Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations

DOI: 10.54254/2753-8818/52/2024CH0138

Stock prediction based on HMM and LSTM model and model selection using SVM

Moxun Deng

Leicester International Institute, Dalian University of Technology, Dalian, 116085, China

2939467211@mail.dlut.edu.cn

Abstract. In the field of time series analysis, stock prediction has always been a hot research topic. Hidden Markov Model (HMM) and Long Short-Term Memory (LSTM) are two powerful models that are very suitable for the stock prediction, but they will have different performance for different datasets. Therefore, selecting the appropriate model for different datasets has become a challenge. For different datasets, their statistical characteristics are an important means of analyzing datasets and the main feature that distinguishes them. Therefore, by analyzing the statistical features of the dataset and using these features for classification, it is possible to determine which model is most suitable for the dataset. This paper will use HMM and LSTM to predict some stock data, consider the effectiveness of the model through two dimensions of Mean Absolute Percentage Error (MAPE), and extract statistical feature values of these data. Finally, Support Vector Machine (SVM) will be used to classify the test dataset and verify whether the predicted model performs better than another.

Keywords: Stock prices, Hidden Markov Model, Long Short-Term Memory, Support Vector Machine.

1. Introduction

With the unprecedented prosperity of the stock market, researchers have never stopped accurately predicting stock prices. The stock model is almost the most complex and influential model in the real world. The price and return of stocks are only observable states from the outside world, and a large number of implicit factors are the main factors leading to stock price fluctuations. Morck et al. indicated that stock prices are related to the current level of economic development in a country [1]. Wen et al. believed that stock prices are related to changes in consumer emotions [2]. Hong et al. thought that prejudice is also one of the important factors affecting stock prices [3]. Song and Yu believed that monetary policy is also an important factor affecting stock prices [4]. It is not difficult to see that the standards proposed above are not quantifiable, which poses a huge obstacle to accurate stock prediction. But as people delve deeper into stock research, more and more models are being used to predict stock models. These models have their own characteristics in predictive performance, making it difficult to choose the best model suitable for all types of data.

In the previous work of Hassan et al. Hidden Markov Model (HMM) was used for stock research, with a focus on aviation stocks, and was used to predict prices [5]. Nguyet introduced Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) into the HMM algorithm for

DOI: 10.54254/2753-8818/52/2024CH0138

state selection [6]. Su and Deng believed that model SVM can be used as a tool for model selection [7]. Hassan et al. combined HMM, Artificial Neural Network (ANN), and Genetic Algorithm (GA) algorithms to form a new algorithm [8]. Nelson et al. used Long Short-Term Memory (LSTM) to predict stock prices [9]. Jiang and Liu combined Auto-Regressive Moving Average Model (ARMA) with LSTM for stock prediction [10].

In summary, the research on stock prices has attracted numerous scholars. This article will implement HMM and LSTM, and compare the prediction results. After comparison, this article will conduct an innovative study: analyzing why the same model performs differently in different datasets by extracting statistical feature values from the dataset, and selecting models by learning statistical features from previous datasets. In the process of seeking the best model, this article adopts Support Vector Machine (SVM) as the algorithm for training the model.

2. Methodology

2.1. Data source

The data for this literature was collected from the website investing.com, which contains a large amount of stock data that can be used for model training. The dataset includes closing prices, opening prices, the highest and lowest prices traded on the day, and the rate of return calculated based on the average price.

The downloaded dataset includes stock data from January 2014 to June 2024, with approximately 2600 pieces of data. In order to maintain the reliability of the prediction results, the author trimmed some of the data to ensure consistent data volume for all participating datasets (Table 1).

Table 1. Sample Dataset

Date	Close	Open	High	Low	Return
2024/6/28	183.42	185.72	186.58	183.34	-1.84%
……	186.86	185.65	187.49	185.54	0.80%
2014/1/29	185.37	184.19	185.93	184	-0.11%

The dataset requires certain preprocessing when making formal model predictions, which will be explained later in this paper. To train the SVM model, people need to perform further operations on the basic dataset. The statistical indicators including mean, weighted mean (calculated using weights assigned to each value), trimmed mean (calculated after removing extreme values), median (median in sorted datasets), skewness (measure of distribution asymmetry), MAD (mean absolute deviation, providing dispersion around the mean), Q1 (first quartile), Q3 (third quartile), IQR (representing the interquartile range of the middle 50% distribution), mode (most common value), range (difference between maximum and minimum values), variance (measure of data distribution), and kurtosis (distribution shape). Not all of these statistical features are useful, so it is necessary to reduce some to improve model performance.

