Neural Networks 50 (2014) 166–174

Contents lists available at ScienceDirect

Neural Networks

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

Large-scale linear nonparallel support vector machine solver

Yingjie Tian [a], Yuan Ping [b,c,*]

[a] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Computer Science and Technology, Xuchang University, Xuchang 461000, China
[c] Information Security Center, Beijing University of Posts and Telecommunications, Beijing 100876, China

HIGHLIGHTS

- A novel nonparallel linear classifier avoids computing the inverses of matrices.
- Two problems of L_1-NPSVM can be solved by the dual coordinate descent method.
- Linear TWSVMs and linear L_1-SVM are the special cases of linear L_1-NPSVM.
- L_1-NPSVM has the similar sparseness with standard SVMs.
- Results show the superiority of L_1-NPSVM on large-scale problems.

ARTICLE INFO

Article history:
Received 4 December 2012
Received in revised form 10 October 2013
Accepted 17 November 2013

Keywords:
Machine learning
Support vector machines
Classification
Nonparallel support vector machine

ABSTRACT

Twin support vector machines (TWSVMs), as the representative nonparallel hyperplane classifiers, have shown the effectiveness over standard SVMs from some aspects. However, they still have some serious defects restricting their further study and real applications: (1) They have to compute and store the inverse matrices before training, it is intractable for many applications where data appear with a huge number of instances as well as features; (2) TWSVMs lost the sparseness by using a quadratic loss function making the proximal hyperplane close enough to the class itself. This paper proposes a Sparse Linear Nonparallel Support Vector Machine, termed as L_1-NPSVM, to deal with large-scale data based on an efficient solver—dual coordinate descent (DCD) method. Both theoretical analysis and experiments indicate that our method is not only suitable for large scale problems, but also performs as good as TWSVMs and SVMs.

1. Introduction

Support vector machines (SVMs), having their roots in statistical learning theory, are useful for pattern classification (Deng & Tian, 2009; Tian, Shi, & Liu, 2012; Vapnik, 1996, 1998). For a binary classification problem with training set

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

where $x_i \in R^n$, $y_i \in \mathcal{Y} = \{1, -1\}$, $i = 1, \ldots, l$, SVM finds the optimal separating hyperplane by maximizing the margin between two parallel support hyperplanes, which involves the minimization of a quadratic programming problem (QPP)

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

where $\xi = (\xi_1, \ldots, \xi_l)^\top$, and $C \geqslant 0$ is a penalty parameter. This SVM is called L_1-SVM since the L_1-loss function $\xi_i = \max(1-y_i((w\cdot x_i) + b), 0)$ is adopted. For this primal problem, L_1-SVM solves its Lagrangian dual problem

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^\top Q\alpha - e^\top\alpha, \\ \text{s.t.} \quad & 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, l, \end{aligned} \tag{3}$$

where $Q \in R^{l\times l}$, and $Q_{ij} = y_iy_j((x_i\cdot x_j) + 1)$. It is also a QPP. An SVM usually maps the training set into a high-dimensional space via a nonlinear function $\phi(x)$, then the kernel function $K(x, x')$ is applied to take instead of the inner product $(\phi(x)\cdot\phi(x'))$, such SVM is called a nonlinear SVM.

* Corresponding author at: Department of Computer Science and Technology, Xuchang University, Xuchang 461000, China. Tel.: +86 15937491700; fax: +86 0374 2968890.
E-mail addresses: tyj@ucas.ac.cn (Y. Tian), pingyuan@bupt.edu.cn, pyuan.lhn@xcu.edu.cn (Y. Ping).

0893-6080/$ – see front matter
http://dx.doi.org/10.1016/j.neunet.2013.11.014

However, in 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 (Bottou, 2007; Chang, Hsieh, & Lin, 2008; Chang & Lin, 2001; Collins, Globerson, Koo, Carreras, & Bartlett, 2008; Hsieh, Chang, Lin, Keerthi, & Sundararajan, 2008; Joachims, 2006; Keerthi & DeCoste, 2005; Lin, Weng, & Keerthi, 2008; Shalev-Shwartz, Singer, & Srebro, 2011; Smola, Vishwanathan, & Le, 2008; Zhang, 2004).

Recently, some nonparallel hyperplane classifiers have been proposed (Jayadeva, Khemchandani, & Chandra, 2007; Mangasarian & Wild, 2006). For the twin support vector machine (TWSVM) (Jayadeva et al., 2007), 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. Experimental results (Jayadeva et al., 2007; Kumar & Gopal, 2008) have shown the effectiveness of TWSVM over the standard SVM on UCI data sets. Furthermore, it is implemented by solving two QPPs smaller than the problem (3), which increases the TWSVM's training speed by approximately fourfold compared with that of SVM. TWSVMs have been studied extensively (Khemchandani, Jayadeva, & Chandra, 2009; Kumar & Gopal, 2009; Peng, 2010; Qi, Tian, & Shi, 2012, 2013; Qi, Tian, & Yong, 2012a, 2012b; Shao, Zhang, Wang, & Deng, 2011).

However, existing TWSVMs have two serious defects which restrict their further studies and real applications: (1) Although TWSVMs solve two smaller QPPs and can be solved by successive overrelaxation (SOR) technique (Shao et al., 2011), they have to compute the inverse of matrices before training, it is in practice intractable for a large dataset; (2) TWSVMs lost the sparseness by using a quadratic loss function making the proximal hyperplane close enough to the class itself.

