J. Oper. Res. Soc. China (2015) 3:1–15
DOI 10.1007/s40305-014-0068-5

ORIGINAL PAPER

Sparse Proximal Support Vector Machine with a Specialized Interior-Point Method

Yan-Qin Bai · Zhao-Ying Zhu · Wen-Li Yan

Received: 26 October 2014 / Revised: 21 November 2014 / Accepted: 17 December 2014 /
Published online: 26 February 2015

Abstract Support vector machine (SVM) is a widely used method for classification. Proximal support vector machine (PSVM) is an extension of SVM and a promising method to lead to a fast and simple algorithm for generating a classifier. Motivated by the fast computational efforts of PSVM and the properties of sparse solution yielded by ℓ_1-norm, in this paper, we first propose a PSVM with a cardinality constraint which is eventually relaxed by ℓ_1-norm and leads to a trade-off $\ell_1 - \ell_2$ regularized sparse PSVM. Next we convert this $\ell_1 - \ell_2$ regularized sparse PSVM into an equivalent form of ℓ_1 regularized least squares (LS) and solve it by a specialized interior-point method proposed by Kim et al. (J Sel Top Signal Process 12:1932–4553, 2007). Finally, $\ell_1 - \ell_2$ regularized sparse PSVM is illustrated by means of a real-world dataset taken from the University of California, Irvine Machine Learning Repository (UCI Repository). Moreover, we compare the numerical results with the existing models such as generalized eigenvalue proximal SVM (GEPSVM), PSVM, and SVM-Light. The numerical results show that the $\ell_1 - \ell_2$ regularized sparse PSVM achieves not only better accuracy rate of classification than those of GEPSVM, PSVM, and SVM-Light, but also a sparser classifier compared with the ℓ_1-PSVM.

Keywords Proximal support vector machine · Classification accuracy · Interior-point methods · Preconditioned conjugate gradients algorithm

This research was supported by the National Natural Science Foundation of China (No. 11371242).

Y.-Q. Bai (✉) · Z.-Y. Zhu · W.-L. Yan
Department of Mathematics, Shanghai University, Shanghai 200444, China
e-mail: yqbai@shu.edu.cn

Z.-Y. Zhu
e-mail: zhaoyingzhu0825@163.com

W.-L. Yan
e-mail: wenliyy@163.com

Mathematics Subject Classification 90C10 · 90C20 · 49M20 · 65K05

1 Introduction

In classification problems, we are given a set of training data $\{(x_1, y_1), \cdots, (x_l, y_l)\}$, where $x_i \in \mathbb{R}^n$ is input and $y_i \in \{+1, -1\}$ is binary output. We wish to find a classifier to separate the training data into two sets, one is with the label "+1" and the other with the label "−1". Meanwhile, when given a new input x, the classifier can assign a label y from $\{+1, -1\}$ to it.

SVM has been proved to be a powerful tool in a wide range of areas to solve classification, pattern recognition, and regression problems. It has drawn much attention in recent years. Both hard and soft margin SVM were first proposed by Vapnik [19,20] based on the principle of structural risk minimization (SRM) principle. The SRM principle was realized by maximizing the margin of two separating parallel hyperplanes. To reduce the computational cost, several variants of SVM have been developed. For example, Suykens et al. [16,17] proposed a least squares support vector machine (LS-SVM), in which the objective function was modified by a least squares error term and the constraints were replaced by equality constraints. Recent study indicates that LS-SVM is efficient for feature selection, linear regression. Moreover, based on the optimization theory, Fung [11] developed proximal SVM (PSVM) to solve classification problem. By contrast, PSVM leads to a fast and simple algorithm for generating a system of linear equations. The formulation of PSVM greatly simplifies the problem with considerably faster computational time than SVM. Moreover, Chen et al. [7] used l_p-norm to replace l_1-norm and presented an l_p-norm proximal SVM and studied its applications. Deng et al. [9] presented a detail study of SVM, including algorithms and extensions. Recently, there are other directions to extend the supervised SVM to semi-supervised SVM, in which the datasets contain two parts, the training set and the test set. For instance, Bai et al. [1,2] developed conic optimization form for semi-supervised SVM.

Furthermore, the idea of ℓ_1 regularization is still receiving a lot of interests nowadays. In signal processing, the idea of ℓ_1 regularization comes up in several contexts including basis pursuit denoising [8] and signal recovery method from incomplete measurements [4,6]. In statistics, the idea of ℓ_1 regularization is used in the well-known Lasso algorithm [18] for feature selection.

We are inspired by the fast computational efforts of PSVM and the properties of sparse solution yielded by ℓ_1-norm. In this paper, we first propose a PSVM with a cardinality constraint which is eventually relaxed by ℓ_1-norm and leads to a trade-off $\ell_1 - \ell_2$ regularized sparse PSVM. Next we convert this $\ell_1 - \ell_2$ regularized sparse PSVM into an equivalent form of ℓ_1 regularized least squares (LSs) and solve it by a specialized interior-point method proposed by Kim et al. [12]. Finally, $\ell_1 - \ell_2$ regularized sparse PSVM is illustrated by means of a real-world dataset taken from the UCI Repository. Moreover, we compare the numerical results with the existing models such as GEPSVM, PSVM and SVM-Light. The numerical results show that the $\ell_1 - \ell_2$ regularized sparse PSVM achieves not only better accuracy rate of classification than those of GEPSVM, PSVM and SVM-Light, but also a sparser classifier compared with the ℓ_1-PSVM.

