An algorithm of Model Selection for Support Vector Regression

Li Xuesi ,Yang Hongqiao, Sun Jing
The 309th Hospital of PLA

Bi Yangang
Tsinghua University

Wu Yuanli
The 309th Hospital of PLA

***Abstract*-To solve the problem of SVR (support vector regression) model selection, this paper proposed a SVM (support vector machine) model parameter optimization algorithm based on gradient descent algorithm. The algorithm obtained the local optimal model parameter by minimizing the model evaluation criteria over the parameter set. Then on the basis of Riemannian geometry, a conformal transformation suitable for SVR was proposed which corrected kernel function in a data-based way. This algorithm can further enhance the generalization ability of SVR. The simulated results are illustrated to show the feasibility and effectiveness of the algorithm.**

Keywords-support vector regression (SVR), model selection, gradient descent, Riemannian geometry

I. INTRODUCTION

Support vector machine (SVM) is a general learning algorithm for small-sample learning problems. It's presented through using structural risk minimization (SRM) based on statistic learning theory. SVM achieves actual risk minimization by seeking SRM, so it can obtain better learning results with fewer samples[1, 2].

SVM is mainly used in the field of pattern recognition, function regression and probability density estimation, etc. Support Vector Regression (SVR) is an important branch of SVM. SVR has been applied to system identification, nonlinear system prediction, but it also has problems to be solved. Model selection for SVR is the key issue in practical application, because it will significantly affect the SVR ability to predictive the system state parameters.

The problem of the SVR model selection can be described as this: for a certain type of process and a SVR algorithm, how to choose the most appropriate type of the kernel function, how to optimize the kernel parameters, and how to correct kernel function for the actual process. Studies have shown that in the absence of priori knowledge(noise, etc.) of the process, Gauss kernel function is better[3]. So, most researches use Gauss kernel, and then study the optimization of parameters.

$$K(x,x_i)=\exp(-\frac{|x-x_i|^2}{2\sigma^2}) \tag{1}$$

In Gauss kernel, σ and the regularization parameter C are to be optimized.

Currently, for the SVM classifier, kernel parameters optimization commonly uses the approaches based on evaluation criteria or Bayesian evidence framework. The approaches based on evaluation criteria which account for a large class are based on the well-known Leave-one-out(LOO) method, by estimating the upper bound of error rates, to find the optimal value of parameters. According to the selected error-bounds, gradient descent method, pattern search method, genetic algorithm[4] can be used to optimize the parameters.

Due to the differences between pattern recognition and function regression, the SVR model selections are also different. The model selection relies on two aspects: selection criteria and search method. This paper proposed an gradient descent algorithm based on minimizing R2w2 to optimize the kernel parameters(R is the smallest radius of the hypersphere containing all samples, w is the parameter of the linear function set). Because the characteristics of the actual process have not been taken into account when selecting the kernel function type, we need to correct the kernel function from the data-based perspective to improve the prediction accuracy of SVR.

II. THE MODEL SELECTION CRITERIA

In this paper ,we discuss the problem of L2 norm nonlinear soft margin SVR. Given training examples $T=\{(x_1,y_1),...,(x_l,y_l)\}$, where $x_i\in\mathbb{R}^n$, $y_i\in\mathbb{R}$, i=1, … ,n, this optimization problem can be converted into a dual form:

$$\max_{\alpha} W(\alpha)=-\sum_{i=1}^{l}\varepsilon_i(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)[K(x_i,x_j)+\frac{1}{C}I]$$
$$s.t.\ \sum_{i=1}^{l}\alpha_i=\sum_{i=1}^{l}\alpha_i^*;\alpha_i,\alpha_i^*\ge 0,i=1,2,...,l \tag{2}$$

Compare (2) and the problem of L2 norm hard margin SVR, we can find (2) has a solution can be introduced as a hard margin SVR problem as (3) is solvable.

$$\min J(\omega,\xi^{(*)})=\frac{1}{2}\omega^T\omega$$
$$s.t.\begin{cases}\omega^T\phi'(x_i)+by_i-y_i\le\varepsilon\\ y_i-\omega^T\phi'(x_i)-b\le\varepsilon\end{cases} \tag{3}$$

The corresponding dual form is as

$$\max_{\alpha} W(\alpha)=-\sum_{i=1}^{l}\varepsilon_i(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K'(x_i,x_j)$$
$$s.t.\ \sum_{i=1}^{l}\alpha_i=\sum_{i=1}^{l}\alpha_i^*;\alpha_i,\alpha_i^*\ge 0,i=1,2,...,l \tag{4}$$

where $K'(x,x')=K(x,x')+1/C$.

