Forecasting Framework Using Hybrid Modeling and Support Vector Regression

Lingjuan Xia¹, Xueyong Liu²* and Lixian Wang³

¹Canvard College, Beijing Technology and Business University, Beijing 101118, China
²School of Management and Engineering, Capital University of Economics and Business, Beijing 100070, China
³Shanxi CBM Exploration and Development Branch of PetroChina Company Limited, Shanxi 046000, China

*Corresponding author: email: liuxueyong43@126.com

Abstract. Time series forecasting for financial market has increasingly attracted the interests of investors and academic researchers. In recent years, some hybrid models have been constructed to improve the predictions since some methods cannot extract useful information from stock time series with noise to conduct prediction. In this research, a prediction framework is proposed to forecast the stock market behavior using the methods of wavelet coherence, multiscale decomposition and support vector regression (SVR). First, a combined method is applied to the raw data and remove noise to get useful information. Then, a SVR model is applied to improve the prediction performance of multidimensional nonlinear data. Furthermore, the comparison experiments were performed with both Shanghai Composite Index and Dow Jones Index to examine the effectiveness of the framework. The results indicate that the proposed framework performs better than other advanced models.

Keywords: Financial time series; Stock price prediction; Wavelet coherence; Support vector regression.

1. Introduction

Investments in stocks and other financial products have gradually become an important part of our daily lives. Because of the extremely attractive value and great theoretical significance, the prediction of stock price has attracted the attention of academia and investors. Due to the environmental noise inherited in the financial market, the stock price series is non-stationary and nonlinear as well as hard to analyze and predict. Besides, financial market is generally affected by a country’s economic policies, political events, and transactions, which makes the prediction of financial time series an even more challenging task.

There are two types of approaches generally employed in the prediction of financial time series: one is the analysis based on certain economic factors; the other is the use of technical methods to analyze historical data to predict the stock prices trend. In other words, the prediction of stock price can be regarded as a problem of pattern recognition. With the rapid development of signal processing method, various artificial intelligence algorithm is used in stock prices prediction. Among these tools, neural networks and machine learning methods have attracted increasing attention with their excellent characteristics of non-linear regression [1].

The first neural network used in stock price prediction was a Feed-forward neural network (FNN) model. Since then, a variety of neural networks have been used to analyze the trend of financial prices. Such as back-propagation (BP) neural network, spiking neural networks, convolutional neural network (CNN)
and the support vector regression (SVR) model. Among these methods, SVR has attracted most attention in the stock price prediction due to its outstanding characteristics of non-linear regression performance. The inherent noisy environment and high volatility of stock prices severely affects the prediction performance. Some researchers have made improvements in these areas. For example, Chang proposed an adaptive AR model based on the fuzzy network system. They used AR model to analyze the volatility of Taiwan's Weighted Index, and then used the results to optimize and refine an ANFIS-based model. Experimental results proved that the model was well-functioned. Essentially, however, this methodology did not preprocess the raw data which would affect the prediction results. The principal component analysis (PCA) method was well-functioned for data preprocessing. However, it is a linear analysis method which requires data to be in line with normal distribution. The stock price data are often neither linear nor normal distribution, which makes the PCA method is not suitable to process stock price series. Then an excellent data preprocessing method called Empirical Mode Decomposition (EMD) has been successfully adopted to process nonlinear signals.

Based on previous studies, we propose a combined method using wavelet coherence and EMD (WTC-EMD) for stock data preprocessing and an SVR method to predict the price of stock market with preprocessed data. The experimental results with both Shanghai Composite Index and Dow Jones Index show that the proposed framework performs better than other methods.

In this paper, we focus on the use of advanced data processing methods to remove noise from raw data and select effective feature indicators for stock price prediction. We also discuss the importance of data preprocessing in the analysis and prediction of stock market price. The structure of this manuscript is as follows: the second part describes techniques and the proposed predicting framework; the third part describes the simulation using data from Shanghai Composite Index and Dow Jones index and conducts a comparative experiment; the fourth part is the experimental results and discussion, analyzes the performance of the model; the fifth part provides a summary and outlook of future research, discusses the advantages of the WTC-EMD method combined with SVR and its application in financial time series.

