WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n3.50
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Combination Forecast of Soybean Futures Data of Dalian Commodity Exchange based on Entropy Weight Method

Jun Wang

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China

ABSTRACT

As financial markets continue to develop, the need for accurate predictions in the futures market is growing. This paper takes the soybean futures of Dalian Commodity Exchange (DCE) as the research object, and uses the induced ordered weighted average (IOWA) operator combined with time series analysis and machine learning methods to construct a combined forecast model. First, multiple single prediction models are established through preprocessing and feature extraction of historical data. Secondly, the IOWA operator is used to integrate the results of each single prediction model to optimize the prediction performance. Experimental results show that compared with traditional prediction methods, the combined prediction model proposed in this paper has significant advantages in prediction accuracy and stability. In addition, this model can effectively capture market dynamics and provide investors with more reliable decision-making support. This article not only enriches the theoretical research on futures market forecasting, but also provides new perspectives and tools for practical operations.

KEYWORDS

Entropy weight method; ARIMA model; LSTM; Combined prediction

1. INTRODUCTION

As an important part of the financial market, the futures market plays a key role in discovering prices, avoiding risks, and allocating resources. Dalian Commodity Exchange (DCE) is an important agricultural product futures trading market in China, and its soybean futures products have attracted widespread attention from domestic and foreign investors. Accurately predicting the trend of soybean futures prices is of great practical significance for market participants to formulate trading strategies, risk management, and for policymakers to conduct market supervision.

However, because the futures market is affected by various factors such as economic conditions, supply and demand, policy changes, market sentiment, etc., futures prices show a high degree of volatility and complexity, which brings great challenges to predictions. Traditional forecasting methods often rely on a single statistical model or econometric model, which is difficult to fully capture the multifaceted nature and dynamic changes of market information. Therefore, there are major limitations in practical applications.

In recent years, combined forecasting methods have gained increasing attention due to their ability to synthesize information from multiple single forecasting models. In particular, the proposal of the induced ordered weighted average (IOWA) operator provides a new idea for nonlinear combination prediction. The IOWA operator introduces the concept of induced value and performs orderly weighting according to the relative importance of each single prediction model at different moments, thereby effectively improving the accuracy and adaptability of the combined prediction.

In view of this, this article aims to explore the combined prediction model of soybean futures data of Dalian Commodity Exchange based on IOWA operator. Through in-depth analysis of historical data, a single prediction model is constructed by combining time series analysis and machine learning technology, and then the IOWA operator is used to integrate the prediction results of each model in order to obtain more accurate and robust prediction results. This research is not only expected to improve the theoretical level of soybean futures price forecasting, but also provide practical decision-making reference for market participants. It has important practical value for improving the risk management mechanism of my country's futures market and improving market forecasting capabilities.

2. LITERATURE REVIEW

Combination forecasting is an effective method to improve forecast accuracy. With the rapid development of big data and artificial intelligence technology, combined forecasting models have shown great potential in many fields such as finance, energy, and transportation. Wang Lin, Liang Jinyu, Zhou Ying et al. (2023) proposed a prediction method for China's export container freight index based on the LSTM-ARIMA combination model. This method combines the LSTM model's ability to process time series data with the ARIMA model's seasonality and trend capturing capabilities, improving the accuracy and reliability of forecasts. Guo Yuchen, Jia Heping, Yu Tao et al. (2023) proposed a carbon price prediction method based on the CNN-LSTM combined model. This method uses the CNN model to extract the characteristics of carbon price data and combines it with the LSTM model to predict carbon prices, achieving good prediction results. Zhang Jian, Zhang Ye et al. (2023) proposed a GRU-LSTM dynamic prediction method for college students' employment based on multi-feature fusion. This method combines GRU and LSTM models, as well as a variety of employment-related features, to improve the accuracy and reliability of dynamic prediction of college students' employment. Wang Xinxin, Shen Xiaopan, Wang Qi et al. (2023) proposed a waterway freight volume prediction method based on LSTM-RBF. This method combines the processing ability of the LSTM model for time series data and the nonlinear fitting ability of the RBF network to improve the accuracy and reliability of waterway freight volume prediction. Zhao Hengzhe, Yang Xiaoying, Shi Yan et al. (2023) proposed a CNC machine tool spindle bearing fault prediction method based on XGBoost-LSTM. This method combines the XGBoost model's ability to classify fault data and the LSTM model's ability to process time series data, improving the accuracy and reliability of fault prediction. Shen Lulu, Huang Jinhao, Hua Min et al. (2024) proposed a solar energy prediction method based on ARIMA-PSO-LSTM. This method combines the ARIMA model's ability to process time series data, the PSO algorithm's ability to optimize model parameters, and the LSTM model's ability to process time series data, improving the accuracy and reliability of solar energy prediction.

