Knowledge-Based Systems 43 (2013) 74–81

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Structural twin support vector machine for classification

Zhiquan Qi *, Yingjie Tian, Yong Shi

Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 13 June 2012
Received in revised form 2 December 2012
Accepted 8 January 2013
Available online 31 January 2013

Keywords:
Classification
Twin support vector machine
Structural information of data
Machine learning
Optimization

ABSTRACT

It has been shown that the structural information of data may contain useful prior domain knowledge for training a classifier. How to apply the structural information of data to build a good classifier is a new research focus recently. As we all know, the all existing structural large margin methods are the common in considering all structural information within classes into one model. In fact, these methods do not balance all structural information's relationships both infra-class and inter-class, which directly results in these prior information not being exploited sufficiently. In this paper, we design a new Structural Twin Support Vector Machine (called $\mathcal{S}$-TWSVM). Unlike existing methods based on structural information, $\mathcal{S}$-TWSVM uses two hyperplanes to decide the category of new data, of which each model only considers one class's structural information and closer to the class at the same time far away from the other class. This makes $\mathcal{S}$-TWSVM fully exploit these prior knowledge to directly improve the algorithm's the capacity of generalization. All experiments show that our proposed method is rigidly superior to the state-of-the-art algorithms based on structural information of data in both computation time and classification accuracy.

1. Introduction

In the last decade, Support Vector Machines (SVMs) [1], as powerful tools for pattern classification and regression, have already successfully applied in a wide variety of fields [2–9]. For the standard support vector classification (SVC), the basic idea is to find the optimal separating hyperplane between the positive and negative examples. The optimal hyperplane may be obtained by maximizing the margin between two parallel hyperplanes, which involves the minimization of a quadratic programming problem (QPP). By introducing kernel trick into the dual QPP, SVC can also solve nonlinear classification problem successfully. As the extension of SVM, many new large margin classifiers based on structural information have been proposed. In fact, the traditional SVM does not sufficiently apply the prior example distribution information within classes and an optimal classifier should be sensitive to the structure of the data distribution. Exploiting clustering algorithms to extract the structural information embedded with classes is one popular strategy [10–12]. The structured large margin machine (SLMM) [10] is a representative work based on the strategy. Firstly, SLMM explores the structural information within classes by Ward's agglomerative hierarchical clustering method on input data [13], and then introduces the related structure information into the constraints. Finally, SLMM is able to be solved by a sequential second order cone programming (SOCP). Experimentally, SLMM is superior to support vector machine minimax probability machine (MPM) [14] and maxi-min margin machine (M4) [15]. However, as we all know, solving the involved SOCP problem is more difficult than the QPP problem as in SVM, so SLMM has more higher computational complexity than traditional SVM. Consequently, a novel structural support vector machine (SRSVM) was proposed by Xue et al. [12]. Unlike SLMM, SRSVM exploits the classical framework of SVM rather than as constraints in SLMM and the corresponding optimization problem is still able to be solved by the QPP. SRSVM has been shown to be theoretically and empirically better in generalization than SVM and SLMM.

In this paper, inspired by the success of TWSVM methods [16–22], we proposed a new Structural Twin Support Vector Machine for classification (called $\mathcal{S}$-TWSVM). Similar to structural SVM methods, $\mathcal{S}$-TWSVM exploits the structural information within classes by some clustering technology, and then introduces the data distributions information into the model of $\mathcal{S}$-TWSVM to construct more reasonable classifier. Besides features above, $\mathcal{S}$-TWSVM has still the following compelling properties.

◇ To our knowledge, $\mathcal{S}$-TWSVM is the first TWSVM implementation based on structural information of the data, which is a useful extension of TWSVM.
◇ We show that the TWSVM and TBSVM [17](TWSVM with the regular terms) are the special cases of our proposed models. This provides an alternative explanation to the success of $\mathcal{S}$-TWSVM.

* Corresponding author.
E-mail addresses: qizhiquan@gucas.ac.cn, qizhiquan@foxmail.com (Z. Qi), tyj@gucas.ac.cn (Y. Tian), yshi@gucas.ac.cn (Y. Shi).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.01.008

◇ Different with all existing structural classifiers such as [14,15,10–12], $\mathcal{S}$-TWSVM only considers one class's structural information for each model. This makes $\mathcal{S}$-TWSVM has the following advantages: (1) further reducing the computational complexity of the related QPP problem; (2) to be able to effectively deal with the condition of the structural information between positive class and negative class existing contradiction and more reasonable to apply the prior structural information within classes; and (3) further improving the flexibility of the algorithm and the model's generalization capacity.

The remaining parts of the paper are organized as follows. Section 2 briefly introduces the background of SLMM and SRSVM; Section 3 describes the detail of $\mathcal{S}$-TWSVM; All experiment results are shown in Section 4; Last section gives the conclusions.

2. Background

For classification about the training data

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\} \in (R^n \times \mathcal{Y})^l, \tag{1}$$

where $x_i \in R^n, y_i \in \mathcal{Y} = \{1, -1\}, i = 1, \ldots, l$.

Suppose there are respectively C_P and C_N clusters in class P and N, i.e., $P = P_1 \bigcup \cdots P_i \bigcup \cdots P_{C_P}, N = N_1 \bigcup \cdots N_j \bigcup \cdots N_{C_N}$.

2.1. SRSVM

By introducing the data distributions of the clusters in different classes into the traditional optimization function of SVM rather than in the constraints. The SRSVM model in the soft margin version can be formulated as [11,12]:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{\lambda}{2} w^\top \Sigma w + C \sum_{i=1}^{l} \xi_i \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, l, \end{aligned} \tag{2}$$

where $\Sigma = \Sigma_{P_1} + \cdots + \Sigma_{P_{C_P}} + \Sigma_{N_1} + \cdots + \Sigma_{N_{C_N}}$, $\Sigma_{P_{C_i}}$ and $\Sigma_{N_{C_j}}$ are the corresponding covariance matrices of clusters in the two classes. λ is an input parameter related with structural information within the clusters. This is a standard QPP problem. Compared with SLMM, SRSVM not only has much lower computational complexity but also holds the sparsity of the solution.