So, the general hard margin non-solution problem can be converted into solving the condition like (3) [5]. (T4)

shows that the regularization parameter C of L2 norm nonlinear soft margin SVR can be optimized as one of the kernel function parameters.

This is a quadratic optimization problem, from which we can get the optimal solution $\alpha^0 = (\alpha_1^0, \ldots, \alpha_l^0, \alpha_1^{*0}, \ldots, \alpha_l^{*0})$, and the training sample which meets $\alpha_i^{*0} - \alpha_i^0 \neq 0$ is the support vector. The regression function consisted by solutions of (T3) is as

$$f(x, \alpha^0) = \sum_{i=1}^{l} (\alpha_i^{*0} - \alpha_i^0) K'(x, x_i) + b \tag{5}$$

where b is determined by the KKT conditions.

Whether the model is appropriate depends on the model selection criteria. Currently using SVR, the judgement is mainly from the statistical results of actual data, such as the average relative error and the average absolute error, etc.

Literature [1] indicates that, for the linear function set (as T4) which used to estimate the real function value, VC dimension h satisfies the following equation:

$$h = \min([\frac{D^2}{\rho^2}] + 1, n), \tag{6}$$

where D=2R, ρ is the interval of the sample set and ρ=1/w, w is a parameter of the linear function set. Symbol [θ] represents the integer part of θ, N is the dimension of the input vector X transformed into the feature space. For the feature space of Gauss kernel is infinite-dimensional, we just have to consider D2/ρ2 (that is, to consider R2w2). According to statistical learning theory, the smaller VC dimension of function set, the smaller confidence interval of the risk experience, and then the better its generalization performance will be. So if we can get a kernel parameter which leads to the smallest R2w2, it is the optimal kernel parameter.

III. SOLVING THE OPTIMAL KERNEL PARAMETER

In this section, firstly, a lemma will be introduced. Then, according to the lemma, we will calculate the gradients of w2，R2 and T = R2w2 related to the kernel parameters. Finally, the optimal kernel parameters will be obtained by using gradient descent algorithm.

A. Lemma 1

Assuming vector νθ and matrix Pθ are a continuous function of θ, consider the following function

$$L(\theta) = \max_{x \in F} x^T v_\theta - \frac{1}{2} x^T P_\theta x \tag{7}$$

where $F = \{x : b^T x = c, x \geq 0\}$, Let $\bar{x} \triangleq \arg\max_x L(\theta, x)$, if it has sole maximum value, then

$$\frac{\partial L(\theta)}{\partial \theta} = \bar{x}^T \frac{\partial v_\theta}{\partial \theta} - \frac{1}{2} \bar{x}^T \frac{\partial P_\theta}{\partial \theta} \bar{x} \tag{8}$$

This lemma shows that, to calculate the gradient of L related to the parameter θ, we only need to calculate the gradients of νθ and Pθ, calculation for X is unnecessary. It also can be proved that even in the case of constraint F cancellation, this lemma is still valid.

B. Gradient calculation of w2

Suppose the vector constituted by kernel parameter and C denoted by θ. According to the characteristics of L2 nonlinear norm hard margin SVR (3), we obtain

$$\|w^2\| = 2 \max_\alpha W(\alpha) = 2W(\alpha^0), \tag{9}$$

and we also have

$$\|w^2\| = \sum_{i,j=1}^{l} (\alpha_i^{0*} - \alpha_i^0)(\alpha_j^{0*} - \alpha_j^0) K(x_i, x_j) \tag{10}$$

the according to lemma 1, we obtain $\frac{\partial \|w\|^2}{\partial \theta} = \sum_{i,j=1}^{l} (\alpha_i^{0*} - \alpha_i^0)(\alpha_j^{0*} - \alpha_j^0) \frac{\partial K(x_i, x_j)}{\partial \theta}$.

C. Gradient calculation of R2

R can be calculated by solving the following optimization problem[1]:

$$R^2 = \max_\beta \sum_{i=1}^{l} \beta_i K(x_i, x_i) - \sum_{i,j=1}^{l} \beta_i \beta_j K(x_i, x_j) \tag{11}$$

$$s.t. \sum_{i=1}^{l} \beta_i = 1; \forall i, \beta_i \geq 0;$$

According to lemma 1, we obtain

$$\frac{\partial R^2}{\partial \theta} = \sum_{i=1}^{l} \beta_i \frac{\partial K(x_i, x_i)}{\partial \theta} - \sum_{i,j=1}^{l} \beta_i \beta_j \frac{\partial K(x_i, x_j)}{\partial \theta}.$$

D. Model selection algorithm

By calculating the gradient of T = R2w2 related to the kernel parameters，

$$\frac{\partial T}{\partial \theta} = \frac{\partial R^2}{\partial \theta} w^2 + \frac{\partial w^2}{\partial \theta} R^2, \tag{12}$$

we can get certain θ that leads the minimum value of T using gradient descent method. Then obtain the optimal kernel parameter.

