The Wavelet Transform with best decomposition Level and Relevant Vector Machine Based Approach for Chaotic Time Series Forecasting

WANG Xiao-LU[1,a *], LIU Jian[1,b], LU Jian-Jun[2,c]

[1]School of Communication and Information Engineering, Xi'an University of Science & Technology, Xi' an, Shaanxi , 710054, China

[2]Department of telecommunications Engineering, Xi'an University of Posts & telecommunications , Xi' an, Shaanxi, 710054, China

[a]greattree_wang@sina.com, [b]edliu@bylink.com.cn, [c]ljj@xupt.edu.cn

Key words: chaotic time series; phase space reconstruction; wavelet transform; RVM; forecasting

Abstract: In order to accurately predict the chaotic time series, a novel approach based on integration of wavelet transform and Relevant Vector Machine (RVM) is proposed. The best wavelet decomposition level is determined with the condition that a certain function space orthogonal projection energy in wavelet MRA, is smaller than the largest energy of the forecasting biases. Delay mapping is introduced to transform the different components into new samples of historical characteristics, after wavelet transform. The different new samples are predicted by their corresponding forecasters, respectively. The final forecasting result is obtained by combining all the predicted results. The sparse relevant support vector and its corresponding hyper parameters are calculated on the new sample space of time series by the Sparse Bayesian learning process. Based on which the prediction results are work out. The results show that the approach only using the SVM or RVM based forecaster the averaged prediction biases is more than 10%. The tracking ability and the dynamic behavior are remarkably improved to the averaged biases of 5.43% by using the wavelet transform with best decomposition Series and RVM based forecaster. It is indicated that the suggested approach is feasible and effective.

Introduction

Time series prediction is widely applied in electrical industry[1,2], communication[3], finance[4], coal mines[5], and many other different fields. Many experts and scholars recently have put forward a lot of methods based on nonlinear theories and its combinations for time series forecasting. The Optimally Pruned Extreme Learning Machine (OP-ELM) and the least-squares-support vector machines (LS-SVMs) are applied to the problem of long-term time series prediction[6, 7]. Rafiul et al introduce a new hybrid of Hidden Markov Model (HMM), Fuzzy Logic and multi-objective Evolutionary Algorithm (EA) for building a fuzzy model to predict non-linear time series data[8]. Bhardwaj et al use the single hidden Markov model (HMM) for clustering and combines it with soft computing techniques (fuzzy inference system/artificial neural network) for the prediction of time series[9]. Chandra presents a competitive Cooperative Coevolution (CC) method for training recurrent neural networks for chaotic time-series prediction[10]. WANG et al present an improved extreme learning machine for online sequential prediction of multivariate time series[11]. The prediction models proposed by these methods has achieved a lot of success in the time series forecasting.

However, the changing rules of time series are influenced by a lot of factors, and these factors interact each other. For example, the time series formed from load of power system are highly sensitive on the weather changed[1,2], and the effect degree on the time series of gas concentration

is changed with mining process in coal mine[5]. Therefore, considering the correlation of input variables in the prediction model is helpful to improve the prediction accuracy.

The relevance vector machine (RVM) based on Bayesian framework learning process is a proposed by M. E. Tipping. This method combines the Bayesian principle, determination of automatic relevance and the priori maximum likelihood theory. Considering the correlation to each input factors, the regression forecasting model can also be established. So that it has been widely used. On the other hand, to predict complicated time series, besides selecting appropriate predictor, effective decomposition is also needed. Decomposed components volatility and complexity reduce, over-fitting can be avoided.

This paper presents a forecasting approach based on wavelet transform and RVM. The chaotic time series are decomposed into the low-frequency component and high-frequency components, and delay mapping is introduced to transform the different components into new samples of historical characteristics. Prediction models based on RVM are respectively built for different components. And the best decomposition level of wavelet transforming are calculated by the orthogonal projection energy in wavelet MRA. A large amount of calculation indicates that the proposed approach can forecast suddenly changed data with high accuracy and remarkable adapting ability.

