Taiwanese Stock Market Forecasting with a Shallow Long Short-Term Memory Architecture

Phuong Ha Dang Bui[1], Toan Bao Tran[2,3], and Hai Thanh Nguyen[1](✉)

[1] College of Information and Communication Technology, Can Tho University, Can Tho 900100, Vietnam
{bdhphuong,nthai}@cit.ctu.edu.vn
[2] Center of Software Engineering, Duy Tan University, Da Nang 550000, Vietnam
tranbaotoan@dtu.edu.vn
[3] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

Abstract. The trading of stock in companies holds an important part in numerous economies. Stock Forecast which is popularly published in the public domain in the forms of newsletters, investment promotion organizations, public/private forums, and scientific forecast services is very necessary to contribute successes in financial for individuals or organizations. Leveraging advancements in machine learning, we propose an approach based on Long Short-Term Memory model and compare the performance to the classic machine learning such as Random Forest model and Support Vector Regression model when we do forecast tasks on Taiwanese stock market. The proposed method with deep learning algorithm shows better performance comparing to the classic machine learning in the tasks of forecasting the stock market in Taiwan.

Keywords: Trading of stock · Machine learning · Forecast tasks

1 Introduction

In recent years, the development of Machine learning and Deep learning technique has an influence on the stock market in participate and the finance predictions in general. The prediction of future stock price movement has been widely researched in numerous studies. The existing approaches have focused on the construction of econometric, statistical based on the selection of variables or forecasting models. The stock market forecasting is not a basic task due to the behavior of a stock time series. By using Machine learning techniques can help the market intermediaries to reveal better forecasting stock prices for short term stock price trends. Furthermore, there are numerous factors that produce uncertainty and high volatility in the market affected stock markets definitely which can be considered as the psychology of investors' political events, general economic conditions, or commodity price index [1]. The computation of the

P. C. Vinh and A. Rakib (Eds.): ICCASA 2020/ICTCC 2020, LNICST 343, pp. 192–202, 2021.
https://doi.org/10.1007/978-3-030-67101-3_16

value of stock groups is based on market capitalization. In each country, the estimated economic status are presented by the prices of stocks with high market investment.

In this paper, we propose an approach based on Long Short-Term Memory model and compare the performance to the classic machine learning such as Random Forest model and Support Vector Regression model to forecast the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) [12]. The proposed method with deep learning algorithm shows better performance comparing to the classic machine learning for forecasting the TAIEX. The rest of this paper is organized as follows. In Sect. 2, we present related work on forecasting the TAIEX. In Sect. 3, we present the forecasting methods based on three models: Random Forest model, Support Vector Regression model and Long Short-Term Memory model. In Sect. 4, we compare the forecasting results of the methods based on three learning models. The conclusion is discussed in Sect. 5.

2 Related Work

Some methods including [2–11] have been presented to forecast the TAIEX [12].

In [2], Cai et al. presented a fuzzy time series model combined with ant colony optimization (ACO) and auto-regression to forecast the TAIEX, where ACO is used to partition the universe of discourse and an auto-regression high-order fuzzy time series model is used to make better using of the historical data. Chen and Chang [3] presented a multi-variable fuzzy forecasting method based on the fuzzy clustering method and fuzzy rule interpolation techniques for forecasting the TAIEX. In [4], Chen and Chen presented a fuzzy time series forecasting method to forecast the TAIEX based on fuzzy variation groups. Chen and Chen [5] presented a method for forecasting the TAIEX based on two-factors second-order fuzzy-trend logical relationship groups and the probabilities of the down-trend, the probabilities of the equal-trend and the probabilities of the up-trend of the two-factors second-order fuzzy logical relationships. Chen and Kao [6] presented a method to forecast the TAIEX based on fuzzy time series, PSO techniques and support vector machines, where the PSO techniques are used to obtain optimal intervals in the universe of discourse and the support vector machine is used to classify the training data set. In [7], Chen et al. presented a method to forecast the TAIEX based on two-factors second-order fuzzy-trend logical relationship groups and PSO techniques, where the PSO techniques are used to get the optimal weighting vector of each group of the fuzzy-trend logical relationship groups. Chen and Phuong [8] presented a method for forecasting the TAIEX based on optimal partitions of intervals in the universe of discourse and optimal weighting vectors of two-factors second-order fuzzy-trend logical relationship groups, where the PSO techniques are used to obtain the optimal partitions of intervals and the optimal weighting vectors simultaneously. Huarng et al. [9] presented a method using a multivariate heuristic model, which can be integrated with univariate fuzzy time-series models to forecast the TAIEX. In

