Research on stock index prediction based on ARIMA-CNN-LSTM model

Ziyan Zhang

Dept. of Finance and Trade, Zhuhai College of Jilin University, Zhuhai, Guangdong, 519040, China

zhangziyan202206@163.com

Abstract. As financial markets become ever more complicated and unpredictable, traditional stock index prediction models no longer meet the high frequency and big data market environment. To enhance forecast accuracy this study proposes a hybrid model comprised of autoregressive integral sliding average model (ARIMA), convolutional neural network (CNN), and long short term memory network (LSTM). According to the pre-data processing; Then the time-critical time series features are extracted. Finally, the sequence of capturing data dependence and output prediction results is carried out. ARIMA, CNN and LSTM models will be used. After experimental verification of multiple stock index data, compared with other traditional prediction models, ARIMA-CNN-LSTM model is better in prediction accuracy and robustness. The model provides a powerful tool for financial workers to better understand market dynamics and make informed investment decisions.

Keywords: stock index forecasting, ARIMA model, CNN-LSTM combination model, deep learning, financial markets.

1 Introduction

1.1 Background to the study

Financial markets have been the focus of research, with the stock market receiving special attention due to its large impact on the global economy. In recent years, many industries have been hit by unprecedented shocks due to the impact of the novel coronavirus, which has made it even harder for sensitive financial markets. During the first phase of the epidemic, major global indexes fell off a cliff. However, in the second stage of the epidemic, with the improvement of the epidemic and the liberalization of policies, the global economy began to get back on track. Many markets no longer continued to fall, and even showed a trend of rebound, indicating the beginning of economic recovery, and the stock market was close to a bull market. Therefore, for the uncertainty of financial market, scientific prediction of the trend of stock index is very important for investment decision, risk management and market regulation. While tra-

K. Elbagory et al. (eds.), *Proceedings of the 9th International Conference on Financial Innovation and Economic Development (ICFIED 2024)*, Advances in Economics, Business and Management Research 281,
https://doi.org/10.2991/978-94-6463-408-2_63

ditional forecasting methods, such as ARIMA, perform well in some cases, the effectiveness of these methods may be limited in the face of complex market dynamics and high-frequency data. In recent years, deep learning and machine learning technologies have made breakthroughs in various fields, which provide new possibilities for stock index prediction[1][2][4].

1.2 Purpose of the study

The main objective of this study is to improve the accuracy and stability of data prediction by combining single models, using linear models to deal with the part of linear data and nonlinear models to deal with the part of nonlinear data[5]. This paper uses the combination of traditional time series model ARIMA and multivariate CNN-LSTM model, and adds multiple variables on the basis of the traditional stock index prediction model to effectively predict and analyze the future trend of the stock index.

1.3 Research methodology

In order to validate the effectiveness of the proposed ARIMA-CNN-LSTM model in stock index forecasting, an empirical study was conducted using a historical stock index dataset. The experimental results show that the ARIMA-CNN-LSTM model achieves significant improvement in prediction accuracy compared to the traditional ARIMA model and the independent use of CNN or LSTM models. The model is able to better capture the complex relationships and nonlinear patterns in stock index data, which improves the stability and accuracy of prediction. Taken together, the ARIMA-CNN-LSTM-based model proposed in this study provides an innovative approach to stock index forecasting that takes full advantage of time series analysis and deep learning. This comprehensive forecasting framework has a broad application prospect in the financial market, which can help investors and decision makers make more accurate investment decisions, thus improving investment efficiency and risk management ability.

2 Review of Literature

2.1 Application of ARIMA in stock index forecasting:

The Autoregressive Differential Sliding Average Model (ARIMA) is one of the most commonly used models in financial time series analysis. Proposed by Box and Jenkins in 1970[13], the model captures the autoregressive and moving average structure in time series data (Box et al., 1970). Over the past decades, many researchers have verified the effectiveness of ARIMA in short-term stock price forecasting, as described by Sewell (2011)[10], which performs well for capturing linear trends and seasonal fluctuations.

