Neurocomputing 138 (2014) 114–119

Contents lists available at ScienceDirect

Neurocomputing

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

Large-scale linear nonparallel support vector machine solver

Yingjie Tian [a], Qin Zhang [a], Yuan Ping [b,*]

[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

ARTICLE INFO

Article history:
Received 4 May 2013
Received in revised form
10 October 2013
Accepted 4 February 2014
Communicated by H. Yu
Available online 8 April 2014

Keywords:
Support vector machines
Nonparallel
Large-scale
Classification
Dual coordinate descent

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 one serious defect restricting their further study and real applications: they have to compute and store the inverse matrices before training, it is intractable for many applications such as that data appear with a huge number of instances as well as features. This paper proposes a Linear Nonparallel Support Vector Machine, termed as L_2-TWSVM, 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 has better generalization performance than linear TWSVMs and linear SVMs.

1. Introduction

Support vector machines (SVMs), having their roots in statistical learning theory, are useful for pattern classification [1–4]. For the binary classification problem with the training set

$$T=\{(x_1,y_1),\ldots,(x_l,y_l)\}, \tag{1}$$

where $x_i\in R^n, y_i\in\{1,-1\}$, $i=1,\ldots,l$, 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):

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

which is called the L_1-SVM since the L_1-loss function $\xi=\max(1-y_i((w\cdot x_i)+b),0)$ is applied, while L_2-SVM solves

$$\begin{aligned}&\min_{w,b,\xi}\ \frac{1}{2}(\|w\|^2+b^2)+C\sum_{i=1}^{l}\xi_i^2,\\&\text{s.t.}\ \ y_i((w\cdot x_i)+b)\geq 1-\xi_i,\ i=1,\ldots,l,\\&\xi_i\geq 0,\ i=1,\ldots,l,\end{aligned} \tag{3}$$

since the L_2-loss function $\xi_i^2=(\max(1-y_i((w\cdot x_i)+b),0))^2$ is applied.

For this primal problem, L_2-SVM solves its Lagrangian dual problem:

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\alpha^\top\overline{Q}\alpha-e^\top\alpha,\\&\text{s.t.}\ \ \alpha_i\geq 0,\ i=1,\ldots,l,\end{aligned} \tag{4}$$

where $\overline{Q}=Q+D$, D is a diagonal matrix, and $Q_{ij}=y_iy_j(\tilde{x}_i\cdot\tilde{x}_j)$, and $D_{ii}=1/(2C)$, here $\tilde{x}_i=(x_i^\top,1)^\top$.

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. However, in some applications such as document classification with the data appearing in a rich dimensional feature space, linear SVM in which the data are not mapped, has the similar performances with nonlinear SVM. For linear SVM, many methods have been proposed in large-scale scenarios [5–15].

Recently, some nonparallel hyperplane classifiers have been proposed [16,17]. For the twin support vector machine (TWSVM) [17], 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 [17,18] have shown the effectiveness of TWSVM over standard SVM on UCI datasets. Furthermore, it is implemented by solving two smaller QPPs than problem (4) which increases the TWSVM training speed by approximately fourfold compared to that of SVM. TWSVMs have been studied extensively [19–25].

However, existing TWSVMs have one serious defect which restricts their further study and real applications: although TWSVMs

* Corresponding author.
E-mail addresses: pingyuan@bupt.edu.cn, pingyuan@bupt.edu.cn (Y. Ping).

http://dx.doi.org/10.1016/j.neucom.2014.02.032
0925-2312/

solve two smaller QPPs and can be solved by successive over-relaxation (SOR) technique [22] or dual coordinate descent (DCD) method [26], they have to compute the inverse of matrices before training, it is in practice intractable for a large dataset.

In this paper, for linear classification problems, we propose a novel linear nonparallel twin support vector machine, termed as linear L_2-TWSVM, for solving very large linear problems. Our L_2-TWSVM has two incomparable advantages including (1) the two problems constructed have the dual problems with elegant formulation and can be solved efficiently by the DCD method, more importantly, we do not need to compute and store the large inverse matrices any more before training; (2) linear TWSVMs and linear L_2-SVM are the special cases of linear L_2-TWSVM, which guarantees theoretically that the linear L_2-TWSVM has superior generalization ability to linear TWSVMs linear L_2-SVM.

The remainder of this paper is arranged as follows: Section 2 briefly introduces the TWSVM and its improved edition TBSVM. Section 3 proposes the linear L_2-TWSVM, then its efficient solver – DCD method. Section 4 details the evaluations with respect to accuracy and efficiency, and Section 5 draws the conclusion.

