ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 4, No.22 , July 2016
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

MS-TWSVM: Mahalanobis distance-based Structural Twin Support Vector Machine

Ramin Rezvani−KhorashadiZadeh[1], Reza Monsefi[2]

[1] Computer Department, Engineering Faculty, Ferdowsi University of Mashhad (FUM), Mashhad, Iran, raminrezvani@stu-mail.um.ac.ir

[2] Computer Department, Engineering Faculty, Ferdowsi University of Mashhad (FUM), Mashhad, Iran, Center of Excellence on Soft Computing and Intelligent Information Processing (SCIIP), monsefi@um.ac.ir

Abstract

The distribution information of data points in two classes as the structural information is inserted into the classifiers to improve their generalization performance. Recently many algorithms such as S-TWSVM has used this information to construct two non-parallel hyperplanes which each one lies as close as possible to one class and being far away from the other. It is well known that different classes have different data distribution in real world problems, thus the covariance matrices of these classes are not the same. In these situations, the Mahalanobis is often more popular than Euclidean as a measure of distance. In this paper, in addition to apply the idea of S-TWSVM, the classical Euclidean distance is replaced by Mahalanobis distance which leads to simultaneously consider the covariance matrices of the two classes. By this modification, the orientation information in two classes can be better exploited than S-TWSVM. The experiments indicate our proposed algorithm is often superior to other learning algorithms in terms of generalization performance.

Keywords: *Structural information, Non-parallel hyperplanes, Non-parallel hyperplanes, Ward's linkage, Twin support vector machine, Structural twin support vector machine.*

1. Introduction

Recently, Support Vector Machines (SVMs) have been known as a popular algorithm in the fields of classification, regression, pattern recognition [2, 28-29]. Classical SVM assumes that two classes can be illustrated with a hyperspherical shape indicating samples in two classes follow the same distribution trend. However two classes have different distribution trends thus using a different hyperellipsoidal shape for each class can better demonstrate its data points. In this situation, Mahalanobis distance can better deal with hyperellipsoidal shapes [12, 23-24]. As all we know Mahalanobis distance uses the covariance matrix of a class which this matrix indicates the distribution trend of data points in the corresponding class. The Euclidean distance which used in many classifiers is a special case of the Mahalanobis distance such that it assumes the covariance matrix of the corresponding class is an identity matrix indicates the distribution trend of data points in the corresponding class is the same in all directions. So due to the nature of hyperellipsoidal shapes which the distribution trend of a class is not the same in different directions, Mahalanobis distance can better perform than Euclidean distance in these situations.

In literature, many efforts have been proposed to illustrate Mahalanobis distance in classifying the data points [1, 3, 4, 14, 15, 22, 26]. On the other hand, in recent years, many structural information based classifiers, have also been proposed, such as structured large margin machine (SLMM) [5], ellipsoidal kernel machine (EKM) [21], mini-max probability machine (MPM) [8] and maxi-min margin machine (M^4) [15] with the higher computational cost than the classical SVM. Recently a structural regularized SVM (SRSVM) [9], has proposed which captures the structural information, based on the cluster granularity.

In [13, 35], an improvement on the speed of SVM has been proposed, called twin support vector machine (TWSVM). It finds two non-parallel hyperplanes instead of two parallel hyperplanes as in SVM. Each TWSVM's hyperplane is obtained by solving two half size QPPs, instead of one large QPP as in the classical SVM. It can be seen in [4], not only the TWSVM' learning speed is higher than that of SVM, but also its test accuracy is also improved. There are also some extensions on TWSVM, such as, the least squares version of TWSVM (LS-TSVM) [18, 19], smooth TWSVM [17], geometric algorithms [30], sparse TWSVM [31], twin support vector regressions (TSVRs) [33, 34].

On the other hand, TWSVMs suffer from lack of the structural information in its optimization problems. Hence, there have been several improvements on TWSVMs to add

this prior knowledge into its learning process. For instance, structural twin support vector machine (S-TWSVM) [36] has been proposed, which introduce the structural information derived by clustering methods into its optimization problems. Each S-TWSVM hyperplane exploit only one class structural information and lies closest to it and simultaneously getting far away from the other class. Another work is the twin Mahalanobis distance based support vector machine (TMSVM) [32]. This algorithm considers the covariance matrices of two classes and introduces them into Mahalanobis distances in its QPPs. Therefore TMSVM not only has faster leaning speed than classical SVM, but also instead of the covariance matrix of the total dataset, it simultaneously considers the covariance matrices of the two classes.

In this paper, we propose an improvement on S-TWSVM which is called Mahalanobis distance-based S-TWSVM (MS-TWSVM) classifier. In MS-TWSVM, not only the data structures of the two classes, based on the cluster granularity [9] introduced into the optimization problems, but also our algorithm substitutes Mahalanobis distance for classical Euclidean distance under the corresponding cluster granularity.

