Efficient and robust TWSVM classifier based on L1-norm distance metric for pattern classification

He Yan
School of Computer Science and Engineering, Nanjing University of Science and Technology
Xiaolingwei 200, Nanjing, 210094, P. R. China
yanhecom@163.com

Tian-An Zhang
Collaborative Innovation Center of Sustainable Forestry in Southern China of Jiangsu Province, Nanjing Forestry University
Longpan Road 159, Nanjing, 210037, P. R. China
dyweld@sina.com

Qiao-Lin Ye
College of Information Science & Technology, Nanjing Forestry University
Longpan Road 159, Nanjing, 210037, P. R. China
yqlcom@njfu.edu.cn

Dong-Jun Yu*
School of Computer Science and Engineering, Nanjing University of Science and Technology
Xiaolingwei 200, Nanjing, 210094, P. R. China
njyudj@njust.edu.cn
*Corresponding Author

Abstract

Twin support vector machine (TWSVM) is a classical distance metric learning method for classification problems. The formulation of TWSVM criterion is based on L2-norm distance, which makes TWSVM prone to being influenced by the presence of outliers. In this paper, to develop a robust distance metric learning method, we propose a new objective for TWSVM classifier using L1-norm distance metric, termed as L1-TWSVM. The optimization strategy is to maximize the ratio of the inter-class distance dispersion to the intra-class distance dispersion by using L1-norm distance rather than L2-norm distance. Besides, we design a simple and valid iterative algorithm to solve L1-norm optimal problems, which is easy to actualize and its convergence to an optimum is theoretically ensured. The efficiency and robustness of L1-TWSVM have been validated by experiments on UCI datasets and artificial datasets. The promising experimental results indicate that our proposals outperform relevant state-of-the-art methods in all kinds of experimental settings.

1. Introduction

Support Vector Machine (SVM) [1-5] is a useful tool for data classification and regression analysis, which is under the constraint that the typical plane is parallel, and the maximum interval classification is implemented by solving the quadratic programming problems (QPPs). SVM is on the basis of structural risk minimization and VC dimension, which has good generalization performance, and it has been extensively used in many practical problems, such as image classification [6], scene classification [7], fault diagnosis [8], bioinformatics [9] and so on.

The advantages of SVM are obvious, but there are still some shortcomings. On one hand, XOR problem cannot be handled smoothly [10]; on the other hand, there is more computational complexity for QPPs [11, 12]. To alleviate these, Mangasarian and Wild proposed a proximal support vector machine via generalized eigenvalues (GEPSVM), which attempts to seek two nonparallel planes by solving a pair of generalized eigenvalue problems instead of complex QPPs, and reduces computation time, with its generalization ability better than PSVM [10]. The advantages of GEPSVM play an important role in its improvement [13-16]. Specially, Jayadeva et al. proposed a twin support vector machine [17] (TWSVM) based on the idea of GEPSVM. TWSVM solves a pair of QPPs (the scale is relatively small compared to that of SVM) to replace generalized eigenvalue problems [10]. Thus, computation cost of TWSVM is only 1/4 of the standard SVM [17]. Currently, many improved methods of TWSVM have been developed based on TWSVM. Ye [18] introduced regularization technique to optimize TWSVM, and proposed a feature selection method for TWSVM via the regularization technique (RTWSVM), which is a convex programming problem to overcome the possible singular problem and improve generalization ability. Kumar et al [19] reformulated the optimization problems of TWSVM by making use of constraints in the form of equalities to replace inequalities to modify the primal QPPs in least squares sense, and proposed a least squares version of TWSVM (LSTWSVM). The solutions of LSTWSVM follow directly from solving two linear equations opposed to solving two QPPs. Therefore, LSTWSVM effectively deals with large samples without any external optimization, and meanwhile, the computational cost is much less than that of TWSVM. Zhang [20] extends projective TWSVM [21] (PTSVM) by smoothing technique and proposed a smoothed projective TWSVM (SPTSVM), which solves two unconstrained differentiable optimization problems rather than two dual QPPs. Besides, an effective fast algorithm is used to optimize SPTSVM, which is called Newton-Armijo algorithm.

2327-0985/17 $31.00 © 2017 IEEE
DOI 10.1109/ACPR.2017.23

It should be noted that GEPSVM and TWSVM are sensitive to outliers, because L2-norm distance exaggerates the effect of outliers by the square operation [22], which reduces classification performance. Li [23] reformulated the optimization problems of nonparallel proximal support vector machine via L1-norm distance (L1-NPSVM). To solve the formulated objective, a gradient ascending (GA) iterative algorithm is proposed, which is simple to execute but may not guarantee the optimality of the solution due to both the need of introduction of a non-convex surrogate function and the difficult selection of step-size. The utilization of L1-norm distance is often considered as a simple and effective way to reduce the impact of outliers, which can also improve the generalization ability and flexibility of the model. For this purpose, we propose a robust TWSVM model using L1-norm distance for binary classification, called L1-TWSVM. It retains the original advantages of TWSVM, but improves the classification performance and robustness.

