An improved method of support vector machine and its applications to financial time series forecasting*

LIANG Yanchun[1]** and SUN Yanfeng[2]

(1. College of Computer Science and Technology, Jilin University, Changchun 130012, China; 2. College of Mathematics, Jilin University, Changchun 130012, China)

Received January 30, 2003; revised April 21, 2003

Abstract A novel method for kernel function of support vector machine is presented based on the information geometry theory. The kernel function is modified using a conformal mapping to make the kernel data-dependent so as to increase the ability of predicting high noise data of the method. Numerical simulations demonstrate the effectiveness of the method. Simulated results on the prediction of the stock price show that the improved approach possesses better forecasting precision and ability of generalization than the conventional models.

Keywords: support vector machine (SVM), data-dependent, information geometry, conformal mapping, financial time series, stock price prediction.

Artificial neural networks have received increasing attention in financial time series forecasting[1,2]. Like most application fields using neural networks, most of the popular neural network models applied to time series forecasting is the feedforward neural network with error back-propagation (BP) algorithm due to its simple architecture yet powerful problem-solving ability. In recent years, a novel neural network algorithm, called support vector machine (SVM), has been introduced by Vapnik et al.[3] to solve machine learning tasks such as regression, pattern recognition and density estimation. The approach is systematic and properly motivated by statistical learning theory. Unlike most traditional neural network models which implement the empirical risk minimization principle, the SVM implements the structural risk minimization principle which seeks to minimize the training error and a confidence interval term. This eventually results in a good performance of generalization. Because of its good properties such as automatic selection on models (parameters and locations of basis functions), trained with quadratic programming (globally optimal solution existed), and good learning ability for small samples, the SVM has received increasing attention in recent years[4~6]. Moreover, the SVM has been successfully applied to the support vector regression (SVR), especially, for modeling nonlinear financial time series, such as prediction on the stock price[7~9].

An improved method of SVM is proposed based on the modification of the kernel function in this paper. A conformal mapping is used to make the kernel data-dependent. Examination on the forecasting precision for the stock prices shows that the proposed method is obviously superior to the traditional SVM in the precision of prediction.

This paper consists of four main sections. Section 1 provides a brief introduction to the SVR and its kernel functions. Section 2 deals with the new data-dependent kernel SVR. The experimental study on stock price prediction is performed in Section 3. The conclusion drawn from this study forms the last section.

1 SVR and its kernel function

Let $\{(\boldsymbol{x}_i, d_i)\}\ (i = 1, \cdots, n)$ be a given set of data points where $\boldsymbol{x}_i$ is the ith input vector and d_i the corresponding desired output. The output of the neural network is

$$y = f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b, \qquad (1)$$

where $\boldsymbol{w}$ is the weight vector, b the bias and $\Phi(\boldsymbol{x})$ the nonlinear mapping from the input space S to the high dimensional feature space F which is the only hidden space in the SVR. $\langle \cdot, \cdot \rangle$ represents the inner

* Supported by the Key Project of National Education Ministry of China (Grant No. 02090)

** To whom correspondence should be addressed. E-mail: ycliang@public.cc.jl.cn

product.

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

$$L_\varepsilon(d_i, y_i) = \begin{cases} |d_i - f(\boldsymbol{x}_i)| - \varepsilon, & |d_i - f(\boldsymbol{x}_i)| \geqslant \varepsilon, \\ 0, & \text{elsewhere.} \end{cases} \tag{2}$$

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

$$R(c, \varepsilon) = \frac{1}{n} c L_\varepsilon(d_i, y_i) + \frac{1}{2} \| w \|^2, \tag{3}$$

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, Eq. (3) is equivalent to a standard quadratic programming (QP) problem and can be solved with QP. When Eq. (3) is optimized, Eq. (1) can be rewritten as[7]

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(\boldsymbol{x}, \boldsymbol{x}_i) + b, \tag{4}$$

where α_i and α_i^* are two kinds of Lagrange multipliers which are denoted using $\alpha_i^{(*)}$, in which if there exists a * in the () it represents α_i^*, otherwise α_i. $k(\boldsymbol{x}, \boldsymbol{x}')$ is a kernel function and satisfies

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle. \tag{5}$$

From Eq. (4), in order to calculate $f(\boldsymbol{x})$, we need only to calculate Lagrange multipliers $\alpha_i^{(*)}$ and kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$ instead of calculating the values of the weight vectors w and the nonlinear mapping $\Phi(\boldsymbol{x})$.