Therefore MS-TWSVM can efficiently inherit the merits of S-TWSVM and TMSVM. In comparison with S-TWSVM, MS-TWSVM can better exploit the orientation information in the two classes, and contrary to other Mahalanobis distance-based methods, for all clusters in the two classes, it considers the sum of them respectively. Thus MS-TWSVM can effectively exploit the structural information in the two classes.

Our proposed algorithm is evaluated with other learning algorithms on both synthetic (OR and XOR classification problems) and UCI benchmark datasets [7]. The results show that MS-TWSVM achieves higher generalization performance than other algorithms in almost all cases.

This paper is organized as the following: section 2 discussed a recent improvement on SVM, S-TWSVM. In next section, our proposed algorithm and its formulation are discussed. In section 5 experimental results on datasets are given and in last section conclusions are discussed.

2. Background

Suppose the training points for two classes are as follow

$$X^{(i)} = [x_1^{(i)}, x_2^{(i)}, ..., x_{N_i}^{(i)}],\ i = 1, 2 \qquad (1)$$

where N_i is the samples with n dimension in class i, such as matrix A with $N_1 \times n$ samples (A_i as ith sample) in class +1 and matrix B with $N_2 \times n$ samples in class -1, where $N_1 + N_2 = N$.

2.1 Structural twin support vector machine

In the last decade, the structural information in data has been focused in classification algorithms and machine learning methods. Many extensions of SVM based on the structural information have been proposed [5, 8, 9, 15, 21, 32, 36]. In these algorithms, the structural information is exploited by some clustering techniques and efficiency introduced to optimizing problem using covariance matrices of clusters and construct reasonable classifiers.

In S-TWSVM, each hyperplane exploits the structure information in one class and lies closest to it and simultaneously being far from the other class. So S-TWSVM can better use this prior information within samples leads to improve its generalization ability. Like other structure-based methods, S-TWSVM has two steps, clustering and model learning. In clustering step, the structure of data distribution in classes, is exploited by some clustering techniques [10, 11, 16, 27]. The clustering methods used in S-TWSVM is Ward's linkage clustering (WIL) [10], the same as other structure-based methods. In model learning step, suppose the two S-TWSVM's hyperplanes for two classes expressed as follow:

$$f_1(x) = w_1^T x + b_1 = 0 \ , \ \ f_2(x) = w_2^T x + b_2 = 0, \qquad (2)$$

where $w_1, w_2 \in R^n, b_1, b_2 \in R$. So two optimization problems of S-TWSVM must be solved

$$\begin{aligned} &\min_{w_1, b_1, \xi} \ \frac{1}{2}\|Aw_1 + e_1 b_1\|_2^2 + c_1 e_2^T \xi + \frac{1}{2} c_2 (\|w_1\|_2^2 + b_1^2) + \frac{1}{2} c_3 w_1^T \Sigma_1 w_1, \\ &s.t. \quad -(Bw_1 + e_2 b_1) + \xi \geq e_2 \ , \ \xi \geq 0, \end{aligned} \qquad (3)$$

for the positive class (1) and

$$\begin{aligned} &\min_{w_2, b_2, \xi} \ \frac{1}{2}\|Bw_2 + e_2 b_2\|_2^2 + c_4 e_1^T \eta + \frac{1}{2} c_5 (\|w_2\|_2^2 + b_2^2) + \frac{1}{2} c_6 w_2^T \Sigma_2 w_2, \\ &s.t. \quad (Aw_2 + e_1 b_2) + \eta \geq e_1 \ , \ \eta \geq 0, \end{aligned} \qquad (4)$$

for the negative class (2), where $c_1, ..., c_6 \geq 0$ are the pre-specified penalty factors, ξ_i and η_i are the slack variables, $\Sigma_1 = \Sigma_{P_1} + ... + \Sigma_{P_{C_P}}, \Sigma_2 = \Sigma_{N_1} + ... + \Sigma_{N_{C_N}}$, Σ_{P_i} denote the covariance matrix of ith and Σ_{N_j} for jth cluster in the two classes, respectively ($i = 1, ..., C_P$, $j = 1, ..., C_N$).

A new test point x is classified according to its distance from the two hyperplanes (2), i.e.,

$$f(x) = \arg\min_{1,2} \{d_1(x), d_2(x)\}, \qquad (5)$$

where

$$d_1(x) = \left|w_1^T x + b_1\right|, \quad d_2(x) = \left|w_2^T x + b_2\right| \qquad (6)$$

and $|.|$ is the perpendicular distance of point x from the hyperplanes $w_1^T x + b_1$ or $w_2^T x + b_2$.

