Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 625 – 631

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Stock Price Prediction Based on SSA and SVM

WEN Fenghua[a*], XIAO Jihong[b], HE Zhifang[a], GONG Xu[a]

a. Business School of Central South University, Changsha, 410081, China
b. School of Economics and Management, Changsha University of Science and Technology, Changsha, 410004 ,China

Abstract

This paper, using the singular spectrum analysis (SSA), decomposes the stock price into terms of the trend, the market fluctuation, and the noise with different economic features over different time horizons, and then introduce these features into the support vector machine (SVM) to make price predictions. The empirical evidence shows that, compared with the SVM without these price features, the combination predictive methods——the EEMD-SVM and the SSA-SVM, which combine the price features into the SVMs perform better, with the best prediction to the SSA-SVM.

Keywords: Stock Price Series; Singular Spectrum Analysis; Support Vector Machine (SVM); Combination Predictive Methods

1. Introduction

Recently, the support vector machine[2] has been widely used in stock price predictions. However, as the stock market is affected by economic, political, financial, social factors and noises, stock prices may have different features over different time horizons. But few studies have introduced the price features into the SVM to make price predictions. Zhang, et al[4] used the ensemble empirical mode decomposition (EEMD) to analyze fundamental features of petroleum price series over different time horizons and pointed out that the decomposed terms can be introduced into the SVM to make predictions. But the EEMD has some limitations in the analysis of stock price series. The EEMD can not effectively extract noise from the price prediction, but the impact of noise is prevalent in the stock market. Therefore, the EEMD can not catpure this feature well.

The SSA, a method for analyzing non-linear, non-stationary time series, was first proposed by Broomhead and King[7]. The SSA is to get a series of singular values which contains the information of the original series through the singular spectral decomposition (SVD). By analyzing singular values of different information, we

*Corresponding author: wfh@amss.ac.cn.

doi:10.1016/j.procs.2014.05.309

can derive time series with different features, and they are often used for extracting noise in time series. Beneki, et al[8] extracted the trend term and economic fluctuation term, and put them into the analyses of tourist revenues series in Britain; Chinese scholar Lian Jijian, et al[9] used order determination and noise reduction based on singular entropy to reduce noise in series. In this paper, we also use the SSA to decompose the stock price.

As mentioned above, to make predictions more accurate, we use the SSA to decompose stock price series into the terms of trend, market fluctuation and noise, and then introduce these terms into the SVM to make price predictions.

2. Introduction to Predictive Methods

2.1 SSA

The core idea of the SSA is to obtain a series of singular values which contain the information of the original series through the singular spectral decomposition (SVD), and then select different singular values to construct series with different constituents. Specific steps are as follows:

Given a time series $X_N = \{x_1, x_2, ...x_N\}$, N is the effective length. Reconstruct the series' phase space, and then we can derive the trajectory matrix:

$$D_m = \begin{bmatrix} x_1, x_2 ..., x_n \\ x_2, x_3 ..., x_{n+1} \\ ..., ..., ..., ... \\ x_m, x_{m+1} ..., x_{n+m-1} \end{bmatrix} \tag{1}$$

where n is the length of the window, m is the embedding dimension, and $2 \le m \le N/2$, $m \le n$,. $N = n + m - 1$.

Take matrix D_m for its singular value decomposition (SVD), $D_m = USV^T$, where U and V are $m \times m$ *and* $n \times n$ matrixes, s is $m \times n$ diagonal matrix, with the diagonal components $\lambda_1, \lambda_2, ...\lambda_p$, and $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_p$. $\lambda_1, \lambda_2, ...\lambda_p$ are the singular values of matrix D_m , and p is the number of the order of the singular value. U and V are the right singular matrix and the left singular matrix. If the time series only contain effective information, then the rank of the matrix s is $k < m$; if the time series contain both effective information and noise, then $k = m$.

The concept of singular entropy[9] is introduced in order to study the law of the amount of information in time series changing with the number of the order of the singular spectrum:

$$E_K = \sum_{i}^{k} \Delta E\left(K < P\right) \tag{2}$$

Where k is the order of singular entropy, ΔE_i represents an increment of singular entropy in order i, by the following formula we can obtain that:

$$\Delta E = -\left(\lambda_i / \sum_{j=1}^{p} \lambda_j \right) \log\left(\lambda_i / \sum_{j=1}^{p} \lambda_j \right) \tag{3}$$

When the increment of the singular entropy falls to the asymptotic value, the effective information of the time series saturates, that is, the information is essentially complete, later increments of the singular entropy are caused by noise, and at this point the number of the order *i* is therefore selected as the one for noise reduction.

