International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.8 (2016), pp.1-12
http://dx.doi.org/10.14257/ijmue.2016.11.8.01

Short-Term Prediction of Stock Index Based on EMD and SVMs

Wen Chen [1,2] and Yixiang Tian [1]

[1.] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu,611731 China
[2.] Sichuan University of Arts and Science, Dazhou ,635000 China

Abstract

In allusion to the prediction of the stock return rate, this paper has proposed nonlinear combined prediction method based on EMD (Empirical Mode Decomposition) and SVMs (Support Vector Machines). The method has divided stock return rate series into several components of different frequencies using EMD technology, getting three new sequences by grouping superposition of each component according to the frequency, which represents items of market volatility, major events, and trend respectively. Based on these three sequences, prediction is made by constructing different SVMs models to obtain the predicted value of each sequence. With SVMs, a combined model is built on the basis of predictive value of each sequence to obtain the final predicted value of stock return rate. Using CSI 300 Index, the validity of this method is verified, and the results show that the proposed model is better than the other models presented in this paper on forecasting CSI 300 Index.

Keywords: *Stock-index; EMD; IMF; SVMs; Combined Prediction*

1. Introduction

Stock market prediction, because of its huge financial returns, has become a hot topic in recent years. Accurate forecasting stock index have great practical significance for investment and academic research. What is more, the prediction of stock index accurately is a very challenging task as the stock index sequence is nonlinear and non-stationary time one. The early stock market prediction was based on the time series analysis techniques which were focused on the statistic, such as ARMA (auto-regression moving average,) [1-2]. With the development of artificial intelligence (AI), artificial neural network (ANN) has been used for financial forecast. A large number of successful applications of this method indicate that ANN is a good tool for time series analysis [3-5]. However, due to the structural constraints, Ann's method itself, it can not demonstrate its performance under a lot of so-called noise from the stock market. What is more, SVM (Support Vector Machine) method has also been used to predict financial time series [6-13]. As a new signal analysis theory and a quantum leap of the linear and steady spectrum analysis on the basis of Fourier transform in recent years, EMD (Empirical Mode Decomposition) has stronger partial properties and adaptability than the wavelet transform [14,15]. Therefore, the integration of EMD and SVR (SVMs-Regress) technology was used to improve the accuracy of forecasting the stock in order to provide timely information to help investors to make the right investment decisions [18-20].

With technical ideas similar to integration of EMD and SVR, this paper has conducted modeling and prediction of stock return rate series based on prediction methods of EMD and SVMs, but the difference is the use of glide time-windows to dynamically update model parameters to achieve the purpose of portraying dynamics of time series more clearly, and the use of the data CSI 300 Index to validate the approach proposed in this work.

In this paper, the overall structure of the study is as follows: In Section 2, the Empirical Mode Decomposition and the principle of Support Vector Machine is introduced. In Section 3&4, we establish the EMD-SVM-SVMs model to predict CSI 300 Index. At last, the conclusion is given in Section 5.

2. Model Principle

2.1. Mode Decomposition

EMD has been proposed by HUANG *et al.* [14], which decomposes the signal into different frequency IMFs (intrinsic mode functions) through a "filter". EMD, a multi-scale analysis, is essentially a smooth process for a signal, As a result, the fluctuations of the signal at different scales are gradual decomposed, and a limited number of data sequences and a residual signal scales with different characteristics are produced. Each sequence is an IMF, whose components must satisfy the following two conditions:

(1) In the entire data series, the number of extreme points must be equal to the number of zero-crossings, or at most a difference of one.

(2) The signal envelope mean defined by local maxima and local minima is zero at any point.

Through a "Sifting" iterative procedure, EMD can make the arbitrary signal decomposed into a finite number of IMF and residual signals. Specific steps are as follows:

(1) Identify all the local maxima and minima of $X(t)$.

(2) Obtain the upper envelope $X_u(t)$ and the lower envelope $X_l(t)$ of the $X(t)$.

(3) Use the upper envelope $X_u(t)$ and the lower envelope $X_l(t)$ to compute the first mean signal series $k_1(t)$, that is, $k_1(t)=[X_u(t)+X_l(t)]/2$.

(4) Evaluate the difference between the original time series $X(t)$ and the mean signal series and get the first IMF $h_1(t)$, that is, $h_1(t)=X(t)-k_1(t)$. Moreover, we see whether $h_1(t)$ satisfies the two conditions of an IMF property. If they are not satisfied, we repeat steps1-3 of the decomposition procedure to eventually find the first IMF.

