

Review on: Twin Support Vector Machines

Yingjie Tian · Zhiquan Qi

Received: 5 June 2014 / Revised: 20 July 2014 / Accepted: 30 August 2014 /
Published online: 8 October 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract Twin support vector machine (TWSVM), an useful extension of the traditional SVM, becomes the current researching hot spot in machine learning during the last few years. For the binary classification problem, the basic idea of TWSVM is to seek two nonparallel proximal hyperplanes such that each hyperplane is closer to one of the two classes and is at least one distance from the other. TWSVM has lower computational complexity and better generalization ability, therefore in the last few years it has been studied extensively and developed rapidly. Considering the many variants of TWSVM, a systematic survey is needed and helpful to understand and use this family of data mining techniques more easily. The purpose of this paper is to closely review TWSVMs and provide an insightful understanding of current developments, at the same time point out their limitations and highlight the major opportunities and challenges, as well as potential important research directions.

Keywords Twin support vector machine · Support vector machine · Machine learning · Data mining · Optimization · Nonparallel · Structural risk minimization principle

1 Introduction

Support vector machines (SVMs) [1–7] using the well established concepts in statistical learning theory (SLT) and optimization theory, maybe the most widely used

Y. Tian (✉) · Z. Qi
Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences,
Beijing 100190, China
e-mail: tyj@ucas.ac.cn

Z. Qi
e-mail: qizhiquan@ucas.ac.cn

techniques for data mining, and have been successfully applied in many fields ranging from text categorization, face verification, speech recognition, information and image retrieval, to remote sensing image analysis, time series forecasting, information security, bankruptcy prediction and etc. [8–20]. For the standard SVM, the principal of maximizing the margin between two parallel support hyperplanes leads to solving a convex quadratic programming problem (QPP), dual theory makes introducing the kernel function possible, then the kernel trick is applied to solve nonlinear cases.

Recently, different with the standard SVM searching for two parallel hyperplanes, some nonparallel hyperplane classifiers have been proposed, such as the generalized eigenvalue proximal support vector machine (GEPSSVM) [21] and the twin support vector machine (TWSVM) [22], which have become one of the researching hot spots in the field of matching learning. For the TWSVM, it seeks two nonparallel proximal hyperplanes such that each hyperplane is closer to one of the two classes and is at least one distance from the other. This strategy results the fundamental difference between TWSVM and the standard SVM, i.e., the former solves two smaller sized QPPs, whereas the latter solves a larger one, which increases the TWSVM training speed by approximately fourfold compared to that of SVM. Furthermore, experimental results in [22] have shown the effectiveness of TWSVM over standard SVM and even GEPSSVM on UCI datasets. Therefore, in the last few years, TWSVM has been studied extensively and developed rapidly. Considering the many variants of TWSVM, a systematic survey is needed and helpful to understand and use this family of data mining techniques more easily although there are several limited reviews about TWSVMs [23, 24]. The goal of this paper is to closely review TWSVMs and provide an insightful understanding of current developments, at the same time point out their limitations and highlight the major opportunities and challenges, as well as potential important research directions.

Section 2 of the paper reviews four representative TWSVMs: original TWSVM, twin bounded support vector machine (TBSVM) [25], improved TWSVM (ITSVM) [26] and nonparallel SVM (NPSVM) [27]. Section 3 describes the variants and extensions of TWSVMs for different learning problems. Section 4 introduces several applications of TWSVMs. Finally, concluding remarks and future research directions are provided in Sect. 5.

2 TWSVM, TBSVM, ITSVM and NPSVM

In this section, we review four representative nonparallel hyperplane classifiers, on which almost all the improved models are based. Here we consider their originally proposed models for the binary classification problem with the training set

$$T = \{(x_1, +1), \dots, (x_p, +1), (x_{p+1}, -1), \dots, (x_{p+q}, -1)\}, \quad (1)$$

where $x_i \in R^n$, $i = 1, \dots, p + q$. Let $l = p + q$ and

$$A = (x_1, \dots, x_p)^\top \in R^{p \times n}, \quad B = (x_{p+1}, \dots, x_{p+q})^\top \in R^{q \times n}. \quad (2)$$

2.1 TWSVM

Unlike the standard SVM solving one QPP and similar to the GEPSVM [21,28], TWSVM [22] constructs two smaller QPPs

$$\begin{aligned} \min_{w_+, b_+, \xi_-} & \quad \frac{1}{2}(Aw_+ + e_+b_+)^\top (Aw_+ + e_+b_+) + c_1e_-^\top \xi_-, \\ \text{s.t.} & \quad -(Bw_+ + e_-b_+) + \xi_- \geq e_-, \quad \xi_- \geq 0, \end{aligned} \tag{3}$$

and

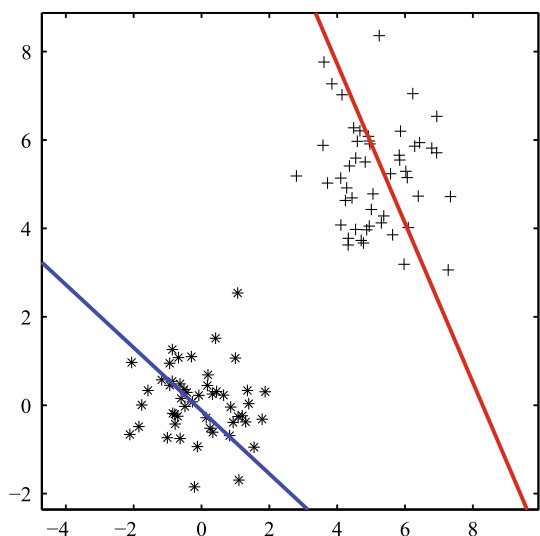
$$\begin{aligned} \min_{w_-, b_-, \xi_+} & \quad \frac{1}{2}(Bw_- + e_-b_-)^\top (Bw_- + e_-b_-) + c_2e_+^\top \xi_+, \\ \text{s.t.} & \quad (Aw_- + e_+b_-) + \xi_+ \geq e_+, \quad \xi_+ \geq 0, \end{aligned} \tag{4}$$

to seek a pair of nonparallel hyperplanes

$$(w_+ \cdot x) + b_+ = 0 \text{ and } (w_- \cdot x) + b_- = 0, \tag{5}$$

such that the positive hyperplane $(w_+ \cdot x) + b_+ = 0$ is proximal to the positive class (measured by the quadratic loss $\frac{1}{2}(Aw_+ + e_+b_+)^\top (Aw_+ + e_+b_+)$) and far from the negative class (measured by the hinge loss $\xi_- = \max\{0, e_- + (Bw_+ + e_-b_+)\}$), and vice visa for the negative hyperplane $(w_- \cdot x) + b_- = 0$, see Fig. 1. $c_i, i = 1, 2$ are the penalty parameters and e_+ and e_- are vectors of ones of appropriate dimensions.

Fig. 1 Geometrical illustration of TWSVM in R^2 : The “+” (class 1) and “*” (class 2) points are generated following two normal distributions respectively, two nonparallel lines (red and blue lines) are obtained from linear TWSVM [26]. (Color figure online)



In order to get the solutions of the above QPPs, their dual problems

$$\begin{aligned} \max_{\alpha} \quad & e_{-}^{\top} \alpha - \frac{1}{2} \alpha^{\top} G (H^{\top} H)^{-1} G^{\top} \alpha, \\ \text{s.t.} \quad & 0 \leq \alpha \leq c_1 e_{-}, \end{aligned} \tag{6}$$

and

$$\begin{aligned} \max_{\gamma} \quad & e_{+}^{\top} \gamma - \frac{1}{2} \gamma^{\top} H (G^{\top} G)^{-1} H^{\top} \gamma, \\ \text{s.t.} \quad & 0 \leq \gamma \leq c_2 e_{+}, \end{aligned} \tag{7}$$

are solved respectively, where

$$H = [A \ e_{+}] \in R^{p \times (n+1)}, \quad G = [B \ e_{-}] \in R^{q \times (n+1)}, \tag{8}$$

therefore the solutions of problems (3) and (4) can be obtained by

$$(w_{+}^{\top}, b_{+})^{\top} = -(H^{\top} H)^{-1} G^{\top} \alpha, \tag{9}$$

$$(w_{-}^{\top}, b_{-})^{\top} = -(G^{\top} G)^{-1} H^{\top} \gamma. \tag{10}$$

Thus an unknown point $x \in R^n$ is predicted to the *Class* by

$$Class = \arg \min_{k=-,+} |(w_k \cdot x) + b_k|, \tag{11}$$

where $|\cdot|$ is the perpendicular distance of point x from the planes $(w_k \cdot x) + b_k = 0, k = -, +$.

