9715 ± 0.0151
WPBC	93	4	101	0.7632 ± 0.0304	0.7801 ± 0.0297
Wine	148	1	29	0.9660 ± 0.0294	0.9778 ± 0.0287

Table 6
1-norm boundary in 1-norm feature subspace.

	Samples	SV	Positive	Boundary	SV1	Rate
Iono	351	173	272	72	67	0.9338
Sonar	208	189	111	97	87	0.7477
Wdbc	569	102	460	107	98	0.9978
Wpbc	198	127	89	108	85	0.9438
Wine	178	68	116	62	56	0.9828

Table 7
2-norm boundary in 1-norm feature subspace.

	Samples	SV	Positive	Boundary	SV1	Rate
Iono	351	173	231	113	86	0.9870
Sonar	208	189	91	117	104	0.7363
Wdbc	569	102	480	88	81	0.9958
Wpbc	198	127	96	102	84	0.9167
Wine	178	68	97	81	60	1

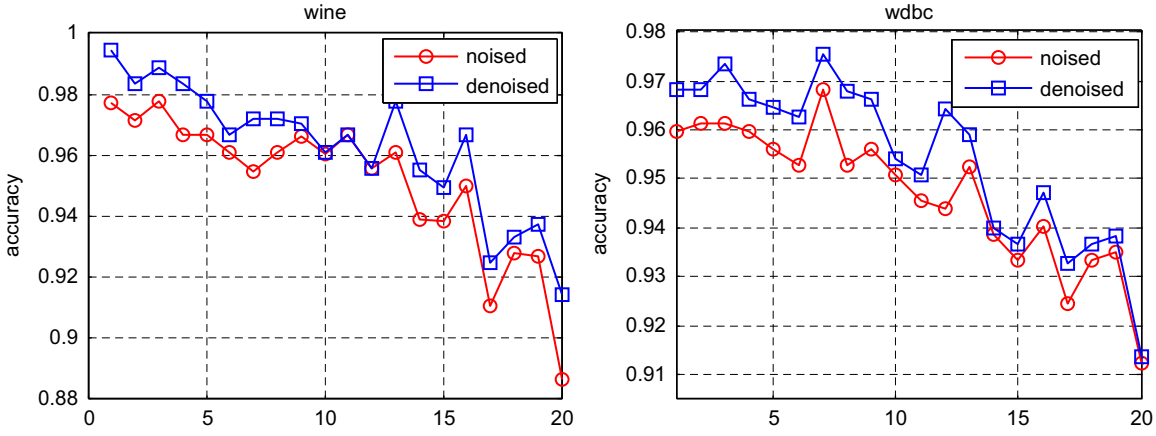


Fig. 4. Comparison the classification accuracies of noise and denoised data varying with the amplitude of noise.

region with high accuracies, comparable with or better than the SVMs trained with the whole samples, shown in Table 2.

The accuracies of samples in positive regions can just show that SVMs trained with boundary samples can recognize the positive region, rather than the whole sample spaces. Therefore the accuracies in Tables 6–9 are too optimistic. In order to test the error, we conduct 10-fold cross validation. The experimental results are shown in Tables 10 and 11. Here, we divide the whole samples into 10 subsets, and combine 9 subsets as a new training sample set, and the left one is the testing sample set. We then select the boundary samples in the new sample set with neighborhood model. The selected boundary samples are used to train SVM and recognize the testing subset. In Table 10, "Boundary" and "SV" are the average numbers of boundary samples and support vectors, respectively, while "time" and "accuracy" are the average values based on 10-fold cross validation in Table 11.

In Table 10, as to wdbc, wpbc and wine, only a minority of the raw data is selected in the boundary, most of the samples are not involved in training. Therefore, the training process will be greatly speeded up with the reduced data. Results in Table 11 validate this analysis. At the same time, we can find that average classification accuracies do not decrease evidently compared with the SVM trained with the whole sample set. It shows that the boundary samples selected with neighborhood model are able to support the optimal classification hyper-plane and SVMs just trained with boundary samples still keep similar classification power as the whole set.