(5) After we obtain the first IMF, repetition of the above steps is necessary to find the second IMF, until we reach the final signal series $r_n(t)$ as a residue component that becomes a monotonic function, which is suggested for stopping the decomposition procedure.

(6) The original time series $X(t)$ can be reconstructed by summing up all the IMF components and one residue component as Eq. (1) ,as follows.

$$X(t)=\sum_{i=1}^{n}h_i(t)+r_n(t) \qquad (1)$$

2.2. Decomposition Component Recombination

The IMFs extracted by the above decomposition are characterized by different time scales; therefore, the IMF components can then be recombined according to high-low frequency. In practice, according to descending order, we have a recombination of IMFs. The steps are as follows:

(1) Calculate the mean IMFs $h_1(t)$ to $h_n(t)$.

(2) Use t-tests to determine significant deviation from the mean zero IMF $h_i(t)$.

(3) Make a simple sum of the first to the *(i-1)-th* IMF to reconstruct the high-frequency part and a simple sum of the *i-th* to the *n-th* IMF to reconstruct the low-frequency part.

2.3. Support Vector Machine

SVM, a new machine learning technique based on statistical learning theory proposed by VAPNIK [16,17], is a great synthesizer of such techniques as the maximum interval hyperplane, mercer nuclear, convex quadratic programming, sparse solution and slack variables. This method has a simple structure, fast learning, global optimization, good generalization performance, *etc.* and can solve problems such as the small sample, nonlinearity, high dimension and local minima. The basic idea of the model is: For a given set of training samples $T=\{(x_i,y_i)\}_{i=1,\ldots,n}$, (which x_i is the input variable, y_i is the output value , n is the total number of samples), SVR performs by a non-linear mapping ϕ ,in which the input variable x of the input data space is mapped to a high dimensional feature space; and then runs by linear regression in the feature space to construct the optimal learner:

$$f(x)=\omega^T\phi(x)+b, \qquad (2)$$

Where, ω and b are estimated by regularization and structural risk criteria.

According to the structure risk minimization, we have:

$$\begin{aligned}
&\min_{w,b} && \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)\ , \\
&s.t. && y_t-\omega^T\phi(x_t)-b\le\varepsilon+\xi_i, \\
& && \omega^T\phi(x_t)+b-y_t\le\varepsilon+\xi_i^*, \qquad (3)\\
& && \xi_i,\xi_i^*\ge 0, i=1,\ldots,n \\
& && C>0
\end{aligned}$$

where ξ_i and ξ_i^* is the slack variable, respectively, ε is the insensitive loss coefficient , C is a penalty factor to be used to limit the minimization of estimation error. To solve this problem we introduce a Lagrange optimization function:

$$L(w,b,\xi_i,\xi_i^*,a,a^*,\gamma,\gamma^*)=$$

$$\frac{1}{2}\|w\|^2-C\sum_{i=1}^{n}(\xi_i+\xi_i^*)-\sum_{i=1}^{n}a_i[\xi+\varepsilon-y_i+f(x_i)]-\sum_{i=1}^{n}a_i[\xi_i^*+\varepsilon+y_i-f(x_i)]-\sum_{i=1}^{n}(\xi_i\gamma_i+\xi_i^*\gamma_i^*)$$

Calculation of the partial derivatives, we have:

$$\frac{\partial L}{\partial w}=0\,,\quad \frac{\partial L}{\partial b}=0\,,\quad \frac{\partial L}{\partial \xi_i}=0\,,\quad \frac{\partial L}{\partial \xi_i^*}=0$$

According to the formula, we have:

$$\begin{cases}
\sum_{i=1}^{n}(a_i-a_i^*)=0 \\
w=\sum_{i=1}^{n}(a_i-a_i^*)x_i \\
C-a_i-\gamma_i=0, i=1,\cdots,n
\end{cases}$$

We can substitute the above formula into the Lagrange function expression, using the least squares method to obtain an expression for nonlinear prediction:

$$f(x)=\sum_{i=1}^{n}(a_i-a_i^*)k(x,x_i)+b \qquad (4)$$

Where $k(x,x_i)$ is called the "kernel function" and must satisfy Mercer's theorem (Vapnik, 1995). The value of the kernel equals the inner product of two vectors, x and x_i, in the feature space $\phi(x)$ and $\phi(x_i)$, that is, $k(x,x_i)=\phi(x)\cdot\phi(x_i)$.

