Prediction of Fund Net Value Based on ARIMA-LSTM Hybrid Model

Peng Zhou, Fangyi Li
zhou.peng@student.zy.cdut.edu.cn, li.fangyi1@student.zy.cdut.edu.cn

Chengdu University of Technology Chengdu, China

Abstract—The net value of fund is affected in many ways, and researchers attempt to quantify these influences in order to predict future net value by developing various models. Current prediction models typically can only reflect the linear variation law, and their nonlinear characteristics are either poorly handled or selectively ignored, resulting in less accurate prediction results. Based on this, the ARIMA-LSTM hybrid model is used in this paper to predict funds. After preprocessing historical data, the ARIMA model is used to filter out the linear data characteristics, followed by the LSTM model to extract the nonlinear characteristics by residual, and finally superposition the respective prediction values of the two models was performed to obtain the hybrid model's prediction results. The paper's methodologies are empirically proven to be more accurate and applicable than typical fund forecast methods.

Keywords-ARIMA model; LSTM model; net fund value; time series

1 INTRODUCTION

The incorporation of financial education and investment into the national education system gradually raises public understanding of financial management [1]. Simultaneously, as a result of the epidemic's impact, market interest rates continue to fall, making the fund market with low-risk high-returns one of the most popular financial options for investors [2]. According to data from cnstock, the average yield of the 2020 partial stock hybrid funds was about 58 percent, and the average yield of equity funds was 60 percent, both of which are the highest levels in nearly 11 years. With its good performance, the fund is very engaging in market trading. Because of the change in the fund's net value is an important aspect in determining the fund's profitability, predicting fund income using net value has also become a popular study area in recent years. In this backdrop, a plethora of research on this field has been launched by several professionals and researchers. Chen Jianing utilized wavelet analysis to estimate the revenue of an Internet money fund [3], Xiang Ying and others used the ARIMA model to predict the fund net value [4], and Meng Guoying attempted to apply machine learning to fund performance prediction [5]. However, because many factors influence the fund in the market, most of these models simply pursue the linear rule of net value change, ignoring the impact of the fund's nonlinear change on prediction accuracy. Based on this, due to the fact that ARIMA

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2326870

model is good at handling linear characteristics in time series and the LSTM model is good at dealing with nonlinear difficulties [6], this paper presents an ARIMA-LSTM hybrid model that combines the two advantages. In comparison, the experimental results showed that the ARIMA-LSTM hybrid model outperforms the standard single model in terms of fitting performance and prediction accuracy.

2 ALGORITHM PRINCIPLE

2.1 ARIMA model

Autoregressive Integrated Moving Averaged Model is a commonly used time series forecasting method. The core idea of the model is to find a suitable mathematical function to fit the linear relationship between the current time value, the past time value, and the random interference amount to infer the future value through the past value [7]. The essence of the ARIMA model is an improvement of the ARMA model, and its mathematical formula is:

$$\begin{aligned} A_t &= \varphi_1 A_{t-1} + \varphi_2 A_{t-2} + \ldots + \varphi_p A_{t-p} + \varepsilon_t \\ &\quad -\theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_z \varepsilon_{t-z} + \varepsilon_t \\ &\quad -\theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_z \varepsilon_{t-z} \end{aligned} \tag{1}$$

Where ε_t is the residual and A_t is a stable time series. It should be noted that the ARIMA model can process only stable time series. If the time series is unstable, it needs to be transformed into a stable series by difference. The processed model is called the ARIMA model, denoted as $ARIMA(p,d,q)$.

Both p and q represent the order of the model. When p=0, the model degenerates to a q-order MA model MA(q), and when q=0, the model degenerates to a p-order AR model AR(p). In addition, d indicates how many differences have passed.

The basic steps of ARIMA model modelling are: analyzing the ADF value of the series, determining the (p, d, q) value of the model, estimating the correlation coefficient of MA and AR, testing the white noise series, and creating a prediction model.

