Neurocomputing 129 (2014) 127–135

Contents lists available at ScienceDirect

Neurocomputing

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

An inexact smoothing-type algorithm for support vector machines

Tie Ni [a], Wei-Zhe Gu [b,*], Jun Zhai [c]

[a] College of Business Administration, Liaoning Technical University, Huludao 125105, PR China
[b] Department of Mathematics, School of Science, Tianjin University, Tianjin 300072, PR China
[c] Department of Mathematics, The University of Chicago, Chicago 60615, USA

ARTICLE INFO

Article history:
Received 22 November 2012
Received in revised form
16 June 2013
Accepted 11 September 2013
Communicated by Xianbin Cao
Available online 12 November 2013

Keywords:
Support vector machines
Non-monotone line search
Smoothing-type algorithm
Global convergence
Superlinear convergence

ABSTRACT

The smoothing-type algorithm has been successfully applied to solve various optimization problems. In this paper, we propose an inexact smoothing-type algorithm for solving the generalized support vector machines based on a new class of smoothing functions. In general, the smoothing-type method is designed based on some monotone line search and solving a linear system of equations exactly at each iteration. However, for the large-scale problems, solving the linear system of equations exactly can be very expensive. In order to overcome these drawbacks, solving the linear system of equations inexactly and the non-monotone line search technique are used in our smoothing-type method. We show that the proposed algorithm is globally and locally superlinearly convergent under suitable assumptions. Preliminary numerical results are also reported.

1. Introduction

Support Vector Machines (SVMs) have been proved to be a successful learning machine in the literature, especially for classification. The SVM is based on statistical learning theory developed by Vapnik [29,30] and has demonstrated outstanding performance in many applications.

We consider the problem of classifying m points in the n-dimensional real space $\Re^n$, represented by $m \times n$ matrix A, according to membership of each point A_i in the class $+1$ or -1 as specified by a given $m \times m$ diagonal matrix D with ones or minus ones along its diagonal. The linear support vector machine attempts to separate these finite points with a hyperplane such that the separation margin is maximized. We aim at determining a hyperplane $\mathcal{H}(\mathbf{w}, b) = \{x | \mathbf{w}^T x = b\}$, where $\mathbf{w} \in \Re^n \backslash \{\mathbf{0}\}$ and $b \in \Re$, which separates the two classes of points.

For the problem considered above, from [16,4] an optimization model to maximize the separation can be expressed as

$$\min_{\mathbf{w},b} \ \tfrac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t.} \quad D(A\mathbf{w} - be) \ge e, \tag{1.1}$$

where $e = (1, \dots, 1)^T \in \Re^n$.

The optimization model (1.1) is called maximum hard-margin method, it requires the perfect separation. However, in most cases such hyperplane may not exist. In other words, some training points that do not satisfy constraints in (1.1) may be allowed to exist. In this case, we introduce a slack vector $\xi = (\xi_1, \dots, \xi_n)^T$ to represent violations of constraints in (1.1), i.e.,

$$\xi := \max\{0, e - D(A\mathbf{w} - be)\}.$$

The corresponding optimization model can be written as

$$\min_{\mathbf{w},b,\xi} \ \tfrac{1}{2}\|\xi\|^2$$
$$\text{s.t.} \quad D(A\mathbf{w} - be) + \xi \ge e,$$
$$\xi \ge 0, \tag{1.2}$$

where the 2-norm of the misclassification error is minimized. Other norms are also used for the misclassification error, which lead to other problem formulations. When the 2-norm is used, the constraint of $\xi \ge 0$ becomes so unnecessary that is dropped from the subsequent discussions.

Thus, combining (1.1) with (1.2) and using a parameter $\nu > 0$ that trades off the two competing goals, maximizing the separation margin and minimizing the misclassification error. The resulting optimization model, called a soft-margin support vector machine, is

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\nu}{2}\|\xi\|^2$$
$$\text{s.t.} \quad D(A\mathbf{w} - be) + \xi \ge e. \tag{1.3}$$

It is obvious that (1.3) is a convex quadratic programming with respect to the variables $(\mathbf{w}, b, \xi)$.