2.2. Method introduction

This section of this paper will focus on the algorithms used in the prediction and selection processes. This section of the paper will concentrate on the algorithms employed in the prediction and selection processes. The paper will use Hidden Markov Models (HMM) and Long Short-Term Memory (LSTM) for predicting stock data, and capturing their Mean Absolute Percentage Error (MAPE). Subsequently, it will compute the statistical characteristic values of these datasets and compile a dataset. Finally, the Support Vector Machine (SVM) will be employed to classify the test dataset, and the classification outcomes will be compared with the actual results.

DOI: 10.54254/2753-8818/52/2024CH0138

2.2.1. The principle of HMM

The hidden Markov model (HMM) is a probabilistic model used for analyzing time series data. Its fundamental component is the hidden Markov chain, a variant of the Markov chain. A standard Markov chain consists of a state transition matrix and a list of observable states. Each state in the list is observable, while the transition matrix records the probabilities of transitioning from one state to another. When using Markov chains for data prediction, it's assumed that the next state depends solely on the current state, known as the Markov hypothesis.

However, practical applications have shown that Markov chains often struggle to accurately predict data. The hidden Markov chain aims to address this limitation by uncovering and studying these hidden variables through observable data, thereby improving predictions based on trends in these hidden factors.

Overall, a hidden Markov chain consists of the following parts. Firstly, it is the observation state, which is a state that people can directly observe during observation, and is usually the state that people ultimately want to obtain results in predicting problems. Next is the hidden state. Hidden states are states that people cannot obtain through observation. Finally, there is a transition matrix between the observed state and the hidden state. In a hidden Markov chain, the observed state is only determined by the hidden state at the current time point, while the current hidden state is only determined by the hidden state at the previous time point. The specific situation is shown in the figure 1:

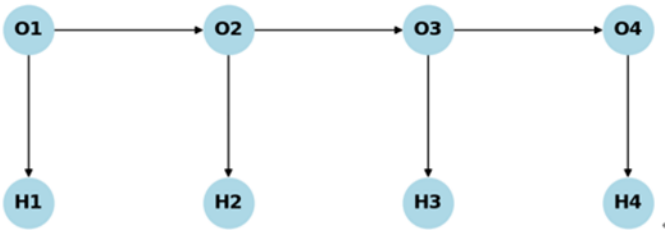

Figure 1. Hidden Markov Chain.

Therefore, although people can calculate the transition matrix of the observed state, in most cases, this transition matrix does not have practical significance.

For Hidden Markov Models, selecting the number of states is another challenge. This paper adopts the Bayesian decision-making method to select the number of states [2]. When selecting a model for prediction, there is a default assumption that there exists an optimal model for any given dataset, which can be found based on known data, and Bayesian Information Criterion is a simple selection criterion. The formula for calculating Bayesian Information Criterion is:

$$BIC = -2\ln(L) + \ln(n) * k \quad (1)$$

Where L is the maximum likelihood under this model, n is the number of data, and k is the number of variables in the model. When selecting the model state, the first step is to set a range for the model state, and then select the optimal situation based on the calculation results of BIC.

2.2.2. The principle of LSTM

Long Short-Term Memory (LSTM) represents an enhancement over the traditional Recurrent Neural Network (RNN) algorithm. It's widely acknowledged that vanilla RNNs struggle with processing long sequences due to issues like gradient explosion and vanishing gradients. Gradient explosion can be managed by capping gradients to prevent NaN outputs during training. Conversely, vanishing gradients are more insidious as they're harder to detect and mitigate.

When vanishing gradients occur, earlier data in the training process gradually lose significance, potentially reducing their contribution to the model's overall training effectiveness or rendering them negligible. Researchers often infer gradient vanishing from poor model training outcomes rather than direct detection.

LSTM addresses the issue of RNN forgetting long-term dependencies by introducing cell state c_t dedicated to retaining long-term information. LSTM typically have three inputs and two outputs. Inputs

DOI: 10.54254/2753-8818/52/2024CH0138

include the previous cell state c_{t-1} and output h_{t-1} from the preceding LSTM, along with the current network input x_t. Outputs consist of the current cell state c_t and output h_t for the subsequent LSTM.

