Knowledge-Based Systems 85 (2015) 224–233

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Ramp loss nonparallel support vector machine for pattern classification

Dalian Liu [a,b], Yong Shi [a,c,d,e,*], Yingjie Tian [c,d,*]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[b] Department of Basic Course Teaching, Beijing Union University, Beijing 100101, China
[c] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
[d] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[e] Collegae of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA

ARTICLE INFO

Article history:
Received 15 February 2015
Received in revised form 16 April 2015
Accepted 8 May 2015
Available online 16 May 2015

Keywords:
Support vector machine
Twin support vector machine
CCCP
Ramp loss
Sparseness

ABSTRACT

In this paper, we propose a novel sparse and robust nonparallel hyperplane classifier, named Ramp loss Nonparallel Support Vector Machine (RNPSVM), for binary classification. By introducing the Ramp loss function and also proposing a new non-convex and non-differentiable loss function based on the ε-insensitive loss function, RNPSVM can explicitly incorporate noise and outlier suppression in the training process, has less support vectors and the increased sparsity leads to its better scaling properties. The non-convexity of RNPSVM can be efficiently solved by the Concave–Convex Procedure and experimental results on benchmark datasets confirm the effectiveness of the proposed algorithm.

1. Introduction

Support vector machines (SVMs), rooted in statistical learning theory (SLT), are computationally powerful tools for pattern classification [1–5]. Recently, a branch of SVM, nonparallel hyperplane SVM, is developed and has attracted many interests. The representative algorithms include the generalized eigenvalue proximal support vector machine (GEPSVM) [6] and the twin support vector machine (TWSVM) [7]. For the binary classification problem, TWSVM 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. It is implemented by solving two smaller quadratic programming problems (QPPs) instead of a larger one, which increases the TWSVM training speed by approximately fourfold compared to that of standard SVM. TWSVMs have been studied extensively [8–23]. Among the extensions of TWSVMs, the nonparallel support vector machine (NPSVM) [21,22] are superior theoretically and overcomes several drawbacks of the existing TWSVMs.

For the standard SVMs, the convex loss functions such as the Hinge loss function are applied, then the convex models are constructed and many convex optimization techniques have been employed to solve them [15,24–28]. However, researchers have shown that classical SVMs are sensitive to the presence of outliers and yield poor generalization performance, since the outliers tend to have the largest margin losses according to the character of the convex loss functions, then are always playing dominant roles in determining the decision hyperplane. There are several methods to construct the robust models [29–36], of which the Ramp loss function has been investigated widely in the theoretical literature in order to improve the robustness of SVMs [34,36]. They constructed a Ramp loss support vector machine (RSVM) by taking the Ramp loss instead of the Hinge loss in the classical SVM, the Ramp loss function limits its maximal loss value distinctly and can put definite restrictions on the influences of outliers so that it is much less sensitive to their presence. However, it will also cause the objective of SVMs losing convexity, as a consequence, the concave–convex programming (CCCP) procedure is applied to solves a sequence of convex problems to produce faster and sparser SVMs.

For the NPSVM, the Hinge loss function and ε-insensitive loss function are applied [22], similarly NPSVM will be also sensitive to the presence of outliers according to the character of the convex loss functions. In this paper, inspired by RSVM, we introduce the Ramp loss function and also propose a new non-convex and non-differentiable loss function based on the ε-insensitive loss function to NPSVM, to construct a novel robust NPSVM, termed as RNPSVM. Compared with the original NPSVM [22], RNPSVM can explicitly incorporate noise and outlier suppression in the training process, has less support vectors and the increased

* Corresponding authors at: Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China.
E-mail addresses: ldlluck@sina.com (D. Liu), yshi@ucas.ac.cn (Y. Shi), tyj@ucas.ac.cn (Y. Tian).

http://dx.doi.org/10.1016/j.knosys.2015.05.008
0950-7051/

sparsity leads to its better scaling properties. RNPSVM is non-convex and the CCCP procedure is applied to solve a sequence of convex QPPs. Experimental results on benchmark datasets confirm the effectiveness of the proposed algorithm.

The rest of this paper is organized as follows. Section 2 briefly dwells on the Hinge loss SVM, Ramp loss SVM, CCCP procedure and TWSVMs. Section 3 proposes the RNPSVM and discusses its properties. Section 4 deals with experimental results and Section 5 contains concluding remarks.

2. Background

In this section, we briefly introduce the Hinge loss SVM, Ramp loss SVM, CCCP procedure and TWSVMs (the standard TWSVM and an improved TWSVM: NPSVM).

