Acta Automatica Sinica
Vol. 39, No. 3 March, 2013

Available online at www.sciencedirect.com
ScienceDirect

Flexible Support Vector Regression and Its Application to Fault Detection

Hui YI[1] Xiao-Feng SONG[1] Bin JIANG[1] Yu-Fang LIU[1, 2] Zhi-Hua ZHOU[2]

Abstract: Hyper-parameters, which determine the ability of learning and generalization for support vector regression (SVR), are usually fixed during training. Thus when SVR is applied to complex system modeling, this parameters-fixed strategy leaves the SVR in a dilemma of selecting rigorous or slack parameters due to complicated distributions of sample dataset. Therefore in this paper we proposed a flexible support vector regression (F-SVR) in which parameters are adaptive to sample dataset distributions during training. The method F-SVR divides the training sample dataset into several domains according to the distribution complexity, and generates a different parameter set for each domain. The efficacy of the proposed method is validated on an artificial dataset, where F-SVR yields better generalization ability than conventional SVR methods while maintaining good learning ability. Finally, we also apply F-SVR successfully to practical fault detection of a high frequency power supply.

Keywords: Support vector regression (SVR), flexible, fault detection, power supply

Support vector regression (SVR), which is known as one of the most powerful tools for function approximation, was proposed by Vapnik in 1995 with the principle of structure risk minimization (SRM)[1]. Compared with tools based on experience risk minimization (ERM), such as neural networks and least square methods, it requires less samples and yields better generalization ability, and hence has been widely used in fields like time series prediction[2−3], process control[4] and fault diagnosis[5]. Although SVR has been well studied and many remarkable achievements have been obtained, the theoretical estimation of regression parameter remains unsolved. There are some practical recommendations on this issue: Cherkassky et al.[6] used VC generalization bounds to control the model complexity in 1999; Chapelle and Vapnik[7−8] penalized the regression based on the ratio of expected training error and generalization error; Scholkopf et al.[9−10] proposed the famous v-SVR that can automatically choose an appropriate epsilon with given prior knowledge, and tackling their work, Kwok and Tsang[11] refined it in case of Gaussian noise.

As there is no general consensus on the setting of proper parameters, re-sampling approaches have wildly been employed in practical applications, such as cross validation[12], gradient descent[13] and swarm intelligence[14−16]. These methods can provide practical parameters, but are expensive in computational cost. In this case, Cherkassky et al.[17] proposed their empirical selection for the SVR parameters which gives a simple yet practical setting of regression parameters without re-sampling. However, all approaches mentioned above consider only in a global fashion while regressing, and in some complicated cases there can hardly find a set of parameters that fit the training samples well.

Recently, some adaptive methods have been proposed for setting regression parameters[18−20]. These methods are capable to adjust the parameters according to the sample distribution, hence perform better than parameter-fixed methods. For example, Hao[21] has changed the fixed parameter ε into a variable in the regression, which efficiently sets the parameter "epsilon" without re-sampling. However, it optimizes only one of the three parameters and is hard to give a better trade-off between over-fitting and under-fitting for some complicated cases.

Inspired by Cherkassky′s[17] and Hao′s[21] work, we proposed a new approach to optimize the process of regression. This approach flexibly divides the samples into several regions, and for each region, it selects proper parameters according to its distribution.

The paper is organized as follows. Section 1 presents the problem of existing approaches. Section 2 describes the proposed flexible support vector regression method. Section 3 presents the performance of the proposed method by comparisons with existing approaches, and applies it to the fault detection of a new kind of power supplies. Finally, conclusions and discussion are given in Section 4.

1 Problem formulation

Given training data $\{(x_1, y_1), \cdots, (x_l, y_l)\}$ with unknown

Manuscript received September 15, 2011; accepted October 14, 2012
Supported by National Natural Science Foundation of China (61034005, 61171191, 61203072, 61273171), the Science Foundation of Jiangsu Province (BK2010500), the Doctoral Fund of Ministry of Education of China (20113218110011), the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions, and the Open Project of State Key Laboratory of Industrial Control Technplogy (ICT1234)
Recommended by Associate Editor Mai-Ying ZHONG
Citation: Hui Yi, Xiao-Feng Song, Bin Jiang, Yu-Fang Liu, Zhi-Hua Zhou. Flexible support vector regression and its application to fault detection. *Acta Automatica Sinica*, 2013, **39**(3): 272−284
1. College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China 2. Guodian Science and Technology Research Institute, Nanjing 210031, China

distribution function, the core technique for SVR is to find a proper parameter vector p_0 and the optimal hyper-plane (the vector w_0):

$$(p_0, w_0) = \arg\min_{p,w} : R_{\mathrm{SRM}}(w,p) = R_{\mathrm{ERM}} + \phi(w) = \arg\min_{p,w} : \frac{1}{l}\sum_{i=1}^{l} Loss(y_i, f(x_i, w, p)) + \frac{1}{2}(w \cdot w), \tag{1}$$