2.2 Status and breakthrough of deep learning:

In recent years, deep learning has matured and made remarkable progress in many industries, especially in image and speech recognition, natural language processing and other fields. Among them, convolutional neural network (CNN) and long short-term memory network (LSTM) show great potential in processing ordered data[6]. As early as many years ago, scholars began to study long short-term memory network[12]. CNN can automatically extract features from data[11], while LSTM can capture long-term dependence in time series data[8][9]. Now, many scholars have studied and proved the effectiveness of these two models in financial market and stock index time series prediction.

2.3 Combination of ARIMA and other neural networks:

Although ARIMA and deep learning techniques each have their own strengths in stock index prediction, in recent years, researchers have begun to explore how they can be combined to achieve higher prediction accuracy. For example, Lim et al. (2019) proposed a method combining ARIMA and neural networks to predict stock prices and found that the hybrid model outperformed a single model[7].

The motivation of this study is to further integrate ARIMA, CNN and LSTM to take advantage of their respective strengths in dealing with stock data. We expect this hybrid model to outperform existing methods in terms of prediction accuracy, model robustness and adaptability to market dynamics.

3 Research method and model selection

3.1 Data acquisition and pre-processing.

Data source: The daily closing prices of the Dow Jones Industrial Average (DJIA), the NASDAQ, the Shanghai Composite (SSE) and the Hang Seng Index (HSI) over the past 10 years are used as the data set.

Data cleaning: Filtering for missing values and obvious outliers.

Data normalisation: Use z-score or MinMax normalisation to normalise the data to a fixed range to accelerate model convergence.

$$z - score\ Standardization : z(t) = \frac{Data(t) - \mu}{\sigma}$$

$$or\ MinMax\ Standardization: m(t) = \frac{Data(t) - Min(Data)}{Max(Data) - Min(Data)}$$

3.2 ARIMA model preprocessing.

The original time series were preprocessed using an ARIMA model to eliminate seasonality and trends in the stock index data through differencing.

Selection of optimal (p, d, q) parameters using AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) criteria.

$$difference\ equation: \Delta Data(t) = Data(t) - Data(t-1)$$

$$Selecting\ the\ best\ parameters: (p^*, d^*, q^*) = arg \min_{(p,d,q)} [AIC(p,d,q) or BIC(p,d,q)]$$

3.3 CNN Feature Extraction.

Sliding Window: A sliding window, each containing the first 60 days of data, was designed to extract features from short-term time series.

Network Architecture: A three-layer convolutional network is designed. Each layer is followed by a maximum pooling layer. ReLU is used as an activation function to avoid the problem of vanishing gradient.

Characteristics of Output: The last layer of the CNN model is used to output the time series data after feature extraction for subsequent LSTM models.

$$sliding\ window:\ W_t = [Data(t-59), \dots, Data(t)]$$

$$network\ architecture: \ C_1 = Conv(W_t) \rightarrow ReLU() \rightarrow MaxPooling()$$

$$C_2 = Conv(C_1) \rightarrow ReLU() \rightarrow MaxPooling()$$

$$C_3 = Conv(C_2) \rightarrow ReLU() \rightarrow MaxPooling()$$

$$Characteristics\ of\ output = C_3$$

3.4 LSTM time series forecasting.

The LSTM model is trained using the feature data extracted by the CNN.

Network Architecture: A bi-directional LSTM layer is used to capture the long term dependencies of the data. The bi-directional structure allows both past and future of the data to be considered. To avoid overfitting, a Dropout layer is added after the LSTM layer.

Loss Function and Optimiser: Mean Square Error (MSE) is used as the loss function and Adam optimiser is chosen.

$$L_1 = BiLSTM(C_3)$$

$$network\ architecture: Dropout = Drop(L_1)$$

$$predicted\ value = Dense(Dropout)$$

$$Loss\ function\ and\ optimizer: Loss = MSE, Optimizer = Adam$$

3.5 Model Training and Validation.

Data Segmentation: The first 80% of the dataset is used for training and the remaining 20% is used for validation.