2.2 Support Vector Machine

In 1998, Vapnik proposed the ε-support vector machine to solve problems in predictions. Here we mainly introduce the ε - SVM.

Given a set of data points, $\{(x_1, y_1), \ldots, (x_i, y_i)\}$, such that $x_i \in R$ is an input and $y_i \in R$ is a target output, the standard form of support vector regression is:

$$\min_{\omega, b, \xi, \xi^*} \frac{\omega^T \omega}{2} + c\sum_{i=1}^{n} \xi_i + c\sum_{i=1}^{n} \xi^*$$

$$s.t: \omega^T \phi(x_i) + b - y_i \le \varepsilon + \xi_i \tag{4}$$

$$\omega^T \phi(x_i) + b - y_i \ge \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0, i = 1, 2 \ldots n$$

To solve the above problem, we need to introduce a Lagrange function, and according to the Duality and the saddle point condition, we can get the dual form :

$$\min_{\partial, \partial^*} \frac{1}{2}(\partial - \partial^*)^T Q(\partial - \partial^*) + \varepsilon \sum_{i=1}^{n} (\partial_i + \partial_i^*) + \sum_{i=1}^{n} y_i (\partial_i - \partial_i^*) \tag{5}$$

$$s.t{:}\sum_{i=1}^{n} (\partial_i - \partial_i^*) = 0, 0 \le \partial_i, \partial_i^* \le c, i = 1, 2 \ldots n$$

$$\omega = \sum_{i=1}^{n} (\partial_i^* - \partial_i) \tag{6}$$

$Q_{ij} = k(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$, $k(x_i, x_j)$ are kernel functions. As long as a function meets the Mercer requirement, it can be used as a kernel function. The main kernel functions are:

(a) Linear kernel $k(x, x_i) = x^T x_i$;

(b) Polymonial kernel $k\left(x,x_i\right)=\left(\gamma x^T x_i+r\right)^p$;
(c) RBF kernel $k\left(x,x_i\right)=\exp\left(-\gamma\left\|x-x_i\right\|^2\right)$;
(d) Sigmoid kernel $k\left(x,x_i\right)=\tanh\left(\gamma x^T x_i+r\right)$.
From the above analysis, we can obtain the optimal classifier:

$$f\left(x\right)=\sum_{i=1}^{n}\left(\partial_i^*-\partial_i\right)k\left(x_i,x\right)+b \tag{7}$$

3. Empirical Research

3.1 Research Steps

1. SSA decomposition: first select a suitable number for the embedding dimension to reconstruct the phase space for the original series to derive the trajectory matrix, and then use the SVD to decompose the trajectory matrix to obtain singular values, determine the number of the order according to the property of the increment for the singular entropy, and finally extract the singular values to reconstruct the trend term, the market fluctuation term, and the noise term;
2. Selection of suitable kernel functions: select with the help of evaluation indexes a kernel function for each decomposed trend term, market fluctuation term, and noise term, and perform SVM predictions for each term using a suitable kernel function to get predictive values for the terms;
3. SVM combination prediction: use the predictive value of each term as the input, and the real price at that time as the output to construct a training model; use the training model to make a prediction; when the input variable is the predictive value for each component, the output is the final predictive value.

3.2 Experiment Data and Evaluation Index

We select the closing price of the Shanghai Stock Exchange (SSE) Composite Index from Jan. 5, 2009 to Sept. 30, 2013 as the empirical research subject, which offers 1150 data points (the data supplied by the Guotai Jun'an CSMAR research databank). We use two-thirds of the research data points (the first 767 closing price) as the training data. In the empirical research, the predictive value is derived one step ahead by moving the window. The evaluation indexes are the mean squared error (MSE), the mean absolute prediction error (MAPE), the directional symmetry (DS), and the correlation coefficient R . The less the values for the MSE and the MAPE are, the better the predictions are; for the DS and R, the greater the value is, the more accurate the prediction is.