2.1. Hinge loss SVM

Consider the binary classification problem with the training set

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

where $x_i \in \mathscr{R}^n, y_i \in \mathscr{Y} = \{1, -1\}, i = 1, \ldots, l$, the standard SVM relies on the classical Hinge loss function (see Fig. 1(b))

$$H_s(z) = \max(0, s - z) \tag{2}$$

where the subscript s indicates the position of the Hinge point, to penalize examples classified with an insufficient margin and results in the following primal problem

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} H_1(y_i f(x_i)), \tag{3}$$

where $f(x)$ is the decision function with the form of $f(x) = (w \cdot \Phi(x)) + b$, and $\Phi(\cdot)$ is the chosen feature map, often implicitly defined by a Mercer kernel $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ [3].

Due to the application of the Hinge loss, standard SVM has the sensitivity to outlier observations since they will normally have the largest hinge loss, thus the decision hyperplane is inappropriately drawn toward outlier samples so that its generalization performance is degraded [37]. Another property of the Hinge loss function is that the number of Support Vectors (SVs) scales linearly with the number of examples [38], and since the SVM training and recognition times grow quickly with the number of SVs, it is obviously that SVMs cannot deal with very large datasets.

2.2. Ramp loss SVM

In order to increase the robustness of SVM and avoid converting the outliers into SVs, the Ramp loss function [34] (see Fig. 1(a)), also known as the Robust Hinge loss

$$R_s(z) = \begin{cases} 0, & z > 1 \\ 1 - z, & s \leqslant z \leqslant 1 \\ 1 - s, & z < s \end{cases} \tag{4}$$

was introduced to replace the Hinge loss function, by making the loss function flat for scores z smaller than a predefined value $s < 1$. $R_s(z)$ can be decomposed into the sum of the convex Hinge loss and a concave loss (see Fig. 1(c)),

$$R_s(z) = H_1(z) - H_s(z), \tag{5}$$

therefore the primal problem of the Ramp loss SVM (RSVM) is formulated as

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} R_s(y_i f(x_i)) = \underbrace{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} H_1(y_i f(x_i))}_{\text{convex}} - \underbrace{C\sum_{i=1}^{l} H_s(y_i f(x_i))}_{\text{concave}}, \tag{6}$$

which can be solved by the "Concave–Convex Procedure" (CCCP) [39].

2.3. The Concave–Convex Procedure

The CCCP procedure is closely related to the "Difference of Convex" (DC) methods, which were successfully applied to a lot of different and various non-differentiable non-convex optimization problems especially in the large-scale setting [40,41]. For such problem (6) with the objective function written as the sum of a convex part $u(x)$ and a concave part $v(x)$, i.e. $u(x) + v(x)$, the CCCP algorithm is an iterative procedure that solves a sequence of convex programs

$$x^{t+1} = \arg \min_x \{u(x) + x^\top \nabla v(x^t)\}. \tag{7}$$

Ref. [34] proposed the CCCP procedure for the RSVM as follows:

Algorithm 1 (*CCCP for RSVM*)

(1) Initialize $\beta^0 = (\beta_1^0, \cdots, \beta_l^0)^\top$, set $t = 1$;
(2) Compute α^t by solving the following convex problem

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{l}\alpha_i, \\ \text{s. t.} \quad & \sum_{i=1}^{l} y_i\alpha_i = 0, \\ & -\beta_i^{t-1} \leqslant \alpha_i \leqslant C - \beta_i^{t-1}, i = 1, \ldots, l, \end{aligned} \tag{8}$$

where $K(x, x')$ is the kernel function;
(3) Compute b^t and construct the decision function as

$$f^t(x) = \sum_{i=1}^{l} y_i\alpha_i^t K(x_i, x) + b^t; \tag{9}$$

(4) Compute $\beta_i^t, i = 1, \ldots, l$ as

$$\beta_i^t = \begin{cases} C, & y_i f^t(x_i) < s \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

(5) If $\beta^t = \beta^{t-1}$, end; else set $t = t + 1$, go to step (2).

2.4. TWSVM

Consider the binary classification problem with the training set

$$T = \{(x_1, +1), \ldots, (x_p, +1), (x_{p+1}, -1), \ldots, (x_{p+q}, -1)\}, \tag{11}$$

where $x_i \in R^n, i = 1, \ldots, p + q$, the TWSVM generates two nonparallel hyperplanes $f^+(x) = (w_+ \cdot x) + b_+ = 0$ and $f^-(x) = (w_- \cdot x) + b_- = 0$, instead of a single one as in conventional SVMs, by solving a pair of smaller-sized QPPs

$$\min_{w_+,b_+} \frac{1}{2}\sum_{i=1}^{p}(f^+(x_i))^2 + C_1\sum_{j=p+1}^{p+q} H_1(-f^+(x_j)), \tag{12}$$

and

$$\min_{w_-,b_-} \frac{1}{2}\sum_{i=p+1}^{p+q}(f^-(x_i))^2 + C_2\sum_{j=1}^{p} H_1(f^-(x_j)) \tag{13}$$

where $C_i, i = 1, 2$ are the penalty parameters.

For the nonlinear case, two kernel-generated surfaces instead of hyperplanes are considered and two other primal problems are constructed.