3. Structural Twin Support Vector Machine ($\mathcal{S}$-TWSVM)

3.1. Extracting structural information within classes

Following the strategy of the SLMM and SRSVM, $\mathcal{S}$-TWSVM also has two steps. The first step is to extract the structural information within classes by some clustering method; the second step is the model learning. In order to compare the main difference of the second step between $\mathcal{S}$-TWSVM and the other two methods, here we also adopt the same clustering method: Ward's linkage clustering (WIL) [13,10–12], which is one of the hierarchical clustering analysis. A main advantage of WIL is that clusters derived from this method are compact and spherical, which provides a meaningful basis for the computation of covariance matrices [10]. Concretely, if S and T are two clusters with means μ_S and μ_T, the Ward's linkage $W(S,T)$ between clusters S and T is computed as [10]

$$W(S, T) = \frac{|S| \cdot |T| \cdot \|\mu_S - \mu_T\|}{|S| + |T|}. \tag{3}$$

Initially, each sample is considered as a cluster. The Wards linkage of two samples x_i and x_j is $W(x_i, x_j) = \|x_i - x_j\|^2/2$. When two clusters are being merged to a new cluster A', the linkage $W(A', C)$ can be conveniently derived from $W(A, C)$, $W(B, C)$ and $W(A, B)$ by [10]

$$W(A', C) = \frac{(|A| + |C|)W(A, C) + (|B| + |C|)W(B, C) - |C|W(A, B)}{|A| + |B| + |C|}. \tag{4}$$

During the hierarchical clustering, the Ward's linkage between clusters to be merged increases as the number of clusters decreases [10]. A relation curve between the merge distance and the number of clusters are able to be drawn to represent this process. The optimal number of clusters is determined by finding the knee point [23]. Furthermore, the WIL can also be extended to the kernel space. More details of WIL are able to be found in [10].

3.2. Model learning

3.2.1. The linearly nonseparable case of $\mathcal{S}$-TWSVM

We obtain two groups of P and N clusters in class P and N by the first step, i.e., $P = P_1 \bigcup \cdots P_i \bigcup \cdots P_{C_p}, N = N_1 \bigcup \cdots N_j \bigcup \cdots N_{C_N}$. Suppose that data points belong to positive class are denoted by $A \in R^{m_1 \times n}$, where each row $A_i \in R^n$ represents a data point. Similarly, $B \in R^{m_2 \times n}$ represents all of the data points belong to negative class, where $m_1 + m_2 = l$. For the linear case, the $\mathcal{S}$-TWSVM determines two nonparallel hyperplanes:

$$f_+(x) = w_+^\top x + b_+ = 0 \quad \text{and} \; f_-(x) = w_-^\top x + b_- = 0, \tag{5}$$

where $w_+, w_- \in R^n$, $b_+, b_- \in R$. Here, each hyperplane is closer to one of the two classes and is at least one distance from the other, at the same time, minimizes the compactness within the class by the structural information obtained by clustering technology. A new data point is assigned to positive class or negative class depending upon its proximity to the two nonparallel hyperplanes. By introducing the data distributions of the clusters in different classes into the object functions of TBSVM, (Notice we only consider one class' structural information for each model. In other words, each model only considers these structural information of which the hyperplane is closer to the class.) the $\mathcal{S}$-TWSVM model can be formulated as

$$\begin{aligned} \min_{w_+, b_+, \xi} \quad & \frac{1}{2}\|Aw_+ + e_+ b_+\|_2^2 + c_1 e_-^\top \xi + \frac{1}{2} c_2 \left(\|w_+\|_2^2 + b_+^2\right) + \frac{1}{2} c_3 w_+^\top \Sigma_+ w_+, \\ \text{s.t.} \quad & -(Bw_+ + e_- b_+) + \xi \geqslant e_-, \xi \geqslant 0, \end{aligned} \tag{6}$$

and

$$\begin{aligned} \min_{w_-, b_-, \eta} \quad & \frac{1}{2}\|Bw_- + e_- b_-\|_2^2 + \frac{1}{2} c_4 e_+^\top \eta + \frac{1}{2} c_5 \left(\|w_-\|_2^2 + b_-^2\right) + \frac{1}{2} c_6 w_-^\top \Sigma_- w_- \\ \text{s.t.} \quad & (Aw_- + e_+ b_-) + \eta \geqslant e_+, \eta \geqslant 0, \end{aligned} \tag{7}$$

where $c_1, \ldots, c_6 \geqslant 0$ are the pre-specified penalty factors, e_+, e_- are vectors of ones of appropriate dimensions, ξ_i is the slack variables, $\Sigma_+ = \Sigma_{P_1} + \cdots + \Sigma_{P_{C_P}}$, $\Sigma_- = \Sigma_{N_1} + \cdots + \Sigma_{N_{C_N}}$, Σ_{P_i} and Σ_{N_j} are respectively the covariance matrices corresponding to the ith and jth clusters in the two classes, $i = 1, \ldots, C_p$, $j = 1, \ldots, C_N$.

The Wolfe dual of the problem (6) is as follow:

$$\begin{aligned} \max_{\alpha} \quad & e_-^\top \alpha - \frac{1}{2} \alpha^\top G (H^\top H + c_2 I + c_3 J)^{-1} G^\top \alpha \\ \text{s.t.} \quad & 0 \leqslant \alpha \leqslant c_1 e_-, \end{aligned} \tag{8}$$

where

$$H = [A \; e_+], \; J = \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix}, \quad G = [B \; e_-], \tag{9}$$

and the augmented vector $\vartheta_+ = \left[w_+^\top \; b_+^\top\right]^\top$ is given by

$$\vartheta_+ = -(H^\top H + c_2 I + c_3 J)^{-1}(G^\top \alpha). \tag{10}$$

I is an identity matrix of appropriate dimensions. According to matrix theory [24], it is very easy to prove that $H^\top H + c_2 I + c_3 J$ is a positive definite matrix.

Similarly, the dual of (7) is

$$\begin{aligned} \max_{\beta} \quad & e_+^\top \beta - \frac{1}{2}\beta^\top P(Q^\top Q + c_5 I + c_6 F)^{-1} P^\top \beta \\ \text{s.t.} \quad & 0 \leqslant \beta \leqslant c_4 e_+, \end{aligned} \tag{11}$$

where

$$P = [A\ e_-],\ \ F = \begin{bmatrix} \Sigma_- & 0 \\ 0 & 0 \end{bmatrix} \quad Q = [B\ e_+], \tag{12}$$

and the augmented vector $\vartheta_- = [w_- b_-]^\top$ given by

$$\vartheta_- = -(Q^\top Q + c_5 I + c_6 F)^{-1} P^\top \beta, \tag{13}$$

where $Q^\top Q + c_5 I + c_6 F$ is a positive definite matric. Once vectors ϑ_+ and ϑ_- are obtained from (10) and (13), the separating planes

$$w_+^\top x + b_+ = 0, \quad w_-^\top x + b_- = 0 \tag{14}$$

are known. A new data point $x \in R^n$ is then assigned to the positive or negative class, depending on which of the two hyperplanes given by (14) it lies closest to, i.e.

$$f(x) = \underset{+,-}{\text{argmin}}\{d_+(x), d_-(x)\}, \tag{15}$$

where

$$d_+(x) = |w_+^\top x + b_+|, \quad d_-(x) = |w_-^\top x + b_-|, \tag{16}$$

where $|\cdot|$ is the perpendicular distance of point x from the planes $w_+^\top x + b_+$ or $w_-^\top x + b_-$.

