QBS
Vol. 39, No. 1, pp. 25-31 (2020)
https://doi.org/10.22283/qbs.2020.39.1.25
pISSN 2288-1344 eISSN 2508-7185 **Original Article**

Quantitative Bio-Science
Official Journal of the Institute of Natural Science at Keimyung University
Available Online at https://qbs.kmu.ac.kr:442/

KOSPI200 Prediction through Low-Pass Filtered Long Short-Term Memory Algorithm

Dong Won Lee[1], Hee Soo Lee[2], Kyong Joo Oh[3,*]

[1]Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea
[2]School of Business, Sejong University, Seoul 05006, Korea
[3]Department of Information & Industrial Engineering, Yonsei University, Seoul 03722, Korea
(Received November 26, 2019; Revised December 15, 2019; Accepted January 15, 2020)

ABSTRACT

Diversification of the modern financial market has led to an increase in the importance of the stock market index. Trend of the stock market at large can be identified through the analysis of the stock market index. Movements of stock indices serve as a key measure for investors while trading individual stocks and play an important role in establishing basic asset allocation strategy. Currently, in the field of computer science, research on data prediction through machine and deep learning is being actively conducted. Using these algorithms, research on financial time-series data is being conducted in the financial domain. Similar to other time series data, stock market indices have typical characteristics such as regularity, wavelength, and noise. In this study, we focus on the time-series characteristics of stock indices by adopting the low-pass filter as a method of denoising data, rather than simply analyzing the index using basic deep learning. Through this research, we aim to increase the predictivity of stock price index using the ensemble model of the low-pass filter and Long Short-Term Memory algorithm (LSTM) and empirically analyze the result through KOSPI200 stock index data. The result of the studies shows that proposed model had surpassed other denoising LSTM models and simple LSTM model in every test period. In conclusion, further studies in denoising data can be resulted in improvement of prediction in financial area.

Key words : Data denoising, Long short-term algorithm, Low-pass filter, Technical indicator, Stock market index

1. Introduction

In financial market, various information exists, and each piece of information interferes with each other and forms a complex market. In modern times, the financial market is becoming more complex. These attempts are based on statistical analysis, using a time series model to predict stock indices [1], Predictability test of k-nearest neighbors (K-NN) algorithm: Application to the KOSPI200 futures [2]. And predictions of the KOSPI 200 index [3]. Recently, using deep learning techniques, stock price indices and various financial data are predicted in various ways, such as a study of the convertible bond, including a deep learning model to improve stock price prediction using RNN and LSTM [4], and genetic algorithm-based scoring model [5], various financial data such as Bond Index and the Investment Performance Using Rough Set [6], predicting debt default of P2P loan borrowers using Self-Organizing Map [7]. Attempts to analyze newly created financial data, such as P2P Lending or Bitcoin [8], are currently in line. However, the prediction of the stock price index, which is still the most important and traditional financial indicator, is also continuing, and the improvement of the algorithm continues. Time series access to financial data is

*Correspondence should be addressed to Dr. Kyong Joo Oh, Department of Information & Industrial Engineering, Yonsei University, Seoul 03722, Korea. Tel: +82-2-2123-5720, Fax: +82-2-364-7807, E-mail: johanoh@yonsei.ac.kr
DW Lee is a Master candidate, Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea. HS Lee is a professor, School of Business, Sejong University, Seoul 05006, Korea.

one of them. Recent academics have attempted to apply the various methods used in speech signaling to finance. Attempts have been made using the EEMD algorithm, the EMD algorithm, and various smoothing algorithms [9-11].

The stock price index is an index representing the flow of stock market. The stock price index is a series of time series data that is constructed over time and consists of vibrations of various periods. The stock price index is composed of noise and data, just like normal time series data. Among various methods of removing noise present on data, a low pass filter is a noise removing filter using a frequency band. This is a kind of filter that can smooth the data by canceling the vibration of the high frequency band in time series. In this study, we want to improve the model's predictive power by eliminating short periods of vibration and smoothing the data and then learning through the LSTM algorithm. Chapter 2 introduces the two key algorithms covered in the study, and Chapter 3 introduces the model to be covered in the text. Chapters 4 and 5 discuss the empirical results and conclusions of the model.