[10] Yu presented a weighted fuzzy time-series method to forecast the TAIEX. Yu and Huarng [11] used neural networks to fuzzy time series forecasting and presented bivariate fuzzy time series models to forecast the TAIEX. In order to increase the forecasting accuracy rates to forecast the TAIEX, we need to propose a forecasting method to obtain better results.

3 Random Forest, Support Vector Regression and Long Short-Term Memory Models for Stock Market Prediction

We use Random Forest [13], Support Vector Regression (SVR) [14] and Long Short-Term Memory (LSTM) [15] models to investigate the performance of stock prices prediction.

3.1 The Random Forest Model

Random Forest is a supervised learning algorithm which uses ensemble learning method for classification and regression. A Random Forest operates by constructing several decision trees at training time and outputting the class or mean prediction of the individual trees. In this paper, we implement Random Forest Regression with the following parameters: the number of trees in the forest is 500, the maximum depth of each tree is 4 and estimate the tree quality by Mean Squared Error (MSE). MSE is the square of RMSE.

3.2 The Support Vector Regression Model

Support Vector Regression is the combination of a Support Vector Machine and Regression. SVR is flexible to define the acceptable error in the model and will find an appropriate hyperplane to fit the data. Furthermore, the stock prices prediction using SVR can gain promising results due to the minimization the error within a certain threshold. In this paper, we implement SVR with the regularization parameter is 0.001, the gamma is 0.1 and Radial Basis Function kernel.

3.3 The Long Short-Term Memory Model

Long Short-Term Memory is a special kind of Recurrent Neural Networks (RNN) and can be used to avoid the long-term dependency problem. The LSTM architecture contains a Long Short-Term Memory layer and a Fully Connected layer. We implement the LSTM with Adam optimizer function [16] and MSE loss function. The learning rate is initiated at 0.001. We train our LSTM on 20 epochs with a batch size of 1. We also illustrate the LSTM in Fig. 1.

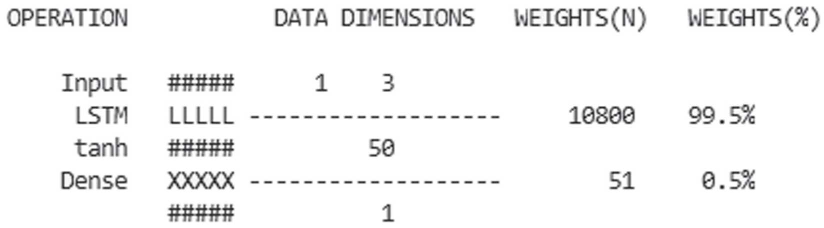

Fig. 1. The visualization of LSTM architecture which includes 2 layers LSTM with tanh activation function

Table 1. The performance of Support Vector Regression model over several training years

No. of training years	Training RMSE	Testing RMSE	Training MAE	Testing MAE
9 years	437.9962	403.0504	361.0728	360.5454
10 years	429.9880	419.9105	353.8121	383.4044
11 years	435.1563	392.1609	361.7393	358.0659

Table 2. The further information of TAIEX dataset

Dataset	From year	To year	Min price	Max price	Avg price	Total number of samples
TAIEX	1990	2004	2560.47	12495.34	6053.86	4086

4 Experimental Results

In this section, we compare the forecasting results of the methods based on three learning models: the Random Forest model, the Support Vector Regression model and the Long Short-Term Memory model.