2. Related works

In this paper, all vectors are column vectors; $\mathbf{e}_1$ and $\mathbf{e}_2$ are identity column vectors; $\mathbf{I}$ is an identity matrix of appropriate dimension. The training sample is indicated by $\mathbf{T}=\left\{\left(\mathbf{x}_j^{(i)},y_i\right)\middle|i=1,2,\ j=1,2,...,m_i\right\}$, where $\mathbf{x}_j^{(i)}$ denotes the i-th class and j-th sample. Suppose that matrix $\mathbf{A}=\left[\mathbf{A}_1^{(1)},\mathbf{A}_2^{(1)},...,\mathbf{A}_{m_1}^{(1)}\right]^T$ with size of $m_1\times n$ represents the data points of class 1, while matrix $\mathbf{B}=\left[\mathbf{B}_1^{(2)},\mathbf{B}_2^{(2)},...,\mathbf{B}_{m_2}^{(2)}\right]^T$ with size of $m_2\times n$ represents the data points of class 2. Matrices $\mathbf{A}$ and $\mathbf{B}$ represent all the training data points, where $m_1+m_2=m$.

3. TWSVM classifier based on L1-norm distance

In TWSVM the distance is measured by L2-norm, it is known that L2-norm distance is sensitive to outliers, which implies that abnormal observations may affect the desired solution of TWSVM. In many literatures [23-25], L1-norm distance is usually considered as a robust alternative to L2-norm distance, which can improve the generalization ability and flexibility of the model. Motivated by the basic idea of L1-norm-based modelling, we embed L1-norm distance metric into the TWSVM model, termed as L1-TWSVM, which inherits the advantages of TWSVM (can overcome the problem of sample imbalance). That is, the ratio of m_1 to m_2 does not influence the classification result. To find the optimum solutions, L1-TWSVM needs to solve the following objective functions:

$$\begin{aligned}&\min_{\mathbf{w}_1,b_1,\mathbf{q}_1}\frac{1}{2}\|\mathbf{A}\mathbf{w}_1+\mathbf{e}_1b_1\|_1+c_1\mathbf{e}_2^T\mathbf{q}_1\\&s.t.\ -\left(\mathbf{B}\mathbf{w}_1+\mathbf{e}_2b_1\right)+\mathbf{q}_1\geq\mathbf{e}_2,\ \mathbf{q}_1\geq 0\end{aligned},\qquad(1)$$

$$\begin{aligned}&\min_{\mathbf{w}_2,b_2,\mathbf{q}_2}\frac{1}{2}\|\mathbf{B}\mathbf{w}_2+\mathbf{e}_2\mathrm{b}_2\|_1+c_2\mathbf{e}_1^T\mathbf{q}_2\\&s.t.\ \left(\mathbf{A}\mathbf{w}_2+\mathbf{e}_1b_2\right)+\mathbf{q}_2\geq\mathbf{e}_1,\ \mathbf{q}_2\geq 0\end{aligned}\qquad(2)$$

Formulas (1) and (2) are convex optimization problems with constraints in the form of inequalities, they have the optimal solutions, we can get two nonparallel optimal planes by solving them:

$$\mathbf{x}^T\mathbf{w}_1+b_1=0,\ \ \mathbf{x}^T\mathbf{w}_2+b_2=0\qquad(3)$$

where augmented vectors $\mathbf{z}_1=\left(\mathbf{w}_1\ b_1\right)^T$, $\mathbf{z}_2=\left(\mathbf{w}_2\ b_2\right)^T$. (1) and (2) can be optimized in the following form:

$$\begin{aligned}&\min_{\mathbf{w}_1,b_1,\mathbf{q}_1}\frac{1}{2}\left(\sum_{i=1}^{m_1}\frac{\left(\mathbf{A}_i\mathbf{w}_1+\mathbf{e}_ib_1\right)^2}{\mathbf{D}_i}\right)+c_1\mathbf{e}_2^T\mathbf{q}_1\\&s.t.\ -\left(\mathbf{B}\mathbf{w}_1+\mathbf{e}_2b_1\right)+\mathbf{q}_1\geq\mathbf{e}_2,\ \mathbf{q}_1\geq 0\end{aligned},\qquad(4)$$