2. Methods

2.1 Wavelet Coherence

A wavelet $\psi_{u,s}(t)$, a real-valued square-integrable function, is defined as:

$$\psi_{u,s}(t) = \frac{\psi\left(\frac{t-u}{s}\right)}{\sqrt{s}}$$

(1)

where $u$ is the location parameter, $s$ presents the scale dilatation parameter. A continuous wavelet transform $W_x(u,s)$ can be obtained via a projection of a wavelet $\psi(.)$. Then, $W_x(u,s)$ is defined as:

$$W_x(u,s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x(t) \psi^*\left(\frac{t-u}{s}\right) dt$$

(2)

where $\psi^*(.)$ is a complex conjugate of $\psi(.)$.

The continuous wavelet method can be generalized for a bivariate case to reveal the linkage between series in both time and across scales. A continuous wavelet transform can be generalized into cross wavelet transform as follow:

$$W_{xy}(u,s) = W_x(u,s) \times W^*_y(u,s)$$

(3)

where $W_x(u,s)$ and $W_y(u,s)$ represent the continuous wavelet transforms of $x(t)$ and $y(t)$, respectively. The * denotes the complex conjugate. The wavelet coherence appears as a very useful tool for capturing the co-movement between two selected time series. According to Torrence and Webster (1999) [2], the wavelet coherence is defined as:

$$R^2_{xy}(u,s) = \frac{|S(s^{-1}W_{xy}(u,s))|^2}{S(s)[W_x(u,s)]^2 S(s)[W_y(u,s)]^2}$$

(4)
where $S$ denotes the smoothing operator. The squared wavelet coherence ranges between 0 and 1, which can be interpreted as a squared correlation in both time and frequency domain.

### 2.2 Empirical Mode Decomposition

Huang et al. proposed a method for signal analysis in 1998, named Hilbert-Huang transforms (HHT). HHT method consists of two parts, EMD and spectral analysis. The EMD method differs from other signal processing methods because the basic functions of EMD are generated adaptively depending on the signal while most signal processing methods require pre-set basis functions. Any complex data sets can usually be decomposed into a few intrinsic mode functions. By applying EMD, the original signal can be decomposed into the following form:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t)$$

where $c_i(t)$ is a set of IMFs. $r_n(t)$ represents residual component after decomposition, namely general trend of the signal. $r_n(t)$ is usually a constant sequence or monotonous sequence. According to theoretical function, the basis function of EMD is different from that of Fourier and wavelet decomposition, which is not pre-determined but adaptively generated from decomposition process. Thus, this method is more suitable for non-periodic and non-stationary data.

In this manuscript, the EMD was used to design a filter to remove noise and redundant information from original data. According to Yang’s research, EMD can be used to design high-pass filter, band-pass filter and low-pass filter. It is especially effective for nonlinear data.

### 2.3 SVR

Support Vector Machine (SVM) method was firstly proposed by Cortes and Vapnik in 1995. It has many unique advantages for classification and prediction when solving the small sample, nonlinear and high dimension problems. SVM is widely used in classification problem (SVC) and Regression problem (SVR). In this research, a SVR model was used to predict the stock price.

For a nonlinear SVR, sample $x$ can be mapped to a high-dimensional feature space $H$, and then the optimal regression function can be solved in space $H$. Specifically, the transformation from the input space $R^n$ to a high-dimensional Hilbert space $H$ is introduced:

$$\{X \in R^n \rightarrow (x \rightarrow \phi(x)) : X \in H\}$$

Through this transformation, the training set $\tilde{T}$ in high dimensional space $H$ can be obtained from training set $T$ in space $R^n$. A nonlinear regression problem in low dimensional space will become linear regression problem in high dimensional space. The optimization problem can be shown as follow:

$$\max \sum_{i=1}^{l} [\alpha_i^* (y_i - \varepsilon) - \alpha_i (y_i + \varepsilon)] - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j)$$

s. t. $0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, 2, ..., l$

The normal vector and regression function are:

$$\omega = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) \phi(x)$$

$$f(x) = \omega \ast K(x_i, x_j)$$

The kernel function is very important to SVR algorithm. The RBF kernel is frequently used because it has fewer parameters and applicable in many situations no matter how small or high dimensions samples are. The RBF kernel can be defined as:

$$K(x_k, x_j) = \exp \left(-\frac{||x_k - x_j||^2}{\delta^2}\right)$$

where $\delta$ represents the free parameter.
2.4 The Proposed Prediction Framework