Table 8
1-norm boundary in 2-norm feature subspace.

	Samples	SV	Positive	Boundary	SV1	Rate
Iono	351	173	323	26	25	0.8669
Sonar	208	189	98	110	105	0.7857
Wdbc	569	102	481	87	80	0.9938
Wpbc	198	127	98	100	83	0.9286
Wine	178	68	133	45	45	0.9925

Table 9
2-norm boundary in 2-norm feature subspace.

	Samples	SV	Positive	Boundary	SV1	Rate
Iono	351	173	287	58	53	0.9373
Sonar	208	189	87	121	108	0.7471
Wdbc	569	102	421	146	105	1
Wpbc	198	127	102	96	79	0.9020
Wine	178	68	89	89	64	0.9888

5. Conclusion

Support vector machine confronts the problem of reducing time and space complexity in training. It is known that the optimal hyper-plane of SVM just depends on part of training samples, namely, support vectors, rather than the whole sample set. Experiences show that support vectors are a minority of the training samples [27]. Therefore, the training time and space will be greatly reduced if one can select the support vectors in preprocessing, and just trains SVM with the selected samples.

In this paper, we show a neighborhood rough set based algorithm for simultaneously selecting sample and features. The neighborhood rough set model segments the samples set into two parts: positive region and boundary. Samples in positive region can be classified into one of the decision without uncertainty, while boundary is the sample subset whose neighborhood comes from multiple classes. Therefore, boundary samples usually distribute near the classification hyper-plane. They are probable the support vectors. One can collect these samples and use them to train SVM. This idea is much similar with the one proposed in Refs. [28,29]. However, the proposed method is built on a

Table 10
Classification results based on 10-fold cross validation (a).

Data	Raw data		1-norm feature		1-norm feature+1-norm boundary		1-norm feature+2-norm boundary	
	Boundary	SV	Boundary	SV	Boundary	SV	Boundary	SV
Iono	351	67	351	111	217	101	171	91
Sonar	208	91	208	130	120	113	142	122
Wdbc	569	40	569	104	95	89	128	95
Wpbc	198	95	198	93	59	51	88	73
Wine	178	41	178	70	86	65	73	61
Spambase	4601	692	4601	1178	2533	972	2761	990
Waveform	5000	931	5000	1238	2976	1094	3406	1175

Table 11
Classification results based on 10-fold cross validation (b).

Data	Raw data		1-norm feature		1-norm feature+1-norm boundary		1-norm feature+2-norm boundary	
	Time(s)	Accuracy	Time(s)	Accuracy	Time(s)	Accuracy	Time(s)	Accuracy
Iono	0.07	0.9306 ± 0.0327	0.11	0.9122 ± 0.0501	0.07	0.9150 ± 0.0477	0.06	0.9122 ± 0.0518
Sonar	0.06	0.8566 ± 0.0834	0.11	0.7783 ± 0.1100	0.03	0.7543 ± 0.1076	0.04	0.7783 ± 0.1100
Wdbc	0.07	0.9648 ± 0.0202	0.22	0.9614 ± 0.0259	0.01	0.9597 ± 0.0287	0.01	0.9597 ± 0.0275
Wpbc	0.06	0.7739 ± 0.0992	0.10	0.7632 ± 0.0304	0.03	0.7632 ± 0.0304	0.03	0.7526 ± 0.0272
Wine	0.23	0.9833 ± 0.0269	0.30	0.9444 ± 0.0524	0.07	0.9444 ± 0.0524	0.05	0.9389 ± 0.0611
Spambase	348	0.9363 ± 0.0118	501	0.8907 ± 0.0158	76	0.8617 ± 0.0296	94	0.8696 ± 0.0149
Waveform	1269	0.8396 ± 0.0206	1949	0.8152 ± 0.1320	509	0.7951 ± 0.0430	763	0.8061 ± 0.0280

systematically theoretical framework, which presents a clear conceptual basic for analyzing decision positive region, boundary and noise.

