An Iterative Modified Kernel for Support Vector Regression

Fengqing Han, Zhengxia Wang, Ming Lei and Zhixiang Zhou
School of Science
Chongqing Jiaotong University
Chongqing City, China

***Abstract*—In order to improve the performance of a support vector regression, a new method for modified kernel function is proposed. In this method the information of whole samples is included in kernel function by conformal mapping. So the Kernel function is data-dependent. With random initial parameter of kernel function, iterative modifying is not stopped until satisfactory effect. Comparing with the conventional model, the improved approach does not need selecting parameters of kernel function. Simulation results show that the improved approach has better learning ability and forecasting precision than traditional model.**

***Keywords*—support vector regression, data-dependent, kernel, iteration**

I. INTRODUCTION

In recent years, a new pattern classification algorithm, called Support Vector Machine(SVM) has proposed by Vapnik[1] is a powerful methodology for solving a wide variety of problems in nonlinear classification, function estimation, and density estimation. Unlike traditional neural network models which minimize the empirical training error, SVM implements the structural risk minimization principle which seeks to minimize the training error and a confidence interval term. This results in a good generalization error. Because of its good properties such as global optimal solution, good learning ability for small samples and so on, the SVM has received increasing attention in recent years[2-4]. Moreover it has been successfully applied to the support vector regression (SVR), especially for nonlinear time series[5-7].

The performance of SVM largely depends on the kernel. Smola[8] elucidated the relation between the SVM kernel method and the standard regularization theory. However, there are no theories concerning how to choose good kernel function in a data-dependent way. In paper [9], the authors propose a method of modifying a kernel to improve the performance of a Support Vector Machine classifier. It is based on the Riemannian geometrical structure induced by the kernel function. In order to increase the separability, the idea is to enlarge the spatial resolution around the separating boundary surface with a conformal mapping. In this method, a primary kernel is used to obtain support vectors. Then the kernel is modified conformally in a data dependent way by using the information of the support vectors. The Final classifier is trained by the modified kernel. Inspired by this idea, liang yanchun[10] apply the modified kernel to SVR and forecast financial time series.

These methods have achieved better performance than the conventional SVM, but we need choose parameters of the modified kernel carefully.

The goal of this paper is to propose a new method to train SVM, which modifies the kernel function repeatedly. Unlike traditional methods which select parameters of the kernel function carefully, the parameters of this algorithm are random. Simulation experiments show that the new method is obviously superior to the traditional SVM in the precision of prediction.

The remainder of this paper is organized as follows. Section 2 briefly introduces the learning of SVR. Section 3 provides a background to data-dependent Kernel. Our method to train SVR by iterative learning will be introduced in Section 4. Section 5 presents some computational results to show the effectiveness of our method. Section 6 concludes our works.

II. SUPPORT VECTOR REGRESSION

Let $\{x_i, y_i\}, i = 1, 2, \cdots, m$ be a given set of training data, where $(x_i, y_i) \in R^n \times R$. The output of the SVR is

$$f(x) = < w, \phi(x) > + b, \tag{1}$$

where w is the weight vector, b the bias and $\phi(x)$ the nonlinear mapping from the input space S to the high dimensional feature space F. $< \cdot, \cdot >$ represent the inner product.

The commonly used ε-insensitive loss function introduced by Vapnik is

$$L_\varepsilon(x_i) = \begin{cases} | f(x_i) - y_i | - \varepsilon, & | f(x_i) - y_i | \geq \varepsilon \\ 0, & | f(x_i) - y_i | < \varepsilon \end{cases}. \tag{2}$$

In order to train w and b, the following function is minimized

$$\min Z = \frac{1}{2}\| w \|^2 + C\sum_i (\xi_i^+ + \xi_i^-)$$

$$s.t.\begin{cases} y_i - < w, \phi(x_i) > -b \le \varepsilon + \xi_i^+ \\ -y_i + < w, \phi(x_i) > +b \le \varepsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \ge 0, i = 1,2,\cdots m \end{cases}, \quad (3)$$

where C is the regularized constant determining the trade-off between the empirical error and the regularization term.

After the introduction of positive slack variables and Lagrange multipliers, (3) is equivalent to a standard quadratic programming (QP) problem which can be solved with QP. After (3) is optimized, (1) can be rewritten as

$$f(x) = \sum_{x_i \in SV} (\alpha_i - \alpha_i')k(x_i, x) + b, \quad (4)$$

where α_i and α_i' are the optimized solution of QP. $k(x, x')$ is the following kernel function

$$k(x, x') = < \phi(x), \phi(x') > . \quad (5)$$

It should be pointed out that not all functions can be taken as kernel function in the SVR. It has been proved that the kernel function should satisfy the conditions of Mercer's theorem. The kernel function plays an important role in the SVR. It has great effect on the predicting precision. In the next section the kernel function is modified by the conformal mapping, which makes the kernel function data-dependent.