$$\begin{aligned}&\min_{\mathbf{w}_2,b_2,\mathbf{q}_2}\frac{1}{2}\left(\sum_{j=1}^{m_2}\frac{\left(\mathbf{B}_j\mathbf{w}_2+\mathbf{e}_jb_2\right)^2}{\mathbf{D}_j}\right)+c_2\mathbf{e}_1^T\mathbf{q}_2\\&s.t.\ \left(\mathbf{A}\mathbf{w}_2+\mathbf{e}_1b_2\right)+\mathbf{q}_2\geq\mathbf{e}_1,\ \mathbf{q}_2\geq 0\end{aligned}\qquad(5)$$

where $\mathbf{D}_i=\left|\mathbf{A}_i\mathbf{w}_1+\mathbf{e}_ib_1\right|\neq 0$, $\mathbf{D}_j=\left|\mathbf{B}_j\mathbf{w}_2+\mathbf{e}_jb_2\right|\neq 0$. Obviously, it is difficult to solve formulas (4) and (5), because they contain absolute value operation, which makes the optimization of the formula (4) intractable. To solve these, we propose an iterative convex optimization strategy. The basic idea of this method is to iteratively update the augmented vector $\mathbf{z}_1$ until its objective function value converges to a fixed value. Assuming that $\mathbf{z}_1^{(p)}$ is the optimal solution for the iteration of p, where p denotes the iteration number. Then, the optimal solution of $\mathbf{z}_1^{(p+1)}$ for the iteration of $p+1$ is defined as the solution to the following problems:

$$\begin{aligned}&\min_{\mathbf{w}_1,b_1,\mathbf{q}_1}\frac{1}{2}\left(\sum_{i=1}^{m_1}\frac{\left(\mathbf{h}_i\mathbf{z}_1\right)^2}{\mathbf{D}_{1i}}\right)+c_1\mathbf{e}_2^T\mathbf{q}_1\\&s.t.\ -\mathbf{G}\mathbf{z}_1+\mathbf{q}_1\geq\mathbf{e}_2,\ \mathbf{q}_1\geq 0\end{aligned},\qquad(6)$$

$$\begin{aligned}&\min_{\mathbf{w}_2,b_2,\mathbf{q}_2}\frac{1}{2}\left(\sum_{j=1}^{m_2}\frac{\left(\mathbf{g}_j\mathbf{z}_2\right)^2}{\mathbf{D}_{2j}}\right)+c_2\mathbf{e}_1^T\mathbf{q}_2\\&s.t.\ \mathbf{H}\mathbf{z}_2+\mathbf{q}_2\geq\mathbf{e}_1,\ \mathbf{q}_2\geq 0\end{aligned}\qquad(7)$$

where $\mathbf{D}_{1i}=\left|\mathbf{h}_i\mathbf{z}_1^{(p)^T}\right|$, $\mathbf{D}_{2j}=\left|\mathbf{g}_j\mathbf{z}_2^{(p)^T}\right|$, $\mathbf{h}_i=\left[\mathbf{A}_i\ \ \mathbf{e}_1^i\right]$, $\mathbf{g}_j=\left[\mathbf{B}_j\ \ \mathbf{e}_2^j\right]$. Then formulas (6) and (7) are rewritten as:

$$\begin{aligned}&\min_{\mathbf{z}_1,\mathbf{q}_1}\frac{1}{2}\mathbf{z}_1^T\mathbf{H}^T\mathbf{D}_1\mathbf{H}\mathbf{z}_1+c_1\mathbf{e}_2^T\mathbf{q}_1\\&s.t.\ -\mathbf{G}\mathbf{z}_1+\mathbf{q}_1\geq\mathbf{e}_2,\ \mathbf{q}_1\geq 0\end{aligned}\qquad(8)$$

$$\begin{aligned}&\min_{\mathbf{z}_2,\mathbf{q}_2}\frac{1}{2}\mathbf{z}_2^T\mathbf{G}^T\mathbf{D}_2\mathbf{G}\mathbf{z}_2+c_2\mathbf{e}_1^T\mathbf{q}_2\\&s.t.\ \mathbf{H}\mathbf{z}_2+\mathbf{q}_2\geq\mathbf{e}_1,\ \mathbf{q}_2\geq 0\end{aligned}\qquad(9)$$

where $\mathbf{D}_1 = diag\left(\mathbf{D}_{11},\mathbf{D}_{12},...,\mathbf{D}_{1m_1}\right)$, $\mathbf{D}_2 = diag\left(\mathbf{D}_{21},\mathbf{D}_{22},...,\mathbf{D}_{2m_2}\right)$.

