Knowledge-Based Systems 73 (2015) 276–288

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Weighted linear loss twin support vector machine for large-scale classification

Yuan-Hai Shao [a], Wei-Jie Chen [a], Zhen Wang [b], Chun-Na Li [a], Nai-Yang Deng [c,*]

[a] Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, PR China
[b] School of Mathematical Sciences, Inner Mongolia University, Hohehot 010021, PR China
[c] College of Science China Agricultural University, Beijing 100083, PR China

ARTICLE INFO

Article history:
Received 7 February 2014
Received in revised form 8 September 2014
Accepted 11 October 2014
Available online 18 October 2014

Keywords:
Pattern recognition
Support vector machines
Twin support vector machines
Large-scale classification
Weighted linear loss function

ABSTRACT

In this paper, we formulate a twin-type support vector machine for large-scale classification problems, called weighted linear loss twin support vector machine (WLTSVM). By introducing the weighted linear loss, our WLTSVM only needs to solve simple linear equations with lower computational cost, and meanwhile, maintains the generalization ability. So, it is able to deal with large-scale problems efficiently without any extra external optimizers. The experimental results on several benchmark datasets indicate that, comparing to TWSVM, our WLTSVM has comparable classification accuracy but with less computational time.

1. Introduction

Traditional support vector machines (SVMs) [5,3,6] such as C-SVC [3] and ν-SVC [32], have already reached many achievements in supervised machine learning [20,15,45,43]. However, their training stage involves solving a quadratic programming problem (QPP) with rather high computational complexity $O(m^3)$, where m is the total size of training data points. This drawback restricts the application of SVM to large-scale problems. There are two ways to address this problem. One aims at solving the QPP in the traditional SVMs more efficiently, e.g. Chunking [42], SMO [24], SVM-Light [12], LIBSVM [4], and LIBLINEAR [8]. The other one aims at establishing new model and finding simpler problem to replace the QPP, e.g. Proximal SVM [9] and Least Squares SVM [39,38], where the QPP is replaced by a linear system of equations since the squared loss function instead of the hinge one is introduced.

It should be mentioned that in the second way there is an interesting approach where two non-parallel hyperplanes are constructed, rather than constructing two parallel supporting hyperplanes in traditional SVMs. It goes back to generalized eigenvalue proximal support vector machine (GEPSVM) [16] which needs to solve generalized eigenvalue problems. Subsequently, the twin support vector machine (TWSVM) [11] is proposed. Different from GEPSVM, TWSVM solves two small related QPPs. Due to its strong generalization ability [13,33], TWSVM and its variants have been studied extensively [25,23,18,41,40,28,29,27]. Specifically, least squares type TWSVM (LSTSVM) [1] has been presented by using the squared loss function instead of the hinge one in TWSVM, leading to very fast training speed since two QPPs are replaced by two systems of linear equations. However, it has been pointed out in [1] that LSTSVM relaxes the constraint "the other class as far as possible from the hyperplane" to "the other class has a distance from the hyperplane", which may result in the reduction of classification ability, and meanwhile, the characteristic of constructing two non-parallel hyperplanes in TWSVM may also be weakened [18].

In this paper, we propose a twin-type support vector machine with weighted linear loss function, called weighted linear loss twin support vector machine (WLTSVM). Following TWSVM, the linear version of our WLTSVM constructs two non-parallel hyperplanes such that each hyperplane is proximal to one class and as far as possible from the other class. However, different from TWSVM, in the linear version of our WLTSVM, a weighted linear loss function is introduced. The main cost of our WLTSVM is solving two systems of linear equations that are much simpler than that of TWSVM, where two QPPs are needed to be solved. Besides, distinct from LSTSVM, our WLTSVM keeps the more reasonable constraint "the other class as far as possible from the hyperplane" in TWSVM. In fact, theoretical analysis shows that our WLTSVM not only