III. DATA-DEPENDENT KERNEL

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

In the traditional SVM and SVR, there are no theories concerning how to choose kernel function in a data-dependent way. While some time series, such as financial time series, weather data etc, are inherently noisy, non-stationary and deterministically chaotic. In order to improve the precision of forecasting, it is necessary to redefine the kernel function from the training data. In this section, a background to data-dependent Kernel is provided. The kernel function is modified based on information geometry[9-12], in a data-dependent way.

From the point of geometry, nonlinear mapping defines an embedding of input space S into feature space F as a curved sub-manifold. Generally, F is a reproducing kernel Hilbert space(RKHS) which is a subspace of Hilbert space. So a Riemannian metric[9] can be induced in the input space, and the Riemannian metric can be expressed in the closed form in terms of the kernel

$$g_{ij}(x) = < \frac{\partial}{\partial x_i}\phi(x), \frac{\partial}{\partial x_j}\phi(x) > = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j'}k(x, x')|_{x'=x} . \quad (6)$$

The volume form in the feature space is defined as

$$dV = \sqrt{g(x)}dx_{(1)}dx_{(2)}\cdots dx_{(n)}, \quad (7)$$

where $g(x) = \det|g_{ij}(x)|$, $x_{(1)}, x_{(2)}\cdots, x_{(n)}$ are the whole elements of x, the factor $\sqrt{g(x)}$ represents how a local area is magnified in F under the mapping $\phi(x)$ [9].

After a conformal mapping is introduced to the kernel function, the new kernel function is defined as

$$\tilde{k}(x, x') = c(x)c(x')k(x, x'), \quad (8)$$

where $c(x)$ is a positive conformal mapping function. It is easy to see that the new kernel function $\tilde{k}(x, x')$ satisfies the conditions of Mercer's theorem.

The conformal mapping is taken as[9]

$$c(x) = \sum_{i \in I} h_i \exp(-\| x - x_i \|^2 / (2\tau^2)), \quad (9)$$

where I is the support vectors set[9].

For the new kernel function $\tilde{k}(x, x')$, the Riemannian metric can be expressed as

$$\tilde{g}_{ij}(x) = \frac{\partial c(x)}{\partial x_i}\frac{\partial c(x)}{\partial x_j} + c(x)^2 g_{ij}(x) . \quad (10)$$

One of the typical kernel functions is Gaussian RBF kernel :

$$k(x, x') = \exp(-\| x - x' \|^2 / 2\sigma^2) . \quad (11)$$

In this case, we have

$$g_{ij}(x) = \delta_{ij} / \sigma^2, \quad (12)$$

where $\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \ne j \end{cases}$.

In the region of a neighborhood of a support vector x_i, we have

$$\sqrt{\tilde{g}(x)} = (h_i^n / \sigma^n)\exp(-nr^2 / (2\tau^2))\sqrt{1 + \sigma^2 r^2 / \tau^4}, \quad (13)$$

where $r = \| x - x_i \|$ is the Euclidean distance between x and x_i.

IV. ITERATIVE MODIFIED ALGORITHM

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

In order to improve the performance of SVM classifier[9], the kernel function is modified such that the factor $\sqrt{\tilde{g}(x)}$ is enlarged around the support vectors. But in SVR, the kernel function is modified such that $\sqrt{\tilde{g}(x)}$ should be compressed nearby support vectors.

In order to make sure the magnification is compressed.

Let

$$c(x)=\sum_{i\in I} h_i \exp(-\| x-x_i \|^2/(2\tau^2)), \tag{14}$$

$$\tilde{k}_s(x,x')=c^s(x)c^s(x')k(x,x'), \tag{15}$$

$$\tilde{g}_{sij}(x)=\frac{\partial}{\partial x_i}\frac{\partial}{\partial x'_j}\tilde{k}_s(x,x')|_{x'=x}. \tag{16}$$

where s is a positive integer and I the whole training data. The summation in (14) runs over all the training data such that the training procedure can be simplified.

In this paper the RBF kernel is adopted. For those point x which is nearby the training data x_i, we have

$$\begin{aligned}\sqrt{\tilde{g}_s(x)}&=\sqrt{\det|\tilde{g}_{sij}(x)|}\\&=(h_i^{ns}/\sigma^n)\exp(-nsr^2/(2\tau^2))\sqrt{1+s^2\sigma^2r^2/\tau^4}\end{aligned}. \tag{17}$$

It is easy to check the following conclusions:

1. when $h_i<\min\{\sigma,1\}$, $\sqrt{\tilde{g}_s(x)}$ is compressed nearby the training data x_i for $s=1,2,\cdots$.