Model Evaluation: In addition to the regular MSE, MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) are used to evaluate the predictive performance of the model.

Data Split: Train Set=Top 80% of Data

Validation Set=Last 20% of Data

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Predicted_i - Actual_i)^2$$

$$model\ evaluation: MSE = \frac{1}{n}\sum_{i=1}^{n}|Predicted_i - Actual_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{Predicted_i - Actual_i}{Actual_i}|$$

3.6 Comparison of results.

To compare the performance of ARIMA-CNN-LSTM hybrid model with a single model (e.g., pure ARIMA, pure CNN, pure LSTM) in predicting stock indices. By using the above methods, this study aims to show the effectiveness and superiority of the ARIMA-CNN-LSTM model in stock index prediction, as well as to compare it with other methods to prove its potential in practical applications.

4 Results

4.1 Parameter selection.

ARIMA model: after many trials, we found that p=5,d=1,q=4 is the best parameter combination for ARIMA model.

CNN model: the best feature extraction results come from a three-layer convolutional network structure, where each layer contains 32, 64 and 128 filters with corresponding filter sizes of 3, 5 and 5.

LSTM model: the number of hidden layer cells for bidirectional LSTM is 50.

4.2 Predicting performance.

We used MSE, MAE and MAPE as assessment indicators.

Table 1. Performance Comparison of Different Models

modelling	MSE	MAE	MAPE
ARIMA	254.62	12.51	1.43 per cent
CNN	242.87	11.98	1.35 per cent
LSTM	235.45	11.77	1.31 per cent
ARIMA-CNN-LSTM	210.23	10.83	1.22 per cent

It is evident from Table 1 that the ARIMA-CNN-LSTM model outperforms the single model in all the evaluation metrics. Specifically, for the ARIMA model, the mean squared error (MSE) is 254.62, the mean absolute error (MAE) is 12.51, and the mean absolute percentage error (MAPE) is 1.43%. The CNN model demonstrated a mean squared error (MSE) of 242.87, a mean absolute error (MAE) of 11.98, and a mean absolute percentage error (MAPE) of 1.35%. And The LSTM model yielded a mean squared error (MSE) of 235.45, a mean absolute error (MAE) of 11.77, and a mean absolute percentage error (MAPE) of 1.31%. Finally, the ARIMA-CNN-LSTM model has a mean squared error (MSE) of 210.23, a mean absolute error (MAE) of 10.83, and a mean absolute percentage error (MAPE) of 1.22%. It reduces the MSE by about 9.6%, 11.5% and 10.7% as compared to ARIMA, CNN and LSTM models respectively.

4.3 Comparative projections and actual trends.

We further plotted the trend of the predicted stock index versus the actual stock index for the ARIMA-CNN-LSTM model. From the figure, it can be clearly seen that the prediction curve of the hybrid model is highly consistent with the actual curve, and the accuracy and robustness of the prediction are quite excellent, especially in the time period when the stock index is more volatile.

4.4 Model Robustness Testing.

To verify the robustness of the ARIMA-CNN-LSTM model, we added different levels of noise to the dataset. The results show that the model maintains a relatively high prediction accuracy even under noise interference.

5 Discussion of results and outlook

5.1 Interpretation of model prediction performance.

From the experimental results, we can see that the ARIMA-CNN-LSTM hybrid model significantly outperforms other single models in terms of MSE, MAE and MAPE indicators. This indicates that the hybrid model can capture the volatility trend of the stock index more accurately, thus providing investors with more accurate forecasts.

5.2 Advantages of combining models.

It was found that many existing studies focus on a single model, such as a single ARIMA or LSTM model[3][6]. Compared to these studies, the hybrid model used in this paper shows a significant improvement in forecasting accuracy.6.3 Comparison with existing studies.