Formula (1) is a convex optimization problem with inequality constraints, and it has a close-form solution. Its Lagrange function is built to solve this, shown as follows:

$$L\left(\mathbf{w}_1,b_1,\mathbf{q}_1,\boldsymbol{\alpha},\boldsymbol{\beta}\right)=\frac{1}{2}\left(\mathbf{A}\mathbf{w}_1+\mathbf{e}_1b_1\right)^T\mathbf{D}_1\left(\mathbf{A}\mathbf{w}_1+\mathbf{e}_1b_1\right)$$
$$+c_1\mathbf{e}_2^T\mathbf{q}_1-\boldsymbol{\alpha}^T\left(-\left(\mathbf{B}\mathbf{w}_1+\mathbf{e}_2b_1\right)+\mathbf{q}_1-\mathbf{e}_2\right)-\boldsymbol{\beta}^T\mathbf{q}_1 \quad (10)$$

where $\boldsymbol{\alpha}=\left(\alpha_1,\alpha_2,\alpha_3...\alpha_{m_2}\right)^T$, $\boldsymbol{\beta}=\left(\beta_1,\beta_2,\beta_3...\beta_{m_1}\right)^T$ are Lagrange multipliers. The partial derivatives of $\mathbf{w}_1$, b_1 and $\mathbf{q}_1$ are solved by Lagrange function L separately, setting their derivatives to be zero. The Karush-Kuhn-Tucker (KKT) conditions can be obtained, shown as follows:

$$\frac{\partial L}{\partial \mathbf{w}_1}=\mathbf{A}^T\mathbf{D}_1\left(\mathbf{A}\mathbf{w}_1+\mathbf{e}_1b_1\right)+\mathbf{B}^T\boldsymbol{\alpha}=0 \quad (11)$$

$$\frac{\partial L}{\partial b_1}=\mathbf{e}_1^T\mathbf{D}_1\left(\mathbf{A}\mathbf{w}_1+\mathbf{e}_1b_1\right)+\mathbf{e}_2^T\boldsymbol{\alpha}=0 \quad (12)$$

$$\frac{\partial L}{\partial \mathbf{q}_1}=c_1\mathbf{e}_2-\boldsymbol{\alpha}-\boldsymbol{\beta}=0 \quad (13)$$

$$-\left(\mathbf{B}\mathbf{w}_1+\mathbf{e}_2b_1\right)+\mathbf{q}_1\geq\mathbf{e}_2,\ \mathbf{q}_1\geq 0 \quad (14)$$

$$\boldsymbol{\alpha}^T\left(-\left(\mathbf{B}\mathbf{w}_1+\mathbf{e}_2b_1\right)+\mathbf{q}_1-\mathbf{e}_2\right)=0,\ \boldsymbol{\beta}^T\mathbf{q}_1=0 \quad (15)$$

We can get $0\leq\boldsymbol{\alpha}\leq c_1\mathbf{e}_2$ from Eq. (13) owing to $\boldsymbol{\alpha}\geq 0$, $\boldsymbol{\beta}\geq 0$. Next, Eq. (11) and (12) are combined to be:

$$\left[\mathbf{A}^T\ \ \mathbf{e}_1^T\right]\mathbf{D}_1\left[\mathbf{A}\ \ \mathbf{e}_1\right]\left[\mathbf{w}_1\ \ b_1\right]^T+\left[\mathbf{B}^T\ \ \mathbf{e}_2^T\right]\boldsymbol{\alpha}=0 \quad (16)$$

where augmented vectors $\mathbf{H}=\left[\mathbf{A}\ \ \mathbf{e}_1\right]$, $\mathbf{G}=\left[\mathbf{B}\ \ \mathbf{e}_2\right]$, with these notations, $\mathbf{z}_1^{(p+1)}$ can be obtained as follows.

$$\mathbf{H}^T\mathbf{D}_1\mathbf{H}\mathbf{z}_1^{(p+1)}+\mathbf{G}^T\boldsymbol{\alpha}=0 \quad (17)$$

Eq. (17) is equivalent to Eq. (18).

$$\mathbf{z}_1^{(P+1)}=-\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H}\right)^{-1}\mathbf{G}^T\boldsymbol{\alpha} \quad (18)$$

In Eq. (18), it needs to calculate inverse matrix $\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H}\right)^{-1}$ to obtain $\mathbf{z}_1^{(p+1)}$. $\mathbf{H}^T\mathbf{D}_1\mathbf{H}$ is a positive semi-definite matrix, which may be ill-conditioned in some situations, so we may acquire an inaccurate or unstable solution. In real application, we use the methods described in literatures [10,17]. Regularization term $\varepsilon\mathbf{I}$ is introduced to solve this, $\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H}+\varepsilon\mathbf{I}\right)$ is a positive definite matrix. We can obtain the final solution of $\mathbf{z}_1^{(p+1)}$.

$$\mathbf{z}_1^{(p+1)}=-\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H}+\varepsilon\mathbf{I}\right)^{-1}\mathbf{G}^T\boldsymbol{\alpha} \quad (19)$$