TWSVM has several advantages such as

- (1) The dual problems (6) and (7) has q and p variables respectively as opposed to $l = p + q$ in the standard SVM. This strategy of solving a pair of smaller sized QPPs instead of a large one makes the learning speed of TWSVM be approximately four times faster;
- (2) By using the quadratic loss function, TWSVM fully considers the prior information within classes in data and is less sensitive to the noise;
- (3) TWSVM is useful for automatically discovering two-dimensional projections of the data;

However, it still has the following drawbacks:

- (I) In the primal problems (3) and (4) of TWSVM, only the empirical risk is minimized, whereas the structural risk is minimized in the standard SVM;
- (II) The inverse matrices of $H^{\top} H$ and $G^{\top} G$ are approximately replaced by $H^{\top} H + \epsilon I$ and $G^{\top} G + \epsilon I$ respectively in order to deal with the singular case and avoid the possible ill conditioning, where I is an identity matrix of appropriate dimensions, ϵ is a small positive scalar to keep the structure of data. This replacements means TWSVM only get the approximate solutions;

- (III) Though TWSVM solves two smaller sized QPPs, it needs to compute the inverse matrices before training the models. So the computational complexity of TWSVM should include two parts: the complexity of computing the inverse matrices and the complexity of solving the dual problems. For a large data set, TWSVM will fail since it is in practice intractable or even impossible to compute the inverse matrices by the classical methods, though Sherman-Morrison-Woodbury [29] formula or the rectangular kernel technique [30] were used in [22];
- (IV) The above TWSVM is only for the linear classification problem since it can not be extended to the nonlinear case directly as standard SVMs usually do. For the nonlinear case [22], two kernel-generated surfaces

$$K(x^\top, C^\top)u_+ + b_+ = 0 \text{ and } K(x^\top, C^\top)u_- + b_- = 0, \tag{12}$$

instead of hyperplanes (5) were considered, where

$$C^\top = [A \ B]^\top \in R^{n \times l}, \tag{13}$$

and K is an appropriately chosen kernel. Two primal problems different with (3) and (4) were constructed and their corresponding dual problems were solved. Similar with the linear case, the nonlinear TWSVM still has the drawbacks (i) (ii) and (iii) and the folloing (V) and (VI) . Furthermore, the nonlinear TWSVM with the linear kernel is not equivalent to the linear TWSVM [26], which is also different with the standard SVM;

- (V) TWSVM needs fast solvers such as the sequential minimization optimization (SMO) [31] algorithm for the standard SVMs;
- (VI) TWSVM loses the sparseness by using the quadratic loss function for each class to make the proximal hyperplane close enough to the class itself, which results that almost all the points in this class contribute to each final decision function.

2.2 TBSVM

The twin bounded support vector machines (TBSVM) [25] was proposed to overcome the drawbacks (I) and (II). For the linear case, two primal problems constructed in TBSVM are

$$\begin{aligned} \min_{w_+, b_+, \xi_-} \quad & \frac{1}{2}c_3(\|w_+\|^2 + b_+^2) + \frac{1}{2}(Aw_+ + e_+b_+)^\top(Aw_+ + e_+b_+) + c_1e_+^\top\xi_-, \\ \text{s.t.} \quad & -(Bw_+ + e_-b_+) + \xi_- \geq e_-, \quad \xi_- \geq 0, \end{aligned} \tag{14}$$

and

$$\begin{aligned} \min_{w_-, b_-, \xi_+} \quad & \frac{1}{2}c_4(\|w_-\|^2 + b_-^2) + \frac{1}{2}(Bw_- + e_-b_-)^\top(Bw_- + e_-b_-) + c_2e_+^\top\xi_+, \\ \text{s.t.} \quad & (Aw_- + e_+b_-) + \xi_+ \geq e_+, \quad \xi_+ \geq 0, \end{aligned} \tag{15}$$

where $c_i, i = 1, 2, 3, 4$ are the penalty parameters and e_+ and e_- are vectors of ones of appropriate dimensions. Compared with the primal problems (3) and (4) of TWSVM, the regularization terms $\frac{1}{2}c_3(\|w_+^2\| + b_+^2)$ and $\frac{1}{2}c_4(\|w_-^2\| + b_-^2)$ are added to minimize that the structural risk, and this modification leads to two dual problems

$$\begin{aligned} \max_{\alpha} \quad & e_-^\top \alpha - \frac{1}{2} \alpha^\top G(H^\top H + c_3 I)^{-1} G^\top \alpha, \\ \text{s.t.} \quad & 0 \leq \alpha \leq c_1 e_-, \end{aligned} \tag{16}$$

and

$$\begin{aligned} \max_{\gamma} \quad & e_+^\top \gamma - \frac{1}{2} \gamma^\top H(G^\top G + c_4 I)^{-1} H^\top \gamma, \\ \text{s.t.} \quad & 0 \leq \gamma \leq c_2 e_+. \end{aligned} \tag{17}$$

with nonsingular matrices $(H^\top H + c_3 I)$ and $(G^\top G + c_4 I)$. Different with the fixed small scalar ϵ in TWSVM, the parameter c_3 or c_4 , is the weighting factor which determines the tradeoff between the regularization term and the empirical risk. For the nonlinear case, two other regularization terms $(\|u_+\|^2 + b_+^2)$ and $(\|u_-\|^2 + b_-^2)$ are introduced to implement the structural risk minimization principle. However, TBSVM still has the drawbacks (III)~(VI). Though it claimed that the successive overrelaxation (SOR) [32] technique was used to solve the dual problems to speed up the training procedure, and seems to be able to deal with the large scale problems, it in fact took no account of the computation of the inverse matrices. Its similar model [33], a coordinate descent margin based TWSVM (CDMTSVM), was presented and a coordinate descent method was proposed for fast training which handles one data point at a time, but the inverse matrices still need to be computed at first.

2.3 ITSVM

The improved twin support vector machine (ITSVM) [26] was proposed to overcome the drawbacks (I)~(V). By introducing the different Lagrangian functions for the primal problems (14) and (15) in the TBSVM, we get the improved dual formulations

$$\begin{aligned} \max_{\lambda, \alpha} \quad & -\frac{1}{2}(\lambda^\top \alpha^\top) \hat{Q}(\lambda^\top \alpha^\top)^\top + c_3 e_-^\top \alpha, \\ \text{s.t.} \quad & 0 \leq \alpha \leq c_1 e_-, \end{aligned} \tag{18}$$

and

$$\begin{aligned} \max_{\theta, \gamma} \quad & -\frac{1}{2}(\theta^\top \gamma^\top) \tilde{Q}(\theta^\top \gamma^\top)^\top + c_4 e_+^\top \gamma, \\ \text{s.t.} \quad & 0 \leq \gamma \leq c_2 e_+, \end{aligned} \tag{19}$$

respectively, where

$$\hat{Q} = \begin{pmatrix} AA^\top + c_3 \hat{I} & AB^\top \\ AB^\top & BB^\top \end{pmatrix} + E, \tilde{Q} = \begin{pmatrix} BB^\top + c_4 \tilde{I} & BA^\top \\ BA^\top & AA^\top \end{pmatrix} + E, \tag{20}$$

and \hat{I} is the $p \times p$ identity matrix, \tilde{I} is the $q \times q$ identity matrix, E is the $l \times l$ matrix with all entries equal to one. The pair of nonparallel hyperplanes (5) are then obtained from the solutions (λ^*, α^*) and (θ^*, γ^*) of (18) and (19)

$$w_+^* = -(A^\top \lambda^* + B^\top \alpha^*), b_+^* = -(e_+^\top \lambda^* + e_-^\top \alpha^*), \tag{21}$$

and

$$w_-^* = -(B^\top \theta^* + A^\top \gamma^*), b_-^* = -(e_-^\top \theta^* + e_+^\top \gamma^*). \tag{22}$$

ITSVM overcomes several drawbacks listed above of the TWSVM. (1) ITSVM implements the structural risk minimization principle since its primal problems are the same with those of TBSVM; (2) ITSVM has no inverse matrices to be computed before training since it only needs to solve the QPPs (18) and (19); (3) Linear ITSVM can be easily extended to the nonlinear case since the dual problems (18) and (19) can be applied with the kernel functions just as the standard SVMs do, i.e., take the $K(A, B^\top)$, $K(A, A^\top)$, $K(B, B^\top)$, $K(B, A^\top)$ instead of AB^\top , AA^\top , BB^\top , BA^\top in \hat{Q} and \tilde{Q} ; (4) SOR technique is also used to solve the dual problems (18) and (19). Here we should point out that the computational complexity of SOR for ITSVM is different with that for TBSVM. Since the dual problems (18) and (19) have $l = p + q$ variables respectively, ITSVM has the same scale with the standard SVM and is almost 2 times larger than TWSVM or TBSVM, which means that ITSVM sacrifices more model training time to skillfully avoids the computation of the inverse matrix. At a first glance, it seems that ITSVM has more computational complexity than TWSVM or TBSVM, in fact when the complexity of computing the inverse matrices are considered in TWSVM or TBSVM, ITSVM is faster since

$$O(p^3) + \#iteration \times O(p) > \#iteration \times O(l), \tag{23}$$

where $O(p^3)$ is the complexity of computing $p \times p$ inverse matrix, and $\#iteration \times O(p)$ is of SOR for p sized problem, $\#iteration \times O(l)$ is of SOR for l sized problem ($\#iteration$ is the number of the iterations, experiments in [32] has shown that $\#iteration$ is almost linear scaling with the problem size) [26].

