Improved Stock Price Prediction by Integrating Data Mining Algorithms and Technical Indicators: A Case Study on Dhaka Stock Exchange

Syeda Shabnam Hasan, Rashida Rahman, Noel Mannan, Haymontee Khan, Jebun Nahar Moni, and Rashedur M. Rahman(✉)

North South University,
Plot 15, Block B, Bashundhara R/A, Dhaka 1229, Bangladesh
Shabnam.4444@yahoo.com, flgn@ymail.com,
{noel.mannan,haymontee.khan,
rashedur.rahman}@northsouth.edu, jnmoni@gmail.com

Abstract. This paper employs a number of machine learning algorithms to predict the future stock price of Dhaka Stock Exchange. The outcomes of the different machine learning algorithms are combined to form an ensemble to improve the prediction accuracy. In addition, two popular and widely used technical indicators are combined with the machine learning algorithms to further improve the prediction performance. To evaluate the proposed techniques, historical price and volume data over the past 15 months of three prominent stocks enlisted in Dhaka Stock Exchange are collected, which are used as training and test data for the algorithms to predict the 1-day, 1-week and 1-month-ahead prices of these stocks. The predictions are made both on training and test data sets and results are compared with other existing machine learning algorithms. The results indicate that the proposed ensemble approach as well as the combination of technical indicators with the machine learning algorithms can often provide better results, with reduced overall prediction error compared to many other existing prediction algorithms.

Keywords: Stock prediction · Machine learning · Regression algorithms · Time series forecast · Technical indicators

1 Introduction

The stock market is considered to be extremely important for the socio-economic stability and progress as well as the overall well-being of the people and the country. Stock investors all over the world collect historical stock price data and analyze them using automated software tools to predict the future stock price. Making accurate prediction is always extremely challenging [1], because of the high degree of non-linearity, uncertainty and volatility [2, 3] present in every stock market all over the world. The individual stocks as well as the overall stock market never follows a straight path, because they operate against a backdrop of continuous noise, news and rumor,

N.T. Nguyen et al. (Eds.): ICCCI 2017, Part I, LNAI 10448, pp. 288–297, 2017.
DOI: 10.1007/978-3-319-67074-4_28

various social, economic, political, geopolitical and miscellaneous factors that affect the stock market every day. In the backdrop of such enormous uncertainty and volatility, the prediction of a particular stock price is always challenging. The Dhaka Stock Exchange (DSE) is no exception, where situation may even become worse because many general investors invest and participate without proper plan, preparation and knowledge that often results in loss of their hard-earned valuable money.

The objective of this work is to introduce a useful guideline for the general investors of DSE on how to employ machine learning tools and algorithms along with technical indicators to predict the future stock price. We have also shown how to combine different prediction algorithms to form an effective ensemble that can predict stock price with higher accuracy. We have selected three prominent and representative stocks of DSE and demonstrated how machine learning algorithms combined with technical indicators can provide improved prediction accuracy on their future price.

The rest of this paper is organized as follows. Section 2 presents a number of recent related works on stock price prediction. Section 3 describes the proposed technique for ensemble design and two popular indicators — MACD and RSI. Section 4 describes the data source, experiment analysis. Finally Sect. 5 concludes the paper with a brief summary and suggestions for future study.

2 Existing Algorithms

To analyze and predict the future behavior of stock markets the most popular techniques employ various technical tools and parametric models in combination [6, 7]. Three most common machine learning approach are Multilayer Perceptron (MLP) or Artificial Neural Network (ANN), Support Vector Machine (SVM) and Gaussian Process based regression (GPR) [4, 5], all of which are widely used for analysis and prediction of time series data [4–8]. Although SVM based regressions (SVR) are the most widely used algorithm for time series prediction, there are some instances where MLP and Gaussian process are reported to perform better than SVR on the benchmark problems [8, 9].

The prediction made by SVR is sufficiently accurate for many prediction tasks, especially with small amount of look-ahead into the future [9–11]. An in-depth comparison between SVR and MLP is presented in [3], which is evaluated on the Hong Kong stock market to predict 1-week and 1-month price predictions. Like most other existing works, the authors in [3] made use of several technical indicators in combination with MLP and SVR.

