Expert Systems With Applications 99 (2018) 126–140

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An effective way to integrate ε-support vector regression with gradients

XiaoJian Zhou[a,b], Ting Jiang[c,*]

[a] School of management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[b] Department of Statistics, Pennsylvania State University, PA 16801, USA
[c] School of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210023, China

ARTICLE INFO

Article history:
Received 6 August 2017
Revised 16 January 2018
Accepted 22 January 2018
Available online 3 February 2018

Keywords:
ε-support vector regression
Metamodel
Gradient information
Machine learning

ABSTRACT

ε-support vector regression (ε-SVR), as a direct implementation of the structural risk minimization principle rather than empirical risk minimization principle, is a new regression method with good generalization ability and can efficiently solve small-sample learning problems. In this work, through incorporating gradient information into the traditional ε-SVR, the gradient-enhanced ε-SVR (GESVR) is developed. The efficiency of GESVR is compared with the traditional ε-SVR by employing analytical function fitting, compared with the gradient-enhanced least square support vector regression (GELSSVR) by using two real-life examples, and tested in a scenario where the exact gradient information is unknown. The results show that GESVR provides more accurate prediction results than the traditional ε-SVR model, and outperforms GELSSVR in some real-life cases.

1. Introduction

Support Vector Machine (SVM) was first introduced by Vapnik (1995). The SVM was originally applied to deal with classification problems and soon extended to regression problems (Smola & Schölkopf, 2004). Compared with other estimation models, such as artificial neural networks (ANN), SVR employs Structure Risk Minimization (SRM) rather than the traditional Empirical Risk Minimization (ERM) principle to address the overfitting problem and thereby has higher generalization ability. When applied in a complex, nonlinear and unstable system, just as presented in Clarke, Griebsch, and Simpson (2005), the SVR can achieve better performance than other methods such as radial basis function (RBF) model (Fang, Li, & Sudjianto, 2006), least interpolating polynomials (Boor & Ron, 1990), inductive learning (Langley & Simon, 1995), multivariate adaptive regression splines (MARS) (Friedman, 1991) and Kriging (Cressie, 1988). Recent applications of ε-SVR into modeling can be traced in Zhou and Ma (2013); Zhou, Ma, and Li (2011); Zhou, Ma, Tu, and Feng (2012).

The construction of the traditional model does not consider the gradient information in the samples. If these gradients can be easily obtained, they should be used to improve the performance of the model. Zhou and Jiang (2016a) used the gradient information to enhance the prediction accuracy of least square support vector regression (LSSVR). Because ε-SVR is the most classical version of SVR, we try to construct the ε-SVR with gradient information in this paper. The process is quite different from the that of LSSVR, which is based on linear equations and therefore loses sparsity.

Just as presented in the work of Zhou and Jiang (2016a), there are a lot of methods to get the gradient information using different programming languages in engineering practice, such as ADOL-C(Griewank, Juedes, & Utke, 1996)/Adic(Bischof, Roh, & Mauer-Oats, 1997)/FADBAD(Bendtsen & Stauning, 1996) for C++, ADOL-F(Shiriaev & Griewank, 1996)/TAF (formerly TAMC)(Giering & Kaminski, 1998)/Adifor(Bischof, Khademi, Mauer, & Carle, 1996)/Tapenade(Hascoët, 2004) for Fortran, and ADMIT(Coleman & Verma, 2000)/ADiMat(Bischof, Bucker, Lang, Rasch, & Vehreschild, 2002) for MATLAB. The last one was adopted by Zhou and Jiang (2016a) to estimate the gradient information in their work. In this paper, we also use ADiMat to evaluate the gradients when the underlying function f required in ADiMat is available, otherwise, we employ the Kriging model to estimate the gradient information.

The remainder of this paper is organized as follows. In the next section, we introduce the construction of traditional ε-SVR. In Section 3, the gradient information in samples is considered in the process of construction of ε-SVR model, and the corresponding algorithm is presented. The presentation and discussion of results are displayed in Section 4. At last, a brief conclusion is discussed in Section 5.

* Corresponding author.
E-mail addresses: xjzhou@njupt.edu.cn (X. Zhou), jiangtinghaha@126.com (T. Jiang).

https://doi.org/10.1016/j.eswa.2018.01.036
0957-4174/

2. The traditional ε-SVR

This section focuses on describing the basic framework of traditional ε-SVR. Considering the work of Smola and Schölkopf (2004) has excellently explained its basic theory, we follow the corresponding part of the work of Smola and Schölkopf (2004) in this section.

