ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913

A stacked forecasting model-CEEMDAN-BPMs model

Yi Chen[1,a,*], Eva Khong[1,b]

[1]Faculty of Finance, City University of Macau, Macau, China
[a]f21092100171@cityu.mo, [b]evakhong@cityu.mo

Abstract: *For the predictive effects of individual models vary, unlike conventional implementations, this paper presents a new more comprehensive approach-CEEMDAN-BPMs model, which aim to the goal to the accuracy improvement of stock price forecasting, especially interesting for analysis nonlinear and nonstationary financial time series. This paper introduced a new model, that is BPMs model, furthermore, this new model is applied to financial time series prediction for high frequency data and low frequency data respectively. The key idea on the BPMs model relies on method of weighted linear stacking. Based on this new model with two decomposition method empirical mode decomposition (EMD) and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) are proposed in this paper. The paper employs the indicator mean absolute percentage error（MAPE), mean absolute error（MAE), mean square error（MSE)and correlation coefficient evaluation criterion and empirical results present that forecasting effects of new model CEEMDAN-BPMs is optimal in forecasting.*

Keywords: *Forecasting, EMD-BPMs forecasting, CEEMDAN-BPMs forecasting*

1. Introduction

Today's world is in the era of intelligence, which gradually presents complicated connectness relationship with external things. In other words, in most cases, the depiction of this peculiar relationship is nonlinearity instead of linearity relationship. This chaotic system plays an important role in complex mathematics-based subjects, such as the atmosphere, aerospace engineering, finance, etc. [1] Additionally, research on this complex system in recent years has mainly focused on how to improve the prediction accuracy [1-3]. Thus, in this era of big data, scholars are keen to use machine learning and deep learning for predicting stock prices and its trends which is concluded in chaotic system for recent research trends in prediction [4]. These chaotic systems including finance system, considering the financial market filled with all kinds of transaction information (fundamentals, news event, rumors, investor sentiment etc.), which is affected by a variety of factors [5]. Thus, basically the image of the financial market is characterized by a dynamic, nonlinearity, non-stationarity, noisy and chaotic, which makes the traditional theoretical methods based on stationarity and linearity assumptions no longer applicable [6-7]. This feature of the financial time series hints the task of prediction is full of challenge. However, scholars have never stopped the pursuit of research of this area of how to improve the prediction accuracy of the model in financial time series prediction, which is the objective of contribution this paper, attempting to understand the world from an alternative nonlinear perspective. Furthermore, this research of this paper has certain significance contribution in practical applications for financial industry, for the neural network intelligent algorithm model proposed by this paper in somewhat extent possibly for reference for some finance companies that employed neural network intelligent algorithm model to forecast time series [8-9].

This article will use a new decomposition and combination model to predict time series. It can be seen that the prediction advantage of combination model with CEEMDAN decomposition is over other single algorithm models and combination models respectively in comparative analysis result given by MAE, MAPE, MSE and R^2.

ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913

2. Methodology-Proposed model

2.1. Empirical mode decomposition (EMD)& Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN)

Empirical mode decomposition (EMD) is one kind of decomposition method, which is an adaptation approach [12]. Such decomposition method is appropriate for non-linear and nonstationary data analyses, producing a collection of intrinsic mode functions (IMFs) and one residue. To be specific, if the initial data is decomposed by EMD, IMF functions and a new residue will be produced after decomposition. If does not exist any extreme value and high frequency oscillation in the residue, EMD will incessantly decompose the residuals into new IMF functions and a new residue till the item of decomposition residual meets the Cauchy standard. The differences between IMFs and residues of EMD decomposition process are given as following [10,11]

$$r_k[t] = r_{k-1}[t] - IMF_k[t], k = 2, \dots, K \tag{1}$$

where $r_k[t]$ denotes the k-th residue at the time t and K denotes the sum of IMFs and residues.

Afterwards, Huang, et.al tackled the flaw of EMD of thoroughly extracting the local characteristics from the blended characteristics of the initial sequence. Hence, they put forward Ensemble Empirical Mode Decomposition (EEMD), adding white noise to the initial time series and performing EMD multiple times [39]. Thus, a new time series is given as following [11]

$$x^i[t] = x[t] + w^i[t], i = 1,2, \dots, N \tag{2}$$

where x[t] represents the initial time series data, $w^i[t]$denotes the i-th white noise, N represents the times of EMD decompsitional process.