None of the above works made use of ensemble to combine multiple machine learning algorithms. There exist very few works, e.g. [12], that emphasizes the use of ensembles to combine multiple machine learning algorithms. However, none of these papers deal with the Dhaka Stock Exchange data. We have found only a few recent works (e.g., [13, 14]) that employ machine learning algorithms on the DSE data set. In [13], the authors employed MLP and fuzzy algorithms to predict stock price in DSE, while the authors in [14] used a number of machine learning techniques for clustering, classification and regression on 30 selected stocks in DSE. However, none of them employ or emphasize an ensemble approach, as proposed in this paper. The unique contribution of this paper is to propose an ensemble architecture that combines four different machine learning algorithms and two popular technical indicators, as elaborated in the following section.

3 Proposed Methodology

The objective of our paper is to introduce a combined machine learning approach with technical indicators to predict future price of particular stocks in the Dhaka Stock Exchange. We have selected four widely used time series prediction algorithms to predict stock price and proposed four different ways to combine the outcomes of the different algorithms. We have divided our work into following three discrete parts:

3.1 Employ Existing Machine Learning Algorithms to Predict Future Stock Price

We have selected four different machine learning algorithms on time series prediction for the task of stock price prediction. The algorithms are- Support Vector Machine Based Regression (SVR), Regression by Multilayer Perceptron (MLP), Linear Regression (LNR) and Gaussian Process based Regression (GPR). These algorithms are provided by the popular machine learning software suite WEKA (Waikato Environment for Knowledge Analysis).

3.2 Ensemble Approach to Combine Multiple Predictions

In order to improve the prediction accuracy of the four machine learning algorithms, we plan to combine their predictions to form an Ensemble. The ensemble will combine the outcomes of its four component predictors using the following methods- Simple Averaging (SA), Weighted Averaging (WA), Voting by SA (VSA) and Voting by WA (VWA). Assuming the predicted stock price are O_1, O_2, O_3 and O_4, the combined output from the ensemble by the different combination methods are as follows.

Ensemble output by Simple Averaging,

$$O_{SA} = \frac{\sum_{i=1}^{n} O_i}{n}$$

Ensemble Output by Weighted Averaging,

$$O_{WA} = \frac{\sum_{i=1}^{n} w_i * O_i}{n}$$

Ensemble output from Voting by Simple Averaging,

$$O_{VSA} = O_j, \text{ where } j = \arg\min_i |O_{SA} - O_i|$$

Ensemble output by Voting by weighted averaging (O_{VWA}),

$$O_{VWA} = O_k, \text{ where } k = \arg\min_i |O_{WA} - O_i|$$

Here, n is the number of component predictors in the ensemble, which is set as $n = 4$. The w_i's in O_{WA} denote the weight values for weighted averaging, which are set as follows: the weight value w_i for a particular predictor algorithm's output O_i is set to the inverse of the Mean Absolute Error (*MAE*) value of the corresponding predictor algorithm on the training data set. Since the outcomes by SVM, MLP, LNR and GPR are denoted by O_1, O_2, O_3 and O_4 respectively, their weight values are set as:

$$w_1 = \frac{1}{MAE_{SVR}},\ w_2 = \frac{1}{MAE_{MLP}},\ w_3 = \frac{1}{MAE_{LNR}},\ w_4 = \frac{1}{MAE_{GPR}}.$$

The selection of one of the four algorithms is made for each row of the training and test sets based on which of the four algorithms produces minimum distance from the computed simple average (SA) and weighted average (WA) values of the four algorithms on the current row. This means for some rows the output of SMOReg may be picked by the ensemble, while for some other rows the output of Perceptron, Linear Regression or Gaussian Process based Regression is picked based on which of them can produce the closest output to the computed SA or WA value. This is why the result of VSA and VWA may not exactly match with one of the four component algorithms or the best performing algorithm, as seen in the Tables 1, 2 and 3. Besides, please note that average RMSE of four methods will not be equal to the RMSE of SA method. This is because for each row the average is done and the error is calculated based on this row's average. This is repeated for a set of 15 months of stock data and RMSE is calculated for SA method.

The motivation behind the ensemble approach is to improve the prediction accuracy by combining the outcomes of the different predictors. Since each predictor has its individual strengths and weaknesses, their effective combination is expected to produce improved prediction performance (i.e., reduced overall *MAE* value).

