Neural Comput & Applic (2014) 25:1303–1311
DOI 10.1007/s00521-014-1611-3

ORIGINAL ARTICLE

Learning with positive and unlabeled examples using biased twin support vector machine

Zhijie Xu · Zhiquan Qi · Jianqin Zhang

Received: 23 December 2013 / Accepted: 28 April 2014 / Published online: 24 May 2014

Abstract PU classification problem ('P' stands for positive, 'U' stands for unlabeled), which is defined as the training set consists of a collection of positive and unlabeled examples, has become a research hot spot recently. In this paper, we design a new classification algorithm to solve the PU problem: biased twin support vector machine (B-TWSVM). In B-TWSVM, two nonparallel hyperplanes are constructed such that the positive examples can be classified correctly, and the number of unlabeled examples classified as positive is minimized. Moreover, considering that the unlabeled set also contains positive data, different penalty parameters for positive and negative data are allowed in B-TWSVM. Experimental results demonstrate that our method outperforms the state-of-the-art methods in most cases.

Keywords Machine learning · Classification · Support vector machine

Z. Xu (✉)
School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, China
e-mail: zhijiexu@163.com

Z. Qi
Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
e-mail: qizhiquan@gucas.ac.cn

J. Zhang
Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
e-mail: zhangjianqin@bucea.edu.cn

1 Introduction

Learning classifiers from a combination of labeled and unlabeled data are an old research subject in machine learning. In some applications, such as text classification and image classification, obtaining a large number of labeled training data is labor intensive and time-consuming. Since learning from labeled and unlabeled data reduces manual labeling effort, it has drawn a lot of attention from these research fields.

In this paper, we study the problem of learning with positive and unlabeled examples, but there is no negative example. Several works have been done for solving this problem in machine learning. The most popular among them are the two-step strategy-based methods, which include S-EM [1], PEBL [2], and Roc-SVM [3]. In these algorithms, a set of reliable negative examples is firstly identified from the unlabeled examples. And in step 2, a set of classifiers is built by iteratively applying a classification algorithm, then a good classifier is selected from the set. The methods for step 1 include the Spy technique [1], the 1-DNF technique [2], the Rocchio algorithm [3, 4], and the naïve Bayesian technique [5]. The methods for step 2 include the Expectation Maximization (EM) algorithm [6] and support vector machines, which contain SVM alone [5], Iterative SVM [2], and Iterative SVM with Classifier Selection [3]. An evaluation of all 16 possible combinations of methods of step 1 and step 2 is also performed in [5], and a benchmark system, called Learning from positive and unlabeled data (LPU) is obtained. In [5], an approach based on a biased formulation of SVM (B-SVM) is proposed for solving this problem. Experimental results show that the method is superior to all the existing two-step techniques.

Different from SVM with two parallel hyperplanes, some nonparallel hyperplane classifiers have been proposed recently, such as the generalized eigenvalue proximal support vector machine (GEPSVM) [7] and the twin support vector machine (TWSVM) [8]. TWSVM seeks two nonparallel proximal hyperplanes such that each hyperplane is closer to one of two classes and as far as possible from the other class. It is implemented by solving two smaller quadratic programming problems (QPPs) rather than a single large QPP in the classical SVM. Experimental results in [8] and [9] have shown that TWSVM outperforms both standard SVM and GEPSVM in the most cases. In addition, TWSVM is excellent at dealing with some certain probability model data (such as Cross Planes data). Thus, the methods of constructing the nonparallel hyperplanes and the extensions of TWSVM have been studied extensively [9–18].

Inspired by the success of B-SVM and TWSVM, in this paper, we would like to solve the problem of learning classifiers from a combination of positive and unlabeled data using TWSVM. This leads to a biased formulation of TWSVM, which is named B-TWSVM. Similar to TWSVM, B-TWSVM constructs two nonparallel hyperplanes, and a new point is assigned to the class whose hyperplane the point is closer to. But there are some differences: (1) TWSVM seeks two nonparallel planes such that each plane is closer to one of two classes and is at least one distance from the other. It is formulated as two QPPs. Whereas in our B-TWSVM, the two hyperplanes are constructed such that each hyperplane is closer to one of two classes, that is, the positive points are closer to the positive plane, and the negative points are closer to the negative plane. It can be implemented by solving a single QPP. (2) Considering that the unlabeled set also contains positive data, B-TWSVM allows to set different penalty parameters for positive data and negative data, so it can weight positive errors and negative errors differently.

The remainder of this paper is organized as follows. In Sect. 2, SVM and TWSVM are introduced briefly. The details of B-TWSVM are described in Sect. 3. In Sect. 4, experimental results are presented and discussed. Finally, conclusions are presented in Sect. 5.

2 Background

In this section, we give a brief outline of SVC and TWSVM.

2.1 Support vector classification (SVC) [19]

The support vector machine (SVM) is a modern learning machine with very good generalization performance in pattern recognition and regression estimation. Here, we describe the ideas of SVM for pattern recognition briefly, and more details can be found in [19, 20]. Support vector classification is formulated to construct binary classifiers. Given the training dataset:

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\}, x_i \in R^n, y_i \in \{-1, 1\}, \; i = 1, \ldots, l, \tag{1}$$

linear SVC constructs an optimal separating hyperplane given by $(w \cdot x) + b = 0$ in the feature space. The computation of this hyperplane relies on the maximization of the margin, which is modeled as follows:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|_2^2, \\ &\text{s.t.} y_i((w \cdot x_i) + b) \geq 0, \; i = 1, \ldots, l. \end{aligned} \tag{2}$$

For nonseparable data, a set of slack variables ξ_i is introduced to allow errors and a penalty parameter C is used to tune the trade-off between allowing errors and maximization of the margin:

$$\begin{aligned} &\min_{w,b,\xi_i} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{l} \xi_i, \\ &\text{s.t.} y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \\ &\quad \xi_i \geq 0, \; i = 1, \ldots, l. \end{aligned} \tag{3}$$

To solve the problem, the dual problem is formulated as:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{j=1}^{l} \alpha_j, \\ &\text{s.t.} \sum_{i=1}^{l} y_i \alpha_i = 0, \\ &\quad 0 \leq \alpha_i \leq C, \; i = 1, \ldots, l, \end{aligned} \tag{4}$$

where $\alpha \in R^l$ are lagrangian multipliers.