The paper is organized as follows. In Sect. 2, we briefly recall the basic concepts of SVM and PSVM, respectively. In Sect. 3, we add the cardinality constraint to the model of PSVM, then reformulate it as the $\ell_1 - \ell_2$ regularized sparse PSVM. In Sect. 4, we describe a specialized interior-point method which uses a preconditioned conjugate gradient algorithm to compute the search direction. Section 5 gives the experimental results on the datasets taken from UCI repository. Finally, conclusive remarks are given in Sect. 6.

2 Preliminaries

In this section, we briefly review the concepts of SVM and PSVM, respectively.

2.1 SVM for Binary Classification

We start by recalling SVM, which is a learning system and uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [20]. Here, we consider the simplest case of linear binary classification to show how an SVM works.

Given a training set $\{(x_1, y_1), \cdots, (x_l, y_l)\} \subseteq \mathbb{R}^n \times \{-1, +1\}$, it is linearly separable if there exists a hyperplane $w^{\mathrm{T}}x + b = 0$ such that

$$\begin{aligned} w^{\mathrm{T}}x_i + b &\geqslant 1, \quad \text{if} \quad y_i = +1, \\ w^{\mathrm{T}}x_i + b &\leqslant -1, \quad \text{if} \quad y_i = -1, \quad \forall i \in \{1, \cdots, l\}. \end{aligned} \tag{2.1}$$

As shown in Fig. 1, the margin or the distance between the two supporting hyperplane is $\frac{2}{\|w\|_2}$, the solid dots and hollow dots show the points of the "+1" class and "−1"

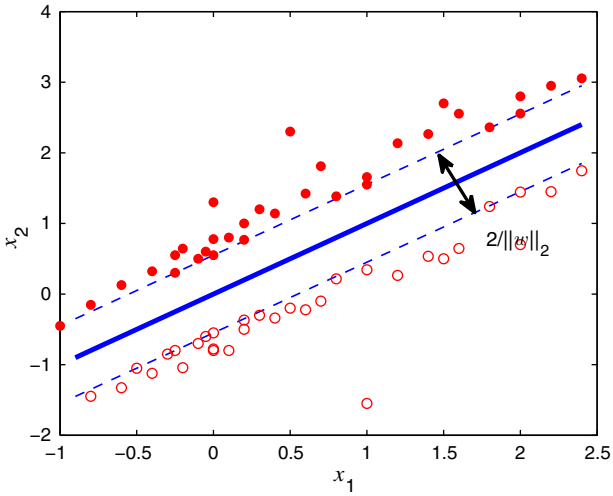

Fig. 1 Hard Margin SVM

class, respectively, while the solid line shows a separation hyperplane $w^{\mathrm{T}}x + b = 0$ between them.

The larger the margin is, the better the separation is. So the SVM aims to find the separation hyperplane which results in the max-margin. This leads to a quadratic programming problem as follows:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y_i \left(w^{\mathrm{T}}x_i + b\right) \geqslant 1, \quad i = 1, \cdots, l. \end{aligned} \tag{2.2}$$

When the set of points is not linearly separable, we generalize the method by relaxing the separability constraints Eq. (2.1). This can be done by introducing positive slack variables $\xi_i, i = 1, \cdots, l$ in the constraints, which then become

$$\begin{aligned} & w^{\mathrm{T}}x_i + b \geqslant 1 - \xi_i, \qquad \text{if} \quad y_i = +1, \\ & w^{\mathrm{T}}x_i + b \leqslant -1 + \xi_i, \qquad \text{if} \quad y_i = -1, \quad \forall i \in \{1, \cdots, l\}. \end{aligned} \tag{2.3}$$

As shown in Fig. 2, it means that the points may locate in the area between the two dashed lines. But each exceeding point must be punished by a misclassification penalty, i.e., an increase in the objective function of Eq. (2.4). Thus, it follows that

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|_2^2 + \frac{C}{2}\sum_{i=1}^{l}\xi_i \\ \text{s.t.} \quad & y_i(w^{\mathrm{T}}x_i + b) \geqslant 1 - \xi_i, \\ & \xi_i \geqslant 0, \quad i = 1, \cdots, l. \end{aligned} \tag{2.4}$$