3.3 Combine Technical Indicators with Machine Learning Algorithms

There exist many technical indicators that reflect and predict different characteristics about the movements of stock price. All these indicators can be divided into two broad categories- Lagging indicators and Leading indicators. A Lagging indicator changes its value after some actual change appears it the stock price and/or volume. In contrast, the Leading indicators change their values before any actual change occurs in the stock's price and volume.

For our research, we have incorporated two technical indicators, one from the leading indicators, Relative Strength Index (RSI) and the other from lagging indicators, Moving Average Convergence Divergence (MACD), with the four machine learning algorithms. MACD is a lagging indicator that follows the ongoing price trend. The MACD value is plotted with a signal line. The crossover of MACD with the signal line gives bullish (buy) or bearish (sell) signal. Whenever MACD drops below the signal line from above, it gives a bearish signal. Similarly, when the MACD value rises above the signal line from below, it gives a bullish signal for the traders. In addition to the crossover signals, MACD may also produce trend reversal signal by divergence from the stock price. RSI is a leading indicator that measures the speed of price movements and predicts possible reversal of current trend by the change of speed of price movements. The value of RSI moves between 0 and 100. Usually a stock is considered overbought if the RSI is above 70, and oversold when RSI is below 30. If the RSI pattern diverges from the price pattern, it indicates a weakness and possible reversal of the price trend.

4 Experimental Study

4.1 Source of Data

The historical data of Dhaka Stock Exchange (DSE) is maintained by its official web site http://www.dsebd.org. However, more organized data sets are readily available at some financial websites, like the http://www.stockbangladesh.com and also, the http://www.lankabd.com. All these data sets can be readily downloaded in text (*.txt), excel (*.xls) and Comma Separated Values (*.csv) format. Currently, DSE enlists more than 400 companies from 22 different financial sectors. From them, three prominent companies have been selected to evaluate our proposed techniques, which are- ACI, Beximco, and GP. For each company, its daily data is collected for the past 15 months (since 01-Jan-2015). The daily data consists of six attributes — Date, Open, High, Low, Close, and Volume. They represent the date, daily open price, high and low price, closing price and volume of a particular stock on that day. The data set has been divided into training set (90%) and test set (remaining 10%). The proposed methods have been evaluated on both the training and test sets. A partial snapshot of the dataset of ACI is presented in Fig. 1.

	A	B	C	D	E	F
1	Date	Open	High	Low	Close	Volume
2	1/3/2016	562.1	575.4	561.9	567.5	115659
3	1/4/2016	569.6	570.9	561.1	561.3	92543
4	1/5/2016	561.7	562	549.9	555.1	192960
5	1/6/2016	557	562.4	550	550.6	127394
6	1/7/2016	550.6	561.7	550.6	557.9	92518
7	1/10/2016	558	561	548.8	551.3	79864
8	1/11/2016	552	554.6	542	544.4	67720
9	1/12/2016	550.3	550.3	540	542.1	57903
10	1/13/2016	545	548.5	538	540.3	113020
11	1/14/2016	542	557.9	540.3	555.2	95556

Fig. 1. ACI dataset over 10 working days.

4.2 Parameter Setup

The default framework of Weka does not provide any forecasting tool. We have installed an additional package (Time Series Forecasting Environment) into Weka which provides additional support for automated regression schemes by creating lagged variables and date-derived periodic variables. The package provides algorithms for Support Vector Machine based Regression, Perceptron based Regression, Linear Regression and Gaussian Process based Regression. We have used the default parameters for these algorithms for this experimental study.

4.3 Experiments and Analysis of Results

We have conducted three different set of experiments for stock price prediction using four machine learning algorithms, which are briefly described below.

Basic Machine Learning Algorithms: We have used four basic machine learning algorithms to predict the future price of three stocks – ACI, Beximco and GP. Results are presented in the Tables 1, 2 and 3, under "Basic Algorithms".

Ensemble Design: We have combined the previously used algorithms to form the ensemble in four different ways- Simple Averaging, Weighted Averaging, Voting by SA, and Voting by WA. The results are shown in Tables 1, 2 and 3 under "Ensemble".

