Association for Information Systems
AIS Electronic Library (AISeL)

MWAIS 2016 Proceedings | Midwest (MWAIS)

Spring 5-19-2016

A Comparative Study of Ensemble-based Forecasting Models for Stock Index Prediction

Dhanya Jothimani
Department of Management Studies, Indian Institute of Technology Delhi India, dhanyajothimani@gmail.com

Ravi Shankar
Department of Management Studies, Indian Institute of Technology Delhi India, ravi1@dms.iitd.ac.in

Surendra S. Yadav
Department of Management Studies, Indian Institute of Technology Delhi India, ssyadav@dms.iitd.ac.in

A Comparative Study of Ensemble-based Forecasting Models for Stock Index Prediction

Dhanya Jothimani
Department of Management Studies,
Indian Institute of Technology Delhi India
dhanyajothimani@gmail.com

Ravi Shankar
Department of Management Studies,
Indian Institute of Technology Delhi India
ravi1@dms.iitd.ac.in

Surendra S. Yadav
Department of Management Studies,
Indian Institute of Technology Delhi India
ssyadav@dms.iitd.ac.in

ABSTRACT

Stock prices as time series are, often, non-linear and non-stationary. This paper presents an ensemble forecasting model that integrates Empirical Mode Decomposition (EMD) and its variation Ensemble Empirical Mode Decomposition (EEMD) with Artificial Neural Network (ANN) for short-term forecasts of stock index. In first stage, the data is decomposed into a smaller set of Intrinsic Mode Functions (IMFs) and residuals using EMD and EEMD. In the next stage, IMFs and residue are taken as the inputs for the ANN model to train and predict the future stock price. The methodology was tested with weekly Nifty data for a period of 8 years. The results suggest that the ensemble forecast model using aggregation of the decomposed series performs better than traditional ANN and Support Vector Regression Models. Further, trading strategies based on EEMD-ANN models yielded better return on investments than Buy-and-Hold strategy.

Keywords

Financial time series prediction, EMD, EEMD, Trading Rules, Nifty, Ensemble forecasting.

INTRODUCTION

Financial series are, often, non-linear and non-stationary. Though there are numerous statistical and soft computing techniques available in the literature for forecasting these series (Atsalakis and Valavanis, 2009; 2013), but it is still considered as a difficult task to predict them.

Popular in machine learning and statistics, ensemble method is based on the idea that use of multiple predictors yields better predictions than any of the base predictors (Opitz and Maclin, 1999). Similar to classification problems, there are two types of ensemble forecasting: competitive ensemble forecasting and cooperative ensemble forecasting. In Competitive forecasting model, different predictors are trained individually with same or different datasets but with different parameters and the results are averaged to obtain the final predictions.

In cooperative ensemble forecasting, the prediction task is divided into various sub-tasks and predictors for each sub-task are chosen. The predicted result is obtained by taking sum of results of all sub-tasks. There are two types of cooperative forecasting: Pre-processing and Post-processing. In pre-processing, dataset is deconstructed into various subset and each subset is predicted using various predictors. Decomposition of time series falls under this category. Post-processing selects the predictors based on the characteristics of data. For instance, AutoRegressive Integrated Moving Average (ARIMA) is used for modelling linear and stationary data; ANN and Support Vector Regression (SVR) are used for modelling non-linear data.

Classical decomposition model works best with the linear time series but it ignores random component and leads to a loss of information, thus, affecting the forecast accuracy. Recently, few signal processing techniques like Discrete Wavelet Transform (DWT) and Empirical Mode Decomposition (EMD) have been used for decomposing the series in time-frequency domain and time domain, respectively.