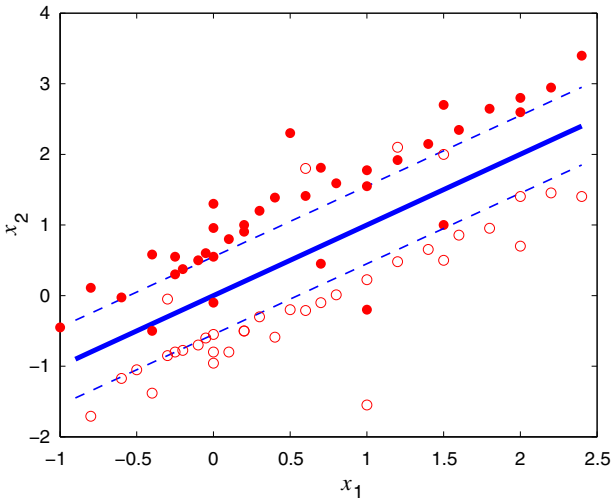

Fig. 2 Soft Margin SVM

2.2 PSVM for Binary Classification

Suppose that a two-class problem of classifying l points in the n-dimensional real space $\mathbb{R}^n$ is considered, the standard PSVM is expressed as follows:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\left(\|w\|_2^2 + b^2\right) + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i\left(w^{\mathrm{T}}x_i + b\right) = 1 - \xi_i, \quad i = 1, \cdots, l, \end{aligned} \tag{2.5}$$

where $C > 0$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $\xi_i \in \mathbb{R}$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$.

The geometric interpretation of Eq. (2.5) is shown in Fig. 3. Compared with Eq. (2.4), this model not only adds advantages such as strong convexity of the objective function, but also changes the nature of optimization problem significantly. The planes $w^{\mathrm{T}}x_i + b = \pm 1$ are not bounding planes any more, but can be thought as "*proximal*" planes, around which points of the corresponding class are clustered and which are pushed as far apart as possible.

3 $\ell_1 - \ell_2$ Regularized Sparse PSVM Model

To obtain a sparse classifier, we add the cardinality constraint $\|w\|_0 \leqslant N$ to the original PSVM model Eq. (2.5), where $\|\cdot\|_0$ denotes the number of non-zero components, $N \in \mathbb{N}^+$. Then a new model is established as follows:

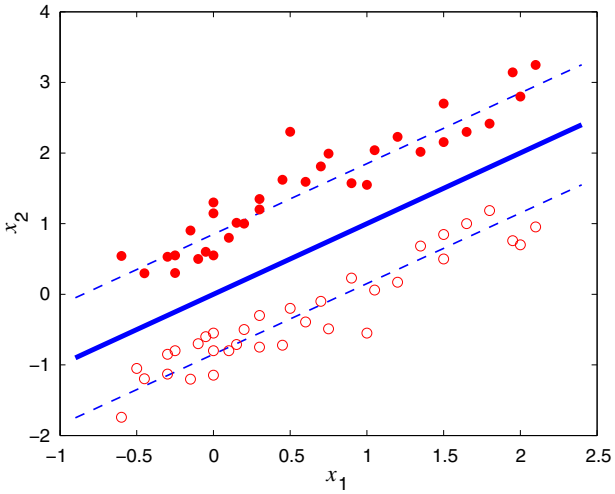

Fig. 3 PSVM

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\left(\|w\|_2^2 + b^2\right) + \frac{C}{2}\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i\left(w^{\mathrm{T}}x_i + b\right) = 1 - \xi_i, \quad i = 1, \cdots, l, \\ & \|w\|_0 \leqslant N, \end{aligned} \tag{3.1}$$

where $C > 0$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $\xi_i \in \mathbb{R}$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$. After adding this cardinality constraint, Eq. (3.1) becomes a combinatorial problem which is difficult to solve.

In the following, we first transform the cardinality constraint into the objective function with a parameter λ. Then we have:

$$\begin{aligned} \min_{w,b,\xi} \quad & \lambda\|w\|_0 + \left(\|w\|_2^2 + b^2\right) + C\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i\left(w^{\mathrm{T}}x_i + b\right) = 1 - \xi_i, \quad i = 1, \cdots, l, \end{aligned} \tag{3.2}$$

where $\lambda \geqslant 0$, $C > 0$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $\xi_i \in \mathbb{R}$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$. In the objective function, the ℓ_0-norm is non-convex, and it is computationally difficult. Not surprisingly, the above problem Eq. (3.2) is also computationally difficult. Furthermore, it is known to be NP-hard.

As the $\|\cdot\|_1$ is, in a certain natural sense, a convexification of the $\|\cdot\|_0$, the following model can be viewed as a convexification of Eq. (3.2), in this paper, we call it an $\ell_1 - \ell_2$ regularized sparse PSVM:

$$\begin{aligned} \min_{w,b,\xi} \quad & \lambda\|w\|_1 + \left(\|w\|_2^2 + b^2\right) + C\sum_{i=1}^{l}\xi_i^2 \\ \text{s.t.} \quad & y_i\left(w^{\mathrm{T}}x_i + b\right) = 1 - \xi_i, \quad i = 1, \cdots, l, \end{aligned} \tag{3.3}$$

where $\lambda \geqslant 0$, $C > 0$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, $\xi_i \in \mathbb{R}$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$. Readers can refer to [5,10] for more details about the relation between ℓ_1-norm and ℓ_0-norm. The objective function of Eq. (3.3) is a trade-off between ℓ_1-norm term and ℓ_2-norm term. The ℓ_2-norm term is responsible for the good classification performance, while the ℓ_1-norm term leads to sparser solutions. When the ith component of w is zero, the ith component of the vector x is irrelevant in deciding the class of x using linear decision function $f(x) = \mathrm{sgn}(w^{\mathrm{T}}x - b)$. The solution of model Eq. (3.3) differs with various parameter λ.