* Corresponding author. Tel./fax: +86 057187313551.
E-mail address: dengnaiyang@cau.edu.cn (N.-Y. Deng).

http://dx.doi.org/10.1016/j.knosys.2014.10.011
0950-7051/

maintains the merits of the TWSVM but also has lower computational cost. Furthermore, the two systems of linear equations in our WLTSVM can be solved efficiently by using the well-known conjugate gradient algorithm, resulting in the ability to deal with large-scale datasets without any extra external optimizers. It should be pointed out that our WLTSVM including its linear version and nonlinear version have been extended to multiple classification problems. Comparing to TWSVM [11,34], LSTSVM [1], NHSVM [35], GEPSVM [16], SVM, and LSSVM, the preliminary numerical experiments on several benchmark datasets show that our WLTSVM gains comparable generalization ability but with remarkable less training time.

This paper is organized as follows. In Section 2, a brief review and the discussion on SVM, LSSVM, TWSVM, and LSTSVM are given. Our WLTSVM is formulated in Section 3. And the numerical experiments are described in Section 4. Finally, Section 5 gives the conclusion.

2. Background

In this section, we consider the following binary classification problem with the training set $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$, where $x_i\in R^n$ are inputs and $y_i\in\{+1,-1\}$ are corresponding outputs. Further, suppose that all of the data points in class $+1$ are denoted by a matrix $A\in R^{m_1\times n}$, where the i-th row $A_i\in R^n$ represents the i-th data point. Similarly, the matrix $B\in R^{m_2\times n}$ represents the data points of class -1, $m-m_1=m_2$. We now give a brief outline of SVM related methods.

2.1. Support vector machine

Linear support vector machine [5,6] searches for a separating hyperplane

$$f(x)=w^\top x+b=0, \tag{1}$$

where $w\in R^n$ and $b\in R$. To measure the empirical risk, the soft margin loss function

$$R_{S+}+R_{S-}=e_2^\top\max(0,e_2+Bw+e_2b)+e_1^\top\max(0,e_1-Aw-e_1b) \tag{2}$$

is used, where e_1 and e_2 are vectors of all 1's of adequate size.

By introducing the regularization term $\frac{1}{2}\|w\|^2$ and the slack variables $\xi_1=(\xi_1,\ldots,\xi_{m_1})$ and $\xi_2=(\xi_1,\ldots,\xi_{m_2})$, the primal problem of SVM can be expressed as

$$\begin{aligned}\min_{w,b,\xi_1,\xi_2}\quad & \frac{1}{2}\|w\|^2+C(e_1^\top\xi_1+e_2^\top\xi_2)\\ \text{s.t.}\quad & Aw+e_1b\geqslant e_1-\xi_1,\quad \xi_1\geqslant 0,\\ & -(Bw+e_2b)\geqslant e_2-\xi_2,\quad \xi_2\geqslant 0,\end{aligned} \tag{3}$$

where $C>0$ is a parameter. Note that the minimization of the regularization term $\frac{1}{2}\|w\|^2$ is equivalent to the maximization of the margin between two parallel supporting hyperplanes $w^\top x+b=1$ and $w^\top x+b=-1$, and the structural risk minimization principle is implemented in this problem. Fig. 1(a) shows the geometric interpretation of this formulation for a toy example. After we obtain the optimal solution of (3), a new data point is classified as $+1$ or -1 according to whether the decision function, $\text{Class}i=\text{sgn}(w^\top x+b)$, yields 1 or -1 respectively.