In order to handle the relationship between these three input values, the concept of gate was introduced in LSTM. LSTM uses two gates to control the content of cell state c_t, one is a forget gate, which determines how much of the previous cell state is retained until the current time c_t. Another is the input gate, which determines how much of the network's input x_t is saved to the cell state c_t at the current time. LSTM uses an output gate to control how much the cell state c_t outputs to the current output value h_t of LSTM.

2.2.3. The principle of SVM

SVM are generally used to solve classification problems, where a given dataset T, find a hyperplane that can separate the dataset, and this optimal hyperplane should maximize the geometric interval between the support vector and the hyperplane (Figure 2).

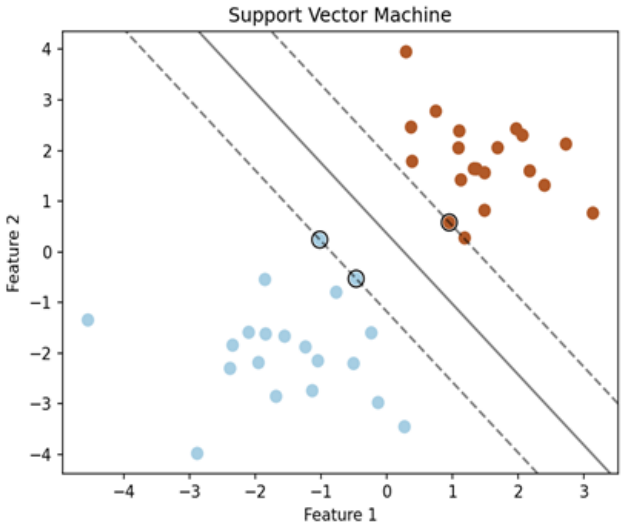

Figure 2. SVM schematic diagram.

In this paper, SVM is used to select the better term from HMM and LSTM in the prediction of the same dataset. As mentioned earlier, when choosing a prediction algorithm, it means that there must be an optimal solution here, and this optimal solution can be found by analyzing known data. For different datasets, algorithms have different performance, and there is no doubt that this change is caused by changes in certain features of the dataset itself. Therefore, analyzing these features can help people find possible algorithms with better performance among various algorithms. The statistical characteristics of the dataset are an important set of features that cannot be ignored. Therefore, SVM can be used for classification learning by combining the performance level of previous predictions on the dataset with the statistical characteristics of the dataset itself. Before making numerical predictions on the new dataset, calculate the statistical characteristics of the dataset, use the already trained SVM to select the algorithm, and use it as the evaluation criterion for algorithm selection.

3. Results and discussion

3.1. Evaluation criterion

This paper uses Mean Absolute Percentage Error (MAPE) to evaluate the performance of the model. MAPE is obtained by dividing the absolute difference between the predicted value and the true value by the true value. In specific model calculations, the MAPE of each step is added and divided by the number of model steps to obtain the value of the entire model's MAPE. It is worth noting that once a value of 0

DOI: 10.54254/2753-8818/52/2024CH0138

occurs in the model, MAPE will overflow and become unusable. Therefore, in order to solve this problem, the raw data needs to undergo certain preprocessing.

In model training, any price will not be zero, but the return may not change. Given that the possible range of returns is (-1, +∞), the author add 1 to the return and make its range (0, +∞) to avoid MAPE overflow. It is worth noting that the processed data is still valuable, and return at this point can be more conveniently calculated for the expected return when investors invest x units of cash.

3.2. HMM prediction results

The prediction results of HMM are shown in the following figure 3, 4 and table 2.

