ISSN 2959-6130

Forecast Analysis of the Stock Market Based on Hidden Markov Model and Long Short-Term Memory Model - Taking the S&P500 Index as an Example

Die Hu

Abstract:

This paper proposes a hybrid method that combines the advantages of HMM and LSTM models to improve the accuracy of stock market predictions, using the S&P500 index as a case study. HMMs are used to identify and model potential market states, while LSTMs predict future stock prices based on these states. Through this integration, this paper hopes to take advantage of HMM's ability to model market conditions and LSTM's advantages in forecasting to provide a more robust forecasting framework. This article is mainly divided into five parts. The first part is an introduction to the research background and objectives. The second part is the review and arrangement of relevant literature. The third part is the elaboration of the theoretical basis and the collection and analysis of data. The fourth part is an explanation of the data analysis results. The fifth part is the conclusion and suggestions of this study.

Keywords: Stock Market, Hidden Markov Model, HMM and LSTM models, prediction, S&P500 Index

1 Introduction

Stock price prediction is an important task in the financial field. It aims to predict future stock price trends by analyzing multi-dimensional data such as market trends, company performance, industry dynamics and macroeconomic factors, and using statistics, machine learning and artificial intelligence technologies (Vijh et al., 2020). This process not only requires an in-depth understanding of the market, but also requires processing a large amount of information to identify the key factors affecting stock prices, thereby providing decision-making support to investors, helping them formulate more precise investment strategies, optimize risk management, and ultimately realize asset realization (Soni et al., 2022). Traditional financial models, such as ARIMA models or linear regression, have been widely used in stock price predictions. However, these models often struggle to capture the nonlinear and time-varying characteristics of financial data, as these data are affected by multiple factors such as market sentiment, economic indicators, and emergencies (Islam & Nguyen, 2020).

Furthermore, among numerous machine learning methods, the long short-term memory (LSTM) model, as a type of recurrent neural network (RNN), has been widely used in time series forecasting due to its ability to effectively process sequence data and preserve long-term dependencies. Despite the significant advantages of LSTM, relying solely on LSTM may not adequately capture underlying market states or the stochastic nature of financial time series (Selvin et al., 2017). This is where Hidden Markov Models (HMM) come into play. HMMs are particularly useful in modeling time series data with hidden states, enabling them to detect underlying market states such as bull or bear markets (Su & Yi, 2022). As one of the important indicators of financial market health, the S&P 500 Index reflects the overall performance of 500 leading public companies in the United States (Frino & Gallagher, 2001). Given its importance, effectively predicting the movement of the S&P 500 Index can provide valuable insights into market trends, thereby assisting strategic decision-making and risk management.

2 Literature Review

2.1 Application of LSTM in stock market prediction

Research by Wang et al. points out that although the stock market is very complex, effective prediction of stock market trends can reduce financial market risks. With the development of machine learning, in addition to traditional machine learning algorithms such as linear regression, polynomial regression, etc., recurrent neural networks such as long short-term memory and spoken short-term memory can also be used to predict time series of the stock market (Wang et al., 2021). Bhandari et al. used

a long-short-term memory model to predict the closing price of the S&P 500 Index for the next day. Based on macroeconomic data and fundamental market data, the authors construct an equilibrium portfolio of nine predictors to capture the behavior of the stock market. Their research found that single-layer LSTM has better fitting and more accurate predictions than multi-layer LSTM (Bhandari et al., 2022). However, LSTM also has some shortcomings. Ma comparatively analyzed the possibility of stock price prediction by the autoregressive moving average model, artificial neural network and long short-term memory model. The study found that although LSTM has the best prediction ability, it is greatly affected by data processing (Ma, 2020). Selvin et al. pointed out that although LSTM alleviates the problems of gradient disappearance and gradient explosion to a certain extent through the gating mechanism, these problems may still occur when processing very long sequences, especially when the model architecture or parameter selection is inappropriate (Selvin et al., 2017).

2.2 Application of HMM in stock market prediction

Gupta and Dhingra combined the time dependence, volatility and complex dependence of the stock market prediction problem and proposed the maximum a posteriori HMM method to predict the next day's stock price based on historical data. They considered fractional changes in stock value while training the model through the stock's intraday high and low values to achieve prediction results (Gupta & Dhingra, 2012). Hassan and Nath proposed a method to use HMM to predict stock prices in cross-correlated markets, specifically applied to the prediction of airline stocks. In this study, the researchers used a trained HMM based on the past data of the selected airlines. Then, the trained HMM is used to search for the behavioral pattern of the target variable in the historical data set, and predictions are made by interpolating the neighboring values of these data sets. The research results show that using HMM for stock market prediction has certain prospects, providing a new paradigm for the field of stock market prediction that has attracted much attention recently (Hassan & Nath, 2005). However, some scholars have suggested that the HMM model has flaws. Research by Rao et al. pointed out that since HMM is mainly used for pattern recognition and classification, when used for prediction, its accuracy is often not as good as specially designed time series prediction models, such as ARIMA, LSTM, etc. Especially in complex and noisy scenarios such as stock market predictions, the performance of HMM may be limited (Rao et al., 2020).