EMD, proposed by Huang et al. (1998), decomposes a signal into a set of adaptive basis functions called Intrinsic Mode Functions (IMFs). It uses Huang-Hilbert Transform (HHT) to decompose the non-stationary and non-linear time series. Unlike DWT, it does not require the a priori information about the series, i.e., scale of decomposition. Though DWT can handle non stationary data, it still requires linear generating process and suffers from leakage between the scales (Crowley 2010). Despite its advantages, EMD suffers from the limitation of mode-mixing problem. To overcome this limitation, a variation of EMD called Ensemble Empirical Mode Decomposition (EEMD) is used for preprocessing of data.

The paper presents two ensemble forecasting models, namely, hybrid EMD-ANN and EEMD-ANN models to predict 1-step ahead forecasts for weekly Nifty price index, where the time series is first decomposed to different sub-series (IMFs) using EMD and EEMD. Then, these sub-series are predicted independently using ANN and are aggregated to obtain the final forecasts. The hybrid EMD-ANN and EEMD-ANN models integrate the benefits of both decomposition and machine learning models. Hybrid EEMD-ANN model overcomes the limitation of EMD. The scope of the paper is limited to cooperative ensemble forecasting models.

Few trading rules have been illustrated to guide the investors to make investment related decisions. Effectiveness of trading rules using ensemble based forecasting models is compared with traditional Buy-and-Hold strategy.

The contributions of the paper are three-fold. It describes and uses cooperative ensemble forecasting model to predict the stock index. It overcomes the limitations of EMD by using EEMD as a data preprocessing technique. It illustrates the use and advantages of trading strategies based on ensemble forecasting model over Buy-and-Hold strategy.

ENSEMBLE FRAMEWORK

The steps are enumerated below:

1. The original series is decomposed into a set of different sub-series using EMD and EEMD for hybrid EMD-ANN and hybrid EEMD-ANN models, respectively.
2. Each sub-series is forecasted separately using ANN.
3. Forecasted sub-series are recombined to get aggregate forecasting, which is then compared with the original series to calculate the error measures.

Empirical Mode Decomposition (EMD)

The procedure for decomposing the original financial time series $X(t)$ is as follows (Huang et al. 1998):

1. All the local minima are identified and interpolated using cubic spline interpolation method to form a lower envelope $L(t)$. Similarly, all the local maxima are interpolated to form an upper envelope $U(t)$.
2. The mean of lower and upper envelopes is calculated using $M(t) = (L(t) + U(t))/2$, which then subtracted from the original series to obtain a local detail $Z(t)=X(t) - M(t)$.
3. Steps 1 and 2 are repeated on $Z(t)$ until: (a) the value of $M(t)$ approaches zero, and (b) the difference between the number of local extrema and zero crossings is at most 1. This process is known as sifting. The first IMF $IMF_1(t)$ equals $Z(t)$ and $R_1(t)=X(t) - Z(t)$ is the residue.
4. The steps 1-3 are repeated on $R_1(t)$ to obtain the second IMF $IMF_2(t)$ and the second residue $R_2(t)$. The process is repeated on $R_i(t)$ to obtain $IMF_{i+1}(t)$ and $R_{i+1}(t)$ until $R_{i+1}(t)$ does not have more than two local extrema, where $i=1, 2, ..., N-1$.

The original series X(t) is expressed as

$$X(t) = \sum_{i=1}^{N} IMF_i(t) + R_N(t)$$

Ensemble Empirical Mode Decomposition (EEMD)

EMD suffers from the limitation of mode mixing problem. Mode mixing refers to a phenomenon where more than one intrinsic mode frequency contains signals in a similar frequency band or an IMF consists of signals spanning a wide band of frequency. Mode mixing is caused by signal intermittency, which could affect the physical meaning of IMF. To overcome the

limitation of mode mixing problem in EMD; Ensemble Empirical Mode Decomposition (EEMD), an ensemble version of EMD was developed (Wu and Huang 2009).

The steps of EEMD are as follows:

1. A collection of noise-added original time series is created.

   $$X^i(t) = X(t) + \acute{U}(t), \quad i \in 1,...,I$$

   where $\varepsilon(t)$ are independent Gaussian white noise, I is the number of trials.