Similar to $\mathbf{z}_2^{(p+1)}$.

$$\mathbf{z}_2^{(p+1)}=\left(\mathbf{G}^T\mathbf{D}_2\mathbf{G}+\varepsilon\mathbf{I}\right)^{-1}\mathbf{H}^T\boldsymbol{\beta} \quad (20)$$

$\mathbf{z}_1^{(p+1)}$ and $\mathbf{z}_2^{(p+1)}$ are brought into Lagrange function (10) separately, Under KKT conditions, the original problems (1) and (2) can be transformed into Wolfe dual problems.

$$\max_{\boldsymbol{\alpha}}\ \mathbf{e}_2^T\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}\left(\mathbf{H}^T\mathbf{D}_1\mathbf{H}\right)^{-1}\mathbf{G}^T\boldsymbol{\alpha}$$
$$s.t.\ \ 0\leq\boldsymbol{\alpha}\leq c_1\mathbf{e}_2 \quad (21)$$

$$\max_{\boldsymbol{\beta}}\ \mathbf{e}_1^T\boldsymbol{\beta}-\frac{1}{2}\boldsymbol{\beta}^T\mathbf{H}\left(\mathbf{G}^T\mathbf{D}_2\mathbf{G}\right)^{-1}\mathbf{H}^T\boldsymbol{\beta}$$
$$s.t.\ \ 0\leq\boldsymbol{\beta}\leq c_2\mathbf{e}_1 \quad (22)$$

We can obtain the Lagrange multipliers $\boldsymbol{\alpha}\in R^{m_2\times 1}$ and $\boldsymbol{\beta}\in R^{m_1\times 1}$ by solving the dual problems. Bring $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ into the Eq. (19) and (20) respectively.

A new point $\mathbf{x}\in \mathrm{R}^n$ is assigned to class 1 or class 2, according to which of the two nonparallel planes in (16) it is closer to the decision function.

$$f\left(\mathbf{x}\right)=\arg\min_{i=1,2}\left(\left|\mathbf{x}^T\mathbf{w}_i+b_i\right|/\left\|\mathbf{w}_i\right\|\right) \quad (23)$$

The iterative procedure of L1-TWSVM is described as follows. In each iteration, $\mathbf{z}_1$ is computed with the current $\mathbf{D}_1$, and then $\mathbf{D}_1$ is updated based on the current calculated $\mathbf{z}_1$. Repeated until the convergence.

Input: Matrices $\mathbf{H}\in R^{m_1\times(n+1)}$ and $\mathbf{G}\in R^{m_2\times(n+1)}$.
Result: $\mathbf{z}_1\in R^{(n+1)\times 1}$.
Set $p=0$, Initialize diagonal matric $\mathbf{D}_1\in R^{m_1\times m_1}$, $\mathbf{D}_1=diag\left(1/\left|\mathbf{H}\mathbf{z}_1\right|\right)$, and initialize $\mathbf{z}_1$, which is a standard solution of TWSVM.
Repeat:
Step 1. Calculate $\mathbf{z}_1^{(P+1)}=-\left(\mathbf{H}^T\mathbf{D}_1^{(p)}\mathbf{H}\right)^{-1}\mathbf{G}^T\boldsymbol{\alpha}$.
Step 2. Update matrix $\mathbf{D}_1^{(p+1)}$ based on the current $\mathbf{z}_1^{(p+1)}$, where the *i*-th element of $\mathbf{D}_1^{(p+1)}$ is $1/\left(\left|\mathbf{h}_i\mathbf{z}_1^{(p+1)^T}\right|\right)$.
Step 3. $p=p+1$.
Step 4. If $\left\|J\left(\mathbf{z}_1^p\right)-J\left(\mathbf{z}_1^{(p+1)}\right)\right\|_1\geq 0.001$ and $p\leq 50$, go on. Otherwise, terminate the run and set $\mathbf{z}_1=\mathbf{z}_1^{(p+1)}$, where $J\left(\mathbf{z}_1\right)$ is the objective function value of formula (14).
Until Convergence
Output: The optimal solution of $\mathbf{z}_1$.

4. Experimental results

To verify the classification performance of L1-TWSVM, it is compared with related algorithms (GEPSVM [10] and TWSVM [17]) on UCI datasets [26]. Experimental environment: Windows10 operating system, an Intel(R) Core(TM) i5-5200u, quad core processor (2.2GHz), 4 GB of RAM. Five various algorithms are implemented in MATLAB 7.1. The parameters are selected by 10-fold cross validation method [27, 28]. Testing accuracy is the average value of results for 10 times. As is known, parameters may influence the accuracy. Thus, to obtain the best generalization performance, all parameters are selected as below. Parameters c_1 and c_2 are in the range of $\{2^i \,|\, i = -7, -6, -5..., 7\}$, while parameter ε is in the range of $\{10^i \,|\, i = -10, -9, -8..., 10\}$.