For the nonlinear case, the two vectors w_1 and w_2 could be expressed in Hilbert space $\mathcal{H}$ as $w_1 = \Sigma_{i=1}^{m_1+m_2} (\lambda_1)_i \varphi(x_i) = \varphi(M)\lambda_1$ and $w_2 = \Sigma_{i=1}^{m_1+m_2} (\lambda_2)_i \varphi(x_i) = \varphi(M)\lambda_2$, respectively, where m_1 and m_2 are the size of data points in the positive and negative classes, respectively. So S-TWSVM computes the kernel-generated hyperplanes.

3. Mahalanobis distance based structural twin support vector machine

Now we express the details of the proposed algorithm which we call as Mahalanobis distance-based structural twin support vector machine (MS-TWSVM) classifier. Similar to S-TWSVM, our proposed algorithm is comprised of the clustering and model learning steps.

3.1 Clustering

In first step, the samples in each class are analyzed to find the points which have the same data distribution and are collected as one cluster. There are many clustering methods that can be used such as nearest neighbor clustering [27], k-means [11] and fuzzy clustering [16]. After clustering, the structural information through the covariance matrices of the clusters, introduced into the optimization problems. We use the Ward's linkage clustering (WIL) [10] technique for clustering the samples which derive the compact and spherical clusters.

Suppose that A and B are two clusters, then Ward's linkage $W(A, B)$ between these two clusters can be calculated as

$$W(A,B) = \frac{|A|.|B| \cdot \|\mu_A - \mu_B\|^2}{|A| + |B|} \tag{7}$$

where μ_A and μ_B are the means of two clusters respectively.

At the First of execution, it assumes one sample as a distinct cluster. Now suppose x_1 and x_2 are two examples, the Ward's linkage between x_1 and x_2 is

$$W(x_1, x_2) = \frac{\|x_1 - x_2\|^2}{2}, \tag{8}$$

when x_1 and x_2 are being merged to construct A' , Ward's linkage between A' and C is calculated as

$$W(A', C) = \frac{(|A| + |C|)W(A, C) + (|B| + |C|)W(B, C) - |C|W(A, B)}{|A| + |B| + |C|}. \tag{9}$$

As can be seen from above, while the clusters are being merged, the ward's linkage between them increases and the cluster amounts decreases [5]. In order to find the optimal number of clusters, we should determine the knee point in the curve with the merge distance in vertical axis and the number of clusters in horizontal and as the knee point is found, the clustering process should be stopped [25].

3.2 Linear MS-TWSVM

We express the positive and negative clusters as $P = \{P_1, ..., P_i, ..., P_{C_P}\}$, $N = \{N_1, ..., N_j, ..., N_{C_N}\}$. Positive samples are represented as $A \in R^{I_1 \times d}$ and $B \in R^{I_2 \times d}$ shows all samples that exists in the negative class ($I_1 + I_2 = l$).

MS-TWSVM expresses two non-parallel hyperplanes:

$$f_+(x) = w_+^T \Sigma_+^{-1} x + b_+ = 0 \ , \ f_-(x) = w_-^T \Sigma_-^{-1} x + b_- = 0 \tag{10}$$

where $w_+, w_- \in R^d, b_+, b_- \in R$.

By employing the Mahalanobis distance instead of Euclidean distance, the following two optimization problems are obtained:

$$\begin{aligned} &\min_{w_+, b_+, \xi_j} \ \frac{1}{2}\sum_{i \in I_1} (w_+^T \Sigma_+^{-1} x_i + e_+ b_+)^2 + c_1 e_-^T \sum_{j \in I_2} \xi_j + \frac{1}{2} c_2 (w_+^T w_+ + b_+^2) + \frac{1}{2} c_3 w_+^T \Sigma_+ w_+, \\ &s.t. \quad -(w_+^T \Sigma_+^{-1} x_j + e_- b_+) + \xi_j \geq e_- \ , \ \xi_j \geq 0, \ j \in I_2 \end{aligned} \tag{11}$$

$$\begin{aligned} &\min_{w_-, b_-, \eta_i} \ \frac{1}{2}\sum_{j \in I_2} (w_-^T \Sigma_-^{-1} x_j + e_- b_-)^2 + c_4 e_+^T \sum_{i \in I_1} \eta_i + \frac{1}{2} c_5 (w_-^T w_- + b_-^2) + \frac{1}{2} c_6 w_-^T \Sigma_- w_-, \\ &s.t. \quad (w_-^T \Sigma_-^{-1} x_i + e_+ b_-) + \eta_i \geq e_+ \ , \ \eta_i \geq 0, \ i \in I_1 \end{aligned} \tag{12}$$

where $c_1, c_2, ..., c_6 \geq 0$ are the pre-specified penalty factors, ξ_j and η_i are the slack variables, $\Sigma_+ = \Sigma_{P_1} + ... + \Sigma_{P_{C_P}}$, $\Sigma_- = \Sigma_{N_1} + ... + \Sigma_{N_{C_N}}$, Σ_{P_i} and Σ_{N_j} are the covariance matrices corresponding to ith and jth clusters in the two classes, respectively ($i = 1, ..., C_P$, $j = 1, ..., C_N$).