2.4 NPSVM

The nonparallel support vector machine (NPSVM) [27] was proposed to further overcome the sparseness drawback (VI). By taking the ε -insensitive loss function instead of the quadratic loss function in TWSVM, TBSVM or ITSVM, NPSVM constructs two primal problems

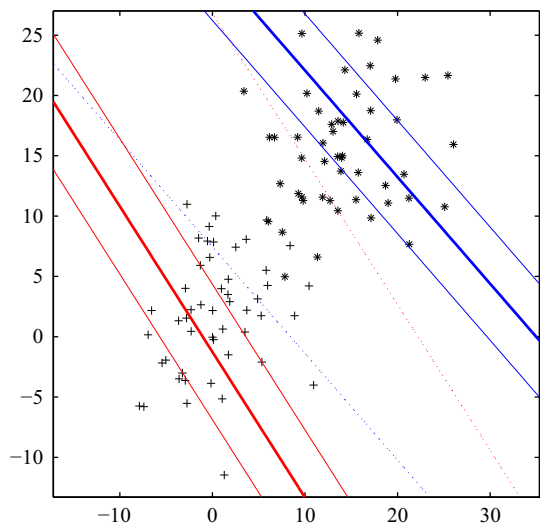
$$\begin{aligned}
 & \min_{w_+, b_+, \eta_+^{(*)}, \xi_-} \quad \frac{1}{2} \|w_+\|^2 + c_1 \sum_{i=1}^p (\eta_i + \eta_i^*) + c_2 \sum_{j=p+1}^{p+q} \xi_j, \\
 & \text{s.t.} \quad (w_+ \cdot x_i) + b_+ \leq \varepsilon + \eta_i, \quad i = 1, \dots, p, \\
 & \quad \quad - (w_+ \cdot x_i) - b_+ \leq \varepsilon + \eta_i^*, \quad i = 1, \dots, p, \\
 & \quad \quad (w_+ \cdot x_j) + b_+ \leq -1 + \xi_j, \quad j = p + 1, \dots, p + q, \\
 & \quad \quad \eta_i, \eta_i^* \geq 0, \quad i = 1, \dots, p, \\
 & \quad \quad \xi_j \geq 0, \quad j = p + 1, \dots, p + q,
 \end{aligned} \tag{24}$$

and

$$\begin{aligned}
 & \min_{w_-, b_-, \eta_-^{(*)}, \xi_+} \quad \frac{1}{2} \|w_-\|^2 + c_3 \sum_{i=p+1}^{p+q} (\eta_i + \eta_i^*) + c_4 \sum_{j=1}^p \xi_j, \\
 & \text{s.t.} \quad (w_- \cdot x_i) + b_- \leq \varepsilon + \eta_i, \quad i = p + 1, \dots, p + q, \\
 & \quad \quad - (w_- \cdot x_i) - b_- \leq \varepsilon + \eta_i^*, \quad i = p + 1, \dots, p + q, \\
 & \quad \quad (w_- \cdot x_j) + b_- \geq 1 - \xi_j, \quad j = 1, \dots, p, \\
 & \quad \quad \eta_i, \eta_i^* \geq 0, \quad i = p + 1, \dots, p + q, \\
 & \quad \quad \xi_j \geq 0, \quad j = 1, \dots, p,
 \end{aligned} \tag{25}$$

where $\varepsilon \geq 0$ is the sparseness parameter, $c_i \geq 0, i = 1, \dots, 4$ are penalty parameters, $\xi_+ = (\xi_1, \dots, \xi_p)^\top$, $\xi_- = (\xi_{p+1}, \dots, \xi_{p+q})^\top$, $\eta_+^{(*)} = (\eta_+^\top, \eta_+^{*\top})^\top = (\eta_1, \dots, \eta_p, \eta_1^*, \dots, \eta_p^*)^\top$, $\eta_-^{(*)} = (\eta_-^\top, \eta_-^{*\top})^\top = (\eta_{p+1}, \dots, \eta_{p+q}, \eta_{p+1}^*, \dots, \eta_{p+q}^*)^\top$, are slack variables. Figure 2 illustrates its geometrical explanation in R^2 , where class 1 locates as much as possible in the ε -band of the positive hyperplane $(w_+ \cdot x) + b_+ = 0$

Fig. 2 Geometrical illustration of NPSVM in R^2 : The “+” (class 1) and “*” (class 2) points are generated following two normal distributions respectively, two nonparallel lines (red and blue bold lines) are obtained from linear NPSVM [27]. (Color figure online)



(measured by the ε -insensitive loss function) and far from class 2 (measured by the hinge loss $\xi_- = \max\{0, e_- + (Bw_+ + e_-b_+)\}$), and vice visa for the negative hyperplane $(w_- \cdot x) + b_- = 0$.

The corresponding dual problems has the similar formulation with that of standard SVM but almost 1.5 times larger

$$\begin{aligned} \min_{\pi} \quad & \frac{1}{2} \pi^\top \Lambda \pi + \kappa^\top \pi, \\ \text{s.t.} \quad & e^\top \pi = 0, \\ & 0 \leq \pi \leq C, \end{aligned} \tag{26}$$

where $\Lambda \in R^{1.5l \times 1.5l}$ is semi-definite positive, therefore (26) can be solved efficiently by the SMO-type technique. NPSVM also implements the structural risk minimization principle, has no inverse matrices to be computed, introduces the kernels directly to deal with the nonlinear case, and most importantly has the sparseness because of the ε -insensitive loss function and hinge loss function adopted simultaneously in each primal problem. NPSVM overcomes all the listed drawbacks of TWSVM and in some sense is a true nonparallel classifier with solid theoretical foundation.

3 Variants and Extensions of TWSVMs

TWSVMs have attracted many interests in recent years and many improved algorithms were proposed. In this section, we will review them based on the problems they solved.

3.1 Variants of Binary TWSVMs

3.1.1 ν -TWSVMs

ν -NPSVM [34] is an equivalent formulation of NPSVM inheriting all the advantages of NPSVM model, but has more excellent properties. It is parameterized by the quantity ν to let ones effectively control the number of support vectors, since the value of ν is the lower bound of the percentage of support vectors. Furthermore, for each class, different sparseness can be obtained by using different parameter ν , which enable us to solve unbalanced classification problems. A ν -TWSVM based on the original TWSVM was proposed in [35], which was interpreted as a pair of minimum generalized Mahalanobis-norm problems on two reduced convex hulls, and an improved geometric algorithm (GA) was developed to improve its efficiency. A rough margin-based ν -TWSVM incorporating the rough set theory [36] was proposed in [37] to give the different penalties to the misclassified points. The twin parametric-margin SVM (TPMSVM) [38], motivated by the TWSVM and the par- ν -SVM [39], determines a pair of parametric-margin hyperplanes, which can automatically adjust a flexible margin and are suitable for the heteroscedastic error structure. Its least squares version can be found in [40]. [41] smoothed the TPMSVM by introducing a quadratic function and suggested a genetic algorithm GA-based model selection for TPMSVM.

3.1.2 Least Squares TWSVMs

A least squares version of TWSVM (LSTWSVM) [42] was proposed in line with the least squares SVM (LSSVM) [43]. The solution of the two modified primal problems reduces to solving just two systems of linear equations and leads to an extremely fast and efficient performance. Weighted LSTWSVM [44] put different weights on the error variables in order to eliminate the impact of noise data and obtain the robust estimation, and also for the imbalanced problem [45]. 1-norm LSTWSVM [46] was designed for automatically selecting the relevant features, by replacing all the 2-norm terms in the regularized LSTWSVM [47] with 1-norm ones and then converting its formulation to a linear programming (LP) problem. A least squares version of twin support hypersphere (TSVH) [48, 49] can be referred to [50], where TSVH aims to find a pair of hyperspheres not hyperplanes by solving two smaller sized QPPs, each QPP is based on the support vector data description (SVDD) [51]. A multiple-surface classification (MSC) algorithm, named projection twin support vector machine (PTSVM) [52–54], seeks projection directions such that the projected samples of one class are well separated from those of the other class in its own subspace, and a recursive algorithm for PTSVM was proposed to further boost the performance. Its least squares version, LSPTSVM, was formulated in [55], where a regularization term was added to ensure the optimization problems are positive definite and for better generalization ability. A nonlinear LSPTSVM for binary nonlinear classification by introducing nonlinear kernel into LSPTSVM was proposed in [56]. FLSPTSVM (feature selection for LSPTSVM) [57] can perform the feature selection and reduce the number of kernel functions required for the classifier. Nonparallel plane proximal classifier (NPPC) [58] was very similar to the LSTWSVM by introducing a technique as used in the proximal support vector machine (PSVM) [30] classifier, which leads to solving two small systems of linear equations in input space.

3.1.3 Localized TWSVMs

Weighted TWSVM with local information (WLTSVM) [59] explored the similarity information between pairs of samples by finding the k -nearest neighbors for all the samples. Ye et al. [60] proposed a reduced algorithm termed localized TWSVM via convex minimization (LCTSVM), which effectively reduced the space complexity of TWSVM by constructing two nearest neighbor graphs in the input space. Wang et al. [61] improved LCTSVM to be a fast LCTSVM suitable for large data sets, in which the number of TWSVMs was decreased. Local and Global Regularized Twin SVM (TWSVMLG) [62] not only exploited the local information of the dataset but also considered the local correlation among each local region, where they first pre-constructed a number of local models of the training set and built the decision functions in each local model with a global regularization. Since the relative density degrees reflect the local geometry of the sample manifold and the scatters of the two classes points, Peng and Xu [63] presented the bi-density TWSVM (BDTWSVM) which incorporated the relative density degrees for all training points using the intra-class graph. [64] incorporated a manifold regularization term into LSTWSVM (ManLSTWSVM) to discover the local geometry inside the samples. LSTWSVM via maximum one-class

within-class variance, termed as (MWSVM) [65], applied the one-class within-class variance to the classifier, and a localized version (LMWSVM) of MWSVM was further proposed to remove the outliers effectively.