2. EMD is applied on each $X^i(t)$ to obtain the decomposed IMFs and residue.

   $$X^i(t) = \sum_{j=1}^{N} C_j^i + r_N^i$$

3. Results of all trials are averaged to reconstruct the original time series. Averaging helps to cancel out the uncorrelated white noise and preserving the meaningful original time series.

   $$X(t) = \frac{1}{I}\left(\sum_{i=1}^{I}\sum_{j=1}^{N} C_j^i + r_N^i\right) + \acute{U}_I$$

   where, $\acute{U}_I = \frac{\acute{U}}{\sqrt{I}}$

DATA, PROCESSING AND PREDICTION

Nifty is the benchmark index of Indian stock market. Nifty consists of 50 companies and covers 22 sectors. The original raw data consisted of weekly closing prices of Nifty. The dataset covered a period of 8 years ranging from September 2007 to December 2015.

EMD AND EEMD

The Nifty closing prices were decomposed using EMD and EEMD. A total of seven relatively stationary IMFs were produced along with the residue component using EMD and EEMD (Figure 1), respectively.

ANN

Each IMF and residual component obtained was predicted using ANN to obtain 1-step ahead forecasts. Since ANN is a supervised machine learning technique, first 70% of dataset was used to train the model and rest 30% was used to test the validity of the model. A three-layer resilient feed forward neural network consisting of input layer, hidden layer and output layer was considered. Resilient Back Propagation was used for training the neural network since it helps to achieve superior performance.

The relationship between the series and its past values, which is estimated as lag parameter, is used as the input to the neural network. Auto Correlation Function and Partial Auto Correlation Function are used to determine the lags in each sub-series. The sub-series IMF3 cuts off at lag 4, which means value of *IMF3* at time *t* is dependent on its past 4 values, hence, the number of neurons in the input layer would be four. The data format of the same can be expressed as:

$$X(t) = f\left[X(t-1), X(t-2), X(t-3), X(t-4)\right]$$

The number of neurons in the output layer is one since the forecasted value is to be obtained as output, which is represented as $X(t)$ in the above equation. The forecasted sub-series are aggregated to obtain final forecast.

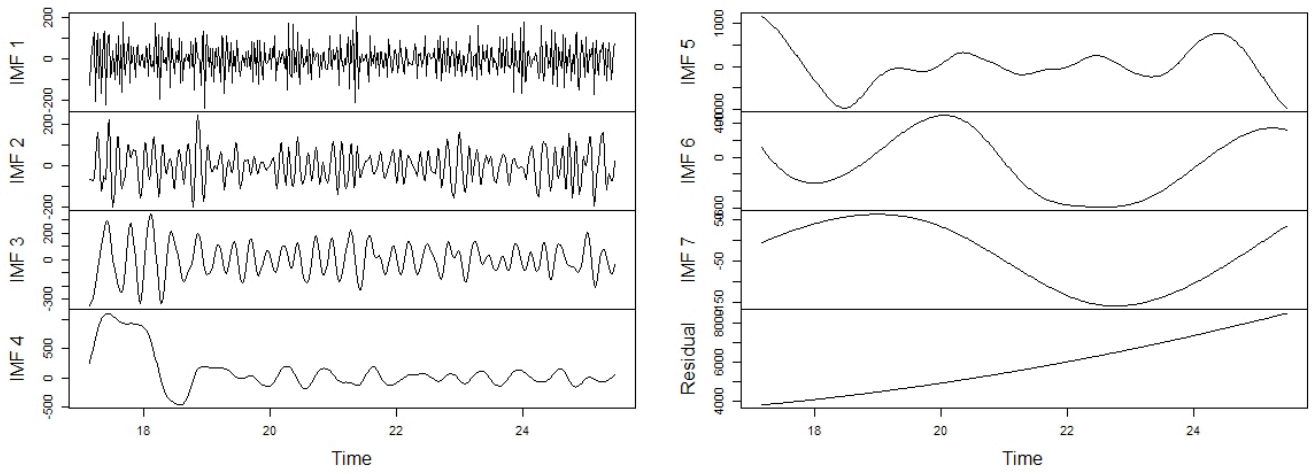

Figure 1. Decomposed Signals Obtained Using EEMD

To check the effectiveness of the ensemble forecast models, 1-step ahead forecasts were also obtained using ANN and SVR.