3.2.2. Nonlinear $\mathcal{S}$-TWSVM

Now we extend the linear $\mathcal{S}$-TWSVM to the nonlinear case.

Similar to linear case, the decision function is written as $f_+(x) = (w_+ \cdot \Phi(x)) + b_+$ and $f_-(x) = (w_- \cdot \Phi(x)) + b_-$, where $\Phi(\cdot)$ is a nonlinear mapping from a low dimensional space to a higher dimensional Hilbert space $\mathcal{H}$. According to Hilbert space theory [25], w_+ and w_- can be expressed as $w_+ = \sum_{i=1}^{m_1+m_2}(\lambda_+)_i \Phi(x_i) = \Phi(M)\lambda_+$ and $w_- = \sum_{i=1}^{m_1+m_2}(\lambda_-)_i \Phi(x_i) = \Phi(M)\lambda_-$, respectively. So the following kernel-generated hyperplane:

$$\begin{aligned} & K(x^\top, M^\top)\lambda_+ + b_+ = 0, \\ & K(x^\top, M^\top)\lambda_- + b_- = 0, \end{aligned} \tag{17}$$

where K is an chosen kernel function: $K(x_i \cdot x_j) = (\Phi(x_i) \cdot \Phi(x_j))$, $M = [A^\top B^\top]$. The nonlinear optimization problem can be expressed as

$$\begin{aligned} \min_{\lambda_+, b_+, \xi} \quad & \frac{1}{2}\|K(A, M^\top)\lambda_+ + e_+ b_+\|^2 + c_1 e_-^\top \xi + \frac{1}{2}c_2(\|\lambda_+\|^2 + b_+^2) \\ & + \frac{1}{2}c_3 \lambda_+^\top \Phi(M)^\top \Sigma_+^\Phi \Phi(M)\lambda_+, \\ \text{s.t.} \quad & -(K(B, M^\top)\lambda_+ + e_- b_+) + \xi \geqslant e_-, \xi \geqslant 0, \end{aligned} \tag{18}$$

and

$$\begin{aligned} \min_{\lambda_-, b_-, \eta} \quad & \frac{1}{2}\|K(B, M^\top)\lambda_- + e_- b_-\|^2 + c_4 e_+^\top \eta + \frac{1}{2}c_5\left(\|\lambda_-\|^2 + b_-^2\right) \\ & + \frac{1}{2}c_6 \lambda_-^\top \Phi(M)^\top \Sigma_-^\Phi \Phi(M)\lambda_-, \\ \text{s.t.} \quad & (K(A, M^\top)\lambda_- + e_+ b_-) + \eta \geqslant e_+, \eta \geqslant 0, \end{aligned} \tag{19}$$

where $\Sigma_+^\Phi = \Sigma_{P_1}^\Phi + \ldots + \Sigma_{P_{C_P}}^\Phi$, $\Sigma_-^\Phi = \Sigma_{N_1}^\Phi + \cdots + \Sigma_{N_{C_N}}^\Phi$, Σ_{P_i} and Σ_{N_j} are respectively the covariance matrices corresponding to the i th and j th clusters in the two classes by the kernel Ward's linkage clustering [10,12], $i = 1, \ldots, C_p$, $j = 1, \ldots, C_N$.

The Wolfe dual of the problem (18) is formulated as follow:

$$\begin{aligned} \max_{\alpha} \quad & e_-^\top \alpha - \frac{1}{2}(\alpha^\top G_\Phi)\left(H_\Phi^\top H_\Phi + c_2 I + c_3 J_\Phi\right)^{-1}(G_\Phi^\top \alpha) \\ \text{s.t.} \quad & 0 \leqslant \alpha \leqslant c_1 e_-, \end{aligned} \tag{20}$$

where

$$\begin{aligned} & H_\Phi = [K(A, M^\top)\ e_+], G_\Phi = [K(B, M^\top)\ e_-] \\ & J_\Phi = \begin{bmatrix} \Phi(M)^\top \Sigma_+^\Phi \Phi(M) & 0 \\ 0 & 0 \end{bmatrix} \end{aligned} \tag{21}$$

and the augmented vector $\rho_+ = [\lambda_+ b_+]^\top$

$$\rho_+ = -\left(H_\Phi^\top H_\Phi + c_2 I + c_3 J_\Phi\right)^{-1}(G_\Phi^\top \alpha). \tag{22}$$

In a similar manner, the dual of (19) is

$$\begin{aligned} \max_{\beta} \quad & e_+^\top \beta - \frac{1}{2}(\beta^\top P_\Phi)\left(Q_\Phi^\top Q_\Phi + c_5 I + c_6 F_\Phi\right)^{-1}(P_\Phi^\top \beta) \\ \text{s.t.} \quad & 0 \leqslant \beta \leqslant c_4 e_+, \end{aligned} \tag{23}$$

where

$$\begin{aligned} & P_\Phi = [K(A, M^\top)\ e_-], Q_\Phi = [K(B, M^\top)\ e_+], \\ & F_\Phi = \begin{bmatrix} \Phi(M)^\top \Sigma_-^\Phi \Phi(M) & 0 \\ 0 & 0 \end{bmatrix}, \end{aligned} \tag{24}$$

and the augmented vector $\rho_- = [\lambda_- b_-]^\top$, which is given by

$$\rho_- = -\left(Q_\Phi^\top Q_\Phi + c_5 I + c_6 F_\Phi^{-1}\right)(P_\Phi^\top \alpha). \tag{25}$$

Once vectors ρ_+ and ρ_- are obtained from (22) and (25), a new data point $x \in R^n$ is then assigned to the positive or negative class, depending on a manner similar to the linear case.