Several solution methods can be used to solve the $\ell_1 - \ell_2$ regularized sparse PSVM Eq. (3.3), for example, the alternating direction method of multipliers (ADMM). Recently, Bai et al. [3] applied ADMM to solve $\ell_1 - \ell_2$ regularized sparse PSVM. Our goal in this paper is to use another well-known solution method: interior-point method to solve Eq. (3.3). Our approach is to convert the $\ell_1 - \ell_2$ regularized sparse PSVM into an equivalent ℓ_1 regularized LS appeared in [12], which is solved by a specialized interior-point method.

Considering the definition of l_2-norm and the constraint $\xi_i = 1 - y_i(w^{\mathrm{T}}x_i + b)$ for $i = 1, \cdots, l$, we can rewrite the objective function as follows:

$$\lambda\|w\|_1 + \left(\|w\|_2^2 + b^2\right) + C\sum_{i=1}^{l}\xi_i^2 = \sum_{i=1}^{l}\left(\sqrt{C}\xi_i\right)^2 + \sum_{i=1}^{n}(w_i)^2 + b^2 + \lambda\|w\|_1$$

$$= \sum_{i=1}^{l}\left(\sqrt{C}\left(y_i x_i^{\mathrm{T}} w + y_i b - 1\right)\right)^2 + \sum_{i=1}^{n} {w_i}^2 + b^2 + \lambda\|w\|_1$$

$$= \left\| \begin{pmatrix} \sqrt{C}\left(y_1 x_1^{\mathrm{T}} w + y_1 b - 1\right) \\ \vdots \\ \sqrt{C}\left(y_l x_l^{\mathrm{T}} w + y_l b - 1\right) \\ e_1^{\mathrm{T}} w \\ \vdots \\ e_l^{\mathrm{T}} w \\ b \end{pmatrix} \right\|_2^2 + \lambda\|w\|_1$$

$$= \left\| \begin{pmatrix} \sqrt{C}y_1 x_1^{\mathrm{T}} \\ \vdots \\ \sqrt{C}y_l x_l^{\mathrm{T}} \\ e_1^{\mathrm{T}} \\ \vdots \\ e_n^{\mathrm{T}} \\ 0 \end{pmatrix} w + \begin{pmatrix} \sqrt{C}y_1 \\ \vdots \\ \sqrt{C}y_l \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} b - \begin{pmatrix} \sqrt{C} \\ \vdots \\ \sqrt{C} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \right\|_2^2 + \lambda\|w\|_1$$

$$= \|Aw + A_1 b - d\|_2^2 + \lambda\|w\|_1,$$

where $A = \begin{pmatrix} \sqrt{C}y_1 x_1^{\mathrm{T}} \\ \vdots \\ \sqrt{C}y_l x_l^{\mathrm{T}} \\ e_1^{\mathrm{T}} \\ \vdots \\ e_n^{\mathrm{T}} \\ 0 \end{pmatrix}_{(l+n+1)\times n}$, $A_1 = \begin{pmatrix} \sqrt{C}y_1 \\ \vdots \\ \sqrt{C}y_l \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}_{(l+n+1)\times 1}$, $d = \begin{pmatrix} \sqrt{C} \\ \vdots \\ \sqrt{C} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}_{(l+n+1)\times 1}$,

$$e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1_{(i)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1}.$$

Therefore, we can convert Eq. (3.3) into an equivalent unconstrained optimization problem with a simple form:

$$\min_{w,b} \quad \|Aw + A_1 b - d\|_2^2 + \lambda \|w\|_1. \tag{3.4}$$

Obviously, we obtain an equivalent expression of $\ell_1 - \ell_2$ regularized sparse PSVM. Therefore, we use a specialized interior-point method to solve Eq. (3.4). The details of the solving process will be described in next section.

4 A Specialized Interior-Point Method

4.1 A Specialized Interior-Point Method and PCG Algorithm

In this section, we use a specialized interior-point method to solve Eq. (3.4), which is the equivalent model of $\ell_1 - \ell_2$ regularized sparse PSVM. We use the preconditioned conjugate gradients (PCG) algorithm to calculate the search direction which is similar to the method in [12]. The objective function of Eq. (3.4) is convex but not differentiable, and we first reformulate it to a convex quadratic problem with linear inequality constraints as follows:

$$\begin{aligned} \min_{w,b,u} \quad & \|Aw + A_1 b - d\|_2^2 + \lambda \sum_{i=1}^{n} u_i \\ \text{s.t.} \quad & -u_i \leqslant w_i \leqslant u_i, \qquad i = 1, \cdots, n, \end{aligned} \tag{4.1}$$

where $w = (w_1 \cdots w_n)$, $b \in \mathbb{R}$, $u \in \mathbb{R}^n$. We define the logarithmic barrier for the bound constraints $-u_i \leqslant w_i \leqslant u_i$,

$$\Phi(w, u) = -\sum_{i=1}^{n} \log(u_i + w_i) - \sum_{i=1}^{n} \log(u_i - w_i),$$

with domain

$$\text{dom}\, \Phi = \{(w, u) \in \mathbb{R}^n \times \mathbb{R}^n \,|\, |w_i| < u_i, \quad i = 1, \cdots, n\}.$$