RESULTS AND DISCUSSION

Error Measures

A comparative analysis of the forecasts of ANN, SVR, EMD-ANN and EEMD-ANN models was performed based on two performance parameters: Root Mean Square Error (RMSE) and Directional Accuracy (DA). RMSE is the square root of mean of errors. Lesser the RMSE value better is the forecast. DA represents the number of times the forecasted values matched the direction specified by the sign followed by the original series. Higher the value of DA, better are the forecasts. Hybrid EEMD-ANN model has better performance, in terms of both RMSE and DA, compared to the remaining models (Table 1(a)). Figure 2 represents the results of the 1-step ahead forecasts obtained using both ensemble models.

	RMSE	DA (%)
EMD-ANN	103.37	52.89
EEMD-ANN	70.31	87.60
ANN	165.38	40.00
SVR	158.34	47.00

Table 1 (a). Error Measures

	Hybrid EMD-ANN		Hybrid EEMD-ANN	
	z	WSRT	z	WSRT
ANN	-4.188	+	-5.990	+
SVR	-4.771	+	-6.109	+

+: EMD-ANN > ANN, EMD-ANN > SVR, EEMD > ANN, EEMD > SVR

=: EMD-ANN = ANN, EMD-ANN = SVR, EEMD = ANN, EEMD = SVR

-: EMD-ANN < ANN, EMD-ANN < SVR, EEMD < ANN, EEMD < SVR

Table 1 (b). Wilcoxon Signed Rank Test (at $\alpha = 0.01$)

