ORIGINAL ARTICLE

A rough margin-based one class support vector machine

Yitian Xu · Chunmei Liu

Received: 20 December 2011 / Accepted: 26 January 2012 / Published online: 16 February 2012

Abstract We propose a rough margin-based one class support vector machine (Rough one class SVM) by introducing the rough set theory into the one class SVM, to deal with the over-fitting problem. We first construct rough lower margin, rough upper margin, and rough boundary and then maximize the rough margin rather than the margin in the one class SVM. Thus, more points are adaptively considered in constructing the separating hyper-plane than those used in the conventional one class SVM. Moreover, different points staying at the different positions are proposed to give different penalties. Specifically, the samples staying at the lower margin are given the larger penalties than those in the boundary of the rough margin. Therefore, the new classifier can avoid the over-fitting problem to a certain extent and yields great generalization performance. Experimental results on one artificial dataset and eight benchmark datasets demonstrate the feasibility and validity of our proposed algorithm.

Keywords Rough set theory · Rough margin · One class SVM · Rough one class SVM

1 Introduction

Support vector machine (SVM), motivated by the Vapnik–Chervonenkis (VC) dimensional theory and the statistical learning theory [1], is a promising machine learning technique. Compared with other machine learning approaches like artificial neural networks [2], SVM has many advantages. First, SVM solves a quadratic programming problem (QPP), which assures that once an optimal solution is obtained, it is the unique (global) solution. Second, SVM derives its sparse and robust solution by maximizing the margin between the two classes. Third, SVM implements the structural risk minimization principle rather than the empirical risk minimization principle, which minimizes the upper bound of the generalization error. At present, SVM has been successfully applied in various aspects ranging from machine learning, data mining, and knowledge discovery [3, 4].

Classical SVM can effectively deal with the binary classification problem, but it requires the class labels of training samples. However, it is sometimes hard or expensive to acquire them in real life, and it leads to unsupervised learning. One class problems are prevalent in real world where positive and unlabeled data are widely available.

One class SVM [5–7] is a novel machine learning algorithm in dealing with one class problem or clustering problem. It first maps the samples into the high-dimensional feature space corresponding to the kernel function, then separates them from the origin with maximum margin. Finally, it returns a decision function f that takes the value $+1$ in a small region capturing most of the data points and -1 elsewhere. Because the classification hyper-plane depends only on a small proportion of the training samples (i.e. support vectors) in one class SVM, hence it is too sensitive to the noise data and outliers, and produces over-fitting problem.

In order to avoid over-fitting problem in one class SVM, some improved algorithms were proposed, for example, in order to reduce the effects of outliers, the fuzzy one class SVM associated with a fuzzy membership to each training sample was proposed in [8], but it was difficult to determine membership of each samples. A weighted one class

Y. Xu (✉) · C. Liu
College of science, China Agricultural University,
Beijing 100083, China
e-mail: xytshuxue@126.com

SVM was proposed in [9]. However, it was also difficult to determine the weighted coefficient. In addition, in order to improve the prediction accuracy of classifier, a support vector classification algorithm based on rough set preprocessing was proposed in [10], and it could eliminate the redundant information, avoid the noise disturbing, and overcome the disadvantage of slowly processing speed caused by classical SVM approach.

Motivated by the above studies, rough set theory [11, 12] is incorporated into the one class SVM, and a rough margin-based one class SVM is proposed in this paper. Thus, more data points are adaptively considered in constructing the separating hyper-plane, and different penalties are proposed to give the samples depending on their positions. The optimal separating hyper-plane is found by maximizing the rough margin from the origin [13, 14] rather than the margin in one class SVM. Therefore, the proposed algorithm can avoid the over-fitting problem to a certain extent.

The effectiveness of the proposed algorithm is demonstrated with the experiments on one artificial dataset and eight benchmark datasets. The experimental results show that our algorithm achieves significant performance in comparison with one class SVM. Moreover, our algorithm does not increase elapsed time.

The paper is organized as follows. Section 2 outlines the rough set theory and one class SVM. A rough margin-based one class SVM is proposed in Sect. 3. Numerical experiments are implemented in Sect. 4. The last section concludes the paper.

2 Preliminaries

2.1 Rough set theory

Rough set theory [11, 12] is an effective tool in dealing with vagueness and uncertainty information, and it deals with information represented by a table called an information system, which consists of objects (or cases) and attributes. An information system is composed of a 4-tuple as follows:

$$S = \langle U, A, V, f \rangle \qquad (1)$$

where U is the universe, a finite set of n objects $\{x_1, x_2, \ldots, x_n\}. A = C \bigcup D, C$ is a set of condition attributes, and D is a set of decision attributes. V is attribute value. $f : U \times A \longrightarrow V$ is the total decision function called the information function.

