Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2014 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-8-2014

Two-Stage Model for Exchange Rate Forecasting by EMD and Random Forest

Han-Chou Lin

Heng-Li Yang

TWO-STAGE MODEL FOR EXCHANGE RATE FORECASTING BY EMD AND RANDOM FOREST

Han-Chou Lin, National Chengchi University, Taiwan, actuallife@hotmail.com
Heng-Li Yang, National Chengchi University, Taiwan, yanh@nccu.edu.tw

ABSTRACT

This study applied random forest (RF) and empirical mode decomposition (EMD) techniques to exchange rate forecasting. The aim of this study is to examine the feasibility of the proposed EMD-RF model in exchange rate forecasting. For this purpose, the original exchange rate series were first decomposed into a finite, and often small, number of intrinsic mode functions (IMFs) and one residual component. Then, a random forest model is constructed to forecast these IMFs and residual value individually, and then all these forecasted values are aggregated to produce the final forecasted value for exchange rates. The daily USD/NTD, USD/JPY, USD/HKD and USD/AUD exchange rates were employed as the data set. The experimental results are that MAPE for the four data sets are, respectively, 0.278%, 1.143%, 0.153% and 5.944%, which shows good performance according to the 10% threshold suggested by Lewis.

Keywords: Random forest, empirical mode decomposition, Hilbert–Huang Transform, intrinsic mode functions, exchange rate forecasting.

INTRODUCTION

Financial time series forecasting is one of the most important and challenging tasks due to its inherent nonlinearity and nonstationary characteristics. In the past decades, this issue has attracted increasing attention by many academic researchers and some forecasting approaches have been developed for dealing with such data. The most popular approach is artificial neural network (ANN) due to their excellent nonlinear modeling capability [2-5]. However ANN models have several limitations, such as local optimization, slow speed of training, and low efficiency, thus have a poor generalization ability. Constructing ANN model would contain a great many parameters. These parameters were always judged by experience, so the model was hard to be established [6].

Recently, there has been a lot of interests in “ensemble learning methods” that generate many classifiers and aggregate their results. Two well-known methods are boosting [7] and bagging [8] of classification trees. In boosting, successive trees give extra weights to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees — each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction. Breiman proposed random forests (RF), which add an additional layer of randomness to bagging [9]. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. This somewhat counterintuitive strategy turns out to perform very well [9]. Diana et al. demonstrated that RF can exhibit better accuracy of prediction and outperformed ANN [10]. In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest).

On the other hand, for better performance, hybrid models have been used. For example, Wang, Yu and Lai developed a hybrid artificial intelligent (AI) system framework by means of a systematic integration of ANN and rule-based expert system, with web text mining [11]. Amin-Naseri and Gharacheh proposed a hybrid AI approach integrating feed-forward neural networks, genetic algorithm, and k-means clustering and obtain satisfactory results [12]. The basic idea of the above hybrid methods is to overcome the drawbacks of individual models and to generate a synergetic effect in forecasting.

Motivated by hybrid methodologies, this study attempts to apply the “divide-and-conquer” principle to construct a novel exchange rate forecasting methodology. In this study, the generic idea of “divide-and-conquer” principle can be understood as “decomposition-and-ensemble”. The main aim of decomposition is to simplify the forecasting task, while the goal of ensemble is to formulate a consensus forecasting on original data. In terms of the above ideas, an empirical mode decomposition (EMD) based random forest ensemble learning is proposed for tourism demand forecasting. EMD based on Hilbert–Huang Transform (HHT), a new technique in dealing with noise and nonlinearity data, will be applied to decompose tourism demand series into a finite and often small number of intrinsic mode functions (IMF). An IMF is defined here as any function having the same number of zero-crossing and extrema, and also having symmetric envelopes defined by the local maxima, and minima respectively. The IMF also thus admits well-behaved Hilbert transforms. The main advantage of selecting EMD as a decomposition tool is very suitable for decomposing nonlinear and nonstationary time series; it has been reported to have worked better, in describing the local time scale instantaneous frequencies, than the wavelet decomposition and Fourier decomposition[13]. In wavelet decomposition, it needs to determine a filter function before decomposition, while in Fourier decomposition, a time series (either linear or nonlinear) can be decomposed into a set of linear components. As the degree of nonlinearity and nonstationarity in a time series increases, the Fourier decomposition often produces large sets of physically meaningless harmonics, when it is applied to nonlinear time series decomposition [13].

THE PROPOSED MODEL

The proposed hybrid approach for exchange rate forecasting, namely, EMD-RF, combines EMD and RF, are described as follows.

