Computers and Mathematics with Applications 66 (2013) 1935–1947

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

Efficient sparse least squares support vector machines for pattern classification

Yingjie Tian *, Xuchan Ju, Zhiquan Qi, Yong Shi

Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Least squares support vector machine
Sparseness
Loss function
Classification
Regression

ABSTRACT

We propose a novel least squares support vector machine, named ε-least squares support vector machine (ε-LSSVM), for binary classification. By introducing the ε-insensitive loss function instead of the quadratic loss function into LSSVM, ε-LSSVM has several improved advantages compared with the plain LSSVM. (1) It has the sparseness which is controlled by the parameter ε. (2) By weighting different sparseness parameters ε for each class, the unbalanced problem can be solved successfully, furthermore, an useful choice of the parameter ε is proposed. (3) It is actually a kind of ε-support vector regression (ε-SVR), the only difference here is that it takes the binary classification problem as a special kind of regression problem. (4) Therefore it can be implemented efficiently by the sequential minimization optimization (SMO) method for large scale problems. Experimental results on several benchmark datasets show the effectiveness of our method in sparseness, balance performance and classification accuracy, and therefore confirm the above conclusion further.

1. Introduction

Support vector machines (SVMs), which were introduced by Vapnik and his co-workers in the early 1990s [1–3], are computationally powerful tools for supervised learning [4,5] and have already outperformed most other methods in a wide variety of applications [6–11]. Least squares support vector machines (LSSVMs) were also proposed [12,13] which only need to solve a linear system instead of a quadratic programming problem (QPP) in standard SVMs, and extensive empirical comparisons [14] show that LSSVMs obtain good performance on various classification and regression problems. LSSVMs have been studied extensively [15–18].

Unfortunately, there are two drawbacks in the plain LSSVMs. (1) Unlike the standard SVM employing a soft-margin loss function for classification and a ε-insensitive loss function for regression, LSSVMs lost the sparseness by using a quadratic loss function. (2) Another obvious limitation of LSSVMs is that although solving a linear system is in principle solvable [19], it is in practice intractable for a large dataset by the classical techniques since their computational complexity is usually of order $O(l^3)$ (l is the size of the training set), which severely limits the utility of LSSVMs in large scale applications.

There are a lot of papers in the literature considering the above two issues so far. As for the fast algorithms for LSSVMs, Suykens et al. [20] presented an iterative algorithm based on the conjugate gradient algorithm, and Chu et al. [21] improved the conjugate gradient algorithm by solving one reduced linear system. Keerthi and Shevade [22] extends the well-known sequential minimization optimization (SMO) [23] algorithm of SVMs for the solution of LSSVMs. For the problems with very large numbers of data points but small numbers of features, Chua [24] proposed a method which involves working with (and storing) matrices that are at most of size $l \times n$ (l is the size of the training set, n is the number of features), and extend the possible range of application for LSSVMs. *However, the resulting solutions of the above methods are not sparse yet.*

* Corresponding author. Tel.: +86 10 82680997.
E-mail addresses: tyj@ucas.ac.cn (Y. Tian), juxuchan10@mails.ucas.ac.cn (X. Ju), qizhiquan@ucas.ac.cn (Z. Qi), yshi@ucas.ac.cn (Y. Shi).

0898-1221/$ – see front matter
http://dx.doi.org/10.1016/j.camwa.2013.06.028

As for the sparse algorithms for LSSVMs, a range of methods are available and can be roughly concluded into two major classes: pruning and fixing size. In the first class, a simple approach to introduce the sparseness is based on the sorted support value spectrum (SVS) and prunes the network by gradually removing points from the training set [25,26]. A sophisticated mechanism was proposed by weighting the support values, the data point with the smallest error introduced after its omission is selected then. This pruning method is claimed to outperform the standard scheme [27]. Hoegaerts et al. [28] suggested an improved selection of the pruning point based on their derived criterion. Zeng and Chen [29] proposed a SMO-based pruning method, the SMO method is introduced into the pruning process and instead of determining the pruning points by errors, the data points that will introduce minimum changes to a dual objective function are omitted. Li et al. [30] selected the reduced classification training set based on $y_i f(x_i)$ ($f(x)$ is the decision function and y the label) instead of the support value. The second class mainly considers the fixed-size LSSVMs for fast finding the sparse approximate solution of LSSVMs, in which a reduced set of candidate support vectors is used in the primal space [13] or in kernel space [31–34]. *However, there are still shortcomings in the existing sparse LSSVMs*. 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 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, we propose a novel LSSVM, termed ε-LSSVM for binary classification. ε-LSSVM introduces the ε-insensitive loss function instead of the quadratic loss function into LSSVM. (1) It has the sparseness which is controlled by the parameter ε. (2) By weighting different sparseness parameters for each class, the unbalanced problem can be solved successfully, furthermore, we also propose an useful choice of the parameter ε. (3) It is actually a kind of ε-support vector regression (ε-SVR) [3–5], the only difference here is that it takes the binary classification problem as a special kind of regression problem. (4) Certainly it can be implemented efficiently by SMO for large scale problems.

The paper is organized as follows. Section 2 briefly dwells on the standard C-support vector machine for classification (C-SVC) and LSSVMs. Section 3 proposes our ε-LSSVM and a weighted ε-LSSVM is given in Section 4. Section 5 deals with experimental results. Section 6 contains concluding remarks.

2. Background

In this section, we give a brief outline of C-SVC and LSSVMs.

2.1. C-SVC

Consider the binary classification problem with the training set

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (R^n \times \mathcal{Y})^l, \tag{1}$$

where $x_i \in R^n$, $y_i \in \mathcal{Y} = \{1, -1\}$, $i = 1, \dots, l$, standard C-SVC formulates the problem as a convex quadratic programming problem (QPP)

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

where $\xi = (\xi_1, \dots, \xi_l)^\top$, and $C > 0$ is a penalty parameter. For this primal problem, C-SVC solves its Lagrangian dual problem

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{l}\alpha_i, \\ &\text{s.t. } \sum_{i=1}^{l} y_i\alpha_i = 0, \\ &0 \leqslant \alpha_i \leqslant C, \quad i=1,\dots,l, \end{aligned} \tag{3}$$

where $K(x, x')$ is the kernel function, which is also a convex QPP and then constructs the decision function.

2.2. LSSVM

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

$$\begin{aligned} &\min_{w,b,\eta} \ \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{l}\eta_i^2, \\ &\text{s.t. } y_i((w\cdot x_i)+b) = 1-\eta_i, \quad i=1,\dots,l. \end{aligned} \tag{4}$$