In the equations (11) and (12), similar to S-TWSVM, by adding terms $w_\pm^T \Sigma_\pm w_\pm$, the compactness of the corresponding classes will be kept. In addition in optimization problem (11) and (12), we use Mahalanobis distances instead of Euclidean distance. As we know, in many real-world classification problems, Mahalanobis distance can better operate than Euclidean distance, so with this substitution, the corresponding covariance matrices of the two classes (usually have different data structures), are simultaneously considered in MS-TWSVM. This improvement can take full advantage of the structural information in two classes of data.

We calculate the Lagrangian functions for (11) and (12) and use the Karush-Kuhn-Tucker (KKT) conditions for (11) and (12), so the QP problems for them are shown as follow:

$$\underset{\alpha}{Max} \quad e_-^T\alpha-\frac{1}{2}\alpha^T G(H^TH+c_2I+c_3J)^{-1}G^T\alpha \qquad (13)$$
$$s.t. \qquad 0\le\alpha\le c_1 e_-$$

where

$$H=[A\Sigma_+^{-1} \quad e_+] \ , \ G=[B\Sigma_+^{-1} \quad e_-], \ J=\begin{bmatrix}\Sigma_+ & 0\\ 0 & 0\end{bmatrix} \qquad (14)$$

and

$$\underset{\beta}{Max} \quad e_+^T\beta-\frac{1}{2}\beta^T P(Q^TQ+c_5I+c_6F)^{-1}P^T\beta \qquad (15)$$
$$s.t. \qquad 0\le\beta\le c_4 e_+$$

where

$$P=[A\Sigma_-^{-1} \quad e_-] \ , \ Q=[B\Sigma_-^{-1} \quad e_+], \ F=\begin{bmatrix}\Sigma_- & 0\\ 0 & 0\end{bmatrix} \qquad (16)$$

After optimizing these two dual QP problems, the vectors $v_+=[w_+^T \ \ b_+^T]^T$ and $v_-=[w_- \ \ b_-]^T$ are computed by following formulas:

$$v_+=-(H^TH+c_2I+c_3J)^{-1}(G^T\alpha) \qquad (17)$$

and

$$v_-=-(Q^TQ+c_5I+c_6F)^{-1}(P^T\beta) \qquad (18)$$

where we denote I as an identity matrix with appropriate dimensions. As seen in matrix theory [6], $H^TH+c_2I+c_3J$ and $Q^TQ+c_5I+c_6F$ are positive definite matrices.

After obtaining the vectors v_+ and v_- from (17) and (18), the hyperplanes

$$w_+^T\Sigma_+^{-1}x+b_+=0 \ , \ w_-^T\Sigma_-^{-1}x+b_-=0 \qquad (19)$$

are known. Similar to S-TWSVM, a new test point $x\in R^n$ gets its class label according to which hyperplanes lie closest to it, i.e.,

$$f(x)=\arg\min_{+,-}\{d_+(x),d_-(x)\}, \qquad (20)$$

where

$$d_+(x)=\frac{\left|w_+^T\Sigma_+^{-1}x+b_+\right|}{\sqrt{w_+^T\Sigma_+^{-1}w_+}} \ \text{ and } \ d_-(x)=\frac{\left|w_-^T\Sigma_-^{-1}x+b_-\right|}{\sqrt{w_-^T\Sigma_-^{-1}w_-}} \qquad (21)$$

3.3 Nonlinear MS-TWSVM

We extend linear MS-TWSVM to nonlinear version. For this purpose the kernel trick is introduced to the optimizing problems causing the samples separable more linearly. So the two primal QPPs of nonlinear MS-TWSVM are as follow:

$$\min_{w_+,b_+,\xi_j} \ \frac{1}{2}\sum_{i\in I_1}(w_+^T\Sigma_+^{-1}\varphi(x_i)+e_+b_+)^2+c_1e_-^T\sum_{j\in I_2}\xi_j+\frac{1}{2}c_2(w_+^Tw_++b_+^2)+\frac{1}{2}c_3w_+^T\Sigma_+w_+,$$
$$s.t. \quad -(w_+^T\Sigma_+^{-1}\varphi(x_j)+e_-b_+)+\xi_j\ge e_- \ , \ \xi_j\ge 0, \ j\in I_2 \qquad (22)$$

$$\min_{w_-,b_-,\eta_i} \ \frac{1}{2}\sum_{j\in I_2}(w_-^T\Sigma_-^{-1}\varphi(x_j)+e_-b_-)^2+c_4e_+^T\sum_{i\in I_1}\eta_i+\frac{1}{2}c_5(w_-^Tw_-+b_-^2)+\frac{1}{2}c_6w_-^T\Sigma_-w_-,$$
$$s.t. \quad (w_-^T\Sigma_-^{-1}\varphi(x_i)+e_+b_-)+\eta_i\ge e_+ \ , \ \eta_i\ge 0, \ i\in I_1 \qquad (23)$$