The most widely used kernel function is the Gaussian radial basis function (RBF), defined as: $k(x,x_i)=\exp(-\|x-x_i\|^2/2\delta^2)$. The RBF kernel is not only easier to implement than alternatives, but it is also capable of nonlinearly mapping the training data into an infinite-dimensional space; thus, it deals suitably with nonlinear relationship problems. Thus, the Gaussian RBF kernel function is used in this work.

3. EMD-SVM-SVMs

EMD can adaptively decompose non-stationary stock return rate series into several IMFs of different frequencies according to their internal characteristics, and IMF can greatly highlight local characteristics of stock return rate after high-to-low regrouping according to frequency. According to many scholars' point of view, stock-index is composed of items of market volatility, major events and trend, and the analysis of the part with specific economic implications in these three groups can more clearly grasp the characteristics of stock-index.

According to the data characteristics of each group, it has predicted using SVMs dynamic model of different types of kernel functions and parameters under migration time window respectively, and at the same time, it has implemented modeling in the syntagmatic relation between predictive values of subentries using SVMs to obtain final predicted value. The basic idea is to take the predictive value of the same time as the input, and actual stock return rate at that time as the output. After learning enough samples, it is easy to establish the arithmetic mapping relation between the predicted value and the actual value of each component. For a well-trained model, its input variables are the predicted value of all subentries, while its outputs are the final predicted value of stock return rate.

In summary, the proposed prediction method can be expressed as EMD-SVM-SVMs model, and the basic steps are as follows:

Step 1 EMD decomposition is performed for stock return rate series to get n IMFs and 1 residual component;

Step 2 IMF is combined into two parts according to the height of frequency, with the remaining components unchanged;

Step 3 SVMs predictive model are established for the three sub-sequences mentioned above to predict;

Step 4 SVMs combined model is obtained by inputting obtained predictive value of each sequence so as to get the final predictive value of stock return rate.

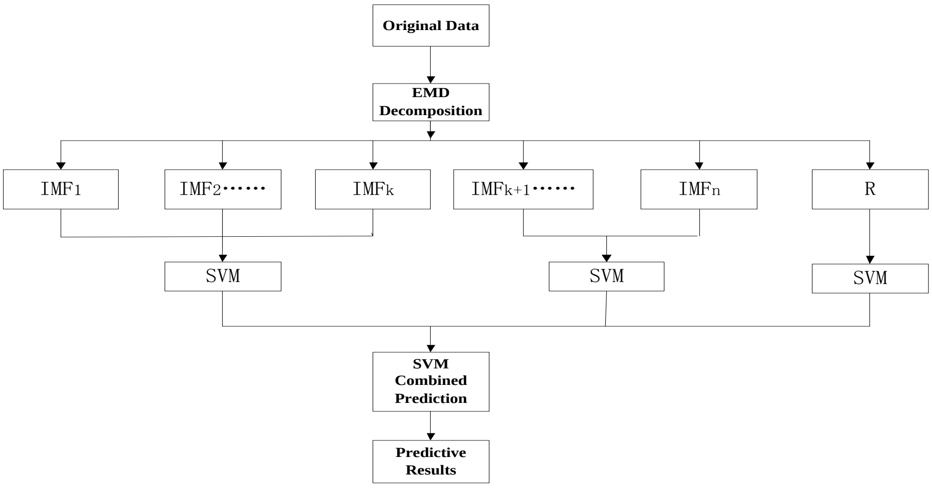

Figure 1. Workflow of EMD-SVM-SVMs Predictive Model

4. Data Experiment

4.1 Experimental Data and Evaluation Criteria

In this paper, the CSI 300 Index series is adopted as the data of test sample, the data is collected from CSI 300 index from April 8, 2005 to May 8, 2015 (a total of 10 years) , thus getting 2449 data of CSI 300 Index (closing quotation). As for CSI 300 Index, the data of closing stock return rate on 1669 trading day is taken as the starting point of the test set, with forward forecast to the data point of 2449. Among them, taken any data *P* in the test set as the border and *L* points of the data point *P* backward (not including point *P*) as the SVM's input, thus point *P* is the SVM's output, namely, one-step forward forecast. In the experiment, the gliding-time-window method is used to obtain one-step-ahead prediction. Prediction accuracy of the model and predictive ability of data directional movements is measured using Root-Mean-Square Error (*RMSE*) and Direction Symmetrical Rate (*DS*).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{x}(t) - x(t))^2} \qquad (5)$$

$$DS = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad (6)$$

$$Where\ d_i = \begin{cases} 1,\ [x(t)-x(t-1)][\hat{x}(t)-x(t-1)] \geq 0 \\ 0,\ [x(t)-x(t-1)][\hat{x}(t)-x(t-1)] < 0 \end{cases}$$

4.2. Experimental Results

The following is the description of the experiment. Firstly, the EMD decomposition is performed for CSI 300 Stock return rate, shown in Figure 2.