2.2 LSTM model

In order to solve the exploding gradient and gradient disappearance of the recurrent neural network (RNN) during the operation, Hochreiter and Schmidhuber proposed an improved method for the recurrent neural network, namely the LSTM neural network model (Long Short-Term Memory) [8]. Unlike the RNN model, the LSTM model resets a cell state in the original hidden layer to preserve long-term memory. The LSTM structure is shown in Figure 1. The internal structure of the model is mainly composed of three control gates: input gate, forget gate and output gate. It is worth noting that tanh is the activation function, C_{t-1} and C_t represent the cell state at $t-1$ and t, respectively. h_t and h_{t-1} are the hidden states of the

cell at t and $t-1$.

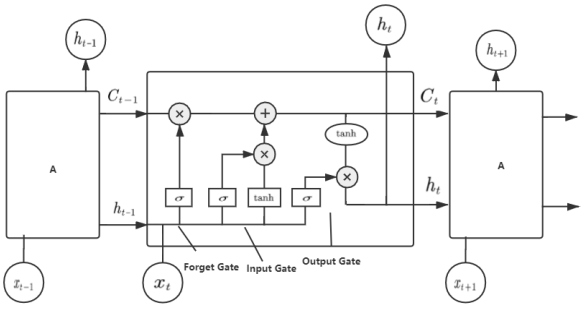

Figure 1. Structure diagram of LSTM model

First, the hidden state h_{t-1} at the time can be determined through the forget gate of the model, and the degree of information retention of input x_t can also be determined. The formula is:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

Second, you can determine how much content in the input variable can be stored in the cell state C_t through the input gate. The formula is:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_{\mathrm{i}}) \tag{3}$$

$$\widetilde{C_t} = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C_t} \tag{5}$$

Finally, the output gate of LSTM outputs the hidden state of each cell, the formula is:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = O_t \cdot \tanh(C_t) \tag{7}$$

In the above formula, W_f, W_i, W_c, W_o are the weight matrix of different control gates; b_f, b_i, b_c, b_o are the bias term of each control gate; C_t and tanh are the corresponding activation functions, which express how much information passes through the different control gates.

3 ARIMA-LSTM FORECASTING MODEL

Changes in fund net worth usually have strong nonlinear and irregular [9], and predictions using only a single model often yield poor results. This paper based on the fund's net value, using the ARIMA-LSTM hybrid model, filter the linear features with the ARIMA model, and then give the nonlinear characteristics stored to the LSTM model for processing, which can ensure the linear and nonlinear characteristics of the data. Finally, we combine the prediction results of two models to obtain the prediction results of the hybrid model. See Figure 2 for its flowchart.

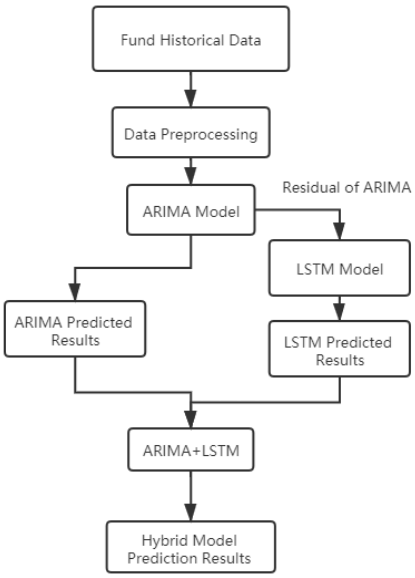

Figure 2. Prediction flow chart of ARIMA-LSTM hybrid model

The time series y_t of the fund's net value can be regarded as consisting of a linear structure L_t and a non-linear structure N_t. The mathematical formula is:

$$y_t = L_t + N_t \tag{8}$$

First, we use the ARIMA model to predict the linear part of the series and get the predicted value $\widehat{L_t}$. Then subtract $\widehat{L_t}$ from the true value to get the residual series e_t.

$$e_t = y_t - \widehat{L_t} \tag{9}$$

Use the LSTM model to process the series obtained in the previous step to predict the non-linear part of the fund's net value to obtain the predicted value $\widehat{N_t}$.