2. Methodology

In this study, we try to improve prediction accuracy of stock market index through following methodology.

2.1 Long-short term memory (LSTM)

LSTM algorithm is proposed to solve vanishing gradient problem in RNN algorithm. In RNN model, gradient is lost or diverted in learning process, which is processed through multiplication. Therefore, by adding gate with summation process, LSTM model solved vanishing/exploding gradient problem. The following figure shows the basic structure of LSTM. You can check the basic structure of LSTM algorithm consists of three gates. LSTM module is consisting of following gates, input gate that determines whether the input is reflected, a forget gate that determines whether the learning result is memorized, and an output gate that finally determines whether the result is printed.

Each operation and learning process has an activation function. As the activation function determines the degree of learning in each operation, various optimizing functions such as Softmax, Sigmoid, and Ada-grid are used in different models. In this paper, Ada-grad optimizing function is used which is the most commonly used and shows the excellent performance

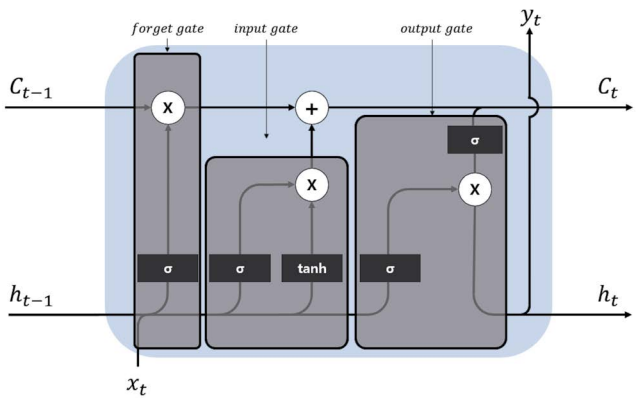

Fig. 1. Cell architecture of LSTM.

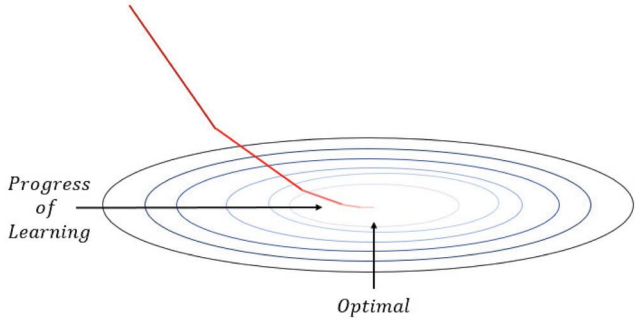

Fig. 2. Ada-grad activation function.

among various activation functions. Ada-grad function is a method of continuous learning by multiplying the slope of the previous learning and continuously progress the learning. It solves the overfitting problem by adjusting the strength of learning.

2.2 Low pass filter

Low-pass filter is a denoising filter that passes a signal of a lower frequency than a selected cutoff frequency and attenuates a signal of a higher frequency among the various frequencies existing in the data. It is generally used in almost all modern electric and electronic applications and can vary in frequency response depending on the design. It is generally used to remove noise of high frequency band in sound analysis, and the opposite concept is high-pass filter that attenuates low frequency band. In the optical domain, the meanings of low-pass and high-pass are interpreted oppositely. But in this study, the filter that plays the role of attenuating the noise of high frequency band is referred to low-pass filter.

Two types of information are required to specify the infor-

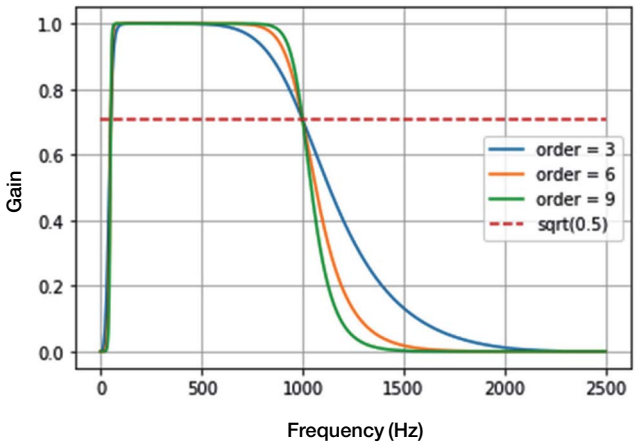

Fig. 3. Attenuation of low-pass filter.