3. THEORETICAL MODEL

3.1. ARIMA Model

The ARIMA model is to obtain a stationary sequence by differential processing of time series data, and then perform autoregression, moving average and other operations on the sequence to obtain the model [3]. Describe the statistical properties and trends of time series data by combining autoregressive (AR) and moving average (MA). The three components of the ARIMA model are expressed as AR(p), I(d), MA(q) respectively, where:

AR(p) represents the autoregressive part, which uses the observed values of the past p moments of time series data to predict the value of the current moment. The AR model is based on the

autocorrelation relationship between time series data, assuming that the value at the current moment is related to the value at the previous p moments.

I(d) represents the difference part. By performing d difference operations on non-stationary time series data, it can be converted into a stationary time series for better analysis and modeling. Stationary time series are easier to analyze and model, and differencing operations can remove trends and seasonality in time series data.

MA(q) represents the moving average part, which uses the prediction errors of the first q moments of time series data to predict the value of the current moment. The MA model is based on the moving average relationship between time series data, assuming that the value at the current moment is related to the prediction error of the previous q moments. The parameters p, d, and q of the ARIMA model represent the autoregressive order, the difference order, and the moving average order respectively, and are used to describe and model the characteristics and patterns of time series data. The ARIMA (p, d, q) model formula is:

$$(1 - \varphi_1 L - \varphi_2 L^2 - \ldots - \varphi_p L^p)(1 - L)^{i-d} Y_t = (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_p L^p)\varepsilon_t$$

Among them, Y_t represents the value of the time series data, L represents the lag operator (that is, $L^i Y_t = Y_t^{-i}$), ε_t represents the white noise error, φ_1, φ_2, ..., φ_p represent the autoregressive coefficients, the moving average coefficients are represented by θ_1, θ_2, ..., θ_p, the autoregressive order is represented by p, the difference order is represented by d, and the moving average order is represented by q.

3.2. LSTM Model

LSTM is a special recurrent neural network (RNN) that is designed to avoid long-term dependency problems and gradient explosion or vanishing problems, which is very important for capturing long-term dynamic relationships in time series data. Remembering information for a long time is the default behavior of LSTM, rather than requiring special efforts as in traditional RNNs. All LSTM units have three gates, namely the forget gate, input gate, and output gate. The role of these gates is to enable the LSTM unit to choose to forget the previous state, choose to update the current state, and choose the output value.

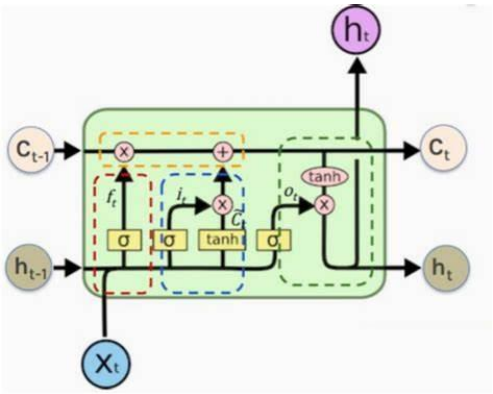

Figure 1. LSTM internal structure diagram

3.3. Forget Gate Structure

First, the current time step input x(t) is concatenated with the previous time step hidden state h(t-1) to obtain $[h\_(t-1), x\_t]$, then transformed through a fully connected layer, and finally activated by the sigmoid function to obtain f(t). We can regard f(t) as the gate value, like the size of a door opening and closing. The gate value will act on the tensor passing through the door. The forget gate value will act on the cell state of the previous layer, representing how much past information is forgotten.