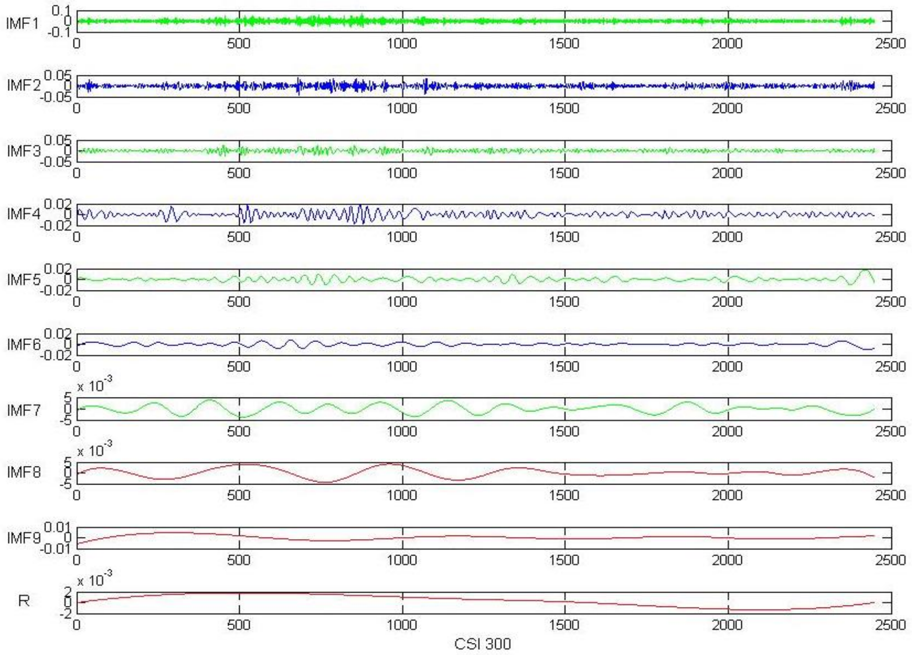

Figure 2. The IMFs and One Residue for CSI 300 Stock Return Rate via EMD

As shown in Figure 2:

1) CSI 300 stock return rate sequence is decomposed into 9 IMFs and 1 residual component, which has reflected the complex multi-time scale and multi-level of the diurnal variation rate;

2) In terms of time series of daily stock return rate, the first IMF is IMF1, a fluctuation with the greatest amplitude, the highest frequency, and the shortest wavelength. Following it, the amplitude of other IMFs decreases sequentially, with gradually decreased frequency and increased wavelength;

3) IMF1 component regards the quasi-two trading day (hereinafter referred to day) as the main cyclical fluctuations, with minor period of fluctuations of quasi-three, quasi-four and quasi quasi-five, *etc.*, and the longer the period is, the smaller the number of occurrences will be. IMF2 component has a cyclical fluctuation of quasi-five trading day; IMF3 component has a main cyclical fluctuation of quasi-twelve day; IMF4 component has a main cyclical fluctuation of quasi-25 day, supplemented by fluctuation cycle of quasi-30 day; IMF5 component has a main cyclical fluctuation of quasi-50 day, with the fluctuation cycle of quasi-60 day nested; IMF6 component has a cyclical fluctuation of quasi-80 to 100 day; IMF7 component has a cyclical fluctuation of quasi-200 day; IMF8 component has a cyclical fluctuation of quasi-400 day; IMF9 component has a cyclical fluctuation of quasi-900 day; the remaining component reflects the change in the overall trend of daily stock return rate time series.