For nonlinear classification, the data are mapped to a higher dimensional feature space, and the optimal separating hyperplane is computed in the feature space using a kernel function. This results in a nonlinear decision function in the input space. Using the kernel function, the optimal separating hyperplane can be determined without any computations in the higher dimensional feature space. The problem is modeled as:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{j=1}^{l} \alpha_j, \\ &\text{s.t.} \sum_{i=1}^{l} y_i \alpha_i = 0, \\ &\quad 0 \leq \alpha_i \leq C, \; i = 1, \ldots, l. \end{aligned} \tag{5}$$

The decision function is given by

$$f(x) = \text{sgn}\left(\sum_{i=1}^{l} \alpha_i^* y_i K(x, x_i) + b^*\right), \tag{6}$$

where α^* is the solution of the dual problem (5), b^* is given by

$$b^* = \frac{1}{N_{sv}}\left(y_j - \sum_{i=1}^{N_{sv}} y_i \alpha_i^* K(x_i, x_j)\right), \tag{7}$$

where N_{sv} represents the number of support vectors satisfying $0 \leq \alpha \leq C$.

2.2 Twin support vector machine (TWSVM) [8]

Given the following training set for the binary classification:

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\}, \tag{8}$$

where (x_i, y_i) is the i-th data point, the input $x_i \in R^n$ is a pattern, the output $y_i \in \{-1, 1\}$ is a class label, $i = 1, \ldots, l$, and l is the number of data points. In addition, let l_1 and l_2 be the number of data points in positive class and negative class, respectively ($l = l_1 + l_2$). Furthermore, the matrix $A \in R^{l_1 \times n}$ and $B \in R^{l_2 \times n}$ consist of the l_1 inputs of positive class and the l_2 inputs of negative class, respectively.

For the linear case, TWSVM seeks to find two nonparallel hyperplanes in n-dimensional input space

$$(w_+ \cdot x) + b_+ = 0 \text{ and } (w_- \cdot x) + b_- = 0, \tag{9}$$

where $w_+, w_- \in R^n, b_+, b_- \in R$. Here, each hyperplane is close to the examples of one class and far away from the examples of the other class. TWSVM is in spirit of GEPSVM [7]. But both of GEPSVM and TWSVM are different from the standard SVM. For TWSVM, each hyperplane is generated by solving a QP problem looking like the primal problem of the standard SVM. The primal problems of TWSVM can be presented as follows:

$$\begin{aligned} \min_{w_+, b_+, \xi} \ & \frac{1}{2}\|A_+ w_+ + e_+ b_+\|_2^2 + C_1 e_-^{\mathrm{T}} \xi, \\ \text{s.t.} \ & -(B w_+ + e_- b_+) + \xi \geq e_-, \\ & \xi \geq 0, \end{aligned} \tag{10}$$

and

$$\begin{aligned} \min_{w_-, b_-, \eta} \ & \frac{1}{2}\|B w_- + e_- b_-\|_2^2 + C_2 e_+^{\mathrm{T}} \eta, \\ \text{s.t.} \ & (A w_- + e_+ b_-) + \eta \geq e_+, \\ & \eta \geq 0, \end{aligned} \tag{11}$$

where C_1 and C_2 are nonnegative parameters, and e_+ and e_- are vectors of ones of appropriate dimensions. In the QP problem (10), the objective function tends to keep the positive hyperplane close to the examples of positive class and the constraints require the hyperplane to be at a distance of at least 1 from the examples of negative class. The QP problem (11) has the similar property.

Once the solutions (w_+, b_+) and (w_-, b_-) of the problems (10) and (11) are obtained, a new point $x \in R^n$ is assigned to class i ($i = +1, -1$), depending on which of the two hyperplanes in (9) is closer to, i.e.,

$$\text{class } i = \arg \min_{k=+,-} \frac{|w_k^{\mathrm{T}} x + b_k|}{\|w_k\|_2}, \tag{12}$$

where $|\cdot|$ is the absolute value. For the nonlinear case, we can refer to [8]. As for K-class classification problem, multi-TWSVM [21] constructs K hyperplanes, each hyperplane close to one of the classes, and at the same time, far from the other classes. A new point will be classified to either of the classes according to the distances to the hyperplanes.

3 Biased twin support vector machine

3.1 Linear case

We now present the proposed biased twin support vector machine for the problem of learning with positive and unlabeled examples. Suppose the training set be

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\}, \tag{13}$$

where x_i is an input vector, $x_i \in R^n$, and y_i is its class label, $y_i \in \{1, -1\}$, $i = 1, \ldots, l$, and l is the number of data points. In addition, assume that the first k examples are positive examples (labeled 1), while the rest $l - k$ examples are unlabeled examples, which we label negative (-1). The goal of B-TWSVM was to find two nonparallel hyperplanes:

$$w_+^{\mathrm{T}} x + b_+ = 0, \ w_-^{\mathrm{T}} x + b_- = 0, \tag{14}$$

such that the label of a new point $x \in R^n$ can be inferred according to which hyperplane it is closer to. From this point of view, the two hyperplanes should represent the two classes of patterns. So we seek these two hyperplanes such that each of them is closer to one of two classes. On the other hand, we expect the positive examples to be classified correctly, so these examples should be closer to the positive hyperplane. In addition, the number of unlabeled examples classified as positive should be minimized, so as many as possible unlabeled examples should be closer to the negative hyperplane. This would give a good classifier, which has been shown theoretically in [1]. Formally, to find the the positive and negative hyperplanes, B-TWSVM considers the following primal problem (no error for positive points but only for unlabeled ones):

$$\min_{w_+,b_+,w_-,b_-,\xi} \frac{1}{2}\|Aw_+ + e_+b_+\|_2^2 + \frac{1}{2}C\|Bw_- + e_-b_-\|_2^2 + C_1e_-^{\mathrm{T}}\xi,$$
$$\text{s.t.} Bw_- + e_-b_- + (Bw_+ + e_-b_+) - \xi \le 0, \xi \ge 0,$$
$$Aw_+ + e_+b_+ + (Aw_- + e_+b_-) \ge 0. \tag{15}$$