Because the forget gate value is calculated by x(t), h(t-1), the whole formula means that the current time step input and the previous time step hidden state h(t-1) determine how much past information carried by the cell state of the previous layer is forgotten. The formula is:

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

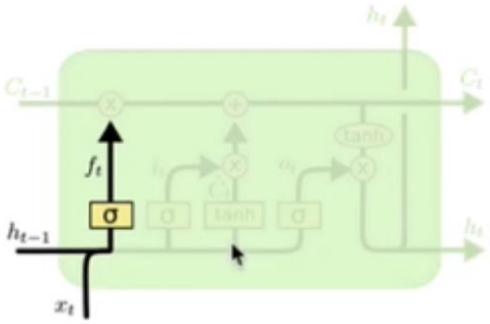

Figure 2. Internal structure of the forget gate

3.4. Input Gate Structure

We can see that there are two calculation formulas for the input gate. The first one is the formula for generating the input gate value. It is almost the same as the forget gate formula. The difference lies in the target they will act on later. This formula means how much input information needs to be filtered. The second formula of the input gate is the same as the internal structure calculation of the traditional RNN. For LSTM, it gets the current cell state, not the implicit state like the classic RNN. Formula:

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

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

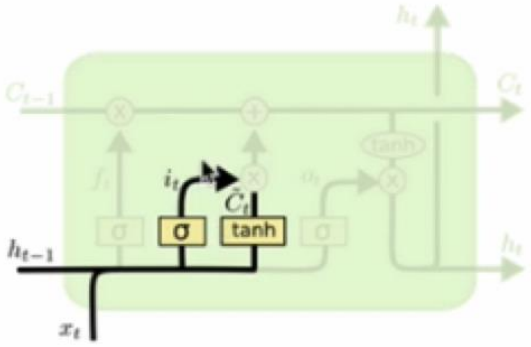

Figure 3. Internal structure of the input gate

3.5. Cell State Update

The structure and calculation formula of cell update are very easy to understand. There is no fully connected layer here. It just multiplies the forget gate value just obtained by C(t-1) obtained in the previous time step, and then adds the result of multiplying the input gate value by the unupdated C̃_t obtained in the current time step. Finally, the updated C(t) is obtained as part of the input of the next time step. The entire cell state update process is the application of the forget gate and the input gate. The formula is:

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

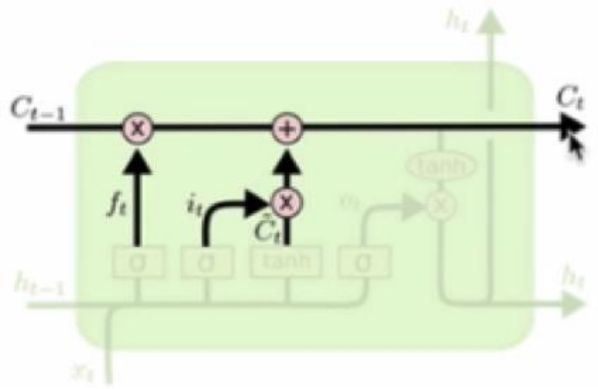

Figure 4. Cell state structure

3.6. Output Gate Structure

There are also two formulas for the output gate. The first is to calculate the gate value of the output gate, which is calculated in the same way as the forget gate and input gate. The second is to use this gate value to generate the hidden state h(t), which will act on the updated cell state C(t) and perform tanh activation, and finally obtain h(t) as part of the input of the next time step. The entire output gate process is to generate the hidden state h(t). The formula is:

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

$$h_t = o_t \cdot \tanh(C_t)$$

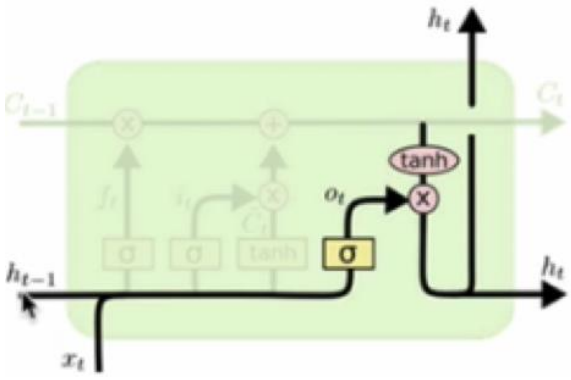

Figure 5. Output gate internal structure

3.7. Entropy Weight Method

In information theory, entropy is a measure of the disorder or chaos of a system. Information is interpreted as a reduction in the disorder of a system, and information is expressed as the variability of a certain indicator of the system. That is, the larger the entropy value of the system, the smaller the amount of information it contains. The smaller the variability of a certain indicator of the system. Conversely, the smaller the entropy value of the system, the greater the amount of information it contains, and the greater the variability of a certain indicator of the system. The steps for determining the coefficients using the entropy method for combined prediction are as follows:

(1) Calculate the relative error proportion P_{it} of the prediction of the tth period of the single prediction method in the i-th:

Where, $\sum_{t=1}^{N} P_{it} = 1, i = 1,2,\cdots,m$

(2) Calculate the entropy value of the relative error of the single prediction method in the i-th item:

$$h_i = -k \sum_{t=1}^{N} P_{it} \ln P_{it}, i = 1,2,\cdots,m$$

Where k is a constant, ln is the natural logarithm, $h_i \geqslant 0, i = 1,2,\cdots,m$

(3) Calculate the degree of variation d_i of the forecast relative error sequence of the i-th single forecast method

$$d_i = 1 - h_i, i = 1,2,\cdots,m$$

(4) Calculate the weighted coefficients $l_1, l_2, \cdots, l_m$ of various prediction methods:

$$l_i = \frac{1}{m-1}\left(1 - \frac{d_i}{\sum_{i=1}^{m} d_i}\right)$$

(5) Calculate the combined prediction value:

$$\hat{x}_\iota \sum_{i=1}^{m} l_i x_{\dot{u}}, t = 1,2,\cdots,N$$

4. EMPIRICAL RESEARCH

4.1. Data Source and Description

This paper selects the soybean futures data of Dalian Commodity Exchange from March 22, 2018 to May 31, 2024 for prediction, with no missing values. The data comes from the financial data terminal of Oriental Fortune Choice.

4.2. ARIMA Model Prediction

(1) Model selection

First, we used EViews software to draw a time series graph and conduct a preliminary visual analysis of the data. As shown in the figure, the analysis results show that the original sequence presents an unsteady characteristic. Specifically, the original data shows a gradually rising trend over time, and gradually tends to fluctuate up and down at a certain point from 2022 to 2024.

Figure 6. Time series graph of raw data

After the ADF test, we failed to reject the null hypothesis that there is a unit root in the time series data. This means that the time series data may be non-stationary. To solve this problem, we processed the original data with the first-order logarithmic difference. The result after processing is shown in

the figure. The time series data after the first-order logarithmic difference meets the stationary requirement. The model can be further established, which provides us with a more accurate basis for analysis and prediction.

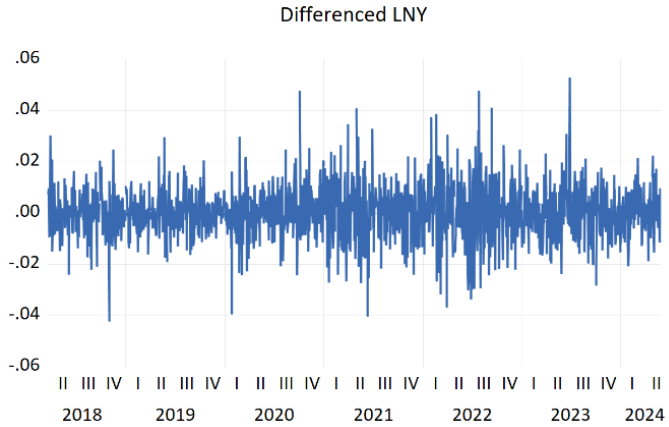

Figure 7. Logarithmic first-order difference plot

By analyzing the autocorrelation and partial autocorrelation plots after logarithmic difference, we found that both autocorrelation and partial autocorrelation showed first-order censoring. Therefore, we established an ARIMA (1,1,1) model and the results showed that at least within 10 % significance level, therefore, we construct an ARIMA (1,1,1) model.