mation of the frequency to be blocked through the low pass filter. First is the frequency range to be blocked, and the second is the level of attenuation in targeted frequency level. When the frequency used as the reference of the cutoff frequency band is Ω_c , the gain of the filter for the input frequency Ω is as follows. According to the following equation, the larger the Ω is, the smaller the output $H_a(\Omega)$, as the denominator of the following equation increases. N, which is the 'order' of the filter, is the other factor that determines the output of the filter. Order is the factor that determines the level of attenuation through filter. As order N increases in the filter, attenuation level increases as shown in Fig. 3.

$$|H_a(\Omega)|^2 = \frac{1}{1 + (\frac{\Omega}{\Omega_c})^{2N}}$$

As we mentioned, Stock index data has all the general characteristics of time series such as wavelength, regularity, and noise. Stock market index trading takes place on very small-time scales. Therefore, high-frequency noise occurs on the stock market index inevitably. If the noise generated from the high frequency data can be removed from the data, the performance of the prediction model will be improved.

3. Proposed Model

In this study, we use the low-pass filter to remove the noise of the data and to improve the prediction performance of the LSTM model. Our suggested model is as follow.

First, stock index data consists of five parts: market price, high price, low price, closing price, and trading volume. Since the target variable to be predicted in this study is the closing

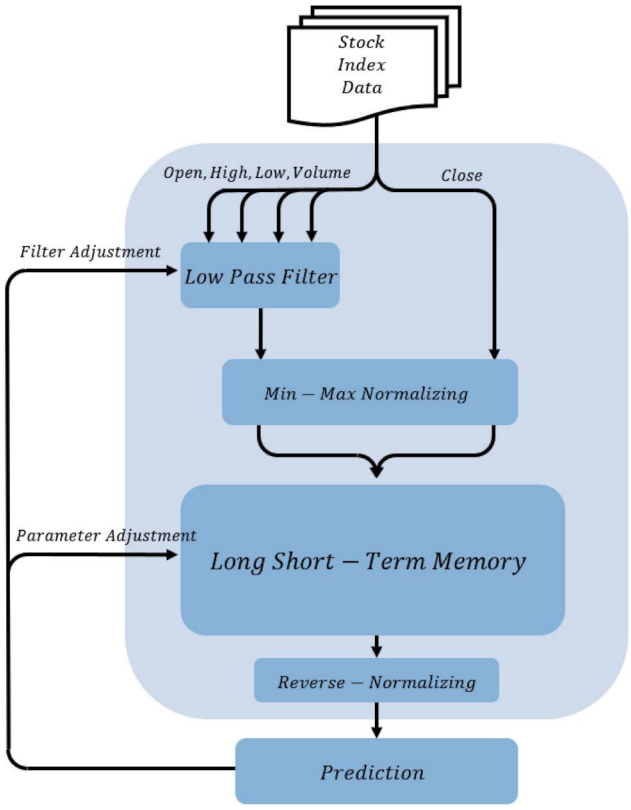

Fig. 4. Proposed model.

price of the stock index, other data except the closing price are preprocessed through the low-pass filter. The preprocessed data is trained in the LSTM model through Min-Max normalization, and then the final normalization is calculated through denormalization. The reason for normalization is that the unit of trading volume is different from other components of the stock market index data. Learning data at different scales without normalization hinders the model's ability to learn and predict. The model then repeats the training according to the preset number of epochs, whereby the gradient and parameters of the LSTM model are adjusted.

To find the effective cutoff frequency of the low-pass filter on the model, we repeated the experiment and adjusted the filter size. The cutoff frequency of the filter is set based on the data, based on 365 days, and various filter sizes are applied according to each data, so that the prediction performance of the model varies according to the filter in different experiment periods.

Hyperparameter, another indicator that determines the performance of LSTM model, is the model with the best parameter of each model based on the training rate of 0.05~0.005 and the epoch count of 2000~5000 in order to compare

benchmark results with other models.