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

$$T=\{(x_1,+1),\ldots,(x_p,+1),(x_{p+1},-1),\ldots,(x_{p+q},-1)\}, \tag{5}$$

where $x_i\in R^n, i=1,\ldots,p+q$. For the linear case, TWSVM [17] seeks two nonparallel hyperplanes

$$(w_+\cdot x)+b_+=0 \quad \text{and} \quad (w_-\cdot x)+b_-=0 \tag{6}$$

by solving two QPPs

$$\begin{aligned}&\min_{w_+,b_+,\xi_-} \ \frac{1}{2}\sum_{i=1}^{p}((w_+\cdot x_i)+b_+)^2+c_1\sum_{j=p+1}^{p+q}\xi_j,\\ &\text{s.t.} \ \ (w_+\cdot x_j)+b_+\le -1+\xi_j, j=p+1,\ldots,p+q,\\ &\xi_j\ge 0,\ j=p+1,\ldots,p+q,\end{aligned} \tag{7}$$

and

$$\begin{aligned}&\min_{w_-,b_-,\xi_+} \ \frac{1}{2}\sum_{i=p+1}^{p+q}((w_-\cdot x_i)+b_-)^2+c_2\sum_{j=1}^{p}\xi_j,\\ &\text{s.t.} \ \ (w_-\cdot x_j)+b_-\ge 1-\xi_j,\ j=1,\ldots,p,\\ &\xi_j\ge 0,\ j=1,\ldots,p,\end{aligned} \tag{8}$$

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

$$\begin{aligned}&\min_{\alpha} \ \frac{1}{2}\alpha^\top G(H^\top H)^{-1}G^\top\alpha-e_2^\top\alpha,\\ &\text{s.t.} \ \ 0\le\alpha\le c_1e_2\end{aligned} \tag{9}$$

and

$$\begin{aligned}&\min_{\gamma} \ \frac{1}{2}\gamma^\top H(G^\top G)^{-1}H^\top\gamma-e_1^\top\gamma,\\ &\text{s.t.} \ \ 0\le\gamma\le c_2e_1\end{aligned} \tag{10}$$

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

We can see that TWSVM solves two smaller QPPs, which claims four times faster than standard SVM [17]. Unfortunately, it needs to compute and store the inverse matrices $(H^\top H)^{-1}$ and $(G^\top G)^{-1}$ before training. Since $H^\top H$ and $(G^\top G)^{-1}$ are all of order $n+1$, therefore TWSVM will fail for the problems with high dimensions, such as document classification with the data appearing in a rich dimensional feature space. 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 scalar, small to keep the structure of data. After solving the dual problems (9) and (10), the solutions of problems (7) and (8) can be obtained by

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

$$(w_-^\top,b_-)^\top=-(G^\top G)^{-1}H^\top\gamma. \tag{12}$$

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

$$Class=\arg\min_{k=-,+}|(w_k\cdot x)+b_k|, \tag{13}$$

where $|\cdot|$ is the perpendicular 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 from problems (7) and (8) are constructed, which can be referred to [17].

2.2. TBSVM

An improved TWSVM, termed as TBSVM, is proposed in [22] whereas the structural risk is claimed to be minimized by adding a regularization term with the idea of maximizing some margin. For the linear case, they solve the following two primal problems:

$$\begin{aligned}&\min_{w_+,b_+,\xi_-} \ \frac{c_3}{2}(\|w_+\|^2+b_+^2)+\frac{1}{2}\sum_{i=1}^{p}((w_+\cdot x_i)+b_+)^2+c_1\sum_{j=p+1}^{p+q}\xi_j,\\ &\text{s.t.} \ \ (w_+\cdot x_j)+b_+\le -1+\xi_j,\ j=p+1,\ldots,p+q,\\ &\xi_j\ge 0,\ j=p+1,\ldots,p+q,\end{aligned} \tag{14}$$

and

$$\begin{aligned}&\min_{w_-,b_-,\xi_+} \ \frac{c_4}{2}(\|w_-\|^2+b_-^2)\\ &\quad+\frac{1}{2}\sum_{i=p+1}^{p+q}((w_-\cdot x_i)+b_-)^2+c_2\sum_{j=1}^{p}\xi_j,\\ &\text{s.t.} \ \ (w_-\cdot x_j)+b_-\ge 1-\xi_j,\ j=1,\ldots,p,\\ &\xi_j\ge 0,\ j=1,\ldots,p.\end{aligned} \tag{15}$$

Their dual problems are

$$\begin{aligned}&\min_{\alpha} \ \frac{1}{2}\alpha^\top G(H^\top H+c_3I)^{-1}G^\top\alpha-e_2^\top\alpha,\\ &\text{s.t.} \ \ 0\le\alpha\le c_1e_2\end{aligned} \tag{16}$$

and

$$\begin{aligned}&\min_{\gamma} \ \frac{1}{2}\gamma^\top H(G^\top G+c_4I)^{-1}H^\top\gamma-e_1^\top\gamma,\\ &\text{s.t.} \ \ 0\le\gamma\le c_2e_1.\end{aligned} \tag{17}$$

Different from problems (9) and (10) with the possibility that $H^\top H$ or $G^\top G$ is singular, problems (16) and (17) are derived without any extra assumption and need not be modified any more. From this point of view, TBSVM is more rigorous and complete than TWSVM. However, TBSVM still need to compute and store the inverse matrices $(H^\top H+c_3I)^{-1}$ and $(G^\top G+c_4I)^{-1}$. More unfortunately, for different c_3 and c_4, they have to compute different inverse matrices. It costs a huge amount of computation. For the nonlinear case, similar to TWSVM two kernel-generated surfaces instead of hyperplanes are considered and two other regularized primal problems are constructed.