Step 1. The original time series $x(t)$, $t = 1, 2,\ldots,N$ is decomposed into n IMF components, $c_j(t)$, $j = 1, 2, \ldots, n$, and one residual component $r_n(t)$ by using EMD.

Step 2. The decomposed datasets are divided into two datasets, namely training datasets and testing datasets. The training datasets are used to build RF models.

Step 3. Forecasting model is set up for each IMFs, and the residue by using RF, respectively. The final prediction results are obtained by compositing the prediction values. In the following F is RF predictor function. The final forecasting result is $\sum_{j=1}^{n-1} F_j(c_j(t)) + F_n(r_n(t))$

The flow chart of the proposed EMD-RF model is shown in Figure. 1.

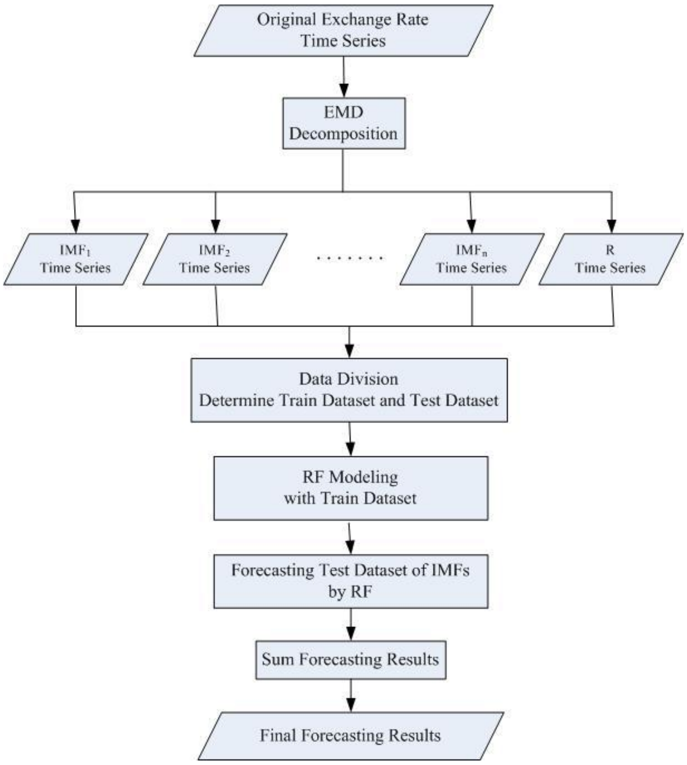

Figure 1 Flow Char of the proposed EMD-RF predictor model

EXPERIMENTAL RESULTS

Daily values of exchange rates USD/NTD, USD/JPY, USD/HKD and USD/AUD extracted from Datastream provided by Morgan Stanley Capital International (MSCI) were used in this study. The whole data set covers the period from January 1, 2007 to December 31, 2013, a total of 2557 observations. The data set is for the period of January, 2001–December, 2010, for a total 2352 observations. In addition, the data are divided into training and testing sets with the ratio 9:1.

To measure the forecasting performance, one criterion is used for evaluation of level prediction forecasting. We select the mean absolute percentage error (MAPE) as the evaluation criterion of level prediction. Typically, the MAPE can be defined by

$$\text{MAPE} = \frac{\sum \left| \frac{x_t - \hat{x}_t}{x_t} \right|}{N} \times 100 \tag{1}$$

where $x(t)$ is the actual value, $x^{\wedge}(t)$ is the predicted value, and N is the number of predictions. The criterion measures the correctness of a prediction in terms of levels and the deviation between the actual and predicted values. The smaller the values, the closer the predicted time-series values will be to the actual values.

The experimental results are that MAPE for the four data sets are, respectively, 0.278%, 1.143%, 0.153% and 5.944%, which shows good performance according to the 10% threshold suggested by Lewis [1]. Figure 2 provides a graphical presentation of forecasting value for the four data sets as following.