2.2. Least squares SVM

Similar to SVM, linear least squares support vector machine (LSSVM) [39,38] also searches for a separating hyperplane (1). To measure the empirical risk, instead of applying the soft margin loss function, the quadratic loss function

$$R_{LS+}+R_{LS-}=\frac{1}{2}\|e_1-Aw-e_1b\|^2+\frac{1}{2}\|e_2+Bw+e_2b\|^2 \tag{4}$$

is used in LSSVM. By introducing the regularization term $\frac{1}{2}\|w\|^2$ and the slack variables ξ_1 and ξ_2, the primal problem of LSSVM can be expressed as

$$\begin{aligned}\min_{w,b,\xi_1,\xi_2}\quad & \frac{1}{2}\|w\|^2+C(\xi_1^\top\xi_1+\xi_2^\top\xi_2)\\ \text{s.t.}\quad & e_1-Aw-e_1b=\xi_1,\\ & e_2+Bw+e_2b=\xi_2,\end{aligned} \tag{5}$$

where $C>0$ is a parameter. Similar to SVM, the minimization of the regularization term $\frac{1}{2}\|w\|^2$ is equivalent to the maximization of the margin between two parallel proximal hyperplanes $w^\top x+b=1$ and $w^\top x+b=-1$. When we obtain the optimal solution of (5), a new data point is classified as $+1$ or -1 according to whether the decision function, $\text{Class}i=\text{sgn}(w^\top x+b)$, yields 1 or -1 respectively.

2.3. Twin support vector machine

Different from SVM, linear twin support vector machine (TWSVM) [11,34] seeks a pair of non-parallel hyperplanes

$$f_1(x)=w_1^\top x+b_1=0\quad\text{and}\quad f_2(x)=w_2^\top x+b_2=0, \tag{6}$$

such that each hyperplane is proximal to data points of one class and as far as possible from the data points of the other class, where $w_1\in R^n$, $w_2\in R^n$, $b_1\in R$ and $b_2\in R$. Here the empirical risks are measured by

$$R_{T1+}+c_1R_{T1-}=\frac{1}{2}\|Aw_1+e_1b_1\|^2+c_1e_2^\top\max(0,e_2+Bw_1+e_2b_1) \tag{7}$$

and

$$R_{T2-}+c_2R_{T2+}=\frac{1}{2}\|Bw_2+e_2b_2\|^2+c_2e_1^\top\max(0,e_1-Aw_2-e_1b_2), \tag{8}$$

where $c_1>0$ and $c_2>0$ are parameters. By introducing the slack variables ξ_1, ξ_2, η_1 and η_2, the primal problems are expressed as

$$\begin{aligned}\min_{w_1,b_1,\xi_1,\xi_2}\quad & \frac{1}{2}\xi_1^\top\xi_1+c_1e_2^\top\xi_2\\ \text{s.t.}\quad & Aw_1+e_1b_1=\xi_1,\\ & -(Bw_1+e_2b_1)+\xi_2\geqslant e_2,\quad \xi_2\geqslant 0,\end{aligned} \tag{9}$$

and

$$\begin{aligned}\min_{w_2,b_2,\eta_1,\eta_1}\quad & \frac{1}{2}\eta_2^\top\eta_2+c_2e_1^\top\eta_1\\ \text{s.t.}\quad & Bw_2+e_2b_2=\eta_2,\\ & (Aw_2+e_1b_2)+\eta_1\geqslant e_1,\quad \eta_1\geqslant 0.\end{aligned} \tag{10}$$

Fig. 1(b) shows the geometric interpretation of this formulation for a toy example. The corresponding dual problems are

$$\begin{aligned}\max_{\alpha}\quad & e_2^\top\alpha-\frac{1}{2}\alpha^\top G(H^\top H)^{-1}G^\top\alpha\\ \text{s.t.}\quad & 0\leqslant\alpha\leqslant c_1e_2,\end{aligned} \tag{11}$$

and

$$\begin{aligned}\max_{\gamma}\quad & e_1^\top\gamma-\frac{1}{2}\gamma^\top H(G^\top G)^{-1}H^\top\gamma\\ \text{s.t.}\quad & 0\leqslant\gamma\leqslant c_2e_1,\end{aligned} \tag{12}$$

where $G=[B\quad e_2]$ and $H=[A\quad e_1]$. In order to deal with the case when $H^\top H$ or $G^\top G$ is singular and avoid the possible ill-conditioning,