4.1. Experiments on artificial datasets

To examine the performance of L1-TWSVM, we did the experiment on XOR datasets (called Cross-plane (60×2)). We know that outliers tend to have a certain influence on the classification performance. Here, two extra outliers are added on the Cross-plane datasets (Cross-plane 1) to assess the robustness of TWSVM and L1-TWSVM. As shown in Fig.1, the results of two classifiers on Cross-plane 1 datasets are given in Fig.2, respectively.

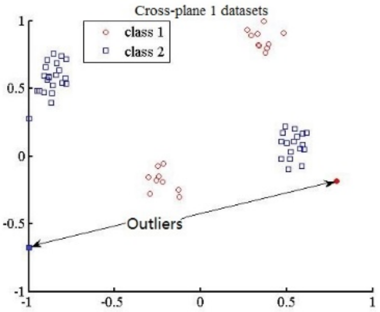
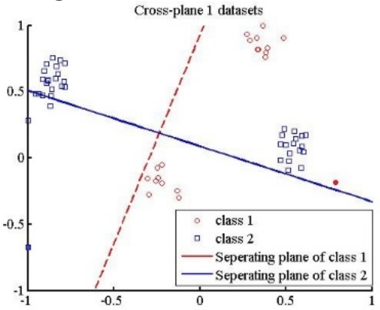
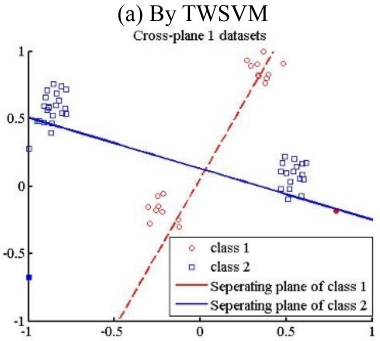


Fig. 1. XOR datasets with outliers



(a) By TWSVM



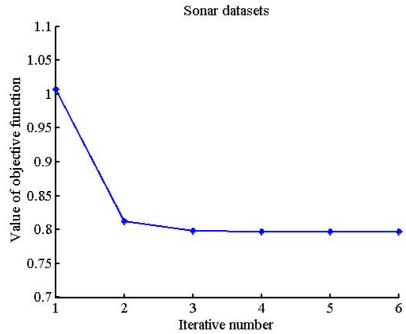
(b) By L1-TWSVM

Fig. 2. The classification results on Cross-plane 1 datasets

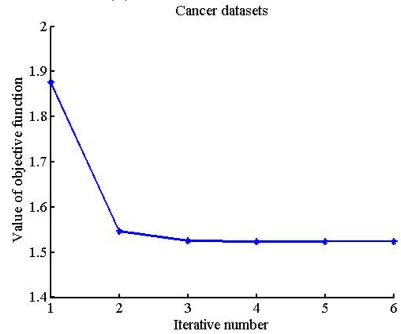
The accuracy of TWSVM and L1-TWSVM are 73.08%, 77.56% respectively. We find that L1-TWSVM achieves the highest classification accuracy after introducing outliers in Fig. 2. This may be attributed to the L1-norm distance embedded in TWSVM. It can be further explained that L2-norm distance may result in the large distance dominating the sum in TWSVM when the outliers appear in datasets, which easily leads to the biased results. This effectively validate the practicability of L1-TWSVM.

4.2. Experiments on UCI datasets

Fig.3 shows the objective function values of L1-TWSVM monotonically decrease with the iteration until it converges to a fixed value, L1-TWSVM can fast converge within about 6 iterations. Horizontal axis represents the number of iterations, and vertical axis represents the value of objective function.



(a) On Sonar datasets



(b) On Cancer datasets

Fig. 3. The objective function values of L1-TWSVM monotonically decrease along with the iterative number

Table 1 is the comparison of accuracy within three algorithms, while Table 2 shows the results on six datasets where 20% Gaussian noise was introduced respectively. Besides, the average training iterative numbers are listed in the tables corresponding to each experiment. The P-Values are from paired t-tests comparing each algorithm to L1-TWSVM. An asterisk (*) denotes the significant difference from L1-TWSVM based on P-Values less than 0.05. The highest accuracy is marked black bold.

We performed paired t-tests comparing L1-TWSVM to related algorithms. In this study, we set the threshold for p-values to be 0.05. For instance, in Table 1, the P-Value of the test comparing L1-TWSVM and TWSVM on Ticdata datasets is 0.00022, and on Monks3 datasets is 0.0097, which are less than 0.05, we conclude that L1-TWSVM and TWSVM have different accuracies on these datasets and L1-TWSVM is significantly better than TWSVM. Similar results can be seen between L1-TWSVM and GEPSVM.