We augment the weighted objective function by the logarithmic barrier and obtain the following function:

$$\Psi_t(w,u,b) = t\|Aw + A_1 b - d\|_2^2 + t\lambda \sum_{i=1}^{n} u_i + \Phi(w,u),$$

where the parameter $t > 0$. This function is smooth, strictly convex, and bounded below. Newton's method is used in minimizing Ψ_t, i.e., the search direction is computed as the exact solution to the Newton system,

$$H \begin{bmatrix} \triangle b \\ \triangle w \\ \triangle u \end{bmatrix} = -g, \tag{4.2}$$

where H is the Hessian and g is the gradient of Ψ_t at the current iterate (b, w, u). The Hessian can be written as

$$H = t \bigtriangledown^2 \|Aw + A_1 b - d\|_2^2 + \bigtriangledown^2 \Phi(w,u) = \begin{pmatrix} 2tA_1^{\mathrm{T}}A_1 & 2tA_1^{\mathrm{T}}A & 0 \\ 2tA^{\mathrm{T}}A_1 & 2tA^{\mathrm{T}}A + D_1 & D_2 \\ 0 & D_2 & D_1 \end{pmatrix},$$

where

$$D_1 = \mathrm{diag}\left(\frac{2\left(u_1^2 + w_1^2\right)}{\left(u_1^2 - w_1^2\right)^2}, \cdots, \frac{2\left(u_l^2 + w_l^2\right)}{\left(u_l^2 - w_l^2\right)^2} \right),$$
$$D_2 = \mathrm{diag}\left(\frac{-4u_1 w_1}{\left(u_1^2 - w_1^2\right)^2}, \cdots, \frac{-4u_l w_l}{\left(u_l^2 - w_l^2\right)^2} \right).$$

Obviously, H is symmetric and positive definite. The gradient can be written as

$$g = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix},$$

where

$$g_1 = \bigtriangledown_b \Psi_t(b,w,u) = 2tA_1^{\mathrm{T}}(Aw + A_1 b - d),$$
$$g_2 = \bigtriangledown_w \Psi_t(b,w,u) = 2tA^{\mathrm{T}}(Aw + A_1 b - d) + \begin{bmatrix} 2w_1/(u_1^2 - w_1^2) \\ \vdots \\ 2w_n/(u_n^2 - w_n^2) \end{bmatrix},$$
$$g_3 = \bigtriangledown_u \Psi_t(b,w,u) = t\lambda \mathbf{1} - \begin{bmatrix} 2u_1/(u_1^2 - w_1^2) \\ \vdots \\ 2u_n/(u_n^2 - w_n^2) \end{bmatrix}.$$

Solving Newton system Eq. (4.2) exactly may be not computationally practical, so we apply the PCG algorithm to the Newton system to compute the search direction approximately (readers can refer to [12] or Sect. 5 in [15] for more details). In this paper, we choose the preconditioner P as follows:

$$P = \operatorname{diag}(t \bigtriangledown^2 \|Aw + A_1 b - d\|_2^2) + \bigtriangledown^2 \Phi(w, u)$$
$$= \begin{pmatrix} 2t\operatorname{diag}(A_1^{\mathrm{T}} A_1) & 0 & 0 \\ 0 & 2t\operatorname{diag}(A^{\mathrm{T}} A) & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & D_1 & D_2 \\ 0 & D_2 & D_1 \end{pmatrix},$$

which is symmetric and positive definite.

4.2 Dual Gap and Stopping Criteria

To derive a Lagrange dual of the model Eq. (3.4), we first introduce a new variable $z \in \mathbb{R}^{l+n+1}$, as well as new equality constraints $z = Aw + A_1 b - d$, to obtain the equivalent problem:

$$\begin{aligned} \min_{w,b,z} \quad & f(w, b, z) = \|z\|_2^2 + \lambda \|w\|_1 \\ \text{s.t.} \quad & z = Aw + A_1 b - d. \end{aligned} \tag{4.3}$$

Associating dual variables $\nu \in \mathbb{R}^{l+n+1}$ with the equality constraints $z = Aw + A_1 b - d$, the Lagrangian is

$$L(w, b, z, \nu) = \|z\|_2^2 + \lambda \|w\|_1 + \nu^{\mathrm{T}}(Aw + A_1 b - d - z).$$

The objective function of Eq. (4.3) is convex but not differentiable, so we use a first-order optimality condition based on subdifferential calculus, and obtain the dual function:

$$g(\nu) = \inf_{w,b,z} L(w, b, z, \nu) = \begin{cases} -(1/4)\nu^{\mathrm{T}}\nu - \nu^{\mathrm{T}} y, & A_1^{\mathrm{T}}\nu = 0, |A^{\mathrm{T}}\nu|_\infty \leqslant \lambda, \\ -\infty, & \text{otherwise.} \end{cases}$$