Given the data set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$(where l denotes the size of the data set), the goal in ε-insensitive SVR is to find a function $f(x)$ that has at most ε deviation from the observed response value y_i for all the training data, and meanwhile, is as flat as possible. We first describe the case of linear functions f, which takes the form

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $\mathbf{w}$ is the weight vector. Flatness of f means minimizing $\|\mathbf{w}\|^2$. Formally $\mathbf{w}$ can be obtained by solving the following convex optimization problem

$$\min \frac{1}{2}\|\mathbf{w}\|^2$$
$$s.t. \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \le \varepsilon \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \le \varepsilon. \end{cases} \tag{2}$$

The tacit assumption in (2) is that the convex optimization problem is feasible. Nevertheless, sometimes this is not the case, and thus we should allow some errors. We can deal with otherwise infeasible constraints of the optimization problem (2) by introducing slacking variables ξ_i^* and ξ_i. Hence we arrive at the following formulation expressed as

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\xi_i^* + \xi_i)$$
$$s.t. \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \le \varepsilon + \xi_i^* \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \le \varepsilon + \xi_i \\ \xi_i^*, \xi_i \ge 0. \end{cases}$$

The user defined constant $C > 0$ controls the trade off between the flatness of f and the degree to which deviations larger than ε are tolerated. This method of tolerating error is known as the ε-insensitive loss function described by

$$L_\varepsilon(\xi) = \begin{cases} 0, & |\xi| < \varepsilon \\ |\xi| - \varepsilon, & other, \end{cases}$$

where ε is another parameter specified beforehand by the user.

The Lagrange function is constructed from the objective function and its corresponding constraints, by introducing Lagrange multipliers, $\alpha_i^*, \alpha_i, \eta_i^*, \eta_i$. Thus we proceed as follows:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\xi_i^* + \xi_i) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* - y_i + \langle \mathbf{w}, \mathbf{x}_i \rangle + b)$$
$$- \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i + y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b) - \sum_{i=1}^{l}(\eta_i^*\xi_i^* + \eta_i\xi_i). \tag{3}$$

The dual variables in (3) should satisfy $\alpha_i^*, \alpha_i, \eta_i^*, \eta_i \ge 0$. This function has a saddle point, which is obtained by minimizing of L with respect to the primal variables and maximizing with respect to the dual variables. The partial derivatives of L with respect to the primal variables ($\mathbf{w}$, b, ξ^*, ξ_i) must vanish for optimality.

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\mathbf{x}_i = 0, \tag{4}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0, \tag{5}$$

$$\frac{\partial L}{\partial \xi_i^{(*)}} = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0. \tag{6}$$

By substituting (4), (5) and (6) into (3), we can get the dual optimization problem.

$$\min_{\alpha^{(*)}} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)y_i$$
$$+ \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i)$$
$$s.t. \begin{cases} 0 \le \alpha_i^*, \alpha_i \le C, i = 1, \ldots, l, \\ \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0. \end{cases},$$

From (4) we obtain

$$\mathbf{w} = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\mathbf{x}_i,$$

and by substituting it into (1), the ε-SVR prediction is found to be

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)\langle \mathbf{x}_i, \mathbf{x} \rangle + b. \tag{7}$$

Intercept b can be computed as follows:

$$b = y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - \varepsilon \text{ for } \alpha_i^* \in (0, C),$$
$$b = y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle + \varepsilon \text{ for } \alpha_i \in (0, C).$$

In order to find a computationally cheaper way, the kernel function is introduced in (7), which may be written as

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)k(\mathbf{x}_i, \mathbf{x}) + b.$$

There are a lot of alternatives for the kernel function $K(\cdot, \cdot)$:

1. $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)$ (linear)
2. $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^m$ (m degree homogeneous polynomial)
3. $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + c)^m$ (m degree inhomogeneous polynomial)
4. $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$ (Gaussian)
5. $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sum_{k=1}^{l}\theta\left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\|^{p_k}))$ (Kriging)

In the above-mentioned kernel functions, the most popular one is the Gaussian kernel, which is also employed in this work.

3. The ε-SVR with gradient information

3.1. The primal optimization problem

After mapping the input space X into a feature space, $F = \{\phi(x) | x \in X\}$, the representation of the input variables can be changed to be:

$$\mathbf{x} = (x_1, \ldots, x_d)^T \mapsto \phi(\mathbf{x}) = (\phi(x_1), \ldots, \phi(x_d))^T.$$

The basic form of ε-SVR prediction will be the function of the type

$$f(\mathbf{x}) = \mathbf{w}^T \cdot \phi(\mathbf{x}) + b.$$

Then the first-order partial derivative of the $f(\mathbf{x})$ can be defined as follows:

$$D_r(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_r} = \mathbf{w}^T \cdot \frac{\partial \phi(\mathbf{x})}{\partial x_r}, \quad r = 1, \ldots, d,$$