2. when $1\le h_i<e$, $\sqrt{\tilde{g}_s(x)}$ is compressed nearby the training data x_i for $s>M$.

3. $\sqrt{\tilde{g}_s(x)}$ is compressed for x which is far away every training data x_i.

In summary, the training process of the new method is as followes.

Step 1. Give random values for $\sigma,\varepsilon,C,h,\tau$ and *error*;

Step 2. Let $k_0(x,x')=k(x,x')$, $j=0$;

Step 3. Training SVR by $k_j(x,x')$, computer the figure of merit $\delta=\sqrt{\sum_i (y_i-f(x_i))^2\Big/\sum_i (y_i-\bar{y})^2}$;

Step 4. If $\delta>error$

Then

$\tilde{k}_{j+1}(x,x')=c(x)c(x')k_j(x,x')$ $j=j+1$, return Step 3

Else

Exit.

After the procedure finished, the modified kernel is

$$\tilde{k}(x,x')=c^s(x)c^s(x')k(x,x'), \tag{18}$$

where s is the final iteration number.

V. SIMULATION

In this section, we do some simulations to evaluate the performance of this method. We have compared the iterative modified SVR and the traditional SVR on the same examples. There have the same training parameters in optimal problem (3). To assess the approximation results and the generalization ability, a figure of merit is defined as

$$\delta=\sqrt{\sum_i (y_i-f(x_i))^2\Big/\sum_i (y_i-\bar{y})^2}, \tag{19}$$

where $\{(x_i,y_i)\,|\,i=1,2,\cdots m\}$ is the test set, $\bar{y}=\sum_i y_i/m$.

A. Approximation of One-dimensional Function

For comparison we show the results with the iterative modified SVR and with the traditional SVR. The solid lines represent the original function and the dashed lines show the approximations.

Fig.1 shows the results for the approximation of the function $f(x)=\sin x+(\sin 3x)/3-2\sin(x/2), x\in[0,2\pi]$.

Approximation results for three different values of h and τ are represented in table I, where the parameters of the traditional SVR and the testing error are $\sigma=0.1$, $\varepsilon=0.1$, $C=0.05$ and $\delta=0.895$ respectively.

The figure of merit for each approximation is computed on a uniformly sampled test set of 126 points. The input x is constructed by the uniform distribution on interval $[0,2\pi]$, The training and test data are composed of 63 points and 126 points respectively.

Fig.2 shows the another approximation for the same function, where the parameters of the traditional SVR and the testing error are $\sigma=0.1$, $\varepsilon=0.1$, $C=0.1$ and $\delta=0.7902$ respectively.

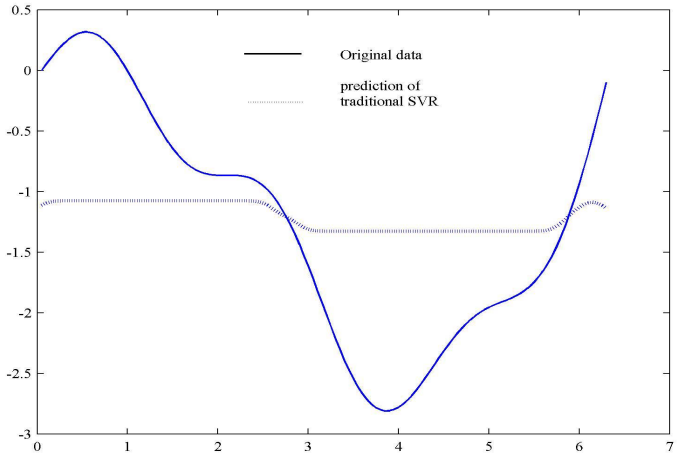
This shows the modified method possesses better performance of generalization than the traditional SVR.