The Lagrange dual of Eq. (4.3) is therefore

$$\begin{aligned} \max_{\nu} \quad & G(\nu) \\ \text{s.t.} \quad & A_1^{\mathrm{T}}\nu = 0, \\ & |A^{\mathrm{T}}\nu|_\infty \leqslant \lambda, \end{aligned} \tag{4.4}$$

where the dual objective $G(\nu)$ is

$$G(\nu) = -(1/4)\nu^{\mathrm{T}}\nu - \nu^{\mathrm{T}} y.$$

First, we define $\bar{b} = \arg\min_b \|Aw + A_1 b - d\|_2^2$, then for an arbitrary w, $A_1^{\mathrm{T}}(Aw + A_1\bar{b} - d) = 0$. So $(w, \bar{b}, \bar{z})$ is primal feasible. Next, we define

$$\begin{aligned} \bar{\nu} &= 2s(Aw + A_1\bar{b} - d), \\ s &= \min\{\lambda/2\|Aw + A_1\bar{b} - d\|_\infty, 1\}. \end{aligned} \tag{4.5}$$

Evidently $\bar{\nu}$ is dual feasible and $G(\bar{\nu})$ is lower bound on p^*, which is the optimal value of the model Eq. (3.4).

Here $(w, \bar{b}, \bar{z})$ is primal feasible and $\bar{\nu}$ is corresponding dual feasible. Assume that $\varepsilon_{\text{abs}} > 0$ is a given absolute accuracy, then we have

$$f(w, \bar{b}, \bar{z}) - p^* \leqslant f(w, \bar{b}, \bar{z}) - G(\bar{\nu}).$$

This guarantees that $(w, \bar{b}, \bar{z})$ is a suboptimal ε_{abs} if the stopping criterion $f(w, \bar{b}, \bar{z}) - G(\bar{\nu}) \leqslant \varepsilon_{\text{abs}}$ holds. The difference between the primal objective value of $(w, \bar{b}, \bar{z})$ and the associated lower bound $G(\bar{\nu})$ is called the duality gap. We denote duality gap η as:

$$\eta = f(w, \bar{b}, \bar{z}) - G(\bar{\nu}). \tag{4.6}$$

We always have $\eta \geqslant 0$, and by the weak duality, the point $(w, \bar{b}, \bar{z})$ is no more than η-suboptimal. At the optimal point, we have $\eta = 0$.

4.3 Algorithm

Algorithm: Specialized IPM for $\ell_1 - \ell_2$ Regularized Sparse PSVM

Given relative tolerance $\varepsilon_{\text{rel}} > 0$.

Initialize $t = 1/\lambda$, $w = 0$, $u = (1, \cdots, 1) \in \mathbb{R}^n$, $b = (A_1^{\mathrm{T}} A_1)^{-1} A_1^{\mathrm{T}} d - (A_1^{\mathrm{T}} A_1)^{-1} A_1^{\mathrm{T}} A w$.

Repeat

1. Compute the search direction $(\Delta b, \Delta w, \Delta u)$ as an approximate solution to the Newton system

$$H(\Delta b, \Delta w, \Delta u)^{\mathrm{T}} = -g.$$

2. Compute the step size $s = \beta^k$ by the backtracking line search. Find the smallest integer $k \geqslant 0$, that satisfies

$$\Phi_t(b + \beta^k \Delta b, w + \beta^k \Delta w, u + \beta^k \Delta u) \leqslant \Phi_t(b, w, u) + \alpha \beta^k \nabla \Phi_t(b, w, u)^{\mathrm{T}} (\Delta b, \Delta w, \Delta u)^{\mathrm{T}}.$$

3. Update the iterate by $(b, w, u) = (b, w, u) + s(\Delta b, \Delta w, \Delta u)$.
4. Set $b = \bar{b} = \arg\min_b \|Aw + A_1 b - d\|_2^2$, the optimal value of the intercept.
5. Construct dual feasible point ν from (14).
6. Update t.

$$t = \begin{cases} \max\{\mu \min\{2n/\eta, t\}, t\}, & s \geqslant s_{\min}, \\ t, & s < s_{\min}. \end{cases}$$

Quit if $\eta / G(\nu) \leqslant \varepsilon_{\text{rel}}$.

Remark The typical values for the line search parameters are $\alpha = 0.01$, $\beta = 0.5$. The update rule for t performs well with $\mu = 2$ and $s_{\min} = 0.5$.