3.3 Empirical Study Process

Let the original series be $x\left(t\right)$, and then take $x\left(t\right)$ for phase space reconstruction. Practically, one-third or one-fourth of the length of the original series are selected to obtain all the information from the original series. This paper selects one-third the length of the original series (approximately 400) as the embedding dimension, decomposes the derived trajectory matrix into the singular value matrix through the SVD, and then constructs the figure of the increment of the singular value, as shown in Fig 1.

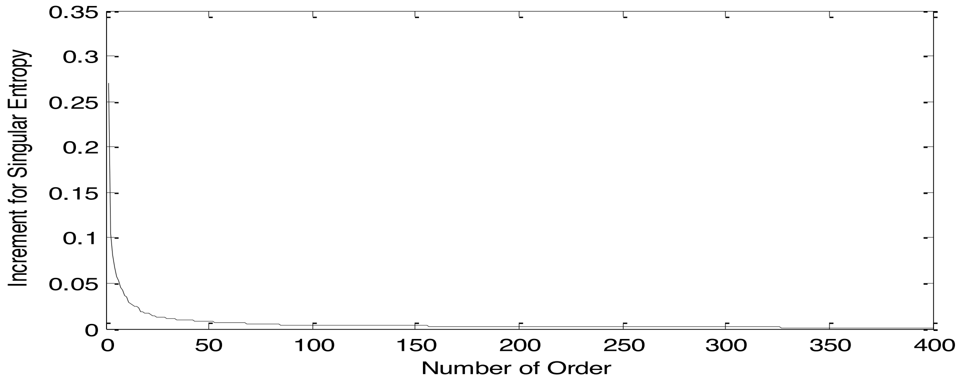

Fig 1. the Increment of the Singular Entropy

As seen in Graph 1, when the number of the order reaches 50, the increment changes little and the effective information saturates. Reconstruct the singular values beyond the number of the order 50 as the noise term $x_3(t)$, and thus the first singular value contains the most fundamental information of the series, use it to reconstruct the trend term $x_1(t)$ and use singular value 2-50 to reconstruct the market fluctuation term $x_2(t)$. Then make a graphical comparison of $x(t)$ $x_1(t)$ $x_2(t)$ and $x_3(t)$, see Fig 2.

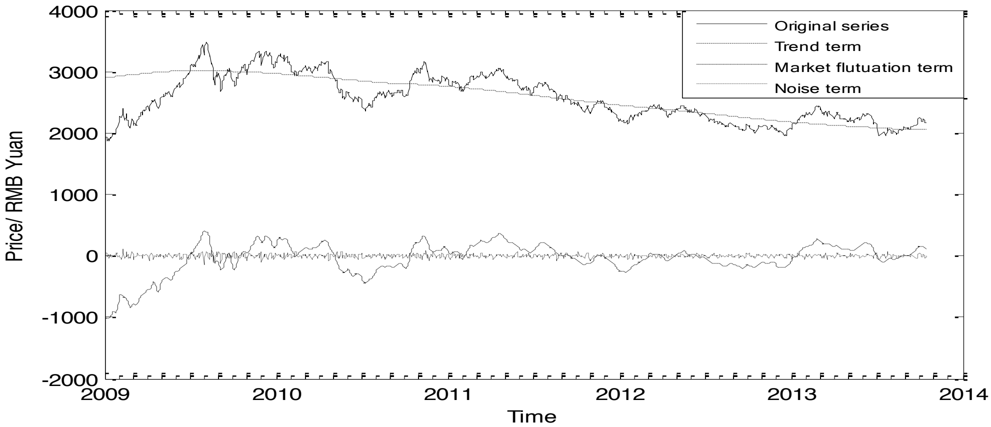

Fig 2. Original Series and Decomposed Terms

Then we get the final predictive value according to the previous steps. The final comparison of the real and predictive values is shown in Fig 3.