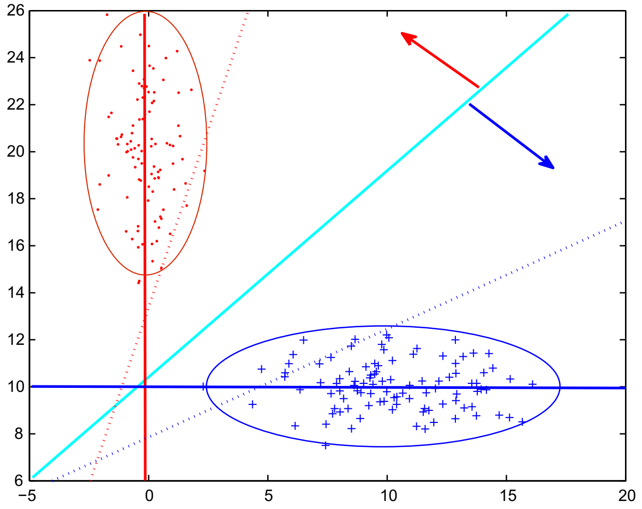


Fig. 1. the geometric interpretation of existing the structural information confliction between positive class and negative class. The red and blue solid line denotes the classifier of $\mathcal{S}$-TWSVM. The red and blue dotted line denotes the classifier of Structural TWSVM of each model consider two class's structural information (for simplify, we called it $\mathcal{SS}$-TWSVM). Obviously, $\mathcal{S}$-TWSVM is able to better predict the data's distribution tendency than $\mathcal{SS}$-TWSVM. The cyan line denotes the classifier based on one hyperplane such as SLMM or SRSVM and the red and blue arrows denotes the tendency of the two class's structural information hoping the classifier to rotate. In the case, the classifier is almost the same as the that's traditional SVM and these structural information does not play a role and make the classifier change. $\mathcal{S}$-TWSVM is obviously superior to SLMM, SRSVM and $\mathcal{SS}$-TWSVM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Now we consider how to compute the kernel matrix J_Φ. Suppose T_{P_i} is a matrix corresponding to the cluster $P_i, T_{P_i} \in R^{P_i \times n}$, in which the kth row is $x_k^\top$. O_{P_i} is a mean matrix of cluster $P_i, O_{P_i} \in R^{P_i \times n}$. Each row of O_{P_i} is the same, i.e.

$$\mu_{P_i} = \frac{1}{P_i}\sum_{x_k \in P_i} x_k. \tag{26}$$

The related covariance matrix for cluster P_i can be expressed as

$$\Sigma^{\Phi}_{P_i} = \frac{1}{P_i}(\Phi(T_{P_i}) - \Phi(O_{P_i}))^\top(\Phi(T_{P_i}) - \Phi(O_{P_i})). \tag{27}$$

So we obtain

$$\begin{aligned}\Phi(M)^\top \Sigma^{\Phi}_{+}\Phi(M) &= \left(\frac{1}{\sqrt{P_i}}(\Phi(T_{P_i}) - \Phi(O_{P_i}))\Phi(M)\right)^\top \\ &\quad \left(\frac{1}{\sqrt{P_i}}(\Phi(T_{P_i}) - \Phi(O_{P_i}))\Phi(M)\right) \\ &= \left(\frac{1}{\sqrt{P_i}}(K(T_{P_i}, M) - K(O_{P_i}, M))\right)^\top \cdot \\ &\quad \left(\frac{1}{\sqrt{P_i}}(K(T_{P_i}, M) - K(O_{P_i}, M))\right).\end{aligned} \tag{28}$$

Similarly, $\Phi(M)^\top\Sigma_M + {}^{\Phi}\Phi(M)$ of F_Φ are computed as

$$\begin{aligned}\Phi(M)^\top \Sigma^{\Phi}_{-}\Phi(M) &= \left(\frac{1}{\sqrt{P_i}}(K(T_{N_i}, M) - K(O_{N_i}, M))\right)^\top \cdot \\ &\quad \left(\frac{1}{\sqrt{P_i}}(K(T_{N_i}, M) - K(O_{N_i}, M))\right).,\end{aligned} \tag{29}$$

where T_{N_i} is a matrix of cluster N_i, O_{N_i} is a mean matrix of cluster N_i.

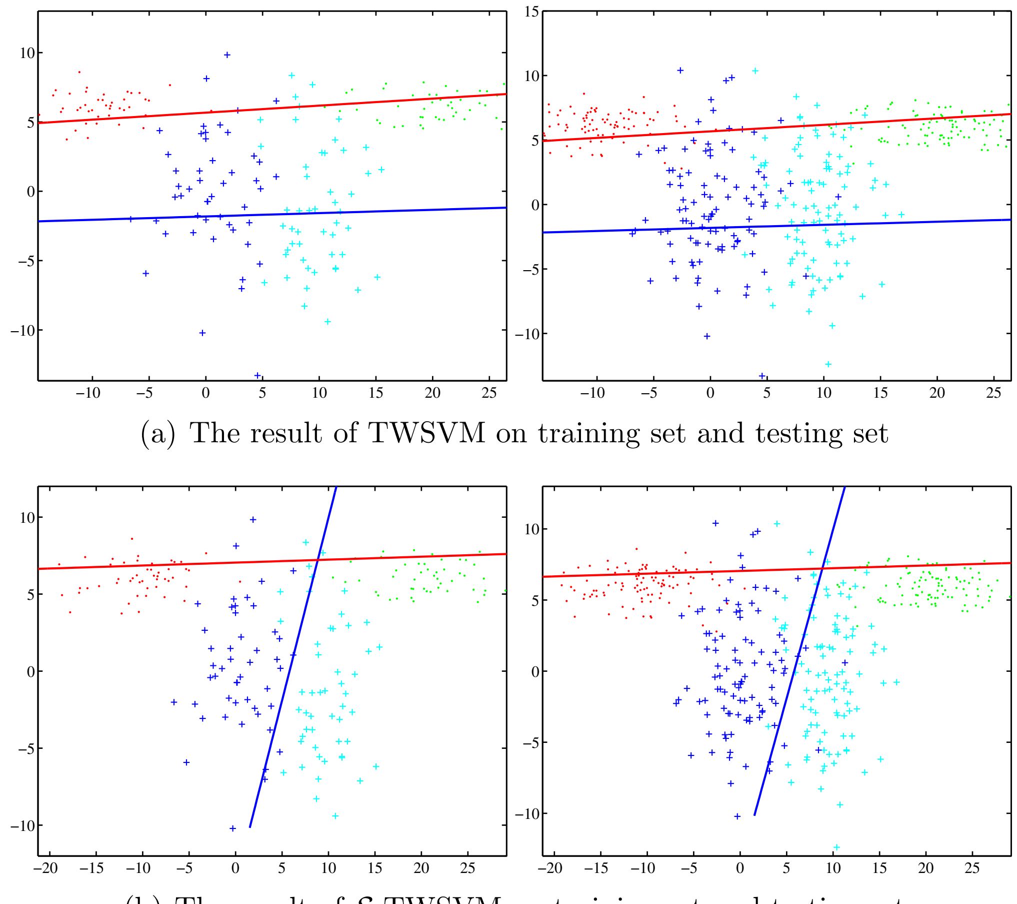
3.3. Discussion

3.3.1. Relationship with TWSVM and TBSVM

$\mathcal{S}$-TWSVM has a similar structure of TWSVM and TBSVM, we can easily proof that TWSVM and TBSVM [17] are the special cases of $\mathcal{S}$-TWSVM. Suppose the variance–covariance matrix of each cluster is $\Sigma_{P_i} = \sigma_{N_j} = I, \quad i = 1, \ldots, C_P, \quad j = 1, \ldots, C_N$. For an example of linear $\mathcal{S}$-TWSVM, the primal optimization problem (6) of $\mathcal{S}$-TWSVM becomes

$$\begin{aligned}\min_{w_+, b_+, \xi} \quad & \frac{1}{2}\|Aw_+ + e_+ b_+\|_2^2 + c_1 e_-^\top \xi + \frac{1}{2}\left((c_2 + c_3 C_P)\|w_+\|_2^2 + c_2 b_+^2\right), \\ \text{s.t.} \quad & -(Bw_+ + eb_+) + \xi \geqslant e_-, \xi \geqslant 0.\end{aligned} \tag{30}$$

It is not difficult to see that the optimization problem (30) is equivalent to one of the primal problem of TBSVM. Specially when $c_3 = 0$, $\mathcal{S}$-TWSVM degenerates to TBSVM. If $c_2 = c_3 = 0$, $\mathcal{S}$-TWSVM becomes TWSVM. $\mathcal{S}$-TWSVM inherits the all virtues of TWSVM and TBSVM and has the natural advantages in the model's training time and generalized capability than traditional SVM methods.