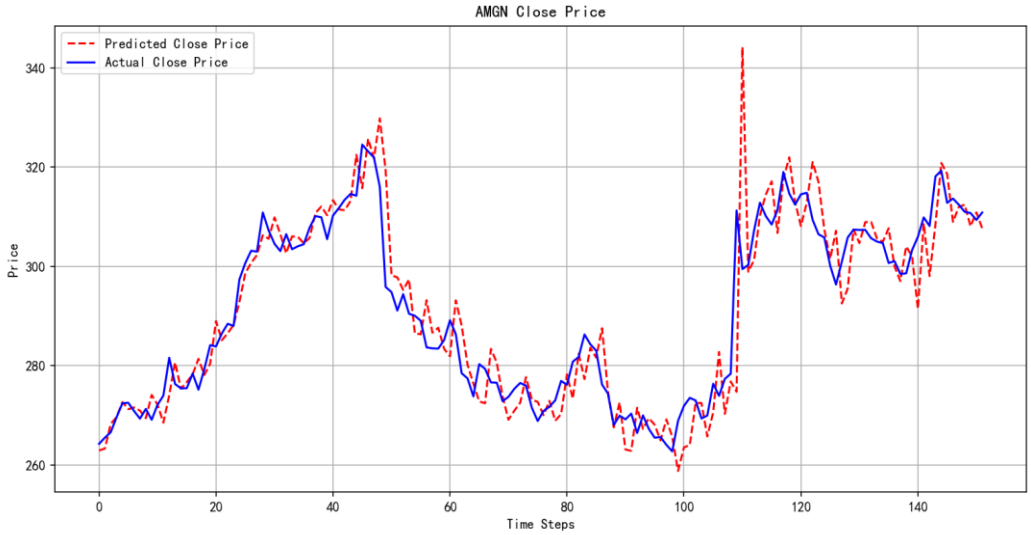

Figure 3. Sample of predictions on stock prices

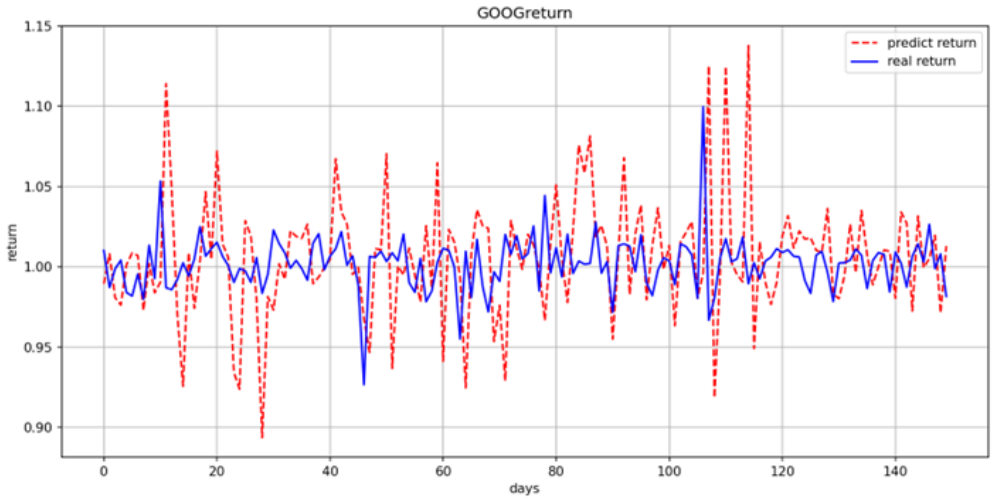

Figure 4. Sample of prediction of returns

It is not difficult to see that HMM performs well in predicting stock values, especially in predicting stock prices. Correspondingly, HMM's prediction of stock returns is not as accurate as price prediction. However, overall, HMM can accurately predict the trend of stock fluctuations, although the prediction is still not accurate to a certain extent, it is within an acceptable range.

DOI: 10.54254/2753-8818/52/2024CH0138

Table 2. MAPE of prediction

Stock	MAPE of Close Price	MAPE of Open Price	MAPE of High Price	MAPE of Low Price	MAPE of Return Rate
GOOG	0.0185	0.0171	0.0178	0.0181	0.0343
TSLA	0.0161	0.0160	0.0154	0.0170	0.0486
NKE	0.0186	0.0178	0.0162	0.0176	0.0514
AMZN	0.0181	0.0172	0.0162	0.0176	0.0277
BA	0.0223	0.0213	0.0191	0.0213	0.0366
CVX	0.0130	0.0124	0.0125	0.0117	0.0207
CAT	0.0138	0.0171	0.0143	0.0145	0.0239
INTC	0.0257	0.0241	0.0228	0.0224	0.0415
MSFT	0.0127	0.0136	0.0116	0.0117	0.0216
DIS	0.0158	0.0155	0.0137	0.0121	0.0283
CSCO	0.0101	0.0110	0.0096	0.0085	0.0167
GS	0.0156	0.0124	0.0136	0.0125	0.0208
KO	0.0076	0.0077	0.0068	0.0076	0.0126
MCD	0.0108	0.0120	0.0094	0.0105	0.0167
MRK	0.0111	0.0118	0.0110	0.0104	0.0205
MMM	0.0157	0.0150	0.0159	0.0155	0.0305
AAPL	0.0148	0.0153	0.0140	0.0143	0.0224
AMGN	0.0151	0.0157	0.0157	0.0144	0.0268

The table 2 above provides a more detailed description of the specific situation when the HMM model predicts stocks. It is not difficult to see that the HMM model can adapt well to stock model prediction, and the MAPE is within 10%, which is an acceptable error.