* Corresponding author. Tel.: +86 13652112869.
E-mail addresses: nitie_4108@yahoo.com.cn (T. Ni), guweizhe@tju.edu.cn (W.-Z. Gu), junzhai0516@uchicago.edu (J. Zhai).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.09.047

From [32,17], the Lagrange dual of (1.3) is

$$\begin{aligned}\min_{x}\quad & \frac{1}{2}x^T DAA^T Dx + \frac{1}{2\nu}x^T x - e^T x\\ \text{s.t.}\quad & e^T Dx = 0,\\ & x \geq 0,\end{aligned} \tag{1.4}$$

It is not difficult to see that (1.4) is a strictly convex quadratic programming with $\nu > 0$.

2. Support vector machines with non-linear kernel

In [18,15], a non-linear separating surface was generated by using a completely arbitrary kernel. So we can obtain the generalized support vector machine by applying a general kernel $K(A, A^T)$ on model (1.3):

$$\begin{aligned}\min_{\mathbf{u},b,\xi}\quad & f(\mathbf{u}) + \frac{\nu}{2}\|\boldsymbol{\xi}\|^2\\ \text{s.t.}\quad & D(K(A, A^T)D\mathbf{u} - be) + \boldsymbol{\xi} \geq e,\end{aligned} \tag{2.1}$$

where f is some convex function in $\Re^m$. The corresponding non-linear separating surface is

$$K(x^T, A^T)Du = b, \tag{2.2}$$

and the corresponding dual is expressed as

$$\begin{aligned}\min_{x}\quad & \tfrac{1}{2}x^T DK(A, A^T)Dx + \tfrac{1}{2\nu}x^T x - e^T x\\ \text{s.t.}\quad & e^T Dx = 0,\\ & x \geq 0.\end{aligned} \tag{2.3}$$

Actually, if we let $K(x^T, A^T) = x^T A^T, \mathbf{w} = A^T D\mathbf{u}$, then (2.2) reduces to $\mathbf{w}^T x = b$. Furthermore, if we set $f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T DAA^T D\mathbf{u}$ and $K(A, A^T) = AA^T$, then (2.1) becomes (1.3) and (2.7) reduces (1.4).

Various numerical optimization methods for solving SVMs have been proposed in the literature, such as, the interior point method [3,31], semismooth method [4], smoothing method [15], the active-set method [20] and so on. The smoothing-type method we will propose can be roughly described as follows: by introducing a new class of smoothing functions, the model (2.3) considered are reformulated as a system of smoothing equations, and hence can be solved by some classical Newton type methods at each iteration and make the smoothing parameter orderly tend to zero as the iteration is going on so that a solution of the original problem can be obtained. In this kind of algorithm, the initial point and the intermediate iteration points are not required to stay in the positive orthant, which is an outstanding difference of this kind of method from interior point algorithms. The smoothing reformulation system can be solved without a sophisticated solver.

The paper is organized as follows. In Section 3, the model (2.3) is reformulated as a system of smoothing equations based a new smoothing reformulation. In Section 4, we propose a general framework of a new smoothing-type algorithm and show that the proposed smoothing-type algorithm is well defined and gives some useful properties of algorithm. In Section 5, we prove the boundedness of the iteration sequence, and then discuss the convergence behavior of the proposed smoothing-type algorithm. Numerical results are reported in Section 6. Some final remarks are given in Section 7.

In our notation, all vectors are column vectors, the subscript T denotes transpose, $\Re^n$ denotes the space of n dimensional real column vectors, and $\Re^n_+$ (respectively, $\Re^n_{++}$) denotes the non-negative (respectively, positive) orthant in $\Re^n$. We use "$:=$" to mean "define" and denote $\mathcal{I} := \{1, 2, \ldots, n\}$. For any vector u, we denote by u_i the ith component of u. A column vector of ones of appropriate dimension will be denoted by e. We denote by $\|u\|$ the 2-norm of u. For any vectors $u, v \in \Re^n$, we write $(u^T, v^T)^T$ as (u, v) for simplicity. We denote by $\Re^{m\times n}$ the space of $m \times n$ real matrices. For $A \in \Re^{m\times n}$, A_i denotes the ith row of A which is a row vector in $\Re^n$. We use N to denote the set of all non-negative integers, i.e., $N := \{0, 1, 2, \ldots,\}$. For any vector $u \in \Re^n$, we also use vec $\{u_i : i \in \mathcal{I}\}$ to denote the vector u, and use diag $\{u_i : i \in \mathcal{I}\}$ to denote the diagonal matrix whose ith diagonal element is u_i. For any $\alpha, \beta \in \Re_+$, $\alpha = O(\beta)$ (or $\alpha = o(\beta)$) means $\limsup_{\beta\to 0}\alpha/\beta < +\infty$ (or $\limsup_{\beta\to 0}\alpha/\beta = 0$). Let $k \geq 0$ denote the iteration index. For any $(\mu, x, y, \gamma), (\mu_k, x^k, y^k, \gamma_k) \in \Re^{1+m+m+1}$, we always use the following notation throughout this paper:

$$\begin{aligned}& z := (\mu, x, y, \gamma), \quad w := (x, \gamma), \quad v := (x, y, \gamma), \quad z := (\mu, v),\\ & z^k := (\mu_k, x^k, y^k, \gamma_k), \quad w^k := (x^k, \gamma_k),\\ & v^k := (x^k, y^k, \gamma_k), \quad z^k := (\mu_k, v^k).\end{aligned}$$

3. Smoothing reformulation based on a new family of smoothing function

The KKT conditions for (2.3) are the following system:

$$\begin{aligned}& x \geq 0, \left(\frac{1}{\nu}I + DK(A, A^T)D\right)x - De\gamma - e \geq 0,\\ & x^T\left(\left(\frac{1}{\nu}I + DK(A, A^T)D\right)x - De\gamma - e\right) = 0,\\ & e^T Dx = 0,\end{aligned} \tag{3.1}$$

where γ is Lagrange multiplier corresponding to the equality constraint $e^T Dx = 0$.

Let $y = (1/\nu I + DK(A, A^T)D)x - De\gamma - e$, the system (3.1) can be equivalently transformed into the following equation system:

$$\begin{aligned}& x \geq 0, \ y \geq 0, \ \text{and } x^T y = 0,\\ & e^T Dx = 0.\end{aligned} \tag{3.2}$$

It is not difficult to see that (3.2) is a mixed complementarity problem. Furthermore, the first row of (3.2) are called complementarity condition. From [5] we know that the complementarity conditions can be replaced by an equation with one NCP-function. The following functions are commonly used NCP-functions:

$$\begin{aligned}& \phi_{FB}(a, b) = \sqrt{a^2 + b^2} - a - b = 0 \Longleftrightarrow a \geq 0, \ b \geq 0, \ ab = 0,\\ & \phi_{\min}(a, b) = \min(a, b) = 0 \Longleftrightarrow a \geq 0, \ b \geq 0, \ ab = 0.\end{aligned}$$

Thus, by reformulation of NCP-functions(in short for $\phi_{FB/min}$), system (3.2) is equivalent to the following non-smooth equations:

$$F(x, y, \gamma) := \begin{bmatrix} y - (\frac{1}{\nu}I + DK(A, A^T)D)x + De\gamma + e\\ \Phi(x, y)\\ e^T Dx\end{bmatrix} = 0, \tag{3.3}$$

where $\Phi(x, y) := (\phi_{FB/min}(x_1, y_1), \ldots, \phi_{FB/min}(x_m, y_m))^T$. Since the function F is not differentiable, some classical iterative methods such as Newton-type methods cannot be directly applied to solve system (3.3). In order to overcome these drawbacks, what we want to do is to make system (3.3) smooth before we can use some classical Newton-type methods. Based on this idea above, in this paper, we propose a new smoothing-type algorithm for solving this SVM, which is one of Newton-type methods.

It is evident that smoothing functions play a very important role in the smoothing-type algorithms. Many smoothing functions have been proposed in the literature. Throughout this paper, we introduce a new class of smoothing functions with one-parameter:

$$\begin{aligned}\phi_\theta(\mu, a, b) = & [\mu a + (1 + \theta\mu)b] + [(1 + \theta\mu)a + \mu b]\\ & - \sqrt{(1 + \theta\mu - \mu)^2(a - b)^2 + 4\mu^2},\end{aligned} \tag{3.4}$$

for any $(\mu, a, b) \in \Re^3$, where $\theta \geq 0$ is a finite number. Fig. 1 intuitively illustrates (3.4) with $\theta = 0.5, \mu = 0.1$ in both two-dimensional and three-dimensional ways. The new class of smoothing functions