EEMD can decompose x[t] into, the actual $\overline{IMF}$euqals the mean of $IMF_k^i[t]$.

$$\overline{IMF} = 1/N \sum_{i=1}^{N} IMF_k^i[t] \tag{3}$$

EEMD can decompose x[t] adding white noise into substantial $IMF_k^i[t]$ Nevertheless, the scholar [40] discovered the flaw of EEMD put forward by [13]. They developed a novel decompositional approach, i.e., CEEMDAN. Thus, the actual IMF function is given as following [13]

$$\overline{IMF} = 1/N \sum_{i=1}^{N} E_1(r_{k-1}[t] + \varepsilon_{k-1} E_{k-1}(w^i[t]) \tag{4}$$

$$r_k[t] = r_{k-1}[t] - \overline{IMF_k} \tag{5}$$

$$x[t] = IMF_k + r_k[t] \tag{6}$$

where the x[t] denotes the targeted time series data, $\overline{IMF}$ denotes the mean decomposition functions

2.2 BPMs model

As a single modeling method can merely forecast a certain tendency, for the sake of acquiring a superior forecast effect, our team utilized the approach of weighted linearity stacking of several models with the aim of combining into a more precise model. We acquired the forecast outcomes of the 4 modeling methods aforesaid:

$$y1(x) = func1(x)$$

$$y2(x) = func2(x)$$

$$\dots$$

$$yn(x) = funcn(x)$$

We aimed to establish a weighting matrix for the purpose of weighting the aforesaid modeling methods and acquire the superimposition forecast outcomes

$$yC\text{omb}(x) = \sum_{i=1}^{n} w(i) yi(x)$$

where yEn denotes the eventual forecast outcome of the aforesaid model combination, and denotes the weight coefficient of the entire modeling methods utilized for weighting, which ought to meet the

ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913

constraints below

We mainly intended to acquire a series of optimum weighting variables, for the sake of minimizing the forecast error. Hence, we established the optimized goal function:

$$\sum_{i=1}^{n} w(i) = 1$$

$$min \quad E\text{mse}=\frac{1}{m}\sum_{k=1}^{m}[\sum_{i=1}^{n} w(i)yi(xcross^k) - yreal(xcross^k)]^2$$

$$s.t \quad w(i) \geq 0, \ \sum_{i=1}^{n} w(i) = 1$$

It's noteworthy that, $xcross^k$ denotes a test input specimen for the m-fold cross-verification partition of the learning specimen, $yreal(xcross^k)$ denotes the real output value in correspondence to the input specimen at present.

We'll briefly introduce m-selective cross-verification. The m-selective cross-verification aims at dividing the entire learning specimens into m parts equally, taking m-1 of them for learning every time and afterwards taking the rest. One to test, this process loops m times, so that every specimen is tested, the specimen established in such way $xcross^k$ denotes an input specimen of m-selective cross-verification. $yi(xcross^k)$ denotes the forecast outcome of the ith model, n denotes the model quantity, and m denotes the cross-verification quantity. Eventually, we acquired a fitness function minimizing the forecast error based on the cross-verification of learning specimens. For the solution of the aforesaid modeling method, the genetic arithmetic was utilized to obtain the solution. The genetic arithmetic input is the aforesaid weight wi, the fitness function is Emse, the fitness equals to Emse in the aforesaid equation; the objective was to realize the minimization of the fitness function. Eventually, we utilized the weights to weight the outcomes of these 4 arithmetics to acquire the optimum forecast outcomes.

2.2. Evaluation criteria

In this paper, an optimal prediction model will be selected by several evolution indicators. The evaluation indicators employed is mean absolute percentage error (MAPE), mean absolute error (MAE), mean square error (MSE) and R square (correlation coefficient), which are defined as respectively as following

$$\text{MAE}=1/\text{n}\sum_{i=1}^{n} | Y_i - \widehat{Y_i}|$$

$$\text{MAPE}=1/\text{n}\sum_{i=1}^{n} | Y_i - \widehat{Y_i}|/Y_i$$

$$\text{MSE}=[1/n\sum_{i=1}^{n}(Y_i - \widehat{Y_i})]^2$$

$$R^2=1\text{-}\frac{\sum_{i=1}^{n}(\widehat{Y_i}-Y_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

where the predicted value $\widehat{Y_i}$ is the result of fitting the model learned from the training set to the test set, the observed value Y_i is the observed value in the test set, and the superscript of n represents the number of predicted observations. MAE is a measure of the average magnitude of forecast error without directional considerations. MSE measure prediction error is based on the average of the squared differences of the distances between the actual and predicted values. MAPE measures prediction error is based on the difference between the actual value and the predicted value as a percentage of the actual value. R^2 is used to measure the degree to which the predicted value fits the regression line. The higher the degree of fit the better the prediction to the actual value.