$$\widehat{N_t} = f(e_{t-1}, e_{t-2}, ..., e_{t-m}) + \varepsilon_t \tag{10}$$

Finally, the ARIMA-LSTM hybrid model's predicted value equals the sum of the two-step predicted values.

$$\hat{y} = \widehat{L_t} + \widehat{N_t} \tag{11}$$

4 EXAMPLE ANALYSIS

4.1 Data processing

In this paper, we select the 1260-day fund net value data of Huabao Hybrid Fund (240008) from June 6, 2016, to July 30, 2021. The data used is derived from the historical net value of funds that have been published on the Tiantian Fund Website. The 1260-day data is divided into three parts, as shown in Table 1, for different model training processes.

Table 1. Classification of training data

Code	*Size*	*Train Size*	*Val Size*	*Test Size*
240008	1260	900	100	260

At the same time, this paper uses the sliding window prediction method [10], as shown in Figure 3. Assuming L is the length of the window to be trained, starting from the leftmost time t, this model predicts the next days' net value and continues to move forward one day until it reaches the rightmost time T. It should be noted that the model will only predict the net value on the next day and will use L-length historical data for analysis before predicting.

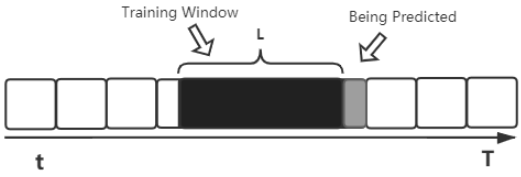

Figure 3 Sliding window online prediction method

4.2 Evaluation indicators

This article selects three common error evaluation indicators to evaluate the prediction accuracy of different models [11]. These three indicators are MSE, MAE, RMSE, and the following are their respective mathematical expressions:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2 \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y_i} - y_i\right| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2} \tag{14}$$

The three indicators have the following characteristics: the smaller the value, the smaller the error between the value predicted by the model and the true value, which means the higher accuracy.

4.3 Result analysis

4.3.1 Constructing ARIMA model

Figure 4 is the original time series chart of the fund. It is not difficult to see that the data changes drastically, and there is no obvious change rule. Besides, the chart rises sharply after 1000 days, proving that the series is a non-stable series, so the difference method is needed to convert the original sequence into a stable series. Figure 5 is the series diagram after the first-order difference processing. It can be seen that the processed series is more stable than the original series. Through the ADF test, it can be determined that the series can become stable after the first-order difference (d=1)

Figure 4. Time series dlagram of fund net value

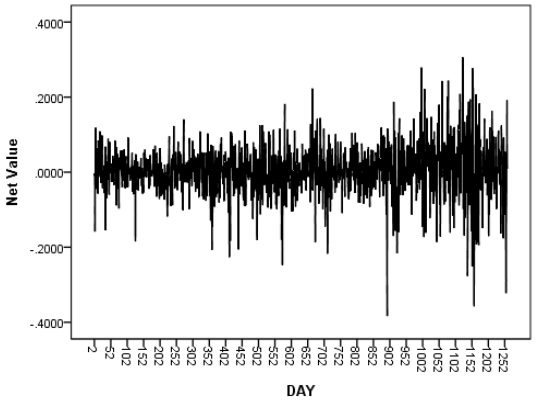

Figure 5. Time series dlagram after FIRST-ORDER difference

The parameters p and q of the ARIMA model can be inferred from the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) diagrams of the series. Figures 6 and 7 are the ACF and PACF diagrams of the fund's net value. After analysis, it is found that the ACF diagram is lagging the first-order truncation, and the PACF diagram is also the first-order truncation. In order to improve the accuracy of the model, this article refers to the AIC values of different (p, d, q) combinations, and the AIC value of the combination (0, 1, 0) is the smallest, which is -2851, indicating the model fit created under this combination of Highest degree and best prediction effect.