4. Data Analysis

In this study, we use high, low, opening, closing price of stock market index of KOSPI200 index, and technical indicators. KOSPI200 Data is from 1990/09/19~2019/09/19, which is daily data. Data is collected from KOSCOM Check Expert server.

The data of each period is divided into 7 to 3 and divided into Training Set and Test Set, respectively. The model trains through the training set in each period and predicts the closing price of the KOSPI200 index from the test set based on the trained model. The LSTM model receives the number Nth of data required for prediction as a hyper parameter and calculates an N + 1th day prediction value through the determined number of data and the gradient of the model.

In each section, the order for determining the attenuation strength of the low pass filter applied to the experiment is fixed to 4. However, the cutoff frequency range and the optimal filter size is adjusted through experiment.

The experiments in this paper were conducted using the Tensorflow library in the Python environment. The version of Python used was 3.7, and the data was Min-Max scaling for learning. The data was divided into 7 : 3 and divided into training data and test data, respectively. Evaluation of the learned data was made by calculating the RMSE and MAPE. As benchmarks of the results, a simple LSTM model, a high pass filtered LSTM model, and a Savitzky-Golay filtered LSTM, which is commonly used as a denoising filter, are used.

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$$RMSE = \sqrt{\frac{\sum(A_t - F_t)^2}{n}}$$

A_t: Actual value

F_t: Predicted value

n: number of data

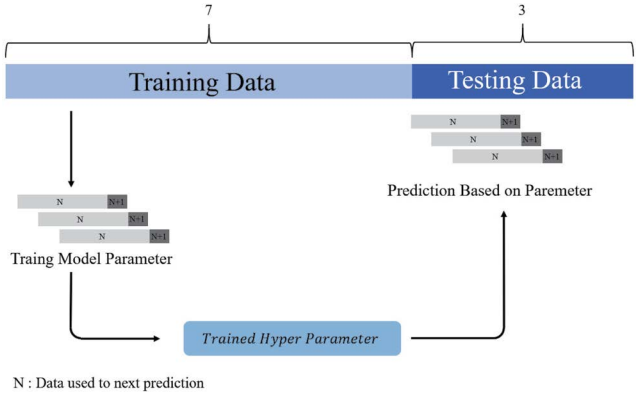

Fig. 5. Data learning and prediction.

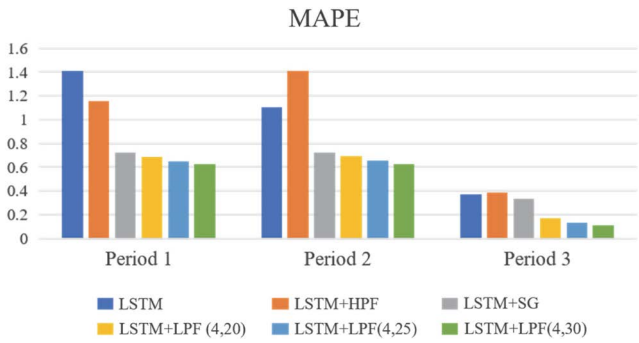

Fig. 6. MAPE of each test period.

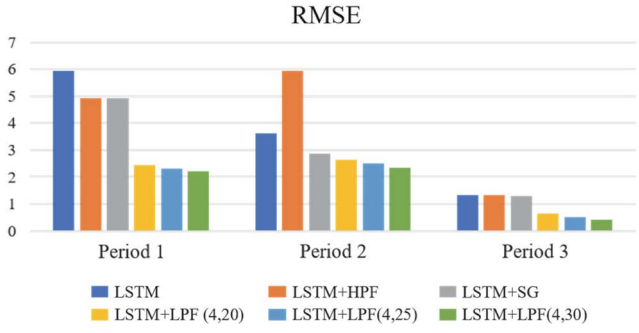

Fig. 7. RMSE of each test period.

