Applied Mathematics and Computation 305 (2017) 32–52

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

# Robust $L_p$-norm least squares support vector regression with feature selection

Ya-Fen Ye[a,b], Yuan-Hai Shao[b,*], Nai-Yang Deng[c], Chun-Na Li[b], Xiang-Yu Hua[d]

[a] College of Economics, Zhejiang University, Hangzhou 310027, China
[b] Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, PR China
[c] College of Science, China Agricultural University, Beijing 100083, PR China
[d] Zhejiang Economic Information Center, Hangzhou 310006, PR China

ARTICLE INFO

*Keywords:*
Support vector regression
Feature selection
$L_p$-norm
Least squares
Robust regression

ABSTRACT

In this paper, we aim a novel algorithm called robust $L_p$-norm least squares support vector regression ($L_p$-LSSVR) that is more robust than the traditional least squares support vector regression(LS-SVR). Using the absolute constraint and the $L_p$-norm regularization term, our $L_p$-LSSVR performs robust against outliers. Moreover, though the optimization problem is non-convex, the sparse solution of $L_p$-norm and the lower bonds for nonzero components technique ensure useful features selected by $L_p$-LSSVR, and it helps to find the local optimum of our $L_p$-LSSVR. Experimental results show that although $L_p$-LSSVR is more robust than least squares support vector regression (LS-SVR), and much faster than $L_p$-norm support vector regression ($L_p$-SVR) and SVR due to its equality constraint, it is slower than LS-SVR and $L_1$-norm support vector regression ($L_1$-SVR), it is as effective as $L_p$-SVR, $L_1$-SVR, LS-SVR and SVR in both feature selection and regression.



## 1. Introduction

Support vector regression (SVR) [1], being one of the computationally powerful tools of support vector machines (SVMs) [2–4], has been successfully applied to various practical problems, such as economy [5,6], finance [7,8], and industry [9,10]. However, using all features in these problems may decrease the regression regressiveness since some features are usually either redundant or even uninformative. Therefore, features should be compressed when there exist some redundant ones. Feature selection is a useful way to select some useful features upon which to focus its attention, while ignoring the rest [11,12]. In general, there are two ways to perform feature selection in regression, having been widely studied [13–19]. One way is that feature selection and regression procedure are realized separately [13,14]. That is to say, in the first stage, important features are selected from the original features; based on these selected features, the regression procedure is then realized in the second stage. Another way is to conduct feature selection and regression simultaneously [15–18], meaning that the solution obtained in the regression procedure has sparse property.

For standard SVR, it may suffer from the presence of redundant features since it always utilizes all the features without discrimination [1–3], leading to the overfitting problem [20–22] easily. In order to overcome this drawback, on the one hand, filter [13] or wrapper [14] feature selection method was adopted before the regression procedure. On the other hand,

* Corresponding author.
*E-mail addresses:* shaoyuanhai21@163.com (Y.-H. Shao), dengnaiyang@cau.edu.cn (N.-Y. Deng).

http://dx.doi.org/10.1016/j.amc.2017.01.062

following the spirit of $L_1$-support vector machine ($L_1$-SVM) [23–26], $L_1$-norm support vector regression ($L_1$-SVR) was adopted to achieve better sparsity [19]. $L_1$-SVR can realize feature selection and regression simultaneously due to its inherent feature selection property [16,27,28]. In order to obtain sparser solution, $L_p$-support vector regression ($L_p$-SVR) ($0 < p < 1$) was studied [15,17,18]. The research results in [15] show that linear $L_p$-SVR can not only do feature selection, but also improve the regression effectiveness. However, $L_p$-SVR [15] has only its linear version, meaning that it is not appropriate when data set is structurally nonlinear. In addition, it has a slower training speed and could not deal with large scale data.