(a) The result of TWSVM on training set and testing set

(b) The result of $\mathcal{S}$-TWSVM on training set and testing set

Fig. 2. The performance comparison of TWSVM and $\mathcal{S}$-TWSVM on XOR Data. The "·" and "+" denote the positive class and negative class, respectively. Different colors represent different clusters. Red and blue lines are two hyperplanes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

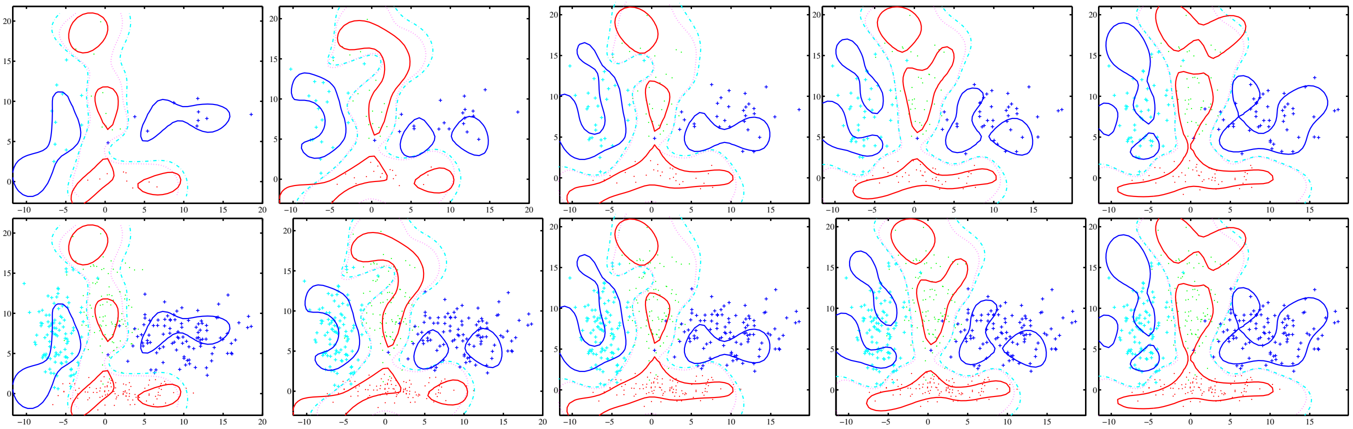


Fig. 3. The performance of $\mathcal{S}$-TWSVM, SLMM and SRSVM in the case of RBF case. The first row and second row are the results on the training set and testing set. Each column are the results on 10%, 20%, 30%, 40% and 50% training sets, respectively. The magenta dotted curve and cyan dash-dot curve denote the hyperplanes of SLMM and SRSVM, respectively; blue and red solid curves are the hyperplanes of $\mathcal{S}$-TWSVM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3.2. Relationship with SLMM and SRSVM

The same as SLMM and SRSVM, $\mathcal{S}$-TWSVM also captures the data structural information within classes by some clustering strategies. Similar to SRSVM, $\mathcal{S}$-TWSVM directly embeds the data distribution information into the TBSVM objective function rather than using as the constraints into SLMM. In addition, the corresponding optimization problem of SLMM is solved by SOCP, but SRSVM and $\mathcal{S}$-TWSVM solved by QPPs.

3.3.3. The advantage of $\mathcal{S}$-TWSVM over other algorithms

From the analysis on the structure of $\mathcal{S}$-TWSVM, it is implemented by solving two smaller QPPs rather than a single large QPP, which makes the learning speed of $\mathcal{S}$-TWSVM is faster than existing classification methods based on structural information of the data. In fact, considering all structural information into one model or one hyperplane (such as MPM [14], M4 [15], SLMM [10], and SRSVM [11,12]) usually losts much useful prior knowledge (see Fig. 1). However, for our proposed method, firstly using two nonparallel hyperplanes to decide the category of a new example is an implementation of exploiting structural information of data. Furthermore, $\mathcal{S}$-TWSVM imbeds the prior structural information for each model. More importantly, $\mathcal{S}$-TWSVM's each model only considers these structural information of which the hyperplane is closer to the class. This makes $\mathcal{S}$-TWSVM try to avoid the confliction of structural information each other and can more fully exploit these prior knowledge to improve the algorithm's generalization performance.

4. Experiments

We compare the $\mathcal{S}$-TWSVM against TBSVM [17], SLMM [10], and SRSVM [11,12] on various data sets in this section.

In order to simplify, let $c_1 = c_4$, $c_2 = c_5$, $c_3 = c_6$ in $\mathcal{S}$-TWSVM. The testing accuracies of all experiments are computed using standard 10-fold cross validation. c_1, c_2, c_3 and RBF kernel parameter σ are all selected from the set $\{2^i | i = -7, \ldots, 7\}$ by 10-fold cross validation on the tuning set comprising of random 10% of the training data. Once the parameters are selected, the tuning set was returned to the training set to learn the final decision function. Sequential Minimal Optimization (SMO) algorithm [26] is used to solve the QP problems in SRSVM and SVM, SeDuMi program[1] to solve the SOCP problem in SLMM, and Successive Over Relaxation (SOR) [17,27] technique to solve the QP problems in TWSVM and $\mathcal{S}$-TWSVM. The "1 vs r" method [26] is used to solve the multi-class classification. All algorithms are implemented by using MATLAB 2010. The experimental environment: Intel Core i7-2600 CPU, 4 GB memory.

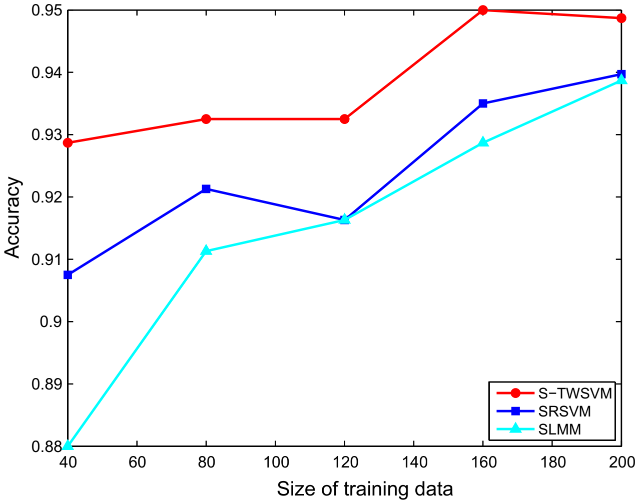


Fig. 4. Accuracy of $\mathcal{S}$-TWSVM, SLMM and SRSVM on the second experiment.

