# Stock Price Forecast Based On EMD-PCA-GRU Neural Network Model

Meijuan YANG [1,a*]

[1]*School of Economics and Management, Harbin Engineering University, Harbin, China*
[a*]*e-mail: 1131808643@qq.com*

Abstract

In order to further improve the accuracy of stock price prediction, this paper combines EMD, PCA and GRU to build a financial time series data prediction model. Firstly, the daily degree and 5-minute closing price of the CSI300 index were substituted into EMD for data denoising, and the decomposed IMF components were substituted into PCA to reduce the dimension of the data, so as to improve the prediction accuracy and efficiency of the model. Finally, the extracted principal components were substituted into GRU model for stock price prediction. The results show that the prediction accuracy of EMD-PCA-GRU model is better than other models. And it is proved that the composite model has a higher degree of fit than the single model. Under the same model, the prediction accuracy of 5 minutes high frequency time series is better.

***Keyword—****stock price forecasting; empirical mode decompos-ition; principal component analysis; gated recurrent unit; deep learning*

## 1. INTRODUCTION

With the development of China's stock market, investors need to accurately change their investment strategies according to the resulting risks. In recent years, with the deepening of research in the field of financial market prediction, the focus of research has also turned to how to combine it with other models to achieve higher prediction performance. Krogh and Vedelsby pointed out that if the single model that constitutes the combined model is highly accurate and diversified, then the prediction results of the combined model must be better than that of the single model[1]. In order to reduce the noise of the original sequence, this paper introduces EMD to process the data before prediction. And decompose to form a multi-level IMF sequence for dimensionality reduction. The LSTM neural network model is widely used in the field of stock price prediction[2-4]. GRU is a slight but excellent variant of LSTM. Chen used different data sets to construct time series, and predicted stock returns with different methods. By comparing the results obtained, it was found that the GRU model had a better fitting degree[5]. This paper constructs a new combined prediction model, in order to further explore whether the prediction accuracy constructed in this paper is further improved, and compare the impact of high frequency data and low frequency data on the prediction model, and whether the prediction accuracy of the fit model is better than that of the single model.

## 2. MATERIALS AND METHODS

### *2.1 Empirical mode decomposition algorithm theory*

EMD is a method of signal stabilization. The realization process of EMD decomposition is to extract stock index components with different frequency characteristics from the original stock index data sequence step by step and treat them as the intrinsic mode function IMF. EMD method does not need to set the basis function before decomposition, but generates the basis function automatically, so it is more adaptive than traditional decomposition methods. Therefore, personal reasons can be avoided to influence the results. The complex and non-stationary closing price data of CSI 300 can be decomposed into relatively simple IMF components and residual R with different frequency characteristics. The original signal can be expressed as the sum of all IMF components obtained by decomposition and R, which can be expressed as:

$$\varphi(t) = \sum_{j=i}^{N} d_{j}(t) + R(t) \tag{1}$$

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 976–981, 2023.
https://doi.org/10.2991/978-94-6463-042-8_139

N is the number of IMF. In general, the number of N is equal to $\log_2 n$ .The IMF sequences were obtained through EMD decomposition, and their dimensions were reduced through PCA.

## *2.2 Principal component analysis*

PCA can reduce the dimension of data and solve the problem of multicollinearity between data. Multiple IMF sequences were obtained through decomposition of the original time series by EMD method. The obtained sequence will form a matrix each time. The principal component analysis method is used to analyze these matrices, and then the first M principal components are selected as the training sample data and input into the GRU, and the cumulative contribution rate is above 85%. If there are many principal components at this time, the principal components with eigenvalue greater than 1 can be selected. The principal components are determined according to the commonly used KMO metrics given by Kaiser. KMO coefficient is adopted to determine the principal component, and the KMO coefficient is shown in the formula:

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} p_{ij}^2} \tag{2}$$

$r_{ij}$ represents the Correlation coefficient of row i and $p_{ij}$ represents the Partial Correlation coefficient of column j.

## *2.3 GRU structural neural network*

RNN has the problem of gradient disappearance, so the processing effect of data information with long time interval is not very ideal. Based on this, the LSTM neural network is extended. GRU is evolved from LSTM. GRU structure has fewer training parameters and a simpler model, while maintaining the prediction effect that LSTM can achieve. The update gate in the GRU determines how much of the past information in the model needs to be passed on, and the reset gate controls the amount of past information that should be forgotten.