It should be pointed out that not all functions can be taken as kernel functions in the SVR. It has been proved theoretically that the kernel functions satisfying the following conditions can be the kernel functions of the SVR.

Theorem 1. (Mercer's theorem)[10].

Let A be a compact subset of $\boldsymbol{R}^N$, $h(\boldsymbol{x}) \in L_2(A)$, and T a positive integral operator satisfying

$$(Th)(\boldsymbol{x}') = \int_A k(\boldsymbol{x}, \boldsymbol{x}') h(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{6}$$

If $k(\boldsymbol{x}, \boldsymbol{x}')$ is a continuous symmetric kernel of T, which satisfies the following equation

$$\iint_{A \times A} k(\boldsymbol{x}, \boldsymbol{x}') h(\boldsymbol{x}) h(\boldsymbol{x}') \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{x}' \geqslant 0, \tag{7}$$

then $k(\boldsymbol{x}, \boldsymbol{x}')$ can be expanded in a uniformly convergent series

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{n_c} \lambda_j \psi_j(\boldsymbol{x}) \psi_j(\boldsymbol{x}'), \tag{8}$$

where λ_j, ψ_j and n_c are the positive eigenvalues, eigenfunctions and the number of the positive eigenvalues, respectively.

The kernel function plays an important role in the SVR. The selection of the kernel function has great effect on the precision of the prediction. In the next section the kernel function is modified by using the method of conformal mapping, which makes the kernel function data-dependent and is desired to improve the effect of the prediction to a given problem.

2 Data-dependent kernel

In the traditional SVR, there are no theories concerning how to choose good kernel functions in a data-dependent way[11,12], while some time series, such as financial time series 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 using the given data. In this section, the kernel is modified, based on the method of information geometry[13], in a data-dependent way in order to improve the prediction ability for highly noisy data.

From the point of geometry, nonlinear mapping $\Phi(\boldsymbol{x})$ defines an embedding of input space S into feature space F as a curved submanifold. Generally, in the SVR, F is a reproducing kernel Hilbert space (RKHS) which is a subspace of Hilbert space. So a Riemannian metric $G_{ij}(\boldsymbol{x})$ can be induced in the input space S, and the Riemannian metric can be expressed in the closed form in terms of the kernel

$$G_{ij}(\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}_i} \frac{\partial}{\partial \boldsymbol{x}_j'} k(\boldsymbol{x}, \boldsymbol{x}') \big|_{\boldsymbol{x}' = \boldsymbol{x}}. \tag{9}$$

In this paper, Gaussian RBF kernel is used

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp(- \| \boldsymbol{x} - \boldsymbol{x}' \|^2 / 2\sigma^2), \tag{10}$$

where σ is a given parameter.

In this case, the Riemannian metric is

$$G_{ij}(\boldsymbol{x}) = \delta_{ij} / \sigma^2, \tag{11}$$

where

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

A conformal mapping is introduced to the kernel functions and the new kernel function is taken as

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') = D(\boldsymbol{x})D(\boldsymbol{x}')k(\boldsymbol{x}, \boldsymbol{x}'). \quad (12)$$

From Eq. (9), the Riemannian metric can be rewritten as

$$\tilde{G}_{ij}(\boldsymbol{x}) = \frac{\partial D(\boldsymbol{x})}{\partial \boldsymbol{x}_i} \cdot \frac{\partial D(\boldsymbol{x})}{\partial \boldsymbol{x}_j} + [D(\boldsymbol{x})]^2 G_{ij}(\boldsymbol{x}). \quad (13)$$

Based on the above discussion, a method to improve the precision of the forecasting is proposed under the condition of enlarging the area around the key data points whereas keeping angles unchanged in the whole space and therefore won't affect much the spatial relationship between the data points. The problem is how to select the key data points and the number of the key data points. The optimal partition algorithm (OPA) is proved to be a novel and effective method in dealing with stock price time series[14]. Therefore, the OPA is used here and the conformal mapping is taken as

$$D(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\exp(-\| \boldsymbol{x} - r_i \|^2/\tau_i^2), \quad (14)$$

where m, r_i and τ_i are the number of the partitioning points, the center and the width of the ith partition, respectively.