If noise in positive dataset is also considered, the following soft margin version of the above problem can be obtained, which uses two parameters C_1 and C_2 to weight negative error and positive error differently:

$$\min_{w_+,b_+,w_-,b_-,\xi} \frac{1}{2}\|Aw_+ + e_+b_+\|_2^2 + \frac{1}{2}C\|Bw_- + e_-b_-\|_2^2 + C_1e_-^{\mathrm{T}}\xi + C_2e_+^{\mathrm{T}}\eta,$$
$$\text{s.t.} Bw_- + e_-b_- + (Bw_+ + e_-b_+) - \xi \le 0, \xi \ge 0,$$
$$Aw_+ + e_+b_+ + (Aw_- + e_+b_-) + \eta \ge 0, \eta \ge 0, \tag{16}$$

where C, C_1 and C_2 are nonnegative parameters, and e_+ and e_- are vectors of ones of appropriate dimensions. Considering that the unlabeled set, which is assumed to be negative, also contains positive data, we can give a big value for C_2 (penalty parameter for positive data)and a small value for C_1 (penalty parameter for negative data). In order to solve the problem (16), we need to derive its dual problem. The Lagrangian corresponding to the problem (16) is given by

$$L(\Theta) = \frac{1}{2}\|Aw_+ + e_+b_+\|_2^2 + \frac{1}{2}C\|Bw_- + e_-b_-\|_2^2 + C_1e_-^{\mathrm{T}}\xi + C_2e_+^{\mathrm{T}}\eta$$
$$+ \alpha^{\mathrm{T}}(Bw_- + e_-b_- + (Bw_+ + e_-b_+) - \xi) - \beta^{\mathrm{T}}\xi$$
$$- \mu^{\mathrm{T}}(Aw_+ + e_+b_+ + (Aw_- + e_+b_-) + \eta) - \gamma^{\mathrm{T}}\eta, \tag{17}$$

where $\Theta = \{w_+, b_+, w_-, b_-, \xi, \eta, \alpha, \beta, \mu, \gamma\}$, $\alpha = (\alpha_1, \ldots, \alpha_{l-k})^{\mathrm{T}}$, $\beta = (\beta_1, \ldots, \beta_{l-k})^{\mathrm{T}}$, $\mu = (\mu_1, \ldots, \mu_k)^{\mathrm{T}}$, $\gamma = (\gamma_1, \ldots, \gamma_k)^{\mathrm{T}}$ are the Lagrange multipliers. The dual problem can be formulated as

$$\max_{\Theta} L(\Theta)$$
$$\text{s.t.} \nabla_{w_+,b_+,w_-,b_-,\xi,\eta} L(\Theta) = 0 \tag{18}$$
$$\alpha, \beta, \mu, \gamma \ge 0.$$

From equation (18), we get

$$\nabla_{w_+} L = A^{\mathrm{T}}((Aw_+ + e_+b_+) + B^{\mathrm{T}}\alpha - A^{\mathrm{T}}\mu = 0 \tag{19}$$

$$\nabla_{b_+} L = e_+^{\mathrm{T}}((Aw_+ + e_+b_+) + e_-^{\mathrm{T}}\alpha - e_+^{\mathrm{T}}\mu = 0 \tag{20}$$

$$\nabla_{w_-} L = CB^{\mathrm{T}}((Bw_- + e_-b_-) + B^{\mathrm{T}}\alpha - A^{\mathrm{T}}\mu = 0 \tag{21}$$

$$\nabla_{b_-} L = Ce_-^{\mathrm{T}}((Bw_- + e_-b_-) + e_-^{\mathrm{T}}\alpha - e_+^{\mathrm{T}}\mu = 0 \tag{22}$$

$$\nabla_{\xi} L = C_1e_- - \alpha - \beta = 0 \tag{23}$$

$$\nabla_{\eta} L = C_2e_+ - \mu - \gamma = 0 \tag{24}$$

Combining (19) and (20), we get

$$[A^{\mathrm{T}}\ e_+^{\mathrm{T}}]^{\mathrm{T}}[A\ e_+][w_+\ b_+]^{\mathrm{T}} + [B^{\mathrm{T}}\ e_-^{\mathrm{T}}]^{\mathrm{T}}\alpha - [A^{\mathrm{T}}\ e_+^{\mathrm{T}}]^{\mathrm{T}}\mu = 0 \tag{25}$$

Let $H = [A\ e_+], G = [B\ e_-]$, and the augmented vector $v_+ = [w_+\ b_+]^{\mathrm{T}}$, the equality can be rewritten as:

$$H^{\mathrm{T}}Hv_+ + G^{\mathrm{T}}\alpha - H^{\mathrm{T}}\mu = 0, \tag{26}$$

$$\text{i.e.} v_+ = -(H^{\mathrm{T}}H)^{-1}(G^{\mathrm{T}}\alpha - H^{\mathrm{T}}\mu). \tag{27}$$

Next, combining (21) and (22) leads to

$$C[B^{\mathrm{T}}\ e_-^{\mathrm{T}}]^{\mathrm{T}}[B\ e_-][w_-\ b_-]^{\mathrm{T}} + [B^{\mathrm{T}}\ e_-^{\mathrm{T}}]^{\mathrm{T}}\alpha - [A^{\mathrm{T}}\ e_+^{\mathrm{T}}]^{\mathrm{T}}\mu = 0. \tag{28}$$

Let the augmented vector $v_- = [w_-\ b_-]^{\mathrm{T}}$, Similarly, the equality can be rewritten as:

$$G^{\mathrm{T}}Gv_- + G^{\mathrm{T}}\alpha - H^{\mathrm{T}}\mu = 0, \tag{29}$$

$$\text{i.e.} \mathrm{v}_- = -(\mathrm{G}^{\mathrm{T}}\mathrm{G})^{-1}(\mathrm{G}^{\mathrm{T}}\alpha - \mathrm{H}^{\mathrm{T}}\mu). \tag{30}$$