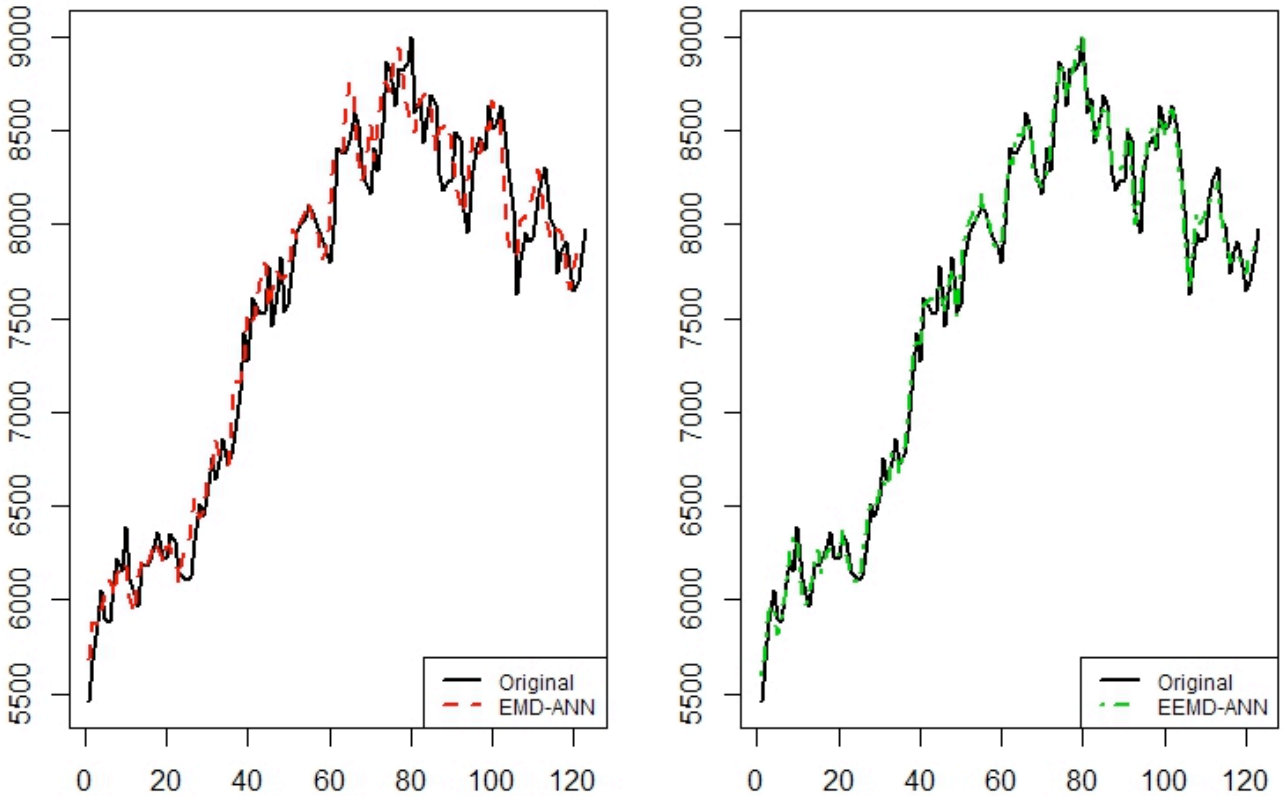

Figure 2: Forecasts Obtained Using EMD-ANN and EEMD-ANN Models

Significance Test

Wilcoxon Signed-Rank Test (WSRT), a non-parametric and distribution-free technique, is used to evaluate the predictive capabilities of two different models based on their signs and ranks (Diebold and Mariano 1995). From Table 1(b), it can be seen that **z** statistics value is beyond (-1.96, 1.96), hence the null hypothesis of two models being same is not accepted. The WSRT results confirm that the ensemble forecasting models outperformed the traditional SVR and ANN models.

Trading Rules

The closing price of Nifty on the first trading day of the following week can be predicted with reasonable accuracy using the discussed ensemble models. The investors can use these predicted values for making investment related decisions with the help of few trading rules.

Let $\hat{y}_k$ and y_k be the forecasted and actual close price on first trading day in the k^{th} trading week, respectively. An error index is used to represent the situation where the close price is expected to rise in the trading week k but it falls or remains same. The index E_k is defined as follows:

$$E_K = \begin{cases} 1 & if \quad \hat{y}_k > y_{k-1} \quad and \quad y_k \le y_{k-1} \quad \forall \quad k = 2,3,...,n \\ 0 & otherwise \end{cases}$$

where n is the total number of weeks.

Based on the error index (Hsu 2014), following three rules are used in this study:

Rule 1: IF ($\hat{y}_{k+1} > y_k$) AND (if an investor is not holding any stock on first trading day in the k^{th} week) THEN (s/he is advised to buy the stock on the next trading day in the k^{th} week)

Rule 2: IF ($\hat{y}_{k+1} < y_k$) AND (if an investor is holding any stock on first trading day in the k^{th} week) THEN (s/he is advised to sell the stock on the next trading day in the k^{th} week)

Rule 3: IF (the investor is holding any stock on first trading day in the k^{th} trading week) AND ($\sum_{j=0}^{2} E_{k-j} = 3$) THEN (s/he is advised to sell the stock on second trading day in k^{th} trading week)

The first rule suggests that an investor should buy a stock if he does not hold any stock on first day of the trading week since the price in the following week is expected to rise. The second rule suggests that if an investor is holding a stock on the first day of the trading week and observes that price is expected to fall the next week, he is advised to sell his stocks. The final rule suggests the investor to sell his held stock if the predictions of rising stock price are completely wrong for three consecutive weeks.

