Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 13, No. 2, 2024

Research on Stock Portfolio Construction Based on Bi-LSTM Neural Networks

Jiawei Li *

School of Statistics, Tianjin University of Finance and Economics, Tianjin, China
* **Corresponding author Email:** 3137704296@qq.com

Abstract: In recent years, the rapid globalisation of China's financial market has provided more opportunities for investors, but also brought about a more complex investment environment. This paper constructs stock portfolios based on Bi-LSTM neural networks, aiming to improve the accuracy of stock price prediction and the optimisation of investment portfolios using deep learning techniques. The theoretical part introduces the portfolio theory, including the mean-variance model and the capital asset pricing model, and explores the advantages of LSTM, Bi-LSTM and ATT-LSTM in processing time series data. The constituent stocks of CSI 300 index are selected in the research design part, and the stocks are screened using entropy weighted TOPSIS method and analysed based on the data from January 2018 to April 2024. The closing price and logarithmic return are predicted by constructing and using LSTM, Bi-LSTM and ATT-LSTM models, and then the trading strategies of EMA, MACD double conditions are determined, and the investment weights are determined by Monte Carlo method for the investment portfolio. The results of the empirical study show that the Bi-LSTM model has the optimal prediction performance, and based on the prediction data of the model, the trading strategy using the dual conditions of EMA and MACD achieves a higher investment return than the strategy using only MACD. In summary, this paper demonstrates the superiority of Bi-LSTM model in stock price prediction through empirical research, and proposes an effective portfolio construction method and trading strategy, which helps investors make more effective decisions in the complex market environment.

Keywords: Bidirectional long and short-term memory neural network; Entropy weight TOPSIS method; Monte Carlo method.

1. Introduction

With the acceleration of globalisation, China's financial market has become increasingly internationalised, providing more opportunities for investors but also bringing about a more complex investment environment. In this context, accurately predicting stock prices and optimising investment portfolios have become key issues. Traditional mean-variance models and capital asset pricing models have limited predictive power in the face of complex time series data. Deep learning, especially Long Short-Term Memory (LSTM) networks, excels in handling time series data. Bidirectional Long Short-Term Memory Networks (Bi-LSTM) further improves the ability to capture both forward and backward information, resulting in more accurate stock price predictions [1].

The purpose of this paper is to construct a stock portfolio model based on Bi-LSTM neural network and explore the application of deep learning techniques in stock price prediction and portfolio optimisation. We select the constituent stocks of CSI 300 index, use the entropy weight TOPSIS method for stock screening, and conduct empirical analysis based on the data from January 2018 to April 2024. By comparing the forecasting performance of LSTM, Bi-LSTM and ATT-LSTM models, and combining the trading strategy with the dual conditions of EMA and MACD, we verify the superiority of the Bi-LSTM model in enhancing investment returns.

This study enriches the portfolio theory and provides new decision-making methods for investors in complex market environments, aiming to promote the development of financial markets and the innovation of investment strategies.

2. Relevant theories

2.1. Portfolio

2.1.1. Portfolio Theory

Portfolio theory, first introduced by the American economist Markowitz in 1952, uses mean and variance to portray return and risk, two key factors that influence investment choices [2]. In a given portfolio, the mean represents the expected return, the variance is the variance of the portfolio's return, and the standard deviation is used to measure systematic risk. This is shown in Figure 1.

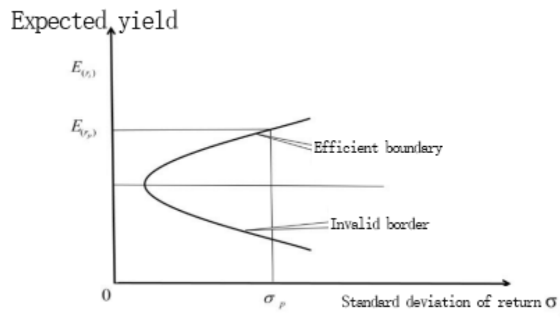

Figure 1. The efficient frontier of the portfolio

2.1.2. Capital Asset Pricing (CAPM) model

The Capital Asset Pricing Model (CAPM) is used to estimate the expected return of an asset and describes the relationship between the expected return of an asset and its risk with the following expression:

$$\bar{r_i} = r_f + \beta_i(\overline{r_m} - r_f) \tag{1}$$

Where $\bar{r_i}$ is the expected return of asset i, r_f is the risk-free rate, β_i is the beta coefficient of asset i, which represents the systematic risk of the asset relative to the market portfolio, and $\overline{r_m}$ is the expected return of the market portfolio.