Forecasting model based on wavelet transform and relevant vector machine

Let *s(k)* to be the given time series. *s(k)* are conducted multi-resolution decomposition by using Mallat algorithm, a low-frequency component a_J and a set of different scales of high-frequency components d_i(i=1,2,3…J) are obtained. Then *s(k)* can be expanded as[14]:

$$s(k) = a_J + \sum_J d_j \quad . \tag{1}$$

It is supposed that the *k* moment value of *s(k)* can be forecasted by the historical data of *u* moments before it, the nonlinear mapping of forecasting can be expressed as $T : \boldsymbol{R}^v \to \boldsymbol{R}$, then moment of the predicted value $\hat{s}(k)$ can be expressed as:

$$\hat{s}(k) = T[s(k\text{-}1), s(k\text{-}2), \ldots s(k\text{-}v\text{-}1), s(k\text{-}v)] \quad , \tag{2}$$

After wavelet decomposition, the forecasting value of *k* moment is:

$$\hat{s}(k) = T[a_J] + \sum_J T[d_j] \quad . \tag{3}$$

Delay functional is defined as $D : I \to I^u$, which denotes that one-dimensional input is mapped into u-dimensional output of 1 to *u* delays, that is:

$$D[s(k)] = [s(k\text{-}1), s(k\text{-}2), s(k\text{-}3), \ldots s(k\text{-}u\text{-}1), s(k\text{-}u)] . \tag{4}$$

Let the u-dimensional output, which are obtained by the delay functional, to be the predictor input. Forecasting model is shown in Fig.1.

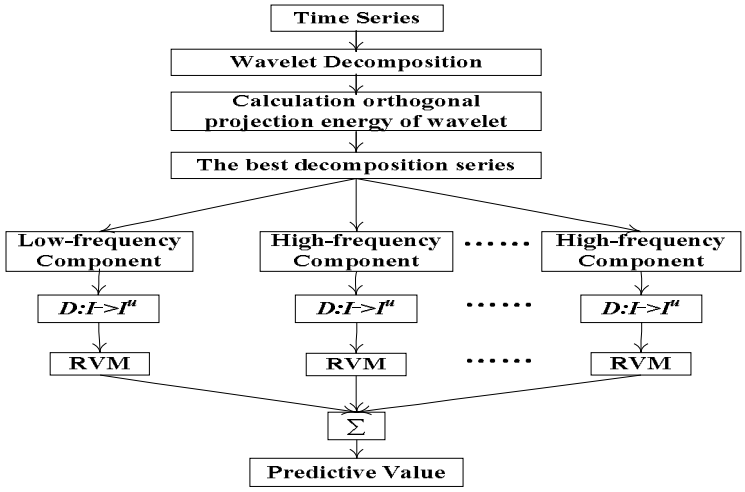


Fig.1 Model for forecasting.

The low-frequency component and high-frequency ones are conducted delay mapping after wavelet decomposition, different delay mapping results are used as input for corresponding RVM predictor, respectively. The best decomposition series are optimized by the calculation orthogonal projection energy based process. Adding the forecasting results of a low-frequency component and a set of high-frequency components together is the prediction value of the original sequence. Using W representation wavelet, this model can be called W_RVM.

Determination of the best composition level of wavelet

It is the key to select decomposition level in wavelet multi-resolution analysis (MRA). If the selected scale is small, it cannot effectively analyze the characteristic of time series, and the selected scale is more, it increases the complexity and difficulty of forecasting. This paper proposed a way to determine the optimal decomposition level from the point of the signal energy.

Definition 1: E_j^A express the energy which is the $f(t)$ orthogonal projection to the scaling function space on V_j for. E_j^D express the energy which is the $f(t)$ orthogonal projection to the corresponding wavelet function space on W_j, $f(t) \in L^2(R)$. E_j^A and E_j^D call the scale energy . As defined by the Parseval theorem and norm definition, there is

$$E_j^A = \left\| f_{A_j}(t) \right\|^2 = \sum_n \left| < f(u), \phi_{j,n}(u) > \right|^2 = \sum_n \left| a_j(n) \right|^2, \quad (5) \qquad E_j^D = \left\| f_{D_j}(t) \right\|^2 = \sum_n \left| < f(u), \psi_{j,n}(u) > \right|^2 = \sum_n \left| d_j(n) \right|^2 . \quad (6)$$

In (5) and (6) A_j and D_j express Orthogonal projection functional on V_j and W_j , respecti vely. $\{\phi_{j,n}(t)\}_{n \in Z}$ and $\{\psi_{j,n}(t)\}_{n \in Z}$ are taken as the orthogonal basis of the scaling function on space of V_j and the wavelet function on space of W_j, respectively. There is $a_j(n)=<f(u),\phi_{j,n}(u)>$, $d_j(n)=<f(u),\psi_{j,n}(u)>$.