3.1.4 Sparse TWSVMs

Sparse nonparallel SVM (SNSVM) [66], a little difference with NPSVM [27] on the regularization terms used, is a sparse TWSVM since it also took the ε -insensitive loss function instead of the quadratic loss function in TWSVM. Peng [67] proposed a rapid sparse TWSVM (STSVM) in the primal space to improve the sparsity and robustness, where for each nonparallel hyperplane, the STSVM employed a back-fitting strategy to iteratively and simultaneously added a support vector. [68, 69] formulated an exact 1-norm linear programming formulation of TWSVM to improve the robustness and sparsity, which led to an extremely simple and fast algorithm.

3.1.5 Structural TWSVMs

Qi et al. [70] designed a novel structural TWSVM (*S*-TWSVM) only considering one class's structural information for each model, which was different with all existing structural classifiers such as [71, 72]. This made *S*-TWSVM further reduce the computational complexity of the related QPPs, and improved the model's generalization capacity. A robust minimum class variance TWSVM (RMCV-TWSVM) [73] introduced the variance matrices for the two classes as the regularization term for better generalization performance. Peng and Xu [74] constructed two Mahalanobis distance-based kernels according to the covariance matrices of two classes of data for optimizing the nonparallel hyperplanes, the Mahalanobis distance-based SVM (TMSVM) is suitable especially for the case that the covariance matrices of two classes of data are obviously different. For the TPMSVM [38], Peng et al. [75] presented a structural version (STPMSVM) by focusing on the structural information of the corresponding classes based on the cluster granularity.

3.2 Other Binary TWSVMs

[76] attempted to improve computing time of TWSVM by converting the primal QPPs of TWSVM into smooth unconstrained minimization problems. The smooth reformulations were solved using the well-known Newton-Armijo algorithm. Shao and Deng [77] considered a margin-based TWSVM with unity norm hyperplanes (UNH-MTSVM), the resulted optimal problems were solved efficiently by converting them into smooth unconstrained minimization problems. Ghorai et al. [78] reformulated TWSVM by considering unity norm of the normal vector of the hyperplanes as the constraints, the resulting nonlinear programming problem was solved by sequential quadratic optimization method. A classifier named norm-mixed TWSVM (NMTWSVM) was presented in [79], its main idea was to replace the hinge loss of the other class in the primal problems with the L_1 -norm-based loss. The geometric analysis showed that the dual problems of NMTWSVM can be interpreted as a pair of

minimum generalized Mahalanobis-norm problems (MGMNPs) on the two reduced affine hulls(RAHs) composed of two classes of points.

A feature selection method based on TWSVM (FTSVM) was designed in [80], where a multi-objective mixed-integer programming problem was solved by a greedy algorithm, in addition, the linear FTSVM was extended to the nonlinear case. Another feature selection method based on LSTWSVM (FLSTWSVM) [81] incorporated a Tikhonov regularization term to the objective of LSTWSVM, and then minimized its 1-norm measure. FLSTWSVM can reduce input features for the linear case and determined few kernel functions for the case. [82] treated the kernel selection problem for TWSVM as an optimization problem over the convex set of finitely many basic kernels.

Shao et al. proposed a TWSVM probability model (PTWSVM) [83] to estimate the posterior probability, they first defined a new ranking continues output and then mapped it into probability by training the parameters of an additional sigmoid function. For imbalanced data classification, [84] used under-sampling to give the adaptive weights for each class, which overcame the bias phenomenon in the TWSVM. Different with standard TWSVMs, [85] constructed two nonparallel hyperplanes simultaneously by solving a single quadratic programming problem, and was consistent between its predicting and training processes. Twin support tensor machine (TWSVM) [86, 87], following the idea of support tensor machine (STM) [88–90], is also an interesting research direction.

3.3 Variants of Twin Support Vector Regressions (TWSVRs)

Some variants of twin support vector machines for regression problems (TWSVRs) are introduced in this section. The training set of the regression problem is given as

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (R^n \times \mathcal{Y})^l, \quad (27)$$

where $x_i \in R^n$, $y_i \in \mathcal{Y} = R$, $i = 1, \dots, l$. [91] proposed twin support vector regression (TSVR) which was similar to TWSVM in spirit, as it derived a pair of ε -insensitive up-bound and down-bound nonparallel planes around the data points. Its least squares version was developed in [92]. [93] improved the TSVR by formulating it as a pair of linear programming problems instead of QPPs. Weighted TSVR [94] assigned the samples in the different positions with different penalties, which can avoid the over-fitting problem to a certain extent. ε -TSVR [95] implemented the structural risk minimization principle by introducing the regularization term in primal problems of TSVR, and the SOR technique was used to solve the optimization problems to speed up the training procedure. [96] adopted the regularization to convert original TSVR into a well-posed problem and employed the L_1 -norm based loss function and regularization to introduce robustness and sparsity simultaneously. Similar to the par- ν -SVR [39, 97] proposed an efficient twin parametric insensitive SVR (TPISVR), which was suitable for the case that the noise is heteroscedastic.

The smoothing technique was applied to convert the original constrained minimization problems of TSVR to unconstrained ones such that Newton algorithm with Armijo inexact stepsize could be adopted effectively [98]. A primal version for TSVR

(PTSVR) was presented in [99] by introducing the smooth function to approximate its loss function, PTSVR directly optimized the QPPs in the primal space based on a series of sets of linear equations. A least squares version for TSVR (PLSTSVR) was also considered in the primal space [100]. [101] introduced a simple and linearly convergent Lagrangian SVM algorithm for the dual of TSVR. [102] proposed the reduced TSVR (RTSVR) using the notion of rectangular kernels to obtain significant improvements in execution time over the TSVR. Regressors based on TSVR for the simultaneous learning of a function and its derivatives were discussed in [103,104], and showed their significant improvements in computational time and estimation accuracy.

3.4 Multiclass TWSVMs

To solve the multiple classification problem with the training set

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\}, \quad (28)$$

where $x_i \in R^n, i = 1, \dots, l, y_i \in \{1, \dots, K\}$ is the corresponding pattern of x_i , there are two types of strategies, one is the “decomposition-reconstruction” strategy [2] by solving a series of smaller optimization problems, while the other one is the “all-together” strategy [105] by solving only one single optimization problem. An interesting model proposed in [106] belonging to the first strategy, termed as multiple birth support vector machine (MBSVM), solved multi-QPPs simultaneously to seek for K nonparallel hyperplanes

$$(w_k \cdot x) + b_k = 0, k = 1, \dots, K. \quad (29)$$

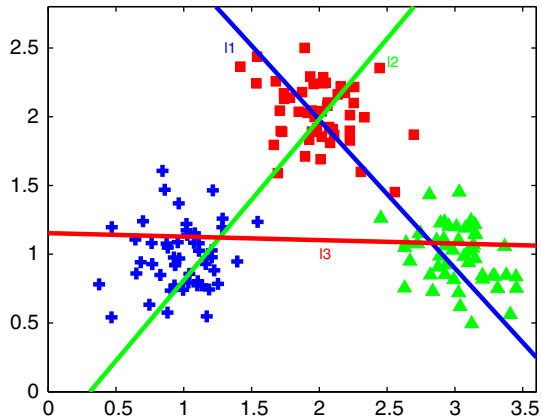
Compared with the straightforward multiclass extension of TWSVM or GEPSVM [107–109], MBSVM took into account the computational complexity, the “min” decision criterion of TWSVM is changed into the “max” one of MBSVM. The geometric interpretation of MBSVM with $x \in R^2$ is shown in Fig. 2, where the k th hyperplane is at least one distance from the k th class points, and is closer to the other classes, a test pattern is assigned to a class depending on which one it lies farthest to (Fig. 3).

[110] combined the decision tree with the TWSVM together (DTTSVM) for the multi-class classification, which can overcome the possible ambiguous occurred in MBSVM. Twin-KSVC took the advantages of both TSVM and K-SVCR (support vector classification-regression machine for k -class classification) [111] and evaluated all the training points into an “1-versus-1-versus-rest” structure [112]. A multi-class LSTWSVM based on optimal directed acyclic graph (DAG) was proposed in [113], an average distance measure and a non-repetitive sequence number rearrangement method were offered to reduce the cumulative errors caused by DAG structure.

3.5 Semi-Supervised TWSVMs

Semi-Supervised Learning (SSL) has attracted an increasing amount of interest in the last decade [114]. For the Semi-supervised binary classification problem with the training set

Fig. 3 A toy example learned by the linear MBSVM [106]



$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \cup \{x_{l+1}, \dots, x_{l+q}\}, \tag{30}$$

where $x_i \in R^n, y_i \in \{-1, 1\}, i = 1, \dots, l, x_i \in R^n, i = l + 1, \dots, l + q$, and the set $\{x_{l+1}, \dots, x_{l+q}\}$ is a collection of unlabeled inputs known to belong to one of the two classes, several novel approaches for making use of the unlabeled data to improve the performance of classifiers have been proposed, in which Laplacian SVM (LapSVM) [115] using the graph Laplacian was a state of the art method. Follow the idea of LapSVM, a novel Laplacian TWSVM (Lap-TSVM) for the SSL problem was proposed in [116], which can exploit the geometry information of the marginal distribution embedded in unlabeled data to construct a more reasonable classifier and be a useful extension of TSVM. Furthermore, by choosing appropriate parameters, Lap-TSVM degenerates to either TSVM or TBSVM. A variant of Lap-TSVM, named as Laplacian smooth TWSVM (Lap-ST SVM) was developed in [117]. They converted the constrained QPPs of Lap-TSVM into unconstrained minimization problems (UMPs) in primal space, and the smooth technique was then introduced to make these UMPs twice differentiable, therefore a fast Newton-Armijo algorithm was further designed to solve the UMPs efficiently, which converges globally and quadratically.