4.1 The TAIEX Dataset

We use the TAIEX dataset in our experiments. The TAIEX dataset includes the stock price on each trading day over several years. The TAIEX were recorded in 15 years, from 1990 to 2004. Further information, e.g. minimum, maximum, average stock prices and the number of total samples of the dataset are described in Table 2. We also visualize the stock prices change of the TAIEX dataset in Fig. 2. The minimum price is 2560.47, whereas the maximum and average are 12495.34 and 6053.86 respectively.

We divide the dataset into two parts, where the first part of dataset is used as the training data and the remaining is used as the testing data. In specific, we use the first 9 years for training, the last 6 years for testing. Besides, we also use the first 10 years for training, the last 5 years for testing and the first 11 years for training, the last 4 years for testing.

We normalize the stock prices by rescaling the values within the range of 0 and 1 due to the differences in the scales across the stock prices. In general, an

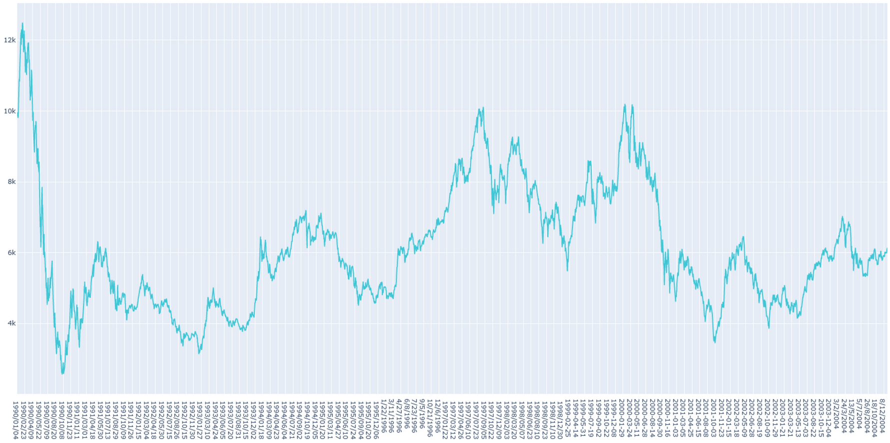

Fig. 2. The visualization of TAIEX dataset from 1990 to 2004

output with a large value may influence negatively to the learning process of the models.

The purposes of data normalization are to enhance the numerical stability of the learning models and to reduce the training time. In other words, the learning models can converge rapidly. Furthermore, accuracy is not affected by normalizing inputs.

4.2 The Metrics for Evaluating the Performance

In order to evaluate the performance of learning models, we use the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). They are both common metrics in regression and define the prediction errors with the range from 0 to ∞. They are negatively-oriented scores, which means lower values are better. The MAE and the RMSE can be computed by the Eq. 1 and Eq. 2.

$$MAE = \frac{\sum_{n}^{i=1} |y_i - \hat{y}_i|}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{n}^{i=1} (y_i - \hat{y}_i)^2}{n}} \tag{2}$$

where n denotes the number of trading days of the historical testing data, y_i denotes the forecasted value of the historical testing datum of the TAIEX on trading day i, and $\hat{y}_i$ denotes the actual value of the historical testing datum of the TAIEX on trading day i, where $1 \leq i \leq n$.

4.3 The Performance of Random Forest Model

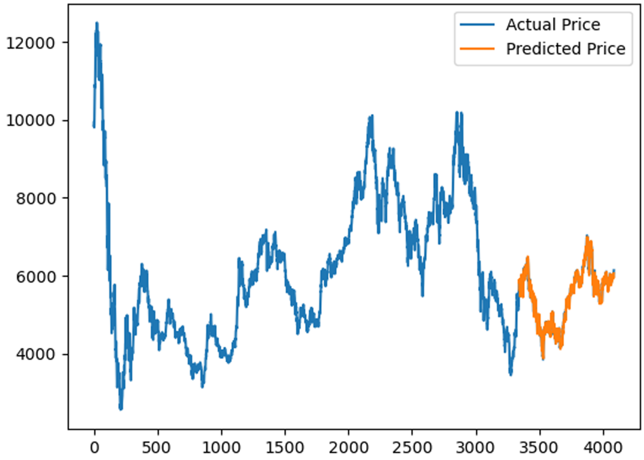

Fig. 3. The visualization of the actual prices and the forecasting prices by using Random Forest model on 11 training years (Color figure online)