Table 1. KOSPI200 data

	Date	Scale	Number of data	Training set	Test set
Period 1	1990/09/10~2019/09/19	1 day	7543	5280	2263
Period 2	2014/09/19~2019/09/19	1 day	1228	859	369
Period 3	2014/09/19~2019/09/19	30 min	10050	7035	3015

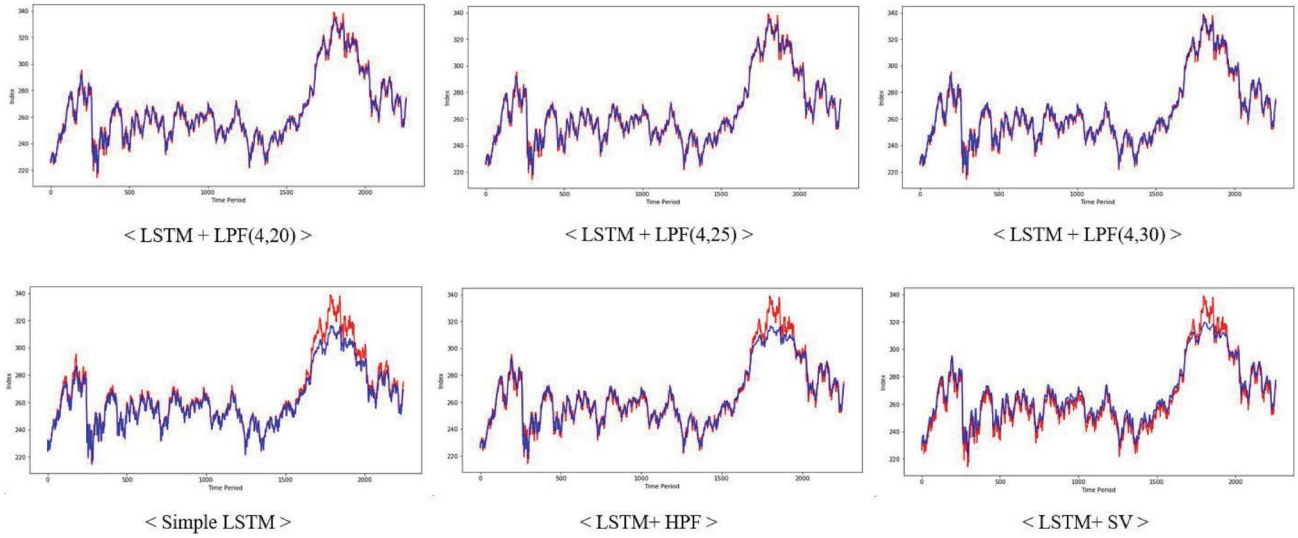

Fig. 8. Prediction graph of each model in test period of 1990/09/19～2019/09/19 daily data.