Since $\beta, \gamma \ge 0$, (23) and (24) turn to be

$$0 \le \alpha \le C_1e_-, \tag{31}$$

$$0 \le \mu \le C_2e_+, \tag{32}$$

According to the dual theory of optimization problem, the Wolfe dual of the problem can be expressed as:

$$\max_{\alpha,\mu} -\frac{1}{2}(\alpha^{\mathrm{T}}G - \mu^{\mathrm{T}}H)[(H^{\mathrm{T}}H)^{-1} + (2-C)(G^{\mathrm{T}}G)^{-1}](G^{\mathrm{T}}\alpha - H^{\mathrm{T}}\mu)$$
$$\text{s.t.} 0 \le \alpha \le C_1e_-,$$
$$0 \le \mu \le C_2e_+, \tag{33}$$

where $H = [A\ e_+], G = [B\ e_-]$. This is a QPP, which can be solved by the quadratic programming solver in the MATLAB Optimization Toolbox. And the augmented vector $v_+ = [w_+\ b_+]^{\mathrm{T}}$ and $v_- = [w_-\ b_-]^{\mathrm{T}}$ can be computed by (27) and (30). Once vectors v_+ and v_- are obtained from (27) and (30), the separating planes

$$w_+^{\mathrm{T}}x + b_+ = 0 \text{ and } w_-^{\mathrm{T}}x + b_- = 0 \tag{34}$$

are known. A new point $x \in R^n$ is then assigned to class i ($i = 1, -1$), depending on which of the two hyperplanes in (14) it is closer to, i.e.,

$$\text{Class } i = \arg\min_{k=+,-} \frac{|w_k^{\mathrm{T}}x + b_k|}{\|w_k\|}, \tag{35}$$

where $|\cdot|$ is the absolute value.

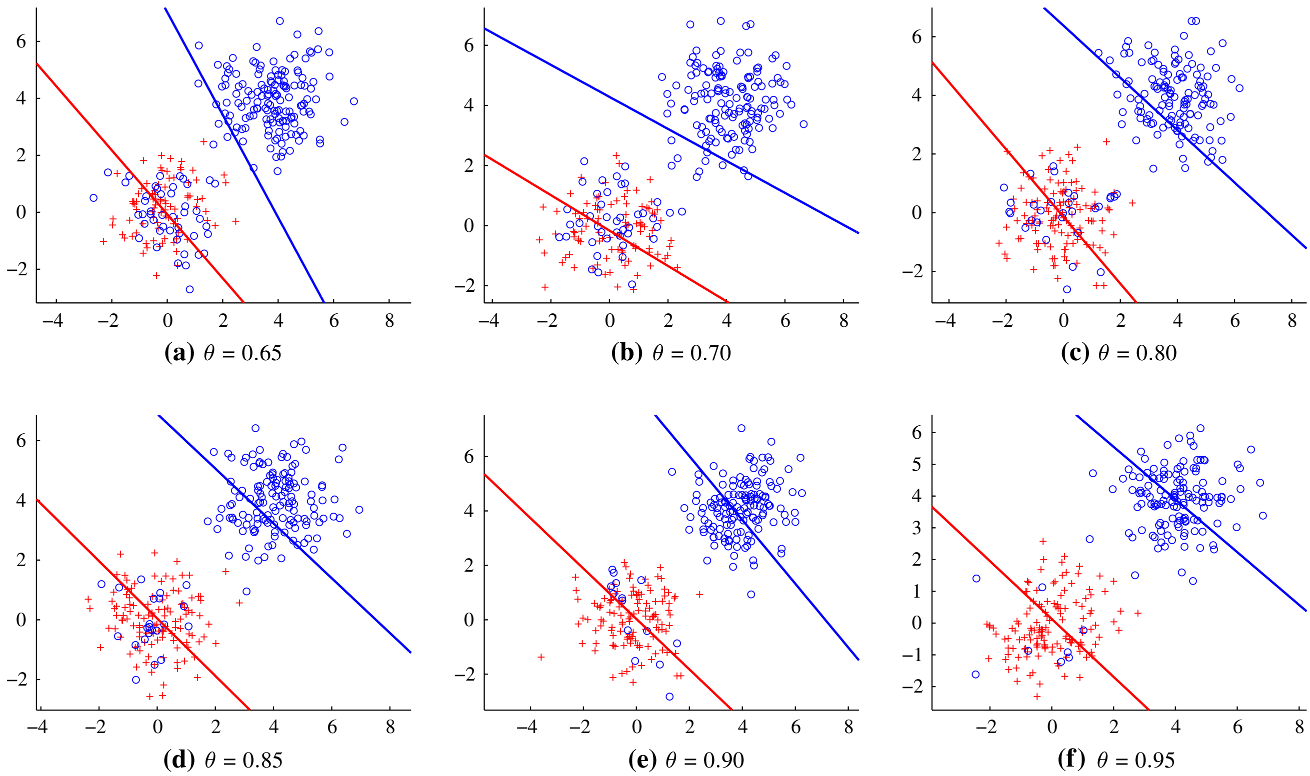


Fig. 1 The results of B-TWSVM on the training set

3.2 Nonlinear case

The above discussion is restricted in the linear case. Here, we will analyze nonlinear biased twin support vector machine by introducing Gaussian kernel function.

$$K(x,x^{\mathrm{T}}) = exp(-\|x - x^{\mathrm{T}}\|^2/2\sigma^2), \tag{36}$$

where σ is a real parameter, and the corresponding transformation:

$$\mathbf{X} = \Phi(\mathbf{x}), \tag{37}$$

where $\mathbf{X} \in \mathbf{H}$, H is a Hilbert space. So the training set becomes

$$\tilde{T} = \{(\Phi(x_1), y_1), \ldots, (\Phi(x_l), y_l)\}, x_i \in R^n, y_i \in \{-1, 1\},\\ i = 1, \ldots, l. \tag{38}$$