The PU problem where the training set consists of a few Positive examples and a large collection of Unlabeled examples, another kind of semi-supervised learning problem, is now gaining more and more attention [118]. Formally, its training set is

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \cup \{x_{l+1}, \dots, x_{l+q}\}, \tag{31}$$

where $x_i \in R^n, y_i = 1$, i.e. x_i is a positive input, $i = 1, \dots, l; x_i \in R^n$, i.e. x_i is an unlabeled input known to belong to one of the two classes, $i = l + 1, \dots, l + q$. For the PU problem, Biased TWSVM (B-TWSVM) based on the Biased SVM (BSVM) [119] was proposed in [120], while [121] applied NPSVM for the PU learning. Both of them firstly converted the PU problem into an unbalance binary classification problem, then applied different weights to the positive examples and the unlabeled examples separately.

3.6 Knowledge-based TWSVMs

In this section, a review of the methods incorporating prior knowledge into TWSVMs, however restricted to the Universum TWSVM, TWSVM with the Privileged information, and Robust TWSVM is given.

Supervised learning problem with Universum samples is a new research subject in machine learning. The concept of Universum sample was firstly introduced in [122], which was defined as the sample not belonging to either class of the classification problem of interest. The authors proposed a new SVM framework, called U -SVM and their experimental results show that U -SVM outperforms those SVMs without considering Universum data. Therefore, TWSVM with Universum (called U -TSVM) was proposed in [123] combining the advantages of TWSVM and U -SVM. Its weighted version can be found in [124]. A nonparallel SVM that can exploit prior knowledge embedded in the Universum (U -NSVM) [125] was different with U -TSVM since it solves one QPP to get two nonparallel hyperplanes simultaneously.

In human learning, there is a lot of teacher information such as explanations, comments, comparisons and so on along with the given examples. Vapnik et al. [126] called this kind of additional prior information as the privileged information, which is only available at the training stage and never available for the test set, i.e., the training set is

$$T = (x_1, x_1^*, y_1), \dots, (x_l, x_l^*, y_l) \quad (32)$$

where $x_i \in R^n, x_i^* \in R^m, y_i \in \{-1, 1\}, i = 1, \dots, l$. They proposed a new learning model, called Learning Using Privileged Information (LUPI), to accelerate the convergence rate of learning specially when the learning problem itself is hard. [127] proposed a fast TWSVM using privileged information (called FTSVMPI), and focused on an interesting topic: how to exploit the privileged information to improve the performance of the Visual Tracking Object (VOT) problem.

Some uncertainty is often present in many real-world problems. For example, when the inputs are subjected to measurement errors, it would be better to describe the inputs by uncertainty sets $\mathcal{X}_i \subset R^n, i = 1, \dots, l$, since all we know is that the input belongs to the set \mathcal{X}_i , so the training set should become

$$T = \{(\mathcal{X}_1, \mathcal{Y}_1), \dots, (\mathcal{X}_l, \mathcal{Y}_l)\}, \quad (33)$$

where \mathcal{X}_i is a set in $R^n, \mathcal{Y}_i \in \{-1, 1\}$. The goal is to explore a robust model which can deal with such data set. There are many methods of constructing the robust SVMs such as [128, 129]. Robust TWSVM (R -TWSVM) [130] via second order cone programming (SOCP) formulations for classification was an improved extension of TWSVM, since there were only inner products about inputs in the dual problems which kernel trick can be directly applied for nonlinear cases and the inverse of matrices were not needed any more.

The prior knowledge in the form of multiple polyhedral sets were incorporated into the linear TWSVM and LSTWSVM, termed as knowledge based TWSVM (KBTWSVM) and knowledge based LSTWSVM (KBLSTWSVM), were formulated

in [131]. A 1-norm regularized KBTWSVM resulted in a linear programming was proposed in [132].

3.7 Multi-instance and Multi-task TWSVMs

Multiple-instance learning (MIL), being useful in many applications including text categorization, natural scene classification, image retrieval and so on, has received intense interests recently in the field of machine learning. In MIL framework, the training set consists of positive and negative bags of instances which are points in the n -dimensional real space R^n , and each bag contains a set of instances. In each positive training bag, there must be at least one positive instance, whereas a negative bag contains only negative instances. The aim of MIL is to construct a classifier learned from the training set to correctly label unseen bags. A number of MIL methods based on SVMs emerged [133, 134]. Based on TWSVM, [135] extended the idea of MI-SVM [133] to construct the MIL nonparallel classifier (MI-NSVM). The method was mainly divided into two steps: The first step was to generate a spare hyperplane and estimate the score of each instance in positive bags. The second step, MI-NSVM sought the “most positive” instance of each positive bag by the information obtained in the first step, and then generated the second hyperplane. MI-TWSVM [136] also aimed at generating a positive hyperplane and a negative hyperplane, such that the former one is close to at least one instance (“witness” instance) in every positive bag and is far from all instances belonging to negative bags, and the latter one is close to all instances belonging to negative bags and is far from the “witness” instances in positive bags. Instead of having QPPs, the MI-TWSVM optimization problem were bilevel programming problems (BLPPs). [137] improved MI-TWSVM by implementing the structural risk minimization (SRM) principle and solving a series of QPPs instead of the BLPPs. [138] proposed a least squares version for MIL based on LSTWSVM which used an iterative strategy by solving two linear systems of equations and a QPP alternatively.

Multi-task learning (MTL) is a learning paradigm which seeks to improve the generalization performance of a task with the help of other tasks [139]. When there are relations between the tasks to learn, it can be advantageous to learn all tasks simultaneously instead of following the more traditional approach of learning each task independently of the others. SVMs have been explored for multi-task learning [140, 141], and [142] embedded multi-task learning into TSVMs and made use of the contribution of correlation of tasks.

3.8 Large scale TWSVMs

Considerable efforts have been devoted to the implementation of efficient optimization method for solving the QPP in standard SVMs, such as the chunking and decomposition methods [1, 31, 143–145]. For TWSVMs, though [25, 33] and other authors applied the SOR technique or the coordinate descent method, they in fact took no account of the computation of the inverse matrices and will fail facing the real large scale prob-

lems. Since NPSVM [27,66] skillfully avoided the computation of the inverse matrices and applied the SOR or SMO to solve the QPPs, they can deal with the large scale problems really.

For some applications such as document classification with the data appearing in a high dimensional feature space, linear SVM in which the data are not mapped, has similar performances with nonlinear SVM. For linear SVM, many methods have been proposed in large-scale scenarios such as [146,147], while for linear TWSVM, L_1 -NPSVM [148] based on the NPSVM [27] and L_2 -TWSVM [149] on ITSVM [26], were designed to deal with large-scale data based on an efficient solver—dual coordinate descent (DCD) method, which is a popular optimization technique, updates one variable at a time by minimizing a single-variable subproblem. Since the dual problems in L_1 -NPSVM or L_2 -TWSVM have the same formulations with that of standard L_1 -SVM or L_2 -SVM, and [147] proposed the DCD method for them, pointed out that the DCD method makes crucial advantage of the linear kernel and outperforms other solvers when the numbers of data and features are both large, therefore the DCD method can be applied for linear TWSVM naturally with several minor modifications.

4 Applications of TWSVMs

As SVMs has many applications, TWSVMs can also be applied into the corresponding fields. This section reviews the limited applications of TWSVMs because of their few years developments.

Speaker recognition is to decide which person is talking from a group of known speakers, [150] combined TWSVMs with Gaussian Mixture Model (GMM) [151] for text-independent speaker recognition where GMMs were used to extract features. [152] addressed the task of gesture recognition using surface electromyogram (sEMG) data, and proved TWSVM to be an effective approach in meeting the learning problem from multi-class data where patterns in different classes arise from different distributions, since TWSVM is a more natural choice for applying to unbalanced datasets. Human activity recognition (HAR) is an important research branch in computer vision, and [153] proposed a framework for HAR with the combination of local space-time features and LSTWSVM. Speech emotion recognition is used to solve the problem of “how to speak”, just like speaker recognition is proposed to solve “who is speaking”, TWSVM was applied into this problem tentatively [154].

Clustered microcalcifications (MCs) in mammogram can be an indicator of breast cancer, [155] presented a mass detection system based on TWSVM to distinguish the masses from normal breast tissues accurately, [156] extracted combined image features from each image block of positive and negative samples, and trained TWSVM with the 164 dimensional features of each sample to test at every location in a mammogram whether the MCs was present or not. [157] boosted the TWSVM for higher accuracy in MCs detection. A predictive model for heart disease diagnosis using feature selection based on LSTWSVM was developed in [158], where the selection of significant features improved the accuracy. Ding et al. [159] applied TWSVM to the intrusion detection to improve the speed of detection and accuracy.