Upper and lower approximations are the important concepts in rough set theory, which are shown in Fig. 1. For a given information system S, a given subset of attributes $R \subseteq A$ determines the approximation space $RS = (U, ind(A))$ in S, and for a given $R \subseteq A$ and $X \subseteq U$ (a concept X), the R-lower approximation $\underline{R}X$ of set X is defined as follows:

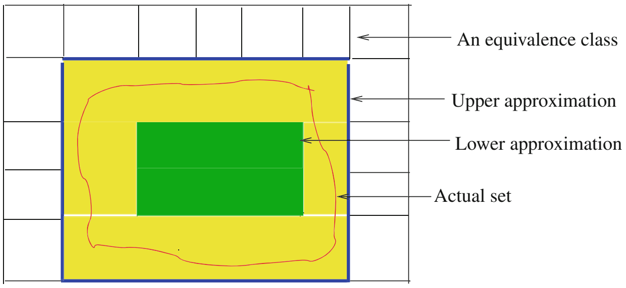

Fig. 1 Lower and upper approximations of a set in RST

$$\underline{R}X = \{x \in U : [x]_R \subseteq X\}. \qquad (2)$$

$\underline{R}X$ is the set of all objects from U which can be certainly classified as elements of X employing the set of attributes R.

The R upper approximation $\overline{R}X$ of set X is defined as follows:

$$\overline{R}X = \{x \in U : [x]_R \bigcap X \neq \phi\} \qquad (3)$$

where $\overline{R}X$ is the set of objects of U which can be possibly classified as elements of X using the set of attributes R, and $[x]_R$ denotes an equivalence class of $Ind(R)$ that contains x (called the indiscernibility relation).

Given an information system S, condition attributes C and decision attributes $D, A = C \bigcup D$, and for a given subset of condition attributes $P \subseteq C$, we can define a positive region

$$pos_p(D) = \bigcup_{X \in U/ind(D)} \underline{P}X. \qquad (4)$$

The positive region pos_p (D) contains all objects in U, which can be classified without error into distinct classes defined by $ind(D)$ based only on information in the $ind(P)$.

If $\underline{R}X \neq \overline{R}X$, then X is a rough set, and its boundary region, $Bnd(X) = \overline{R}X - \underline{R}X$, is correspondingly nonempty.

2.2 One class support vector machine

One class SVM was proposed by Schölkopf for estimating the support of a high-dimensional distribution [6]. Given a training set without any class information, one class SVM constructs a decision function that takes the value +1 in a small region capturing most of the data points, and −1 elsewhere. The strategy in this technique is to map the input vectors into a high-dimensional feature space corresponding to a kernel and construct a linear decision function in this space to separate the dataset from the origin with maximum margin. Via the freedom to utilize different

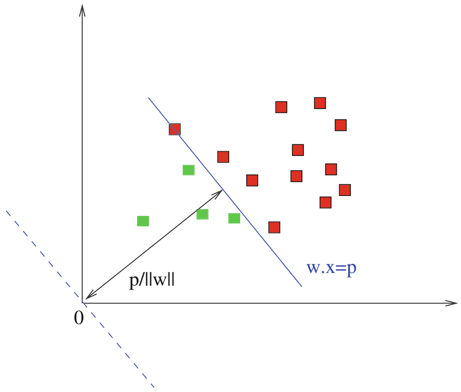

Fig. 2 Illustration of the one class SVM

types of kernel, the linear decision functions in the feature space are equivalent to a variety of nonlinear decision functions in the input space. A parameter $v \in (0, 1]$ was introduced into the one class SVM to control the trade-off between the fraction of data points in the region and the generalization ability of the decision function.

Given a training dataset without any class information

$$T = (x_1, x_2, \ldots, x_l), x_i \in R^n \tag{5}$$

To separate the dataset from the origin, which is illustrated in Fig. 2, one class SVM solves the following QPP.

$$\begin{aligned} \min_{w,\xi,\rho} \quad & \frac{1}{2}\|w\|^2 - \rho + \frac{1}{vl}\sum_{i=1}^{l}\xi_i \\ \text{s.t.} \quad & (w \cdot \phi(x_i)) \geq \rho - \xi_i, \\ & \xi_i \geq 0, i = 1, 2, \ldots, l. \end{aligned} \tag{6}$$

Here, ρ is a threshold parameter, and ξ_i is a slack variable. $v \in (0, 1]$ is a parameter chosen a priori, and it has the following property.

