Journal of Computational Science 14 (2016) 61–68

Contents lists available at ScienceDirect

Journal of Computational Science

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

# Ramp loss least squares support vector machine

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

[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] College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA
[f] Tourism Institute of Beijing Union University, Beijing 100101, China

ARTICLE INFO

*Article history:*
Received 15 August 2015
Received in revised form 28 January 2016
Accepted 3 February 2016
Available online 4 February 2016

*Keywords:*
Least squares support vector machine
Sparse
Ramp loss
CCCP
Classification

ABSTRACT

In this paper, we propose a novel sparse least squares support vector machine, named ramp loss least squares support vector machine (RLSSVM), for binary classification. By introducing a non-convex and non-differentiable loss function based on the $\varepsilon$-insensitive loss function, RLSSVM has several improved advantages compared with the plain LSSVM: firstly, it has the sparseness which is controlled by the ramp loss, leading to its better scaling properties; secondly, it can explicitly incorporate noise and outlier suppression in the training process, and thirdly, the non-convexity of RLSSVM can be efficiently solved by the Concave-Convex Procedure (CCCP). Experimental results on several benchmark datasets show the effectiveness of our method.



## 1. Introduction

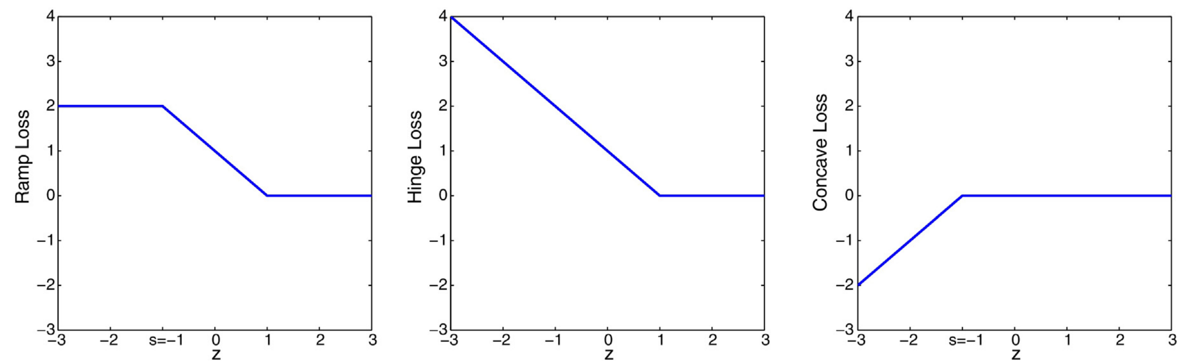
Support vector machines (SVMs), being computationally powerful tools for supervised learning, are successfully and widely used in classification and regression in a variety of real-world applications [1–5]. One efficient extension of SVMs is the least squares support vector machine (LSSVM), see [6,7], which only needs to solve a linear system instead of a quadratic programming problem (QPP) in standard SVMs. Extensive empirical comparisons [8] show that LSSVMs obtain good performance on various classification and regression problems. LSSVMs have been studied extensively, see for example, [9–12].

However, there are still several disadvantages in the standard SVMs or LSSVMs: First, for the standard SVMs or LSSVMs, the convex loss functions such as the Hinge loss function or the quadratic loss function are applied, then the convex models are constructed and many convex optimization techniques have been employed to solve them [13–18]. However, researchers have shown that classical SVMs or LSSVMs are sensitive to the presence of outliers and yield poor generalization performance, since the outliers tend to have the largest losses according to the character of the convex loss functions, then are always playing dominant roles in determining the decision hyperplane. Second, another obvious limitation of LSSVMs is that, unlike the standard SVM employing soft-margin loss function for classification and $\varepsilon$-insensitive loss function for regression, LSSVMs usually looses the sparseness by using a quadratic loss function.

There are a lot of papers in the literature considering the above two issues so far. As for the robustness, there are several methods to construct the robust models [19,6,20–25], of which the ramp loss function has been investigated widely in the theoretical literature in order to improve the robustness of SVMs [23,25]. 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 solve a sequence of convex problems to produce faster and sparser SVMs. As for the sparse one, a range of methods for LSSVMs are available and can be roughly divided into two major classes: Pruning and Fixing size ones, which are summarized in [26]. For the first class, it imposes the sparseness by gradually omitting the least important data from the training set and re-estimating the LSSVMs, which is time consuming; For the second class, it is assumed that the weight

* Corresponding authors at: Key Laboratory of Big Data Mining and Knowledge Management, 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), xiankai@buu.edu.cn (X. Huang).

http://dx.doi.org/10.1016/j.jocs.2016.02.001
1877-7503/




(a) The Ramp Loss (b) The Hinge Loss (c) The Concave Loss

**Fig. 1.** The Ramp Loss function (left) can be decomposed into the sum of the convex Hinge Loss (middle) and a concave loss (right).

vector $w$ can be represented as a weighted sum of a limited number (far less than the size of the training set) of basis vectors, which is a rough approximation and not theoretically guaranteed.

In this paper, by introducing a non-convex and non-differentiable loss function instead of the quadratic loss function to LSSVM, a robust and sparse LSSVM is constructed and named RLSSVM. Compared with the original LSSVM, RLSSVM 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. Similar to RSVM, RLSSVM 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 and LSSVM. Section 3 proposes the RLSVM and discusses its properties. Section 4 presents the experimental results and Section 5 contains concluding remarks.

## 2. Background

In this section, we briefly introduce the Hinge loss SVM, Ramp Loss SVM and LSSVM.

### 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 $x_i \in \mathcal{R}^n, y_i \in \mathcal{Y} = \{1, -1\}, i = 1, \ldots, l$, 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

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

where $f(x)$ is the decision function with the form of $f(x) = (\mathrm{w} \cdot \Phi(x)) + b$, and $\Phi(\cdot)$ is the chosen feature map, often implicitly defined by a Mercer kernel $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ [3]. For the choice of the kernel function $K(x, x')$, one has several possibilities: $K(x, x') = (x \cdot x')$ (linear kernel); $K(x, x') = ((x \cdot x') + 1)^d$ (polynomial kernel of degree $d$); $K(x, x') = \exp(-\parallel x - x' \parallel^2/\sigma^2)$ (RBF kernel); $K(x, x') = \tanh(\kappa(x \cdot x') + \theta)$ (Sigmoid kernel), etc.

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 [27]. Another property of the Hinge Loss function is that the number of Support Vectors (SVs) scales linearly with the number of examples [28], 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 [23] (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) can be formulated as

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

which can be solved by the CCCP Procedure [29].

### 2.3. LSSVM

For the given training set (1), the primal problem of standard LSSVM to be solved is

$$\min_{\mathrm{w},b,} \quad \frac{1}{2}\|\mathrm{w}\|^2 + \frac{C}{2}\sum_{i=1}^{l} Q(y_i f(x_i) - 1), \tag{7}$$