Table 3. The performance of Random Forest model over several training years

No. of training years	Training RMSE	Testing RMSE	Training MAE	Testing MAE
9 years	50.5450	115.8137	35.0288	85.0645
10 years	52.4664	94.5714	36.4398	72.0539
11 years	51.6092	91.1585	35.9066	69.8861

The performance of Random Forest model on 9 training years, 10 training years and 11 training years are presented in Table 3. The performance of Random Forest model on 11 training years is better than the others. In specific, the lowest errors obtained are 91.1585 and 69.8861 for the RMSE and the MAE respectively. The performance of that on 10 training years and 9 training years follows closely, with the RMSE are 94.5714 and 115.8137 respectively, and the MAE are 72.0539 and 85.0645 respectively. We also visualize the forecasting prices and the actual prices by using Random Forest model on 11 training years in Fig. 3. The blue line in Fig. 3 stands for the actual prices, whereas the orange line denotes the predicted prices.

4.4 The Performance of Support Vector Regression Model

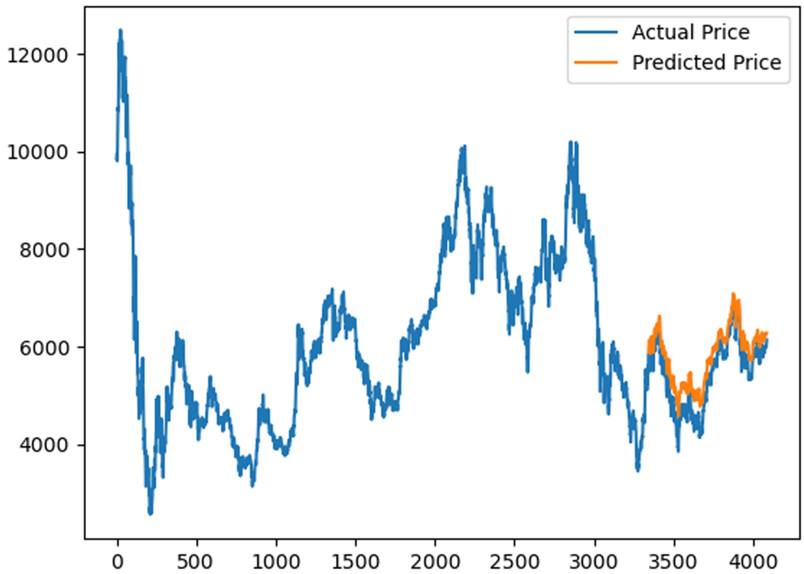

Fig. 4. The visualization of the actual prices and the forecasting prices by using Support Vector Regression model on 11 training years (Color figure online)

We also investigate the performance of SVR model on 9 training years, 10 training years and 11 training years. The SVR model on 11 training years reached the highest performance with the RMSE is 392.1609 and the MAE is 358.0659, followed by that on 9 training years, with the RMSE and the MAE obtained are 403.0504 and 360.5454 respectively. We also present the performance details of SVR model in Table 1. Furthermore, Fig. 4 visualizes the actual prices and forecasting prices by using SVR model on 11 training years. The actual prices is visualized by the blue line, whereas the orange line stands for the forecasting prices.