5 Remarks and Future Directions

This paper has offered an extensive review of TWSVMs, including four basic models: original TWSVM, the improved versions TBSVM, ITSVM and NPSVM. Based on the four models, the extensions and variants of TWSVMs are reviewed, such as the ν -TWSVMs, least squares TWSVMs, localized TWSVMs, sparse TWSVMs, structural TWSVMs, and linear programming TWSVMs for the binary classification problems; TWSVRs for the regression problems; Multi-class TWSVMs for the multi-class problems; Semi-supervised TWSVMs for the semi-supervised problems; Knowledge-based TWSVMs for the problems with the Universum set, Privileged information, or uncertain information; Multi-instance and Multi-task TWSVMs; and also the large scale TWSVMs. Some of these models have already been used in several real-life applications, such as speaker recognition, intrusion detection and etc. Researchers and engineers in data mining, especially in SVMs can benefit from this survey in better understanding the essence of the TWSVMs. In addition, it can also serve as a reference repository of such approaches.

In this paper, we can see the development of TWSVMs following the way of modifying the corresponding optimization models, since the essence of TWSVMs is to construct appropriate optimization models (linear programming, quadratic programming, nonlinear programming, second order cone programming) for different learning problems. TWSVMs have better generalization ability, are more flexible than TBSVMs because of searching for the nonparallel hyperplanes. However, there's no such thing as a free lunch. Just as we pointed out for the ITSVM and NPSVM models in Sect. 2, ITSVM sacrifices more model training time than TWSVM to skillfully avoids the computation of the inverse matrices, while NPSVM sacrifices further more model training time than ITSVM to get the sparsity. We can not count on two smaller QPPs to get better performance than one large OPP as the declaration in the original TWSVM, since the costly computation of inverse matrices are not considered.

As we can see, the extensions or variants of TWSVMs were mostly based on the original TWSVMs or TBSVM, so there is a great space for development based on the ITSVM and NPSVM. And new practical problems remaining to be explored present new challenges to TWSVMs to construct new optimization models. These models should also have the same desirable properties as the models in this paper including: good generalization, scalability, simple and easy implementation of algorithm, robustness, as well as theoretically known convergence and complexity [160].

Acknowledgments This work has been partially supported by grants from National Natural Science Foundation of China (Nos. 11271361, 61472390, 61402429, 71331005), Major International (Regional) Joint Research Project (No. 71110107026), the Ministry of water resources' special funds for scientific research on public causes (No. 201301094).

References

1. Cortes C, Vapnik V (1995) Support vector networks. *Mach Learn* 20(3):273–297
2. Vapnik VN (1996) *The nature of statistical learning theory*. Springer, New York
3. Vapnik VN (1998) *Statistical learning theory*. Publishing House of Electronics Industry, New York

4. Cristianini N, Taylor JS (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
5. Fung GM, Mangasarian OL (2001) Multicategory proximal support vector machine classifiers. *Mach Learn* 59(1—2):77–97
6. Deng NY, Tian YJ (2009) Support vector machines: theory, algorithms and extensions. Science Press, Beijing
7. Deng NY, Tian YJ, Zhang CH (2012) Support vector machines: optimization based theory, algorithms and extensions. CRC Press, Chapman and Hall, Boca Raton
8. Joachims T (1999) Text categorization with support vector machines: learning with many relevant features. In: Proceedings of 10th European conference on machine learning, pp 137–142
9. Lodhi H, Cristianini N, Shawe-Taylor J, Watkins C (2000) Text classification using string kernels. *Adv Neural Inf Process Syst* 13:563–569
10. Jonsson K, Kittler J, Matas YP (2002) Support vector machines for face authentication. *J Image Vis Comput* 20(5):369–375
11. Tefas A, Kotropoulos C, Pitas I (2001) Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. *IEEE Trans Pattern Anal Mach Intell* 23(7):735–746
12. Ganapathiraju A, Hamaker J, Picone J (2004) Applications of support vector machines to speech recognition. *IEEE Trans Signal Process* 52(8):2348–2355
13. Gutta S, Huang JRJ, Jonathon P, Wechsler H (2000) Mixture of experts for classification of gender, ethnic origin, and pose of human. *IEEE Trans Neural Netw* 11(4):948–960
14. Shin KS, Lee TS, Kim HJ (2005) An application of support vector machines in bankruptcy prediction model. *Expert Syst Appl* 28(1):127–135
15. Melgani F, Bruzzone L (2004) Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans Geosci Remote Sens* 42(8):1778–1790
16. Kim KJ (2003) Financial time series forecasting using support vector machines. *Neurocomputing* 55(1):307–319
17. Liu Y, Zhang D, Lu GG, Ma WY (2007) A survey of content-based image retrieval with high-level semantics. *Pattern Recogn* 40(1):262–282
18. Adankon MM, Cheriet M (2009) Model selection for the LS-SVM application to handwriting recognition. *Pattern Recogn* 42(12):3264–3270
19. Borgwardt KM (2011) Kernel methods in bioinformatics. *Handbook of statistical bioinformatics*, Part 3, pp 317–334
20. Khan NM, Ksantini R, Ahmad IS, Boufama B (2012) A novel SVM plus NDA model for classification with an application to face recognition. *Pattern Recogn* 45(1):66–79
21. Mangasarian OL, Wild EW (2006) Multisurface proximal support vector machine classification via generalized eigenvalues. *IEEE Trans Pattern Anal Mach Intell* 28(1):69–74
22. Jayadeva RK, Khemchandani R, Chandra S (2007) Twin support vector machines for pattern classification. *IEEE Trans Pattern Anal Mach Intell* 29(5):905–910
23. Ding SF, Hua XP, Yu JZ (2014) An overview on nonparallel hyperplane support vector machine algorithms. *Neural Comput Appl* 25:975–982. doi:[10.1007/s00521-013-1524-6](https://doi.org/10.1007/s00521-013-1524-6)
24. Ding SF, Yu JZ, Qi BJ, Huang HJ (2014) An overview on twin support vector machines. *Artif Intell Rev* 42(2):245–252
25. Shao YH, Zhang CH, Wang XB, Deng NY (2011) Improvements on twin support vector machines. *IEEE Trans Neural Netw* 22(6):962–968
26. Tian YJ, Ju XC, Qi ZQ, Shi Y (2013) Improved twin support vector machine. *Sci China Math* 57(2):417–432
27. Tian YJ, Qi ZQ, Ju XC, Shi Y, Liu XH (2013) Nonparallel support vector machines for pattern classification. *IEEE Trans Cybern* 44(7):1067–1079
28. Guarracino MR, Cifarelli C, Seref O, Pardalos PM (2007) A classification method based on generalized eigenvalue problems. *Optim Methods Softw* 22(1):73–81
29. Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. The John Hopkins University Press, Baltimore
30. Fung G, Mangasarian OL (2001) Proximal support vector machine classifiers. In: Proceedings KDD-2001: knowledge discovery and data mining, San Francisco, CA, pp 77–86
31. Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: *Advances in kernel methods—support vector learning*. MIT Press, Cambridge, pp 185–208