3.3. LSTM prediction results

The prediction results of LSTM are shown in the following figure 5, 6 and table 3.

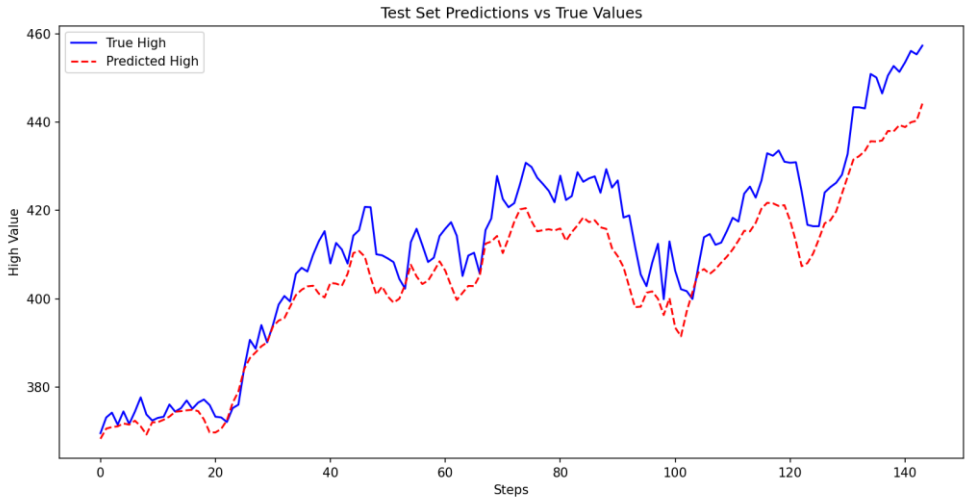

Figure 5. Sample of predictions on stock prices

DOI: 10.54254/2753-8818/52/2024CH0138

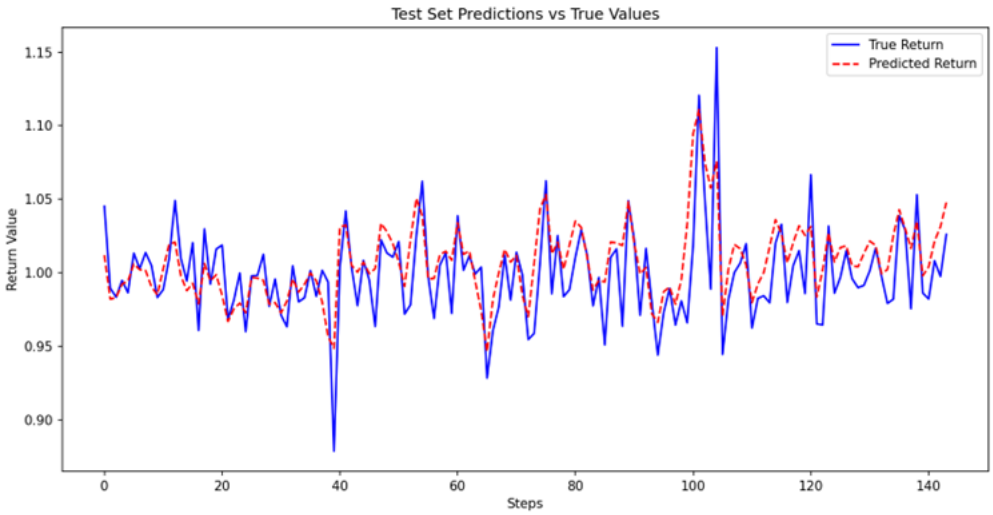

Figure 6. Sample of prediction of returns

It is not difficult to see that LSTM performs well in predicting stock values, especially in predicting stock returns. However, at the same time, LSTM cannot accurately predict stock prices, although the error is within an acceptable range. Compared with HMM, it is not difficult to find that these two different algorithms have different predictive effects on different forms of data. Therefore, it can be considered to find a more suitable algorithm for predicting data by calculating statistical features.