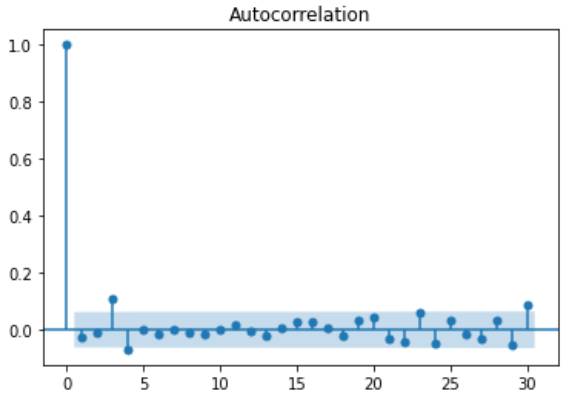

Figure 6. ACF dlagram

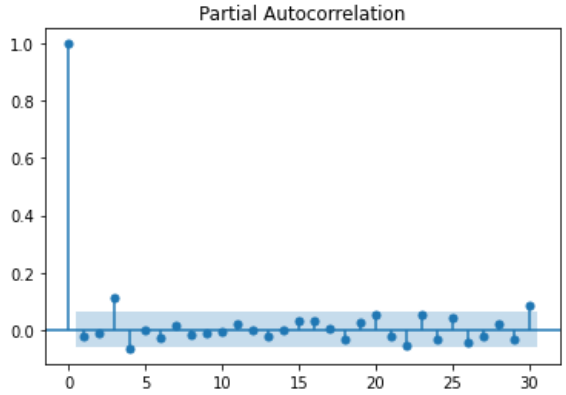

Figure 7: PACF diagram

According to the experimental data, the prediction formula of the ARIMA model (p=0, d=1, q=0) can be written as:

$$(1-L)y_t = \alpha_0 + \varepsilon_t \tag{15}$$

$$y_t = \alpha_0 + y_{t-1} + \varepsilon_t \tag{16}$$

If only use ARIMA single model to predict the fund's net value, the obtained model prediction chart is shown in Figure 8. It can be seen that the ARIMA model has a low prediction accuracy of the fund's net value after 1000 days, which means the model is not suitable for use in actual fund forecasting

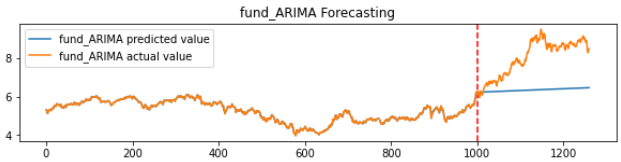

Figure 8. Net value forecast chart of arima model

4.3.2 LSTM processing ARIMA residuals

Finally, use Python to build a suitable LSTM model. After many parameter adjustments, the final model structure is determined as follows: the number of layers is 3, the input and output dimensions are both set to 1, the learning rate is 0.005, the training iteration is 100 times, and the batch_size is 64. Figure 9 and Figure 10 show the fit of the target fund and the forecast of future net value using the hybrid model. It can be seen that the value predicted by the ARIMA-LSTM hybrid model is roughly the same as the real trend, and the degree of fit is significantly better than that of the ARIMA model.