2.3 Combination of HMM and LSTM

Liu et al. applied the Hidden Markov Model to stock market prediction and explored four improvement methods: GMM-HMM, XGB-HMM, GMM-HMM+LSTM and XGB-HMM+LSTM. The study discussed the effects of these methods through experiments and analyzed the advantages and disadvantages of different models. Through experiments and analysis, this article determined that GMM-HMM+LSTM is the optimal model and proved the effectiveness of this model in stock market prediction and can provide investors with more accurate market timing judgments and strategic suggestions in practical applications (Liu et al., 2021). Hashish et al. proposed a novel hybrid model that combines hidden Markov models and long short-term memory networks to analyze cryptocurrency price fluctuations from a descriptive and predictive perspective. Specifically, HMM is used to describe the historical trend of cryptocurrencies, while LSTM is used to predict future price movements. To verify the effectiveness of the proposed model, the author selected 2-minute frequency Bitcoin data from the Coinbase trading market and compared it with the traditional time series prediction model ARIMA and the traditional LSTM model. The results show that the proposed hybrid model outperforms traditional forecasting methods in terms of forecast accuracy (Hashish et al., 2019).

2.4 Literature gap

Although both LSTM and HMM have their own advantages in stock market prediction, they still have certain limitations when used alone. Although LSTM can handle time series data, it has a strong dependence on data preprocessing, and gradient disappearance or explosion problems may still occur when processing very long sequences. HMM has advantages in pattern recognition and classification, but its prediction accuracy is not as good as modern models such as LSTM in complex stock market environments. Although existing research has explored hybrid models that combine the two, most studies only focus on a specific market or data set and lack in-depth exploration of the versatility and adaptability of hybrid models in different market environments. Therefore, further research on how to optimize the combination of HMM and LSTM in a wider market environment to improve prediction accuracy and model robustness is an important unresolved issue in the current literature.

3 Methodology

3.1 Research design

This study aims to predict the S&P 500 Index by combining hidden Markov models and long short-term memory

networks. The research data covers the period from January 1, 1994, to January 1, 2024. The research uses the Python programming language for data processing and model construction. The main goal of this research is to evaluate the performance of HMM combined with LSTM in stock market prediction and compare it with the prediction results using LSTM alone. By comparing the prediction effects of the two models, this study hopes to prove the advantages of the hybrid model in capturing market volatility and trend prediction. In the study, a variety of indicators will be used to evaluate the prediction effect of the model, including mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R^2). These indicators will help quantify the prediction error and fitting effect of the model and provide a basis for comparison of model performance.

3.2 Data collect and preparation

The data for this study comes from Yahoo Finance, and the data is downloaded through Python's yfinance library. The closing price (Close) of the S&P 500 Index (S&P 500) is selected, and its corresponding identifier is ^GSPC. The time range of the data is from January 1, 1994, to January 1, 2024, covering 30 years of historical data. In the data preparation stage, this paper needs to first preprocess the original data for subsequent model training and prediction. First, this study extracts the data of the Close column as a NumPy array and uses reshape (-1, 1) to convert it into a two-dimensional array. This is to meet the requirements of subsequent normalization processing and model input. Next, since neural network models such as LSTM are sensitive to the numerical range of the input data, the data needs to be normalized before inputting to the model. The purpose of normalization is to scale the data to a unified range, usually [0, 1], which helps improve the training effect and prediction accuracy of the model. This article uses MinMaxScaler to normalize the closing price data. The normalized data scaled_prices will be used as the input of the LSTM model to ensure that the model has better stability and convergence speed when processing data.

3.3 Data analysis

First, this paper builds an LSTM model. This study uses a 60-day time window to construct training data for the LSTM model. This study divides the data into training set and test set. 80% of the data is used for training and 20% of the data is used for testing, without disturbing the time sequence of the data to maintain the continuity of the time series. Then, this study constructs a model containing two layers of LSTM and one fully connected layer. The first layer of LSTM has 50 units and returns sequences; the second layer of LSTM also has 50 units but does not return sequences. Finally, a fully connected layer is used for output. Next, this study compiled using the Adam optimizer and mean square error (MSE) as the loss function and trained the model for 20 epochs. During the training process, 10% of the data is used for verification. This paper then predicts the test set and renormalizes the prediction results and actual values to restore the original price range. Finally, this paper calculated the mean square error (MSE), mean absolute error (MAE) and coefficient of determination (R^2) to evaluate the prediction effect of the model (Table 1), and drew a comparison chart between the actual price and the LSTM predicted price to intuitively show the prediction effect (Figure 1).