where $\Sigma_+=\Sigma_{P_1}+...+\Sigma_{P_{c_p}}$, $\Sigma_-=\Sigma_{N_1}+...+\Sigma_{N_{c_N}}$, Σ_{P_i} and Σ_{N_j} are the kernel covariance matrices for the i[th] and j[th] clusters in the two positive and negative classes, respectively ($i=1,...,C_P, j=1,...,C_N$). These covariance matrices are computed as $\Sigma_{P_i}=\varphi(A_{P_i})J_{P_i}J_{P_i}^T\varphi(A_{P_i})^T$ and $\Sigma_{N_j}=\varphi(B_{N_j})J_{N_j}J_{N_j}^T\varphi(B_{N_j})^T$. So the covariance matrix Σ_+ of the positive class can be obtained as follow:

$$\Sigma_+=\Sigma_{P_1}+...+\Sigma_{P_{c_p}}=\sum_{i=1}^{c_p}\varphi(A_{P_i})J_{P_i}J_{P_i}^T\varphi(A_{P_i})^T=$$
$$[\varphi(A_{P_1})....\varphi(A_{P_{c_p}})]\times\begin{bmatrix}J_{P_1} & & \\ & \ddots & \\ & & J_{P_{c_p}}\end{bmatrix}\times\begin{bmatrix}J_{P_1} & & \\ & \ddots & \\ & & J_{P_{c_p}}\end{bmatrix}\times\begin{bmatrix}\varphi(A_{P_1})^T\\ \vdots \\ \varphi(A_{P_{c_p}})^T\end{bmatrix}= \qquad (24)$$
$$\varphi(A)J_+J_+^T\varphi(A)^T$$

where

$$\varphi(A)=[\varphi(A_{P_1})...\varphi(A_{P_{c_p}})] \ \text{ and } \ J_+=\begin{bmatrix}J_{P_1} & & \\ & \ddots & \\ & & J_{P_{c_p}}\end{bmatrix}. \qquad (25)$$

Similarly for the negative class, the covariance matrix Σ_- is computed as

$$\Sigma_-=\varphi(B)J_-J_-^T\varphi(B)^T \qquad (26)$$

where

$$\varphi(B)=[\varphi(B_{N_1})...\varphi(B_{N_{c_N}})] \text{ and } J_-=\begin{bmatrix}J_{N_1} & & \\ & \ddots & \\ & & J_{N_{c_N}}\end{bmatrix}. \qquad (27)$$

As we know, the terms $\Sigma_\pm^{-1}$ are usually ill-conditioned. So by adding a small positive number σ, the terms $(\sigma I+\Sigma_\pm^{-1})^{-1}$ could be positive definite matrices.

Consider the Woodbury matrix identity [20], $(U+VV^T)^{-1}=U^{-1}-U^TV(I+V^TU^{-1}V)^{-1}V^TU^{-1}$, we set $U=\sigma I$ and $V=\varphi(A)J_+$, so we have

$$(\sigma I+\Sigma_+)^{-1}=[\sigma I+\varphi(A)J_+J_+^T\varphi(A)^T]^{-1}= \qquad (28)$$
$$\sigma^{-1}I-\sigma^{-1}\varphi(A)J_+(\sigma I+J_+^TK_AJ_+)^{-1}J_+^T\varphi(A)^T$$

where $K_A = \varphi(A)^T \varphi(A) = k(A,A)$. Similarly we obtain

$$(\sigma \mathrm{I} + \Sigma_-)^{-1} = [\sigma \mathrm{I} + \varphi(B) J_- J_-^T \varphi(B)^T]^{-1} = \sigma^{-1}\mathrm{I} - \sigma^{-1}\varphi(B) J_- (\sigma \mathrm{I} + J_-^T K_B J_-)^{-1} J_-^T \varphi(B)^T \quad (29)$$

where $K_B = \varphi(B)^T \varphi(B) = k(B,B)$. Thus, the Mahalanobis distance-based kernels in the feature space H , can be computed by these two covariance matrices, as follow:

$$\begin{aligned} K_+(x_i, x_j) &= \left\langle \varphi(x_i).\varphi(x_j) \right\rangle_+ \approx \varphi(x_i)^T (\sigma I + \Sigma_+)^{-1} \varphi(x_j) \\ &= \varphi(x_i)^T [\sigma^{-1} I - \sigma^{-1}\varphi(A) J_+ (\sigma I + J_+^T K_A J_+)^{-1} J_+^T \varphi(A)^T] \varphi(x_j) \\ &= \sigma^{-1} K(x_i, x_j) - \sigma^{-1} K(x_i, A) J_+ (\sigma I + J_+^T K_A J_+)^{-1} J_+^T K(A, x_j). \end{aligned} \quad (30)$$

