Stock Index Closing Price Prediction Based on KPCA-EMD-CEEMDAN-LSTM

Qingru Wu[1, *], Duqiao Han[2], Jingyi Li[3]

[1]School of Finance, Central University of Finance and Economics, Beijing, 102206, China

[2]School of Marine Science and Technology, Harbin Institute of Technology, Weihai, Shandong, 264209, China

[3]School of Mechanical and Automobile Engineering, South China University of Technology, Guangzhou, Guangdong, 510000, China

*Corresponding author. Email: wuqingru0304@163.com

Keywords: CSI 300, KPCA (Kernel principal component analysis), EMD (Empirical Mode Decomposition), CEEMDAN, LSTM (long short-term memory).

Abstract: A stock market index is a reference number compiled by a stock exchange or financial service institution to indicate changes in the stock market. Investors use artificial intelligence to build a model to predict the time series data to judge stock market trends. This paper proposes KPCA-EMD-LSTM-regular further to improve the accuracy of stock index price prediction. This paper chooses the more market-representative CSI 300 stock index data. In selecting indicators, the technical indicators are included in the model's input variables, and the impact of fundamental market indicators and technical indicators on the closing price of stock index futures is comprehensively considered. This paper uses the CSI 300 index data for empirical analysis. First, the kernel principal component analysis (KPCA) method is used to reduce the dimensionality of the data. Then a two-level decomposition model EMD-CEEMDAN model is constructed to decompose the closing price and regularized extended short-term memory network (LSTM-regular) model for prediction. The empirical results show that the hybrid model has a small error in predicting the closing price of stock index futures than the individual LSTM, EMD, and KPCA models and has achieved better prediction accuracy.

1. Introduction

The stock market is a complex nonlinear system, and the stock price is affected by many economic and social factors. Therefore, it is difficult for traditional linear or near-linear prediction models to efficiently and accurately predict the price trend of stock indexes. As we all know, deep learning transforms the feature representation of samples in the original space into a new feature space through layer-by-layer feature transformation and extracts features of a large amount of original time series data, thereby making prediction easier. The functional relationship between input and output is established through the learning and tuning of the network. Then, the relationship between reality is as close as possible. Our automated requirements for stock price prediction can be achieved using a successfully trained network model.

In the past ten years, artificial intelligence in the financial field has become a hot topic of discussion in academia and the financial industry. Many studies have also been published today, and various models have been generated. Huang Youwei et al. [1] used the long short-term memory neural network (LSTM) in the recurrent neural network (RNN) to build a financial forecasting model and found that the LSTM model has a good effect on the prediction of time series data. Later, given the shortcomings of the single model, scholars tried to add optimization algorithms to LSTM to make the prediction effect better. Liu Chong et al. [2] introduced an attention mechanism based on a deep LSTM network to distinguish the influence of different moments of the output layer on the prediction results of the current moment and found that the introduction to the attention mechanism can improve the performance of the model. Zhang Lu [3] used the traditional trend forecasting model and deep

learning LSTM to forecast financial time series data onto pattern mining and fitting data perspectives, which have potential complementarity in the forecasting effect. LSTM-PPEM combined prediction model. The results show that the combined model has lower error and lag than the single model. Ji Guangyue et al. [4] constructed a CEEMDAN-LSTM network cyanobacterial bloom prediction model using the Adaptive Noise Complete Ensemble Empirical Mode Decomposition (CEEMDAN) optimization algorithm and predicted the dissolved oxygen content of the Xijiang River. The empirical mode decomposition (EMD) of the data is beneficial for obtaining the eigenfunction, which is convenient for extracting the features of the subsequent deep learning. It significantly improves the prediction effect: Liu Me et al. [5] used the CSI 300 index, Shanghai Securities Composite Index (SSEC index), and Shenzhen Securities Component Index (SZI index). The EMD-LSTM model is constructed exponentially and compared with ARIMA, SVR, and LSTM models. The EMD-LSTM model has obvious advantages for predicting the training set and test set.

Kernel principal component analysis (KPCA) can extract the information about indicators to the maximum extent. It can improve the prediction effect when combined with machine learning models: Liu Huiping [6] used the principal kernel component analysis method to reduce the dimension of the petrochemical production process failure data input length. Finally, the proposed method is proved by using the Tennessee dataset, which shortens the training time and improves the effectiveness of the diagnosis. The main contribution of Bao Wei et al. [7] is the first attempt to introduce the SAEs method to extract the deep invariant daily features of financial time series and build WSAEs-LSTM, WLSTM, LSTM, and RNN, of which WSAEs-LSTM has the best predictability. LSTM-regular is an LSTM model that adds regular effects to the fully connected layer. Ren Jun et al. [8] applied the regularized LSTM model to the forecast of the Dow Jones index. The experimental comparison showed that this method's root means the square error was the smallest, and the prediction fitting degree was the highest.