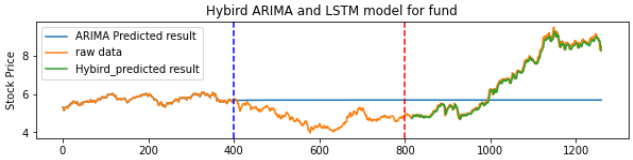

Figure 9. Fitting dlagram of hybrid model

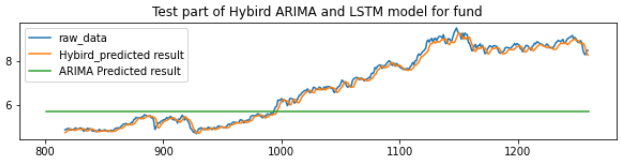

Figure 10. Prediction dlagram of ARIMA-LSTM hybrid model

4.3.3 Comparison of three prediction indicators

Table 2 displays the prediction outcomes of the three models based on the identical study data. The statistics show that the LSTM model's MSE, MAE, and RMSE values are significantly lower than those of the ARIMA model, and the ARIMA-LSTM hybrid model's values are also lower than those of the other two single models. The quantitative examination of the indicators revealed that the ARIMA model has the worst prediction result, the LSTM model has a better prediction effect, and the hybrid model has the best forecast effect. To sum up, the ranking of the prediction effects of the three models from high to low is the ARIMA-LSTM hybrid model, the LSTM model and the ARIMA model. In conclusion, the ARIMA-LSTM hybrid model is a more reliable time series analysis model, which is more suitable for predicting the fund's net value in real life than the independent model.

Table 2. Error indicators table of prediction results

Model	***MSE***	***MAE***	***RMSE***
ARIMA	3.61	1.69	1.90
LSTM	0.13	0.31	0.36
ARIMA-LSTM	0.01	0.09	0.12

5 CONCLUSION

The fund's net value change has both linear and non-linear features. Traditional methods for predicting net value are difficult to use due to non-linearity, resulting in low forecast accuracy. Although machine learning prediction approaches have considerable advantages when dealing with nonlinear issues, they are prone to overfitting when dealing with limited data samples, resulting in low prediction accuracy. The hybrid model separates the two traits and combines their benefits. It excels at dealing with complex time series difficulties, such as the fund's net worth, and has shown to be a more dependable analytical and forecasting tool. However,

whether the model can solve the gradient problem in the face of longer time series should be studied in the future.

Acknowledgments. We thank Chunmei Li for critical reading and comments during the preparation of the paper. We would also like to thank Ms.Wushuang from Chengdu University of Technology for her guidance on academic writing.

REFERENCES

[1] Lin Xiaoling. research on financial consumer education in the era of internet finance [J]. journal of Beijing agricultural vocational college, 2014,28(05):86-90.

[2] Kuang Yicheng, Qu Bo. Study on the Impact of COVID-19 Epidemic on China's Securities Investment Funds [J]. China Price, 2021(02):74-77.

[3] Chen Jianing. Forecast Analysis of Internet Money Found Return Based on Wavelet Analysis: Taking Yu'ebao as an example [D]. Shandong University Technology, 2019.

[4] Xiang Ying, Wang Yaping. Application of ARIMA model in fund net value prediction [J]. Neijiang Science and Technology, 2009,30(11):130+139.

[5] Meng Guoyao. Performance prediction model of private equity fund based on machine learning [D]. Southwestern University of Finance and Economics, 2019.

[6] Xiang Xudong, Hu Xueping, Yu Xiaomei, Deng Xin. Prediction of Antarctic monthly mean surface temperature based on ARIMA model and LSTM model [J]. Journal of Science of Teachers' College and University, 2021,41(06):33-37.

[7] Aladağ Erdinç. Forecasting of particulate matter with a hybrid ARIMA model based on wavelet transformation and seasonal adjustment[J]. Urban Climate, 2021, 39.

[8] Gao Wenjun, Li Zhi, Min Xing, Zhang Yulong. Response Time Forecasting of Application System Based on ARIMA-LSTM Hybrid Model [J]. Computer and Digital Engineering, 2021,49(05):880-885.

[9] Mou Yu. The Variation and Forecast Analysis of the Net Value of Fund Unit Based on H-P Filtering Method—— Taking E-Fund Defense Mixed Military Industry Fund as an example [J]. Management and Technology of Small and Medium-sized Enterprises (next issue), 2021(06):104-107.

[10] L. Mohamad Hanapi et al. A Novel Fuzzy Linear Regression Sliding Window GARCH Model for Time-Series Forecasting[J]. Applied Sciences, 2020, 10(6).

[11] Guangqi Qiu and Yingkui Gu and Junjie Chen. Selective health indicator for bearings ensemble remaining useful life prediction with genetic algorithm and Weibull proportional hazards model[J]. Measurement, 2020, 150.