$x_t$ is the input, the GRU unit can be obtained by formula (3)- (6):

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{3}$$

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{4}$$

$$\tilde{h}_t = \tanh(r_t^* U h_{t-1} + W x_t) \tag{5}$$

$$h_t = (1-z_t)^* \tilde{h}_t + z_t^* h_{t-1} \tag{6}$$

Where, $z_t$ represents the update gate; $r_t$ represents the reset gate; $h_{t-1}$ is the output of the previous layer; $h_t$ is the summary of input $x_t$ and $h_{t-1}$ ; $\sigma$ represents the Sigmoid function; $U^{(z)}$ , $W^{(z)}$ , $U^{(r)}$ , $W^{(r)}$ ,U and W are training parameter matrix; $z_t^* h_{t-1}$ represents the composition of $z_t$ and $h_{t-1}$ . This paper will use the GRU to process the time series data, stock market stock price forecast.

## *2.4 Construction of EMD-PCA-GRU model*

Figure 1 is the frame diagram of the composite model of stock price forecast. The collected raw data of CSI 300 stock index were decomposed into several IMF components through EMD, then PCA was used for dimension reduction, and the processed principal components were input as GRU data to achieve prediction.

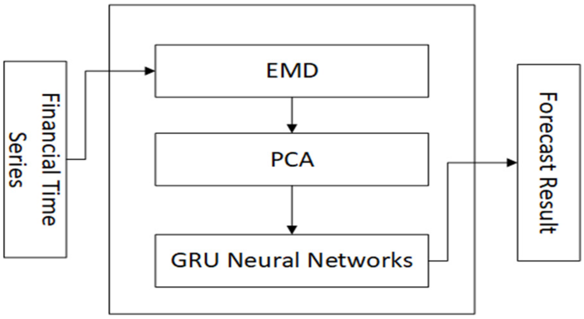

**FIGURE 1**. Composite model framework of stock price forecast

# 3. RESULT AND DISCUSSION

## *3.1 Data sources and ADF test*

### *3.1.1 Data sources*

In this paper, the daily closing price of the CSI 300 index from January 3, 2017 to August 31, 2020 and the 5-minute closing price from October 15, 2019 to August 30, 2020 are selected as examples. The time span of the sample covers many important economic pieces and is representative enough. Excluding the influence of holidays and other factors, there are 892 daily data and 10391 5-minute data. Data source: Wind database. 70% of the data is used as the training set and 30% as the test set.

### *3.1.2 Data test*

To verify that the financial data described above is non-stationary. This paper uses ADF method to test whether the sequence is stationary. Table 1 show the processing results of 5-minute data and daily low-frequency data respectively:

**TABLE 1**. ADF TEST

| | | | |
|---|---|---|---|
| ADF test of daily time series | | | |
| | | t-Statistic | Prob.* |
| ADF test statistic | | -0.939442 | 0.7757 |
| Test critical values | 1% level | -3.437475 | |
| | 5% level | -2.864574 | |
| | 10% level | -2.568439 | |
| Conclusion | | Non-stationary | |
| ADF test of 5 minutes time series | | | |
| | | t-Statistic | Prob.* |
| ADF test statistic | | 2.889148 | 1.0000 |
| Test critical values | 1% level | -3.430805 | |
| | 5% level | -2.861626 | |
| | 10% level | -2.566857 | |
| Conclusion | | Non-stationary | |

According to the table, within the confidence interval of 1%, 5% and 10%, neither of the two sets of data can reject the null hypothesis, that is to say, they are both non-stationary. Therefore, it is necessary to preprocess the data with EMD before prediction.

## 3.2 *Experimental setup and model test*

Results And analysis experiments were performed using Python3, SPSS and Keras. The selected data were substituted into EMD for decomposition, and the IMF sequence obtained by EMD decomposition was synthesized into a matrix, and then the dimension was reduced through PCA. The first step in the dimension reduction process is to calculate and judge whether the KMO value is greater than 0.5. Then, the principal components whose cumulative contribution rate is greater than 85% or eigenvalue is greater than 1 are extracted and substituted into the GRU model for training. The neural network set the time delay to 60. A three-layer GRU model is built in Python for training and prediction. The Sequential model commonly used in Keras is used during compilation. The daily data is smaller than the 5-minute high-frequency data, so the number of samples for one training (batch_size) is set to 20, and the number of samples (batch_size) for one training on 5-minute data is 100. In order to obtain better training effect, the epochs was set to 500 at first and then increased gradually to find the optimal training times. Use mean_squared_error as the loss function. The activation function of the Dense layer is set as linear function, and the optimizer is defined as "adam". MAE and MSE were used to judge the fitting degree of the model. Experiment and result analysis

## 3.3 *Empirical results*

### 3.3.1 *The IMF component of the CSI 300 stock index*

As mentioned above, the low-frequency diurnal data are decomposed by EMD method. Figure 2 shows the IMF component of the stock index with daily data.