We consider the following kernel-generated hyperplanes:

$$\begin{aligned} K(x^{\mathrm{T}}, O^{\mathrm{T}})w_+ + b_+ = 0,\\ K(x^{\mathrm{T}}, O^{\mathrm{T}})w_- + b_- = 0, \end{aligned} \tag{39}$$

where $O^{\mathrm{T}} = [A\ B]^{\mathrm{T}}$ and K is a chosen kernel function. The nonlinear optimization problem can be expressed as

$$\begin{aligned} &\min_{w_+,b_+,w_-,b_-,\xi,\eta} \frac{1}{2}\|K(A,O^{\mathrm{T}})w_+ + e_+b_+\|_2^2 + \frac{1}{2}C\|K(B,O^{\mathrm{T}})w_- + e_-b_-\|_2^2\\ &+ C_1e_-^{\mathrm{T}}\xi + C_2e_+^{\mathrm{T}}\eta,\\ &\text{s.t.} K(B,O^{\mathrm{T}})w_- + e_-b_- + (K(B,O^{\mathrm{T}})w_+ + e_-b_+) - \xi \le 0, \xi \ge 0,\\ &K(A,O^{\mathrm{T}})w_+ + e_+b_+ + (K(A,O^{\mathrm{T}})w_- + e_+b_-) + \eta \ge 0, \eta \ge 0 \end{aligned} \tag{40}$$

The Wolfe dual of the problem (40) can be expressed as:

$$\begin{aligned} &\max_{\alpha,\mu} -\frac{1}{2}(\alpha^{\mathrm{T}}G_\Phi - \mu^{\mathrm{T}}H_\Phi)[(H_\Phi^{\mathrm{T}}H_\Phi)^{-1} + (2 - C)(G_\Phi^{\mathrm{T}}G_\Phi)^{-1}]\\ &\quad (G_\Phi^{\mathrm{T}}\alpha - H_\Phi^{\mathrm{T}}\mu)\\ &\text{s.t.} 0 \le \alpha \le C_1e_-,\\ &\quad 0 \le \mu \le C_2e_+, \end{aligned} \tag{41}$$

where $H_\Phi = [K(A,O^{\mathrm{T}})\ e_+]$, $G_\Phi = [K(B,O^{\mathrm{T}})\ e_-]$, and the augmented vector $v_+ = [w_+\ b_+]^{\mathrm{T}}$, $v_- = [w_-\ b_-]^{\mathrm{T}}$ are given by

$$v_+ = -(H_\Phi^{\mathrm{T}}H_\Phi)^{-1}(G_\Phi^{\mathrm{T}}\alpha - H_\Phi^{\mathrm{T}}\mu), \tag{42}$$

$$v_- = -(G_\Phi^{\mathrm{T}}G_\Phi)^{-1}(G_\Phi^{\mathrm{T}}\alpha - H_\Phi^{\mathrm{T}}\mu). \tag{43}$$

Once vectors v_+ and v_- are obtained from (42) and (43), a new data point $x \in R^n$ is then assigned to class i $(i = 1, -1)$ by