The corresponding model selection algorithm is as follows:

Step1. Initialization of the parameter θ

Step2. Using aforementioned SVR algorithm, we can optimize quadratic W(α) and R2(β), calculate the parameters $\alpha^0(\theta) = \arg\max_\alpha W(\alpha, \theta)$ and $\beta^0(\theta) = \arg\max_\beta R^2(\beta)$ to get the maximum W(α) and R2(β).

Step3. Calculate $\theta = \arg\min_\theta T(\alpha^0, \theta)$ by using gradient descent algorithm.

Step4. If T has reached the minimum value, stop; else, go to step2.

E. Algorithm description

Lemma 1 guarantees the gradient of model evaluation R2w2 related to the kernel parameters not depend on α and β, so it ensures that the updates of α and β in step2 will not affect T violently in step3. Therefore, T's gradient descent algorithm is monotonic decline.

IV. KERNEL FUNCTION CORRECTION BASED ON RIEMANNIAN GEOMETRY

A. Amari's kernel function correction method

According to Riemannian structure of input induced space, Japanese scholars Amari [7] proposed a method that correct the kernel function based on data.

Amari discussed the geometry problem of kernel-based non-linear SVM in [7], the Riemann matrix and kernel function are related as follows:

$$g_{ij}(x)=\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}K(x,x')|_{x'=x} \tag{13}$$

In the feature space, the relationship between the local volume differential and the input space is as follows:

$$dV=\sqrt{g(x)}dx_1dx_2...dx_d \tag{14}$$

Where g(x) = det| gij(x)|, and amplification factor $\sqrt{g(x)}$ expresses the way of the input space's local area to be magnified in the feature space F under the mapping φ.

According to (13) and (14), we know that local amplification factor is closely related with the kernel function at the point. The basic idea of this algorithm is increasing the space decomposition force of SVM's boundary, by changing the volume elements in different regions, and reducing the amplification factor in other regions at the same time.

B. Conformal transformation

Definition.1(1999,Amari[7])

$$\tilde{K}(x,x')=D(x)D(x')K(x,x') \tag{15}$$

is called a conformal transformation of a kernel by factor D(x), $K(x,x')$ is the modified function of SVM.

Amari proposed a conformal transformation, and Feng Jun proposed two new ones based on Amari's work [8], and obtained better experiment results. But all the aforementioned methods can not be used in SVR because that SVR has no priori classification label. Therefore, we propose a new conformal transformation as follows to improve it.

$$D_1(x)=\sum_{i\in sv}\exp(\frac{-\|x-x_i\|^2}{\tau_i^2}),\ \tau_i^2=\|x_m-x_i\|^2 \tag{16}$$

where

$$x_m=\begin{cases}x_m^+, y_i-(w,\phi(x))-b>0\ (1)\\ x_m^-, y_i-(w,\phi(x))-b<0\ (2)\end{cases} \tag{17}$$

$$x_m^+=\frac{1}{n_{sv}^+}\sum_{i=1}^{n_{sv}^+}x_i,\ x_m^-=\frac{1}{n_{sv}^-}\sum_{i=1}^{n_{sv}^-}x_i, \tag{18}$$

n_{sv}^+ is the number of the vectors which satisfy <T17-1>, x_m^+ is the mean eigenvalue of the support vectors; n_{sv}^- is the number of the vectors which satisfy <T17-2>, x_m^- is the mean eigenvalue of the support vectors. So τ_i^2 is the Euclidean distance between the support vector Xi and the eigenvector centrality of the class which Xi is belonging to.

According to the nature of the kernel function [2], K1(x,x')= D(x)D(x') is a positive definite kernel, and Gauss kernel K2 is a positive definite kernel. So $\tilde{K}(x,x')=K_1(x,x')K_2(x,x')$ is a positive definite kernel too.

C. The kernel function correction algorithms based on data

Setp1. According to the problem, initially determine the type of kernel function. Then optimize the kernel parameters by using the algorithm this paper proposed.

Step2. To obtain the information of the support vectors by training SVR using kernel function K, we correct it according to (15), (16), (17) and (18) to obtain $\tilde{K}$, then let $K=\tilde{K}$.

Step3. Repeat step 2 until the prediction accuracy is good enough.

V. SIMULATION EXAMPLES AND ANALYSIS

In this section, an auto-regressive test of function (19) will be given to verify the effectiveness of the algorithm.

$$f(t)=2\sin(0.03t)\mathrm{sinc}(0.8t-3)+0.5e^{0.05t}+e(t) \tag{19}$$

where t ∈ [0,8], e(t) is uniformly distributed white noise which amplitude is 0.05 and auto-regressive length is 3. Train it using L2 norm hard margin SVR(4), and take the first 70% of the data as training samples, the remain 30% as prediction samples. Use Gauss kernel and the initial values are C=0.1, σ =5.