$$\begin{aligned} K_-(x_i, x_j) &= \left\langle \varphi(x_i), \varphi(x_j) \right\rangle_- \approx \varphi(x_i)^T (\sigma I + \Sigma_-)^{-1} \varphi(x_j) \\ &= \varphi(x_i)^T [\sigma^{-1} I - \sigma^{-1}\varphi(B) J_- (\sigma I + J_-^T K_B J_-)^{-1} J_-^T \varphi(B)^T] \varphi(x_j) \\ &= \sigma^{-1} K(x_i, x_j) - \sigma^{-1} K(x_i, B) J_- (\sigma I + J_-^T K_B J_-)^{-1} J_-^T K(B, x_j). \end{aligned} \quad (31)$$

Now we obtain the following MS-TWSVM hyperplanes (32) by employing the above kernels:

$$w_+^T K_+(C,x) + b_+ = 0 \text{ and } w_-^T K_-(C,x) + b_- = 0 \quad (32)$$

Where Mahalanobis distance-based kernels $K_+(.,.)$ and $K_-(.,.)$ can be computed by (30) and (31). So Mahalanobis distance-based QPPs are as follow:

$$\begin{aligned} \min \ & \frac{1}{2}\sum_{i\in I_1} (w_+^T K_+(C,x_i) + e_+ b_+)^2 + c_1 e_-^T \sum_{j\in I_2} \xi_j + \\ & \frac{1}{2} c_2 (w_+^T w_+ + b_+^2) + \frac{1}{2} c_3 w_+^T \Sigma_+ w_+, \\ s.t. \ & -(w_+^T K_+(C,x_j) + e_- b_+) + \xi_j \geq e_- \quad , \quad \xi_j \geq 0, \quad j \in I_2 \end{aligned} \quad (33)$$

$$\begin{aligned} \min \ & \frac{1}{2}\sum_{j\in I_2} (w_-^T K_-(C,x_j) + e_- b_-)^2 + c_4 e_+^T \sum_{i\in I_1} \eta_i + \\ & \frac{1}{2} c_5 (w_-^T w_- + b_-^2) + \frac{1}{2} c_6 w_-^T \Sigma_- w_-, \\ s.t. \ & (w_-^T K_-(C,x_i) + e_+ b_-) + \eta_i \geq e_+ \quad , \quad \eta_i \geq 0, \quad i \in I_1 \end{aligned} \quad (34)$$

By some simple computing, the Wolfe Duals of (33) and (34) can be shown as:

$$\max \ e^T \alpha - \frac{1}{2}\alpha^T R^T (SS^T + c_2 I + c_3 Z)^{-1} R\alpha \quad (35)$$
$$s.t. \quad 0 \leq \alpha \leq c_1 e$$

$$\max \ e^T \beta - \frac{1}{2}\beta^T M^T (NN^T + c_5 I + c_6 Y)^{-1} M\beta \quad (36)$$
$$s.t. \quad 0 \leq \beta \leq c_4 e$$

where α and β are the Lagrangian vectors and

$$S = [K_+(A,C),\ e_+]^T \quad , \quad R = [K_+(B,C),\ e_-]^T \quad (37)$$

$$M = [K_-(A,C),\ e_-]^T \quad , \quad N = [K_-(B,C),\ e_+]^T \quad (38)$$

By solving the problems (35) and (36), the vectors u and v are obtained as follows:

$$u = -(SS^T + c_2 I + c_3 Z)^{-1} R\alpha \quad (39)$$

$$v = -(NN^T + c_5 I + c_6 Y)^{-1} M\beta \quad (40)$$

Where

$$Z = \begin{bmatrix} \Sigma_+ & 0 \\ 0 & 0 \end{bmatrix} \text{ and } Y = \begin{bmatrix} \Sigma_- & 0 \\ 0 & 0 \end{bmatrix} \quad (41)$$

$$\Sigma_+ = \varphi(A) J_+ J_+^T \varphi(A)^T = k(C,A) J_+ J_+^T k(A,C) \quad (42)$$

$$\Sigma_- = \varphi(B) J_- J_-^T \varphi(B)^T = k(C,B) J_- J_-^T k(B,C) \quad (43)$$

4. Experiments

To compare the performance of MS-TWSVM, S-TWSVM, TMSVM and TSVM, we execute these algorithms on UCI benchmark datasets [7]. To evaluate the classification accuracy of MS-TWSVM in comparison to other algorithms, several benchmark datasets in the UCI database are used. In the feature space, we use the Gaussian kernel to compare the algorithms. In all experiments, for simplicity we set $c_1 = c_4$, $c_2 = c_5$, $c_3 = c_6$ and use ten-fold cross validation procedure to measure the testing accuracy. All parameters c_1 , c_2 , c_3 and γ in the Gaussian kernel, are selected by ten-fold cross validation from $\{2^i \,|\, i = -7, ..., 7\}$ on 10% of training samples. Samples are normalized in the range [-1 1]. All methods are implemented on PC with 2.4 GHz Intel core i7 and 4 GB of memory.