In this research, we establish a financial index prediction model using WTC-EMD and SVR methods. The model consists of two parts: data processing using WTC-EMD and the SVR model for prediction. The input data are the indicators of the daily price of stock index (including open price, high price, low price, close price and volume). The output is the open price of the next day. The steps of the proposed prediction model are as follows:

**Step 1: Data preparation.**
Collect raw data and construct a data matrix. Each column contains indicators of the stock price of a day and the rows represent the trading day.

**Step 2: Feature extraction.**
Perform wavelet coherence method to find the regions in time-frequency space where the financial time series (high, low, close price and volume) co-vary with open price.
According to the results of wavelet coherence analysis, design corresponding filter based on EMD method to remove noise and select indicators.

**Step 3: Prediction with SVR algorithm.**
Build the SVR prediction model. SVR is trained with de-noised data. Then the trained SVR can be used for prediction with the testing data.

**Step 4: Calculate the metrics.**
Several metrics are calculated based on the output results.

3. Empirical Study

3.1 Data

There are three research data sets used in this research. As shown in Table 1, data set 1 is Shanghai Composite Index which is used to verify the validity of the proposed method. The samples include 3627 trading days from January 4th, 2000 to December 1st, 2014. Each sample consists of daily information. The overall data are split into two sections: January 4th, 2000 to December 31st, 2011 and January 4th, 2012 to December 31st, 2014.

In addition, we also conduct a comparative study to demonstrate the effectiveness of this method to an advanced prediction model [3] where data set 2 and data set 3 are collected.

| Data set          | Category     | Sample              | Data number |
|-------------------|--------------|---------------------|-------------|
| Shanghai Composite Index (Data set 1) | Training     | 04/01/2000-31/12/2011 | 2900        |
|                   | Testing      | 04/01/2012-31/12/2014 | 727         |
| Shanghai Composite Index (Data set 2) | Training     | 04/01/2000-31/12/2004 | 721         |
|                   | Testing      | 04/01/2005-31/12/2005 | 242         |
| Dow Jones Index (Data set 3) | Training     | 02/01/2003-31/12/2004 | 507         |
|                   | Testing      | 01/01/2005-31/12/2005 | 252         |
3.2 Simulation Results and Discussion
The data set 1 shown in Table 1 is used to verify the validity of the methodology proposed in this study. Five indicators were contained in data set 1, namely open price, high price, low price, close price and trading volume. The overall data set 1 is split into two sections: January, 2000 to December, 2011 and January, 2012 to December, 2014. The former data set including 2900 daily samples is used as the training set while the latter including 727 daily samples is used as testing set.

The WTC method is adopted on training data set to find regions in time-frequency space where two indexes co-movement. Figure 1 shows the results of WTC analysis. In Figure 1, there are four wavelet coherence results where red represents a high degree of correlation and blue represents a low degree of correlation. It is obvious that the blue part appears in a high frequency of wavelet coherence due to WTC: open-high, the same results for open-low and open-close. It means that the noise mainly exists in high frequency portion of data. Due to the WTC results of open-volume, we can find that the noise distributes in all frequency range. Therefore, feature extraction should include removing high frequency noise existing in data (open, high, low and close price) and minor impact factor (volume).

According to the results of wavelet coherence, a low-pass filter is designed based on EMD method. Perform EMD method on raw data, high price for example, to obtain several IMF components and a residual.

The correlation between each IMF and original data are calculated to obtain correlation coefficient \( \mu_i \) (\( i = 1, 2, \ldots, n \)). Remove the IMF where \( \mu \) value is under a threshold, then reconstruct the signal with the remaining IMF components. The threshold is set as:

\[
\lambda = \frac{\max(\mu_i)}{\kappa}, (i = 1, 2, \ldots, n)
\]

where \( \kappa \) is a proportionality factor, experience value is 10. \( \lambda \) obtained in this experiment is 0.09.