We use $err=\frac{1}{n}\sum_{i=k}^{n}|\frac{y_i-y_i^*}{y_i}|$ as the prediction error indicators, where yi is the actual value of the prediction sample, y*i is corresponding calculated one.

Using the kernel parameter optimization algorithm proposed in this paper, we obtain the kernel parameter values as C=57.7, σ =5.64, the prediction error decrease from the initial 11.3% to 4.1% in the final. Corresponding prediction error and model evaluation R2w2 increasing trend with iteration steps are shown in Figure. 1.

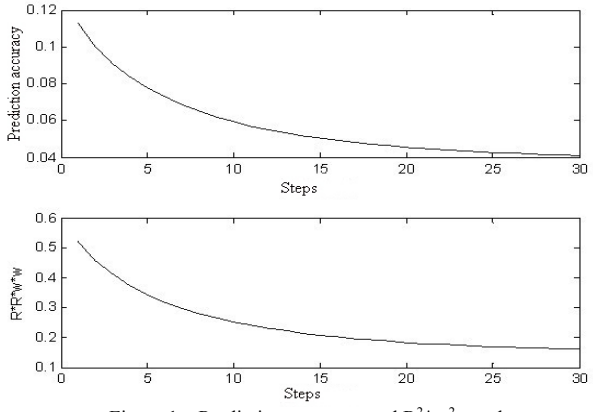

Figure.1. Prediction accuracy and R^2*w^2 trend

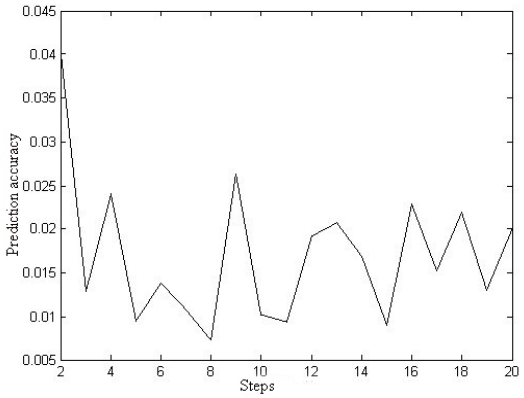

Figure.2. Kernel function correction results

After obtain the aforementioned local optimum kernel function, we correct it iteratively with the conformal transformation that proposed in this paper, the result is shown in Figure. 2. The prediction error decrease from 4.1% to 2%, the prediction accuracy of SVR is further improved.

VI. CONCLUSION

This paper proposes an automatic model selection algorithm for SVR, it has the advantage that the calculation process is monotonic decline; it also has disadvantage that in each iteration step it should solve two single-objective optimization problems, and the calculation speed will be slow when the sample set is large. The algorithm solves the model selection problem which is very important in practical engineering; the simulation results show that this joint algorithm is very effective.

REFERENCES

[1] VAPNIK V N. Statistical learning theory [M]. New York: Wiley, 1998.

[2] DENG Nai-yang; TIAN Ying-jie. New methods in data mining-Support Vector Machine [M]. Beijing: Science Press, 2004.

[3] SMOLA A.J. Learning with kernels [D]. Berlin: Technical University of Berlin, 1998.

[4] LIU Huan-Jun; WANG Yao-Nan; et al. A Method to choose kernel function and its parameters for support vector machines [C]. Machine Learning and Cybernetics, ICMLC2005. roceddings.2005, pp4277-4280.

[5] CRISTIANINI N; SHAWE T J. An introduction to support vector machines [M]. Cambridge: Cambridge University Press, 2004.

[6] CHAPELLE O; VAPNIK V. Choosing multiple parameters for support vector machines [J]. Machine learning. 2002, 46(1):131-159

[7] AMARI S; WU Si. Improving support vector machine classifiers by modifying kernel functions [J]. Neural networks, 1999, 12(2), pp783-789.

[8] FENG Jun; CHEN Zhi-jun; LI Li-rong. Support Vector Machines Classifier Based on Modifying Kernel Function [J]. System Simulation. 2006, 18(3), pp570-576.

[9] YUAN Xiao-fang; WANG Yao-nan. Parameter selection of support vector machine for function approximation based on chaos optimization [J]. Journal of Systems Engineering and Electronics, 2008, 19(1), pp191-197.

[10] HUANG C; DUN J. A distributed PSO-SVM hybrid system with feature selection and parameter optimization [J]. Applied Soft Computing. 2008, 8(4), pp138l-1391.