B. Approximation of Multi-dimensional Functions

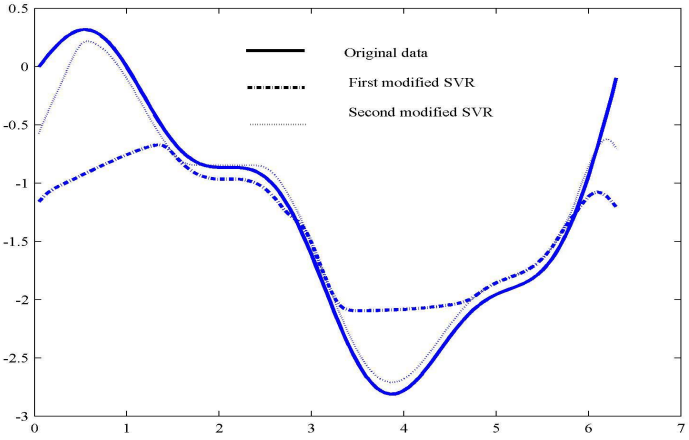
Table II shows the approximation results of the earthquake wave, which is regarded as time series. The training data set is $\{(x_i, y_i) \in R^3 \times R \mid i = 1,2,\cdots,50\}$, a historical lag with order 3. The figure of merit for each approximation is computed on a uniformly sampled test set of 50 points. Fig.3 illustrates part of original data, its approximation by the traditional SVR and modified SVR. From comparison we can see the modified method possesses better performance of generalization than the traditional SVR.

TABLE I. APPROXIMATION RESULTS AND PARAMETERS OF SINGLE VARIA BLE FUNCTION

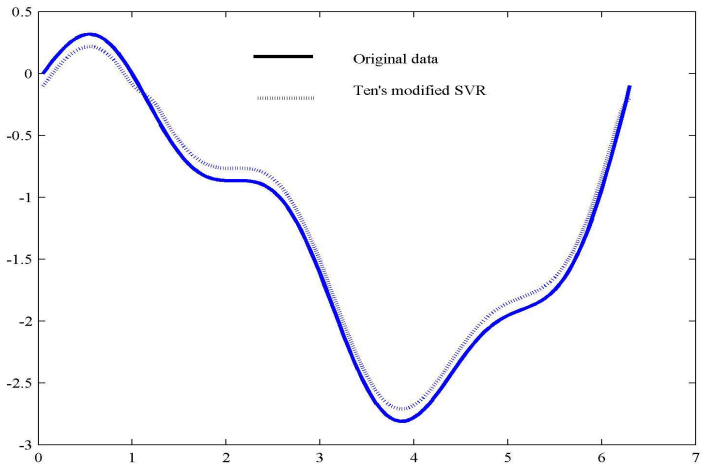
Parameters & testing error of the traditional SVR	**Parameters of the modified SVR**	**testing error of the first modified SVR**	**testing error of the second modified SVR**	**testing error of the 10's modified SVR**
$\sigma = 0.1$ $\varepsilon = 0.1$ $C = 0.05$ $\delta = 0.895$	$h_i = 0.1,\ \tau = 1$	0.5719	0.1598	0.1014
	$h_i = 0.1,\ \tau = 0.8$	0.6821	0.3491	0.1014
	$h_i = 0.08,\ \tau = 0.8$	0.7841	0.6099	0.1439
$\sigma = 0.1,\ \varepsilon = 0.1$ $C = 0.1$ $\delta = 0.7902$	$h_i = 0.8,\ \tau = 1$	0.1044	0.1038	0.1038



a. The approximation result with the traditional SVR.



b. The approximation results with the first and second modified SVR.



c. The approximation result with the ten's modified SVR.

Figure 1. Approximation results of one-dimensional function.

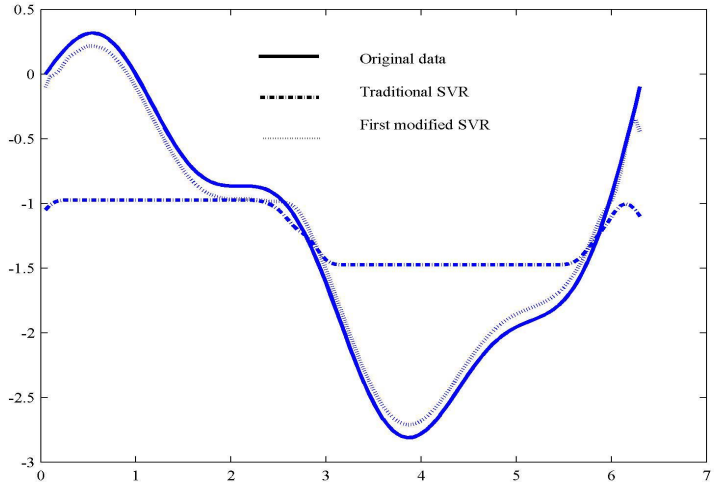


Figure 2. Approximation results with the traditional and first modified SVR.