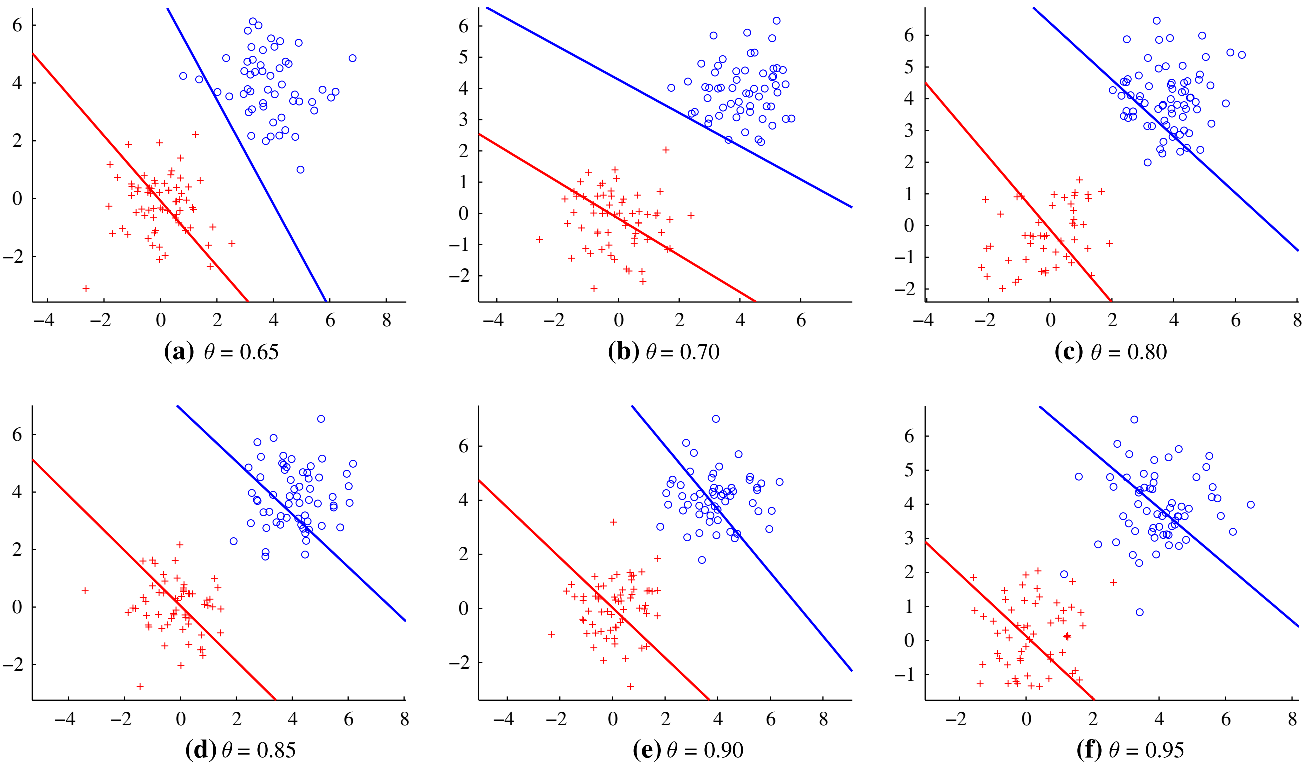


Fig. 2 The results of B-TWSVM on the testing set

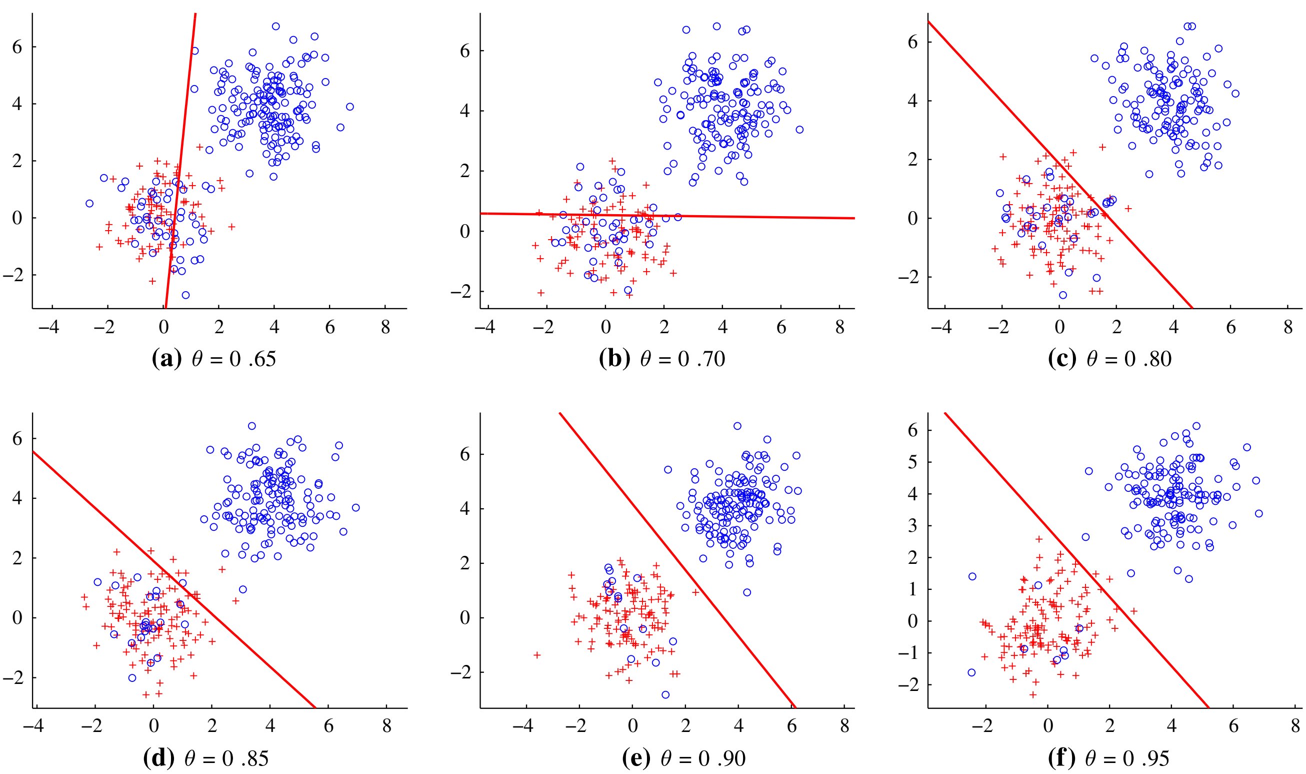
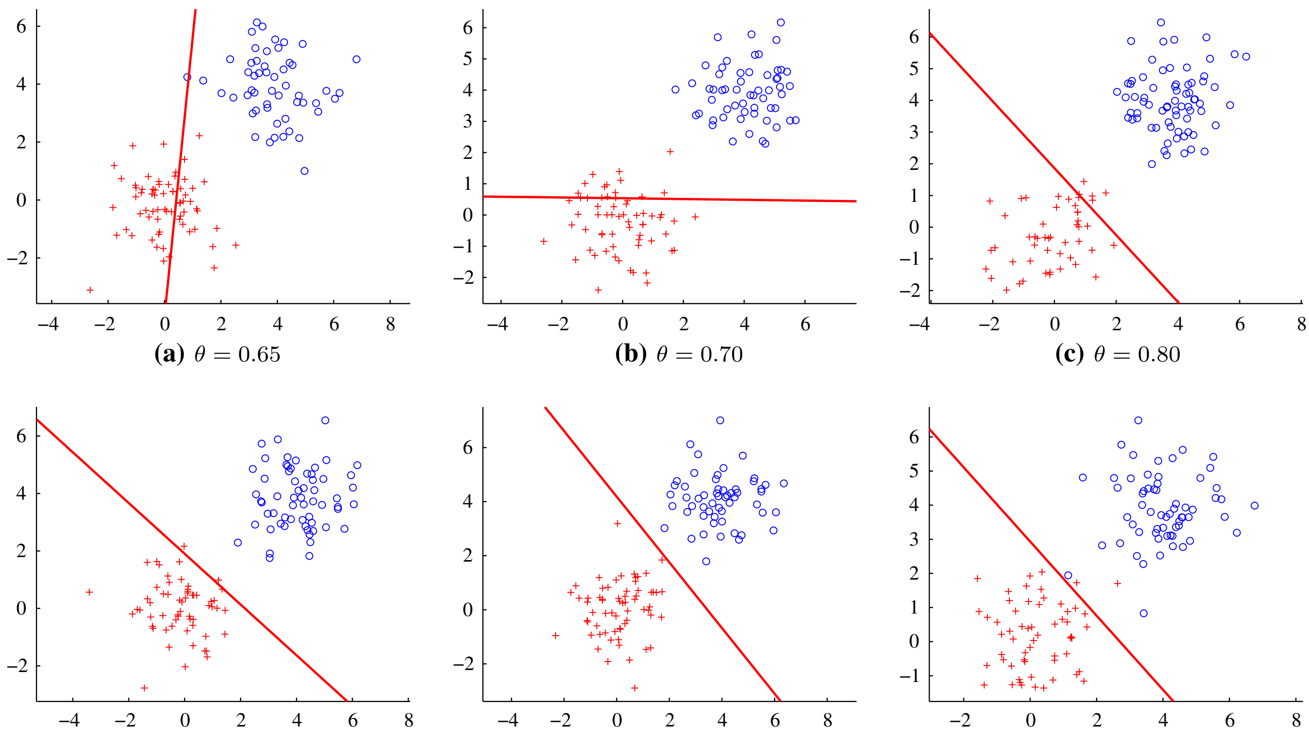