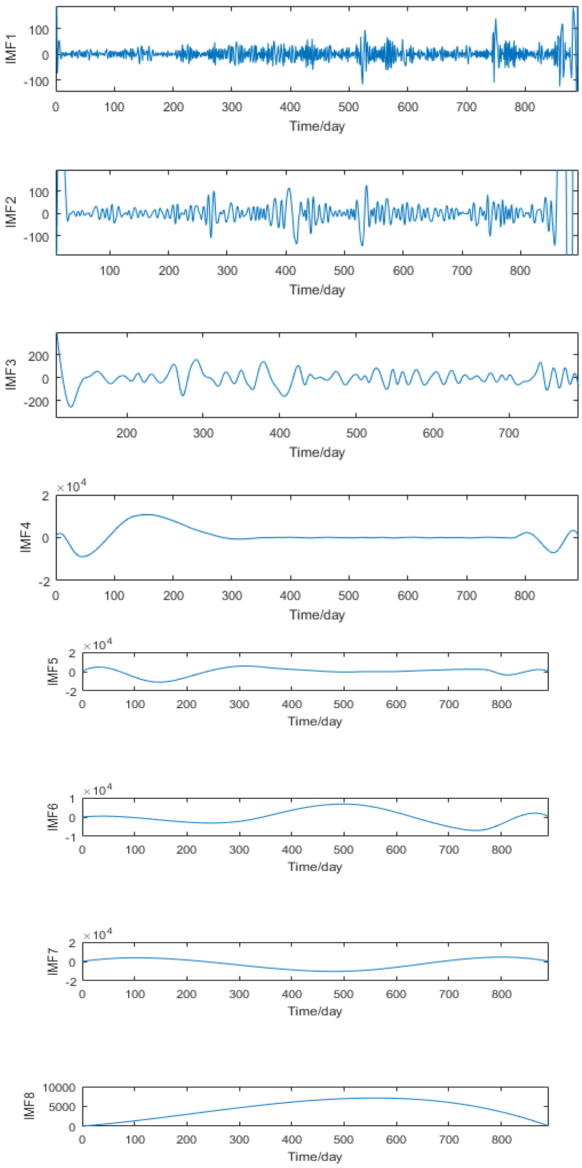

**FIGURE 2**. IMF Component Diagram of Daily Data

As can be seen from Figure 2, IMF1 has the highest fluctuation frequency, which contains the main information of the original data and captures the high-frequency fluctuation characteristics of the closing price. The frequency and amplitude of IMF1-8 gradually decrease, which contains the local characteristics of the original sequence and the data gradually becomes stable. Figure 3 shows the IMF component of the stock index with 5 minutes of data. It can be seen that IMF series has many layers, among which the 5-minute data IMF1 captures the high-frequency fluctuation characteristics of the closing price, and IMF12 reflects the low-frequency fluctuation of the closing price.