A proof of the new kernel function defined by using Eqs. (12) and (14) satisfying Mercer's condition will be performed in the following theorem.

Theorem 2. The kernel function $\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')$ satisfies Mercer's condition.

Proof. From Eq. (12), the continuity and symmetry of the kernel $\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')$ is obvious. Next, let us prove the positive semi-definite property of the kernel $\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')$. The symbols in Theorem 1 are still used here. Since $D(\boldsymbol{x}) > 0$, and there exists a positive β such that $D(\boldsymbol{x}) \geqslant \beta > 0$, we have

$$\begin{aligned}\int_A\int_A \tilde{k}(\boldsymbol{x}, \boldsymbol{x}')h(\boldsymbol{x})h(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' &= \int_A\int_A D(\boldsymbol{x})D(\boldsymbol{x}')k(\boldsymbol{x}, \boldsymbol{x}')h(\boldsymbol{x})h(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \\ &\geqslant \beta^2\int_A\int_A k(\boldsymbol{x}, \boldsymbol{x}')h(\boldsymbol{x})h(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \geqslant 0. \quad (15)\end{aligned}$$

This completes the proof.

In summary, the training process of the improved method consists of the following three steps: (i) Determine the number of the partitioning points, center and width of the partitions using the OPA; (ii) calculate the basic kernel function using Eq. (10) and modify it using Eqs. (12) and (14); (iii) train the SVR using the modified kernel function $\tilde{k}(\boldsymbol{x}, \boldsymbol{x}')$.

3 Applications to stock price prediction

In order to examine the effectiveness of the proposed algorithm based on the modified kernel, we apply it to the stock price forecasting. The data used are the indexes in the stock exchanges, including the Standard and Poor's 500 (S&P 500) index and the composite index in Shanghai Stock Exchange (CISSE). The data of S&P 500 and CISSE used in this paper are those with 353 continuous trading days from 01/02/2001 and 01/04/1999, respectively. The normalization to the data is performed first. Denote the stock price series as $x_i (i = 1, 2, \cdots, 353)$. Let $x_{\max} = \max\limits_{1 \leqslant i \leqslant 353}(x_i)$. Denote the normalized data as $x_i' (i = 1, 2, \cdots, 353)$ where $x_i' = x_i / x_{\max}$.

To the above two groups of data, the numbers of the input nodes are both taken as 3 in this paper, that is, a historical lag with order 3 is considered in the simulation. The number of the output nodes is 1, that is, a single step prediction is taken. The original S&P 500 and CISSE data are first formed into 350 input-output data pairs. Then the data are divided into two parts, respectively, to form the training data set (175) and the test data set (175). The error function is defined as

$$E = \frac{1}{n}\sum_{i=1}^{n}(d_i - y_i)^2. \quad (16)$$

There are four parameters needed to be determined in the improved SVR method. Table 1 shows the values of the parameters. Table 2, Figs. 1 and 2 show the comparison between the actual and the simulation values of the prediction on the indexes when the improved SVR and traditional SVR (simplified as SVR in the table and figures) are applied to the S&P 500 and CISSE data, respectively. From Table 1 it can be seen that there is an obvious difference between the two methods for the training and test precision. The training errors in S&P 500 data and CISSE data of the modified method are around 1 and 3 times less than those of the SVR, respectively. While the test errors in S&P 500 data and CISSE data of the modified method are around 3 and 6 times less than those of the SVR, respectively. This shows the modified method possesses better performance of generalization than the SVR. From Figs. 1 and 2 it can also

be seen that the forecasting effect of test data is superior to that of training data both in S&P500 data and CISSE data, showing again that the improved method can make SVR have a better ability of generalization.

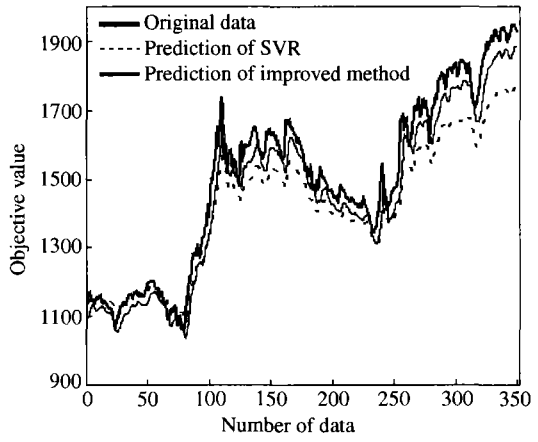

Fig. 1. Fitting results using SVR and improved method for CISSE.

Table 1. Parameter values and numerical results

Parameters and classes of results	SVR		Improved method	
	S&P 500	CISSE	S&P 500	CISSE
C	25.0	25.0	25.0	25.0
σ	8.0	23.5	8.0	23.5
ε	0.0004	0.04	0.0004	0.04
m	Null	Null	10	10
Number of SV	97	77	95	78
Training error	0.000087	0.0014	0.000076	0.0005
Testing error	0.000671	0.0038	0.000209	0.0006

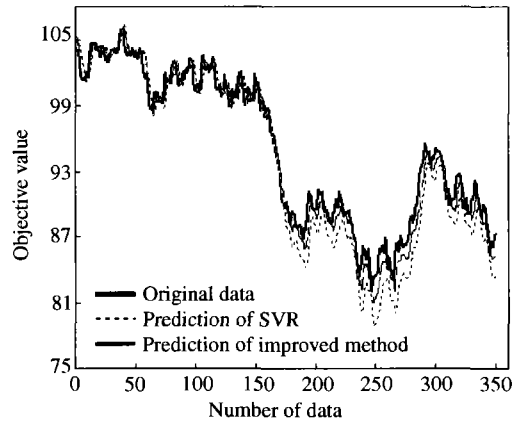

Fig. 2. Fitting results using SVR and improved method for S&P 500.

In order to further examine the improved method, we employ four statistical metrics to evaluate the prediction performance. The four metrics include weighted directional symmetry (WDS), modified directional symmetry (MDS), normalized mean squared error (NMSE) and mean absolute error (MAE)[15]. The definitions of these criteria are stated below. The first two metrics are directional metrics which need to be maximized, while the last two are precision metrics which need to be minimized. WDS measures both the magnitude of the prediction error and the direction. It penalizes the errors related to the incorrectly predicted direction and rewards those associated with the correctly predicted direction. The larger the value of WDS is, the better the forecasting performance is in terms of both magnitude and direction. MDS takes into consideration all the correctly predicted directions (upward, downward, and no change), as well as computer truncation errors. NMSE and MAE are the measures of the deviation between the actual and predicted values. The smaller the values of NMSE and MAE are, the closer the predicted time series values are to the actual values.

The definition of the WDS metric is as follows:

$$\mathrm{WDS} = \frac{\sum_{i=1}^{n} a_i \mid y_i - d_i \mid}{\sum_{i=1}^{n} a_i' \mid y_i - d_i \mid}, \tag{17}$$

where

$$a_i = \begin{cases} 1, & \text{if } (y_i - y_{i-1})(d_i - d_{i-1}) \geqslant 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$a_i' = \begin{cases} 0, & \text{if } (y_i - y_{i-1})(d_i - d_{i-1}) \geqslant 0, \\ 1, & \text{otherwise.} \end{cases}$$

The MDS is defined as

$$\mathrm{MDS} = \frac{100}{n} \sum_{i=1}^{n} a_i, \tag{18}$$

where

$$a_i = \begin{cases} 1, & \text{if } (y_i - y_{i-1})(d_i - d_{i-1}) > 0 \\ & \quad \text{and } (\mid y_i - y_{i-1} \mid > \mu \\ & \quad \text{and } \mid d_i - d_{i-1} \mid > \mu) \\ & \quad \text{or } (\mid y_i - y_{i-1} \mid < \mu \\ & \quad \text{and } \mid d_i - d_{i-1} \mid < \mu)), \\ 0, & \text{otherwise.} \end{cases}$$

in which μ is a small constant related to the numerical precision in the computation.

The NMPE is defined as

$$\mathrm{NMSE} = \frac{1}{n\delta^2} \sum_{i=1}^{n} (y_i - d_i)^2, \tag{19}$$

where $\delta = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (d_i - \bar{d})^2}$, $\bar{d}$ is the means of $\{d_i\}_{i=1}^{n}$.

The MAE is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - d_i | . \qquad (20)$$

Table 2 shows the numerical simulation results for the S&P 500. From the table it can be seen that the improved method has smaller NMSE and MAE but larger WDS and MDS than SVR. The obvious improvements of the MAE and NMSE in the proposed method for the test data show that the improved method can make SVR have a better ability of generalization.

Table 2. Comparison of statistical metrics for training and test data

	WDS		MDS		MAE		NMSE	
	Training	Test	Training	Test	Training	Test	Training	Test
SVR	1.64	1.12	25.287	32.571	0.008	0.023	0.116	0.795
Improved method	1.91	1.39	25.291	33.714	0.007	0.012	0.101	0.248

4 Conclusions and discussions

A novel SVR method is presented. It is based on the modification of the kernel function by using the conformal mapping in information geometry, which makes the kernel function data-dependent. When dealing with regression problems and when data being highly noisy, the use of support vectors is unfit. In this paper, the key points in the data are used in the construction of the conformal mapping instead of support vectors, which enables the proposed method to fit the regression problems and the data with high noise. The effectiveness and generalization ability of the proposed method are demonstrated using the stock price forecasting. However, there are still some complicated situations to be considered in the stock prediction, such as multi-step forecasting where the predicted values are iteratively used as the inputs for the next forecasting.

References

1 Wu, W. et al. Prediction of ups and downs of stock market by BP neural networks. Journal of Dalian University of Technology (in Chinese), 2001, 41(1): 9.

2 Wu, W. et al. On improvement of BP algorithms for prediction of stock market. Journal of Dalian University of Technology (in Chinese), 2001, 41(5): 518.

3 Vapnik, V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

4 Flake, G. W. et al. Efficient SVM regression training with SMO. Machine Learning, 2002, 46: 271.

5 Zhang, X. G. Introduction to statistical learning theory and support vector machines. Acta Automatica Sinica (in Chinese), 2000, 26(1): 37.

6 Liu, J. H. et al. Support vector machine training algorithm: a review. Information and Control (in Chinese), 2002, 31(1): 45.

7 Cao, L. J. et al. Financial forecasting using support vector machines. Neural Computing & Applications, 2001, 10 (2): 184.

8 Francis, E. H. et al. ε-Descending support vector machines for financial time series forecasting. Neural Processing Letters, 2002, 15 (2): 179.

9 Yang, H. Q. et al. Support vector machine regression for volatile stock market prediction. In: Proceedings of Intelligent Data Engineering and Automated Learning. Berlin: Springer-Verlag, 2002: 391.

10 Colin, C. Kernel methods: A survey of current techniques. Neurocomputing, 2002, 48: 63.

11 Amari, S. et al. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 1999, 12: 783.

12 Wu, S. et al. Conformal transformation of kernel functions: a data-dependent way to improve support vector machine classifiers. Neural Processing Letters, 2002, 15: 59.

13 Amari, S. et al. An information-geometrical method for improving the performance of support vector machine classifiers. In: Proceedings of International Conference on Artificial Neural Networks'99. London: IEEE Conference Publication, 1999, 85.

14 Sun, Y. F. et al. Optimal partition algorithm of RBF neural network and its applications to stock price prediction. In: Proceedings of International Conference on Intelligent Information Technology. Beijing: Posts & Telecom Press, 2002, 448.

15 Sara, M. et al. Performance metrics for financial time series forecasting. Journal of Computational Intelligence in Finance, 1999, 8: 5.