Proposition 1 *Assume the solution of QPP* (6) *satisfies* $\rho \neq 0$. *The following statements hold:*

(1) *v is an upper bound on the fraction of outliers.*
(2) *v is a lower bound on the fraction of support vectors.*

The proof can be found in [6]. The solution to the QPP (6) is transformed into its dual problem by the saddle point of the Lagrange function,

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i, x_j) \\ \text{s.t.} \quad & \sum_{i=1}^{l}\alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{vl}, i = 1, 2, \ldots, l. \end{aligned} \tag{7}$$

Once the solution $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)^T$ to the QPP (7) has been found, the decision function can be expressed as follows:

$$f(x) = sgn\left(\sum_{i=1}^{l}\alpha_i K(x_i, x) - \rho\right). \tag{8}$$

We can recover threshold ρ by exploiting that for any such $0 < \alpha_j < \frac{1}{vl}$, the corresponding pattern x_j satisfies

$$\rho = (w \cdot \phi(x_j)) = \sum_{i=1}^{l}\alpha_i K(x_i, x_j). \tag{9}$$

Here, $K(x_i, x_j)$ is a kernel function that gives the dot product $(\phi(x_i)\cdot\phi(x_j))$ in the higher-dimensional space.

3 A rough margin-based one class SVM

We know that the separating hyper-plane depends only on a small proportion of the training samples (i.e. support vectors) in one class SVM, and it is very sensitive to the noises and outliers. In order to overcome the over-fitting problem when noises and outliers exist in one class SVM, we propose a rough margin-based one class SVM in this section, where more data points are adaptively considered in constructing the separating hyper-plane. Following the rough set theory, the rough margin can be expressed as the lower margin and the upper margin. The training samples locating within the lower margin (corresponding to the positive region) are considered as outliers, the samples locating within the upper margin (corresponding to the negative region) are not outliers and are regarded as target class samples, and the samples staying at the boundary of the rough margin (corresponding to the boundary region) are possible outliers. It is more reasonable to give different penalties to the points depending on their positions in the process of learning the optimal hyper-plane. Specifically, we give the major penalties to the training points that lie within the lower margin and give the minor penalties to the points that lie within the boundary of the rough margin.

The rough margin is illustrated in Fig. 3. The blue line denotes the hyper-plane $(w \cdot \phi(x)) = \rho_\mu$, and the red line stands for the hyper-plane $(w \cdot \phi(x)) = \rho_l$. The region satisfying $(w \cdot \phi(x)) > \rho_\mu$ corresponds to the upper margin, the region satisfying $(w \cdot \phi(x)) < \rho_l$ corresponds to the lower margin, and the region satisfying $\rho_l < (w \cdot \phi(x)) < \rho_\mu$ corresponds to the boundary region.

3.1 Rough one class supper vector machine

Motivated by the above studies, a rough margin-based one class SVM is designed as follows [15],

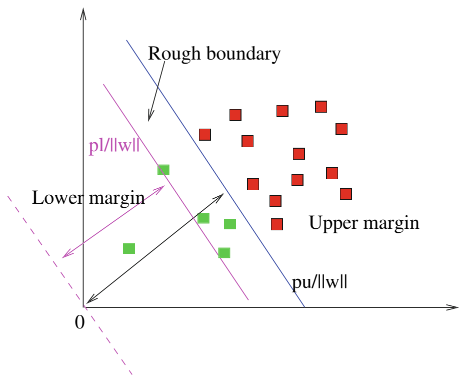

Fig. 3 Illustration of the Rough one class SVM

$$\begin{aligned}
\min_{w,\xi,\rho}\quad & \frac{1}{2}\|w\|^2-(\rho_l+\rho_\mu)+\frac{1}{vl}\sum_{i=1}^{l}(\xi_i+\sigma\xi_i^*)\\
\text{s.t.}\quad & (w\cdot\phi(x_i))\geq\rho_\mu-\xi_i-\xi_i^*,\\
& 0\leq\xi_i\leq\rho_\mu-\rho_l,\\
& \rho_l,\rho_\mu\geq 0,\\
& \xi_i^*\geq 0, i=1,2,\ldots,l,
\end{aligned}\tag{10}$$

where ξ_i and ξ_i^* are slack variables. ρ_μ and ρ_l are trade-off parameters. Parameter σ (>1) is chosen a priori, which implies that larger penalty to the point locating in the rough low margin, as it has larger effect on the separating hyperplace than other points. In order to solve the QPP (10), we first introduce the following Lagrangian function,

$$\begin{aligned}
L=&\frac{1}{2}\|w\|^2-(\rho_l+\rho_\mu)+\frac{1}{vl}\sum_{i=1}^{l}(\xi_i+\sigma\xi_i^*)\\
&-\sum_{i=1}^{l}\alpha_i((w\cdot\phi(x_i))-\rho_\mu+\xi_i+\xi_i^*)\\
&-\sum_{i=1}^{l}\beta_i(\rho_\mu-\rho_l-\xi_i)-\sum_{i=1}^{l}\gamma_i\xi_i-\mu_1\rho_l-\mu_2\rho_\mu\\
&-\sum_{i=1}^{l}\eta_i\xi_i^*,
\end{aligned}\tag{11}$$

where $\alpha_i\geq 0,\ \beta_i\geq 0,\ \gamma_i\geq 0,\ \eta_i\geq 0,\ \ \mu_1\geq 0,\ \ \mu_2\geq 0$ are Lagrangian multipliers. According to Karush–Kuhn–Tucker (KKT) conditions, parameters satisfy the following conditions:

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}\alpha_i\phi(x_i)=0,\tag{12}$$

$$\frac{\partial L}{\partial\xi_i}=\frac{1}{\upsilon l}-\alpha_i+\beta_i-\gamma_i=0,\tag{13}$$

$$\frac{\partial L}{\partial\xi_i^*}=\frac{\sigma}{\upsilon l}-\alpha_i-\eta_i=0,\tag{14}$$

$$\frac{\partial L}{\partial\rho_l}=-1+\sum_{i=1}^{l}\beta_i-\mu_1=0,\tag{15}$$

$$\frac{\partial L}{\partial\rho_\mu}=-1+\sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}\beta_i-\mu_2=0,\tag{16}$$

$$\alpha_i((w\cdot\phi(x_i))-\rho_\mu+\xi_i+\xi_i^*)=0,\tag{17}$$

$$\beta_i(\rho_\mu-\rho_l-\xi_i)=0,\tag{18}$$

$$\gamma_i\xi_i=0,\quad \eta_i\xi_i^*=0,\quad \mu_1\rho_l=0,\quad \mu_2\rho_\mu=0.\tag{19}$$

We can derive the dual problem of the QPP (10) as follows:

$$\begin{aligned}
\max_{\alpha}\quad & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j)\\
\text{s.t.}\quad & \sum_{i=1}^{l}\alpha_i\geq 2,\\
& 0\leq\alpha_i\leq\frac{\sigma}{vl}, i=1,2,\ldots,l.
\end{aligned}\tag{20}$$

For the sake of convenience, we first give an equivalent formulation of the QPP (20). The optimal trade-offs ρ_μ and ρ_l in the QPP (10) are actually larger than zero. On the conditions of $\rho_\mu,\ \rho_l>0$, we give the following Proposition.

Proposition 2 *The QPP* (20) *can be transformed into the following QPP.*

$$\begin{aligned}
\max_{\alpha}\quad & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_jK(x_i,x_j)\\
\text{s.t.}\quad & \sum_{i=1}^{l}\alpha_i=2,\\
& 0\leq\alpha_i\leq\frac{\sigma}{vl}, i=1,2,\ldots,l.
\end{aligned}\tag{21}$$

The QPPs (20) and (21) differ in the first constraint condition. Namely, the inequality constraint $\sum_{i=1}^{l}\alpha_i\geq 2$ in (20) becomes an equality constraint $\sum_{i=1}^{l}\alpha_i=2$ in (21).

Proof According to assumptions $\rho_l,\ \rho_\mu>0$ and KKT conditions $\mu_1\rho_l=0,\ \ \mu_2\rho_\mu=0$, we can obtain Lagrangian multipliers $\mu_1=0$ and $\mu_2=0$. After combining (15) and (16), we can get an equality constraint $\sum_{i=1}^{l}\alpha_i=2$. □

3.2 Lagrangian multiplier α_i and support vector

Once the optimal solution $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_l)^T$ to the QPP (20) has been found, the position of a training point is determined by the value of α_i, and we can then draw the following conclusions.

Proposition 3 *Different values of α_i are corresponding to the different positions of the points. We can divide the samples into five cases according to the values of α_i.*

(1) *If $\alpha_i = 0$, then the point lies within the upper margin and satisfies $(w \cdot \phi(x_i)) \geq \rho_u$. It is corresponding to a point of the target class; namely, it is an usual point.*

(2) *If $0 < \alpha_i < \frac{1}{vl}$, then the point lies on the border of the upper margin and satisfies $(w \cdot \phi(x_i)) = \rho_u$. It is corresponding to a support vector on the border of the upper margin.*

(3) *If $\alpha_i = \frac{1}{vl}$, then the point lies inside the boundary of the rough margin and satisfies $\rho_l < (w \cdot \phi(x_i)) < \rho_u$. It is corresponding to a support vector in the rough boundary.*

(4) *If $\frac{1}{vl} < \alpha_i < \frac{\sigma}{vl}$, then the point lies on the border of the lower margin and satisfies $(w \cdot \phi(x_i)) = \rho_l$. It is corresponding to a support vector on the border of the lower margin.*

(5) *If $\alpha_i = \frac{\sigma}{vl}$, then the point lies within the lower margin and satisfies $(w \cdot \phi(x_i)) < \rho_l$. It is usually corresponding to an outlier.*

Proof

(1) If $\alpha_i = 0$, we can get Lagrangian multiplier $\gamma_i > 0$ from (13), and $\eta_i > 0$ from (14). We further obtain $\xi_i = 0$ and $\xi_i^* = 0$ from KKT condition (19). Finally, we can achieve $(w \cdot \phi(x_i)) \geq \rho_u$ when we substitute ξ_i and ξ_i^* into constraint $(w \cdot \phi(x_i)) \geq \rho_\mu - \xi_i - \xi_i^*$ in (10).