Table 1 Prediction effect of LSTM only

LSTM ONLY MSE	5528.27174991739
LSTM ONLY MAE	57.21070296864576
LSTM ONLY R^2	0.988393723456274

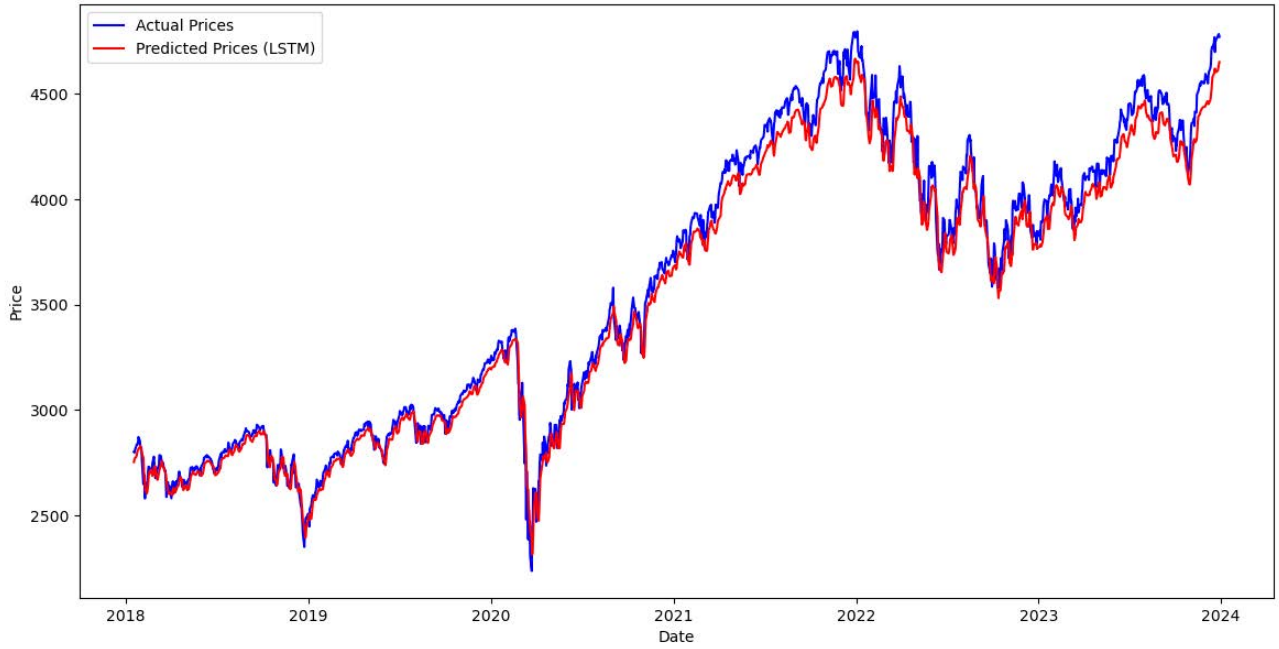

Figure 1 Actual price vs. predicted price of LSTM only

Then, this paper builds an HMM model. This paper uses Gaussian Hidden Markov Model (GaussianHMM) to train standardized price data, sets the number of states, and fits the model. According to the trained HMM, the hidden state is predicted, and then the hidden state obtained by the hidden Markov model is spliced with the standardized price data to generate new feature data. Next, this paper uses the trained HMM to infer the future time series and obtain the hidden state sequence. The hidden state sequence of HMM is used as an input feature of the LSTM model. In this way, LSTM can not only utilize traditional stock price time series data, but also consider changes in market conditions. Likewise, this paper calculated the mean square error (MSE), mean absolute error (MAE) and coefficient of determination (R^2) to evaluate the prediction effect of the combine model (Table 2), and drew a comparison chart between the actual price and the LSTM predicted price to intuitively show the prediction effect (Figure 2).