Table 1 Experimental results of three algorithms

Datasets (N×n)	GEPSVM Test±Std (%) Time(s) P-Value	TWSVM Test±Std (%) Time(s) P-Value	L1-TWSVM Test±Std (%) Time(s) Iterative number
Heart	72.59±8.15*	**82.96±4.74**	82.59±5.98
(270×13)	0.0064	0.0209	0.1343
	0.0108	0.5924	5
Monks3	78.51±4.20*	81.62±6.93*	**87.91±3.24**
(554×6)	0.0062	0.0675	0.3642
	3.8023e-004	0.0097	5
Cancer	95.76±2.66	96.79±1.92	96.63±1.86
(683×9)	0.0059	0.0715	0.4533
	0.4171	0.8528	6
Ticdata	65.87±3.18*	65.97±3.68*	**71.30±3.42**
(958×9)	0.0061	0.5432	10.8275
	0.0106	2.1681e-004	13
Pidd	74.74±4.06*	76.30±3.15*	**77.35±2.01**
(768×8)	0.0060	0.0959	0.4255
	0.0311	0.0491	5
Sonar	74.12±8.13	74.02±7.48	71.71±8.44
(208×60)	0.0096	0.0217	0.1033
	0.5002	0.5015	6

Table 2 Introduce 20% Gaussian noise, results of three algorithms

Datasets (N×n)	GEPSVM Test±Std (%) Time(s) P-Value	TWSVM Test±Std (%) Time(s) P-Value	L1-TWSVM Test±Std (%) Time(s) Iterative number
Heart	66.30±7.85*	75.93±6.25	**79.55±6.63**
(270×13)	0.0066	0.0373	0.1053
	0.0073	0.6274	7
Monks3	78.52±3.94*	81.08±7.19*	**87.02±4.22**
(554×6)	0.0061	0.0961	0.4715
	0.0015	0.0214	5
Cancer	95.61±2.78	**97.37±1.42**	96.93±1.90
(683×9)	0.0061	0.0881	0.5112
	0.2054	0.3419	6
Ticdata	65.66±2.85*	66.29±4.13*	**71.41±3.63**
(958×9)	0.0065	0.5911	10.0191
	0.0075	0.0031	13
Pidd	74.74±3.81*	73.95±3.64*	**77.08±3.66**
(768×8)	0.0063	0.0296	0.1222
	0.0483	0.0417	5
Sonar	**76.02±7.89**	71.57±10.37	68.17±13.50
(208×60)	0.0096	0.0325	0.1483
	0.2259	0.2228	6

From the data in Table 1, we find the accuracy of L1-TWSVM is comparable to that of other competing algorithms, and even higher than that of others in some scenario. This indicates the performance of L1-TWSVM is better. Besides, from the columns of L1-TWSVM in the Table 1 above, it can be found that L1-TWSVM can fast converge within about 6 iterations.

According to the tables in results, the performance degradations of our method is very small when 20% Gaussian noise is introduced, even the accuracy is improved by the proposed algorithm compared with TWSVM. In addition, the accuracy of L1-TWSVM has a little change compared to other methods, especially when Gaussian noise is introduced. This may be attributed to the embedding of L1-norm distance, which makes L1-TWSVM more robust to outliers than others. This further demonstrates L1-norm distance is useful for data classification, especially for the samples with outliers.

It is worth noting that the training time of L1-TWSVM is the highest. L1-TWSVM requires to calculate two QPPs and the iterative algorithm we develop needs to iteratively compute the optimal solutions, which takes a long time to do these, but surpasses other methods in accuracy.

5. Conclusions

An efficient and robust TWSVM classifier based on L1-norm distance for binary classification is proposed in this paper, called L1-TWSVM, which makes full use of the robustness of L1-norm distance to noises and outliers. Specially, we design a simple and valid iterative algorithm to solve the L1-norm optimal problems, which is easy to implement. Experimental results indicate that L1-TWSVM can effective ease the impact of outliers, and improves the generalization ability and flexibility of the model. But L1-TWSVM still needs to solve the QPPs, to obtain the optimal solutions must iteratively compute the QPPs. According to the experimental results above, the computational cost of L1-TWSVM is the highest compared with other relative algorithms under the same scenario. This is difficult to effectively handle large data samples. In summary, L1-TWSVM has better performance and robustness than other algorithms, especially when Gaussian noise is introduced.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61373062, 61772273), the Natural Science Foundation of Jiangsu (No. BK20141403), the Fundamental Research Funds for the Central Universities (No.30916011327), and the National Key Research and Development Program: Key Projects of International Scientific and Technological Innovation Cooperation between Governments (No. S2016G9070).