Fig. 3 The results of B-SVM on the training set



(a) $\theta = 0.65$ **(b)** $\theta = 0.70$ **(c)** $\theta = 0.80$

(d) $\theta = 0.85$ **(e)** $\theta = 0.90$ **(f)** $\theta = 0.95$

Fig. 4 The results of B-SVM on the testing set

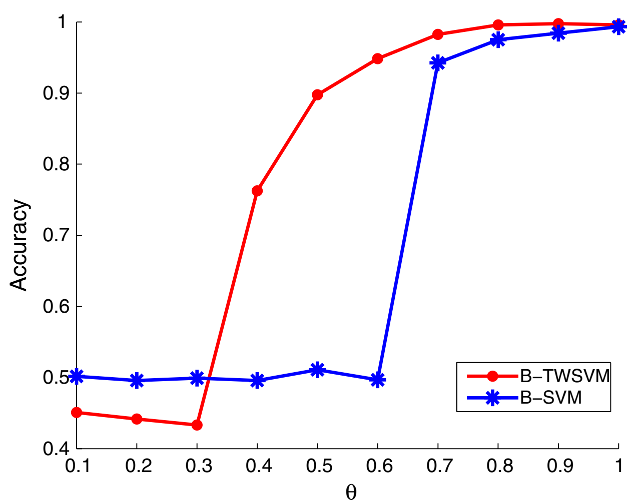


Fig. 5 The accuracy of B-TWSVM and B-SVM for different θ on the toy data

$$\text{Class } i = \arg\min_{k=+,-} \frac{|K(x^{\mathrm{T}}, O^{\mathrm{T}})w_k + b_k|}{\sqrt{w_k^{\mathrm{T}} K(O, O^{\mathrm{T}}) w_k}}, \tag{44}$$

where $|\cdot|$ is the absolute value.

4 Experiments

In this section,we compare B-TWSVM against B-SVM on several datasets.

4.1 Experimental setup

Datasets We use a 2-D toy data and the UCI Machine Learning Repository datasets in our experiments.The first toy dataset contains 200 positive data and 200 negative data. All points are generated by Gaussian distribution: positive points (the mean of $\mu = [0, 0]$, the standard deviation of $\sigma = [1, 1]$) and negative points ($\mu = [4, 4]$, $\sigma = [1, 1]$). The second one is the UCI Machine Learning Repository datasets, which is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [22]. We select 9 datasets from it for experiments. For each dataset, we select one class as the positive class, and the others as the negative class. Then, we randomly select the same number of data from positive class and negative class to compose a dataset.

For each dataset, 30 % of the points are randomly selected for testing, and the rest (70 %) are used to create training sets as follows: θ percent of the points from the positive class is first selected as the positive set . The rest of the positive points and negative points are used as unlabeled set. We vary θ from 0.1 to 1 to create a wide range of scenarios.

Experimental systems All methods are implemented using MATLAB R2011b on a PC with an Intel core i3 CPU (3.2 GHz) with 4.0 GB RAM. We use the optimization

Table 1 The testing accuracy(%) and F-score on UCI datasets in the case of RBF and $\theta = 0.9$

Dataset (Ins × Fea × Class)	B-TWSVM Acc (%) F-score	B-SVM Acc (%) F-score	TWSVM Acc (%) F-score
Australian	**80.86 ± 2.42**	76.38 ± 3.06	74.59 ± 2.10
(690 × 14 × 2)	**0.8940 ± 0.0150**	0.8658 ± 0.0198	0.8543 ± 0.0139
Diabetes	**67.27 ± 4.30**	64.47 ± 3.85	64.72 ± 3.55
(768 × 8 × 2)	**0.8036 ± 0.0309**	0.7834 ± 0.0281	0.7853 ± 0.0260
Heart	**71.25 ± 5.36**	70.14 ± 4.50	69.17 ± 5.77
(270 × 13 × 2)	**0.8311 ± 0.0366**	0.8238 ± 0.03.8	0.8165 ± 0.0405
Ionosphere	**84.34 ± 5.70**	83.29 ± 4.85	76.32 ± 7.82
(351 × 34 × 2)	**0.9141 ± 0.0338**	0.9081 ± 0.0289	0.8547 ± 0.1207
Glass	70.24 ± 6.94	**72.62 ± 9.54**	72.51 ± 9.63
(214 × 9 × 6)	0.8234 ± 0.0493	**0.8381 ± 0.0667**	0.8375 ± 0.0643
Segment	92.56 ± 2.64	95.68 ± 2.58	**97.19 ± 1.04**
(960 × 19 × 7)	0.9612 ± 0.0144	0.9778 ± 0.0135	**0.9857 ± 0.0053**
Vehicle	68.24 ± 1.76	66.94 ± 3.13	**71.88 ± 4.48**
(846 × 18 × 4)	0.8111 ± 0.0124	0.8016 ± 0.0222	**0.8356 ± 0.0307**
Wine	94.44 ± 5.40	**95.00 ± 5.52**	94.28 ± 4.15
(178 × 13 × 3)	0.9707 ± 0.0290	**0.9736 ± 0.0299**	0.9654 ± 0.0218
Vowel	**91.72 ± 4.36**	89.31 ± 5.25	81.03 ± 12.51
(528 × 10 × 11)	**0.9564 ± 0.0235**	0.9428 ± 0.0290	0.8904 ± 0.0781

toolbox QP in MATLAB to solve the related optimization problems in this paper.