Table 2 Prediction effect of LSTM & HMM

LSTM+HMM MSE	3642.6475786215224
LSTM+HMM MAE	47.89082744616203
LSTM+HMM R^2	0.9923524788466762

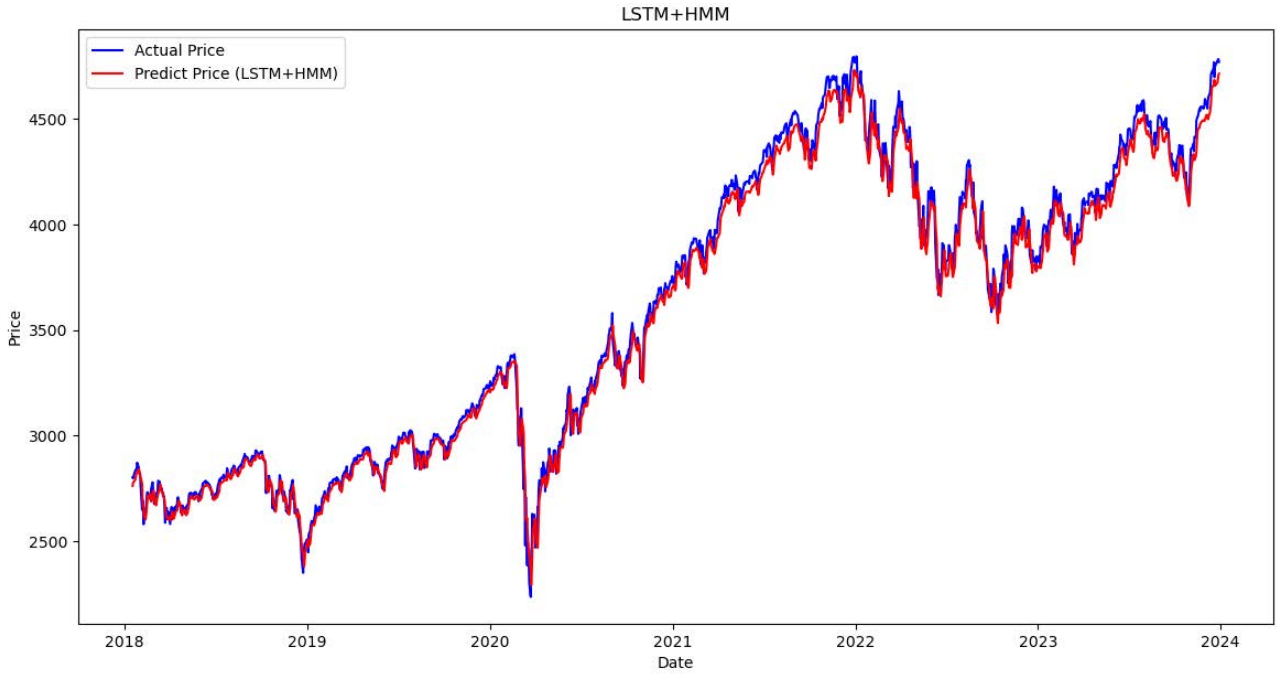

Figure 2 Actual price vs. predicted price of LSTM & HMM

3.4 Ethical issues

This research may involve some ethical issues that require attention. First, the complexity and black-box nature of the model makes its decision-making process difficult to explain and transparent. Investors may rely on these models to make major financial decisions, and the workings of the models are not understood by the average user. This information asymmetry may lead investors to rely too much on technology and ignore a comprehensive understanding and rational judgment of the market. Secondly, the potential systemic risks brought by model predictions cannot be ignored. If most market participants rely on similar prediction models to make decisions, market behavior may converge, thereby increasing market volatility and even triggering a financial crisis. Such risks require special attention in the context of highly interconnected global financial markets.

4 Discussion

It can be seen from the results that the joint model of LSTM and HMM shows better performance in all evaluation indicators than the model using only LSTM. Specifically: the MSE of the joint model of LSTM and HMM (3642.65) is significantly lower than the MSE of the LSTM model alone (5528.27). This shows that the joint model is more accurate in terms of prediction errors and better able to capture changing trends in the data. The MAE of the joint model (47.89) is also lower than the MAE of the LSTM-only model (57.21), indicating that the absolute deviation between the prediction results of the joint model and the actual value is smaller and the prediction accuracy is higher. The R^2 value of the joint LSTM and HMM model (0.992) is higher than the R^2 value of the LSTM only model (0.988). An increase in the R^2 value indicates that the joint model can explain more data variability and has better prediction results. Overall, the combined model of LSTM and HMM is better than the LSTM model alone in terms of prediction accuracy and model fitting effect. This shows that combining HMM with LSTM can better capture the hidden patterns in time series data, thereby improving the accuracy of prediction. This improvement has important practical significance for price prediction in the stock market and can provide a more accurate reference for investment decisions.