By signal decomposition process of the MRA[14], there is

$$E^A{}_j > E^D{}_{j+1}, \quad E^A{}_j > E^A{}_{j+1}, \quad (7) \qquad \left\| f(t) \right\|^2 = E^D{}_j + E^D{}_{j+1} + E^D{}_{j+2} + \ldots + E^D{}_{j+J-1} + E^A{}_{j+J} . \quad (8)$$

After normalization E_j^A and E_j^D, they are expressed as E_j^a and E_j^d, respectively. So there is $E_j^a = E_j^A \Big/ \left\| f_{A_j}(t) \right\|^2$ 和 $E_j^d = E_j^D \Big/ \left\| f_{A_j}(t) \right\|^2$.

Now consideration if the scale energy of a certain decomposition level is small than the energy of the maximum error of the permit prediction accuracy. Which means that this decomposition series can meet the forecasting accuracy requirements. And then the one more decomposition series

must can meet the prediction accuracy requirements, no longer need to decompose.

Lets *u(k)* as the sample, and *L* as the capacity of the sample. The error energy meeting the prediction accuracy is expressed as the sum of square errors, it is taken as η .The normalized energy of maximum prediction error is taken as $\hat{\eta}$ for the prediction error in the range of $\alpha\%$.

$$\hat{\eta} = \sum_{L} [u(k) \times \alpha\%]^2 \Big/ \sum_{L} [u(k)]^2 . \tag{9}$$

Definition 2: *R(j)* are taken as the grading index, which is used to determine the best decomposition level for MRA decomposition algorithm. When $R(j-1) \geq 1,\ 0 < R(j) \leq 1$ The *j*+1 of resolution is the best resolution, that is the best decomposition level of wavelet

$$R(j) = min(E_j^{\ a} / \hat{\eta}, E_j^{\ d} / \hat{\eta}) . \tag{10}$$

RVM regression estimation method

Lets $\{x_n\}_{n=1}^N$ to be a given set of training sample input data and $\{t_n\}_{n=1}^N$ to be a given set of output, the estimating model with noise is

$$t_i = y(x_i, w) + \varepsilon_i . \tag{11}$$

Where, the sample noise ε obeys the Gauss distribution with the mean zero and the variance σ^2, then $p(t_i \mid x_i) = N(t_i \mid y(x_i, w), \sigma^2)$. The output for the RVM model is[11]

$$y(x, w) = \sum_{i=1}^{N} w_i K(x, x_i) + w_0 \tag{12}$$

Where, $w_i (i = 1, ..., N)$ is the weight coefficient, *N* is the number of sample, and the $K(x, x_i)$ is the kernel function. λ is the width of Gauss kernel.

For the independent output set, the likelihood function for the whole of the sample is [11]

$$p(t \mid w, \sigma^2) = \prod_{i=1}^{N} N(t_i \mid y(x_i, w), \sigma) = (2\pi\sigma^2)^{-N/2} \exp(-\| t - \phi(x) w \|^2 / 2\sigma^2 \tag{13}$$

Where, $t = (t_1, t_2, ..., t_N)$, $w = [w_1, w_2, ..., w_N]^T$, $\phi = [\phi(x_1), \phi(x_2), ,..., \phi(x_N)]^T$, $\phi(x_N) = [1, K(x_1, x_N),$ $K(x_2, x_N), ..., K(x_n, x_N)]^T$.

If it has directly been solved the *w* and σ^2 by using of the maximum likelihood method, then it usually leads to the seriously over fitting results. To avoid this phenomenon, the *w* has been given the prior distribution of Gauss with zero mean by the sparse Bayesian principle, that is

$$p(w \mid \alpha) = \prod_{i=1}^{N} N(w_i \mid 0, \alpha_i^{-1}) \tag{14}$$

Where, α is the hyper parameters of the corresponding to weight of *w*. Introduction of the hyper parameters $\beta = \frac{1}{\lambda^2}$, Each weight of *w* are individual corresponding to a hyper parameters.

After defining the prior probability distribution and the likelihood distributions, the all the unknown parameters of the posterior probability distribution have been obtained according to Bayesian principles[11]. They are

$$. p(w \mid t, \alpha, \beta) = \frac{p(t \mid w, \beta) p(w, \alpha)}{p(t \mid \alpha, \beta)} . \tag{15}$$

Both $p(t \mid w, \beta)$ and $p(w, \alpha)$ are the Gauss distribution, $p(t \mid \alpha, \beta)$ are regarded as the normalized coefficient While it does not contain *w*. And then the posterior distribution of *w* can be further expressed as

$$p(t \mid \alpha, \beta) = N(w \mid \mu, \Sigma) \tag{16}$$

where, μ is the mean, $\mu = \beta \Sigma \phi^T t$, and Σ is the covariance, $\Sigma = (\beta \phi^T \phi + A)^{-1}$, $A = diag(\alpha_0, \alpha_2, ..., \alpha_N)$. The hyper parameters can be optimized through maximization the marginal of likelihood function $p(t|\alpha, \beta)$, and then the maximum of the posteriori distribution of *w* can be obtained. The object function is also obtained by $p(t|\alpha, \beta)$ taken negative logarithm. The object function are derive on the hyper parameters α and β, and make the derivative formula equal 0, then the following iteration formula can be got[11].