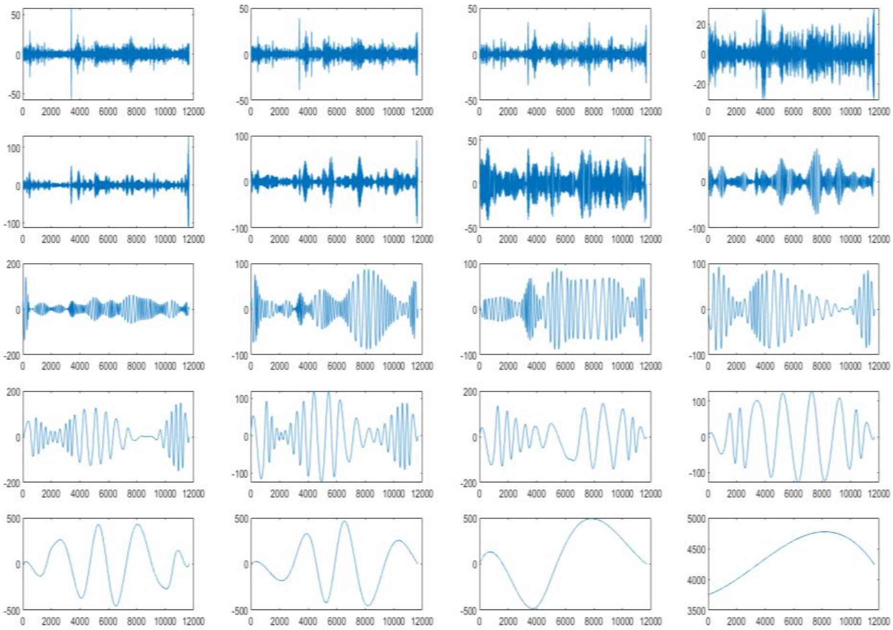
3. Experimental results and discussions

The experimental data in this paper comes from the Wind database, and the 15-minute-high frequency data and the daily closing price low frequency data of the CSI 300 Index are used as the research objects. The interval of the selected time series is March 18, 2019 to March 17, 2022. During the modeling process, 80% of the data will be used to train the model, and the remaining 20% of the data will be used to test the predicted performance of the model.

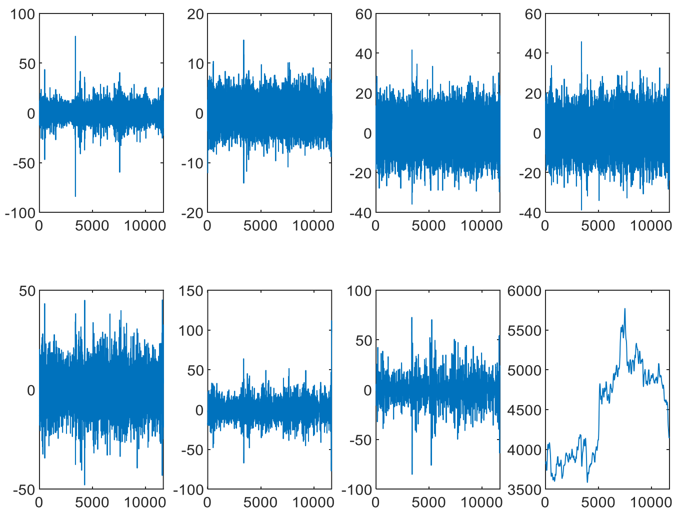
ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913

3.1 The results of data decomposition by EMD and CEEMDAN for high frequency and low frequency data

This part presents the decomposition result of EMD and CEEMDAN for the high-frequency and low-frequency data of the CSI 300 index respectively. We can see that Fig.1. (a) shows the decomposition of 15-minute high-frequency time series data that after EMD decomposition, the high-frequency time series is decomposed into 20 IMFs. Fig.1. (b) presents the decomposition of CEEMDAN, which is decomposed into 8 IMFs. We set the noise amplitude to 0.2, the number of noise additions to 500, and the maximum number of iterations to 5000 in the parameter setting. For the low frequency data, we can see that Fig.2. (a) shows the decomposition result of the time series data of the stock index daily data that decomposed by EMD method, presenting that the high frequency time series is decomposed into 11 IMFs. Fig.2. (b) presents the result that is decomposed into 8 IMFs by CEEMDAN. In the parameter setting, we also set the noise amplitude to 0.2, the number of noise additions to 500, and the maximum number of iterations to 5000.