Furthermore, neighborhood rough set divides the feature set into strongly relevant features, weakly relevant and indispensable features, weakly relevant and superfluous features, and irrelevant features. Based on the neighborhood model, we develop an attribute reduction algorithm to select the strongly relevant features, weakly relevant and indispensable features.

Experiments are conducted with toy examples and UCI data sets. The results show that the proposed method can exactly discover boundary samples of complex classification problems. The attribute reduction algorithm based on neighborhood rough sets is able to select minority of features and keep the similar classification power. To summarize, the proposed method can reduce data sets in terms of both patterns and features. SVM trained with the selected samples in reduced subspaces get the comparable classification performance with the raw data.

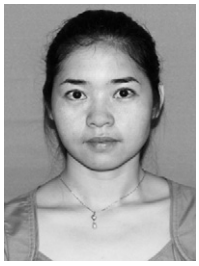
Acknowledgement

This work is partly supported by National NSFCs (60703013, 60903088, 60903089, 61070242), Development Program for Outstanding Young Teachers in Harbin Institute of Technology (HITQNJS.2007.017), by the natural science foundation of Hebei Province (F2009000231, F2009000227, F2010000323) and by the Scientific Research Project of Department of Education of Hebei Province (2009410).

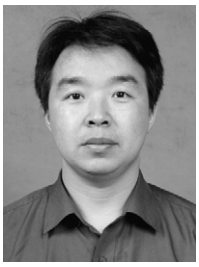
References

[1] M.B. Almeida, A. Braga, J.P. Braga, SVM-KM: Speeding SVMs learning with a priori cluster selection and k-means, in: Proceedings of the Sixth Brazilian Symposium on Neural Networks, 2000, pp. 162–167.
[2] M. Beynon, M.J. Peel, Variable precision rough set theory and data discretisation: an application to corporate failure prediction, Omega 29 (2001) 561–576.
[3] M. Beynon, Reducts within the variable precision rough sets model: a further investigation, European Journal of Operational Research 134 (2001) 592–605.
[4] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.
[5] C. Cortes, V.N. Vapnik, Support vector networks, Machine Learning 20 (1995) 273–297.
[6] M. Dash, H. Liu, Consistency-based search in feature selection, Artificial Intelligence 151 (2003) 155–176.
[7] B. Fei, J.B. Liu, Binary tree of SVM: a new fast multiclass training and classification algorithm, IEEE Transactions on Neural Networks 17 (2006) 696–704.
[8] H. Fröhlich, O. Chapelle, B. Schölkopf, Feature selection for support vector machines by means of genetic algorithms, in: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, 2003.
[9] I. Guyon, J. Weston, S. Barnhill, et al., Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.
[10] M.A. Hearst, B. Scholkopf, S. Dumais, E. Osuna, J. Platt, Trends and controversies-support vector machines, IEEE Intelligent Systems 13 (1997) 18–28.
[11] L. Hoegaerts, J.A.K. Suykens, J. Vandewalle, B. De Moor, Subset based least squares subspace regression in RKHS, Neurocomputing 63 (2005) 293–323.
[12] Q.H. Hu, D.R. Yu, Z.X. Xie, Neighborhood classifiers, Expert Systems with Applications 34 (2008) 866–876.
[13] 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.
[14] T. Joachims, Making large scale support vector machine learning practical, Advances in Kernel Methods: Support Vector Machines, MIT Press, Cambridge, MA, 1999, pp. 169–184.
[15] R. Koggalage, S. Halgamuge, Reducing the number of training samples for fast support vector machine classification, Neural Information Processing 2 (3) (2004) 57–65.
[16] Y.J. Lee, O.L. Mangasarian, RSVM: Reduced Support Vector Machines, Data Mining Institute Technical Report 00-07, July, 2000, in: Proceedings of the First SIAM International Conference on Data Mining, Chicago, 2001.
[17] K.M. Lin, C.J. Lin., A study on reduced support vector machines, IEEE Transactions on Neural Networks 14 (2003) 1149–1159.
[18] T.Y. Lin, Neighborhood systems and relational database, in: Proceedings of 1988 ACM 16th Annual Computer Science Conference, 1998, pp. 23–25.