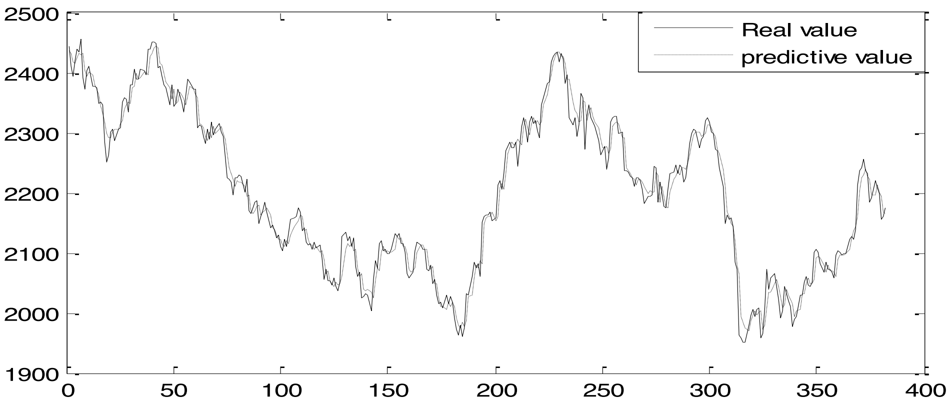

Fig 3. Final Comparison of the Real and Predictive Values

Fig 3 shows that the final predictive value provides a good fitting to the trend, and the change of the wave shapes of the actual price, and thus making a good predictive result.

With the same sample data, the prediction made only with the SVM, the EEMD-SVM combination prediction and the prediction made with our method (SSA-SVM) are compared with their results tabulated in Table 1.

Table 1. Prediction Error

	SVM	EEMD-SVM	SSA-SVM
MSE	0.002315	0.001972	0.001785
MAPE	0.008138	0.007696	0.007245
DS	52.493%	62.729%	67.979%
R	96.542%	97.056%	97.343%

From Table 1, we can see that making combination predictions by decomposing the original index series into series with economic implications is more desirable than making mere SVM predictions, and SSA-SVM combination predictions are better than EEMD-SVM combination predictions, which indicates that SSA decomposition of the original index series can better grasp the features of the original index series, and thus get better predictive results.

4. Conclusion

Based on the SSA and the research subject selected from the closing price of the Shanghai Stock Exchange (SSE) Composite Index from Jan. 5, 2009 to Sept. 30, 2013, this paper decomposes the original stock index into the trend terms, the market fluctuation terms, the noise terms and time series with different economic features. Then we introduce the index features into the SVM method for prediction. the final predictive value provides a good fitting to the trend, and the change of the wave shapes of the actual price, and thus making a good predictive result. Finally we compare the predictive effect of the SSA-SVM combination prediction with the SVM prediction and the EEMD-SVM combination prediction, and find that the combination prediction which combines the decomposition of original index into series with certain economic implications to the SVM is more effective than the SVM prediction; and the SSA can better grasp the features of the original index series than the EEMD, while the SSA-SVM combination prediction have better predictive effect than that of the EEMD-SVM combination prediction.

References

1. Ince H, Trafalis T. Short term forecasting with support vector machines and application to stock price prediction. *International Journal of General Systems*; 2008, 37(6), p. 677-687.
2. Vapnik V. Statistical Learning Theory. Wiley, New York, NY; 1998.
3. Sai Ying, Zhang Fengting, Zhang Tao. China's stock index futures regression prediction research based on SVM. *China Journal of Management Science*; 2013, 03, p. 35-39.
4. Zhang Xuegong. About the statistical learning theory and support vector machine. *ACTA AUTOMATICA SINICA*; 2000, 01, p. 36-46.
5. Zhang X, Lai k k, Wang S Y. A New Approach for Crude Oil Price Analysis Based on Empirical Mode Decomposition. *Energy Economics*; 2008, 30(3), p. 905-918.
6. Yang Yunfei, Bao Yukun, Hu Zhongyi, Zhang Rui. Rrude oil price forecasting method based on EMD and SVMS. *Journal of Management*; 2010, 12, p. 1884-1889.
7. Broomhesd, D. S, King, G. P. Extracting Qualitative Dynamics from Experimental Data. *Physica D:Nonlinear Phenomena*; 1986, 20(2-3), p. 217-236.
8. Beneki,C.,Eeckels,B.,and Leon,L.,Signal Extraction and Forecasting of the UK Tourism Income Time Series: A Singular Spectrum Analysis Approach. *Journal of Forecasting*; 2012, 31(5), p. 351-400.
9. Lian Jijian, Li Huokun, Zhang Jianwei. Hydranlic structure vibration modal ERA identification method based on the singular entropy order noise reduction. *SCI CHINA SER*; 2008, 09, p. 1398-1413.