Dependent Variable: DLNY
Method: ARMA Maximum Likelihood (OPG - BHHH)
Date: 06/03/24 Time: 08:47
Sample: 3/23/2018 5/31/2024
Included observations: 1500
Convergence achieved after 24 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000454	0.000266	1.709330	0.0876
AR(1)	-0.720743	0.368089	-1.958069	0.0504
MA(1)	0.700107	0.378162	1.851339	0.0643
SIGMASQ	0.000108	2.84E-06	37.98831	0.0000

R-squared	0.000886	Mean dependent var	0.000454
Adjusted R-squared	-0.001117	S.D. dependent var	0.010402
S.E. of regression	0.010408	Akaike info criterion	-6.289911
Sum squared resid	0.162042	Schwarz criterion	-6.275742
Log likelihood	4721.433	Hannan-Quinn criter.	-6.284633
F-statistic	0.442382	Durbin-Watson stat	1.973463
Prob(F-statistic)	0.722731		

Inverted AR Roots	-.72
Inverted MA Roots	-.70

Figure 8. ARIMA model results

(2) Residual test

An excellent fitting model can extract all relevant information from the observation data, that is, the residual sequence is a white noise sequence. In this article, we conduct a significance test on the residual sequence of the model. The results show that the residual sequence of the model The sequence is a white noise sequence, therefore, we can reject the null hypothesis and the model has extracted all valid information. That is, the model fit is good.

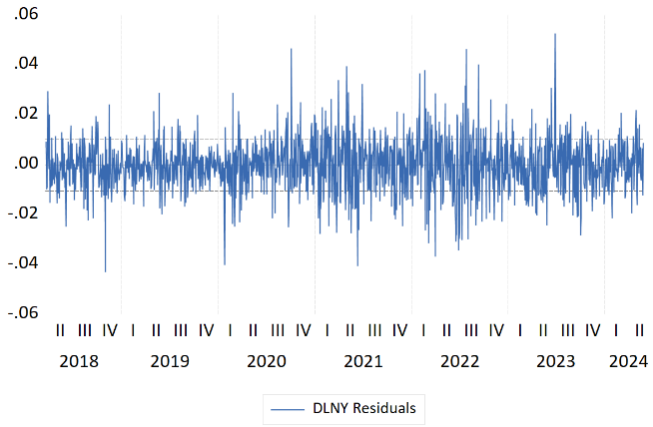

Figure 9. Residual sequence diagram

We make a prediction diagram of the original data and the predicted data, and we can see that the model prediction results are good.

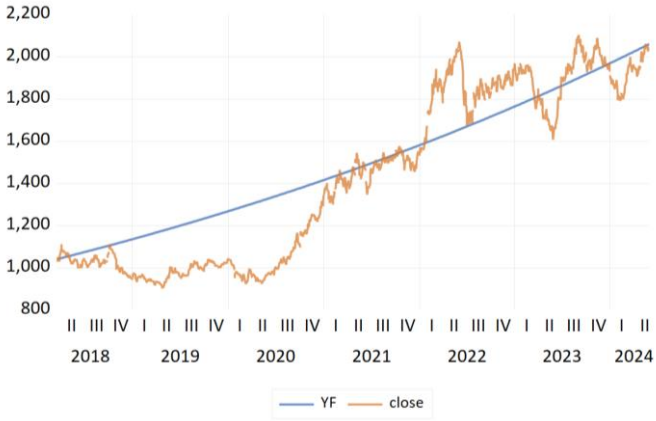

Figure 10. Schematic diagram of fitting results

(3) Model prediction accuracy

Based on the above model, we get the model prediction accuracy and model prediction error

Table 1. ARIMA model prediction error rate

Date	Forecasted value	Actual value	Forecast error
5/28/2024	2054.987	2053.49	0.07%
5/29/2024	2055.921	2047.82	0.40%
5/30/2024	2056.855	2024.24	1.61%
5/31/2024	2057.789	2043.25	0.71%

4.3. LSTM Model Prediction

This paper selects the soybean futures data of Dalian Commodity Exchange from March 22, 2018 to May 31, 2024 for prediction, with no missing values. Among them, the test set and the training set are divided into 8:2 ratios, and the model output variable is the closing price. This experiment uses the Tensorflow framework to establish an LSTM model.