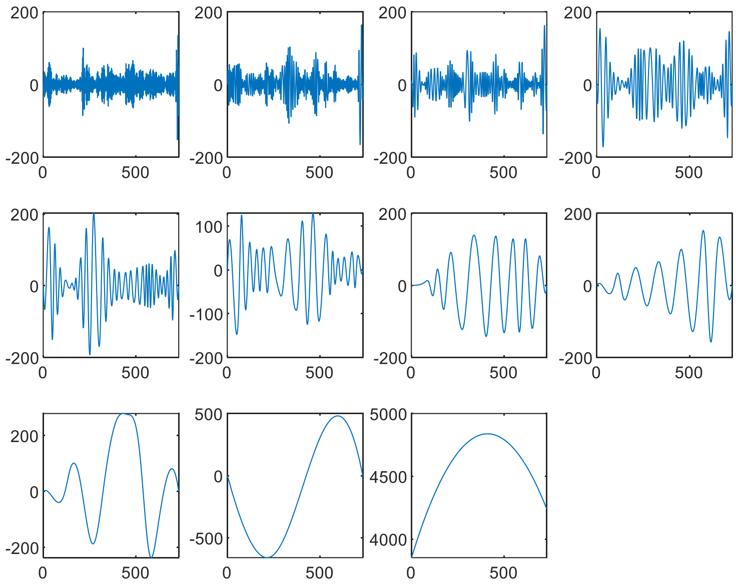
(a) 15 minutes High Frequency Data Decomposed by EMD



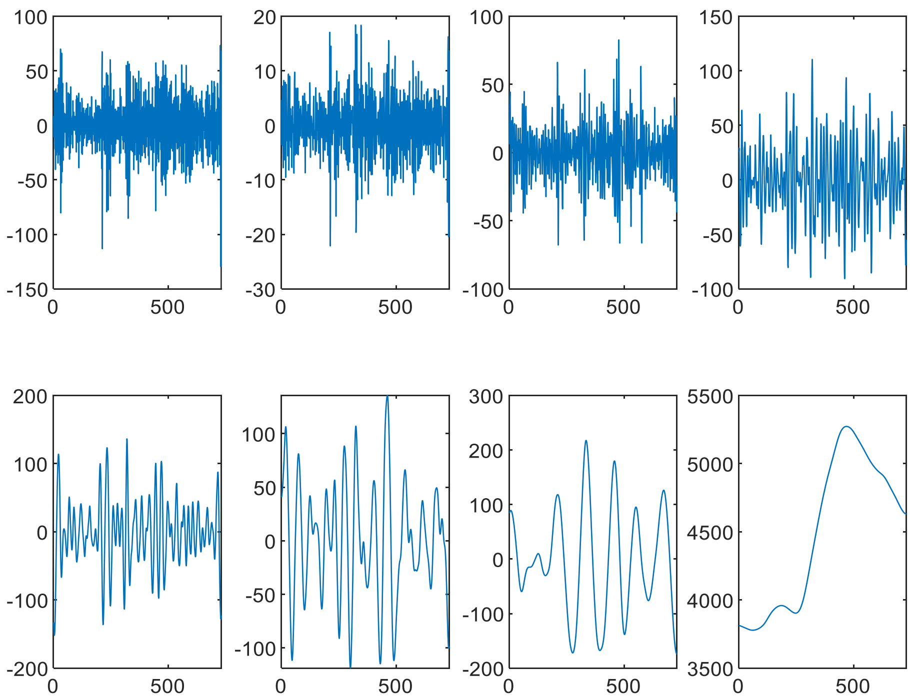
(b) 15 minutes High Frequency Data Decomposed by CEEMDAN

Figure 1: Decomposition results of CSI300 index for 15 minutes Data.

ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913



(a) Daily Data Low Frequency Data Decomposed by EMD



(b)Daily Data Low Frequency Data Decomposed by CEEMDAN

Figure 2: Decomposition results of CSI300 index for Daily Data.

3.2 Model prediction errors comparison for high frequency data and low frequency data

In this section, we would show how to choose an optimal forecasting model. Thus, the following will show the comparison results of the prediction performance of a single model and a combined model, which is evaluated by Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Square Error (MSE) and R-squared (Correlation coefficient) indicators to choose an optimal forecasting model. We adhere to a principle of selecting the optimal prediction model, that is, the value of evaluation the smaller the better.

ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913

Table 1: Comparison of the experimental results of the prediction error between the stacked model and the pure single model.

Panel A: Result of 15 minutes high frequency data				
Model	MAE	MAPE	MSE	R^2
BP	46.308288	0.009750	4127.812086	0.953282
LSTM	21.580714	0.004497	653.640832	0.996130
KELM	9.243126	0.001950	184.163850	0.997610
SVM	8.63899	0.001825	163.105792	0.997839
BPMs	8.656891	0.001826	163.515182	0.997864
EMD-BPMs	5.905360	0.001287	154.527893	0.998024
CEEMDAN-BPMs	**6.704520**	**0.001414**	**94.928101**	**0.998759**
Panel B: Result of daily data low frequency data				
Model	MAE	MAPE	MSE	R^2
BP	57.729137	0.012124	5120.422006	0.932427
LSTM	44.592704	0.009377	3385.740622	0.963151
KELM	41.333389	0.008731	2903.165490	0.964174
SVM	40.710606	0.008606	2881.230556	0.963809
BPMs	40.559036	0.008572	2895.082520	0.963735
EMD-BPMs	33.772346	0.007151	1838.364624	0.981951
CEEMDAN-BPMs	**24.841278**	**0.005214**	**901.286255**	**0.995212**

From the comparison results given by Panel A and Panel B in Table 1, the prediction performance of the CEEMDAN-BPMs model is the optimal model in prediction of high-frequency data and low-frequency data.

Following the analysis is given of Panel A of the prediction error of 15 minutes high frequency data. Firstly, we could make comparison of the purely model BP, LSTM, KELM, SVM and the newly stacked model BPMs model. From the comparison error results, the best prediction performance is the SVM model among other purely models, because its error indictors are smaller than previous purely model. Then we would compare the single best model SVM model with the BPMs model, and we find that the new stacked BPMs model is slight inferior to the SVM model, for most of the indicators are not as good as the single optimal model. Even though the fitting degree of the BPMs model is higher than the SVM model, which is 0.997864. Therefore, we consider improve the prediction performance of the new combination model by decomposition method. We can clearly see that the prediction performance of the BPMs model after decomposition is significantly improved, which can defeat the SVM model and its stacked BPMs model. Specifically, we could make comparison with SVM model, BPMs mode and EMD-BPMs model. The prediction error indicators of EMD-BPMs for MAE, MAPE, MSE and R square for 5.905360,0.001287,154.527893 and 0.998024 respectively, which is better than SVM model of its MAE, MAPE, MSE for 2.73363,0.000538 and 8.577899 respectively, and for R square, EMD-BPMs model is higher 0.000185 than SVM model. When compare with BPMs model, EMD-BPMs model prediction error has improved significantly its MAE, MAPE and MSE indictor is smaller for 2.751531,0.000539 and8.987289 respectively and its R square has improved 0.00016.

On the other hand, for the prediction error of CEEMDAN-BPMs for MAE, MAPE, MSE and R square is 6.704520,0.001414,94.928101and 0.998759 respectively, and finally we have reason to believe that the CEEMDAN-BPMs model is the optimal model in forecasting, for comparing with the EMD-BPMs model with its the high goodness of fit and its MSE value is smaller.

Next, we would give an analysis of precision performance of models in low frequency data given by Panel B. The analysis logic is the same as mentioned above. We firstly start to analyze and compare the single model BP, LSTM, KELM, SVM and the newly stacked BPMs model. From this result given by Panel B, the results will not be too confusing. Thus, obviously, the newly stacked BPMs model has a relatively small prediction error and a high degree of fit. Therefore, there is no doubt that the BPMs model is the best prediction model in the first stage comparing with other purely models. Then, taking into improve prediction performance of the BPMs model, the low-frequency time series is to decompose in the second stage. Surprisingly, the result is satisfactory that is what we expected. The prediction result of CEEMDAN-BPMs does not disappoint for its MAE, MAPE, MSE is 24.841278,0.005214 and 901.286255, the R square is 0.995212, which significantly beat other models in comparison of prediction performance.