4.1 UCI datasets

Now we illustrate the result of executions for four algorithms MS-TWSVM, S-TWSVM, TMSVM and TMSVM on UCI datasets [12]. The dimensions and sizes of training and testing data for each dataset are shown in Table [3]. Model's parameters selection is performed by the method of ten-fold cross validation on 10% of the training set.

The test accuracies and CPU training times are shown in the Table [1]. The following results are given from Table [1]; first MS-TWSVM's test accuracies on almost all datasets, is higher than other three algorithms. This is because MS-TWSVM's optimization problems conclude the corresponding data structures of two classes, and by substituting Mahalanobis distance for Euclidean distance, it can better capture the orientation information in each class and causes further improvement on the generalization performance. Second the CPU training time of MS-TWSVM is higher than the others. This is a reasonable result because MS-TWSVM needs to perform clustering phase to exploit the data structures, and also needs to manipulate matrix inversions. As seen in Table [1], The CPU training time of TMSVM is less than S-TWSVM's. This more training time is spent on the clustering step in S-TWSVM which TMSVM doesn't have it. However, these two algorithms are comparable to TSVM in terms of

classification accuracy for almost all datasets. On the other hand, TSVM is the fastest algorithm among these algorithms which it doesn't have the clustering step to exploit the structural information and the distance measure is used in TSVM is Euclidean distance which requires less computation efforts than Mahalanobis distance. The classification accuracy of TSVM, since it doesn't need to exploit the orientation information or the data structures in two classes, is less than the others.

Table 1: The result of nonlinear MS-TWSVM and S-TWSVM on benchmark datasets.

Datasets	***MS-TWSVM*** ***Acc.*** ***Exe. (s)***	***S-TWSVM*** ***Acc.*** ***Exe. (s)***	***TSVM*** ***Acc.*** ***Exe. (s)***	***TMSVM*** ***Acc.*** ***Exe. (s)***
Banana	71.89±0.1125 7.88	64.34±0.1525 4.94	62.92±0.1500 1.40	69.20±0.126 2.81
Breast Cancer	73.90±0.051 0.98	73.64±0.0465 0.80	72.47±0.036 0.29	73.64±0.037 0.43
Diabetes	75.33±0.064 6.99	74.34±0.1208 3.80	72.23±0.049 1.97	71.00±0.059 2.53
Flare	65.75±0.042 18.31	64.25±0.0510 4.94	61.76±0.052 2.34	62.26±0.054 10.98
German	71.73±0.033 9.93	71.30±0.0341 5.03	70.40±0.021 3.57	71.13±0.025 4.64
Heart	79.60±0.059 0.6339	78.10±0.0963 0.5843	77.50±0.081 0.1237	75.20±0.085 0.2639
Image	82.34±0.040 22.81	71.00±0.0507 17.61	70.77±0.037 13.12	80.61±0.0608 15.95
Ringnorm	67.72±0.073 2.75	66.55±0.1139 2.03	60.55±0.123 1.80	61.69±0.110 2.67
Splice	84.63±0.094 11.62	78.00±0.0103 10.79	73.98±0.130 2 5.67	83.45±0.007 7.45
Thyroid	69.33±0.109 0.8098	67.20±0.1638 0.5224	64.27±0.035 0.1162	69.33±0.1461 0.1658
Titanic	78.45±0.057 0.4807	77.77±0.0828 0.4619	77.52±0.067 0. 0876	77.60±0.078 0.1070

5. Conclusions

We proposed an extension of the structural twin support vector machine called MS-TWSVM algorithm. The idea behind the proposed algorithm is to improve capturing the orientation information in two classes of data, by substituting Mahalanobis distance for Euclidian distance in S-TWSVM. As all we know, for many real-world problems, Mahalanobis distance can be comparable to Euclidean distance and leading MS-TWSVM to simultaneously consider the corresponding covariance matrices of the two classes. So MS-TWSVM can efficiently capture the orientation information of the two classes. As mentioned earlier, our proposed method aims to sum the covariance matrices all clusters in two classes respectively which leads to the structural information of two classes be different. As seen in the experiments, MS-TWSVM can better exploit the structural information of two classes and improve the classification accuracy. One of the future works is to experiment our proposed method on the large scale problems. On the other hand our method exploits the structural information based on the cluster granularity. So using the point granularity instead of the cluster granularity is a new topic.

References

[1] A. Ruiz, P.L.-T. Pedro, Nonlinear kernel-based statistical pattern analysis, IEEE Transactions on Neural Networks 12 (2001) 16–32.