The CAPM, as one of the cornerstones of capital market theory, provides an important theory for the study of asset pricing, which helps investors to understand the relationship between asset return and risk, and provides a basic explanation for the operation of financial markets.

2.2. Bi-Directional Long Short-Term Memory

In order to deal with the problems of gradient vanishing and gradient explosion that exist in traditional RNN, Hochreiter and Schmidhuber proposed a special type of RNN in 1997, i.e., Long Short-Term Memory Neural Network (LSTM) [3], which can make up for the deficiencies of RNN due to the unique network structure that LSTM possesses. The LSTM, in comparison with RNN, has three gating units, i_t and o_t, which solves the problem of long time sequence dependency in deep learning. Compared with RNN, LSTM adds three gating units, namely, the forgetting gate f_t, the input gate i_t and the output gate o_t, which solves the problem of long time sequence dependency in deep learning, and the structure is shown in Figure 2:

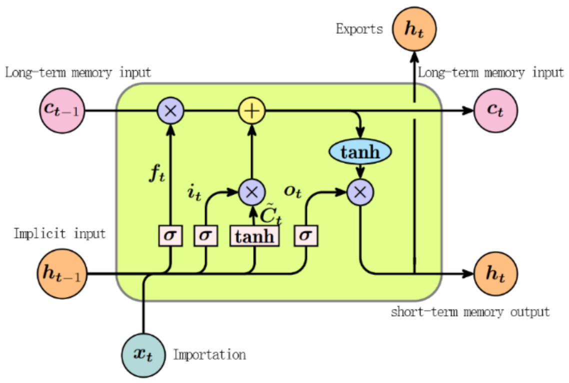

Figure 2. LSTM network structure

2.3. Bi-Directional Long Short-Term Memory Neural Network

An unavoidable problem in LSTM modeling is the inability to encode back-to-front information. However, bi-directional LSTMs can extract the information before and after the data at the same time, which in turn mines deeper features and better captures the dependencies of bi-directional information. Two LSTMs are combined to form a single-layer Bi-LSTM, where one LSTM can process the input sequence in the forward direction; the other LSTM can process it in the reverse direction, and the outputs of the two are spliced together to form the final Bi-LSTM output after all the time steps are completed. The forward LSTM generates a result vector after processing all time steps, and the reverse LSTM also gets another result vector. These two result vectors are combined as the result of Bi-LSTM [4], which is then input into the subsequent neural network for regression or classification to obtain the final output features, the structure of which is shown in Figure 3:

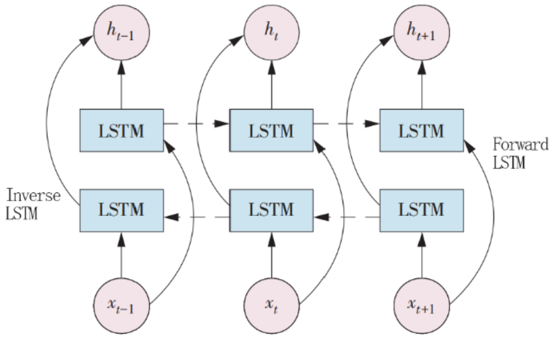

Figure 3. Bi-LSTM structure

2.4. The Long Short-Term Memory Neural Network based on Attention Mechanism

In daily life, human beings tend to selectively receive useful information because of their limited ability to process information. The ATT-LSTM model is similar to the process of human beings processing information [5], and it can effectively selectively receive part of the information when the model learns and receives a large amount of information, and the structure is shown in Figure 4:

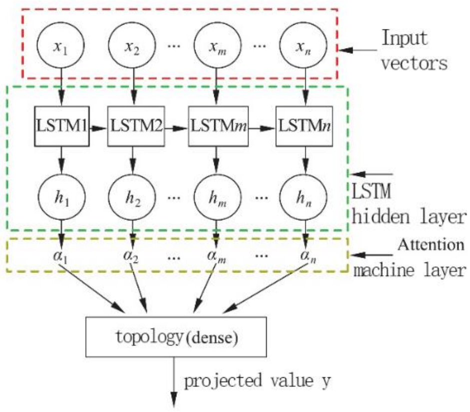

Figure 4. ATT-LSTM structure

3. Research design

3.1. Basic processing of sample data

This paper uses the data of CSI3 This paper uses the data of CSI300 index constituents are historical daily data, containing three major categories of a total of 31 characteristic indicators, the time interval is from January 1, 2018 to April 1, 2024, the indicator data are from tushare financial data interface, and the use of Python software for the subsequent study. 00 index constituents are historical daily data, the Containing a total of 31 characteristic indicators in three categories, the time interval is from January 1, 2018 to April 1, 2024, and the indicator data are sourced from tushare financial data interface, using Python software for subsequent research.

Cumulative/distribution points are obtained based on relevant data and formulae; dividend yields are missing for some stocks, which are treated as zero for companies with excessively long dividend distributions; negative P/E ratios according to tushare are treated as missing values for removal.

3.2. LSTM neural network modeling applications

Based on the relevant theoretical learning, it can be seen that LSTM solves the gradient disappearance and gradient explosion problem of RNN model, which is more suitable for the prediction of time-series data, so this paper constructs the LSTM neural network model to carry out the subsequent research.

1)Model Input and Output Settings

In this paper, Keras is used to build an LSTM neural network model. The inputs to the model are three-dimensional tensors are: number of samples, the number of samples in each training batch; time step, the number of time steps or sequence lengths for each sample; and number of features, the number of input features to the model. The output of the model is the predicted value of the closing price of the stock versus the logarithmic return.

2)Selection of hyperparameters

In this paper, based on literature study and practical attempts, the LSTM layer is finally determined to be 1 layer, the number of hidden units is 50, the time step is 1, the number of samples being fed in each training is 32, and the number of training rounds is 100. other hyperparameters are the default settings in Keras. The Dropout layer is also created to reduce the possibility of overfitting, with a parameter of 0.2 i.e., there is a 20% probability of randomly dropping the output of a neuron at each time step during training.

3)Model construction

In this paper, we build LSTM, Bi-LSTM and ATT-LSTM models, the input features are all basic trading information, the input feature processing methods are all normalization, and the predicted values are closing price and logarithmic yield.

3.3. Portfolio Construction

In this paper, portfolio construction is performed based on predicted stock closing prices, MACD is calculated and trading signals are identified, and then 100,000 simulations are performed using Monte Carlo method to determine the investment weights of each stock.

3.3.1. Principles for determining trading signals

Smoothed Array of Moving Averages (MACD): Uses the difference between short-term (12-day) and long-term (26-day) exponential moving averages to determine overbought and oversold points in the market, as well as market turnaround points.

For the closing price of day t is set as C_t, the moving average is denoted as EMA_t, the deviation is defined as $MACD\_DIF_t$, and the relationship between them is as follows:

$$EMA(n)_t = EMA(n)_{t-1} \times \frac{n-1}{n+1} + C_t \times \frac{2}{n+1} \quad (2)$$

$$MACD\_DIF_t = EMA(n)_t - EMA(m)_t \quad (3)$$

Based on the calculation of the h day EMA of $MACD\_DIF_t$, i.e., the mean value of the deviation $MACD\_DEA(h)_t$:

$$MACD\_DEA(h) = MACD\_DEA(h)_{t-1} \times \frac{h-1}{h+1} + MACD\_DIF_t \times \frac{2}{h+1} \quad (4)$$

Finally, the $MACD$ of day t is obtained, i.e. $MACD_t$:

$$MACD_t = 2 \times (MACD\_DIF_t - MACD\_DEA_t) \quad (5)$$

In summary, combining the dual conditions of EMA and MACD, the trading signal conditions are shown below:

Buy signal:

$$\begin{cases} EMA_t \geq EMA_{t-1} \\ MACD_t > 0 \\ MACD_{t-1} < 0 \end{cases} \quad (6)$$

Sell Signal:

$$\begin{cases} EMA_t \leq EMA_{t-1} \\ MACD_t < 0 \\ MACD_{t-1} > 0 \end{cases} \quad (7)$$

