Pattern Recognition 78 (2018) 167–181

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse L_q-norm least squares support vector machine with feature selection

Yuan-Hai Shao[a,*], Chun-Na Li[b], Ming-Zeng Liu[c], Zhen Wang[d], Nai-Yang Deng[e,*]

[a] School of Economics and Management, Hainan University, Haikou, 570228, PR China
[b] Zhijiang College, Zhejiang University of Technology, Hangzhou, 310024, PR China
[c] School of Mathematics and Physics Science, Dalian University of Technology at Panjin, Dalian, 124221, PR China
[d] School of Mathematical Sciences, Inner Monggolia University, Hohhot 010021, PR China
[e] College of Science, China Agricultural University, Beijing 100083, PR China

ARTICLE INFO

Article history:
Received 31 October 2016
Revised 11 January 2018
Accepted 22 January 2018

Keywords:
Least squares support vector machine (LS-SVM)
L_q-norm
Feature selection
Sparse approximation
Global optimality

ABSTRACT

Least squares support vector machine (LS-SVM) is a popular hyperplane-based classifier and has attracted many attentions. However, it may suffer from singularity or ill-condition issue for the small sample size (SSS) problem where the sample size is much smaller than the number of features of a data set. Feature selection is an effective way to solve this problem. Motivated by this, in the paper, we propose a sparse L_q-norm least squares support vector machine (L_q-norm LS-SVM) with $0 < q < 1$, where feature selection and prediction are performed simultaneously. Different from traditional LS-SVM, our L_q-norm LS-SVM minimizes the L_q-norm of weight and releases the least squares problem in primal space, resulting in that feature selection can be achieved effectively and small enough number of features can be selected by adjusting the parameters. Furthermore, our L_q-norm LS-SVM can be solved by an efficient iterative algorithm, which is proved to be convergent to a global optimal solution under some assumptions on the sparsity. The effectiveness of the proposed L_q-norm LS-SVM is validated via theoretical analysis as well as some illustrative numerical experiments.

1. Introduction

In pattern recognition, we are generally given a set of samples of input vectors along with corresponding class labels, and the task is to find a deterministic function that best represents the relation between input vectors and class labels. As a powerful tool, support vector machine (SVM) [1,2] has been successfully applied and studied from various aspects [3–7]. Specifically, in order to reduce the time complexity, least squares support vector machine (LS-SVM) [8,9] was proposed, which attempts to minimize the least squares error on the training samples while simultaneously to maximize the margin between two classes. Extensive empirical comparisons show that LS-SVM is able to obtain good performance on many problems with a fast training speed [10–14].

However, LS-SVM is not suitable for small sample size (SSS) problem, where the sample size is smaller than the number of features. The reason is that LS-SVM lacks the ability of feature selection while feature selection is an effective way. The aim of this paper is to present a modified version of LS-SVM with rather strong feature selection ability.

In general, there are two ways to perform feature selection for SVM classifiers. One way is to conduct a separate feature selection procedure before classification, such as univariate ranking [15] and recursive feature elimination [16]. Another way is to conduct feature selection and classification simultaneously. The latter way is usually preferable since better performance could be achieved, as pointed by Fan and Li [17,18]. This statement has been confirmed and demonstrated by subsequently published papers in recent years [19–21]. For example, L_1-norm SVM [24] with LP formulation was proposed to accomplish feature selection and prediction at the same time, where the standard LP packages could be used [25]. As an improvement of L_1-norm SVM, Zou [26] considered the hybrid SVM that penalizes the empirical hinge loss by the adaptively weighted L_1-norm function, where the weights were computed with the L_2-norm SVM. Inspired by L_1-norm SVM, the L_q-norm SVM [20] ($0 < q < 1$) was proposed [21]. Numerical experimental results have shown that the employment of L_q-norm not only makes the classifier more suitable for selecting features but also improves the classification accuracy [20–23].

[*] Corresponding authors.
E-mail addresses: shaoyuanhai@hainu.edu.cn (Y.-H. Shao), dengnaiyang@cau.edu.cn (N.-Y. Deng).

https://doi.org/10.1016/j.patcog.2018.01.016
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

Encouraged by the successful use of L_q-norm, we investigate the SSS problem and present a new sparse least squares support vector machine, called L_q-norm LS-SVM ($0<q<1$), where the sparse formulation is constructed in the primal space. Different from traditional sparse least squares support vector machines in the dual space, our L_q-norm LS-SVM obtains sparse solution in the primal space and is able to select useful features. Due to the use of L_q-norm, L_q-norm LS-SVM is hard to solve. Therefore, we introduce a regularized version of L_q-norm problem and solved it by an iterative algorithm with convergence to a global optimal solution. We summarize the main advantages of our L_q-norm LS-SVM as follows:

(i) feature selection is achieved effectively by minimizing the L_q-norm of weight in the primal least squares SVM for small sample size problem;
(ii) the number of selected features in L_q-norm LS-SVM can be adjusted by choosing the parameters according to the practical requirement;
(iii) the corresponding optimal problem of our L_q-norm LS-SVM is solved by an efficient iterative algorithm, which is proved to be convergent to a global optimal solution under some assumptions;
(iv) preliminary experimental results show its validity in both classification performance and feature selection ability.

The paper is organized as follows. Section 2 briefly dwells on LS-SVM. Section 3 proposes an L_q-norm minimization problem which is the foundation of our model. Our L_q-norm LS-SVM and its theoretical analysis are presented in Section 4. Section 5 makes some experimental comparisons between our L_q-norm LS-SVM with LS-SVM, SVM, L_1-norm SVM, and L_q-norm SVM, and concluding remarks are given in Section 6.