5 Numerical Tests

In this section, we take six datasets (including Pima Indians, Heart, Australian, Mushroom, Spambase, and Sonar) from the UCI repository [14] and one synthetic dataset. Table 1 gives their characteristics. We first use the synthetic dataset, which is generated by Gaussian distribution, to test the classification effectiveness of our model. Then we compare the results of $\ell_1 - \ell_2$ regularized sparse PSVM with various λ to demonstrate the properties of the trade-off between the ℓ_1-norm and ℓ_2-norm terms. We also compare the numerical results of $\ell_1 - \ell_2$ regularized sparse PSVM model with GEPSVM, PSVM, and SVM-Light which has been calculated in [13]. At last, we try to figure out the effectiveness of our model in finding sparse solutions and compare it with ℓ_1-PSVM [7]. The $\ell_1 - \ell_2$ regularized sparse PSVM is implemented on a PC with an Intel Core i5, 2.50 GHz CPU, 2.00 GB RAM.

First, we perform our model on synthetic data to demonstrate the classification effectiveness of our model. The results are shown in Fig. 4. The experimental results verify the validity of our model and algorithm.

Table 1 One synthetic dataset and six real-world datasets from UCI Repository

Dataset	Classes	Attributes	Instances
Synthetic dataset	2	2	1 000
Pima Indians	2	8	768
Heart	2	13	270
Australian	2	14	690
Mushroom	2	22	8 124
Spambase	2	57	4 601
Sonar	2	60	208

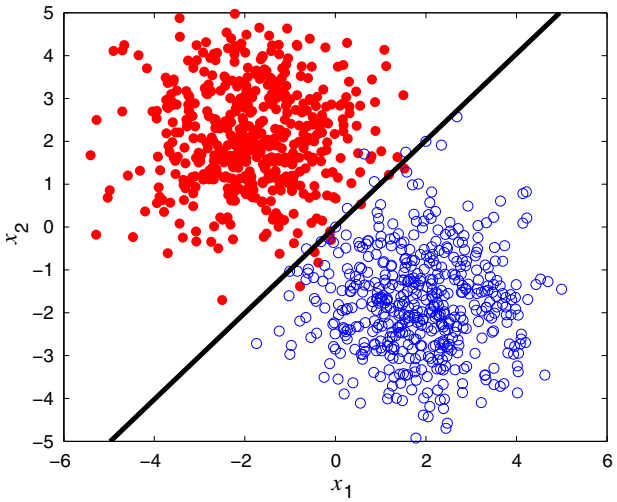

Fig. 4 Performance of the $\ell_1 - \ell_2$ regularized sparse PSVM with synthetic dataset

From the point view of the optimization problem, initially, C is a penalty parameter which penalizes the constraint $\xi > 0$ into the objective function. In the point view of classification, C is also a weight which is used to balance the maximum margin and the minimum classification error. Numerically, we have compared the numerical results of $\ell_1 - \ell_2$ regularized sparse PSVM with various parameter $C \in \{10, 1, 0.1, 0.01\}$, the results show that the model always obtains higher classification accuracy with $C = 0.1$. So we fixed $C = 0.1$ in this section.

For each dataset from UCI repository, to compare the effectiveness of the trade-off between ℓ_1-norm and ℓ_2-norm in $\ell_1 - \ell_2$ regularized sparse PSVM, we choose the parameter $\lambda \in \{0, 1, 10\}$ and $C = 0.1$. The classification accuracy and the execution time are summarized in Tables 2 and 3, respectively.

From Table 2, we observe that for a given C, the classification accuracy of the $\ell_1 - \ell_2$ regularized sparse PSVM performs better when $\lambda = 0$ (the model without ℓ_1-norm term), while from Table 3, we can observe that the model succeeds in decreasing the execution time when $\lambda > 0$ (ℓ_1-norm term is added). The numerical results verify that the ℓ_2-norm term is responsible for the good classification performance while ℓ_1-norm term plays an important role on decreasing the execution time.

To demonstrate the performance of our model, we compare the classification accuracy of the $\ell_1 - \ell_2$ regularized sparse PSVM with that of GEPSVM, PSVM and SVM-Light [13].

The testing accuracy of $\ell_1 - \ell_2$ regularized sparse PSVM is evaluated with a 10-fold cross-validation and the performance is the average misclassification error over 10 folds. In 10-fold cross-validation, the total dataset is divided into ten parts. Each part is chosen once as the test set while the other nine parts form the training set.

Table 4 shows the comparison of classification accuracy. The numbers listed in bold show the better classification accuracy for each dataset. We conclude that $\ell_1 - \ell_2$ regularized sparse PSVM is more efficient than other three SVMs.

Table 2 Classification accuracy of $\ell_1 - \ell_2$ regularized sparse PSVM with varying λ

Dataset	$\lambda = 0$/%	$\lambda = 1$/%	$\lambda = 10$/%
Pima Indians	77.21	76.82	76.82
Heart	86.30	85.56	80.00
Australian	86.67	86.38	86.26
Mushroom	93.17	92.61	86.96
Spambase	88.83	88.70	87.96
Sonar	85.10	83.65	79.33

Table 3 Execution time (sec) of $\ell_1 - \ell_2$ regularized sparse PSVM with varying λ

Dataset	$\lambda = 0$/s	$\lambda = 1$/s	$\lambda = 10$/s
Pima Indians	2.121	0.146	0.163
Heart	1.839	0.273	0.250
Australian	1.803	0.438	0.292
Mushroom	4.693	0.933	1.150
Spambase	7.230	1.588	2.500
Sonar	3.418	1.342	2.054

Finally, we want to test the effect of $\ell_1 - \ell_2$ regularized sparse PSVM in finding sparse solutions and compare it with ℓ_1-PSVM which has been calculated in [7]. Tuning parameter λ, keep the classification accuracy of $\ell_1 - \ell_2$ regularized sparse PSVM approximately equals to the results of ℓ_1-PSVM in [7], then compare the number of zero variables in w. Table 5 shows the numerical results, where $\sharp$ denotes the number of zero variables in w.