3.3.2. Determination of investment weights

In this paper, Monte Carlo method is used to simulate the weights and calculate the effective frontier curve of the portfolio under each weight [6], where: r_i is the return of the i stock; $\bar{r}_i$ is the expected return of the ith stock; σ_i is the standard deviation of r_i; $P[r_1,\ldots, r_n]$ is the risky portfolio, which consists of n risky assets according to a certain weight ratio; The weight of each asset in P is w_i and satisfies the following requirement:

$$\sum_{t=1}^{n} w_i = 1 \quad (8)$$

Assume that the market is completely free to trade without restriction, $w_i \in R$. Based on the returns on individual assets, the return on the portfolio asset P can be calculated as:

$$r_p = \sum_{i=1}^{n} w_i r_i \quad (9)$$

The expected return and variance of the portfolio is:

$$E(r_p) = E(\sum_{i=1}^{n} w_i r_i) = \sum_{i=1}^{n} w_i E(r_i) = \sum_{i=1}^{n} w_i \bar{r}_i (10)$$

$$Var(r_p) = E[r_p - E(r_p)] = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j Cov(r_i, r_j) \quad (11)$$

Construct the MPT model: assume that the fixed expected return is μ , the weights $w = (w_1, w_2, \ldots, w_n)$, and the portfolio of risky assets P satisfies that the variance of P is minimized among all the portfolios that can achieve the expected return μ, i.e:

$$\min_w Var(r_p) = \min_w \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j Cov(r_i, r_j) \quad (12)$$

$$s.t. \begin{cases} \sum_{i=1}^{n} w_i \bar{r}_i = \mu \\ \sum_{i=1}^{n} w_i = 1 \end{cases} \quad (13)$$

4. Empirical studies

4.1. Model evaluation based on entropy weight TOPSIS method

In this section, the monthly average data (March-April 2024) of CSI 300 index constituents are normalized and then principal component analysis (KOM value of 0.775, p-value of 0.000) is used to downsize 31 indicators into 10 indicators (explaining more than 90% of the information of the overall indicators), and then entropy-weighted TOPSIS is used to screen out five stocks for LSTM prediction afterwards [7].

4.1.1. Entropy weighting

In the first step, the normalized data for each principal component (X_{ij}^*) is used to calculate the weight of the value of the i stock under the j indicator in the sum of all values (P_{ij}).

$$P_{ij} = \frac{X_{ij}^*}{\sum_{i=1}^{n} X_{ij}^*} \quad (14)$$

In the second step, the information entropy value of the j indicator (e_j) is calculated.

$$e_j = -k * \sum_{n=1}^{n} (P_{ij} - lnP_{ij}) \quad (15)$$

In the third step, the information entropy redundancy (d_j) is calculated.

$$d_j = 1 - e_j \quad (16)$$

In the fourth step, the weights of the evaluation indicators are calculated (W_j).

$$W_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \quad (17)$$

The specific weights of each principal component are shown in Table 1 below.

Table 1. Summary of the results of weight calculation by entropy method

Term	information entropy$e$	information utility value$d$	weighting factor$w$
Pca10	0.9955	0.0045	4.85%
Pca9	0.9939	0.0061	6.60%
Pca8	0.9876	0.0124	13.46%
Pca7	0.9933	0.0067	7.27%
Pca6	0.9922	0.0078	8.49%
Pca5	0.9951	0.0049	5.28%
Pca4	0.9944	0.0056	6.04%
Pca3	0.9901	0.0099	10.74%
Pca2	0.9987	0.0013	1.38%
Pca1	0.9670	0.0330	35.90%

4.1.2. TOPSIS

In the first step, the weighted normalized decision matrix (Z_{ij}) is determined.

$$Z_{ij} = W_j X_{ij}^* \tag{18}$$

In the second step, determine the positive ideal solution (Z_j^+) and the negative ideal solution (Z_j^-).

$$\begin{cases} Z_j^+ = max\{Z_{ij}\} \quad (j = 1,2,\dots 10) \\ Z_j^- = min\{Z_{ij}\} \quad (j = 1,2,\dots 10) \end{cases} \tag{19}$$

In the third step, the distance of the evaluation object from the positive and negative ideal solutions is calculated (D_i^+), (D_i^-).

$$\begin{cases} D_i^+ = \sqrt{\sum_{j=1}^{m} \left(Z_{ij} - Z_j^+\right)^2} \\ D_i^- = \sqrt{\sum_{j=1}^{m} \left(Z_{ij} - Z_j^-\right)^2} \end{cases} \tag{20}$$

In the fourth step, the closeness of the evaluation object to the positive and negative ideal solutions is calculated (C_i).

$$C_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{21}$$

The results of the TOPSIS evaluation calculations are shown in Table 2.