32. Mangasarian OL, Musicant DR (1999) Successive overrelaxation for support vector machines. *IEEE Trans Neural Netw* 10(5):1032–1037
33. Shao YH, Deng NY (2012) A coordinate descent margin based-twin support vector machine for classification. *Neural Netw* 25:114–121
34. Tian YJ, Zhang Q, Liu DL (2014) ν -nonparallel support vector machine for pattern classification. *Neural Comput Appl*. doi:[10.1007/s00521-014-1575-3](https://doi.org/10.1007/s00521-014-1575-3)
35. Peng XJ (2010) A ν -twin support vector machine (ν -TSVM) classifier and its geometric algorithms. *Inf Sci* 180:3863–3875
36. Pawlak Z (1982) Rough sets. *Int J Comput Inf Sci* 11:341–356
37. Xu YT, Wang LS, Zhong P (2012) A rough margin-based ν -twin support vector machine. *Neural Comput Appl* 21(6):1307–1317
38. Peng X (2011) TPMSVM: a novel twin parametric-margin support vector for pattern recognition. *Pattern Recogn* 44(10–11):2678–2692
39. Hao PY (2010) New support vector algorithms with parametric insensitive margin model. *Neural Netw* 23(1):60–73
40. Shao YH, Wang Z, Chen WJ, Deng NY (2013) Least squares twin parametric-margin support vector machine for classification. *Appl Intell* 39:451–464
41. Wang Z, Shao YH, Wu TR (2013) A GA-based model selection for smooth twin parametric-margin support vector machine. *Pattern Recogn* 46(8):2267–2277
42. Kumar MA, Gopal M (2009) Least squares twin support vector machines for pattern classification. *Expert Syst Appl* 36(4):7535–7543
43. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. World Scientific Pub. Co., Singapore
44. Chen J, Ji GG (2010) Weighted least squares twin support vector machines for pattern classification, vol. 2. In: The 2nd international conference on computer and automation engineering, pp 242–246
45. Tomar D, Singhal S, Agarwal S (2014) Weighted least square twin support vector machine for imbalanced dataset. *Int J Database Theory Appl* 7(2):25–36
46. Gao SB, Ye QL, Ye N (2011) 1-norm least squares twin support vector machines. *Neurocomputing* 74(17):3590–3597
47. Xu Y, Xi W, Lv X, Guo R (2012) An improved least squares twin support vector machine. *J Inf Comput Sci* 9:1063–1071
48. Peng XJ, Xu D (2013) A twin-hypersphere support vector machine classifier and the fast learning algorithm. *Inf Sci* 12–27
49. Peng XJ, Xu D (2014) Twin support vector hypersphere (TSVH) classifier for pattern recognition. *Neural Comput Appl* 24(5):1207–1220
50. Peng XJ (2010) Least squares twin support vector hypersphere (LS-TSVH) for pattern recognition. *Expert Syst Appl* 37(12):8371–8378
51. Tax D, Duijn R (2004) Support vector data description. *Mach Learn* 54:45–66
52. Chen XB, Yang J, Ye QL, Liang J (2011) Recursive projection twin support vector machine via within-class variance minimization. *Pattern Recogn* 44(10–11):2643–2655
53. Shao YH, Wang Z, Chen WJ, Deng NY (2013) A regularization for the projection twin support vector machine. *Knowl-Based Syst* 37:203–210
54. Hua XP, Ding SF (2012) Matrix pattern based projection twin support vector machines. *Int J Digital Content Technol Appl* 6(20):172–181
55. Shao YH, Deng NY, Yang ZM (2012) Least squares recursive projection twin support vector machine for classification. *Pattern Recogn* 45(6):2299–2307
56. Ding SF, Hua XP (2014) Recursive least squares projection twin support vector machines for nonlinear classification. *Neurocomputing* 130:3–9
57. Guo JH, Yi P, Wang RL, Ye QL, Zhao CX (2014) Feature selection for least squares projection twin support vector machine. *Neurocomputing*. doi:[10.1016/j.neucom.2014.05.040](https://doi.org/10.1016/j.neucom.2014.05.040)
58. Ghorai S, Mukherjee A, Dutta PK (2009) Nonparallel plane proximal classifier. *Signal Process* 89(4):510–522
59. Ye QL, Zhao CX, Gao SB, Zheng H (2012) Weighted twin support vector machines with local information and its application. *Neural Netw* 35:31–39
60. Ye QL, Zhao CX, Ye N, Chen XB (2011) Localized twin SVM via convex minimization. *Neurocomputing* 74(4):580–587

61. Wang YN, Tian YJ (2012) Fast localized twin SVM. In: 8th international conference on natural computation, pp 74–78
62. Wang YN, Zhao X, Tian YJ (2013) Local and global regularized twin SVM, vol. 18. In: International conference on computational science, pp 1710–1719
63. Peng XJ, Xu D (2013) Bi-density twin support vector machines for pattern recognition. *Neurocomputing* 99:134–143
64. Wang D, Ye QL, Ye N (2010) Localized multi-plane twsvm classifier via manifold regularization, vol. 2. In: International conference on intelligent human–machine systems and cybernetics, pp 70–73
65. Ye QL, Zhao CX, Ye N (2012) Least squares twin support vector machine classification via maximum one-class within class variance. *Optim Methods Softw* 27(1):53–69
66. Tian YJ, Ju XC, Qi ZQ (2013) Efficient sparse nonparallel support vector machines for classification. *Neural Comput Appl* 24(5):1089–1099
67. Peng XJ (2011) Building sparse twin support vector machine classifiers in primal space. *Inf Sci* 181(18):3967–3980
68. Tanveer M (2013) Smoothing technique on linear programming twin support vector machines. *Int J Mach Learn Comput* 3(2):240–244
69. Tanveer M (2014) Robust and sparse linear programming twin support vector machines. *Cognitive Comput*. doi:10.1007/s12559-014-9278-8
70. Qi ZQ, Tian YJ, Shi Y (2013) Structural twin support vector machine for classification. *Knowl-Based Syst* 43:74–81
71. Kzhuang KH, Yang H, King I (2004) Learning large margin classifiers locally and globally. In: The twenty-first international conference on machine learning (ICML-2004), pp 401–408
72. Xue H, Chen S, Yang Q (2011) Structural regularized support vector machine: a framework for structural large margin classifier. *IEEE Trans Neural Netw* 22(4):573–587
73. Peng XJ, Xu D (2013) Robust minimum class variance twin support vector machine classifier. *Neural Comput Appl* 22(5):999–1011
74. Peng XJ, Xu D (2012) Twin Mahalanobis distance-based support vector machines for pattern recognition. *Inf Sci* 200:22–37
75. Peng XJ, Wang YF, Xu D (2013) Structural twin parametric margin support vector machine for binary classification. *Knowl-Based Syst* 49:63–72
76. Kumar MA, Gopal M (2008) Application of smoothing technique on twin support vector machines. *Pattern Recogn Lett* 29(13):1842–1848
77. Shao YH, Deng NY (2013) A novel margin-based twin support vector machine with unity norm hyperplanes. *Neural Comput Appl* 22:1627–1635
78. Ghorai S, Hossain SJ, Mukherjee A, Dutta PK (2010) Unity norm twin support vector machine classifier. In: Annual IEEE India conference, pp 1–4
79. Peng XJ, Xu D (2013) Norm-mixed twin support vector machine classifier and its geometric algorithm. *Neurocomputing* 99:486–495
80. Bai L, Wang Z, Shao YH, Deng NY (2014) A novel feature selection method for twin support vector machine. *Knowl-Based Systems* 59:1–8
81. Ye QL, Zhao CX, Ye N, Zheng H, Chen XB (2012) A feature selection method for nonparallel plane support vector machine classification. *Optim Methods Softw* 27(3):431–443
82. Khemchandani R, Jayadeva, Chandra S (2009) Optimal kernel selection in twin support vector machines. *Optim Lett* 3(1):77–88
83. Shao YH, Deng NY, Yang ZM, Chen WJ, Wang Z (2012) Probabilistic outputs for twin support vector machines. *Knowl-Based Syst* 33:145–151
84. Shao YH, Chen WJ, Zhang JJ, Wang Z, Deng NY (2014) An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recogn* 47:3158–3167
85. Shao YH, Chen WJ, Deng NY (2014) Nonparallel hyperplane support vector machine for binary classification problems. *Inf Sci* 263:22–35
86. Zhang XS, Gao XB, Wang Y (2009) Twin support tensor machines for MCS detection. *J Electron (China)* 26:318–325
87. Zhao XB, Shi HF, Lv M, Jing L (2014) Least squares twin support tensor machine for classification. *J Inf Comput Sci* 11(12):4175–4189
88. Cai D, He XF, Wen JR, Han J, Ma WY (2006) Support tensor machines for text categorization. Department of Computer Science Technical Report No. 2714, University of Illinois at Urbana—Champaign (UIUCDCS-R-2006-2714)

89. Kotsia I, Patras I (2011) Support tucker machines. In: Proceedings of IEEE conference on computer vision and pattern recognition, Colorado, USA, pp 633–640
90. Kotsia I, Guo WW, Patras I (2012) Higher rank support tensor machines for visual recognition. *Pattern Recogn* 45:4192–4203
91. Peng XJ (2010) TSVR: an efficient twin support vector machine for regression. *Neural Netw* 23(3):365–372
92. Zhao YP, Zhao J, Zhao M (2013) Twin least squares support vector regression. *Neurocomputing* 118:225–236
93. Zhong P, Xu YT, Zhao YH (2012) Training twin support vector regression via linear programming. *Neural Comput Appl* 21(2):399–407
94. Xu YT, Wang LS (2012) A weighted twin support vector regression. *Knowl-Based Syst* 33:92–101
95. Shao YH, Zhang CH, Yang ZM, Jing L, Deng NY (2013) ϵ -twin support vector machine for regression. *Neural Comput Appl* 23(1):175–185
96. Chen XB, Yang J, Chen L (2014) An improved robust and sparse twin support vector regression via linear programming. *Chin Conf Pattern Recogn*. doi:[10.1007/s00500-014-1342-5](https://doi.org/10.1007/s00500-014-1342-5)
97. Peng XJ (2012) Efficient twin parametric insensitive support vector regression model. *Neurocomputing* 79:26–38
98. Chen XB, Yang J, Liang J, Ye QL (2012) Smooth twin support vector regression. *Neural Comput Appl* 21(3):505–513
99. Peng XJ (2010) Primal twin support vector regression and its sparse approximation. *Neurocomputing* 73(16–18):2846–2858
100. Huang HJ, Ding SF, Shi ZZ (2013) Primal least squares twin support vector regression. *J Zhejiang Univ Sci C* 14(9):722–732
101. Balasundaram S, Tanveer M (2013) On Lagrangian twin support vector regression. *Neural Comput Appl* 22(1):257–267
102. Singh M, Chadha J, Ahuja P, Jayadeva, Chandra S (2011) Reduced twin support vector regression. *Neurocomputing* 74(9):1471–1477
103. Khemchandani R, Karpatne A, Chandra Suresh (2013) Twin support vector regression for the simultaneous learning of a function and its derivatives. *Int J Mach Learn Cybern* 4(1):51–63
104. Jayadeva, Khemchandani R, Chandra S (2006) Regularized least squares twin svr for the simultaneous learning of a function and its derivative. In: 2006 international joint conference on neural networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, pp 1192–1197
105. Allwein EL, Schapire RE (2001) Reducing multiclass to binary: a unifying approach for margin classifiers. *J Mach Learn Res* 1:113–141
106. Yang Z, Shao Y, Zhang X (2013) Multiple birth support vector machine for multiclass classification. *Neural Comput Appl* 22(1):153–161
107. Wu ZD, Yang CF (2009) Study to multi-twin support vector machines and its applications in speaker recognition. In: International conference on computational intelligence and software engineering, pp 1–4
108. Zhen W, Jin C, Ming Q (2010) Non-parallel planes support vector machine for multi-class classification. *Int Conf Logistics Syst Intell Manag* 1:581–585
109. Jayadeva R, Khemchandani, Chandra S (2007) Fuzzy multi-category proximal support vector classification via generalized eigenvalues. *Soft Comput* 11(7):679–685
110. Shao YH, Chen WJ, Huang WB, Yang ZM, Deng NY (2013) The best separating decision tree twin support vector machine for multi-class classification. *Proc Comput Sci* 17:1032–1038
111. Angulo C, Parra X, Catal A (2003) K-SVCR: a support vector machine for multi-class classification. *Neurocomputing* 55:57–77
112. Xu YT, Guo R, Wang LS (2013) A twin multi-class classification support vector machine. *Cogn Comput* 5(4):580–588
113. Chen J, Ji GG (2010) Multi-class ltsvm classifier based on optimal directed acyclic graph, vol. 3. In: The 2nd international conference on computer and automation engineering, pp 100–104
114. Zhu XJ (2006) Semi-supervised learning literature survey. Computer Sciences TR 1530, University of Wisconsin
115. Belkin M, Niyogi PP, Sindhvani VV (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J Mach Learn Res* 7:2399–2434
116. Qi ZQ, Tian YJ, Shi Y (2012) Laplacian twin support vector machine for semi-supervised classification. *Neural Netw* 35:46–53