TABLE II. APPROXIMATION RESULTS AND PARAMETERS OF MULTI-DIMENSIONAL FUNCTION

Parameters of the traditional SVR	**Parameters of the modified SVR**	**testing error of the traditional SVR**	**testing error of the first modified SVR**	**testing error of the second modified SVR**
$\sigma = 0.04$ $\varepsilon = 0.01$ $C = 2$	$h_i = 4$, $\tau = 0.06$	0.9753	0.1931	0.1229

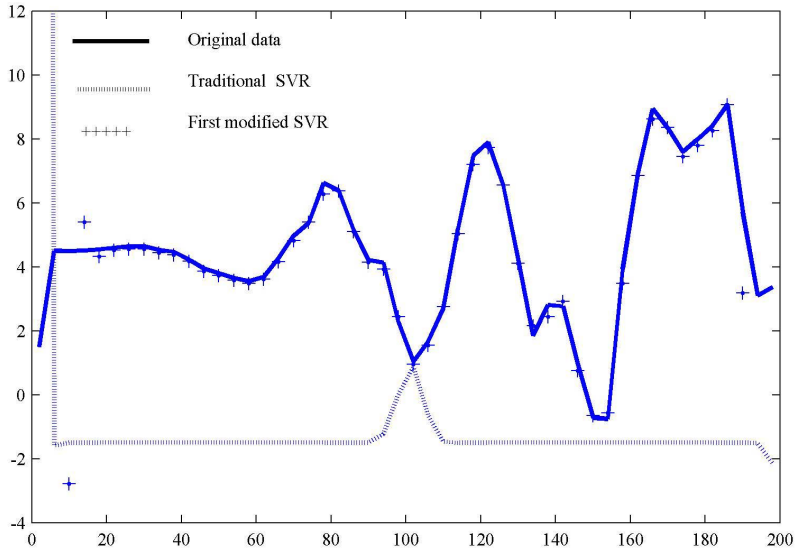


Figure 3. Approximation results of multi-dimensional function.

VI. CONCLUSION

In this paper we present an iterative modified kernel to improve the performance of SVR. It is based on the conformal mapping in information geometry, which results in data-dependent kernel. The whole training data are used in the construction of the conformal mapping instead of support vectors. The idea is to compress the spatial resolution such that the performance of forecasting is improved. It is important that the kernel is modified by iteration and the parameters of the kernel could be set as random values. Simulation results show the effectiveness and generalization ability of this method.

ACKNOWLEDGMENT

This work is supported by NSF Grant #50578168, Natural Science Foundation Project of CQ CSTC 2007BB2396 and KJ070403.

REFERENCES

[1] Vapnik V, Nature of Statistical Learning Theory [M], Translated by Zhang Xue-gong. Beijing: Tsinghua University Press, 2000, pp. 29–136.

[2] Scholkopf B, Sung K, Burges C, "Comparing support vector machines with gaussian kernels to radial basis function classifiers,". IEEE Trans Signal Processing, 1997, vol. 45, pp. 2758–2765.

[3] Perez-Cruz F , Navia-Vazquez A , Figueiras-Vidal A R, "Empirical risk minimization for support vector classifiers," IEEE Trans on Neural Networks , 2003 , vol. 14, pp. 296–303.

[4] Belhumeur P N, Hespanha J P, Kriegman D J, "Eigenfaces vs fisherfaces : recognition using class specific linear projection," IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, vol. 19, pp. 711–720.

[5] Cao Li-juan, Tay Francis E H, "Financial forecasting using support vector machines," Neural Computing &Applications, 2001, vol. 10, pp. 184–192.

[6] Tay Francis E H, Cao Li-juan, "E-descending support vector machines for financial time series forecasting," Neural Processing Letters, 2002, vol. 15, pp.179–195.

[7] Yang Hai-qin, Chan Lai-wan, "King Irw in support vector machine regression for volatile stock market prediction," Proceedings of Intelligent Data Engineering and Automated Learning. Berlin: Springer-Verlag, 2002, pp.319–396.

[8] Smola, A.J., Schölkopf, B. &MÜLler, K. R., "The connection between regularization operators and support vector kernels," Neural Network, 1998, vol. 11, pp.637–649.

[9] Amari S, W u Si, "Improving support vector machine classifiers by modifying kernel functions," Neural Networks, 1999, vol. 12, pp.783–789.

[10] Liang Yan-chun, Sun Yan-feng, "An improved method of support vector machine and its applications to financial time series forcesting," Progresss in Natural Science, 2003, vol. 13, pp.696–700.

[11] Colin Campbell, Kernel Methods: "A survey of current techniques," Neurocomputing , 2002, vol. 48, pp.63–84.

[12] Amari S, W u Si, "An information-geometrical method for improving the performance of support vector machine classifiers," Proccedings of International Conference on Artificial Neural Network s'99. London: IEEE conference Publication No. 470, 1999, pp. 85–90.