Integrating Technical Indicators: We have included two major technical indicators – the MACD and RSI as additional columns into the data set for ACI. The resulting data set is used by the basic machine learning algorithms to predict the future price of ACI, which is shown in Tables 4 and 5. Please note that, in the 'Target Selection' list box of the 'Forecast' tab of WEKA, we have selected all of Open, High, Low and Close prices as the 'Targets' of the prediction algorithms for the results in these Tables 4 and 5, which is quite different from Tables 1, 2 and 3, where we have selected only the Close price as the 'Target' in WEKA. Since the Open, High, Low and Close prices are intricately related, predicting them together results in smaller RMSE error value, as demonstrated by the 5-day and 22-day RMSE values in Tables 4 and 5, compared to the much higher RMSE error value of the same stock over the same prediction-lengths in the previous Table 1.

Table 1. RMSE of 1-day, 5-day and 22-day-ahead predictions on ACI dataset

Data set	Forecast length	Basic algorithms				Ensemble			
		SMOReg	Perceptron	Linear regression	Gaussian process	Simple averaging	Weighted averaging	Voting by simple avg.	Voting by weighted avg.
Training set	1-day	6.5	7.4	6.5	10.8	6.6	**6.4**	6.6	6.7
	5-day	15	17.4	15.7	19.3	**12.9**	13	14	14.2
	22-day	24.9	114.3	52.5	**20.7**	36.9	24.3	36.6	22.9
Test set	1-day	13.9	27.6	18.8	40.2	**12.4**	13.2	13.7	13.7
	5-day	60.7	118.9	111.5	84.7	51.7	52.7	51	**50.9**
	22-day	120.2	142.2	2077	60.4	71.6	**35.6**	120.1	60.3
Mean RMSE		40.2	71.3	380.4	39.4	32	**24.2**	40.3	28.1
		132.8				**31.2**			

Table 2. RMSE of 1-day, 5-day and 22-day-ahead predictions on Beximco dataset

Data set	Forecast length	Basic algorithms				Ensemble			
		SMOReg	Perceptron	Linear regression	Gaussian process	Simple averaging	Weighted averaging	Voting by simple avg.	Voting by weighted avg.
Training set	1-day	0.5	0.5	**0.4**	0.6	0.5	**0.4**	0.5	**0.4**
	5-day	0.9	1.3	1.1	1.3	**0.8**	**0.8**	0.9	**0.8**
	22-day	1.3	3.6	**0.9**	1.6	1.0	**0.9**	1.0	1.0
Test set	1-day	**1.2**	5.0	1.6	3.7	1.4	1.4	1.4	1.4
	5-day	2.8	8.3	6.3	10.4	**2.0**	2.5	2.8	2.8
	22-day	8.6	8.2	27.7	5.3	4.7	**4.3**	7.6	7.6
Mean RMSE		2.6	4.5	6.3	3.8	**1.7**	**1.7**	2.4	2.3
		4.3				**2.0**			

Table 3. RMSE of 1-day, 5-day and 22-day-ahead predictions on Grameen Phone (GP) dataset

Data set	Forecast length	Basic algorithms				Ensemble			
		SMOReg	Perceptron	Linear regression	Gaussian process	Simple averaging	Weighted averaging	Voting by simple avg.	Voting by weighted avg.
Training set	1-day	2.4	3.0	2.3	8.5	2.9	2.3	2.4	**2.2**
	5-day	**4.6**	10.1	**4.6**	15.3	5.5	**4.6**	5.1	4.8
	22-day	6.9	52.5	**6.4**	10.9	14.6	7.5	7.7	7.3
Test set	1-day	7.8	14.6	4.7	4.9	5.5	6.0	**4.5**	4.6
	5-day	19.6	33.6	16.0	16.6	**15.2**	**15.2**	15.4	20.6
	22-day	23.2	122.7	33.3	**9.3**	21.9	24.1	23.2	23.2
Mean RMSE		10.8	39.4	11.2	10.9	10.9	9.9	**9.7**	10.4
		18.1				**10.3**			

Table 4. Comparison of SMOReg and Linear regression, before and after combining MACD with them.