[19] T.Y. Lin, Neighborhood systems—a qualitative theory for fuzzy and rough sets, in: P. Wang (Ed.), Advances in machine intelligence and soft computing, vol. 4, 1997, pp. 132–155.
[20] T.Y. Lin, Neighborhood systems—application to qualitative fuzzy and rough sets, in: P. Wang (Ed.), Advances in Machine Intelligence and Soft-Computing, 1997 pp. 132–155.
[21] A. Lyhyaoui, M. Martinez, I. Mora, M. Vazquez, et al., Sample selection via clustering to construct support vector-like classifiers, IEEE Transactions on Neural Networks 10 (1999) 1474–1481.
[22] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, IEEE Transactions on Neural Networks 12 (2001) 181–201.
[23] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Machine Learning 61 (2005) 129–150.
[24] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, Adv. Kernel Methods: Support Vector Machines, MIT Press, Cambridge, MA, 1999, pp. 185–208.
[25] F. Rossi, N. Villa, Support vector machine for functional data classification, Neurocomputing 69 (2006) 730–742.
[26] V.D. Sanchez, Advanced support vector machines and kernel methods, Neurocomputing 55 (2003) 5–20.
[27] B. Scholkopf, C. Burges, V. Vapnik, Extracting support data for a given task, in: Proceedings of First International Conference on Knowledge Discovery and Data Mining. AAAI press, Menlo Park, CA, 1995, pp. 252–257.
[28] H. Shin, S. Cho, Fast pattern selection for support vector classifiers, in: Proceedings. of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Lecture Notes in Artificial Intelligence (LNAI 2637), Seoul, Korea, 2003, pp. 376–387.
[29] H. Shin, S. Cho, Invariance of neighborhood relation under input space to feature space mapping, Pattern Recognition Letters 26 (2005) 707–718.
[30] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (1999) 293–300.
[31] F.E.H. Tay, L.J. Cao, Modified support vector machines in financial time series forecasting, Neurocomputing 48 (2002) 847–861.
[32] V. Tresp, Scaling kernel-based systems to large data sets, Data Mining and Knowledge Discovery 5 (2001) 197–211.
[33] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[34] W.Z. Wu, W.X. Zhang, Neighborhood operator systems and approximations, Information Sciences 144 (2002) 201–217.
[35] Y.Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Information Sciences 111 (1998) 239–259.
[36] Y.Y. Yao, Neighborhood systems and approximate retrieval, Information Sciences 176 (2006) 3431–3452.
[37] N. Zhong, J. Dong, S. Ohsuga, Using rough sets with heuristics for feature selection, Journal of Intelligent Information Systems 16 (2001) 199–214.
[38] W. Ziarko, Variable precision rough sets model, Journal of Computer and System Sciences 46 (1993) 39–59.

Qiang He received his B.Sc. and M.Sc. degrees in mathematics from Hebei University, Baoding, China, in 2000 and 2003, respectively. From 2003 to now, he worked as a lectuer in the Faculty of Mathematics and Computer Science, Hebei University. In 2004 and 2008, he worked as a Research Assistant at the Department of Computing, Hong Kong Polytechnic University, Kowloon. His main research interests include inductive learning, genetic algorithms and statistical learning theory.



Zongxia Xie received her B.E. from Dalian Maritime University in 2003, and M.E. and Ph.D. from Harbin Institute of Technology in 2005, and 2010, respectively. Now she is a postdoctoral fellow with Shenzhen Graduate School, Harbin Institute of Technology. Her major interests include machine learning and pattern recognition with rough sets and SVM, solar image processing and knowledge discovery. She has published more than 20 conference and journal papers on related topics.



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

Congxin Wu was the former dean of Department of Mathematics and the former head of institution of Mathematics, Harbin Institution of Technology. He is a board member of China System Engineer Academy, board member of the China Fuzzy Mathematics and Fuzzy System Committee, Board member of China Mathematics Society. His main research interests are in functional analysis, fuzzy mathematics.