As can be seen from Figure 1, the LSTM model successfully tracks the actual price trend especially during periods when the trend is largely obvious. Nonetheless, in some areas with large fluctuations, the fluctuation range of the red curve is slightly smaller than the actual price, indicating that the LSTM model may have certain prediction bias in some cases. As can be seen in Figure 2, the joint LSTM and HMM model performs more closely in tracking actual prices, especially in areas with large price fluctuations. Compared with the first picture, the gap between the red curve and the blue curve is smaller, and the prediction effect is significantly better. In general, although the LSTM model can predict the price of the S&P500 index well, its performance is slightly insufficient when dealing with complex fluctuations and sudden changes. The LST-

M+HMM joint model performs better when predicting S&P500 index prices, especially in the case of complex fluctuations and sharp changes. The joint model can better capture the characteristics of price fluctuations and make the prediction results closer to the actual trend.

5 Conclusion & Recommendation

This study compares the prediction effect of the S&P500 index price using the LSTM model alone and the combined LSTM and HMM model. The results show that although the LSTM model can better capture the overall trend of market prices, its prediction effect is lacking when dealing with violent fluctuations and complex changes in prices. The joint model of LSTM and HMM significantly improves the accuracy of prediction by combining the temporal feature learning ability of LSTM and the hidden state modeling ability of HMM. Specifically, the joint model is better than the individual LSTM model in terms of mean square error (MSE), mean absolute error (MAE), and R^2 value, and has a higher degree of fit between the actual price trend and the predicted price.

Although the joint model of LSTM and HMM performs well, there is still room for improvement. It is recommended to explore more complex hybrid model structures, such as adding a convolutional neural network (CNN) to extract richer features or try a combination of variational autoencoders (VAE) and LSTM to further improve the prediction effect. In addition, the current model mainly relies on historical price data for prediction. In the future, you can consider introducing more external factors, such as macroeconomic indicators, market sentiment data, and global economic events, to enhance the model's sensitivity and predictive ability to market changes. Finally, although the joint model of LSTM and HMM has significantly improved the prediction accuracy, its complexity also increases the "black box" of the model. Future research can explore how to improve the interpretability of the model and help investors better understand the decision-making process of the model, so that they can trust the prediction results of the model more in practical applications.

References

Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., & Khatri, R. K. (2022). Predicting stock market index using LSTM. *Machine Learning with Applications*, *9*, 100320.

Frino, A., & Gallagher, D. (2001). Tracking S&P 500 index funds. *Journal of portfolio Management*, *28*(1).

Gupta, A., & Dhingra, B. (2012, March). Stock market prediction using hidden markov models. In *2012 Students Conference on Engineering and Systems* (pp. 1-4). IEEE.

Hashish, I. A., Forni, F., Andreotti, G., Facchinetti, T., & Darjani, S. (2019, September). A hybrid model for bitcoin prices prediction using hidden Markov models and optimized LSTM networks. In *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 721-728). IEEE.

Hassan, M. R., & Nath, B. (2005, September). Stock market forecasting using hidden Markov model: a new approach. In *5th international conference on intelligent systems design and applications (ISDA'05)* (pp. 192-196). IEEE.

Islam, M. R., & Nguyen, N. (2020). Comparison of financial models for stock price prediction. *Journal of Risk and Financial Management*, *13*(8), 181.

Liu, M., Huo, J., Wu, Y., & Wu, J. (2021). Stock market trend analysis using hidden Markov model and long short term memory. *arXiv preprint arXiv:2104.09700*.

Ma, Q. (2020). Comparison of ARIMA, ANN and LSTM for stock price prediction. In *E3S Web of Conferences* (Vol. 218, p. 01026). EDP Sciences.

Rao, P. S., Srinivas, K., & Mohan, A. K. (2020). A survey on stock market prediction using machine learning techniques. In *ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications* (pp. 923-931). Springer Singapore.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017, September). Stock price prediction using LSTM, RNN and CNN-sliding window model. In *2017 international conference on advances in computing, communications and informatics (icacci)* (pp. 1643-1647). IEEE.

Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction: a systematic review. In *Journal of Physics: Conference Series* (Vol. 2161, No. 1, p. 012065). IOP Publishing.

Su, Z., & Yi, B. (2022). Research on HMM-Based Efficient Stock Price Prediction. *Mobile Information Systems*, *2022*(1), 8124149.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia computer science*, 167, 599-606.

Wang, Q., Kang, K., Zhang, Z., & Cao, D. (2021). Application of LSTM and conv1d LSTM network in stock forecasting model. *Artificial Intelligence Advances*, *3*(1), 36-43.