Weighted Support Vector Regression Algorithm Based on Data Description

Weimin Huang   Leping Shen
School of Business Administration, South China University of Technology, Guangzhou, 510640, China
weimin.h@163.com ,lepingshen@163.com

Abstract

In order to overcome the overfitting problem caused by noises and outliers in support vector regression (SVR) ,a weighted coefficient model based on support vector data description (SVDD) is presented in this paper. The weighted coefficient value to each input sample is confirmed according to its distance to the center of the smallest enclosing hypersphere in the feature space. The proposed model is applied to weighted support vector regression (WSVR) for 1-dimensional data set simulation . Simulation results indicate that the proposed method actually reduces the error of regression and yields higher accuracy than support vector regression (SVR) does.

1. Introduction

In 1995, Vapnik V and the others put forward the support vector machine [1,2] (called SVM for short) based on statistical learning theory of finite sample; because it has firm theoretical basis and good generalization performance, and effectively solves a series of problems such as nonlinear and curse of dimensionality and so on, therefore it wins the wide attention once appearing, and is well applied in some fields (such as image recognition, genome sequencing analysis, isolated point detecting) [3,4] . Vapnik V put forward ε support vector regression, ε-SVR algorithm based on giving the definition of ε-insensitive loss function [1].

In the standard ε-SVR algorithm, the selection of design data C and ε is very important to construct the regression functions. The data ε indicates the error expectation (requirement to error) of the system having upon the estimation function in the sample data point. The smaller the ε, the less error requirement of the estimation functions in the sample data point and the higher degree of accuracy estimated by the function. The data C is the penalty for the sample data with its estimation function error larger than ε; the larger the C, the greater the penalty. In the standard ε-SVR algorithm, all the C and ε to which the sample corresponding are same with each other, i.e. for different sample data, the requirement to the degree of accuracy and the penalty for deviating the requirement to the degree of accuracy is undiscriminating, in which the same C leads to the very sensitivity of SVM to the isolated point, and therefore generates the over-fitting phenomenon[5-7]. Therefore, the scholars put forward the Weighted SVR, WSRR Algorithm to solve the problems above. For example, aiming at the problems of pattern recognition, the document [8] thinks that the requirement to the error of the first sample in sample set is lowest, being set as the value less than 1, the requirement to the error of the last sample in sample set is highest, being set as the value 1, and work out the weighted coefficient of the other samples through linear interpolation method; aiming at the problems of financial time series forecasting with drastic trending change, the documents [9,10] think that the requirement to the error of near-term is far higher than the that to the error of the long-term and adopts the indexical function of time order to represent the requirement to the error of every sample; the document [7] adopts measuring the requirement to the error of samples through detecting the Euclidean distance between the sample and training sample. However, the research above does neither give the systematic description upon the so-called method of WSVR, nor give the detailed description of optimization and training method, therefore it is necessary to seek the new WSVR method to solve the problem of sample data isolated point.

This paper puts forward a WSVR Algorithm based on Support Vector Data Description [11] (SVDD). Firstly, get the data domain description mode of the training set sample, and then endue different weighted coefficient upon its penalty factor C in accordance with the degree of each sample deviating from the data domain. Finally, apply the algorithm putting forward to the one-dimension artificial forecasting problem with the isolated point, and also compare with the traditional Support Vector Regression Algorithm.

2. ε Support Vector Regression Algorithm

Given l independent identically distributed data samples $(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_l,y_l)$, $x_i \in R^n$, $i=1,\cdots,l.$, based on giving the definition of ε insensitive loss function, ε-SVR algorithm seeks for an

978-0-7695-3290-5/08 $25.00 © 2008 IEEE
DOI 10.1109/CCCM.2008.25

IEEE computer society

optimized function $f(x)=(w\bullet x)+b$, in which $w,x\in R^n, b\in R$, so the forecasted expectation risk $R[f]$ is the minimum [2]:

$$R[f]=\frac{1}{2}\|w\|^2+C\bullet R_{emp}^{\varepsilon}[f]. \tag{1}$$

In which $\|w\|^2$ is the structural risk, representing the complexity of the mode, $R_{emp}^{\varepsilon}[f]=\frac{1}{2}\sum_{i=1}^{l}|y_i-f(x_i)|_{\varepsilon}$ is called the experience risk, representing the error of regression mode, C is the relaxation penalty factor, being used to balance the structural risk and experience risk. Minimum (1) is equal to the quadratic programming problem below:

$$\min_{w,\xi,\xi^*}\Phi=\frac{1}{2}\|w\|^2+C\frac{1}{l}\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$

$$s.t.\begin{cases} y_i-((w\bullet x_i)+b)\le\varepsilon+\xi_i \\ (w\bullet x_i)+b-y_i\le\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*\ge 0, i=1,\cdots,l \end{cases}. \tag{2}$$

In which ξ_i,ξ_i^* is the relaxation quantum, the domain between $y_i-((w\bullet x_i)+b)=\varepsilon$ and $y_i-((w\bullet x_i)+b)=\varepsilon$ is called regression interval [12].

Introduce the Lagrange coefficient a_i, a_i^*, and finally transfer the quadratic programming problem as the dual problem below:

$$\max W=\sum_{i=1}^{l}(a_i-a_i^*)y_i-\varepsilon\sum_{i=1}^{l}(a_i+a_i^*) \\ -\frac{1}{2}\sum_{i,j=1}^{l}(a_i-a_i^*)(a_j-a_j^*)\bullet(x_i\bullet x_j) \tag{3}$$

Find the solution of the quadratic programming problem above, and get the optimized Lagrange coefficient a_i, a_i^* and the threshold value b, the sample to which $a_i, a_i^*>0$ corresponding to is called the support vector (SV).

Under the condition of nonlinear, introduce the transformation $\Phi: R^n\rightarrow H$, and map the sample from the inputting space R^n to a high-dimension feature space H, and get the optimized function in H in order to get the minimum definite risk function, according to Mercer condition [12], exist the mapping Φ and kernel function $K(\bullet,\bullet)$, so $K(x_i,x_j)=\Phi(x_i)\bullet\Phi(x_j)$, regress the function when introducing the kernel function as:

$$f(x)=\sum_{i=1}^{l}(a_i-a_i^*)K(x_i,x)+b. \tag{4}$$

The frequently used kernel function is linear kernel $K(x_i,x_j)=x_i\bullet x_j$, ,polynomial kernel $K(x_i,x_j)=((x_i\bullet x_j)+1)^d$, and Gaussian radius basis function $K(x_i,x_j)=e^{\frac{-\|x_i-x_j\|^2}{\sigma^2}}$ and so on.

3. Weighted Support Vector Regression Algorithms

In the ε-SVR algorithm, the sample deviating the regression interval is endued with the same penalty factor C, every input sample is treated indiscriminately, but because the loss function value of the samples will be different due to the different degree from deviating the regression interval, the contribution to the experience risk $R_{emp}^{\varepsilon}[f]$ will be different, the sample further deviating from the regression interval, the more influence it exerts upon the regression function, therefore the fixed penalty factor C will make the regression function very sensitive to the isolated point, and then to mat lab these isolated points. Figure 1 is the schematic diagram indicating the change of regression interval before and after adding the isolated points. Seeing from the figure, the regression interval moves in the direction near to the isolated points because of the adding of isolated points, so the error is generated.

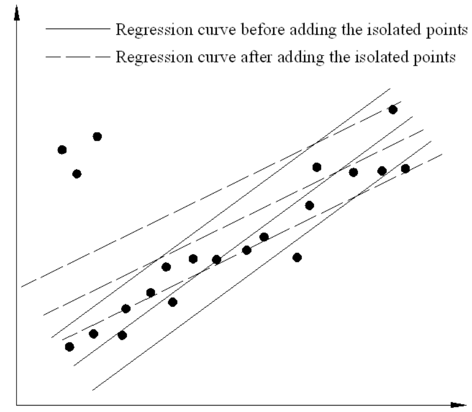
Regression curve before adding the isolated points
Regression curve after adding the isolated points

Figure1.Over fitting generated in identically distributed isolated points in ε-SVR

Introduce the weighted coefficient s_i for every sample, the inputting sample set after being weighted is $((x_1,y_1,s_1),\cdots,(x_i,y_i,s_i),\cdots,(x_l,y_l,s_l)), 0<s_i\le 1$, the object function in (2) is rewritten as

$$\min_{w,\xi,\xi^*}\Phi=\frac{1}{2}\|w\|^2+C\frac{1}{l}\sum_{i=1}^{l}s_i(\xi_i+\xi_i^*). \tag{5}$$

Introduce the Lagrange coefficient $a_i, a_i^*, \eta_i, \eta_i^*$ and construct the Lagrange function:

$$\begin{aligned} L = \frac{1}{2}\|w\|^2 + C\frac{1}{l}\sum_{i=1}^{l} s_i(\xi_i + \xi_i^*) \\ -\sum_{i=1}^{l} a_i(\varepsilon + \xi_i - y_i + w \bullet \Phi(x_i) + b) \\ -\sum_{i=1}^{l} a_i^*(\varepsilon + \xi_i^* + y_i - w \bullet \Phi(x_i) - b) \\ -\sum_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*). \end{aligned} \tag{6}$$

Let the partial differential coefficient of L for w, b, ξ_i, ξ_i^* is equal to zero, the equation below is generated:

$$\begin{cases} \dfrac{\partial L}{\partial w} = w - \sum_{i=1}^{l}(a_i - a_i^*)\Phi(x_i) = 0 \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{l}(a_i - a_i^*) = 0 \\ \dfrac{\partial L}{\partial \xi_i} = Cs_i - (a_i + \eta_i) = 0 \\ \dfrac{\partial L}{\partial \xi_i^*} = Cs_i - (a_i^* + \eta_i^*) = 0 \end{cases} . \tag{7}$$

Substitute the formula into (6), get its dual value of Wolfe:

$$\begin{aligned} \max W = \sum_{i=1}^{l}(a_i - a_i^*)y_i - \varepsilon\sum_{i=1}^{l}(a_i + a_i^*) \\ -\frac{1}{2}\sum_{i,j=1}^{l}(a_i - a_i^*)(a_j - a_j^*) \bullet (x_i \bullet x_j) \end{aligned} \tag{8}$$

$$s.t. \begin{cases} \sum_{i=1}^{l}(a_i - a_i^*) = 0 \\ 0 \le a_i, a_i^* \le s_i C, i = 1, \cdots, l. \end{cases}$$

When introduce the kernel function, the regression decision function is:

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x) + b. \tag{9}$$

Seeing from the constraint of (8), the upper bound of the Lagrange coefficient a_i, a_i^* is the function of weighted coefficient, therefore (9) is called the decisive function of WSVR.

4 .Confirmation of weighted coefficient

For confirming the form of weighted coefficient function, the extent the sample lean to the regression interval should be evaluated. The paper adopts method of support vector data description, mapping the data sample to a high-dimensional feature space, then seeking the smallest enclosing hypersphere in the high-dimensional space, and confirming the weighted coefficient value according to the distance from the sample to the center of hypersphere.

4.1 Support Vector Data Description

To present the training sample set $\chi = \{x_1, \cdots x_i, \cdots x_l\}$ inputting the space, of which. $x_i \in R^n, i = 1, \cdots, l$ For setting up data domain description model of sample, map the sample inputting the space to a high-dimensional feature space F through the map $\Phi : R^n \to F$, then seek the smallest enclosing hypersphere in the feature space, which can be concluded as solution of following quadratic programming:

$$\begin{aligned} \min W(\xi_i, R, a) = R^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \begin{cases} \|\Phi(x_i) - a\|^2 \le R^2 + \xi_i \\ \xi_i \ge 0, i = 1, 2, \cdots, l \end{cases} . \end{aligned} \tag{10}$$

Where R is the radius of the smallest enclosing hypersphere in feature space, a is the center of hypersphere, ξ is slack variable, and C is penalty factor. To introduce Lagrange coefficient β_i, η_i, to define Lagrange function as:

$$\begin{aligned} \max L = R^2 + \sum_{i=1}^{l}(R^2 + \xi_i - \|\Phi(x_i) - a\|^2)\beta_i \\ -\sum_{i=1}^{l}\xi_i\eta_i + C\sum_{i=1}^{l}\xi_i \end{aligned} \tag{11}$$

$$s.t. \begin{cases} \|\Phi(x_i) - a\|^2 \le R^2 + \xi_i \\ \xi_i \ge 0, \beta_i \ge 0, \eta_i \ge 0, i = 1, \cdots, l \end{cases} .$$

To make partial derivative of L for variable equal to zero and then following equations can be obtained:

$$\begin{cases} \sum_{i=1}^{l}\beta_i = 1 \\ a = \sum_{i=1}^{l}\beta_i\Phi(x_i) \\ \beta_i = C - \eta_i \end{cases} \tag{12}$$

Make use of the above equations for objective function in equation (11), and then Wolfe dual is solved as:

$$\max Q = \sum_{i=1}^{l} K(x_i, x_i)\beta_i - \sum_{i,j=1}^{l} \beta_i \beta_j K(x_i, x_j). \tag{13}$$
$$s.t. 0 \le \beta_i \le C, i = 1, \cdots, l$$

To solve the above quadratic programming can obtain the best Lagrange and the data domain description of sample in the feature space. Sample corresponding to non-zero Lagrange coefficient is support sector, its map in the feature space is on the sphere of enclosing hypersphere or outside the hypersphere.

4.2 Weighted coefficient function model based on data domain description

The distance from the map $\Phi(x_i)$ of point x_i inputting feature space to the center of the smallest enclosing hypersphere is defined as $D^2(x_i) = \left\| \Phi(x_i) - a \right\|^2$, considering that $a = \sum_{i=1}^{l} \beta_i \Phi(x_i)$ has

$$D^2(x_i) = \sum_{i,j=1}^{l} \beta_i \beta_j K(x_i, x_j) + K(x_i, x_j) \\ -2\sum_{j=1}^{l} K(x_i, x_i)\beta_j, i, j = 1, \cdots, l \tag{14}$$

To define $\chi_{NBSV} = \{x_1, \cdots, x_k, \cdots, x_m\}$ as the subset inputting the space, of which x_k is the non-edge support vector in the sample, $(0 < \beta_i \le C)$ is the number of non-edge support vector, the radius of the smallest enclosing hypersphere in the feature space is $R = D(x_i) \mid x_i \in \chi_{NBSV}$, then after fixing R and a, the data domain description of data set can be obtained, formulas in definition (15) are respective the max and min distances from the sample to the center of the smallest enclosing hypersphere.

$$D_{\max} = \max(D(x_i) \mid x_i \in \chi) \\ D_{\min} = \min(D(x_i) \mid x_i \in \chi). \tag{15}$$

To define weighted coefficient function as follows:

$$s_i = \begin{cases} (1 - \dfrac{D(x_i) - D_{\min}}{D_{\max} - D_{\min}})^f + u, R < D(x_i) \le D_{\max} \\ 1 - \dfrac{D(x_i) - D_{\min}}{D_{\max} - D_{\min}}, D_{\max} \le D(x_i) \le R \end{cases} \tag{16}$$

Where $u < 1$ is the sufficient small positive real number, and $f \ge 2$. In case of $f = 2$, the figured diagram of weighted coefficient function is as figure 2.

As to point x_i inputting the space, the distance from its map in the feature space to the center a of the smallest enclosing hypersphere is meet the expression $D_{\min} \le D(x_i) \le D_{\max}$, in case of $D_{\min} \le D(x_i) \le R$, x_i is in the support vector regression to meet data domain description, and the sample is near the regression interval, as sample 1, 2, 3, 4, 5 in figure 3 show, its weighted coefficient will be linearly decreased with the increase of $D(x_i)$, while it should be paid attention to that the sample 1 is non-support vector, sample 2 is non-edge support vector, then with the KKT condition[12] in the quadratic optimation, their slack variable are all zero, in case of $R < D(x_i) \le D_{\min}$,the sample x_i will lean to regression interval; as sample 6, 7, 8 in figure 3 show, its weighted coefficient is the quadratic function of $D(x_i)$, with the increase of $D(x_i)$, their weighted coefficients will rapidly decrease, in case $D(x_i)$ nearly equal to $D_{\max}$, its weighted coefficient will be a smaller positive real number u, then affection from these points on the regression function can be reduced.

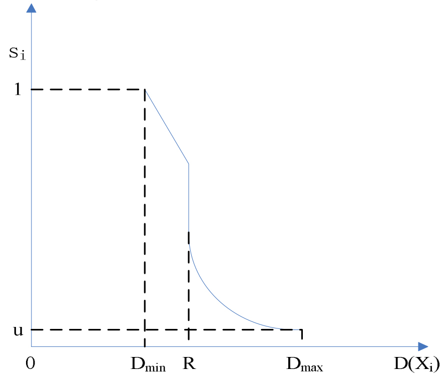

Figure2. weighted coefficient function based on SV data domain Description

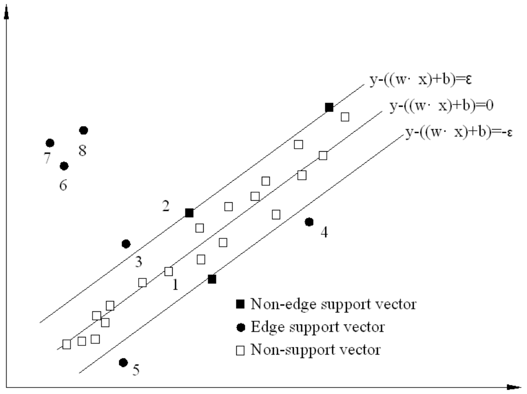

Figure 3. SVR schematic diagram

5. Simulation Experience and Analysis

Take approximated one-dimension function

$f(x)=\frac{\sin x}{x}+\zeta$ as an example, equably take 100 data from x interval of definition and add 12 isolated point data, of which $E\zeta=0, E\zeta^2=(0.05)^2$ .For choices of ε-SVR parameters, Gaussian radius kernel is chosen by kernel function, of which $\sigma=2, C=10$, $\varepsilon=0.075$.the program is made with MATLAB6.5. As regression result of ε-SVR shows, mean Square error is 0.042. The conclusion can be made from the figure, the existence of isolated point make the regression line (curve in the regression interval line) lean to direction of isolated point, and then cause the error.