In this paper, for linear classification issues, we propose a novel nonparallel linear classifier, termed as linear L_1-NPSVM, to solve very large linear problems. Our L_1-NPSVM has incomparable advantages including: (1) The two problems constructed have the elegant formulation and can be solved efficiently by the dual coordinate descent (DCD) method, more importantly, we do not need to compute the inverses of the large matrices any more before training; (2) It has the valuable sparseness similar with the standard SVMs; (3) L_1-NPSVM degenerates to TWSVMs when the corresponding parameters are chosen, and L_1-SVM is a special case of L_1-NPSVM.

The paper is organized as follows. Section 2 briefly introduces the initial TWSVM and its improved edition TBSVM (Twin Bounded Support Vector Machine) (Shao et al., 2011). Section 3 proposes the linear L_1-NPSVM and its corresponding multi-class model, then its efficient solver—DCD method is proposed. Section 4 deals with experimental results and Section 5 contains concluding remarks.

2. Background

In this section, we briefly introduce two variations of the TWSVM.

2.1. TWSVM

Consider the binary classification problem with the training set

$$T = \{(x_1, +1), \dots, (x_p, +1), (x_{p+1}, -1), \dots, (x_{p+q}, -1)\}, \tag{4}$$

where $x_i \in R^n$, $i = 1, \dots, p+q$. For the linear case, TWSVM (Jayadeva et al., 2007) seeks two nonparallel hyperplanes

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

by solving two QPPs

$$\begin{aligned} \min_{w_+, b_+, \xi_-} \quad & \frac{1}{2}\sum_{i=1}^{p}((w_+ \cdot x_i) + b_+)^2 + c_1 \sum_{j=p+1}^{p+q} \xi_j, \\ \text{s.t.} \quad & (w_+ \cdot x_j) + b_+ \leqslant -1 + \xi_j, \\ & j = p+1, \dots, p+q, \\ & \xi_j \geqslant 0, \quad j = p+1, \dots, p+q, \end{aligned} \tag{6}$$

and

$$\begin{aligned} \min_{w_-, b_-, \xi_+} \quad & \frac{1}{2}\sum_{i=p+1}^{p+q}((w_- \cdot x_i) + b_-)^2 + c_2 \sum_{j=1}^{p} \xi_j, \\ \text{s.t.} \quad & (w_- \cdot x_j) + b_- \geqslant 1 - \xi_j, \quad j = 1, \dots, p, \\ & \xi_j \geqslant 0, \quad j = 1, \dots, p, \end{aligned} \tag{7}$$

where $c_i,\ i = 1, 2$ are the penalty parameters. The solutions (w_+, b_+) and (w_-, b_-) are derived by solving their dual problems

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^\top G(H^\top H)^{-1}G^\top \alpha - e_2^\top \alpha, \\ \text{s.t.} \quad & 0 \leqslant \alpha \leqslant c_1 e_2 \end{aligned} \tag{8}$$

and

$$\begin{aligned} \min_{\gamma} \quad & \frac{1}{2}\gamma^\top H(G^\top G)^{-1}H^\top \gamma - e_1^\top \gamma, \\ \text{s.t.} \quad & 0 \leqslant \gamma \leqslant c_2 e_1 \end{aligned} \tag{9}$$

where $\alpha = (\alpha_1, \dots, \alpha_q)^\top \in R^q$, $\gamma = (\gamma_1, \dots, \gamma_p)^\top \in R^p$, $H = [A, e_1] \in R^{p\times(n+1)}$, $G = [B, e_2] \in R^{q\times(n+1)}$, $e_1 = (1, \dots, 1)^\top \in R^p$, $e_2 = (1, \dots, 1)^\top \in R^q$, $A = (x_1, x_2, \dots, x_p)^\top \in R^{p\times n}$, and $B = (x_{p+1}, x_{p+2}, \dots, x_{p+q})^\top \in R^{q\times n}$.

We can see that TWSVM solves two smaller QPPs, which claims 4 times faster than the standard SVM (Jayadeva et al., 2007). Unfortunately, it needs to compute and store the inverse matrices $(H^\top H)^{-1}$ and $(G^\top G)^{-1}$ before training. Since both $H^\top H$ and $(G^\top G)^{-1}$ are all of order $n + 1$, TWSVM fails frequently in dealing with problems of high dimensions, such as document classification. Furthermore, in order to deal with the case when $H^\top H$ or $G^\top G$ is singular and avoid the possible ill conditioning, the inverse matrices $(H^\top H)^{-1}$ and $(G^\top G)^{-1}$ are approximately replaced by $(H^\top H + \epsilon I)^{-1}$ and $(G^\top G + \epsilon I)^{-1}$ respectively, where I is an identity matrix of appropriate dimensions, ϵ is a positive and small scalar to keep the structure of data. After solving the dual problems (8) and (9), the solutions of problems (6) and (7) can be obtained by

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

$$(w_-^\top, b_-)^\top = -(G^\top G)^{-1}H^\top \gamma. \tag{11}$$

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

$$Class = \arg\min_{k=-,+} |(w_k \cdot x) + b_k|, \tag{12}$$

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

For the nonlinear case, two kernel-generated surfaces instead of hyperplanes are considered and two other primal problems different with problems (6) and (7) are constructed, which can refer to Jayadeva et al. (2007).

2.2. TBSVM

An improved version of TWSVM, termed as TBSVM, is proposed in Shao et al. (2011) whereas the structural risk is claimed to be minimized by adding a regularization term with the idea of