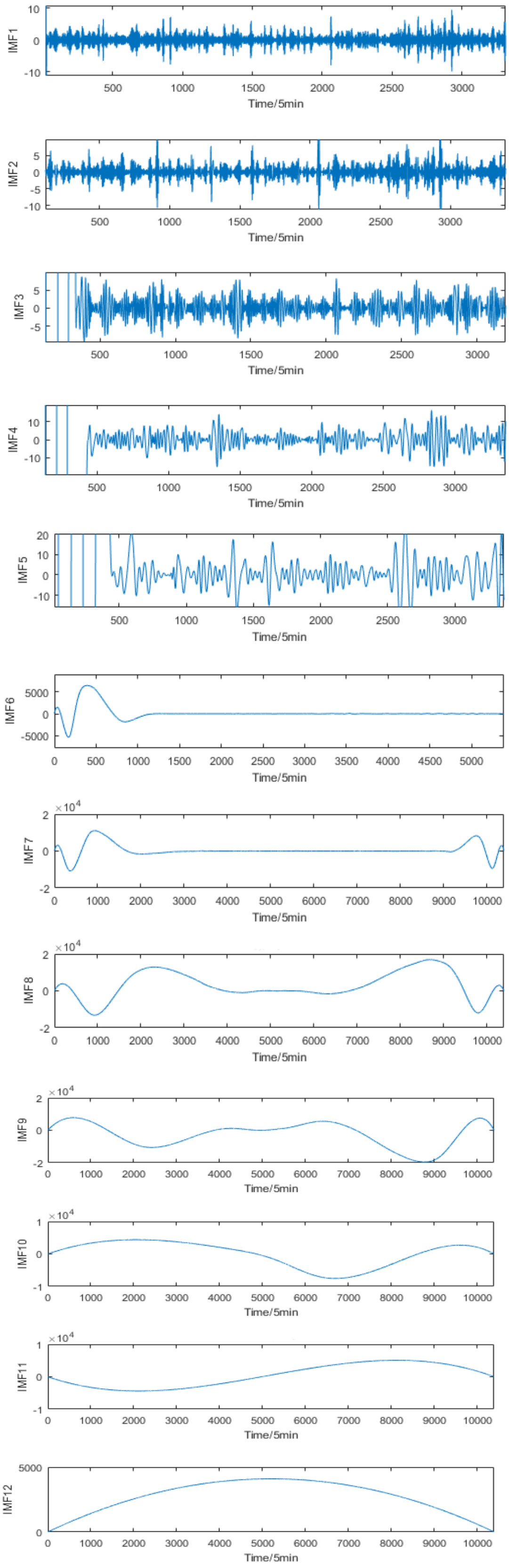

**FIGURE 3**. IMF Component Diagram of 5 minutes Data

Generally speaking, the prediction accuracy will improve as EMD decomposes more IMF sequences. However, the error caused by the decomposition will increase with the increase of IMF components, and the prediction accuracy will inevitably decrease. So choosing the right number of layers of decomposition is very important. As can be seen from the figure, the frequency and amplitude of all IMF vary with time, and the amplitude of IMF decreases from high frequency to low frequency.

### *3.3.2 Principal component selection after principal component analysis*

The KMO value of the IMF sequence of daily data of CSI300 is 0.819 and that of the IMF sequence of 5-minute high-frequency data is greater than 0.82. In the figure below, the principal components whose eigenvalues are greater than 1 and whose cumulative contribution rate exceeds 85% are selected. As can be seen from Table 2, although the cumulative contribution rate of principal component 1 of the daily data IMF series of CSI300 is 83.106% and less than 85% through principal component analysis, only its characteristic value exceeds 1, so principal component 1 is selected and substituted into GRU model. And the cumulative contribution rate of principal components 1 and 2 in the IMF series of 5-minute data of CSI300 after PCA processing is 89.297%, and the characteristic value is more than 1. In order to achieve a better prediction effect, this paper substituted the principal component 1 of the daily data and the principal component 1 and 2 of the 5-minute high-frequency data into the GRU model.

**TABLE 2**. CUMULATIVE CONTRIBUTION RATE OF PRINCIPAL COMPONENTS OF DAILY DATA AND 5-MINUTE DATA IMF SERIES

| Element | Initial characteristic value of 5 minutes data | | |
|---|---|---|---|
| | Sum | Variance% | Accumulation % |
| 1 | 9.422 | 78.518 | 78.518 |
| 2 | 1.293 | 10.779 | 89.297 |
| Element | Initial characteristic value of daily data | | |
| | Sum | Variance% | Accumulation% |
| 1 | 6.648 | 83.106 | 83.106 |
| 2 | .870 | 10.872 | 93.978 |

## ***3.4 Prediction results of EMD-PCA-GRU combined model***

### *3.4.1 Prediction results of daily data*

The extracted principal components are substituted into the GRU neural network for operation, and then the data is used to train the model to convergence to make the model better. It can be learned during the training process that the loss tends to decrease and is relatively stable after about 600 iterations. The convergence value is 0.0023, indicating that the model is well trained. In this paper, the daily data training times are set as 1000 times. Then the test set is substituted into the trained model. Figure 4 shows the prediction results of daily data in the EMD-PCA-GRU model.

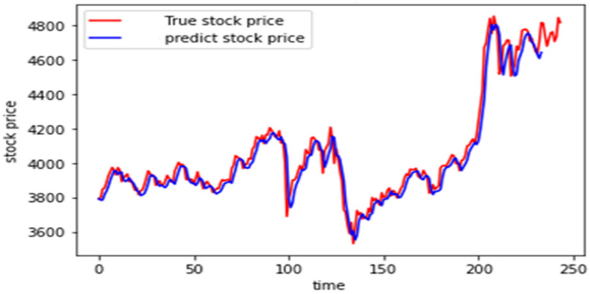

**FIGURE 4**. Daily EMD-PCA-GRU forecast results

### *3.4.2 5-minute data prediction results*

The loss of 5-minute data tends to decrease during training, and it is relatively stable after about 500 iterations. The convergence value is 0.0010, indicating that the model is well trained. Due to the large number of 5-minute high-frequency data test data, 175 prediction data were selected to represent the final prediction effect diagram for the convenience of observation. The final prediction results are shown in Figure 5.