4) It can be seen from Figure 2 that IMF1~IMF4 components have obvious "persistence", that is, long-term memory, reflected in the change in fluctuation range of the daily stock return rate series. From the point of the average, it shows that the sharp fluctuation trend of IMF component in the past means that there will be a great fluctuation trend in the future, while the smaller fluctuation trend of IMF component in the past means that there will be a minor fluctuation trend in the future, which shows the aggregation characteristics of the fluctuation unless key events changing this trend arise, and also implies that IMF fluctuation component sequence has showed certain nongaussianity.

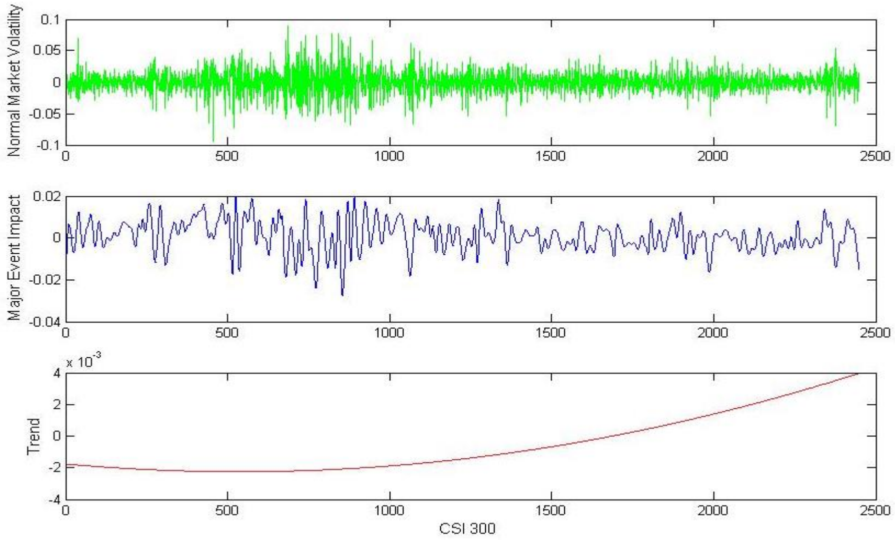

Figure 3. Re-construction of the Data from the IMF Components

The average of each IMF of CSI 300 stock return rate series is checked, and it has been found that the first IMF with the average significantly deviating 0 is IMF5. By adding IMF1 to IMF4, that is, the high-frequency components with the mean of near 0, we can obtain the so-called the high-frequency part, which stands for item of normal market volatility of stock return rate series. Similarly, by adding IMF5 to IMF9, the low-frequency part can be obtained, which stands for item of major events impact, and the unchanged remaining components represent the item of long-term trend of stock return rate. After regrouping, the results are shown in Figure 3.

As what can be seen from Figure 2-3, the item of trend is a major component of stock return rate series, having a decisive impact on the long-term trend of the stock return rate. From a historical point of view the stock return rate, despite the fact that stock return rate will undergo acute fluctuation under the influence of some major events, as the incident disappears, the stock return rate will return to the vicinity of stock return rate trend item. Major-events item is composed of the low-frequency IMFs, an important component of stock return rate, and violent fluctuations will appear under the influence of major events. In addition to being influenced by trend item and major events, stock return rate is also affected by many other factors, such as weather, people's emotions and rumors, and so forth. These factors have a short time of impact and can be classified into high-frequency market volatility item, which is composed of the high-frequency IMFs. The item of normal market volatility has little effect on stock return rate, but because of its frequent fluctuation, the item has a vital role in the short-term forecasting of stock return rate. According to the change law of market volatility item, major events influence item and trend item respectively, different SVM functions and parameters are selected to establish gliding-time dynamic prediction model, and then the combination forecasting is implemented. Selections of kernel functions and parameters are shown in Table 1.

Table 1. Kernel Function & Parameters of SVMs Model of Subentries

Kernel Function	High Frequency	Low Frequency	Trend
	RBF	RBF	Polynomial
Parameters Training and Setup	C_1 = 0.5	C_1 =0.5	C_1 =0.57435
	C_2 = 0.0625	C_2 =0.5	C_2 =0.57435
	C_3 = 1.0	C_3 =0.5	C_3 =0.57435
	C_4 = 0.70711	C_4 =0.5	C_4 =0.57435
	δ_1 = 16	δ_1 =0.0625	δ_1 = 0.0039063
	δ_2 = 0.70711	δ_2 =0.0625	δ_2 = 0.0039063
	δ_3 = 0.17678	δ_3 =0.0625	δ_3 = 0.0039063
	δ_4 = 0.25	δ_4 =0.0625	δ_4 = 0.0039063
	ε =0.01	ε =0.01	ε =0.01

NOTE: Table 1 has listed the actual value used for the first four predictions in the model parameters trained in the one-step forward prediction with the starting point of 1669th trading day, when the width of time window is L = 22, in which C_i represents penalty factor, δ_i indicates parameter of kernel function, ε stands for insensitive loss factor, and i represents the model subscript. Subscript i = 1,2,3,4 indicates training parameters in one-step forward prediction for the first, second, third and fourth time respectively. Due to space limitations, the remaining values are omitted.