Table 2. TOPSIS evaluation calculations

Term	positive ideal solution distanceD^+	Negative ideal solution distanceD^-	relative proximity$C$	Sorting results
Subject of evaluation55	1.498	13.628	0.901	1
Subject of evaluation104	8.068	5.880	0.422	2
Subject of evaluation12	8.210	5.646	0.407	3
Subject of evaluation49	8.637	5.234	0.377	4
Subject of evaluation141	9.060	4.779	0.345	5

Find out the stock names of the corresponding rows based on the serial numbers of the evaluation objects, i.e., evaluation object 55 is Guizhou Moutai, evaluation object 104 is Kingsoft Office, evaluation object 12 is Beifang Huachuang, evaluation object 49 is Gujing Gongjiu, and evaluation object 141 is Ningde Times.

4.2. LSTM models and comparisons

According to the feature importance ranking graph, open, high, low, pre_close, change, pct_chg, vol, amount, Ln_return, are used as input features to the LSTM model (where pct_chg is the raw data of yield). Because there is a Dropout layer to prevent overfitting, the program is run more often to find more accurate model results, and the time is in reverse order because the data is pulled directly from tushare. Here, the Guizhou Maotai stock is used as an example to compare the three models. The results are shown in Figure 5, Figure 6 and Figure 7.

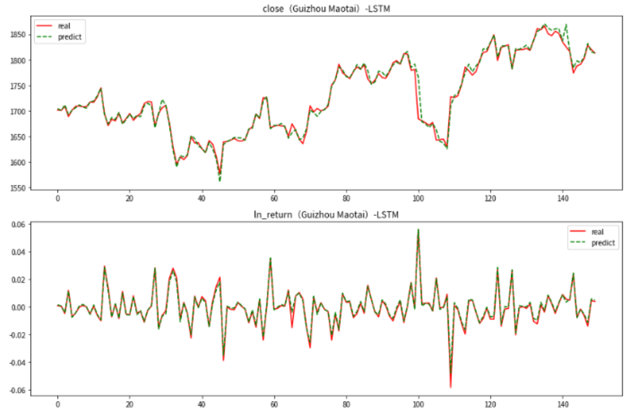

Figure 5. LSTM model to predict the closing price and logarithmic yield of Guizhou Moutai

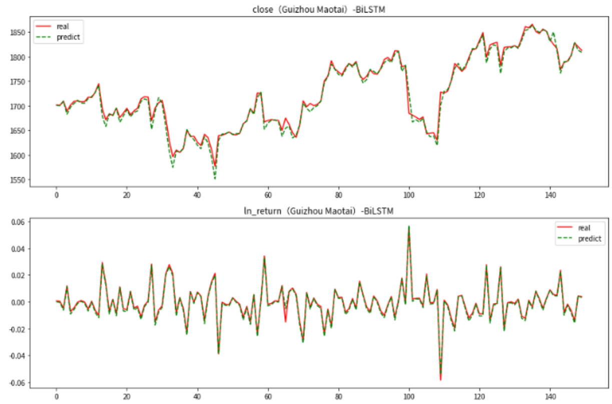

Figure 6. Bi-LSTM model to predict the closing price and logarithmic yield of Guizhou Moutai

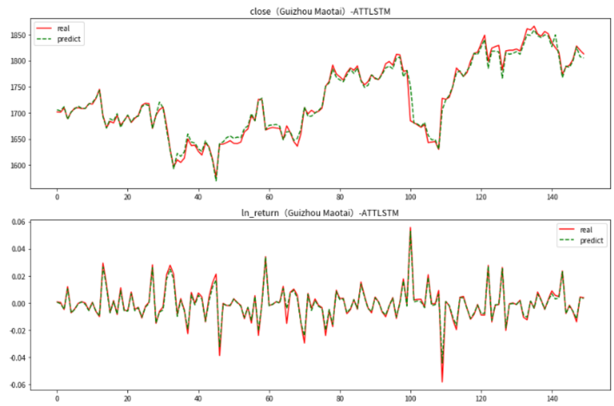

Figure 7. ATT-LSTM model to predict the closing price and logarithmic yield of Guizhou Moutai

According to the four evaluation indexes of MAE, RMSE, MAPE and R2 of the three models, the comparison results of the evaluation indexes of each model in Table 3 are obtained.