$$\alpha_j = \frac{\lambda_j}{\mu_j^2} \quad , \beta = \frac{N - \sum_j \lambda_j}{\left\| t - \phi \mu \right\|^2} \tag{17}$$

Where, $j = 0,1,2,..., N$, μ_j is the first j elements of the posterior mean vector μ . $\lambda_j \equiv 1 - \alpha_j \Sigma_{jj}, \Sigma_{jj}$ is the article j diagonal elements of the posterior covariance matrix Σ. the formula (17) have been followed by iterative calculation, until all the parameters converge or the number of iterations reach to the maximum number of training [11].

The hyper parameters α_{MP} and σ_{MP}^2 has been got by maximum likelihood method[11]. Given the input value x_*, the corresponding predicted values is

$$y_* = \mu^T \phi(x_*) . \tag{18}$$

Application example

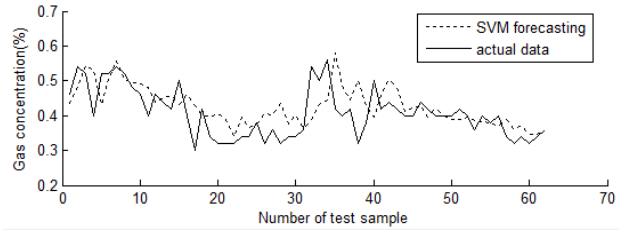
In this section, 242 samples of gas concentration from June, 2006 to March, 2007 in Changjiashan coal 416 working face are used as an example, and according to the second proposed RVM method as a predictor to expound the application of the proposed approach.

As comparison，the RVM and SVM are taken as time series predictors, respectively, and the parameter is set as following: SVM: balance parameter *C*=1, insensitive loss function ε =0.01, Gauss kernel function parameters λ_{SVM} =0.01, RVM：the width of Gauss kernel function λ_{RVM} =0.001. The first 180 of the 242 gas concentration samples are taken as the training sample, and the remaining 62 sample data as test samples to validate the prediction performance of the proposed method.

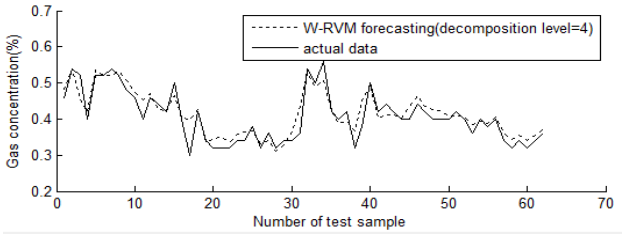
Using the compactly supported biorthogonal wavelet db7 as mother wavelet, then the 180 training sample are decomposed the a set of low frequency components and a high frequency ones. According to the formula (5) and (6) , the each level scale energy are calculated . Then According to the formula (9) and (10)，$\hat{\eta}$ =0.166，*R*(2)= 1.03 are work out. So the best decomposition level is 4.

The number of delay unit in the delay functional D is set 8 in accordance with the experience.

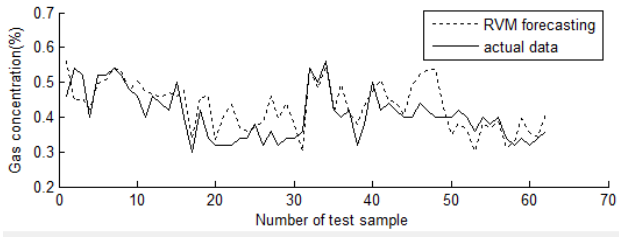
For comparison, the SVM and RVM are direct using to predict the test samples, respectively. And the decomposition level is set 4 and 5 in wavelet transform, then combined with RVM (W-RVM) prediction method to predict the test samples, respectively, the results of four kinds of prediction methods are obtained as shown in Fig.2.