References

[1] X.J. Liu, S.C. Chen, H.J. Peng, Computer keystroker verification based on support vector machines, Journal of Computer Research & Development, 39 (2002) 1082-1086.

[2] S.F. Tian, H.K. Huang, Database learning algorithms based on support vector machine, Journal of Computer Research & Development, 37 (2000) 17-22.
[3] V. Vapnik, The nature of statistical learning theory, 1995), pp. 988 - 999.
[4] P.S. Bradley, O.L. Mangasarian, Massive data discrimination via linear support vector machines, Optimization Methods & Software, 13 (2000) 1-10.
[5] C. Cortes, V. Vapnik, Support Vector Network, 20 (1995) 273-297.
[6] Q. Song, W. Hu, W. Xie, Robust support vector machine with bullet hole image classification, IEEE Transactions on Systems Man & Cybernetics Part C, 32 (2002) 440-448.
[7] H. Yin, X. Jiao, Y. Chai, B. Fang, Scene classification based on single-layer SAE and SVM, Expert Systems with Applications, 42 (2015) 3368–3380.
[8] V. Muralidharan, V. Sugumaran, V. Indira, Fault diagnosis of monoblock centrifugal pump using SVM, Engineering Science & Technology An International Journal, 17 (2014) 152-157.
[9] A. Subasi, Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders, Computers in Biology & Medicine, 43 (2013) 576-586.
[10] O.L. Mangasarian, E.W. Wild, Multisurface proximal support vector machine classification via generalized eigenvalues, IEEE Transactions on Pattern Analysis & Machine Intelligence, 28 (2006) 69-74.
[11] N. Deng, Y. Tian, C. Zhang, Support vector machines. Optimization based theory, algorithms, and extensions, Crc Press, (2012).
[12] [12] C.C. Chang, C.J. Lin, LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems & Technology, 2 (2011) 389-396.
[13] Y.H. Shao, W.J. Chen, N.Y. Deng, Nonparallel hyperplane support vector machine for binary classification problems, Information Sciences, 263 (2014) 22-35.
[14] Y.H. Shao, N.Y. Deng, W.J. Chen, Z. Wang, Improved generalized eigenvalue proximal support vector machine, IEEE Signal Processing Letters, 20 (2013) 213-216.
[15] Q. Ye, N. Ye, Improved Proximal Support Vector Machine via Generalized Eigenvalues, International Joint Conference on Computational Sciences and Optimization, Cso 2009, Sanya, Hainan, China, 24-26 April2009), pp. 705-709.
[16] M.R. Guarracino, C. Cifarelli, O.S. Pardalos, P. M., A classification method based on generalized eigenvalue problems, Optimization Methods & Software, 22 (2007) 73-81.
[17] Jayadeva, R. Khemchandani, S. Chandra, Twin support vector machines for pattern classification, Pattern Analysis & Machine Intelligence IEEE Transactions on, 29 (2007) 905-910.
[18] Q. Ye, C. Zhao, Xiaobo, Chen, A feature selection method for twsvm via a regularization technique, Journal of Computer Research & Development, 48 (2011) 1029-1037.
[19] M.A. Kumar, M. Gopal, Least squares twin support vector machines for pattern classification, Expert Systems with Applications, 36 (2009) 7535-7543.
[20] X. Zhang, L. Fan, Application of smoothing technique on projective tsvm, International Journal of Applied Mathematics and Machine Learning, 2 (2015) 27-45.
[21] X. Chen, J. Yang, Q. Ye, J. Liang, Recursive projection twin support vector machine via within-class variance minimization, Pattern Recognition, 44 (2011) 2643-2655.
[22] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis & Machine Intelligence, 30 (2008) 1672-1680.
[23] C.N. Li, Y.H. Shao, N.Y. Deng, Robust L1-norm non-parallel proximal support vector machine, Optimization, 65 (2015) 1-15.
[24] C. Ding, D. Zhou, X. He, H. Zha, R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization, International Conference on Machine Learning2006), pp. 281-288.
[25] J. Gao, Robust L1 principal component analysis and its Bayesian variational inference, Neural Computation, 20 (2008) 555.
[26] K. Bache, M. Lichman, UCI machine learning repository, (2013).
[27] Q. Ye, C. Zhao, S. Gao, H. Zheng, Weighted twin support vector machines with local information and its application, Neural Networks the Official Journal of the International Neural Network Society, 35 (2012) 31-39.
[28] S. Ding, X. Hua, J. Yu, An overview on nonparallel hyperplane support vector machine algorithms, Neural Computing & Applications, 25 (2013) 975-982.