(2) If $0 < \alpha_i < \frac{1}{vl}$, we can get $\eta_i > 0$ from (14), and then substitute it into $\eta_i \xi_i^* = 0$, we have $\xi_i^* = 0$. In addition, we can get $\gamma_i > 0$ from (13), and then we also get $\xi_i = 0$ from KKT condition (19). Finally, we can achieve $(w \cdot \phi(x_i)) = \rho_u$ when we substitute α_i, ξ_i and ξ_i^* into (17).

(3) If $\alpha_i = \frac{1}{vl}$, we can get $\eta_i > 0$ from (14), and then substitute it into $\eta_i \xi_i^* = 0$, we further get $\xi_i^* = 0$. By substituting $\alpha_i > 0$ and $\xi_i^* = 0$ into (17), we can achieve $(w \cdot \phi(x_i)) = \rho_u - \xi_i (\xi_i > 0)$, that is $\rho_l < (w \cdot \phi(x_i)) < \rho_u$.

(4) If $\frac{1}{vl} < \alpha_i < \frac{\sigma}{vl}$, we can get $\eta_i > 0$ from (14), and then substitute it into $\eta_i \xi_i^* = 0$, we further obtain $\xi_i^* = 0$. In addition, we can get $\beta_i > 0$ by substituting α_i into (13). We can get $\rho_l = \rho_u - \xi_i$ from (18), and then substitute them into (17), we have $(w \cdot \phi(x_i)) = \rho_l$.

(5) If $\alpha_i = \frac{\sigma}{vl}$, we can get $\beta_i > 0$ from (13) and further obtain $\rho_l = \rho_u - \xi_i$ from (18). When we substitute them into (17), we can get $(w \cdot \phi(x_i)) = \rho_l - \xi_i^*$, $(\xi_i^* > 0)$, that is $(w \cdot \phi(x_i)) < \rho_l$. Point x_i is usually corresponding to an outlier. □

3.3 Threshold ρ and decision rules

We can recover ρ_μ by exploiting that for any such $0 < \alpha_j < \frac{1}{vl}$, the corresponding pattern x_j satisfies

$$\rho_\mu = (w \cdot \phi(x_j)) = \sum_{i=1}^{l} \alpha_i K(x_i, x_j). \tag{22}$$

We calculate ρ_μ by following formula to increase the robustness of the proposed algorithm,

$$\rho'_\mu = (w \cdot \phi(x_j)) = \frac{1}{l_1} \sum_{j=1}^{l_1} \sum_{i=1}^{l} \alpha_i K(x_i, x_j), \tag{23}$$

where l_1 denotes the number of support vectors lying on the boundary of lower margin, which corresponds to $0 < \alpha_j < \frac{1}{vl}$.

In addition, we also recover $\rho_ł$ by exploiting that for any such $\frac{1}{vl} < \alpha_j < \frac{\sigma}{vl}$, the corresponding pattern x_j satisfies

$$\rho_l = (w \cdot \phi(x_j)) = \sum_{i=1}^{l} \alpha_i K(x_i, x_j). \tag{24}$$

We also calculate the robust ρ_l according to the following formula,

$$\rho_l^{'} = (w \cdot \phi(x_j)) = \frac{1}{l_2} \sum_{j=1}^{l_2} \sum_{i=1}^{l} \alpha_i K(x_i, x_j), \tag{25}$$

where l_2 denotes the number of support vectors that lying on the boundary of upper margin, which satisfies $\frac{1}{vl} < \alpha_j < \frac{\sigma}{vl}$.

Based on the above studies, we present the following double decision functions,

$$f(x) = sgn\left(\sum_{i=1}^{l} \alpha_i K(x_i, x) - \rho'_u\right), \tag{26}$$

$$f'(x) = sgn\left(\sum_{i=1}^{l} \alpha_i K(x_i, x) - \rho'_l\right). \tag{27}$$

Given a testing point x_t, once $f(x_t)$ and $f'(x_t)$ are obtained from (26) and (27), we determine its class label according to the following proposition.

Proposition 4 *Once $f(x_t)$ and $f'(x_t)$ are obtained from (26) and (27) for a testing point x_t, we determine its class label by*

(1) *If $f(x_t) > 0$, then the testing point x_t belongs accurately to the target class; namely, it is an usual point.*

(2) *If $f'(x_t) < 0$, then the testing point x_t belongs accurately to an outlier.*

(3) *If $f'(x_t) > 0$ and $f(x_t) < 0$, then the testing point x_t belongs to a possible outlier. Namely, the given information is not sufficient to determine its class label.*

3.4 Property of parameters σ and v

Like one class SVM, parameter v and the new introduced parameter σ in our proposed algorithm also have their properties, and they are described as follows.