4.5 The Performance of Long Short-Term Memory Model

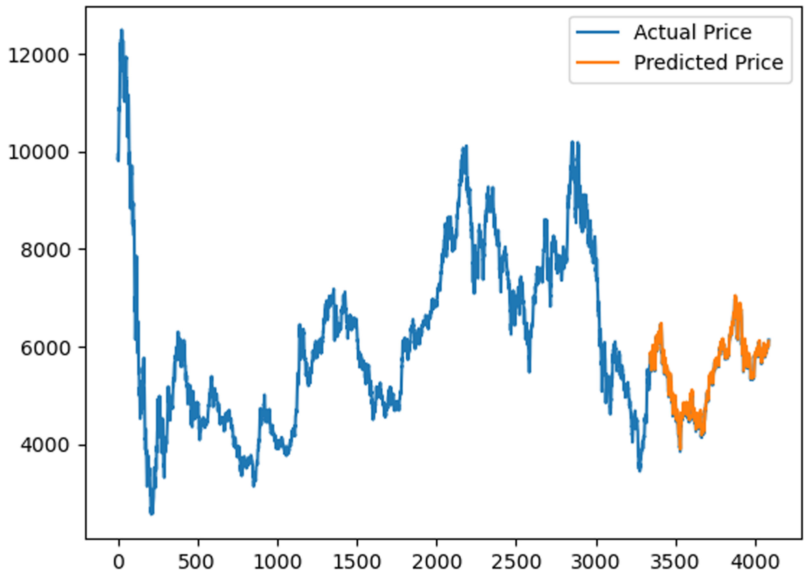

Fig. 5. The visualization of the actual prices and the forecasting prices by using Long Short-Term Memory model on 11 training years (Color figure online)

The optimal performance of LSTM model obtained by using 11 training years with the RMSE is 88.7651 and the MAE is 67.7686, followed by the performance of that by using 10 training years with the RMSE and the MAE are 89.7668 and 69.1916 respectively. We also present the performance details of LSTM model in Table 4. The actual prices and the forecasting prices by using LSTM model on 11 training years are visualized in Fig. 5. The actual prices is visualized by blue line, whereas the orange line stands for the forecasting prices.

Table 4. The performance of Long Short-Term Memory model over several training years

No. of training years	Training RMSE	Testing RMSE	Training MAE	Testing MAE
9 years	122.3369	109.4436	86.2959	80.0706
10 years	125.5805	89.7668	89.2567	69.1916
11 years	121.4129	88.7651	88.2835	67.7686

Table 5. The comparison between the performance of Random Forest model, Support Vector Regression model, and Long Short-Term Memory model in the testing phase on 11 training years

The learning model	RMSE	MAE
LSTM	88.7651	67.7686
Random Forest	91.1585	69.8861
SVR	392.1609	358.0659

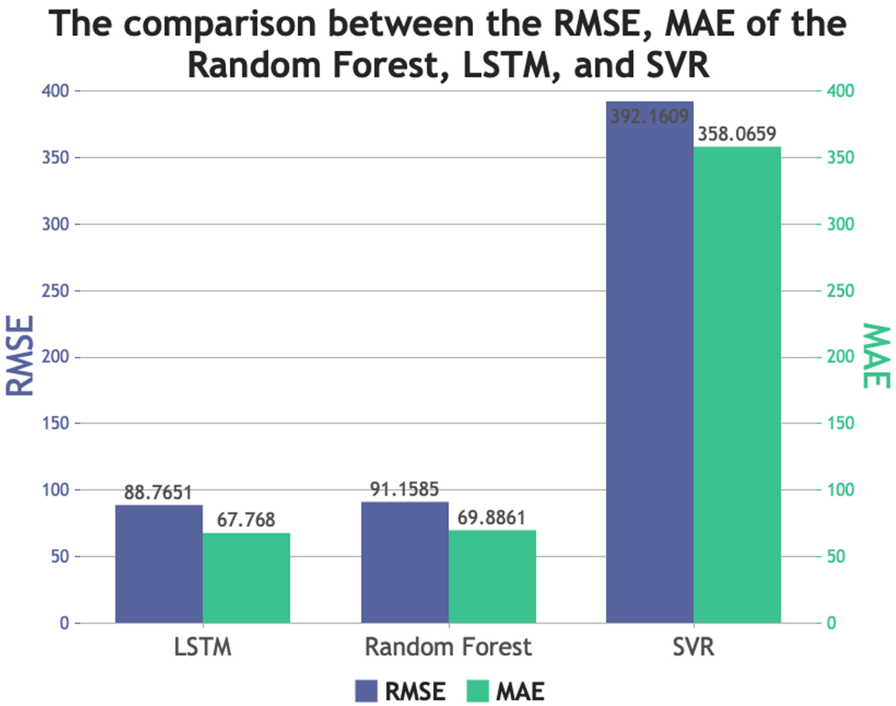

Fig. 6. The visualization of the RMSE and the MAE for comparison between Random Forest model, LSTM model, and SVR model on 11 training years