To realize feature selection in regression more efficiently, we follow the spirit of least squares support vector machine (LSSVM) [29] and propose a novel algorithm called $L_p$-least squares support vector regression ($L_p$-LSSVR) in this paper. Our $L_p$-LSSVR uses an equality constraint instead of inequality ones in $L_p$-SVR, cutting down the computational cost largely. Moreover, in order to improve the robustness of the conventional LSSVM using a sum squared error cost function [29], here the ordinary error cost function with the absolute constraint is used in our $L_p$-LSSVR. In addition, our linear $L_p$-LSSVR can perform feature selection and improve the regression effectiveness simultaneously by adjusting parameter $p$. When the structure of selected features is nonlinear we can adopt our nonlinear $L_p$-LSSVR to realize regression effectively. This implies that our $L_p$-LSSVR conducts nonlinear regression with feature selection by two stages. Therefore, our $L_p$-LSSVR has the ability to perform feature selection in both linear and nonlinear regression. To solve the non-convex $L_p$-norm problem, We adopt a convergent successive linear algorithm (SLA) [30–33], which could obtain an approximate local solution to our $L_p$-LSSVR. Furthermore, lower bounds for the absolute value of nonzero components are used in every approximate local solution, ensuring that all useful features can be selected [34]. Experimental results on both artificial and real-world data sets show the superiorities of our $L_p$-LSSVR. By using the absolute equality constraint, our $L_p$-LSSVR is more robust to outliers than LS-SVR, and much faster than $L_p$-SVR and SVR. In particular, compared results to those for $L_1$-SVR, LS-SVR, and SVR, our $L_p$-LSSVR not only selects fewer features but also has good regression effectiveness.

This paper is organized as follows. Section 2 of this paper briefly dwells on the SVR and LS-SVR. Section 3 proposes our linear and nonlinear $L_p$-LSSVR. Section 4 describes artificial and real-world data sets experiments and Section 5 concludes the paper.

## 2. Background

Starting with our notation, consider a regression problem in $n$-dimensional real vector space $R^n$. For column vector $x \in R^n$, $[x]_i$ denotes the $i$th component of $x, i = 1, 2, \dots, n$. $|x|$ denotes the vector in $R^n$ of absolute values of the components of $x$. $p$-norm $\|x\|_p$ ($0 < p < 1$) is defined as $(|[x]_1|^p + |[x]_2|^p + \cdots + |[x]_n|^p)^{\frac{1}{p}}$. A training set is denoted by $\{A, Y\}$, where $A$ is an $l \times n$ matrix, and the $i$th row $A_i \in R^n$ represents the $i$th training sample, $i = 1, 2, \dots, l$. $Y = (y_1, y_2, \dots, y_l)^T$ in $R^l$ denotes the response vector.

We next briefly review some algorithms that are closely related to our $L_p$-LSSVR. For simplicity, it is only concerned with the linear version. The optimal linear regression function is constructed as follows:

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

where $w \in R^n$ and $b \in R$.

### 2.1. Support vector regression

In conventional SVR [1], parameters in the linear regression function (1) are estimated by solving the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi,\eta} \quad & \frac{1}{2}||w||^2 + C(e^T\xi + e^T\eta) \\ s.t. \quad & Y - (Aw + eb) \le \varepsilon e + \xi, \xi \ge 0, \\ & (Aw + eb) - Y \le \varepsilon e + \eta, \eta \ge 0, \end{aligned} \tag{2}$$

where $||\cdot||^2$ represents the $L_2$-norm, $C > 0$ is a parameter determining the trade-off between the regularization term and empirical risk, $\xi$ and $\eta$ are slack variables, and $e$ is a vector of ones of appropriate dimensions. Generally, conventional SVR may use all features without discrimination for training because SVR solution $w$ lacks of sparseness.

To overcome this drawback, [19] replaced the $L_2$-norm penalty in (2) with $L_1$-norm penalty [16,27,28], and considered the $L_1$-support vector regression ($L_1$-SVR):

$$\begin{aligned} \min_{w,b,\xi,\eta} \quad & ||w||_1 + C(e^T\xi + e^T\eta) \\ s.t. \quad & Y - (Aw + eb) \le \varepsilon e + \xi, \xi \ge 0, \\ & (Aw + eb) - Y \le \varepsilon e + \eta, \eta \ge 0, \end{aligned} \tag{3}$$

where $\|\cdot\|_1$ represents the $L_1$-norm, $C$ is a positive parameter, and $\xi$ and $\eta$ are slack vectors. By using the $L_1$-norm, a small enough $C$ will drive some coefficients of $w_i$ towards zero [16,27,28]. This implies that $L_1$-SVR has an inherent feature selection property.

To improve the sparseness, Zhang et al. [15] proposed $L_p$-support vector regression ($L_p$-SVR) in the spirit of $L_p$-support vector machine ($L_p$-SVM) [34–38]. The primal $L_p$-SVR problem is formulated as follows: