TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 06/10 pp706–715
DOI: 10.26599/TST.2018.9010092
Volume 24, Number 6, December 2019

Residuals-Based Deep Least Square Support Vector Machine with Redundancy Test Based Model Selection to Predict Time Series

Yanhua Yu* and Jie Li

Abstract: In this paper, we propose a novel Residuals-Based Deep Least Squares Support Vector Machine (RBD-LSSVM). In the RBD-LSSVM, multiple LSSVMs are sequentially connected. The second LSSVM uses the fitting residuals of the first LSSVM as input time series, and the third LSSVM trains the residuals of the second, and so on. The original time series is the input of the first LSSVM. Additionally, to obtain the best hyper-parameters for the RBD-LSSVM, we propose a model validation method based on redundancy test using Omni-Directional Correlation Function (ODCF). This method is based on the fact when a model is appropriate for a given time series, there should be no information or correlation in the residuals. We propose the use of ODCF as a statistic to detect nonlinear correlation between two random variables. Thus, we can select hyper-parameters without encountering overfitting, which cannot be avoided by only cross validation using the validation set. We conducted experiments on two time series: annual sunspot number series and monthly Total Column Ozone (TCO) series in New Delhi. Analysis of the prediction results and comparisons with recent and past studies demonstrate the promising performance of the proposed RBD-LSSVM approach with redundancy test based model selection method for modeling and predicting nonlinear time series.

Key words: time series prediction; information redundancy; residuals-based deep Least Squares Support Vector Machine (LSSVM); Omni-Directional Correlation Function (ODCF)

1 Introduction

As one of data mining technologies, time series prediction has been widely used in many applications, such as financial market prediction, electrical utility load forecasting, weather and environmental state prediction, and communication network traffic volume prediction. Models such as the well-known auto-regressive moving average, Auto-Regressive Integrated Moving Average (ARIMA), and seasonal ARIMA can only be applied for linear time series[1, 2]. However, many time series in the real world are typically nonlinear. In this case, more advanced time series prediction algorithms, such as neural network[3–5] and Support Vector Machine (SVM)[6–11], need to be applied.

Recently, some researches have ensembled several models to enhance the accuracy of time series prediction[5, 7]. In Ref. [5], Han and Xu ensembled two models to predict the time series. Finally, they added the predicted values of the two models to obtain the final predicted value of the ensemble model. Tang et al. proposed another ensemble model of multiple Least Squares Support Vector Machines (LSSVMs)[7]. The algorithm creates multiple LSSVMs predictive models via a method of iterative error correction, and the predicted values of each LSSVM are added. They conducted experiments on two benchmark chaos series with known Lorenz equation and Mackey-Glass

• Yanhua Yu and Jie Li are with the School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: yuyanhua@bupt.edu.cn; jli@bupt.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2018-01-10; revised: 2018-02-22; accepted: 2018-02-26

equation and obtained good performance. However, this ensemble model is not appropriate for some other time series, such as annual sunspot series. When the original sunspot series is modeled by the first LSSVM model, the residuals cannot be modeled by another LSSVM, and the Mean Squared Error (MSE) of the first residuals on the test set is still large. This is because although there are multiple models in the ensemble model, each constituent model is trained separately by minimizing the MSE of the validation set. However, the best model in every step may not be the best for the ensemble model as a whole.

To solve this problem, we propose a novel model of deep SVM, called Residuals-Based Deep Least Squares Support Vector Machine (RBD-LSSVM). The RBD-LSSVM is composed of multiple LSSVMs, and the LSSVMs are sequentially connected. Like the models in the ensemble model in Ref. [7], the first LSSVM is employed to model the original time series, while the second LSSVM is used to model the residual of the first LSSVM, and so on. In the deep SVM, we train the combined model as a whole; thus, the RBD-LSSVM can avoid the problem of underfitting, which occurs in Ref. [7].

Furthermore, we propose an approach to select the optimal RBD-LSSVM model; this approach involves using the Omni-Directional Correlation Function (ODCF) to test for information redundancy. The approach is based on the fact that inherent correlations exist between the observations in a time series, which is totally different from general regression problems where the samples are assumed to be independent of each other. If a model is appropriate for the time series, it should extract all information from the series. Thus, there should be no redundant information in the residuals. To check for information redundancy, we propose the use of the ODCF statistic, which was proposed by Zhu and Zhang[12] and Zhang et al.[13] to check validity of identified neural network in modeling of nonlinear systems. The ODCF can be categorized into Omni-Directional Auto-Correlation Function (ODACF) and Omni-Directional Cross Correlation Function (ODCCF). The ODACF can be used to check if there is correlation (linear or nonlinear) in the residuals, while the ODCCF can be used to check if there is correlation between the residuals and original time series. In Ref. [10], we proposed using the ODCF to select hyper-parameters for a single SVM. When the RBD-LSSVM is used to model a time series, simply using cross validation with the validation set may not prevent overfitting. However, the information redundancy test based approach can prevent overfitting.

This paper is organized as follows. Section 2 introduces the RBD-LSSVM and information redundancy test based model selection method and ODCF statistic. In Section 3, the shortcomings of the current ensemble models are analyzed, and the RBD-LSSVM is described. The optimal model selection procedure and methods based on the ODCF are also presented. Section 4 describes experiments conducted on two benchmark time series: annual sunspot number and monthly Total Column Ozone (TCO) in New Delhi. Finally, the conclusions are presented in Section 5.

2 Materials and Methods

2.1 Introduction to LSSVM

SVM is a machine learning approach based on structural risk minimization and Vapnik-Chervonenkis dimension, proposed by Vapnik[14] and Deng and Tian[15]. The LSSVM is a kind of SVM developed by Suykens et al.[16] The LSSVM can be employed for both classification and regression.

The goal of regression is to estimate a function $f(x)$ based on a finite number set of noisy samples (x_i, y_i), $i = 1, 2, \ldots, n$, with m-dimensional input $x_i \in \mathbf{R}^m$ and output $y_i \in \mathbf{R}$. The function $f(x)$ that best characterizes the data-generating process has the following form:

$$y_i = f(x_i) + e_i \tag{1}$$

Here, $f(x)$ is the target function, and e_i represents additive zero mean noise with variance σ^2. Equation (2) generally defines the function $f(x)$ for linear and nonlinear regression applications:

$$f(x) = (w \cdot \phi(x)) + b \tag{2}$$

where w is the coefficient vector, and $b \in \mathbf{R}$ is the bias. This equation means $f(x)$ can be viewed as linear in the feature space after input x in the input space is mapped to the higher-dimensional feature space via function $\phi(\cdot)$. To obtain the optimal function $f(x)$, SVMs not only try to minimize empirical loss on training set, but also try to reduce model complexity by minimizing $||w||^2$. Based on this principle, SVM regression is formulated as follows:

$$\min_{w\in\mathbf{R}^n,\, b\in\mathbf{R},\, \xi\in\mathbf{R}^n} R(w,b,\xi)=\frac{1}{2}||w||^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2,$$

$$\text{s.t.}\quad w\cdot\phi(x_i)+b=y_i-\xi_i,\quad i=1,2,\ldots,n \tag{3}$$

Here, C (>0) is the penalty parameter determining the tradeoff between the empirical loss and complexity. The Lagrangian function for this problem is

$$L(w,b,\xi,\alpha)=\frac{1}{2}||w||^2+\frac{C}{2}\sum_{i=1}^{n}\xi_i^2+\sum_{i=1}^{n}\alpha_i(y_i-\xi_i-(w\cdot\phi(x_i))-b) \tag{4}$$

where $\alpha=[\alpha_1,\alpha_2,\ldots,\alpha_n]^{\mathrm{T}}$ is the Lagrange multiplier vector. We can obtain conditions for optimality by partial-differentiating (4) with w, b, ξ_i, and α_i as follows:

$$\begin{aligned}
&\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i\phi(x_i),\\
&\frac{\partial L}{\partial b}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i=0,\\
&\frac{\partial L}{\partial \xi_i}=0\Rightarrow C\xi_i=\alpha_i,\\
&\frac{\partial L}{\partial \alpha_i}=0\Rightarrow y_i-\xi_i-(w\cdot\phi(x_i))-b=0
\end{aligned} \tag{5}$$

The solution of Eq. (5) is as follows:

$$\begin{bmatrix}0 & e^{\mathrm{T}}\\ e & Q+C^{-1}I\end{bmatrix}\times\begin{pmatrix}b\\ a\end{pmatrix}=\begin{pmatrix}0\\ y\end{pmatrix} \tag{6}$$

where $e=[1,1,\ldots,1]^{\mathrm{T}}$, Q is the matrix with element $K_{ij}=\phi(x_i)\phi(x_j)$, $i=1,2,\ldots,n$, I is the unit matrix, and $y=[y_1,y_2,\ldots,y_n]$. After obtaining α and b from the above equation, we can obtain the target function $f(x)$:

$$f(x)=\sum_{i=1}^{n}\alpha_i(\phi(x_i)\cdot\phi(x))+b \tag{7}$$

However, $\phi(x)$ is not easily determined; therefore, kernel function $K(x_i,y_i)=\phi(x_i)\phi(x_j)$ is introduced and $f(x)$ is expressed as Eq. (8),

$$f(x)=\sum_{i=1}^{n}\alpha_i K(x_i,x)+b \tag{8}$$

Typical kernel functions include linear kernel, polynomial kernel, and Gaussian radial basis function. In this paper, we use the Gaussian radial basis function:

$$K(x_i,x_j)=\exp\left(-\frac{(x_i-x_j)^2}{2\sigma^2}\right) \tag{9}$$

2.2 Information redundancy test based on ODCF

Thus far, most of the methods check the performance of the LSSVM using only the MSE on the validation set. For example, the cross validation approach, which is applied in most applications, only uses the MSE on the validation set; this method may not prevent overfitting when applied in the RBD-LSSVM. In this paper, we propose using the ODCF statistic to select the best model based on information redundancy test. In Ref. [10], we proposed this method to select the best model for a single SVM. Next, we briefly describe the information redundancy test principle and ODCF statistic used in the RBD-LSSVM.

There is a significant difference between time series prediction and other regression problems: The observations in a time series have inherent correlations with each other, while in other regression problem, the samples are assumed to be independent of each other. Based on this characteristic, for time series prediction, an Auto-Regression (AR) model can be constructed to predict the succeeding data using the previous data. Therefore, a model is valid if the residuals are reduced to contain no redundant information. Information redundancy means autocorrelation in the residual or correlation between residual and delayed outputs. This concept is similar to the **nonlinear dynamical system identification**, where model validation is an important procedure[12, 13]. Model validation is based on the principle that if the model structure is correct, the residuals should form an independent random sequence and should be unpredictable from all past inputs, outputs, and residuals. Since time series is a kind of dynamical system without input, we use this principle in time series prediction. Furthermore, for nonlinear system identification, Zhu and Zhang[12], Zhang et al.[13], Ljung[17], and Mao and Billings[18] have developed validation procedures to check the quality of identified neural network based on a correlation test. In the present study, we use the ODCF developed by them, which is simple and convenient, compared to other statistics such as the Brock-Dechert-Scheinkman statistic[19], which can only be used to check for nonlinear autocorrelation. Next, we briefly describe how to use the ODCF to test for information redundancy. As a dynamical system without input information, a time series model can be expressed by

the following equation:

$$y(t) = f(y^{t-1}) + e(t) = f(y(t-1), \ldots, y(t-t_y)) + e(t) \quad (10)$$

where $y(t)$ denotes the outputs of the dynamical system, $f(\cdot)$ is a nonlinear function, and $e(t)$ denotes an additive noise such as measurement error with zero mean and certain variance. A model $\hat{f}(\cdot)$ can be obtained by training, which is used to approximate the relationship between the data sequences observed from system. The one-step-ahead predicted outputs and residuals can be expressed as

$$\hat{y}(t) = \hat{f}(y^{t-1}) \quad (11)$$

$$\varepsilon(t) = y(t) - \hat{y}(t) \quad (12)$$

The model is valid if residuals are reduced to a random noise expressed as $\varepsilon(t) \to e(t)$. However, the model is invalid if the residuals contain predictable redundant information, which can be expressed as follows:

$$\begin{aligned}\varepsilon(t) &= \hat{g}(y^{t-1}, e^{t-1}) + e(t) = \\ &\hat{g}(\ldots, y(t-p), \ldots) + e(t) = \\ &\hat{g}(\ldots, f(y^{(t-p-1)}) + \varepsilon(t-p), \ldots) + e(t)\end{aligned} \quad (13)$$

From this equation, we can see that if residuals contain redundant information, there will be a correlation between the residuals and delay as well as between the residuals and the original time series. Therefore, these two correlations should be tested. To test these correlations, which may be nonlinear, we use the ODCF, which is easy to compute and is appropriate to test both linear and nonlinear correlations[12, 13]. The ODCF can be categorized into ODCCF and ODACF. The latter can be viewed as a special case of the former when the two series are identical. Next, the computation equations for ODCCF are listed. Assume $a(t)$ and $b(t)$ are two time series. Let

$$\begin{aligned}\alpha(t) &= |a'(t)| = |a(t) - \frac{1}{n}\sum_{t=1}^{n} a(t)|, \\ \beta(t) &= |b'(t)| = |b(t) - \frac{1}{n}\sum_{t=1}^{n} b(t)|, \\ \alpha'(t) &= \alpha(t) - \frac{1}{n}\sum_{t=1}^{n} \alpha(t), \\ \beta'(t) &= \beta(t) - \frac{1}{n}\sum_{t=1}^{N} \beta(t)\end{aligned} \quad (14)$$

Then four correlation coefficient functions can be calculated as follows:

$$r_{\alpha\beta}(\tau) = \frac{\sum_{t=\tau+1}^{n} (\alpha'(t-\tau)\beta'(t))}{\sqrt{\sum_{t=1}^{n}(\alpha'(t))^2 \sum_{t=1}^{n}(\beta'(t))^2}} \quad (15)$$

$$r_{\alpha b}(\tau) = \frac{\sum_{t=\tau+1}^{n} (\alpha'(t-\tau)b'(t))}{\sqrt{\sum_{t=1}^{n}(\alpha'(t))^2 \sum_{t=1}^{n}(b'(t))^2}} \quad (16)$$

$$r_{ab}(\tau) = \frac{\sum_{t=\tau+1}^{n} (a'(t-\tau)b'(t))}{\sqrt{\sum_{t=1}^{n}(a'(t))^2 \sum_{t=1}^{n}(b'(t))^2}} \quad (17)$$

$$r_{a\beta}(\tau) = \frac{\sum_{t=\tau+1}^{n} (a'(t-\tau)\beta'(t))}{\sqrt{\sum_{t=1}^{n}(a'(t))^2 \sum_{t=1}^{n}(\beta'(t))^2}} \quad (18)$$

where τ denotes time delay. By combining the above four coefficients, the ODCCF, denoted by $\rho_{ab}(\tau)$, can be acquired as follows. If

$$|\max(r_{\alpha\beta}(\tau), r_{\alpha b}(\tau), r_{ab}(\tau), r_{a\beta}(\tau))| > |\min(r_{\alpha\beta}(\tau), r_{\alpha b}(\tau), r_{ab}(\tau), r_{a\beta}(\tau))|,$$

then

$$\rho_{ab}(\tau) = \max(r_{\alpha\beta}(\tau), r_{\alpha b}(\tau), r_{ab}(\tau), r_{a\beta}(\tau)) \quad (19)$$

else

$$\rho_{ab}(\tau) = \min(r_{\alpha\beta}(\tau), r_{\alpha b}(\tau), r_{ab}(\tau), r_{a\beta}(\tau)) \quad (20)$$

When $a(t) = b(t), \rho_{ab}(\tau)$ is referred to as ODACF. According to the central limit theorem, when the model is valid, the following two equations hold.

$$\begin{cases}\rho_{\varepsilon\varepsilon}(\tau) = 1, & \tau = 0, \\ \rho_{\varepsilon\varepsilon}(\tau) \sim N(0, \frac{1}{n}), & \text{otherwise}\end{cases} \quad (21)$$

$$\begin{cases}\rho_{y\varepsilon}(\tau) \neq 0, & \tau = 0, \\ \rho_{y\varepsilon}(\tau) \sim N(0, \frac{1}{n}), & \text{otherwise}\end{cases} \quad (22)$$

where $N(0, 1/n)$ denotes a normal distribution with zero mean and variance of $1/n$. Therefore, a statistical hypothesis test is employed to check the following three conditions:

(1) if $\rho_{\varepsilon\varepsilon}(\tau)$ follows $N(0, 1/n)$, where $\tau \neq 0$;

(2) if $\rho_{y\varepsilon}(\tau)$ follows $N(0, 1/n)$, where $\tau \neq 0$;

(3) if $\rho_{y\varepsilon}(\tau) \neq 0$, where $\tau \neq 0$.

If one or more than one of the above three conditions are not satisfied, there is information redundancy in the residuals, which indicates the model is not valid.

3 RBD-LSSVM and Model Selection Based on ODCF

3.1 Phase space reconstruction for original time series

Given a set of time series data $\{x(t_1), x(t_2), \ldots, x(t_i), \ldots, x(t_n)\}$, to obtain the underlying model for the given time series, a transformation is performed as follows:

$$X = [x_1, x_2, \ldots, x_l]^{\mathrm{T}} = \begin{bmatrix} x(t_1) & x(t_{1+\tau}) & \cdots & x(t_{1+(m-1)\tau}) \\ x(t_2) & x(t_{2+\tau}) & \cdots & x(t_{2+(m-1)\tau}) \\ \vdots & \vdots & \ddots & \vdots \\ x(t_l) & x(t_{l+\tau}) & \cdots & x(t_{l+(m-1)\tau}) \end{bmatrix} \tag{23}$$

$$Y = \begin{bmatrix} x(t_{1+m\tau}) \\ x(t_{2+m\tau}) \\ \vdots \\ x(t_{l+m\tau}) \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix} \tag{24}$$

Here, the parameter m is embedding dimension, and τ is delay. l is the number of samples in training set, and $l = n - m\tau$. Our task is to find the function $f(\cdot)$ between input X and output Y. In the process, the delay τ and embedding dimension m are parameters to be determined.

3.2 RBD-LSSVM

Ensemble learning is an effective method to enhance prediction accuracy[20]. Recently, several scholars have proposed some ensemble methods to predict classical nonlinear time series[5, 7]. In Ref. [5], Han and Xu ensembled ARIMA and Echo State Network (ESN) to predict time series. First, they used ARIMA to capture the linear features, and then built a regularized ESN to capture the dynamic nonlinear features. Finally, the predicted values of the two models were added to obtain the final predicted value of the ensemble model. In Ref. [7], Tang et al. proposed another ensemble model composed of multiple LSSVMs. In the algorithm, multiple LSSVM predictive models were created via a method of iterative error correction, and prediction performance was significantly improved. They first used one LSSVM to model the original time series and obtained the prediction residuals of the training set. Then, the second LSSVM was employed to model the residuals of the first LSSVM, and new residuals were computed. Then, the residuals of the second LSSVM would train the third LSSVM, and so on. Finally, the predicted values of each LSSVM in the ensemble model were added to obtain the final predicted value. They conducted experiments on two benchmark chaos series with Lorenz and Mackey-Glass deterministic differential equations and realized good performance. However, this ensemble model is not appropriate for some other time series, such as the annual sunspot time series. When the original sunspot time series is modeled by the first LSSVM model, the residuals cannot be modeled by another LSSVM, and there is still a big error on validation set. This is because although there are multiple models in the ensemble model, each constituent model is separately trained by minimizing the MSE of the validation set. However, the best model in one step may not be the best one for the whole model.

To solve this problem, we propose a novel model, the RBD-LSSVM, which is composed of multiple LSSVMs sequentially connected. Like the models in the ensemble model in Ref. [7], in the proposed model, the first LSSVM is employed to model the original time series, the second LSSVM is used to model the residuals of the first LSSVM, and so on. However, different from the model in Ref. [7], in the RBD-LSSVM, we train the multiple LSSVMs as a whole. The structure of the RBD-LSSVM is illustrated in Fig. 1.

In Fig. 1, $x^2(1)$ denotes the value of the first dimension of input object x in the second layer; x_1^2 denotes the input vector of the first sample in the second layer; l_1 denotes the number of samples in the training set of the first layer; m_1 denotes the number of dimensions of the input in the first layer; r_t^2 denotes the residual at time point t in the second layer; and $\hat{y}_t^2$ denotes the forecasted value of target at time point t in the second layer.

In this network, the first LSSVM models the original time series and obtains the fitting residuals on the training set. The second LSSVM models the residuals of the first LSSVM and obtains the second residuals.

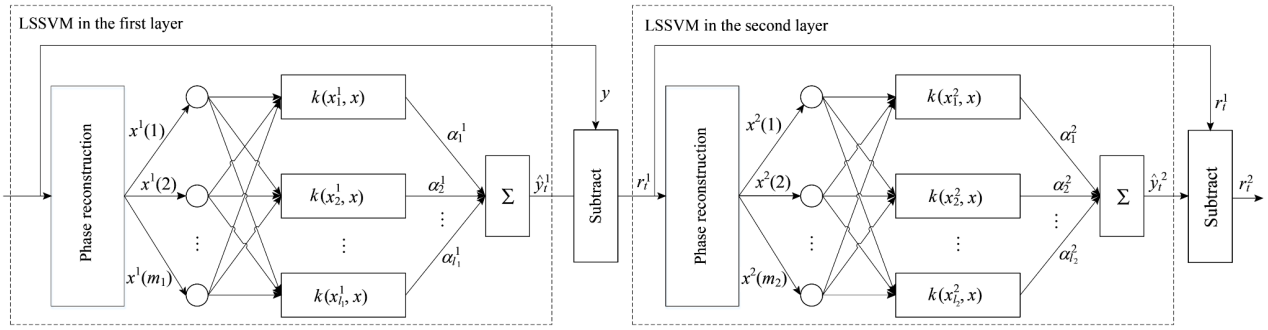

Fig. 1 RBD-LSSVM with two layers.

The third LSSVM models the residuals of the second LSSVM and obtains the third residuals as the input of the fourth LSSVM, and so on. Finally, the last LSSVM in the RDB-LSSVM models the residuals of the previous LSSVM. The predicted value of the RBD-LSSVM is the sum of the predicted values of all the LSSVMs in the RBD-LSSVM. The proof is given as follows:

$$\begin{aligned} y^2(t) &= r^1(t) = y^1(t) - \hat{y}^1(t), \\ y^3(t) &= r^2(t) = y^2(t) - \hat{y}^2(t), \\ &\vdots \\ y^d(t) &= r^{d-1}(t) = y^{d-1}(t) - \hat{y}^{d-1}(t), \\ y^{d+1}(t) &= r^d(t) = y^d(t) - \hat{y}^d(t) \end{aligned} \tag{25}$$

where the superscript denotes the number of layers. By separately adding the quantities in the left and right sides of the above equations and cancelling those that appear in both sides, we obtain the following equation:

$$r^d(t) = y^1(t) - (\hat{y}^1(t) + \hat{y}^2(t) + \hat{y}^3(t) + \cdots + \hat{y}^d(t)) \tag{26}$$

From Eq. (24), we can estimate the original value at time point t $y^1(t)$ by adding the predicted values in all layers of the RBD-LSSVM.

3.3 Model selection based on ODCF and MSE on validation set

Because of the powerful fitting ability of the RBD-LSSVM, simply using the minimum MSE on validation set may not prevent overfitting. The proposed method is described as follows:

(1) Model the original time series using one-layer LSSVM. Hyper-parameters are selected through ODCF test based on a grid search. If there are models validated by ODCF test, compute the MSE by n-fold cross validation on the training set.

(2) Model the original time series using two-layer RBD-LSSVM. Hyper-parameters are selected through ODCF test based on a grid search on two-layer RBD-LSSVM. If there are models validated by ODCF test, compute the MSE of the validation set by n-fold cross validation.

(3) Continue to model the time series using three-layer and four-layer RBD-LSSVMs. Hyper-parameters are selected through ODCF test based on a grid search. If there are models validated by ODCF test, compute the MSE of the validation set using n-fold cross validation.

(4) Continue to model the time series using five-layer RBD-LSSVM, and so on.

Select the RBD-LSSVMs that are validated by ODCF test and have the least MSE obtained by n-fold cross validation; if the MSEs of two RBD-LSSVMs with different layers are similar, select the RBD-LSSVM with the fewer layers.

4 Experimental Results and Discussion

To evaluate the proposed RBD-LSSVM model, comprehensive experiments based on two nonlinear benchmark time series were conducted. The first case study applied the proposed approach to the annually recorded sunspot time series, which is a real-world time series that is commonly used as a benchmark for evaluating time series prediction approaches. In the second case study, another real-world time series, the monthly TCO in New Delhi, measured in Dobson, was predicted.

Note that in all case studies, no separate validation set was used. N-fold cross validation was used. The model validated by ODCF test and with the lowest validation error was selected as the best model and was used to predict test data.

For numerical assessment of the prediction accuracy, the following error criteria were used.

(1) Normalized Mean Squared Error (NMSE)

$$\begin{aligned}\text{NMSE} &= \frac{1}{\delta^2 n}\left(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right),\\ \delta^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n}(y_i-\overline{y})^2\right),\\ \overline{y} &= \frac{\sum_{i=1}^{n} y_i}{n}\end{aligned} \tag{27}$$

(2) Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right)} \tag{28}$$

(3) Pearson correlation coefficient

$$r_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{D_X}\sqrt{D_Y}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{29}$$

4.1 Annual sunspot number

Annual sunspot number is a nonlinear time series that has long served as a benchmark to assess statistical and prediction methods. To make our method comparable with those of other works[21–24], the dataset from 1700 to 1979 was used and was divided into two parts. The data points for years 1700–1955 were used to train the models, and those for years 1956–1979 formed the test set. In the experiment, Matlab2016.a and Matlab-based LSSVM toolbox LSSVMlabv1_8_R2009b_R2011alssvm were employed. To enhance the training efficiency, time delay was set as $\tau = 1$. Therefore the three hyper-parameters to be selected were the penalty parameter C in Eq. (4), kernel parameter σ^2 in Eq. (9), and embedding dimension m in Eq. (23). We selected the optimal model based on a grid search. In the end, the best model was a four-layer RBD-LSSVM, with hyper-parameter values as shown in Table 1.

The MSE computation on validation set and information redundancy test based on ODCF were combined to select the best RBD-LSSVM.

(1) First, we modeled the original time series using one-layer LSSVM. Considering that the training set was not large, we used a confidence interval of 2.8 times standard deviation. No one-layer RBD-LSSVM was validated by ODCF test.

Table 1 Hyper-parameter values for the four-layer RBD-LSSVM validated using ODCF test.

Layer	σ^2	C	m	MSE
1st layer	18	244	6	
2nd layer	11	310	14	
3rd layer	11	1	4	
4th layer	241	1–394 001	4	0.3–0.55

(2) We modeled the original time series using a two-layer RBD-LSSVM. Some models were validated by ODCF test. For the first layer $m = 6, \sigma^2 = 21, C = 51$, and for the second layer, $m = 14, \sigma^2 = 61, C = 51 - 2301$. The models were validated by ODCF test, and 10-fold cross validation MSEs were between 150 and 350.

(3) We modeled the original time series using a three-layer RBD-LSSVM. Some models were validated by ODCF test. For the first layer, $m = 6, \sigma^2 = 21, C = 251$; for the second layer, $m = 14, \sigma^2 = 21, C = 401$; and for the third layer, $m = 14, \sigma^2 = 61, C = 1201 - 2201$. The models were validated by ODCF test, and 10-fold cross validation MSE were between 4.5–5.5.

(4) We modeled the original time series using a four-layer RBD-LSSV as listed in Table 1.

(5) We modeled the original time series using a five-layer RBD-LSSVM. Models were validated by ODCF test with nearly the same value of MSE on the validation set. Therefore, a four-layer RBD-LSSVM is correct.

According to the model selection guidelines in Section 3.3, we selected the four-layer RBD-LSSVM in Table 1. For a certain value of σ^2, as the value of C increased, the model became more complex. This means the model will be changed from underfitting to overfitting. Based on this analysis, we selected $C = 197\,001$, which is the median of range from 1 to 394 001, as is shown in Table 1. The target and predicted outputs for test series of this RBD-LSSVM model is illustrated in Fig. 2.

Considering that only NMSE is used to evaluate the prediction result in Refs. [4, 10, 21–24], we only used NMSE in this experiment. A comparison based on the NMSE with several approaches is presented in Table 2. As shown, the proposed RDB-LSSVM outperforms the other models.

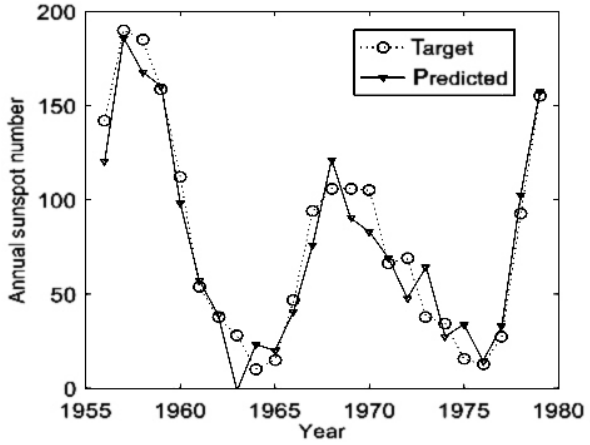

Fig. 2 Predicted and target test series of annual sunspot number for years 1956–1979.

Table 2 Performance comparison for sunspot number series prediction from 1956 to 1979.

Method	NMSE (%)
RBD-LSSVM	6.5
Local LSSVM[4]	7.8
ODCF-SVM[10]	13.5
Neural network[21]	15.1
Benchmark[22]	15.4
Benchmark[23]	35.0

The NMSE metric indicates the overall deviation of the predicted from actual data. To give the distribution of prediction errors, Pearson correlation coefficient was calculated. A correlation of $r = 0.967$ indicates that the predicted and actual data have the same distribution and trend.

4.2 Monthly total column zone in New Delhi

Here, we discuss the TCO observed in New Delhi (latitude 28.65°N, longitude 77.22°E Dobson), which is a real-world highly-complex time series. Wang et al. have predicted this time series in Ref. [9]. We used the same training set and test set as in Ref. [9]. A total of 522 data points from July 1957 to Dec. 2000 were used to train the model, and 48 data points from Jan. 2001 to Dec. 2004 formed the test set. Data were downloaded from world ozone ultraviolet radiation data center with website http://woudc.org/data/explore.php?lang=en.

Its Auto-Correlation Function (ACF) was computed and is depicted in Fig. 3.

From Fig. 3, it can be seen that a seasonality of 12 exists in the original TCO series $\{x_i\}$. Taking a difference of lag-12 with equation $y_i = x_{i+12} - x_i$, the ACF of the resulting time series y_i is shown in Fig. 4.

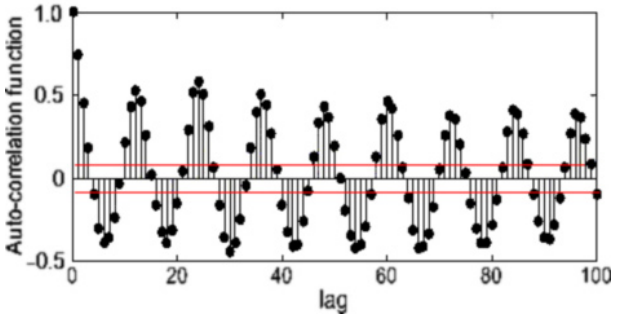

Fig. 3 ACF for original monthly TCO in New Delhi.

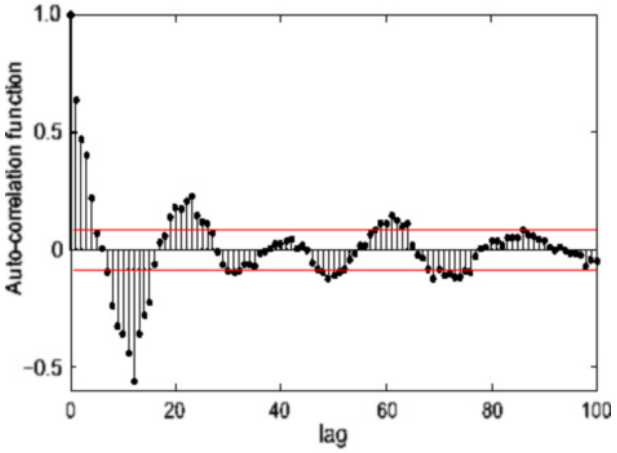

Fig. 4 ACF for seasonally differenced monthly TCO in New Delhi.

Figure 4 shows that the seasonality has been eliminated. We use RBD-LSSVM to model the seasonally differenced time series y_i.

After the similar process with sunspot time series, a four-layer RBD-LSSVM was selected. For this time series, 2.85 times standard deviation was used to test if ODCCF $\rho_{y\epsilon}(\tau)$ and ODACF $\rho_{\epsilon\epsilon}(\tau)$ follow normal distribution with zero means and $1/n$ variance when $\tau \neq 0$. Furthermore, 3 times standard deviation was used to test if ODCCF $\rho_{y\epsilon}(\tau)$ was not zero when $\tau \neq 0$. The RBD-LSSVM validated by ODCF test is listed in Table 3.

Table 3 shows a series of models by ODCF test. We selected $C = 541$, which is the median of the range 31–

Table 3 Hyper-parameter values for the four-layer RBD-LSSVM validated using ODCF test.

Layer	σ^2	C	m
1st layer	3041	6501	49
2nd layer	1	121	6
3rd layer	1	361	2
4th layer	81	31–1051	17

1051 as the value of penalty parameter of the fourth layer of the best RBD-LSSVM. The target and predicted test data of monthly TCO are shown in Fig. 5.

The training and test NMSEs and comparisons provided in Table 4 indicate that the proposed RDB-LSSVM had far better predictions compared to the approaches used in Refs. [7, 9].

5 Conclusion

Time series prediction has been widely used in many areas. SVM and neural networks are two promising methods to predict nonlinear time series. To solve the problems existing in the current ensemble methods based on neural networks or SVM, we propose a novel method, the RBD-LSSVM. Furthermore, to prevent the possibility of overfitting, we propose the use of information redundancy test based on ODCF and MSE on validation set. We present the structure and principles of the RBD-LSSVM, as well as the procedures and methods on how to comprehensively use ODCF and MSE on validation set to select the best hyper-parameters of the RBD-LSSVM. Experiments conducted on two typical nonlinear time series—annual sunspot number and monthly TCO in New Delhi, demonstrate the advantage of the proposed approach.

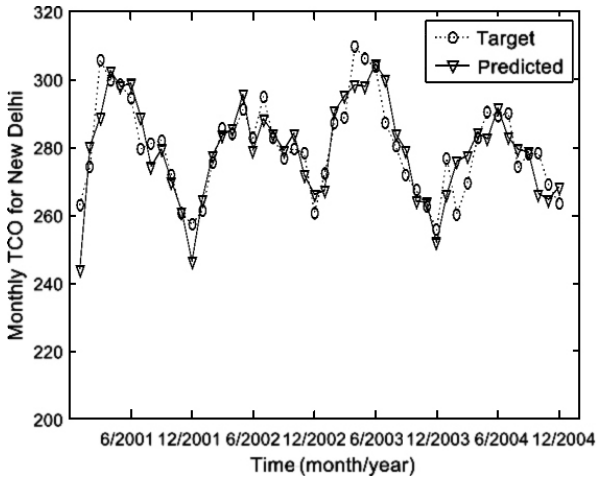

Fig. 5 Target and predicted test series for TCO in New Delhi (48 one-step-ahead predictions).

Table 4 Performance comparison for total ozone column in New Delhi.

Method	RMSE	Pearson correlation coefficient	NMSE
RBD-LSSVM	7.14	0.87	0.2429
Iterative error correction based on direct search in Ref. [7]	8.42	0.80	0.3810
ϵ-SVM in Ref. [9]	10.3	0.68	0.5054

References

[1] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Beijing, China: Posts & Telecom Press, 2005.

[2] H. X. Chen, S. Feng, X. Pei, Z. Zhang, and D. Y. Yao, Dangerous driving behavior recognition and prevention using an autoregressive time-series model, *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 682–690, 2017.

[3] J. Wang, L. Y. Tang, and Y. Y. Luo, A weighted EMD-based prediction model based on TOPSIS and feed forward neural network for noised time series, *Knowledge Based System*, vol. 132, pp. 167–178, 2017.

[4] M. Arash and A. Majid, Developing a local least-squares support vector machines-based neuro-fuzzy model for nonlinear and chaotic time series prediction, *IEEE Transactions on Neural Networks and Learning System*, vol. 24, no. 2, pp. 207–218, 2013.

[5] M. Han and M. L. Xu, A hybrid prediction model of multivariate chaotic time series based on error correction, *Acta Phys. Sin*., vol. 62, no. 12, p. 120510, 2013.

[6] Z. Q. Guo, H. Q. Wang, and Q. Liu, Financial time seriesforecasting using LPP and SVM optimized by PSO, *Soft Computing*, vol. 17, no. 5, pp. 805–818, 2013.

[7] Z. J. Tang, F. Ren, T. Peng, and W. B. Wang, A least square support vector machine prediction algorithm for chaotic time series based on the iterative error correction, *Acta Phys. Sin.*, vol. 63, p. 050505, 2014.

[8] N. I. Sapankevych and R. Sankar, Time series prediction using support vector machines: A survey, *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 24–38, 2009.

[9] G. L. Wang, P. C. Yang, and Y. Q. Mao, On the application of non-stationary time series prediction based on the SVM method, *Acta Phys. Sin.,* vol. 57, no. 2, pp. 714–719, 2009.

[10] Y. H. Yu and J. D. Song, Redundancy-test-based hyperparameters selection approach for support vector machines to predict time series, *Acta Phys. Sin.,* vol. 61, no. 17, p. 170516, 2012.

[11] T. T. Chen and S. J. Lee, A weighted LS-SVM based learning system for time series forecasting, *Information Sciences*, vol. 299, pp. 99–116, 2015.

[12] Q. M. Zhu and L. F. Zhang, Development of omni-directional correlation functions for nonlinear model validation, *Automatica*, vol. 43, no. 9, pp. 1519–1531, 2007.

[13] L. F. Zhang, Q. M. Zhu, and A. Longden, A correlation-test-based validation procedure for identified neural networks, *IEEE Trans. on Neural. Netw*., vol. 20, no. 1, pp. 1–13, 2009.

[14] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd Ed. New York, NY, USA: Springer, 1999.

[15] N. Y. Deng and Y. J. Tian, *Support Vector Machines: Theory, Algorithm and Extension* (in Chinese). Beijing, China: Science Press, 2009.

[16] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, Least squares support vector machines, *International Journal of Circuit Theory & Applications*, vol. 27, no. 6, pp. 605–615, 2002.

[17] L. Ljung, *System Identification: Theory for the User*, 2nd Ed. Prentice-Hall Inc., 1999.

[18] K. Z. Mao and S. A. Billings, Multi-directional model validity tests for non-linear system identification, *International Journal of Control*, vol. 73, no. 1, p. 32, 2000.

[19] W. A. Brock, W. D. Dechert, J. Scheinkman, and B. LeBaron, A test for independence based on the correlation dimension, *Econometric Reviews*, vol. 15, pp. 197–235, 1996.

[20] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Taylor & Francis, 2012.

[21] S. Yilmaz and Y. Oysal, Fuzzy wavelet neural network models for prediction and identification of dynamical systems, *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1599–1609, 2010.

[22] L. J. Cao, Support vector machines experts for time series forecasting, *Neurocomputing*, vol. 51, pp. 321–339, 2003.

[23] A. S. Weigend, B. A. Huberman, and D. E. Rumelhart, Predicting the future: A connectionist approach, *Int. J. Neural Systems*, vol. 1, pp. 193–209, 1990.

[24] H. Tong and K. S. Lim, Threshold autoregressive, limit cycles and cyclical data, *Journal of Royal Statistic Society*, vol. 42, pp. 245–292, 1980.

Yanhua Yu received the BEng, MEng, and PhD degrees from Beijing University of Posts and Telecommunications, China, in 1995, 1998, and 2008, respectively. She is currently an associate professor with the School of Computer, Beijing University of Posts and Telecommunications. She has published more than 30 papers, including more than 10 SCI/EI indexed articles. Her current research interests include time series prediction, network management, data mining, and natural language processing.

Jie Li received the BEng degree from Shenyang University of Aviation and Aerospace in 1999, MEng degree from Sichuan University in 2004, and PhD degree from Beijing University of Posts and Telecommunications in 2009. He has published more than 10 papers. He is now a lecturer with the School of Computer at Beijing University of Posts and Telecommunications. His research interests include cloud computing, mobile network, and data mining.