Evaluation measure In our experiments, we use the Accuracy and the F-score on the positive class as the evaluation measure. F-score takes into account of both Recall (r) and Precision (p), $F = 2pr/(p + r)$. Recall, Precision and Accuracy are defined as follows. Recall = T P/(T P +FN), Precision = T P/(T P+FP), Accuracy = (T P+T N)/(T P+FP+T N+FN), where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively.

4.2 Experimental results

In the first experiment, we vary θ from 0.1 to 0.9. The Linear kernel is used. For every θ, the experiment is repeated 10 times. The average accuracies of B-TWSVM and B-SVM for different value of θ are shown in Fig. 5. It is seen there that performance of both methods improves obviously as we increase the value of θ (the proportion of the points selected from the positive class as the positive dataset). Our B-TWSVM's average classification accuracy improves from 43 to 76 % when θ increases from 0.3 to 0.4, while B-SVM's average classification accuracy is around 50 % until θ increases to 0.7. It is also shown that B-TWSVM gives better accuracy than B-SVM when θ is 0.4 to 0.9, and the accuracy of them is the same when θ is 1. Figures 1, 2, 3, and 4 sketch the performance of B-TWSVM and B-SVM on the training set and testing set, respectively, in some cases. We see that when θ is 0.65 or 0.7, B-TWSVM achieves better performance on both training set and testing set. In the second experiment, we select 9 datasets from the UCI Machine Learning Repository datasets. We specify the parameter $\theta = 0.9$. For each dataset, the experiment is repeated 10 times. These average accuracies and F-scores of B-TWSVM, B-SVM, and TWSVMare summarized in Table 1. In Table 1, the best accuracy is shown by bold figures. From Table 1, it is easy to see that B-TWSVM outperforms B-SVM and TWSVM in the most cases.

5 Conclusions

For problem of learning a classifier from a combination of positive examples and unlabeled examples, a new method based on TWSVM (B-TWSVM) was proposed in this paper. The main contribution is that for improving classification accuracy, B-TWSVM applies two nonparallel hyperplanes instead of a single one in B-SVM, then a new data point is assigned to the positive or negative class depending upon its proximity to the two nonparallel hyperplanes. But it is different from TWSVM in that B-TWSVM finds two hyperplanes by solving only one QPP, while in TWSVM they are obtained by solving two different QPPs. In addition, different penalty parameters for positive dataset and unlabel dataset are allowed in our B-TWSVM, so we can give a big penalty parameter for positive examples and a small one for negative examples, considering that the unlabeled set also contains positive data. Computational comparisons between our B-TWSVM, B-SVM, and TWSVM have been made on several datasets, indicating that our B-TWSVM outperforms B-SVM and TWSVM in most cases.

Acknowledgments This work has been partially supported by grants from Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR201107123), the Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation Foundation (20111216N), Beijing Municipal Commission of Education, Science and Technology Development (KM201210016014), and Scientific Research Foundation of Beijing University of Civil Engineering and Architecture (00331609054).

References

1. Liu B, Lee WS, Yu PS, Li X (2002) Partially supervised classification of text documents. In: Machine learning-international workshop then conference, Citeseer, pp 387–394
2. Yu H, Han J, Chang KC-C (2002) Pebl: positive example based learning for web page classification using svm. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 239–248
3. Li X, Liu B (2003) Learning to classify texts using positive and unlabeled data. In: International joint conference on artificial intelligence, vol 18. Lawrence Erlbaum Associates Ltd, pp 587–594
4. Rocchio Jr J (1971) Relevance feedback in information retrieval. In: The SMART system experiments in automatic document processing. Prentice Hall, New York, pp 313–323
5. Liu B, Dai Y, Li X, Lee WS, Yu PS (2003) Building text classifiers using positive and unlabeled examples. In: Data mining, 2003. ICDM 2003. Third IEEE international conference on, IEEE, pp 179–186
6. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B (Methodological) 39(1):1–38
7. Mangasarian OL, Wild EW (2006) Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Trans Pattern Anal Mach Intell 28(1):69–74
8. Khemchandani R, Chandra S et al (2007) Twin support vector machines for pattern classification. IEEE Trans Pattern Anal Mach Intell 29(5):905–910
9. Kumar MA, Gopal M (2008) Application of smoothing technique on twin support vector machines. Pattern Recogn Lett 29(13): 1842–1848
10. Khemchandani R, Chandra S et al (2009) Optimal kernel selection in twin support vector machines. Optim Lett 3(1):77–88
11. Arun Kumar M, Gopal M (2009) Least squares twin support vector machines for pattern classification. Expert Syst Appl 36(4):7535–7543
12. Ghorai S, Mukherjee A, Dutta PK (2009) Nonparallel plane proximal classifier. Signal Process 89(4):510–522
13. Peng X (2010) Primal twin support vector regression and its sparse approximation. Neurocomputing 73(16):2846–2858
14. Peng X (2012) Efficient twin parametric insensitive support vector regression model. Neurocomputing 79:26–38
15. Shao Y-H, Deng N-Y (2012) A coordinate descent margin based-twin support vector machine for classification. Neural Netw 25:114–121
16. Shao Y-H, Zhang C-H, Wang X-B, Deng N-Y (2011) Improvements on twin support vector machines. IEEE Trans Neural Netw 22(6):962–968
17. Shao YH, Chen WJ, Deng NY (2014) Nonparallel hyperplane support vector machine for binary classification problems. Inf Sci 263:22–35
18. Tian Y, Qi Z, Ju X, Shi Y, Liu X (2014) Nonparallel support vector machines for pattern classification. IEEE Trans Cybern. doi:10.1109/TCYB.2013.2279167
19. Vapnik V (1999) The nature of statistical learning theory. Springer, Berlin
20. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2(2):121–167
21. Wang Z, Chen J, Qin M (2010) Nonparallel planes support vector machine for multi-class classification. In: 2010 international conference on logistics systems and intelligent management (ICLSIM) vol 1, pp 581–585
22. Asuncion A, Newman DJ (2007) UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html