(1) Data preprocessing

Before inputting the data, this paper performs standard processing on the closing price data of Dalian Commodity Exchange soybean commodity futures and maps it to [0, 1]. The formula is:

$$x_i = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Among them, x_i represents the standardized data; x_{max} and $x_{\min}$ are the data of each closing price in the futures closing price data set after different production data are standardized to eliminate the dimension effect and make the data comparable.

(2) Network training

In this paper, we use the LSTM neural network model to predict the soybean futures price of Dalian Commodity Exchange. The model contains two layers of neurons, with 32 neurons in the first layer and 50 neurons in the second layer. In order to evaluate the prediction performance of the model, we use the prediction error analysis method to measure the degree of fit of the model by comparing the difference between the predicted value and the actual value. In each training process, the model is iterated 100 times to reach the convergence state and stabilize the prediction results.

We divide the data set into training set and test set in a ratio of 8:2, that is, 1200 training samples and 301 test samples. In addition, in order to avoid overfitting during the model training process, we use the dropout mechanism. In this way, the model can generalize better during the training process and improve the accuracy of the prediction.

The prediction results are shown in the figure. It can be seen from the figure that the model training results are good and the fitting effect is good. This shows that the LSTM neural network model has a higher predictive performance in predicting the prices of soybean futures on the Dalian Commodity Exchange.

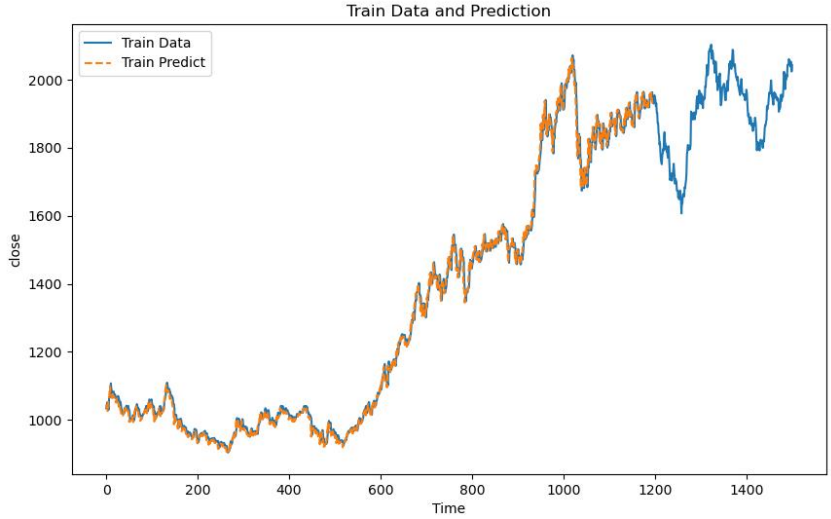

Figure 11. Comparison between test set and actual value

(3) Model prediction accuracy

Based on the above model, we obtain the model prediction accuracy and model prediction error as shown in the following table.

Table 2. LSTM model prediction error rate

Date	Forecasted value	Actual value	Forecast error
5/28/2024	2053.49	2026.863	1.30%
5/29/2024	2047.82	2038.88	0.44%
5/30/2024	2024.24	2033.742	0.47%
5/31/2024	2043.25	2014.243	1.42%

(4) Combination model prediction

In the field of predictive analysis, a single prediction model often fails to fully demonstrate its advantages, especially when faced with complex and volatile data. To overcome this limitation, researchers usually adopt a combination prediction method to obtain more accurate and robust prediction results by integrating the prediction results of multiple models.

The core idea of combined prediction is to combine multiple prediction models and take advantage of their respective advantages to improve the accuracy and reliability of the prediction as a whole. In this method, each model contributes to the final prediction result, and the final result is based on the combination of all model predictions.

By analyzing the fitting effect and error of the ARIMA model and the LSTM model, the entropy weight method is used to construct a combined prediction model to improve the prediction accuracy.

(5) Comparison of prediction model effects

In predictive analysis, a single prediction model may not be able to fully capture the complexity and variability of the data. Therefore, the accuracy of the prediction can be improved by combining multiple prediction models. This paper uses the entropy weight method to combine the ARIMA model and the LSTM model for prediction, successfully combining the advantages of the two models and improving the accuracy and reliability of the prediction. The results of this paper show that the combined prediction effect is the best and is better than the single prediction model. This shows that the combined prediction of ARIMA model and LSTM model by entropy weight method can better capture the complexity and variability of data, thereby improving the accuracy and reliability of prediction.