After reviewing the existing literature, it is found that the three single models of KPCA, EMAD, and LSTM-regular have potential complementarity in the prediction effect. For the combination model constructed by using two of the single models to predict the trend of a stock index, such as EMD-LSTM, domestic and foreign scholars have made some research results, but few combination models constructed by three single models have been proposed. This paper proposes the KPCA-EMD-LSTM model and uses the CSI 300 index to test its optimization effect.

The main innovations of this paper include: (1) The EMD-CEEMDAN double decomposition is introduced to replace the traditional empirical mode decomposition method to decompose the financial time series trend. (2) Combined with actual investment experience, relatively more comprehensive and reasonable characteristic factors were selected, including technical analysis indicators of the CSI 300 Index and corresponding fundamental analysis indicators, and other transactions in domestic and overseas financial markets were considered. The influence of the CSI 300 index varieties is selected, and the market data extraction characteristics of these trading varieties are selected. (3) The LSTM model is regularized, and the LSTM-regular model is established to predict the trend of the CSI 300 index on the next trading day and obtain a better prediction accuracy than LSTM.

The rest of the chapters are arranged as follows. The second chapter first introduces the data dimension reduction method KPCA, the third chapter introduces the relevant theoretical basis of the EMD-CEEMDAN model and the LSTM model, and the fourth chapter uses the CSI 300 index data onto the empirical analysis. The empirical process mainly includes four steps: feature extraction, data processing, model calculation, and model verification. The feature extraction combines the actual investment experience and comprehensively selects the technical analysis indicators and fundamental analysis indicators of the CSI 300 index. The model calculation uses six models: LSTM, LSTM-regular, KPCA-LSTM, KPCA-LSTM-regular, KPCA-EMD-LSTM, KPCA-EMD-LSTM-regular to test the effect of stock index closing price prediction and select the most accurate model according to the prediction results.

2. Model

2.1 KPCA

To better deal with nonlinear data, this paper uses KPCA to reduce the dimension of the initial data, that is, introduces a nonlinear mapping function, mines the nonlinear information contained in the data set, and maps the data in the original space to a high-dimensional space. The specific steps are: calculating the kernel matrix, centralizing the kernel matrix, decomposing the eigenvalues, standardizing the eigenvectors, selecting the number of principal components, and calculating the principal nonlinear components. Data dimension reduction or feature (principal nonlinear component) extraction can be achieved through these steps. In addition, data reconstruction based on KPCA generally reconstructs the acquired nonlinear principal components into the original space by solving the approximate inverse mapping.

2.2 EMD-CEEMDAN

In this experiment, the EMD-CEEMDAN double decomposition is innovatively used to replace the traditional EMD decomposition method. Firstly, the experimental data are decomposed by EMD, and the abnormal IMF components are selected by screening. Then, the normal IMF components are obtained by noise reduction based on the EMD algorithm, and the CEEMDAN decomposition is directly carried out after the reconstruction of residual IMF. The energy of each IMF component is extracted as a feature to construct a feature set.

Empirical mode decomposition (EMD) is a new method for non-stationary processing signals proposed by Dr. Norden e. Huang, a Chinese-American scientist of NASA in 1998. It has obvious advantages in processing non-stationary and nonlinear data and is suitable for analysing nonlinear and non-stationary signal sequences with high signal-to-noise ratios. EMD decomposition model has been very mature and widely used. This paper will no longer do a specific introduction.