After the forecast of all subentries is made with selected corresponding SVMs, the prediction of error is shown in Table 2.

Table 2. Prediction Error of Subentries

	RMSE	*DS %*
High Frequency	0.012798	75.3442
Low Frequency	0.0016959	87.7805
Trend	0.0000025457	100.000

As what can be seen from Table 2, SVMs model has pretty excellent predictive ability of low-frequency items of major events and trend, inferior performances in terms of high-frequency market volatility, which is mainly due to the fact that high-frequency market volatility item is greatly affected by market speculation, also affected by factors of weather, people's emotions and rumors and so on, which has undoubtedly increased the difficulty of forecasting. IMFs are combined into two items of high and low frequency to forecast, which helps to avoid the accumulation of errors in prediction process, and can improve the prediction accuracy of the final combination. In combined forecasting, sigmoid kernel function is used. Results of the final combined prediction of CSI 300 Stock return rate are shown in Figure 4.

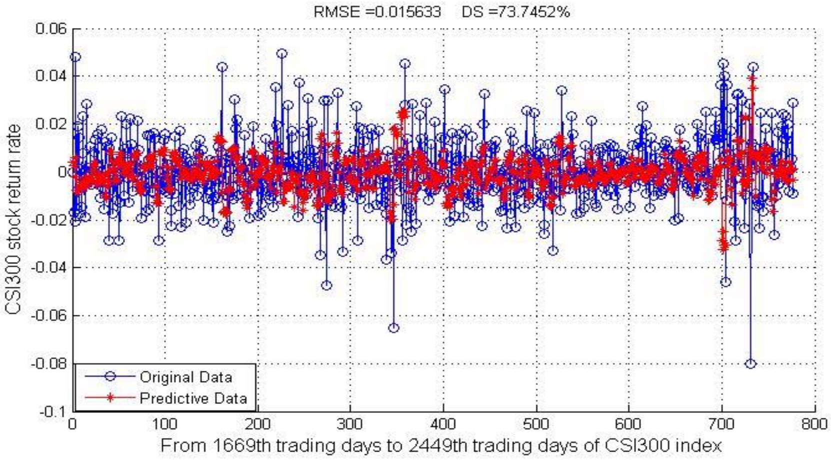

Figure 4. Final Prediction of CSI 300 Stock Return on 1669th-2449th Trading Days

Table 3. Comparison of Predictive Stock Return Rate

	RMSE	*DS*|%
EMD-SVM-SVM	0.015633	73.7452
Single SVM	0.019892	62.6289
Single BPNN	0.017300	53.610

Note: 780 steps forward

Based on the same data set, this paper has selected single BPNN (BP neural network) model and single SVM model to make comparison, and the predicted results are shown in Table 3. The results show that *DS* and *RMSE* of predicting of EMD-SVM-SVMs is significantly better than single BPNN and single SVM model, thereby indicating that the model proposed in this paper has a preferred predictive ability.

5. Conclusion

The non-stationary stock return rate sequence is decomposed with EMD technology, and the decomposed sequence is regrouped according to different frequencies, so as to constitute time series with different economic implications. Different predictive models are established in allusion to these sequences using SVMs, which has reduced complexity of modeling on one hand, and can more accurately portray the data characteristics of time series with different economic implications on the other hand. This contributes to better illustrating the predictive results.

However, we believe this work is still at its early phase as there is still large space of possibilities to be explored. One of the most promising research directions is to investigate the possibility of forecasting the movement direction of stock market with other hybrid SVM. Another possibility is to introduce new technology (such as HMM et) into this model.

Acknowledgements

This research was supported by National Social Science Fund under Grant No. 14BJY174, and The Ministry of education of Humanities and Social Science project under Grant No 12YJA790125.

References

[1] H. J. Douglas, "Time series analysis", Princeton: Princeton university press, vol. 2, **(1994)**.