The correlation coefficients less than 0.09 are IMF1 and IMF2. After removing these two components, reconstruct the signal as:

\[
y(t) = \sum_{i=3}^{n} \text{IMF}_i(t) + r(t)
\]

The SVR prediction model consists of two parts: training and testing. The reconstructed signal will be used for training the SVR model. After that, testing data will be used for SVR prediction. Three tests were conducted to verify the proposed framework. Figure 3 shows the prediction results and Table 2 are quantized prediction results.
Figure 2. The actual and predicted values of Shanghai Composite Index.

NOTES: “(5)” represents that input indicators contain trading volume.
“(4)” represents that input indicators don’t contain trading volume.

Table 2. Index on data set Shanghai Composite Index.

| Prediction Model       | MAPE  | RMSE  | MAE   | MSE      | $r$     |
|------------------------|-------|-------|-------|----------|---------|
| SVR (5)                | 0.015 | 53.059| 33.916| 2815.257 | 0.954   |
| WTC-EMD-SVR (5)        | 0.008 | 42.920| 20.242| 1842.126 | 0.971   |
| WTC-EMD-SVR (4)        | 0.007 | 19.328| 14.414| 373.572  | 0.991   |

Figure 2 and Table 2 clearly show that the proposed feature extraction method especially de-noising method based on WTC-EMD can improve forecast accuracy, furthermore, indicator selection (removing the volume variable) also helps to improve forecasting accuracy. The WTC-EMD-SVR (4) prediction framework get better results on the metrics of MAPE, RMSE, MAE, MSE and $r$ then SVR prediction framework without feature extraction.

To further test the effectiveness of the above-mentioned method, an advanced hybrid model was also performed to test the efficacy of the framework. The data set 2 and data set 3 used in the comparison study were shown in Table 1. The prediction results of comparison study are shown in Figure 3, Table 3 and Table 4. Experimental results show that the method proposed in this research reduced the metrics of MAPE, RMSE, MAE, and MSE. Meanwhile increased the metric of $r$.

Table 3. Results on data set 2 of Shanghai Composite Index.

| Prediction Model       | ICA-CCA-SVR | WTC-EMD-SVR |
|------------------------|-------------|-------------|
| MAPE                   | 0.011       | 0.005       |
| RMSE                   | 16.540      | 7.716       |
| MAE                    | 12.638      | 5.781       |
| MSE                    | 273.572     | 59.537      |
| $r$                    | 0.952       | 0.991       |
Table 4. Results on data set 3 of Dow Jones Index.

| Prediction Model | ICA-CCA-SVR | WTC-EMD-SVR |
|------------------|-------------|-------------|
| MAPE             | 0.011       | 0.005       |
| RMSE             | 16.54       | 7.716       |
| MAE              | 12.638      | 5.781       |
| MSE              | 273.572     | 59.537      |
| $r$              | 0.952       | 0.991       |

Figure 3. Comparison study on the data of DJI and Shanghai Composite Index.

4. Conclusions
A novel stock time series forecasting framework is proposed in this research. The framework combines WTC and EMD method with SVR algorithm to forecast the trend of stock index. We carry out empirical analysis research on Shanghai Composite index and selects five indicators of daily trading as feature vector. We use the hybrid WTC-EMD method to extract effective features on non-stationary financial data and processed data for SVR training as well as the SVR model for prediction. Feature extraction is a very important step in prediction models while raw data preprocessing is not being paid enough attention by many models. We proposed a WTC-EMD approach in order to remove the redundant information and noise existing in raw stock market data. Comparative results show that the forecasting framework proposed in this research obtains better performance in extracting valid stock market information.

Acknowledgements
This study was supported by the Fundamental Research Funding for the Beijing Universities (No. XRZ2020017).

References
[1] Y.J. Wang, H. Liu, Q. Guo, et al., “Stock volatility prediction by hybrid neural network,” IEEE Access vol. 7, 154524-154534, 2019.
[2] C. Torrence, and P.J. Webster, “Interdecadal changes in the ENSO-Monsoon system,” Journal of Climate vol. 12, no. 8, 2679-2690, 1999.
[3] Z.Q. Guo, H.Q. Wang, Q. Liu. et al., “A feature fusion based forecasting model for financial time series,” PLoS ONE vol. 9, no. 6, article ID e101113, 2014.