It is assumed that the security (stocks or index) can be traded (sold and bought) at the opening price on the next trading day of the week as soon as buying/selling decisions are taken. The transactions based on the trading rules using the results of EMD-ANN model are illustrated in Table 2.

Date	Close Price	Forecasted Close Price[1]	Transaction	Transaction Date	E_k	Rule
06-01- 2014	6171.45	6284	Buy at 6203.9	07-01-2014	1	1
13-01-2014	6261.65	6283.212				0
20-01-2014	6266.75	6211.375	Sell at 6320.15	21-01-2014	2	0
27-01-2014	6089.5	6096.991	Buy at 6131.85	28-01-2014	1	1
03-02-2014	6063.2	6221.944				1
10-02-2014	6048.35	6278.925	Sell at 6072.45	11-02-2014	3	1
17-02-2014	6155.45	6420.686	Buy at 6071.3	18-02-2014	1	0
24-02-2014	6276.95	6537.067				0
03-03-2014	6526.65	6437.396	Sell at 6216.75	04-03-2014	2	0

Table 2. Illustration of Trading Rules

The close price on January 6, 2014 is 6171.45, which is less than the forecasted value of first trading day of next week (6284.00). Hence, based on Rule 1, the investor is advised to buy the stock on January 7, 2014. On January 20, 2014, since the forecasted price for next week (January 27, 2014) is lower than the current price, then the investor is advised to sell his stock on January 21, 2014. On February 10, 2014, it was observed that prediction of stock price rising went wrong for the third time, hence using Rule 3, the investor is advised to sell his security on February 11, 2014. The rules were applied to rest of the test data. Similarly, return on investment (ROI) was calculated using trading rules for EEMD-ANN model. It was found that ROI obtained using predictions of EEMD-ANN and trading rules was higher than that of EMD-ANN and Buy-and-Hold strategy.

CONCLUSION

The paper presented a Cooperative Ensemble forecasting model that integrates EMD, Ensemble EMD and ANN. The model first uses EMD and EEMD to decompose the financial time series. Then, it uses ANN to predict the series separately and

[1] Forecasted values of first trading day of next week

aggregates the forecasted sub-series. The presented ensemble forecasting models showed a consistent superior performance in predicting the weekly Nifty index, as compared to both ANN and SVR. Further, it was observed that EEMD-ANN model outperformed EMD-ANN model.

In addition, three trading strategies based on EMD-ANN, EEMD-ANN and Buy-and-Hold strategies were evaluated to determine the timing for buying and selling the securities. It was found that the trading strategies based on the results of EEMD-ANN model yielded better ROI than that of EMD-ANN model and Buy-and-Hold strategies.

As a part of future direction, the model can be tested for high frequency intraday stock index data. A combination of cooperative and competitive ensemble forecasting techniques can be used for improving the forecasting accuracy.

REFERENCES

1. Atsalakis G, Valavanis K (2009) Surveying stock market forecasting techniques- Part II: Soft computing methods. *Expert Systems with Applications,* 36(3, Part 2),5932 - 5941.
2. Atsalakis G, Valavanis K (2013) Surveying stock market forecasting techniques- Part I: Conventional methods. Zopounidis C, ed., *Computation Optimization in Economics and Finance Research Compendium*, 49- 104 (New York: Nova Science Publishers, Inc).
3. Crowley P (2010) Long cycles in growth: Explorations using new frequency domain techniques with US data. *Bank of Finland Research Discussion Paper* No. 6/2010, org/10.2139/ssrn.1573641.
4. Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253- 265.
5. Hsu CM (2014) An integrated portfolio optimisation procedure based on data envelopment analysis, artificial bee colony algorithm and genetic programming. *International Journal of Systems Science,* 45, 12, 2645 – 2664
6. Huang N, Shen Z, Long S,Wu M, Shih H, Zheng Q, Yen N, Tung C, Liu H (1998) 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,* 454,1971,903 - 995
7. Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research,* 11,169 - 198.
8. Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis,* 1,1,1-41.