[2] H. S. Kim, R. Eykholt, and J. D. Salas, "Nonlinear dynamics, delay times, and embedding windows", Physica D: Nonlinear Phenomena, vol. 127, no. 1, **(1999)**, pp. 48-60.

[3] T. Kimoto, K. Asakawa, M. Yoda and M. Takeoka, "Stock market prediction system with modular neural networks", 1990 IJCNN International Joint Conference on Neural Networks, **(1990)**.

[4] S. Ramesh and R. B. Patil, "Connectionist approach to time series prediction: an empirical test", Journal of Intelligent Manufacturing, vol. 3, no. 5, **(1992)**, pp. 317-323.

[5] C. Wei, W. Wagner, and C. H. Lin, "Forecasting the 30-year US treasury bond with a system of neural networks", NeuroVe $ t Journal, **(1996)**, pp. 10-16.

[6] J. C. B. Christopher, "A tutorial on support vector machines for pattern recognition", Data mining and knowledge discovery, vol. 2, no. 2, **(1998)**, pp. 121-167.

[7] C. Corinna and V. Vapnik, "Support-vector networks", Machine learning, vol. 20, no. 3, **(1995)**, pp. 273-297.

[8] E. H. T. Francis and L. Cao, "Application of support vector machines in financial time series forecasting", Omega, vol. 29, no. 4, **(2001)**, pp. 309-317.

[9] K. K. Jae, "Financial time series forecasting using support vector machines", Neurocomputing, vol. 55, no. 1, **(2003)**, pp. 307-319.

[10] U. Thissen, R. V. Brakel, A. P. De Weijer, M. J. Melssen and L. M. C. Buydens, "Using support vector machines for time series prediction", Chemometrics and intelligent laboratory systems, vol. 69, no. 1, **(2003)**, pp. 35-49.

[11] H. T. Jung, H. F. Hsiao and W. C. Yeh, "Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithm", Neurocomputing, vol. 82, **(2012)**, pp. 196-206.

[12] W. Baohua, H. Huang and X. Wang, "A support vector machine based MSM model for financial short-term volatility forecasting", Neural Computing and Applications, vol. 22, no. 1, **(2013)**, pp. 21-28.

[13] X. Zhikun, Y. Gao and Y. Jin, "Application of an Optimized SVR Model of Machine Learning", International Journal of Multimedia & Ubiquitous Engineering, vol. 9, no. 6, **(2014)**.

[14] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, The Royal Society, vol. 454, no. 1971, **(1998)**.

[15] C. Li, X. Wang, Z. Tao, Q. Wang and S. Du, "Extraction of time varying information from noisy signals: An approach based on the empirical mode decomposition", Mechanical Systems and Signal Processing, vol. 25, no. 3, **(2011)**, pp. 812-820.

[16] V. Vapnik, S. Golowich and A. Smola, "Support Vector Method for function estimation", Regression estimation and Signal processing, vol. 9, **(1997)**.

[17] V. Vladimir, "The nature of statistical learning theory", Springer Science & Business Media, **(2013)**.

[18] W. H. Bo and Q. B. Zhu, "Trend prediction of non-stationary vibration signals based on Empirical Mode Decomposition and Least Square Support Vector Machine", Computer Engineering and Applications, vol. 16, **(2008)**, pp. 049.

[19] L. C. Sin, S. H. Chiu and T. Y. Lin, "Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting", Economic Modelling, vol. 29, no. 6, **(2012)**, pp. 2583-2590.

[20] Y. F. Yang, Y. K. Bao, Z. Y. Hu and R. Zhang, "Crude Oil Price Prediction based on Empirical Mode Decomposition and Support Vector Machines", Chinese Journal of Management, vol. 12,; **(2010)**, pp. 022.

Authors

Wen Chen, He received his MS degree in 2007 from University of Electronic Science and Technology of China. He is currently working toward PhD degree in UESTC Since 2011, he has been working as a teacher at Sichuan University of Arts and Science, and since 2009, he was a lecturer in school of Mathematics and Finance of Sichuan University of Arts and Science. His research interests are machine learning and financial engineering.

Yixiang Tian, He received the BS degree from Sichuan Normal University (1985), and received the MS degree from Sichuan University (1993), and received PhD degree from Huazhong University of Science and Technology, China. He is currently a professor and doctoral supervisor in School of Management and Economics, University of Electronic Science and Technology of China. His research interests are in the areas of econometrics, financial risk and the data mining.