The CEEMDAN algorithm is a new signal decomposition algorithm proposed by Torres M E. et al. in 2011. When the overall average calculates the first-order IMF component obtained by CEEMDAN, the final first-order IMF component is obtained. Then, the residual part of the overall average calculation is repeated, which solves the modal aliasing phenomenon in empirical mode decomposition (EMD). The specific algorithm process of CEEMDAN is as follows:

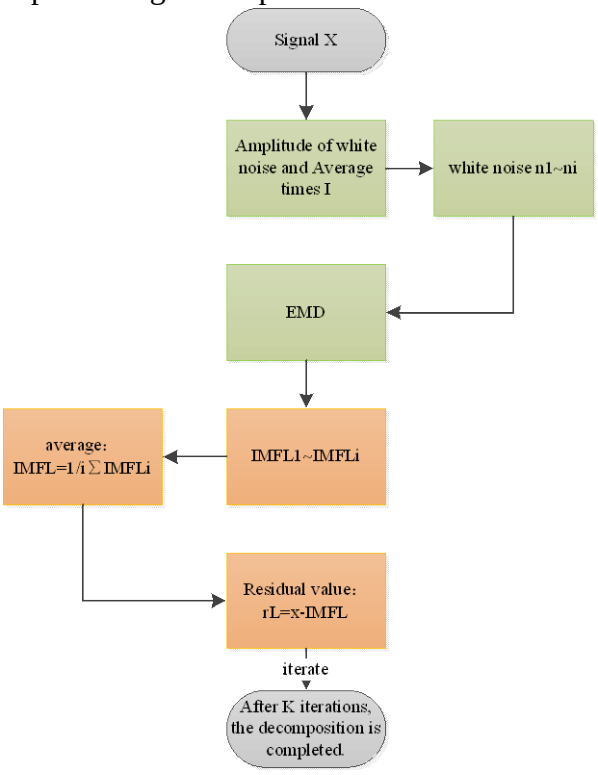

Figure 1. Ceemdan algorithm process

2.3 Regularized LSTM

Long Short-Term Memory (LSTM) was proposed by Sepp Hochreiter and Jürgen Schmid Huber in 1997, and many variants such as No Input Gate (NIG), No Forget Gate (NFG), and Gated Recurrent Unit (GRU) have been derived so far. They are widely used in language processing, image analysis, disease prediction, and synthetic music. LSTM is used in this paper. Its structure is shown in Figure 2:

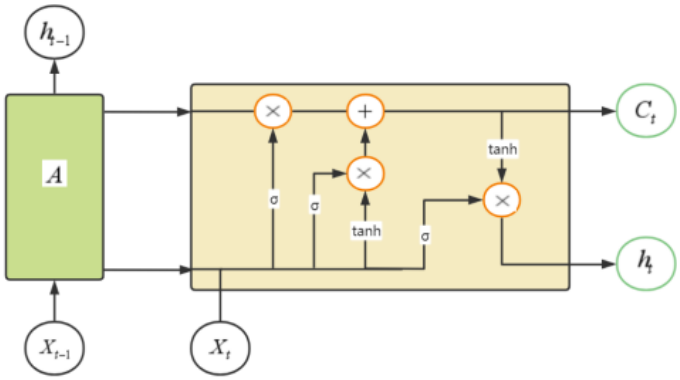

Figure 2. LSTM structure

LSTM has a chain structure. The data and the previous state are inputted at time t, the new hidden layer and memory state are obtained and continue to be passed down. LSTM is divided into four parts:

① Forget Gate: Forget Gate determines what information to forget from the cell state. The specific expressions are as follows (is the sigmoid function, is Forget Gate weight, is Forget Gate bias):

$$f_t = \sigma[W_f * (h_{t-1}, x_t) + b_f] \tag{1}$$

② Input Gate: Input Gate updates the information through a sigmoid layer and creates a candidate value vector through a tan layer. The specific expressions are as follows:

$$i_t = \sigma[W_i * (h_{t-1}, x_t) + b_i] \tag{2}$$

$$\tilde{C}_t = \tanh[W_c * (h_{t-1}, x_t) + b_c] \tag{3}$$

③ Update old cell state:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

④ Output Gate: The output part of the cell state determined by a sigmoid layer; is processed by a tan layer and multiplied by the output of the sigmoid layer. The expression is as follows:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

Deep learning is a black-box model, so the model is sometimes too complex and overfitting, resulting in poor prediction effect of the trained model. Therefore, this paper uses the L2 regularization method to add a penalty term to the original objective function to punish the model

with high complexity, reducing the interference characteristics that make the model too complex, improving the overfitting problem, and improving the prediction efficiency of the model. The expression of the L2 penalty term is as follows:

$$\Omega(w) = \| w \|_2^2 = \sum_i w_i^2 \tag{7}$$

The objective function expression under the L2 regularization method is as follows:

$$\tilde{J}(w; X, y) = J(w; X, y) + \alpha\Omega(w) \tag{8}$$

3. Calculation

3.1 Data Sources

The Shanghai Composite Index considers the market value of all listed companies according to its compilation method, regardless of whether these listed companies represent the theme of the Chinese economy. Moreover, with the successive listing of some heavyweight stocks, the dominant role of large-cap stocks on the Shanghai Composite Index has become more and more evident. The Shanghai Composite Index uses the sum of the changes in the market value of all listed stocks to reflect the changes in the stock market, which makes the Shanghai Composite Index constantly distorted and jeopardizes investors' and economic management departments' grasp of the trend of the stock market. [9] Therefore, this article uses the relevant data of the CSI 300 index for modeling analysis instead of The Shanghai Composite Index, which is obtained from the share financial terminal (www.tushare.com). The data range selected in this paper is from May 16, 2016, to January 21, 2021, and the sample size is 1387. The data includes the market data, the technical analysis indicators, and the fundamentals-related data. The market data includes closing price, highest price, lowest price, opening price, trading volume, and so on.

The fundamental data for the CSI 300 index includes CSI 300 futures IF main contract market Data, dynamic PE, dynamic PB, and so on. Moreover, to follow the cyclical changes in trends and build an effective machine learning model to predict the daily trend of the CSI 300 Index, this paper constructs 30 technical analysis indicators from Python's third-party Talib, such as DEMA EMA, and KAMA, to enrich our factor library. The technical analysis indicators of the index itself are based on the market data and fundamental data of the CSI 300 Index. We performed a routine test of time series on the closing prices we wanted to predict. In the unit root test, the undifferentiated data failed to reject the null hypothesis that the series was not stationary with a P-value of 0.577. This provides a reason for us not to adopt the classic ARMA model that has strict requirements on series stationarity and also reflects our forward-looking adoption of machine learning models.

This paper performs a routine test of time series on the closing prices we wanted to predict. In the unit root test, the undifferentiated data failed to reject the null hypothesis that the series was not stationary with a P-value of 0.577. This provides a reason for us not to adopt the classic ARMA model that has strict requirements on series stationarity, and it also reflects our forward-looking adoption of machine learning models.

3.2 Decomposition Results

In this paper, the EMD-CEEMDAN algorithm is used to divide the closing price of the Shanghai and Shenzhen 300 stock index into eight different IMF series and one residual series according to the data types of fluctuation. As shown in Figure 3:

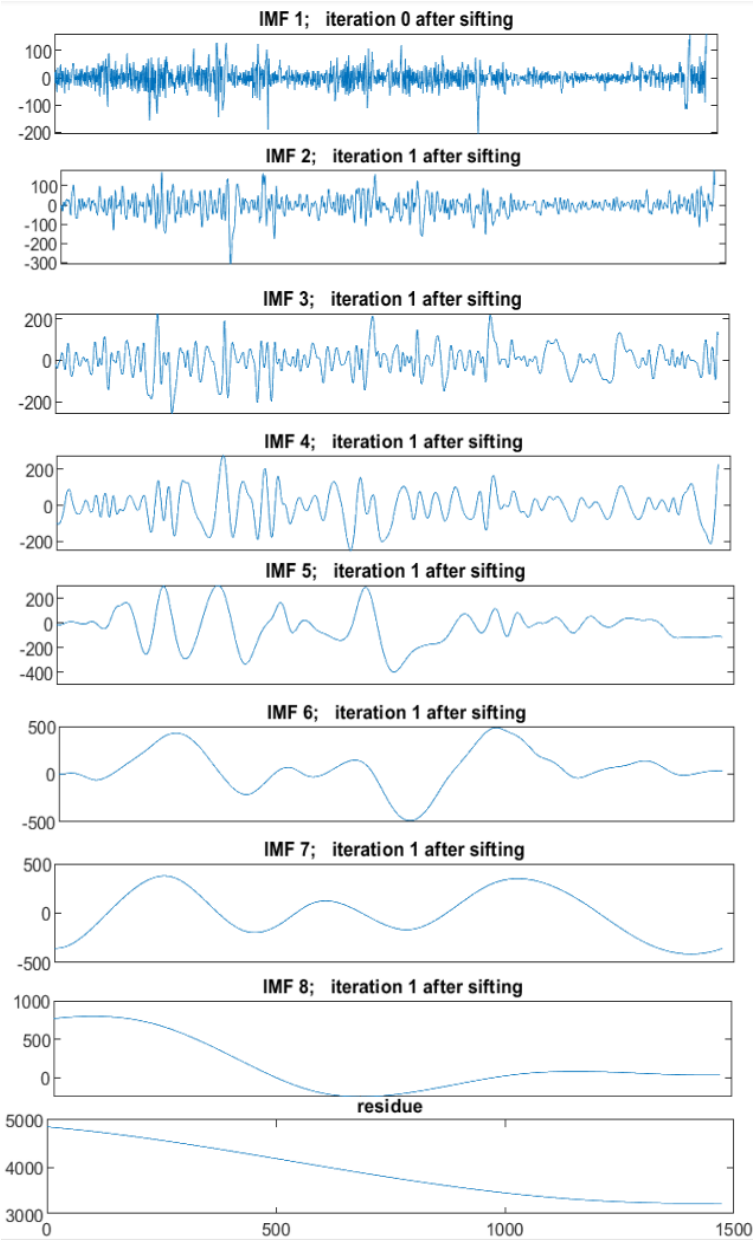

Figure 3. IMF decomposition results and residual value sequence diagram

3.3 Model Calculation

This paper uses the CSI 300 index closing price and related basic analysis factors, technical analysis factors, with LSTM, LSTM-regular, KPCA-LSTM, KPCA-LSTM-regular, KPCA-EMD-LSTM, KPCA-EMD-LSTM-regular these six models to test the stock index closing price prediction effect. The experiment uses a sliding time window to build samples, the width of the sliding window is 10, that is, using the data of the first 10 days to predict the closing price data of the 11th day. The training takes the first 90 % of the samples as the training set and the last 10 % as the test set. The software used in this paper is Pycharm, and the selected deep learning library is Keras. There are 256 neurons in LSTM hidden layer and the optimization algorithm is adam. The calculation method of error is the mean square error. The epoch of model training is 150. The batch size is 100 and the Dropout is 0.01. KPCA is a kernel principal component analysis method, and the final dimension is 2-dimensional. EMD is an improved ensemble empirical mode decomposition (EMD-CEEMDAN). LSTM-regular is to add the L2 regularization command to the full connection layer and set kernel _ regularizer = l2 (0.0001) to limit the weight of the layer. Model predictions are shown in Figure 4 and Table 1 (In the figure, r is the abbreviation of regular, and the prediction results of each model are the best results in multiple experiments).

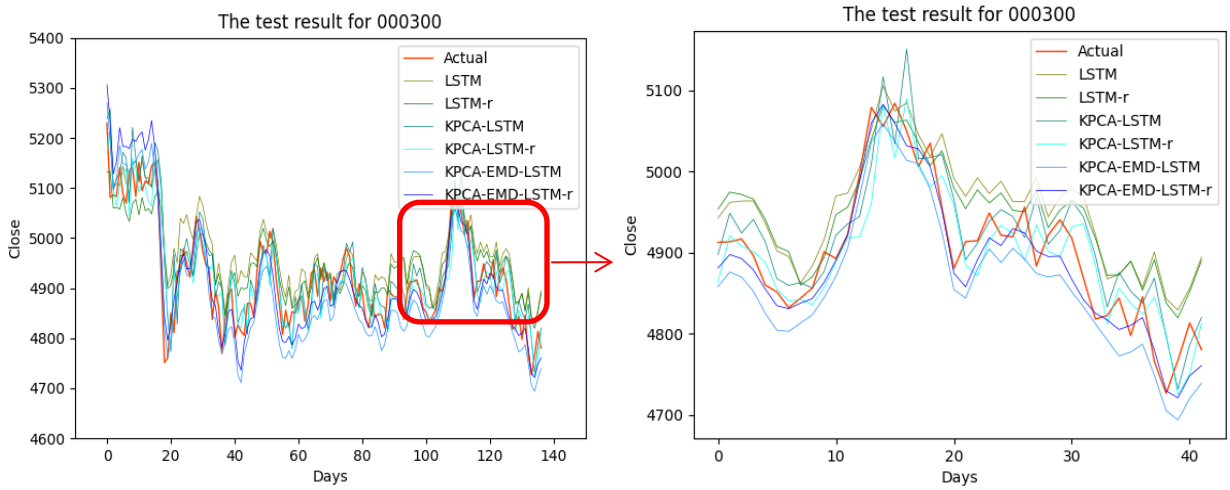

Figure 4. Prediction curves of each model

Table 1. Evaluation of prediction results of each model

MODEL	RMSE (point)	MAPE	MAE (point)	R^2
LSTM	71.962	1.173	58.294	0.477
LSTM-regular	62.276	1.005	49.731	0.608
KPCA-LSTM	59.100	0.939	46.528	0.647
KPCA-LSTM-regular	55.019	0.844	41.590	0.694
KPCA-EMD-LSTM	52.593	0.928	45.191	0.721
KPCA-EMD-LSTM-regular	43.710	0.685	33.936	0.807