Table 3. MAPE of prediction

Stock	MAPE of Close Price	MAPE of Open Price	MAPE of High Price	MAPE of Low Price	MAPE of Return Rate
GOOG	0.0288	0.0244	0.0272	0.0357	0.0303
TSLA	0.0293	0.0272	0.031	0.0258	0.0387
NKE	0.0198	0.0156	0.0175	0.017	0.0309
AMZN	0.0202	0.0168	0.017	0.0182	0.0331
BA	0.0217	0.0162	0.0193	0.0181	0.0232
CVX	0.0123	0.0092	0.0107	0.0107	0.0337
CAT	0.025	0.023	0.0246	0.0275	0.0285
INTC	0.0242	0.019	0.0222	0.0215	0.0247
MSFT	0.0239	0.0219	0.0216	0.0275	0.0138
DIS	0.0168	0.0139	0.016	0.0169	0.0155
CSCO	0.0135	0.0105	0.0138	0.0107	0.0112
GS	0.0151	0.0118	0.0135	0.0158	0.0132
KO	0.0077	0.0067	0.0067	0.007	0.008
MCD	0.0144	0.0109	0.0124	0.0148	0.0167
MRK	0.0216	0.0174	0.0171	0.0217	0.0083
MMM	0.0175	0.0153	0.0174	0.0164	0.02
AAPL	0.0185	0.0156	0.0175	0.0164	0.0165
AMGN	0.0216	0.0171	0.0186	0.0199	0.0127

DOI: 10.54254/2753-8818/52/2024CH0138

The table 3 above provides a more detailed description of the specific situation when the LSTM predicts stocks. It is not difficult to see that the LSTM can adapt well to stock model prediction, and the MAPE is within 10%, which is an acceptable error.

3.4. SVM prediction results

When using SVM for model selection, it is necessary to determine which statistical feature values are useful. Obviously, the mean, median, and quartile range do not affect the effectiveness of model training. Therefore, these factors will not be reflected in the model. The results of SVM are shown in the following table 4. The trained SVM model has the ability to select acceptable models.

Table 4. Accurately predicting the better algorithm

Dataset	Accuracy
Dataset1	89%
Dataset2	67%
Dataset3	89%
Dataset4	56%
Average	75%

4. Conclusion

This article selected approximately 2500 data samples from 180 datasets from 2014 to 2024, totaling approximately 460000 data samples. The author used HMM and LSTM to predict stock data, recorded their MAPE, and selected the model with better prediction performance for each dataset as the classification criterion for the dataset. Then, by calculating the statistical eigenvalues of these data, SVM is used to learn from the obtained eigenvalues and previous training results, and the trained model is used to classify the test dataset. This article first demonstrates the good performance of HMM and LSTM algorithms in stock prediction. Based on the classification results, it is evident that using statistical features and SVM algorithm for model selection is feasible.

As a further research direction, the author believes that more models should be introduced first to predict stock data, in order to determine whether SVM algorithm can still be used for accurate model selection under multiple model conditions. Secondly, the classification algorithm should be further optimized by comparing multiple classification algorithms to find models with higher classification accuracy. Finally, a feasible direction is trying to expand the application scope of the algorithm and explore whether algorithms that use statistical features as classification criteria are still effective in other prediction fields, such as weather forecasting.

References

[1] Morck, et al. 2000 The information content of stock markets: why do emerging markets have synchronous stock price movement?, Journal of Financial Economics.
[2] Wen F, et al. 2014 Research on the Impact of Investor Emotional Characteristics on Stock Price Behavior, Journal of Management Science, 3, 60-69.
[3] Hong H, et al. 2008 The Only Game in Town: Stock-Price Consequences of Local Bias. Journal of Financial Economics, 1, 20-37.
[4] Song Q and Yu S, 2002 On the Relationship between Monetary Policy and Stock Market, Wuhan Finance.
[5] Hassan M R and Nath B, 2005 Stock market forecasting using hidden Markov model: a new approach, 5th International Conference on Intelligent Systems Design and Applications, 192-196.
[6] Nguyet N, 2018 Stock Price Prediction using Hidden Markov Model, Int. J. Financial Stud., 6(2).
[7] Su G and Deng F, 2006 Model Selection for Support Vector Regression, Science and Technology Bulletin, 22(2), 5.

DOI: 10.54254/2753-8818/52/2024CH0138

[8] Hassan M R, et al. 2007 A fusion model of HMM, ANN and GA for stock market forecasting. Expert systems with Applications, 33(1), 171-180.
[9] Nelson D M Q, et al. 2017 Stock market's price movement prediction with LSTM neural networks, 2017 International joint conference on neural networks (IJCNN) IEEE, 1419-1426.
[10] Jiang Q and Liu C 2018 A stock prediction method based on ARMA-LSTM model. Journal of Management Science.