[2] B. Boser, L. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the 5th Annual Workshop on Computational Learning Theory, ACM Press, Pittsburgh, (1992) 144–152.

[3] B. Haasdonk, E. Pekalska, Classification with kernel Mahalanobis distance classifiers, in: Advances in Data Analysis, Data Handling and Business Intelligence Studies in Classification, Data Analysis, and Knowledge Organization, (2010) Part 5 351–361.

[4] D. Wang, D.S. Yeung, E.C.C. Tsang, Weighted the Mahalanobis distance kernels for support vector machines, IEEE Transactions on Neural Networks 18 (2007) 1453–1462.

[5] D. Yeung, D. Wang, W. Ng, E. Tsang, X. Zhao, Structured large margin machines: sensitive to data distributions, Machine Learning 68 (2) (2007) 171–200.

[6] F.R. Gantmacher, Matrix Theory, New York, Chelsea, (1990).

[7] G. Rätsch, Benchmark Repository, datasets, http://ida.first.fhg.de/projects/bench/benchmarks.htm, (2000).

[8] G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, M.I. Jordan, A robust minimax approach to classfication, Journal of Machine Learning & Research 3 (2002) 555–582.

[9] 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.

[10] J. H. Ward, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association 58 (301) (1963) 236–244.

[11] J.A. Hartigan, M.A. Wong, A k-means clustering algorithm, Applied Statistics 28 (1) (1979) 100–108.

[12] J.A. Schinka, W.F. Velicer, I.B. Weiner, Research methods in psychology, Wiley, New york, (2003).

[13] 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.
[14] K. Huang, H. Yang, I. King, M.R. Lyu, L. Chan, The minimum error minimax probability machine, Journal of Machine Learning Research 5 (2004) 1253–1286.
[15] K. Huang, H. Yang, I. King, M.R. Lyu, Maxi-min margin machine-learning large margin classifiers locally and globally, IEEE Transactions on Neural Networks (2008) 260–272.
[16] L.A. Zadeh, Fuzzy sets, Information Control 8 (1965) 338–353.
[17] M.A. Kumar, M. Gopal, Application of smoothing technique on twin support vector machines, Pattern Recognition Letter 29 (6) (2008) 1842–1848.
[18] M.A. Kumar, M. Gopal, Least squares twin support vector machines for pattern classification, Expert Systems with Applications 36 (4) (2009) 7535–7543.
[19] M.A. Kumar, R. Khemchandani, M. Gopal, S. Chandra, Knowledge based least squares twin support vector machines, Information Sciences 180 (16) (2010) 4606–4618.
[20] M.A. Woodbury, Inverting modified matrices, Memorandum Rept. 42,
Statistical Research Group, Princeton University, Princeton, NJ, (1950).
[21] P.K. Shivaswamy, T. Jebara, Ellipsoidal kernel machines, in Proceeding of 12th International Workshop on Artificial Intelligence Statistic, (2007) 1–8.
[22] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, B. Kijsirikul, A new kernelization framework for Mahalanobis distance learning algorithms, Neurocomputing 73 (10–12) (2010) 1570–1579.
[23] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemometrics and Intelligent Laboratory Systems 50 (2000) 1–18.
[24] R. Gnanadesikan, J.R. Kettenring, Robust estimates, residuals, and outlier detection with multiresponse data, Biometrics 28 (1972) 81–124.
[25] S. Salvador, P. Chan, Determining the Number of Clusters/Segments in
Hierarchical Clustering/Segmentation Algorithms, Tech. Rep., (2003).
[26] S. Xiang, F. Nie, C. Zhang, Learning a Mahalanobis distance metric for data clustering and classification, Pattern Recognition 41 (2008) 3600–3612.
[27] S.-Y. Lu, K.S. Fu, A sentence-to-sentence clustering procedure for pattern analysis, IEEE Transactions on Systems Man & Cybernetics 8 (5) (1978) 381–389.
[28] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, (1998).
[29] V.N. Vapnik, The Natural of Statistical Learning Theory, Springer, New York, (1995).
[30] X. Peng, A ν-twin support vector machine (ν-TSVM) classifier and its geometric algorithms, Information Sciences 180 (8) (2010) 3863–3875.
[31] X. Peng, Building sparse twin support vector machine classifiers in primal space, Information Sciences 181 (11) (2011) 3967–3980.
[32] X. Peng, D. Xu, Twin Mahalanobis distance-based support vector machines for pattern recognition, Information Sciences 200 (1) (2013) 22–37.
[33] X. Peng, Primal twin support vector regression and its sparse approximation, Neurocomputing 73 (16–18) (2010) 2846–2858.
[34] X. Peng, TSVR: an efficient twin support vector machine for regression, Neural Networks 23 (3) (2010) 365–372.
[35] 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.
[36] Z. Qi, Y. Tian, Y. Shi, Structural twin support vector machine for classification, Knowledge-Based Systems 43 (2013) 74–81.