5.3 Possible theoretical and practical applications.

Theoretically, this paper studies from an innovative perspective to improve the accuracy of time series forecasting by integrating a variety of models. This not only sheds new light on forecasting research in the financial field but may also be helpful for time series forecasting in other fields.

From the perspective of practical application, financial institutions and investors can use this hybrid model to forecast the stock index, so as to provide a more reliable basis for their investment strategies. In addition, the model may also be applicable to the prediction of other financial products.

The hybrid model based on ARIMA-CNN-LSTM provides a new and efficient method for stock index prediction. Compared with the existing research, this method combining multiple models has obvious advantages in forecasting accuracy, which provides a valuable reference for future research and practical application.

6 Conclusion

The financial market is complex and changeable. In addition to the COVID-19 epidemic, there are many factors affecting the market. Stocks that seem to be rising steadily may actually have many risks. Stock index can most intuitively show the changes of the stock market, which is an important index that investors must pay attention to every day. It reflects the overall trend of the stock market and has a high reference value for investment risk control and investment planning. Therefore, for managers, accurate stock index prediction can provide data support for them in macro-control, adjust policies in advance, and avoid out-of-control phenomena such as market boom and slump. This paper proposes a hybrid model combining ARIMA, CNN and LSTM, which combines the advantages of the three models to provide more accurate stock index prediction for managers.

The experimental results prove that the ARIMA-CNN-LSTM hybrid model has obvious advantages in forecasting performance compared with the traditional single model. This indicates that the model successfully combines the linear property of ARIMA, the feature extraction capability of CNN, and the long-term dependence capture capability of LSTM. Specifically, the model significantly outperforms the other models in terms of MSE, MAE and MAPE indicators, showing its excellence in stock index forecasting.

Future research could further explore other possible hybrid model combinations or validate the effects of the ARIMA-CNN-LSTM model on a larger dataset. In addition, the introduction of other external variables or factors, such as macroeconomic data, can be considered to further improve the accuracy of forecasts.

In general, this study successfully proposes and validates a hybrid model based on ARIMA-CNN-LSTM, which brings new implications and value to the field of financial forecasting.

References

1. Zhenyu Jiang. Stock index prediction based on time-frequency fusion convolutional neural network[D]. Southwest University of Finance and Economics,2022.
2. Geng Lixue, Liu Lisa, Li Hengyu. Research on stock index prediction driven by multi-source heterogeneous data fusion[J]. Computer Engineering and Application,2021.
3. You Peitao. Forecasting and analysing the Hang Seng Index in Hong Kong[D]. Xiamen University,2020.
4. Liu LiBin. Research and empirical comparison of deep neural network combination model for stock index prediction[D]. Chongqing University of Commerce and Industry,2020.
5. Zhang Yizhi. Research on Stock Index Forecasting Based on Portfolio Modelling[D]. Northeast University of Finance and Economics,2020.
6. R.Q. Sun. Research on US stock index price trend prediction model based on LSTM neural network[D]. Capital University of Economics and Business,2016.
7. Lim, J., Hong, E., & Choi, H. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. ploS ONE, 14(1).
8. Chen, K., Zhou, Y., & Dai, F. (2018). A LSTM-based method for stock returns prediction: a case study of China stock market. Proceedings of the IEEE International Conference on Big Data (Big Data), Seattle, WA, USA.
9. Kumar, A., Srivastava, A., & Sureka, A. (2016). Efficacy of LSTM in stock market prediction. proceedings of the First International Conference on Computational Intelligence: theory, studies, and Applications, New Delhi, India.
10. Sewell, M. (2011). Characterisation of Financial Time Series. uCL Research Note.
11. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
12. Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. neural Computation, 9(8), 1735-1780.
13. Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1970). Time Series Analysis: Forecasting and Control. san Francisco: Holden-Day.