Proposition 5 *Suppose there are l_3 points that satisfy $0 < \alpha_i \le \frac{\sigma}{vl}$, that is, l_3 denotes the total number of support vectors, then $\frac{l_3}{l} \ge \frac{2v}{\sigma}$. That is to say, $\frac{2v}{\sigma}$ is a lower bound on the fraction of support vectors.*

Proof Suppose there are l_3 support vectors that satisfy $0 < \alpha_i \le \frac{\sigma}{vl}$, then $\sum_{i=1}^{l_3} \alpha_i \le \frac{\sigma}{vl} \cdot l_3$, and $\sum_{i=1}^{l_3} \alpha_i = \sum_{i=1}^{l} \alpha_i = 2$, hence $\frac{\sigma}{vl} \cdot l_3 \ge 2$, that is $\frac{l_3}{l} \ge \frac{2v}{\sigma}$. Namely, $\frac{2v}{\sigma}$ is a lower bound on the fraction of support vectors. □

Proposition 6 *Suppose there are l_4 outliers that correspond to $\alpha_i = \frac{\sigma}{vl}$, then $\frac{2v}{\sigma}$ is an upper bound on the fraction of outliers.*

Proof Suppose there are l_4 outliers, and they satisfy $\alpha_i = \frac{\sigma}{vl}$, then $l_4 \alpha_i = \frac{l_4 \sigma}{vl}$. Moreover $l_4 \alpha_i \le \sum_{i=1}^{l} \alpha_i = 2$. Then we get $\frac{l_4 \sigma}{vl} \le 2$, that is $\frac{2v}{\sigma} \ge \frac{l_4}{l}$. That is to say, $\frac{2v}{\sigma}$ is an upper bound on the fraction of outliers. □

Both parameters v and σ control the number of outliers and the width of the boundary of the rough margin. The smaller v value is, the narrower rough margin is, which implies smaller proportion of points cannot be actually determined their class labels.

4 Numerical experiments

In this section, we conduct experiments on one artificial dataset and eight benchmark datasets. In the artificial experiment, we verify the property of parameters v, σ in Rough one class SVM. In the eight benchmark experiments, we investigate the validity of the proposed algorithm from both accuracy and time aspects.

4.1 Experiment on artificial dataset

We randomly generate two classes data points, and they follow Gaussian distributions. The first class samples $X_1 \sim N(3.5, 1)$, and the second class samples $X_2 \sim N(1, 1)$. There are 100 points in all in the first class and 8 points in the second class. Their distributions are shown in Fig. 4. In the process of learning, the 100 points are considered usual points, and the other points are considered outliers.

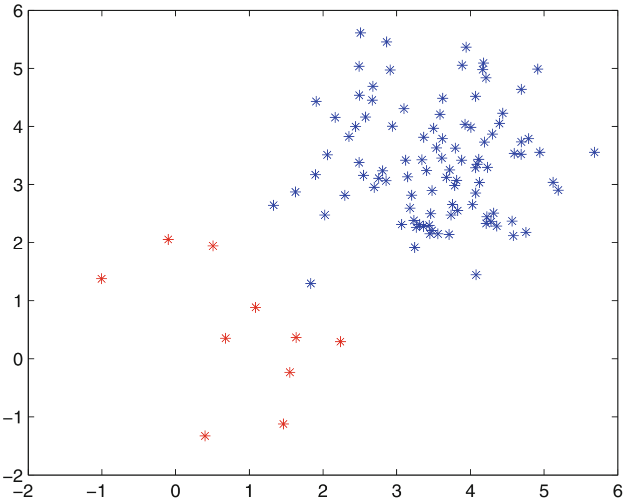

Fig. 4 The distribution of artificial data

In Fig. 5, the x-axis denotes the values of parameter σ, and the y-axis denotes the average accuracy of all of the testing samples. From Fig. 5, we can find that Rough one class SVM produces great generalization performance when parameter σ ranges from 2 to 128. However, the proposed algorithm produces over-fitting problem when a larger value is set to parameter σ. It is helpful to the choice of parameter σ in our proposed algorithm. The relationship between the average accuracy of the Rough one class SVM and parameter σ is shown in Fig. 5.

In Fig. 6, the x-axis denotes the values of parameter σ. The green curve denotes the changing curve between the fraction of support vectors and parameter σ. The red line denotes the changing curve between the values of $\frac{2v}{\sigma}$ and