Table 3. Evaluation of the effect of combined model and single model

Serial number	True value	ARIMA	LSTM	Combined forecast	ARIMA Error Value	LSTM Error value	Combined prediction error value
1	1940.16	1913.598	2034.551675	1949.203841	4.87%	-1.37%	0.47%
2	1956.15	1915.124	2035.476125	1950.552847	4.06%	-2.10%	-0.29%
3	1978.96	1942.658	2040.104681	1971.34358	3.09%	-1.83%	-0.38%
4	2001.34	1972.138	2045.672842	1993.784777	2.22%	-1.46%	-0.38%
5	2015.04	1990.086	2046.602345	2006.722783	1.57%	-1.24%	-0.41%
6	2008.47	1996.726	2047.532271	2011.682321	1.94%	-0.58%	0.16%
7	2043.25	2040.454	2057.789383	2045.55683	0.71%	-0.14%	0.11%

5. CONCLUSION

This article conducts an in-depth analysis of the closing price data of soybean futures on the Dalian Commodity Exchange and uses the entropy weight method to conduct a combined forecast of the ARIMA model and the LSTM model. The research results show that the combined forecast of ARIMA model and LSTM model through entropy weight method can significantly improve the forecast accuracy and provide more reliable and accurate forecast results for the soybean futures market. In addition, the research methods of this article also provide useful references and inspirations

for other time series forecasting problems. In future research, the application of combined forecast models in different markets and forecast scenarios can be further explored, with a view to providing investors and market participants with more comprehensive decision-making support.

REFERENCES

[1] Jin Feifei, Chen Huayou, Zhou Ligang, et al. Optimal combination prediction model of IOWA operator based on maximum-minimum closeness [J]. Practice and Understanding of Mathematics, 2013(07):112-118.

[2] Wang Lin, Liang Jinyu, Zhou Ying. Forecast of China's export container freight index based on LSTM-ARIMA combination model [C]//China Association for Science and Technology, Ministry of Transport, Chinese Academy of Engineering, Hubei Provincial People's Government. 2023 World Transportation Proceedings of the Conference (WTC2023) (Volume 2). People's Communications Press Co., Ltd., 2023: 6. DOI: 10.26914/c.cnkihy.2023.019493.

[3] Guo Yuchen, Jia Heping, Yu Tao, et al. Carbon price prediction method based on CNN-LSTM combined model [J]. Science and Technology Management Research, 2023, 43(11): 200-206.

[4] Zhang Jian, Zhang Ye. GRU-LSTM dynamic prediction of college students' employment based on multi-feature fusion [J]. Computer Science, 2023, 50(S1):916-921.

[5] Wang Xinxin, Shen Xiaopan, Wang Qi, et al. Waterway freight volume prediction based on LSTM-RBF [J]. Science, Technology and Engineering, 2023, 23(18): 7995-8001.

[6] Zhao Hengzhe, Yang Xiaoying, Shi Yan, et al. Research on CNC machine tool spindle bearing failure prediction method based on XGBoost-LSTM [J]. Modern Manufacturing Engineering, 2023, (08): 155-160.DOI: 10.16731/j.cnki .1671-3133.2023.08.022.

[7] Lu Zhiping, Yu Xiaojing, Lu Chengyu. New energy vehicle sales forecast method integrating VMD and LSTM models [J]. Journal of Wuhan University of Technology (Information and Management Engineering Edition), 2023, 45(04): 546-551 .

[8] Fan Yuanjie, Sui Yiping, Zhang Lei, et al. Shanxi thermal coal price prediction based on LSTM-SVR combined model [J]. Mining Research and Development, 2024, 44(04): 252-258. DOI: 10.13827/j .cnki.kyyk.2024.04.032.

[9] Shen Lulu, Huang Jinhao, Hua Min, et al. Solar energy prediction based on ARIMA-PSO-LSTM [J/OL]. Radio Communication Technology, 1-11 [2024-06-20]. http://kns.cnki.net/kcms/detail/13.1099.TN.20240618.0938.008.html