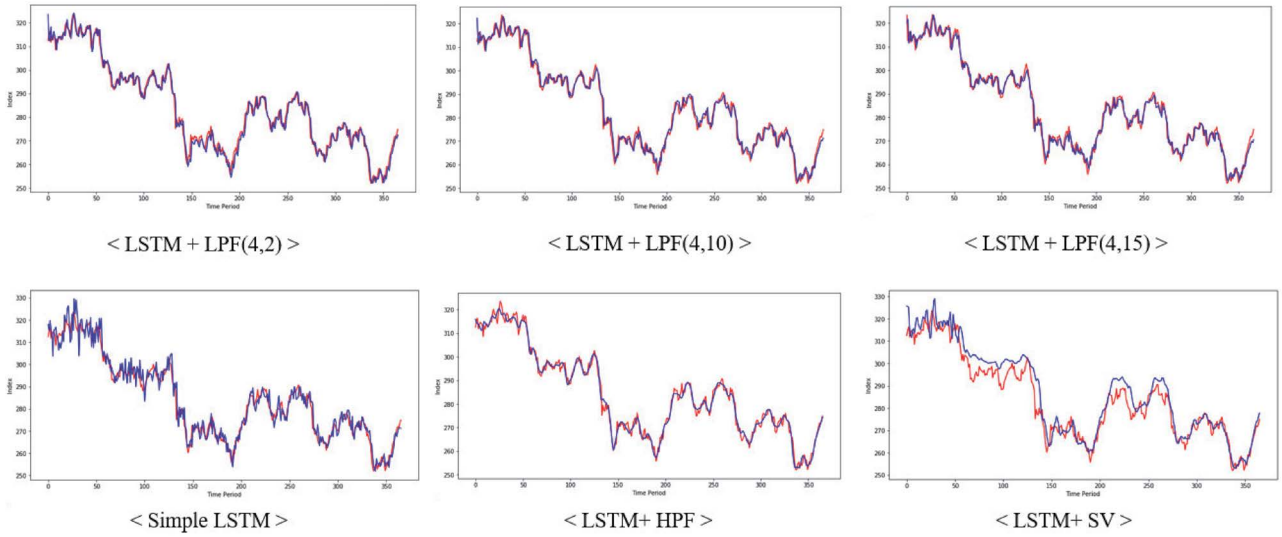

Fig. 9. Prediction graph of each model in test period of 2014/09/19～2019/09/19 daily data.

In three test periods, the model proposed in this paper showed overall good performance. It showed higher predictive power in the 30-year experiment with large number of data and the 30-minute peak test with a higher frequency of data.

Of the six models tested in the first test period, LSTM + LPF (4,15) model showed the best predictive power. Except for the proposed model, the LSTM + SG filtered model showed the greatest difference compared to the LSTM + SG Filter model, which showed a 48.6% improvement in RMSE and 54.3% in MAPE. The LPF (4,15) model had the best predictive power. The proposed model showed a big difference against LSTM + SG Filter model, which showed the highest predictive power.

In the 5-year period data experiment, the Savitzky-Golay filtered model showed the best performance among the benchmark models. Compared with the LPF model, the RMSF and MAPE of the LPF models (4,15) were improved by 18.3% and 13.2%, respectively. The gap between models was reduced compared to the experiment in 29 years period, but still shows a significant difference. Compared with other models, we can see that the performance of the Savitzky-Golay filtered model is

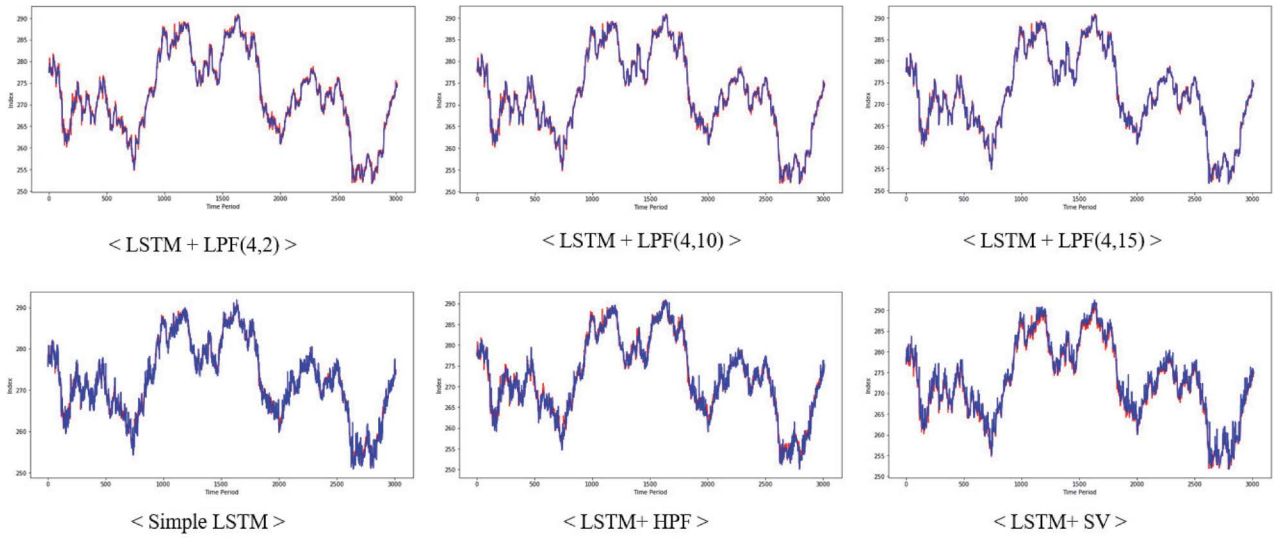

Fig. 10. Prediction graph of each model in test period of 2014/09/19~2019/09/19 30 minutes data.