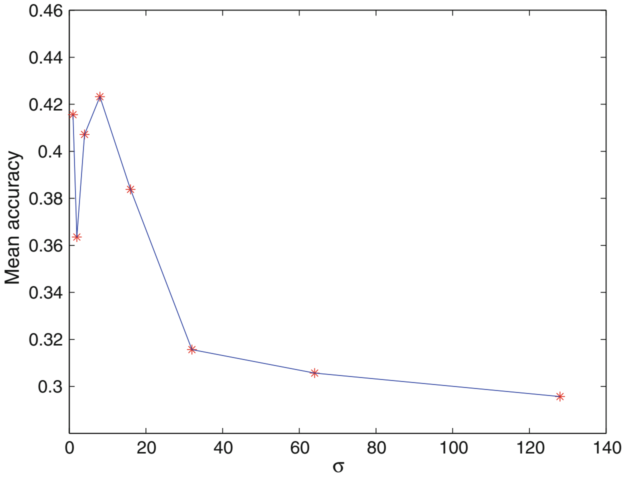

Fig. 5 The relationship between the performance of the Rough one class SVM and parameter σ

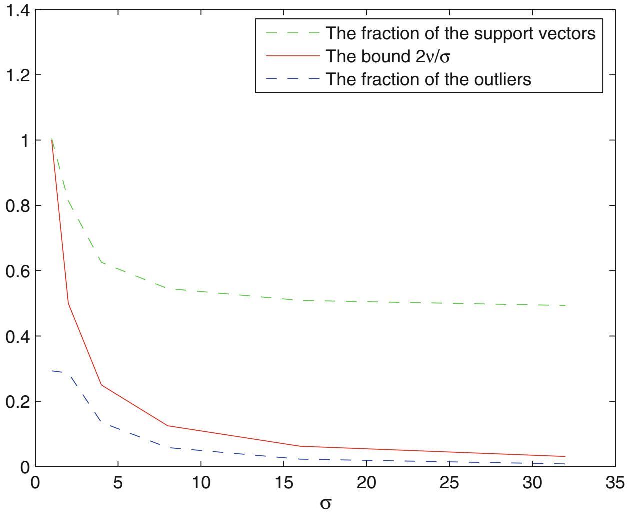

Fig. 6 The *red curve* denotes a bound($\frac{2v}{\sigma}$). The *green curve* denotes the fraction of support vectors. The *blue curve* denotes the fraction of outliers in Rough one class SVM (color figure online)

Table 1 Eight benchmark datasets

Dataset	Feature	Positive samples	Negative samples
Breast Cancer1	33	148	46
Spectf heart	44	212	55
Breast Cancer2	31	357	212
Ionosphere	34	225	126
Pima	8	500	268
Liver disorder	6	200	145
Heart	13	150	120
Sonar	60	97	111

parameter σ. The blue line denotes the changing curve between the fraction of outliers and parameter σ.

Figure 6 shows the correctness of Propositions 5 and 6. That is to say, $\frac{2v}{\sigma}$ is a lower bound on the fraction of support vectors and an upper bound on the fraction of outliers.

4.2 Experiments on benchmark datasets

In this section, we conduct experiments on eight benchmark datasets to investigate the effectiveness of the Rough one class SVM. The datasets come from UCI machine learning repository,[1] which are Breast Cancer1, Spectf heart, Breast Cancer2, Ionosphere, Pima, Liver disorder, heart, and Sonar showed in Table 1. For the experiment on each dataset, we use fivefold cross-validation to evaluate the performance of our proposed algorithm and one class SVM. That is to say, the dataset is split randomly into five subsets, and one of those sets is reserved as a test set; this process is repeated five times. All of the algorithms are implemented in Matlab 7.9 (R2009b).

From Table 1, we can find that most datasets exist the class imbalance between the two classes samples; therefore, one class SVM is employed and Rough one class SVM is researched in this paper. In all of the experiments, we take the positive samples as target class samples and the negative samples as outliers.

We know that the performance of the algorithm depends heavily on the choices of parameters. In our experiments, we only consider the Gaussian kernel function

$$k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/p^2)$$

for these datasets. We selected the optimal values for the parameters by the grid search method. In our experiments, the Gaussian kernel parameter p was selected from the set $\{2^i | i = -4, -3, \ldots, 10\}$. The optimal values for v were selected from the range $\left\{\frac{i}{10} | i = 1, 2, \ldots, 9\right\}$. The penalty parameter σ was selected from the set $\{2^i | i = 1, 2, \ldots, 7\}$.

The experimental results are shown in Table 2. Where "Accuracy" denotes the mean value of five times testing results and plus or minus the standard deviation, "Accuracy$_1$" denotes the testing accuracy of all of the samples, "Accuracy$_2$" denotes the testing accuracy of the target samples, and "Accuracy$_3$" denotes the testing accuracy of the outliers, respectively. "Time" denotes the mean value of five experimental times, each experimental time consists of training time and testing time.

5 Result comparisons

We compare the performance of Rough one class SVM with one class SVM and can easily draw the following conclusions:

(1) From the perspective of prediction accuracy, we can find that Rough one class SVM yields better prediction performance than conventional one class SVM except Breast Cancer1 dataset. They yield the comparable performance on Ionosphere dataset.

(2) In terms of computation time, although a new parameter σ was introduced into the Rough one class SVM, it does not increase the CPU time. Namely, the new parameter σ does not increase the computational complexity of our proposed algorithm.