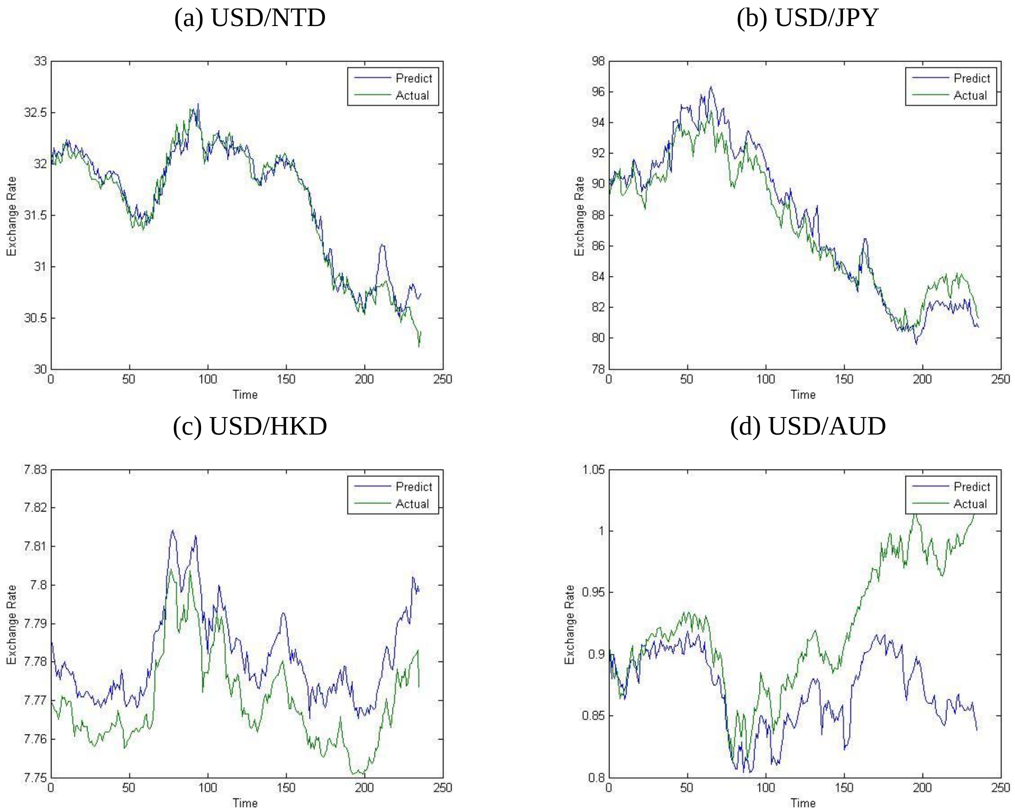

Figure 19 Forecasting value for the four exchange rates

REFERENCES

[1] Lewis, E.B. (1982) Control of Body Segment Differentiation in Drosophila by the Bithorax Gene Complex. Embryonic Development, Part A: Genetics Aspects. New York.

[2] Chen, A.S., Leung, M.T., & Daouk, H. (2003) ‘Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index’, *Computer Operation Research,* Vol. 30, No. 6, *s*, pp. 901-923.

[3] Zhang Y., & Wu L. (2009) ‘Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network’. *Expert System Application*, Vol. 36, No. 5, pp. 8849-8854, doi:10.1016/j.eswa.2008.11.028.

[4] Chen, C.F., Lai, M.C., & Yeh, C.C. (2012) ‘Forecasting tourism demand based on empirical mode decomposition and neural network’. *Knowledage-Based Systems*, Vol. 26, pp. 281-287, doi:10.1016/j.knosys.2011.09.002.

[5] Jaeger, H., & Haas, H. (2004) ‘Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication’. *Science*, Vol. 304, No. 5667, pp. 78-80.

[6] Lapedes, A., & Farber, R. (1987) Nonlinear signal processing using neural networks: Prediction and system modelling. United States.

[7] Bartlett, P., Freund, Y., Lee, W.S., & Schapire, R.E. (1998) ‘Boosting the margin: a new explanation for the effectiveness of voting methods’. *Annals of Statistics*, Vol. 26, No. 5, pp. 1651-1686.

[8] Breiman, L. (1996) ‘Bagging predictors’. *Machine Learning*, Vol.24, No.2, pp.123-140.

[9] Breiman, L. (2001) ‘Random forests’. *Machine Learning*, Vol. 45, No.1, pp. 5-32.

[10] Diana, P-A, Francisco, E.M., Aneta, K., Juan, A.R., & John, B. (2008) ‘Random forests, a novel approach for discrimination of fish populations using parasites as biological tags.’ *International Journal of Parasitol*, Vol. 38, pp.1425-1434.

[11] Wang, S.Y., Yu, L., & Lai, K.K. (2004) ‘A novel hybrid AI system framework for crude oil price forecasting’. *Lecture Notes in Computer Science*, Vol. 3327, pp. 233-242.

[12] Amin-Naseri, M.R., & Gharacheh, E.A. (2007) 'A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series', *Proceedings of the 10th International Conference on Engineering Applications of Neural Networks*, pp. 160-167.

[13] Huang, W., Shen, Z., Huang, N.E., & Fung, Y.C. (1999) 'Nonlinear indicial response of complex nonstationary oscillations as pulmonary hypertension responding to step hypoxia'. *Proc of the National Academy of Sciences*, Vol. 96, pp.1934-1839.