4.1. Toy data

In the subsection, we use a 2-D toy data to show the intuitive performance of $\mathcal{S}$-TWSVM. The 2-D toy data are the synthetic XOR dataset [12], which is a typical linear nonseparable problem in classification and randomly generated under two Gaussian distributions in each class. In practice, samples in each class are designed to two clusters P_1, P_2 and N_1, N_2 (the number of samples in each cluster is equal), and each Gaussian distribution contains 100 samples. We carry out two different experiments on the dataset. For the first experiment, we use the half of the data in each cluster for training and others for testing. The comparative results of $\mathcal{S}$-TWSVM and TWSVM are given in Fig. 2. For the second experiment, we respectively use 10%, 20%, 30%, 40%, 50% of data in each cluster as the training set, and others for testing. The comparative results of $\mathcal{S}$-TWSVM and SRSVM are shown in Fig. 3.

[1] http://sedumi.i.e.lehigh.edu/.

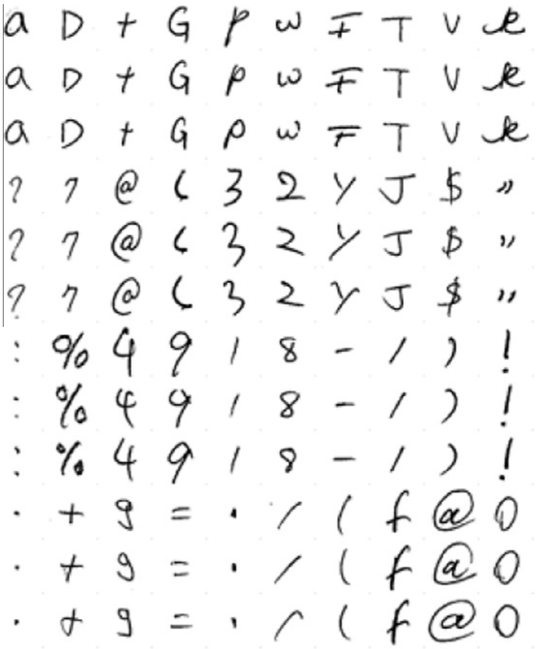
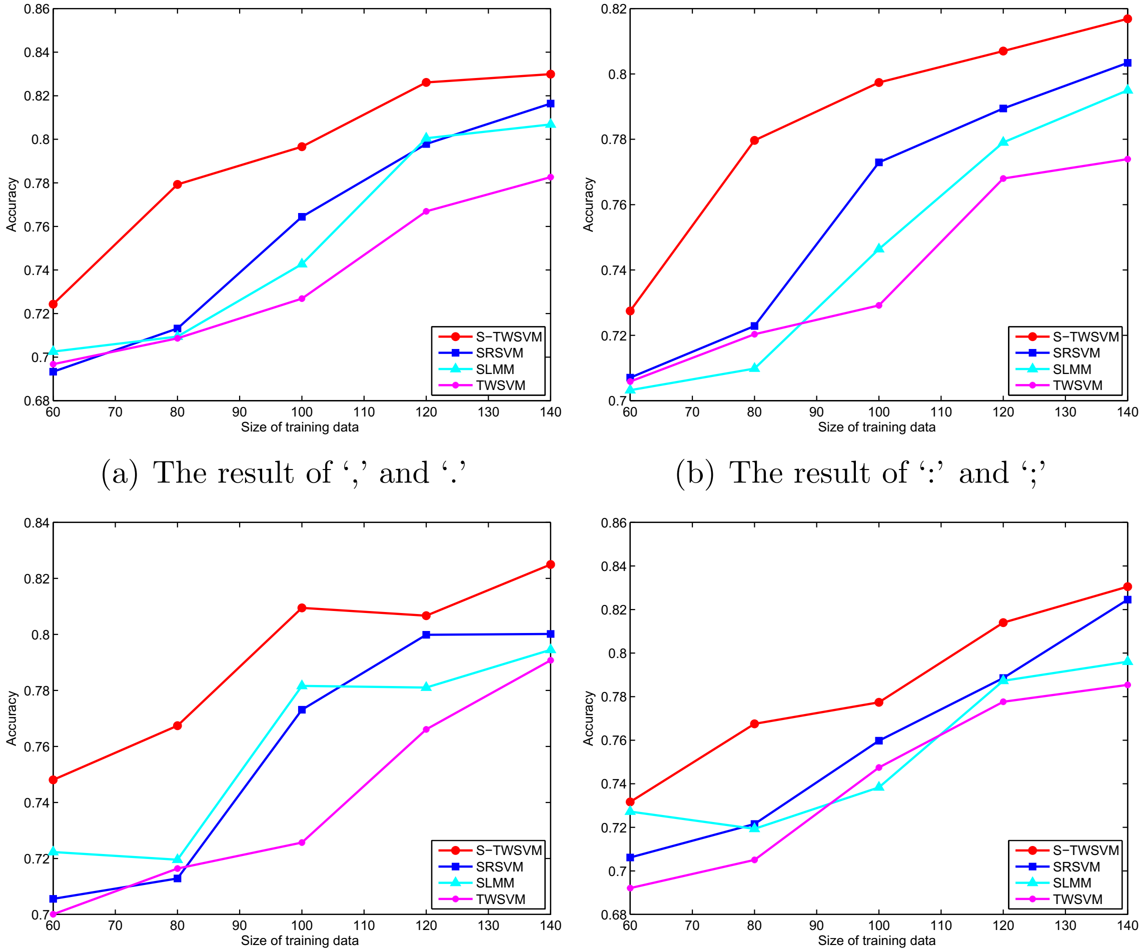


Fig. 5. Samples from ABCDETC dataset.

From Fig. 2, we can see that the positive class has two horizontal distributions and the negative class has two vertical distributions. In this condition, the structural information within classes is very important for classification. Due to TWSVM neglecting this information and only focusing on the separability between the classes, the final classifier cannot cater to the future trend of data distribution. However, due to the object function of $\mathcal{S}$-TWSVM adding the structural information within the classes obtained by some clustering algorithm, it can make a tradeoff between the structural information within classes and the discriminative information. So $\mathcal{S}$-TWSVM obtains a better result than TWSVM.