Table 2. RMSE in each experiment

	LSTM	LSTM + HPF	LSTM + SV	LSTM + LPF (4,25)	LSTM + LPF (4,50)	LSTMLPF (4,100)
Period 1	5.9414	4.9164	4.9186	2.4501	2.2972	2.2069
Period 2	3.6196	5.9414	2.8829	2.6452	2.4955	2.3553
Period 3	1.3189	1.3359	1.2956	0.6315	0.5075	0.4203

Table 3. MAPE in each experiment

	LSTM	LSTM + HPF	LSTM + SV	LST + LPF(4,25)	LST + LPF(4,50)	LSTM + LPF(4,100)
Period 1	1.4140	1.1582	0.7212	0.6861	0.6495	0.6247
Period 2	1.1055	1.4141	0.7212	0.6949	0.6528	0.6261
Period 3	0.3677	0.3875	0.3332	0.1694	0.1305	0.1098

significantly improved due to the characteristics of the filter that fits the data through polynomial.

As a result, the highest level of performance improvement was found in the 30 minutes data. It was confirmed that a large volatility occurs in the three models that were used as benchmarks, whereas the proposed LPF model was able to follow the original data well without any significant variation in the predicted data. Through this, we can see that the data with higher frequency and longer term shows better performance improvement. Among the benchmark models, the Savitzky-Golay filtered model had the best predictive power, and the model with the best predictive power among the proposed models improved the RMSE by 67.6% and MAPE by 67.0%, compare to Savitzky-Golay filtered model.

5. Conclusion

This study improves the predictive performance of LSTM model by eliminating high frequency band noise in financial time series through low-pass filter, widely used in sound analysis. The predictive performance was evaluated by comparing with the similar models and the model showed a significant improvement.

The experimental results show that the proposed model shows better predictive power than the benchmark model in all three test intervals, and the performance is significantly improved compared to the general LSTM model. The model shows a greater performance improvement for high frequency, long-term data, which shows that the proposed algorithm

is less constrained by performance degradation for the length and frequency of the data.

In addition, by adjusting the filter size in the process of applying the low-pass filter, it is inferred that noise exists in the high frequency band of data. And, information of time series exists in certain band of frequency, which is inferred through the fact that the model's prediction performance is lowered when the filter size in increased and data over a certain band of frequency is lost by the filter. Based on the study, various possibilities of adjusting the order, and the bandwidth of the frequency filter can be explored through further study. In future work, many empirical studies are expected to be conducted in variety of financial data with different characteristics.

References

1. Park IC, Kwon OJ, Kim TW. Using a time series model to predict stock indices. J Korean Data Info Sci Soc 2009;28:287-295.
2. Kim MH, Lee SH, Shin DH. Predictability test of k-nearest neighbors (K-NN) algorithm: Application to the KOSPI200 futures. Korea J Bus Admin 2015;28:2613-2633.
3. Lee HS. Forecasting the Prices of KOSPI200 Index. Korea J Bus Admin 2014;27:2165-2179.
4. Shin DH, Choi KH, Kim CB. Deep learning model for prediction rate improvement of stock price using RNN and LSTM. J Korean Data Info Sci Soc 2017;15:109-116
5. Cho KH, Oh KJ. Scoring model to determine trade timing based on genetic algorithm. J Korean Data Info Sci Soc 2018;29: 735-745.
6. Lee SY, Yang JH, Jeong BJ, Oh KJ. A study of the convertible bond index and the investment performance using rough set. QBS 2019;38:23-31.
7. Park JH, Lee HJ, Oh KJ. Predicting debt default of P2P loan borrowers using Self-Organizing Map. QBS 2019;38:63-71
8. Seo YB, Hwang CH. Predicting bitcoin market trend with deep learning models. QBS 2018;37:65-71
9. Wu YX, Wu QB, Zhu JQ. Improved EEMD-based crude oil price forecasting using LSTM networks. Physica A 2019;516: 114-124
10. Xu M, Shang PJ, Lin A. Cross-correlation analysis of stock markets using EMD and EEMD. Physica A 2016;442:82-90.
11. Umer UM, Sevil T, Sevil G. Forecasting performance of Smooth Transition Auto-Regressive (STAR) model on travel and leisure stock index, J Fin Data Sci 2019;5:12-21.