Combining the three prediction models, Bi-LSTM and ATT-LSTM model prediction and evaluation results are better than the LSTM model, which indicates that the addition of bidirectional and attention mechanism layer makes the neural network prediction more accurate, and further comparison reveals that the Bi-LSTM model prediction and evaluation results are the best among the three models, so using Bi-LSTM model prediction data for portfolio construction.

Table 3. Evaluation of LSTM, Bi-LSTM and ATT-LSTM model prediction for Kweichow Moutai

Model	Target audience	Average absolute error MAE	Rms error$RMSE$	Mean absolute percentage error$MAPE$	Coefficient of determinationR^2
LSTM	closing price	5.9449	9.0114	0.34%	0.9839
Bi-LSTM	closing price	4.2523	7.1815	0.25%	0.9898
ATT-LSTM	closing price	5.1381	7.4302	0.30%	0.9891
LSTM	logarithmic yield	0.0015	0.0024	55.26%	0.9653
Bi-LSTM	logarithmic yield	0.0008	0.0014	22.14%	0.9891
ATT-LSTM	logarithmic yield	0.0011	0.0017	21.82%	0.9828

5. Conclusions

Through this study, we validate the superiority of stock portfolio models based on Bi-LSTM neural networks in stock price prediction and portfolio optimisation. Empirical analyses show that the Bi-LSTM model significantly outperforms the traditional LSTM and ATT-LSTM models in terms of prediction accuracy when dealing with complex time series data. Combined with the dual-conditional trading strategy of EMA and MACD, the predicted data based on the Bi-LSTM model achieves higher investment returns with fewer trades, which further proves the effectiveness of the method in practical applications.

This study not only enriches the content of portfolio theory, but also provides investors with new tools to make scientific investment decisions in a complex market environment. By introducing deep learning technology, we effectively improve the accuracy of stock price prediction and optimise the portfolio construction method. This provides financial market participants with new ideas to help cope with the increasingly complex investment environment.

Future research can validate the robustness of the model over a wider dataset and longer time horizon, while exploring more applications of deep learning models in the financial market. Through continuous improvement and innovation, it is believed that deep learning technology will play a greater role in the financial field.

References

[1] Yan Wenxin. Application of machine learning based in stock prediction[J]. Information Systems Engineering,2024, (04):40-43.

[2] Zhu Huainian,Chen Zhuoyang,Binning.Robust nonzero and stochastic differential portfolio games for two persons under Heston model[J/OL]. Journal of Sun Yat-sen University (Natural Science Edition) (in Chinese and English),1-12[2024-06-04]. https: // doi. org/ 10. 13471/ j. cnki. acta. snus.2022A062.

[3] Yang Chao,Mao Junkui,Yang Yue,et al. Multi-stage turbine transition state tip clearance prediction based on long and short-term memory neural network[J/OL]. Propulsion Technology,1-12[2024-06-04] .https://doi.org/10.13675/j.cnki.tjjs.2403004.

[4] Yu Qianqian,Zhang Lei,Hu Xiaoxiang. Research on screening algorithm of abnormal power usage behavior based on CNN and Bi-LSTM[J]. Electronic Design Engineering, 2024, 32(09):96-100. DOI:10.14022/j.issn1674-6236.2024.09.020.

[5] Chenxi Wang. Research on ultra-short-term wind power prediction based on IWHO-CNN-ATT-LSTM[D]. Liaoning University of Engineering and Technology, 2023. DOI: 10. 27210/ d. cnki. glnju. 2023.000039.

[6] Q. Zhang,L. Wang,J.S. Xing. Optimization of GRU neural network based on Monte Carlo method for cogeneration load forecasting[J]. Journal of Beihua University(Natural Science Edition),2 024, 25(04):545-551.

[7] REN Peng,LI Chong,TAO Peng,et al. Vulnerability assessment of power grid nodes based on weighted entropy TOPSIS method[J]. Journal of Electric Power Science and Technology, 2019, 34(03):143-149.

DOI:10.19781/j.issn.1673-9140.2019.03.017.