From the data in the table, the goodness of fit of the regularized LSTM model is 27.53 % higher than that of the traditional model. In the latter two more optimized models (KPCA-LSTM and KPCA-EMD-LSTM), the improvement effect of regularization on the goodness of fit is reduced, which is 7.27 % and 11.99 %, respectively. The goodness of fit of the LSTM model after KPCA dimension reduction is 35.70 % higher than that of the traditional model. In the more optimized regularized LSTM model, the improvement effect of KPCA on the goodness of fit is reduced to 14.13. The KPCA-LSTM model with empirical mode decomposition operation increases the goodness of fit by 11.34 % and 16.24 % in irregular and regular states. Finally, the KPCA-EMD-LSTM-regular model has the best prediction effect. Its root mean square error is 43.71, the average absolute percentage error is 0.69, the average absolute error is 33.94, and the goodness of fit is 0.81, which is 69.21 % higher than that of the traditional LSTM model.

4. Conclusion

The unstable characteristics of financial time series have adverse effects on its prediction effect. The traditional EMD decomposition method provides ideas to alleviate this adverse effect. In this paper, EMD-CEEMDAN dual decomposition is used to replace the traditional EMD decomposition, and KPCA factor dimension reduction and model regularization are supplemented to construct a more accurate prediction model KPCA-EMD-LSTM-regular than the traditional LSTM model. This paper fully considers the relevant factors of the CSI 300 index and constructs a complete factor library. In the experiment, the sliding time window is used to construct the sample, and the width of the sliding window is 10. Finally, the optimization of the model regularization, KPCA factor dimension reduction, EMD-CEEMDAN dual decomposition operation, and the superiority of the prediction model KPCA-EMD-LSTM-regular are verified.

Based on the results of our paper, we propose the following suggestions to technical researchers related to financial investment. First, it is suggested to integrate various factors such as technical indicators and fundamental indicators instead of monotonous market data when predicting price trends, and using methods such as the nuclear principal component analysis to reduce the complexity

of the model regression caused by the data dimension; Second, using regularization to reduce the unnecessary complexity of the model may increase the prediction effect of the model; Third, using methods to decompose the sequence to be predicted can improve the accuracy, although it increases the operational complexity. Combining these three operations can obtain a more accurate forecast value, and constructing an investment strategy based on the forecast value can increase investors' confidence in the expected return value, thereby reducing the risk of return uncertainty.

References

[1] Huang Liming, Chen Weizheng, Yan Hongfei, Chen Chong, et al.A stock Prediction Method Based on Recurrent Neutral and Deep Learning. Software Guide, 2019, 18 (1): 7.

[2] Liu Chong, Du Junping, et al.Financial Data Prediction Method Based on Deep LSTM and Attention Mechanism [J]. Computer Science, 2020, 47 (12): 6.

[3] Zhang Lu Research on stock market trend forecasting method based on combination model [D]. University of Chinese Academy of Sciences (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences).

[4] Wang Yaqian, Zhu Jiaming. Research on stock index futures arbitrage strategy based on GARCH model [J]. Journal of Chongqing University of Science and Technology: Social Science Edition, 2016 (6): 3.

[5] Liu Ming, Shan Yuying. Prediction of stock index closing price based on EMD-LSTM model [J]. Journal of Chongqing University of Technology (Natural Science), 2021, 35 (12): 269 - 276.

[6] Liu Huiping, Li Dazi. Fault diagnosis of the petrochemical production process based on the KPCA-LSTM model [J]. Petrochemical Automation, 2021, 57 (6): 6-9, 36. DOI: 10.3969/j.issn.1007-7324.2021.06.002.

[7] Wei B, Jun Y, Yulei R, et al. A deep learning framework for financial time series using stacked autoencoders and long-short term memory [J]. PLoS ONE, 2017, 12 (7): e0180944.

[8] Ren Jun, Wang Jianhua, Wang Chuanmei, et al. Stock Index Prediction Based on Regularized LSTM Model [J]. Computer Applications and Software, 2018, 35 (4): 6.

[9] Yao Nan, Zhang Haiping. Feasibility analysis of abolishing Shanghai Composite Index and using CSI 300 Index to improve decision-making level [J]. Science and Technology Information, 2008 (15): 2.