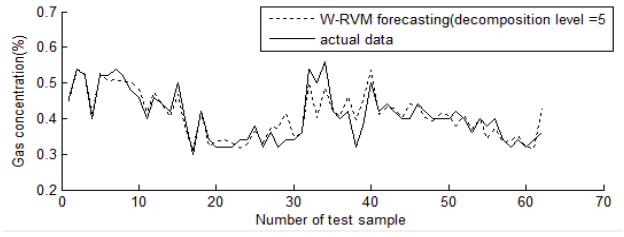
(a) Forecasting result of SVM



(c) Forecasting result of W-RVM (decomposition level =4)



(b) Forecasting result of RVM



(d) Forecasting result of W-RVM (decomposition level =5)

Fig.2 Results of forecasting Shown

form Fig.2, When the gas concentration changed, especially sudden changed, the predicted results by the RVM or SVM can close to the actual data after long time, it is shown in Fig.2 (a) and (b). After the wavelet transform with the best decomposition level, prediction performance was improved obviously by using the RVM prediction, such as shown in Fig.2 (c). But the decomposition level setting 5, which is one more than the best wavelet transform decomposition level in wavelet transform, then using the RVM prediction, there is almost little change in prediction performance, the prediction results are shown in Fig.2 (d). It can be ignored the contribution of the prediction performance for adding wavelet decomposition level under the meeting prediction accuracy.

The above 4 methods of their prediction performance for gas concentration time series prediction are shown in Table1.

It is shown from the Table1 that RVM and SVM have high precision of prediction and training. Comparison with SVM in the condition of same sample, the number of vectors of RVM is less, RV=119, SV=149. Which means that the RVM method has faster computing speed and less memory in the practical application.

Table1. Forecasting capability

Prediction Method	SVM	RVM	W-RVM Decomposition level = 4	W-RVM Decomposition level = 5
Max error (%)	56.49	36.70	23.64	31.90
Min error (%)	0.12	0.05	0.05	0.10
Mean error (%)	12.27	11.64	5.43	5.42
Mean square error	0.50	0.46	0.22	0.24
Number of Support vector	149	119	102	102

Shown from Fig.2 (c) and Table1, it can track the change of gas concentration data by using wavelet transform and the RVM predictor, the average prediction error is 5.43%. Using wavelet components obtained the best decomposition level, which ensure the prediction accuracy, at the same time, the required components are the most suitable. It is help to streamline the prediction model, and reduce the amount of calculation.

Conclusion

（1）The ratio of the scale energy and the energy of the maximum error with meeting the forecasting precision, which can determine the best wavelet decomposition level in prediction process.

（2）the adapting ability and dynamic behavior are better improved by using the predictor based on RVM for the complicated chaotic time series prediction.

（3）The results of application show that the chaotic time series is decomposed by wavelet transform with the best decomposition level, then the RVM is taken as predictor. The performance of the forecasting is remarkably improved, and the obtained averaged biases is 5.43%, which indicates that the proposed approach is feasible and effective.

Acknowledgment

This work was supported by the department of science and technology project of Shaanxi province in China (No.2012K06-13), The department of education project of Shaanxi province in China (No.2013JK1156), the Ph.D. starting project of Xi'an University of Science and Technology （No.2013QDJ022）.

References

[1]K. WAN and R. LIU: Power System Technology, Vol.36 (11) (2012), p76, In Chinese

[2]X. WANG and L. MENG: Power System Protection and Control, Vol.43(1)(2015), p.61, In Chinese.

[3]L. WEND, J. YANK and I. KEHA: Journal of Electronics & Information Technology, Vo1.33, No.4, p.885, 2011.

[4]B. ZHOU and H. YAN: Systems Engineering—Theory & Practice, Vo1.33(10)(2013), p.2654, In Chinese.

[5]J. CHENG，J. BAI and J. QIAN: Journal of China University of Mining & Technology，Vo1.37(2)(2008), p.231.

[6]G. Alexander, M. Yoan and V. Anne-Mari : Neural Networks, Vo1.51(2014), p.50.

[7]L.J. Herrera, H. Pomares and I. Rojas: Neurocomputing, Vol.70(2007), p.2870.

[8]H. Rafiul, N B Md and M Kirley: Neurocomputing, Vo1.81(2012), p.1-11.

[9]S. Bhardwaj, S. Srivastava and J.R.P Gupta: Computational Intelligence, Vo1.31(1)(2015), p.106.

[10]R. Chandra: IEEE Transactions on Neural Networks and Learning Systems, (2015, March 5), p.128.

[11]X. WANG and M. HAN: Engineering Applications of Artificial Intelligence, Vo1.40(2015), p.28, In Chinese.

[12]Tipping M E: Journal of Machine Learning Research; Vol.1(3)(2001), p.211.

[13] X. WANG: Intelligent Forecasting Approaches of Development Trends of Coal Gas Monitoring Data[Dissertation]. Xi'an University of Science and Technology,2011, In Chinese.

[14]X. TANG: Wavelet analysis and application[M].Chongqing University Press，2006, In Chinese.