2. Least squares support vector machine

Consider a binary classification problem in the n-dimensional real space R^n. The set of training points is represented by $T=\{(x_i,y_i)|i=1,2,\ldots,m\}$, where $x_i\in R^n$ is the input and $y_i\in\{+1,-1\}$ is the corresponding output ($i=1,2,\ldots,m$). We further organize the m inputs by a matrix $X\in R^{m\times n}$ and the m outputs by a diagonal matrix $Y\in R^{m\times m}$ with its (i, i)th element y_i.

Least squares support vector machine (LS-SVM) [8,9] searches for a hyperplane

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

that separates two classes. The optimization problem of LS-SVM is formulated as

$$\begin{aligned}\min_{w,b,\xi}\quad & \tfrac{1}{2}||w||^2+\tfrac{\gamma}{2}\xi^\top\xi\\ \text{s.t.}\quad & Y(Xw+eb)+\xi=e\end{aligned} \tag{2}$$

by minimizing the structural risk and the sum-of-squares empirical risk, where $||\cdot||$ denotes the L_2-norm, $e=(1,\ldots,1)^T\in R^m$, $\gamma>0$ is a parameter, and $\xi\in R^m$ is a slack variable.

Introducing the Lagrange multiplier $\alpha=(\alpha_1,\ldots,\alpha_m)^\top$ for the equality constraint gives

$$L(w,b,\xi,\alpha)=\tfrac{1}{2}\|w\|^2+\tfrac{\gamma}{2}\xi^\top\xi-\alpha^\top(Y(Xw+eb)+\xi-e). \tag{3}$$

The optimality conditions are

$$\begin{cases}\dfrac{\partial L}{\partial w}=0 \implies w=X^\top Y^\top\alpha,\\ \dfrac{\partial L}{\partial b}=0 \implies \alpha^\top Ye=0,\\ \dfrac{\partial L}{\partial \xi}=0 \implies \alpha=\gamma\xi,\\ \dfrac{\partial L}{\partial \alpha}=0 \implies Y(Xw+eb)+\xi-e=0.\end{cases} \tag{4}$$

This leads to the following linear system of equations

$$\begin{bmatrix} e^\top Y & 0\\ YXX^\top Y^\top+\frac{1}{\gamma}I & Ye\end{bmatrix}\begin{bmatrix}\alpha\\ b\end{bmatrix}=\begin{bmatrix}0\\ e\end{bmatrix}. \tag{5}$$

After obtaining the solution α^* and b^* of (5) and computing $w^*=X^TY^T\alpha^*$, a new sample is classified as Class +1 or Class −1 according to whether the decision function, Class $j=\text{sgn}(w^{*\top}x+b^*)=\text{sgn}(\alpha^{*T}YXx+b^*)$, yields +1 or −1 respectively.

3. Sparse approximation of LS-SVM with L_q-norm in primal space

As mentioned before, LS-SVM lose sparsity. So, many computing sparse solutions of LS-SVM has appeared in the literatures, e.g., [27–29], but all of them are not concerned with feature selection as discussed in this paper because they are motivated by solving problem (5), where the problem size of (5) is $(m+1)*(m+1)$, and the feature selection can not be implemented for SSS problem at the same time.

In contrast to solve the problem (5), we consider to solve the primal problem of LSSVM. Firstly, we reformulate (2) as

$$\min_{w,b} f(w,b)=\tfrac{1}{2}||w||^2+\tfrac{\gamma}{2}\|Y(Xw+eb)-e\|^2. \tag{6}$$

Setting the gradient of (6) with respect to w and b to zero gives

$$\begin{cases}\dfrac{\partial f}{\partial w}=0\implies w+\gamma X^\top Y^\top(Y(Xw+eb)-e)=0,\\ \dfrac{\partial f}{\partial b}=0\implies \gamma e^\top Y^\top(Y(Xw+eb)-e)=0.\end{cases} \tag{7}$$

By arranging (7) in matrix form, we have

$$\begin{bmatrix}X^\top X+\frac{1}{\gamma}I & X^\top e\\ e^\top X & e^\top e\end{bmatrix}\begin{bmatrix}w\\ b\end{bmatrix}=\begin{bmatrix}X^\top\\ e^\top\end{bmatrix}Ye. \tag{8}$$

Let

$$H=H(\gamma)=\begin{bmatrix}X^\top X+\frac{1}{\gamma}I & X^\top e\\ e^\top X & e^\top e\end{bmatrix}, \tag{9}$$

$$d=\begin{bmatrix}X^\top\\ e^\top\end{bmatrix}Ye \text{ and } u=\begin{bmatrix}w\\ b\end{bmatrix}, \tag{10}$$

then (8) can be written as

$$Hu=d, \tag{11}$$

where $H=H(\gamma)\in R^{(n+1)\times(n+1)}$ is a symmetric matrix, $u\in R^{n+1}$, and $d\in R^{n+1}$.

When the number of samples is much smaller than the number of features, the matrix H may be ill-conditioned and the solution of (11) may not be unique. Therefore, it is natural to sparsify problem (11) by introducing the L_0-norm, and the new problem is given by

$$\begin{aligned}\min_{u}\quad & ||u||_0\\ \text{s.t.}\quad & \|Hu-d\|^2\le\delta,\end{aligned} \tag{12}$$

where $\delta>0$ is a tolerance measure and $||u||_0$ denotes the number of nonzero elements of vector $u\in R^{n+1}$ to be estimated. The