In the second experiment, the positive class and negative class have both the horizontal distribution and the vertical distribution. How to fully exploit these prior knowledge will be a very difficult task. From Figs. 3 and 4, we can find that SRSVM's discriminant boundaries basically enclose those of SLMM, which means that SRSVM has better generalization performance than SLMM. The conclusion of [12] is confirmed again. Meanwhile, we also find that the result of $\mathcal{S}$-TWSVM is much better than that of the LMM and SRSVM (The average accuracy of SRSVM is 4.48% higher than that of SLMM, but $\mathcal{S}$-TWSVM 7.26% higher than that of SRSVM). This show that $\mathcal{S}$-TWSVM can fully exploit to these prior structural information to design a more reasonable classifier.



(a) The result of ',' and '.'

(b) The result of ':' and ';'

(c) The result of '!' and '?'

(d) The result of 'a' and 'b'

Fig. 6. The ROC curves of TWSVM, SLMM, SRSVM and $\mathcal{S}$-TWSVM on ABCDETC dataset.

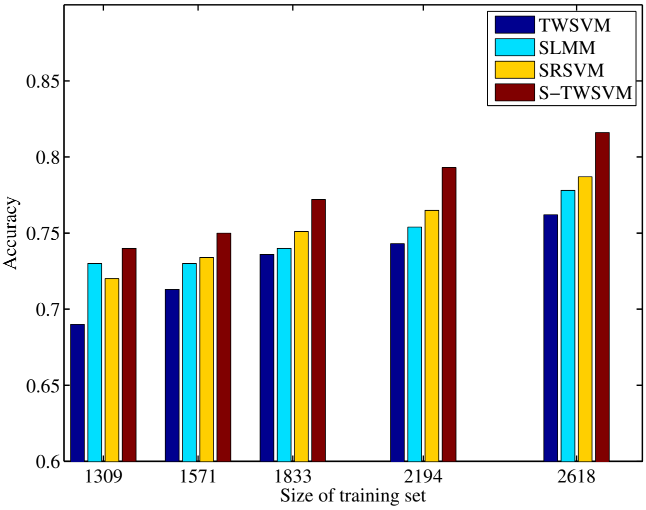


Fig. 7. The average accuracy on PASCAL 2006 dataset.

4.2. Image datasets

In this subsection, we use image datasets to evaluate $\mathcal{S}$-TWSVM and other algorithms. Because our goal is only to compare the performance between $\mathcal{S}$-TWSVM) and other algorithms, all experiment are carried out on raw pixel features.

(1) ABCDETC dataset [28] (see Fig. 5): there are 78 classes in the dataset of 19646 images. These categories are respectively: lowercase letters ('a–z'), uppercase letters ('A–Z'), digits ('0–9') and other symbols (' , . : ; ! ? + - = / $ % ( ) @ " '). Those symbols are finished in pen by 51 subjects, that wrote 5 versions for each symbol on a single gridded sheet, and then saved as 100 × 100 binary images. To improve the computing speed, we shrunk into the original images into 30 × 30 pixels by the bilinear interpolation method. The ',' and '.', '!' and '?', ':' and ';', 'a' and 'b' are taken as the labeled data respectively. The sizes of training data are set to 60, 80, 100, 120 and 140, respectively; 140 other samples are as testing set. All results are shown in the Fig. 6.

(2) PASCAL 2006 dataset [29]: it contains 10 object categories (cats, bicycles, cows, motorbikes, cars, dogs, buses, sheep, people, horses) and 5304 images. We use 1309, 1571, 1833, 2194, 2618 images as training data and others for testing. All images are resized to be gray images of 80 × 100 with 256 gray levels and are normalized to [0, 1]. Each example is expressed as a 8000 × 1 vector on raw pixel features. All results can be found in Fig. 7.

From Fig. 7, we firstly find that the advantage of the algorithms based on the structural information of the data distribution is more evident on ABCDETC dataset, which possibly implies that the samples of characters obey some more regular probability distribution. Next, we find that TWSVM is superior to SLMM or SRSVM in some cases (see Fig. 6). This shows TWSVM which uses some prior structural information itself is a competitive method. In addition, with the increase of training samples, the accuracies of the four algorithm also grows and simultaneously the difference among these accuracies of the four algorithm becomes much smaller.

4.3. UCI datasets

In this subsection, we perform these methods on the UCI datasets [30]. For each dataset, we randomly select the same number of data from different classes to compose a dataset. Fifty percent of each extracted dataset are for training, 50% for testing. The parameters' selection of models uses 10-fold cross validation method mentioned above. All data are normalized to [0, 1]. The final results are shown in the Table 1. From the Table 1, we can draw the conclusion as follows: (1) $\mathcal{S}$-TWSVM, SRSVM and SLMM have the better predictive ability than TWSVM in most cases. This shows that these priori structural information embedded in classes has a great help to improve the classification performance of the classifier. (2) SRSVM is superior to SLMM in most cases. However, $\mathcal{S}$-TWSVM is superior to the SLMM and SRSVM on the all UCI datasets. This further confirms that applying two models respectively to consider these structural information of each class is a reasonable approach, which is able to try to avoid such case of the structural information conflicting between positive class and negative class. This also shows that using two nonparallel hyperplane can more fully exploit to these prior knowledge to improve the algorithm's generalization performance. (3) SRSVM's training time is smaller than SLMM. This is because that SLMM needs to solve an SOCP problem but SRSVM only to solve a standard QPP. More importantly, the training speed of $\mathcal{S}$-TWSVM is much faster than SRSVM and SLMM. This is because $\mathcal{S}$-TWSVM is implemented by solving two smaller QPPs rather than a single large QPP.[2]

Table 1
The testing accuracy and training times on UCI datasets.

Datasets	$\mathcal{S}$-TWSVM Accuracy Time (s)	SRSVM Accuracy Time (s)	SLMM Accuracy Time (s)	TWSVM Accuracy Time (s)
Hepatitis	80.63 ± 1.32	79.83 ± 1.27	80.12 ± 2.25	78.47 ± 3.12
(155 × 19)	0.34	1.08	16.56	–
Australian	70.54 ± 0.89	69.32 ± 2.31	68.92 ± 3.54	68.43 ± 1.19
(690 × 14)	1.33	3.98	130.68	–
BUPA liver	69.75 ± 1.81	68.96 ± 1.09	69.42 ± 1.28	68.44 ± 1.17
(345 × 6)	0.81	1.93	78.81	–
CMC	66.66 ± 2.34	65.27 ± 2.35	64.76 ± 1.76	64.12 ± 3.31
(844 × 9)	1.51	3.42	190.24	–
Credit	77.26 ± 2.80	76.11 ± 2.61	76.25 ± 3.40	75.18 ± 2.77
(690 × 19)	1.21	3.98	143.43	–
Diabetis	64.28 ± 0.71	64.19 ± 2.43	63.74 ± 3.02	62.93 ± 4.27
(768 × 8)	0.98	4.12	111.66	–
Flare-Solar	60.17 ± 0.91	58.43 ± 2.77	58.24 ± 2.55	59.23 ± 1.56
(1066 × 9)	1.84	6.32	132.33	–
German	64.15 ± 1.76	63.84 ± 1.88	63.44 ± 2.12	63.92 ± 3.47
(1000 × 20)	2.41	7.01	210.37	–
Heart-Statlog	76.22 ± 1.92	76.04 ± 2.47	75.79 ± 3.18	75.86 ± 1.38
(270 × 14)	0.44	1.71	63.21	–
Image	84.28 ± 2.75	83.44 ± 1.40	84.72 ± 1.75	83.76 ± 3.42
(2310 × 18)	5.45	17.27	501.24	–
Ionosphere	76.82 ± 1.11	76.55 ± 2.63	75.44 ± 1.86	76.21 ± 2.33
(351 × 34)	0.85	2.75	77.78	–
Spect	74.40 ± 0.77	74.12 ± 1.29	73.52 ± 1.51	73.11 ± 2.41
(267 × 44)	0.71	2.13	65.24	–