For training in the data set $(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_l,y_l)$ by using WSVR, the input of radius R for the smallest enclosing hypersphere and $D_{\min}, D_{\max}$ in the sample set $(x_1,\cdots,x_i,\cdots x_l)$ should be confirmed at first. The SVDD is adopted for training, the Gaussian radius kernel serves as kernel function, of which $\sigma=2, C=10$. After the training, $R=1.068$, $D_{\min}=0.994, D_{\max}=1.179$, and select $f=2$, $u=0.005$, the weighted coefficient function can be:

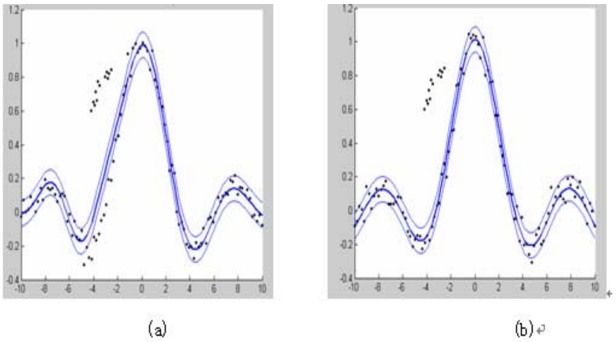

Figure 4. Regression results of ε-SVR and WSV

$$s_i=\begin{cases}\left(1-\dfrac{(D(x_i)-0.994)}{(1.179-0.994)}\right)^2+0.005, \\ \quad 1.068<D(x_i)\leq 1.179 \\ \left(1-\dfrac{(D(x_i)-0.994)}{(1.179-0.994)}\right), \\ \quad 0.994\leq D(x_i)\leq 1.068\end{cases} \qquad (17)$$

To adopt WSVR to train, the Gaussian radius kernel serves as kernel function, $\sigma=2, C=10, \varepsilon=0.075$. The result is shown as figure 4(b), regression mean square error is 0.003, and the conclusion can be made that affection from isolated point on the regression interval line is reduced, and the regression error is significantly minished.

6.Conclusions

The paper puts forwards a weighted coefficient function model based on support vector data domain description, mapping the data sample to a high-dimensional feature space. The weighted coefficient value is confirmed according to its distance to the center of the smallest enclosing hypersphere in the feature space. Simulation experiment proofs this method can reduce the overfitting problem caused by existence of isolated point in the data sample of ε-SVR,and minish regression error to enhance anti-noise ability of support vector regression.

Reference

[1] Vapnik V. (1999).An overview of statistical learning theory. *IEEETrans. On NN,* Vol. 10, No. 3, pp: 988-999.

[2] Cortes C, Vapnik V. (1995). *Support vector networks.* Machine Learning, Vol. 20, No. 4, pp: 273~297.

[3] Osuna E, Freund R, Girosi F.(1997). T*raining support vector machines: an application to face detection.* Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. LosAlamitos,CA,USA: IEEE Computer Society, pp:130~136.

[4] Guyon I, Weston J, Barnhill S, et. al. (2002). *Gene selection for cancer classification using support vector machines*.Machine Learning, Vol.46, No.6, pp: 389～422.

[5] Burges C J C.(1998). *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery, Vol. 2, No. 2,pp:1~47.

[6] Zhang X G. (1999).*Using class-center vectors to build support vector machines*. Neural Networks for Signal Processing IX - Proceedings of the 1999 IEEE Workshop. Wisconsin: IEEE, pp.33~37.

[7] Du Shuxin, Wu Tiejun. (2004) .*Weighted support vector machines for regression and its application.* Journal of Zhejiang University (Engineering Science),Vol.38, No.3,pp:302~306 (in Chinese)

[8] Lin CF, Wang S D.(2002).*Fuzzy support vector machines*. IEEE Transon Neural Networks, pp: 464~471.

[9] Toy F E H, Cao L J. (2002).*Descending support vector machines for financial time series forecasting.* Neural Processing Letters, Vol. 15, No. 2,pp:179~195.

[10] Toy F E H,Cao L J. (2002).Modified support vector machines in financial time series forecasting. Neuro computing , Vol. 48,pp:847~861.

[11] Tax D M J,Duin R P W. (2004).Support vector data description. Machine Learning,Vol. 54, No. 1,pp:45~66.

[12] Deng Naiyang, Tian Yingjie. (2004)New Data Mining Method- SVM. Beijing Science Press, 2.pp:96~166. (in Chinese) .