In conclusion, for the prediction and analysis of high-frequency data, the combined model needs to be decomposed to improve the prediction performance. For the prediction and analysis of low-frequency data, whether the stacked model in decomposition or not, its forecasting performance is better than those of a single model. Therefore, prediction performance would be improved potentially after decomposition. we could conclude that CEEMDAN-BPMs model is the optimal model in prediction of high frequency data and low frequency respectively.

4. Conclusion

CEEMDAN-BPMs model is used to predict historical data of stock index namely CSI300 CSI300 from emerging stock markets in this paper. Firstly, the original data for high frequency data and low frequency data is decomposed by the CEEMDAN and EMD method respectively. The paper creates a new model-BPMs model and adds SVM, BP, LSTM and KELM model as comparative models, and uses MAE, MAPE, MSE and R-square to evaluate the accuracy of single model and hybrid model, then concludes that the CEEMDAN-BPMs model is optimal no matter for high frequency data or low frequency data.

The method of decomposition prediction is used to construct the financial time series prediction model. It can be found that the prediction accuracy of the decomposed superimposed BPMs model is significantly improved. Among them, the prediction accuracy of the decomposed BPMs model is satisfactory for the prediction accuracy of high-frequency data and low-frequency data respectively. Therefore, for the prediction and analysis of high-frequency data, the combined model needs to be decomposed to improve the prediction performance. For the prediction and analysis of low-frequency data, the prediction performance of the stacking model is better than that of the single model with or without decomposition. Therefore, the prediction performance after decomposition may be improved. We can conclude that the CEEMDAN-BPMs model is the optimal model for predicting high-frequency data and low-frequency data, respectively.

References

[1] Yanan, G., Xiaoqun, C., Bainian, L., & Kecheng, P. (2020). Chaotic Time Series Prediction Using LSTM with CEEMDAN. Journal Of Physics: Conference Series, 1617(1), 012094. https://doi.org/10.1088/1742-6596/1617/1/012094

[2] Fan, H., Jiang, J., Zhang, C., Wang, X., & Lai, Y. C. (2020). Long-term prediction of chaotic systems with recurrent neural networks. arXiv preprint arXiv: 2004.01258.

[3] Wang, R., Peng, C., Gao, J., Gao, Z., & Jiang, H. (2019). A dilated convolution network-based LSTM model for multi-step prediction of chaotic time series. Computational And Applied Mathematics, 39(1). https://doi.org/10.1007/s40314-019-1006-2

[4] Wu, D., Wang, X., & Wu, S. (2021). A Hybrid Method Based on Extreme Learning Machine and Wavelet Transform Denoising for Stock Prediction. Entropy, 23(4), 440. https://doi.org/10.3390/e23040440

[5] Zhou, F., Zhou, H., Yang, Z., & Yang, L. (2019). EMD2FNN: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. Expert Systems With Applications, 115, 136-151. https://doi.org/10.1016/j.eswa. 2018.07.065

[6] Abu-Mostafa, Y., & Atiya, A. (1996). Introduction to financial forecasting. Applied Intelligence, 6(3), 205-213. https://doi.org/10.1007/bf00126626

[7] Lin, H., Sun, Q., & Chen, S. (2020). Reducing Exchange Rate Risks in International Trade: A Hybrid Forecasting Approach of CEEMDAN and Multilayer LSTM. Sustainability, 12(6), 2451. https://doi.org/10.3390/su12062451

[8] Hassoun, M. (1996). Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems [Book review]. IEEE Transactions On Neural Networks, 7(5), 1316. https://doi.org/10.1109/tnn.1996.536328

[9] Ruan, D., Zhou, C., & Gupta, M. (2003). Fuzzy set techniques for intelligent robotic systems. Fuzzy Sets And Systems, 134(1), 1-4. https://doi.org/10.1016/s0165-0114(02)00226-9

[10] Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., & Zheng, Q. et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings Of The Royal Society Of London. Series A: Mathematical, Physical And Engineering Sciences, 454(1971), 903-995. https://doi.org/10.1098/rspa. 1998.0193

[11] Zhou, Y., Li, T., Shi, J., & Qian, Z. (2019). A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices. Complexity, 2019, 1-15. https://doi.org/10.1155/2019/4392785

ISSN 2616-5775 Vol. 5, Issue 9: 82-89, DOI: 10.25236/AJCIS.2022.050913

[12] Wu, Z., & Huang, N. (2009). Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. Advances In Adaptive Data Analysis, 01(01), 1-41. https://doi.org/10.1142/s1793536909000047

[13] Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. (2011, May). A complete ensemble empirical mode decomposition with adaptive noise. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4144-4147). IEEE.