5. Conclusion

In this paper, we proposed a new Structural Twin Support Vector Machine (called $\mathcal{S}$-TWSVM), which is sensitive to the structure of the data distribution. In the view of structural information, we firstly point out the shortcomings of the existing algorithms based on structural information. Next, we design a new $\mathcal{S}$-TWSVM algorithm and analysis its advantages and relationships with other algorithms. Theoretical analysis and all experimental results show $\mathcal{S}$-TWSVM can more fully exploit these prior structural information to improve the classification accuracy. At the same time the training time of $\mathcal{S}$-TWSVM is obviously superior to the state-of-the-art

[2] As our goal is only to compare our algorithm with others based on the structural information, we do not give the training time of TWSVM.

structural algorithms. In the future work, how to further accelerate the algorithm is under our consideration. In addition, the extension of semi-supervised learning and multi-instance classification are also interesting.

Acknowledgment

This work has been partially supported by Grants from National Natural Science Foundation of China (NO.70921061, NO.11271361), the CAS/SAFEA International Partnership Program for Creative Research Teams, Major International (Ragional) Joint Research Project (NO.71110107026), the President Fund of GUCAS.

References

[1] C. Cortes, V.N. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.
[2] W.S. Noble, Support Vector Machine Applications in Computational Biology, MIT Press, 2004.
[3] B. Schölkopf, I. Guyon, J. Weston, Statistical Learning and Kernel Methods in Bioinformatics, Tech. Rep., 2000.
[4] M.M. Adankon, M. Cheriet, Model selection for the LS-SVM. Application to handwriting recognition, Pattern Recognition 42 (12) (2009) 3264–3270.
[5] N. Khan, R. Ksantini, I. Ahmad, B. Boufama, A novel SVM+NDA model for classification with an application to face recognition, Pattern Recognition 45 (1) (2012) 66–79.
[6] Y.-C. Wu, Y.-S. Lee, J.-C. Yang, Robust and efficient multiclass SVM models for phrase pattern recognition, Pattern Recognition 41 (9) (2008) 2874–2889.
[7] R. Liu, Y. Wang, T. Baba, D. Masumoto, S. Nagata, SVM-based active feedback in image retrieval using clustering and unlabeled data, Pattern Recognition 41 (8) (2008) 2645–2655.
[8] Y. Tian, Y. Shi, X. Liu, Recent advances on support vector machines research, Technological and Economic Development of Economy 18 (1) (2012) 5–33.
[9] J. Tan, Z. Zhang, L. Zhen, C. Zhang, N. Deng, Adaptive feature selection via a new version of support vector machine, Neural Computing and Applications. doi:10.1007/s00521-012-1018-y.
[10] D. Yeung, D. Wang, W. Ng, E. Tsang, X. Wang, Structured large margin machines: sensitive to data distributions, Machine Learning 68 (2) (2007) 171–200.
[11] H. Xue, S. Chen, Q. Yang, Structural support vector machine, in: The 15th International Symposium on Neural Networks, 2008, pp. 501–511.
[12] H. Xue, S. Chen, Q. Yang, Structural regularized support vector machine: a framework for structural large margin classifier, IEEE Transactions on Neural Networks 22 (4) (2011) 573–587, http://dx.doi.org/10.1109/TNN.2011.2108315.
[13] J. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (301) (1963) 236–244, http://dx.doi.org/10.2307/2282967.
[14] G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, M.I. Jordan, A robust minimax approach to classification, Journal of Machine Learning Research 3 (2002) 555–582.
[15] K.H. Kzhuang, H. Yang, I. King, Learning large margin classifiers locally and globally, in: The Twenty-First International Conference on Machine Learning (ICML-2004), 2004, pp. 401–408.
[16] Jayadeva, R. Khemchandani, S. Chandra, Twin support vector machines for pattern classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 905–910.
[17] Y.-H. Shao, C.-H. Zhang, X.-B. Wang, N.-Y. Deng, Improvements on twin support vector machines, IEEE Transactions on Neural Networks 22 (6) (2011) 962–968.
[18] Y.-H. Shao, Nai, A coordinate descent margin based-twin support vector machine for classification, Neural Networks 25 (2012) 114–121.
[19] Z. Qi, Y. Tian, S. Yong, Robust twin support vector machine for pattern classification, Pattern Recognition 46 (1) (2013) 305–316, http://dx.doi.org/10.1016/j.patcog.2012.06.019.
[20] Z. Qi, Y. Tian, S. Yong, Twin support vector machine with universum data, Neural Networks 36C (2012) 112–119, http://dx.doi.org/10.1016/j.neunet.2012.09.004.
[21] Z. Qi, Y. Tian, S. Yong, Laplacian twin support vector machine for semi-supervised classification, Neural Networks 35 (2012) 46–53. http://dx.doi.org/10.1016/ j.neunet. 2012.07.011.
[22] Z. Yang, Y. Shao, X. Zhang, Multiple birth support vector machine for multi-class classification, Neural Computing and Application. doi:10.1007/s00521-012-1108-x.
[23] S. Salvador, P. Chan, Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms, Tech. Rep., 2003.
[24] F.R. Gantmacher, Matrix Theory, New York, Chelsea, 1990.
[25] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, Adaptive Computation and Machine Learning, MIT Press, 2002.
[26] N. Deng, Y. Tian, C. Zhang, Support Vector Machines Optimization based Theory, Algorithms and Extensions, Taylor and Francis, 2012.
[27] O.L. Mangasarian, D.R. Musicant, Successive overrelaxation for support vector machines, IEEE Transactions on Neural Networks 10 (5) (1999) 1032–1037, http://dx.doi.org/10.1109/72.788643.
[28] J. Weston, R. Collobert, F. Sinz, L. Bottou, V. Vapnik, Inference with the Universum, in: ICML '06: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 1009–1016.
[29] M. Everingham, A. Zisserman, C.K.I. Williams, L. Van Gool, The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. <http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf>.
[30] A. Asuncion, D. Newman, UCI Machine Learning Repository (2007).