where $p = \{C, \varepsilon, K(\cdot)\}$, C is the penalty parameter, ε is the insensitive parameter, $K(\cdot)$ is the kernel function, and $Loss(y_i, f(x_i, w, p)) = C_i \cdot \left| y_i - (\sum_{i=1}^{l} \beta_i K(x, x_i) + b) \right|_{\varepsilon}$. The intrinsic difference between support vector regression and conventional regressions is that SVR considers not only learning ability but also VC confidence (generalization ability). It can be solved by constructing the Lagrangian:

$$\min : L(\alpha) = \frac{1}{2}(w, w) + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(y_i - (w, x_i) - b + \varepsilon + \xi_i) - \sum_{i=1}^{l}\alpha_i^*(-y_i + (w, x_i) + b + \varepsilon + \xi_i^*) - \sum_{i=1}^{l}(\beta_i\xi_i + \beta_i^*\xi_i^*), \tag{2}$$

where α_i and $\beta_i \geq 0$ are the Lagrange multipliers

$$\begin{cases} \alpha_i, \beta_i > 0, & (x_i, y_i) \in \text{supportvectors}, \\ \alpha_i, \beta_i = 0, & \text{else}, \end{cases} \tag{3}$$

and ξ_i is the slack variable.

$$\xi_i = \left| y_i - \left(\sum_{i=1}^{l} \beta_i K(x, x_i) + b \right) \right|_{\varepsilon}. \tag{4}$$

Thus, the dual problem for (1) could be described as to maximize the following quadratic form

$$W(\alpha) = -\sum_{i=1}^{l}\varepsilon(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j), \tag{5}$$

subject to the constraints

$$\sum_{i=1}^{l}\alpha_i^* = \sum_{i=1}^{l}\alpha_i, \quad 0 \leq \alpha_i \leq C,\ 0 \leq \alpha_i^* \leq C, \quad i = 1, \cdots, l. \tag{6}$$

Here RBF kernel is employed for SVR:

$$K(x_i, x_j) = \mathrm{e}^{-\frac{\left\| x - x' \right\|^2}{2\sigma^2}}. \tag{7}$$

From (2)$\sim$(4), we can see that support vectors are termed to be the only objects providing information for approximating the unknown function $f(x)$. And the parameters $p : \{\varepsilon, \sigma, C\}$ may greatly affect the optimization for (5) and (6), and determine the number of support vectors. However, vector w_0 can be easily obtained by solving the above quadratic programming (QP) problem, whereas the theoretical estimation of parameter p_0 remains unsolved. More support vectors provide more information, but according to Occam′s Razor, no support vector should be adopted unless it is necessary. In Cristianini′s book[22], the following lemma has been proposed to illustrate how support vectors affect the generalization ability for regression.

Lemma 1. Considering thresholding real-valued linear function Γ with unit weight vectors on an inner production space X, for any probability distribution D on $X \times \{-1, 1\}$, with probability $1 - \delta$ over l random examples, the maximal margin hyperplane has error no more than:

$$\underset{\wp}{err} \leq \frac{1}{l - d}\left(d\log\frac{el}{d} + \log\frac{l}{\delta} \right), \tag{8}$$

where $d = \#sv$ is the number of support vectors.

This lemma indicates that the generalization ability for SVM learning is determined by the number of support vectors, and less support vectors yield a better generalization performance.

Generally, the challenge for SVR is the determination of support vectors. Over-needed support vectors may lead to over-fitting, otherwise, to under-fitting. But in some cases, it is hard to give a strike trade-off between over-fitting and under-fitting. Considering Fig. 1 as an example, samples in this figure could be divided into 3 parts $X = \{X_A, X_B, X_C\}$ according to the distribution complexity. In both A and C parts, the distribution is simple, and just a small quantity of support vectors are required for approximation. Oppositely the distribution in part B is much more complicated, and more support vectors should be employed to give a sufficient learning of potential information.

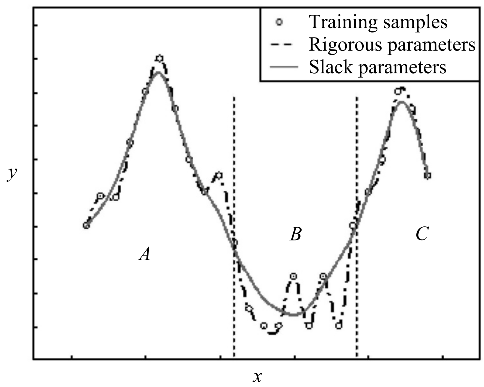

Fig. 1 Dilemma of SVR parameter setting

If a slack set of parameters p_1 is selected, the regression