(3) We also compare our experimental results with those shown in [15], we find that our experimental results are lower than those in [15]. The main reason lies in that our experimental results are obtained under the case of unsupervised learning, but the results in [15] are obtained under the case of supervised learning.

[1] http://archive.ics.uci.edu/ml/datasets.html.

Table 2 Performance comparisons between one class SVM and Rough one class SVM

Datasets	One class SVM					Rough one class SVM				
	(p, ν)	Accuracy$_1$	Accuracy$_2$	Accuracy$_3$	Time	(p, ν, σ)	Accuracy$_1$	Accuracy$_2$	Accuracy$_3$	Time
Breast Cancer1	(1024, 0.1)	74.73 ± 3.43	95.95 ± 3.88	6.44 ± 5.27	0.87	(1024, 0.1, 4)	72.16 ± 3.03	92.57 ± 3.87	6.44 ± 5.27	0.82
Spectf heart	(128, 0.1)	69.63 ± 4.86	87.68 ± 5.96	0.00 ± 0.00	1.72	(512, 0.3, 32)	75.64 ± 2.23	95.26 ± 2.62	0.00 ± 0.00	1.79
Breast Cancer2	(64, 0.2)	89.66 ± 2.88	90.22 ± 3.80	88.72 ± 2.91	22.80	(128, 0.4, 16)	90.20 ± 3.56	95.27 ± 2.66	81.69 ± 5.47	15.70
Ionosphere	(2, 0.1)	80.33 ± 2.66	80.44 ± 4.30	80.12 ± 10.17	3.64	(2, 0.1, 2)	80.33 ± 2.66	80.44 ± 4.30	80.12 ± 10.17	3.71
Pima	(128, 0.1)	67.60 ± 3.31	80.60 ± 3.26	43.47 ± 7.80	31.46	(128, 0.1, 2)	68.00 ± 3.23	81.20 ± 2.85	43.47 ± 7.80	32.26
Liver disorder	(8, 0.6)	58.55 ± 4.05	20.00 ± 10.32	86.50 ± 3.74	4.02	(8, 0.8, 2)	58.96 ± 2.49	11.72 ± 7.74	92.00 ± 3.67	4.17
Heart	(4, 0.9)	59.26 ± 3.88	69.33 ± 7.11	46.67 ± 11.90	2.31	(64, 0.5, 4)	61.85 ± 5.18	84.00 ± 12.0	34.17 ± 22.88	1.82
Sonar	(0.5, 0.9)	53.92 ± 9.06	7.00 ± 11.66	95.00 ± 4.47	1.27	(1, 0.5, 2)	55.42 ± 6.66	5.0 ± 10.0	100.0 ± 0.0	0.93

6 Conclusion

A novel rough margin-based one class SVM is proposed to avoid the over-fitting problem in this paper. More data points are adaptively considered in constructing the separating hyper-plane rather than few data points in one class SVM. Moreover, different penalties are proposed to give the samples depending on their positions since the points in lower margin have more effects than those in the boundary of rough margin. Finally, the proposed algorithm yields higher prediction accuracy than one class SVM. Certainly, the combination of rough set and SVM needs further research.

Acknowledgments The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. This work was supported by the National Natural Science Foundation of China (Grant No. 61153003, 11171346).

References

1. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
2. Ripley B (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
3. Manevitz LM, Yousef M (2001) One-class SVMs for document classification. J Mach Learn Res 2(1):139–154
4. Adankon MM, Cheriet M (2010) Genetic algorithm–based training for semi-supervised SVM. Neural Comput Appl 19(8): 1197–1026
5. Ben-Hur A, Horn D, Siegelmann HT, Vapnik V (2002) Support vector clustering. J Mach Learn Res 2:125–137
6. Schölkopf B, Platt J, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13(7):1443–1471
7. Choi Y-S (2009) Least squares one-class support vector machine. Pattern Recogn Lett 30(13):1236–1240
8. Hao P (2008) Fuzzy one-class support vector machines. Fuzzy Sets Syst 159(18):2317–2336
9. Bicego M, Figueiredo MAT (2009) Soft clustering using weighted one-class support vector machines. Pattern Recogn 42(1):27–32
10. Xu Y (2009) Classification algorithm based on feature selection and samples selection. Lect Notes Comput Sci 5552:631–638
11. Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11:341–356
12. Pawlak Z (2002) Rough sets and intelligent data analysis. Inf Sci 147(1):1–12
13. Asharaf S, Shevade SK, Narasimha murty M (2005) Rough support vector clustering. Pattern Recogn 38(10):1779–1783
14. Zhang J, Wang Y (2008) A rough margin based on support vector machine. Inf Sci 178(9):2204–2214
15. Xu Y, Wang L (2011) A rough margin-based ν-twin support vector machine. Neural Comput Appl. doi:10.1007/s00521-011-0565-y