117. Chen WJ, Shao YH, Ning H (2014) Laplacian smooth twin support vector machine for semi-supervised classification. *Int J Mach Learn Cybern* 5(3):459–468
118. Geurts P (2011) Learning from positive and unlabeled examples by enforcing statistical significance. *Int Conf Artif Intell Stat* 15:305–314
119. Liu B (2006) *Web data mining: exploring hyperlinks, contents, and usage data*. Springer, Berlin
120. Xu ZJ, Qi ZQ, Zhang JQ (2014) Learning with positive and unlabeled examples using biased twin support vector machine. *Neural Comput Appl*. doi:[10.1007/s00521-014-1611-3](https://doi.org/10.1007/s00521-014-1611-3)
121. Zhang Y, Tian YJ, Ju XC (2014) Nonparallel hyperplane support vector machine for pu learning. In: *The 2014 10th international conference on natural computation (ICNC 2014)*
122. Weston J, Collobert R, Sinz F, Bottou LL, Vapnik V (2006) Inference with the universum. In: *Proceedings of the 23rd international conference on machine learning*, pp 1009–1016
123. Qi ZQ, Tian YJ, Shi Y (2012) Twin support vector machine with universum data. *Neural Netw* 36:112–119
124. Lu SX, Tong L (2014) Weighted twin support vector machine with universum. *Adv Comput Sci* 3(2):17–23
125. Qi ZQ, Tian YJ, Shi Y (2014) A nonparallel support vector machine for a classification problem with universum learning. *J Comput Appl Math* 263:288–298
126. Vapnik V, Vashist A (2009) A new learning paradigm: learning using privileged information. *Neural Netw* 22:544–557
127. Qi ZQ, Tian YJ, Shi Y (2014) A new classification model using privileged information and its application. *Neurocomputing* 129:146–152
128. Pannagadatta SK, Bhattacharyya C, Smola AJ (2006) Second order cone programming approaches for handling missing and uncertain data. *J Mach Learn Res* 7:1283–1314
129. Zhong P, Fukushima M (2007) Second order cone programming formulations for robust multi-classification. *Neural Comput* 19(1):258–282
130. Qi ZQ, Tian YJ, Shi Y (2013) Robust twin support vector machine for pattern classification. *Pattern Recognit* 46(1):305–316
131. Kumara MA, Khemchandani R, Gopala M, Chandrad S (2010) Knowledge based least squares twin support vector machines. *Inf Sci* 180(23):4606–4618
132. Ju XC, Tian YJ (2011) A novel knowledge-based twin support vector machine. *International conference on data mining workshops*, pp 429–433
133. Andrews S, Tsochantaridis I, Hofmann T (2003) Support vector machines for multiple instance learning. In: *Neural information processing systems*, pp 561–568
134. Yang ZX, Deng NY (2009) Multi-instance support vector machine based on convex combination. In: *The eighth international symposium on operations research and its applications*, pp 481–487
135. Qi ZQ, Tian YJ, Yu XD, Shi Y (2014) A multi-instance learning algorithm based on nonparallel classifier. *Appl Math Comput* 241:233–241
136. Shao YH, Yang ZX, Wang XB, Deng NY (2010) Multiple instance twin support vector machines. *The ninth international symposium on operations research and its applications*, pp 433–442
137. Zhang Q, Tian YJ, Liu DL (2013) Nonparallel support vector machines for multiple-instance learning. *Procedia Comput Sci* 17:1063–1072
138. Liu LY, Zhao YH, Zhong P (2012) Multiple instance classification based on least squares twin support vector machine. *J Convergent Inf Technol* 7(6):72–77
139. Caruana R (1997) Multitask learning. *Mach Learn* 28:41–75
140. Evgeniou T, Micchelli C, Pontil M (2006) Learning multiple tasks with kernel methods. *J Mach Learn Res* 6(1):615–623
141. Evgeniou T, Pontil M (2004) Regularized multitask learning. In: *KDD*, pp 109–117
142. Xie XJ, Sun SL (2012) Multitask twin support vector machines. *Neural Inf Process* 7664:341–348
143. Joachims T (1998) *Making large-scale svm learning practical*. In: *Advances in kernel methods-support vector learning*, MIT Press, Cambridge, pp 169–184
144. Chang CC, Lin C-J (2011) Libsvm: a library for support vector machines. *ACM Trans Intell Syst Technol* 2(3):27
145. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. *J Mach Learn Res* 9:1871–1874
146. Joachims T (2006) Training linear svms in linear time. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'06*, ACM, New York, pp 217–226

147. Hsieh CJ, Chang KW, Lin CJ, Keerthi SS, Sundararajan S (2008) A dual coordinate descent method for large-scale linear svm. In Proceedings of the 25th international conference on machine learning, ICML '08, ACM, New York, pp 408–415,
148. Tian YJ, Ping Y (2014) Large-scale linear nonparallel support vector machine solver. *Neural Netw* 50:166–174
149. Tian YJ, Zhang Q, Ping Y (2014) Large-scale linear nonparallel support vector machine solver. *Neurocomputing* 138:114–119
150. Cong HH, Yang CF, XRP (2008) Efficient speaker recognition based on multi-class twin support vector machines and gmms. In: 2008 IEEE conference on robotics, automation and mechatronics, pp 348–352
151. Liu M, Xie Y, Yao Z, Dai B (2006) A new hybrid gmm/svm for speaker verification. In: International conference on pattern recognition (ICPR'06)
152. Naik GR, Kumar DK, Jayadeva (2010) Twin svm for gesture classification using the surface electromyogram. *IEEE Trans Inf Technol Biomed* 14(2):301–308
153. Mozafari K, Nasiri JA, Charkari NM, Jalili S (2011) Action recognition by space-time features and least squares twin svm. In: The first international conference on informatics and computational intelligence, pp 287–292
154. Yang CF, Ji LP, Liu GS (2009) Study to speech emotion recognition based on twinssvm. *Fifth Int Conf Nat Comput* 2:312–316
155. Si X, Jing L (2009) Mass detection in digital mammograms using twin support vector machine-based cad system. *WASE Int Conf Inf Eng* 1:240–243
156. Zhang XS, Gao XB, Wang Y (2009) Mcs detection with combined image features and twin support vector machines. *J Comput* 4(3):215–221
157. Zhang XS (2009) Boosting twin support vector machine approach for mcs detection. *Asia-Pac Conf Inf Process* 1:149–152
158. Tomar D, Agarwal S (2014) Feature selection based least square twin support vector machine for diagnosis of heart disease. *Int J Bio-Sci Bio-Technol* 6(2):69–82
159. Ding XJ, Zhang GL, Ke YZ, Ma BL, Li ZC (2008) High efficient intrusion detection methodology with twin support vector machines. *Int Symp Inf Sci Eng* 1:560–564
160. Tian YJ, Shi Y, Liu XH (2012) Recent advances on support vector machines research. *Technol Econ Develop Econ* 18(1):5–33



Yingjie Tian received the First degree in mathematics in 1994, the masters degree in applied mathematics in 1997, and the Ph.D. degree in management science and engineering. He is currently a Professor with the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing, China. He has published four books about support vector machines (SVMs), one of which has been cited over 1000 times. His current research interests include SVMs, optimization theory and applications, data mining, intelligent knowledge management, and risk management.



Zhiquan Qi received a Master degree and Ph.D. degree in College of Science from China Agricultural University in 2006 and 2011. Currently he is a research assistant of research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences. His research interests include object detection, object tracking, change detecting and machine learning. He has published tens of papers about support vector machines, boosting, object detection, and so on.