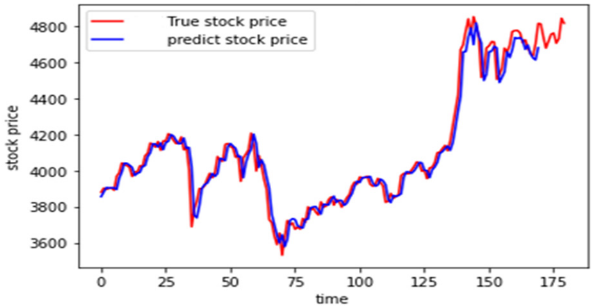

**FIGURE 5**. 5 minutes EMD-PCA-GRU forecast results

### *3.4.3 Comparison results between different models*

**TABLE 3**. COMPREHENSIVE EVALUATION RESULT

| | MAE | MSE |
|---|---|---|
| GRU（Low frequency data） | 0.0497 | 0.0038 |
| EMD-PCA-LSTM（Low frequency data） | 0.0341 | 0.0023 |
| EMD-PCA-GRU（Low frequency data） | 0.0287 | 0.0015 |
| EMD-PCA-GRU（High frequency data） | 0.0204 | 0.0010 |
| VMD-EEMD-LSTM（Low frequency data） | 0.4613 | —— |
| LSTM model based on transfer learning（Low frequency data） | —— | 0.015 |

As mentioned above, the daily low frequency data and the 5-minute high frequency data were respectively substituted into the EMD-PC-GRU combination model for prediction, and it can be concluded from Table3 that the 5-minute high frequency data had better prediction accuracy under the same conditions. Under the same daily data condition, MAE and MSE of the test set are both low, which indicates that the prediction accuracy of EMD-PCA-GRU is better than that of EMD-PCA-LSTM. By comparing the VMD-EEMD-PCA model, the MAE value is 0.4613[6]. And the MSE value of LSTM method based on transfer learning is 0.015[7]. It can be seen from the above table that the MODEL EMD-PCA-GRU model constructed in this paper is superior to other models and the single GRU model. In addition, it can be learned that under the same model, the model fitting degree of high-frequency data is better.

## 4. CONCLUSION

This paper proposes a new deep learning-based prediction model EMD-PCA-GRU, which has strong robustness and good reference value for quantitative investment. The research conclusions of this paper include: 1. The model proposed in this paper effectively combines the respective advantages of EMD,PCA and GRU to further improve the prediction ability of the model. 2. Due to the large amount of data selected in the training set, the model fitting effect is good and the accuracy of prediction is improved; 3. This paper uses high-frequency data for empirical analysis. Compared with low-frequency data, high-frequency data can reflect stock market information and stock price fluctuations more fully and timely, and the prediction effect of high-frequency data is better. 4. The application field of the financial time series prediction model proposed in this paper is not only limited to the financial field, but also can be used in the time series prediction field in other fields, with certain universality.

## REFERENCES

[1] Krogh, A., Vedelsby, J.(1994) Neural network ensembles, cross validation and active learning. In: Proceedings of the 7tn International Conference Neural Information Processing Systems. Cambridge. MA.USA.

[2] Kaur, R., Sharma, D., Yogesh, K., Bhatt D. P. (2021) Measuring Accuracy of Stock Price Prediction Using Machine Learning Based Classifiers. IOP Conference Series: Materials Science and Engineering.1099(1).

[3] Si, W. L., Ha, Y. K. (2020) Stock market forecasting with super-high dimensional time-series data using ConvLSTM, trend sampling, and specialized data augmentation. Expert Systems With Applications.161.

[4] Abba, S. G., Fernando, L., Daniel T. (2020) Stock Price Prediction Using LSTM and Search Economics Optimization. IAENG International Journal of Computer Science.47(4).

[5] Chen, K., Zhou, Y., Dai, F.Y. (2015) A LSTM-based method for stock returns prediction: A case study of China stock market. In: IEEE International Conference on Big Data. Jeju. Korea(south).2823-2824.

[6] Guo, J. L. (2020) Research on CSI 300 Index forecast based on VMD-EEMD-LSTM model. Modern Finance & Economics.40(08):31-44.

[7] Xie, F., Pan, B.X. (2020) High precision prediction method of LSTM Internet Finance Index based on transfer learning. Journal of Southwest Minzu University.41(07):129-134.