From Table 5, we observe that $\ell_1 - \ell_2$ regularized sparse PSVM succeeds in finding more sparse solutions with higher classification accuracy than ℓ_1-PSVM.

6 Conclusions

In this paper, we have proposed a PSVM with cardinality constraint and converted it into an $\ell_1 - \ell_2$ regularized sparse PSVM, then we have solved the equivalent model of the $\ell_1 - \ell_2$ regularized sparse PSVM by a specialized interior-point method with PCG algorithm to compute the search direction. We have implemented the $\ell_1 - \ell_2$ regularized sparse PSVM by a real-world dataset taken from the UCI repository. The classification accuracy and execution time are tested by choosing different parameter λ. The numerical results show that the $\ell_1 - \ell_2$ regularized sparse PSVM outperforms the others with more accuracy for classification. Moreover, it succeeds in finding sparse solutions with higher accuracy than or almost the same as ℓ_1-PSVM.

We have only considered the binary linear classification problems in this paper. Our future research will extend to the multi-class classification and nonlinear classification. Moreover, compared with the performance in paper [3], our numerical results of the execution time are slower than that in [3], where the ADMM was used. We will modify the PCG so as to speed up the solution method.

Table 4 Comparison of classification accuracy

Dataset	$\ell_1 - \ell_2$ PSVM/%	GEPSVM/%	PSVM/%	SVM-Light/%
Pima Indians	**76.82** ($\lambda = 1$)	73.6	75.9	75.7
Spambase	**84.87** ($\lambda = 1$)	76.8	77.1	77.1
Mushroom	**85.62** ($\lambda = 1$)	81.1	80.9	81.5

Table 5 Comparison of the effectiveness in finding sparse solutions

Dataset	$\ell_1 - \ell_2$ PSVM/%	ℓ_1-PSVM/%
Heart	80.00 ($\lambda = 10, \sharp = 5$)	79.63 ($\sharp = 3$)
Australian	86.67 ($\lambda = 15, \sharp = 6$)	85.94 ($\sharp = 2$)
Sonar	76.92 ($\lambda = 16, \sharp = 37$)	75.62 ($\sharp = 36$)

References

[1] Bai, Y.Q., Chen, Y., Niu, B.L.: SDP relaxation for semi-supervised support vector machine. Pac. J. Optim. **8**(1), 3 (2012)
[2] Bai, Y.Q., Niu, B.L., Chen, Y.: New SDP models for protein homology detection with semi-supervised SVM. Optimization **62**(4), 561 (2013)
[3] Bai, Y.Q., Shen, Y.J., Shen, K.J.: Consensus proximal support vector machine for classification problems with sparse solutions. J. Oper. Res. Soc. China **2**, 57–74 (2014)
[4] Cands, E.J.: Compressive sampling. Proc. Int. Congress Math. **3**, 1433–1452 (2006)
[5] Cands, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. Inf. Theory IEEE Trans. **52**(2), 489–509 (2006a)
[6] Cands, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006b)
[7] Chen, W.J., Tian, Y.J.: L_p -norm proximal support vector machine and its applications. Proc. Comput. Sci. **1**, 2417–2423 (2012)
[8] Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
[9] Deng, N.Y., Tian, Y.J., Zhang, C.H.: Support Vector Machines. CRC Press, Taylor and Francis Group, Boca (2013)
[10] Donoho, D.L.: Compressed sensing. Inf. Theory IEEE Trans. **52**(4), 1289–1306 (2006)
[11] Fung, G., Mangasarian, O.L.: Proximal support vector machine classifiers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM, pp. 77–86 (2001)
[12] Kim, S.J., Koh, K., Boyd, S.P.: An interior-point method for large-scale ℓ_1-regularized least square. J. Sel. Topics Signal Process. **12**, 1932–4553 (2007)
[13] Mangasarian, O.L., Wild, E.W.: Multisurface proximal support vector machine classification via generalized eigenvalues. Pattern Anal Mach. Intell. IEEE Trans. **28**(1), 69–74 (2006)
[14] Murphy, P.M., Aha, D.W.: UCI machine learning repository, www.ics.uci.edu/mlearn/MLRepository.html, 1992
[15] Nocedal, J., Wright, S.: Numerical Optimization. Springer Series in Operations Research, New York (2006)
[16] Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process Lett. **9**, 293–300 (1999)
[17] Suykens, J.A.K.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
[18] Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B, 267–288, 1996
[19] Vapnik, V.N., et al.: The Nature of Statistical Learning Theory. Springer, New York (1995)
[20] Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)