4.6 The Comparison Between the Learning Models

The comparison between the learning models on 11 training years is presented in Table 5. The LSTM model shows the highest performance with the RMSE is 88.7651 and the MAE is 67.7686, followed by the Random Forest model, with the RMSE and the MAE are 91.1585 and 69.8861 respectively. The performance of SVR model is not good enough in comparison to the others. We also visualize the RMSE and the MAE for comparison between Random Forest model, LSTM model, and SVR model on 11 training years in Fig. 6.

5 Conclusion

In this paper, we propose the approach with a shallow deep learning architecture based on Long short-term memory for forecasting the TAIEX. In order to evaluate the performance for forecasting prices, we divide the dataset into two phases. The first one is the preceded 9–11 years which fetched into the model for the learning and the remaining is for the testing phase. As observed from the tables, we can see that with more than time for training set, we obtain lesser errors in testing phase.

We also compare the performance between deep learning approach and the classic machine learning method for forecasting tasks. The obtained results show that deep learning with Long short-term memory (although only with a shallow architecture) performs an efficient prediction performance comparing to the classic machine learning such as Random Forest model and Support Vector Regression model.

We continue the work with further research on sophisticated architectures to improve the performance in forecasting.

References

1. Miao, K., Chen, F., Zhao, Z.: Stock price forecast based on bacterial colony RBF neural network. J. Qingdao Univ. (Nat. Sci. Ed.) **2**(11) (2007)
2. Cai, Q., Zhang, D.F., Zheng, W., Leung, S.C.H.: A new fuzzy time series forecasting model combined with ant colony optimization and auto-regression. Knowl. Based Syst. **74**, 61–68 (2015)
3. Chen, S.M., Chang, Y.C.: Multi-variable fuzzy forecasting based on fuzzy clustering and fuzzy interpolation techniques. Inf. Sci. **180**(24), 4772–4783 (2010)
4. Chen, S.M., Chen, C.D.: TAIEX forecasting based on fuzzy time series and fuzzy variation groups. IEEE Trans. Fuzzy Syst. **19**(1), 1–12 (2011)
5. Chen, S.M., Chen, S.W.: Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups and the probabilities of trends of fuzzy logical relationships. IEEE Trans. Cybern. **45**(3), 405–417 (2015)
6. Chen, S.M., Kao, P.Y.: TAIEX forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines. Inf. Sci. **247**, 62–71 (2013)
7. Chen, S.M., Manalu, G.M.T., Pan, J.S., Liu, H.C.: Fuzzy forecasting based on two-factors second-order fuzzy-trend logical relationship groups and particle swarm optimization techniques. IEEE Trans. Cybern. **43**(3), 1102–1117 (2013)
8. Chen, S.M., Phuong, B.D.H.: Fuzzy time series forecasting based on optimal partitions of intervals and optimal weighting vectors. Knowl. Based Syst. **118**, 204–216 (2017)
9. Huarng, K., Yu, T.H.K., Hsu, Y.W.: A multivariate heuristic model for fuzzy time-series forecasting. IEEE Trans. Syst. Man Cybern. Part B Cybern. **37**(4), 836–846 (2007)
10. Yu, T.H.K.: Weighted fuzzy time-series model for TAIEX forecasting. Physica A **349**(3–4), 609–624 (2004)
11. Yu, T.H.K., Huarng, K.H.: A bivariate fuzzy time series model to forecast the TAIEX. Expert Syst. Appl. **34**(4), 2945–2952 (2008)

12. TAIEX. http://www.twse.com.tw/en/products/indices/tsec/taiex.php
13. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
14. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: Mozer, M.C., Jordan, M.I., Petsche, T. (eds.) Advances in Neural Information Processing Systems, vol. 9, pp. 155–161. MIT Press, Cambridge (1997)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
16. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization (2014). arXiv:1412.6980v9