Data set	Forecast length	Algorithm			
		SMOReg	SMOReg with MACD	Linear regression	Linear reg. with MACD
Training set	1-day	6.5	**6.4**	6.5	**6.1**
	5-day	15.0	**14.6**	15.7	**13.9**
	22-day	24.9	**22.5**	52.5	**28.3**
Test set	1-day	12.0	12.0	18.8	**13.7**
	5-day	32.3	**28.0**	111.5	**42.4**
	22-day	**79.3**	85.3	2039.0	**264.6**
MEAN RMSE		28.3	**28.1**	374.0	**61.5**

Analysis of Results: We have made a number of observations, which are summarized in the following few points.

1. Among the four basic machine learning algorithms, Support Vector Machine based Regression (SMOReg) performs overall best, especially for two stocks- Beximco and GP (Tables VI and VII, with Root Mean Squared Error, i.e., RMSE of 2.6 and 10.8 for these two stocks), followed by the Gaussian Process based Regression.

Table 5. Comparison of SMOReg and Multilayer Perceptron, before and after combining RSI with them.

Data set	Forecast length	Algorithm			
		SMOReg	SMOReg with RSI	Multilayer perceptron	Multilayer perceptron with RSI
Training set	1-day	6.5	6.5	7.4	**6.4**
	5-day	15.0	**14.7**	17.4	**12.2**
	22-day	24.9	**24.4**	114.3	**12.9**
Test set	1-day	12.0	12.0	27.6	**18.3**
	5-day	32.3	**30.5**	118.9	**91.6**
	22-day	79.3	**73.8**	101.9	**109.1**
MEAN RMSE		28.3	**27.0**	64.6	**41.8**

2. Among the four different ensemble approach, the weighted averaging performs overall best, followed by the voting by weighted averaging technique.
3. If we compare the basic algorithms with the ensemble techniques, we can easily observe much improved results by the ensemble methods. For example, ensembles produce mean RMSE = 31.2 for the ACI dataset, compared to the much larger RMSE = 132.8 by the basic algorithms. Similarly, for Beximco, and GP data sets, ensemble produce overall RMSE of only 2.0 and 10.3, compared to the much higher 4.3 and 18.1 respectively.
4. The overall better prediction performance by the ensemble approach is demonstrated. The RMSE values of the four basic algorithms are much higher compared to the smaller values of the ensemble techniques.
5. The predictions made by the four ensemble techniques closely follow the actual price pattern, as demonstrated by Fig. 2.

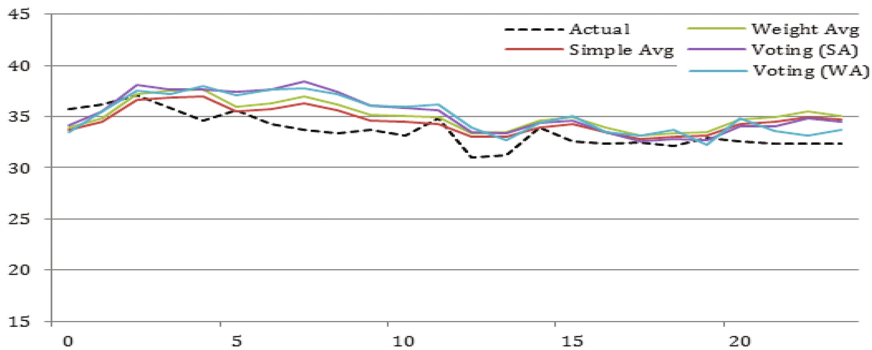

Fig. 2. Plot of actual price and predicted price of the Beximco data set by the four ensemble techniques.

6. When the technical indicators MACD and RSI are combined with the machine learning techniques, the prediction accuracy is improved, as demonstrated by lower
7. RMSE value in Tables 4 and 5. After including MACD with SMOReg and Linear Regression, the RMSE value drops to 28.1 and 61.5 from previous higher 28.3 and 374.0, respectively (Table 4). Inclusion of RSI reduces the RMSE error value into 27.0 and 41.8, compared to RMSE of 28.3 and 64.6, respectively (Table 5).
8. The improved prediction performance by including technical indicators MACD and RSI is also demonstrated. The RMSE values are much higher without MACD and RSI, but significantly smaller when MACD and RSI are included with SMOReg, Linear Regression and Multilayer Perceptron.
9. Summarizing our observations, we can say that the overall prediction performance of machine learning algorithms can be improved further by combining the outcomes of multiple algorithms into an ensemble, as well as by combining them with technical indicators for stock price.

5 Conclusion and Future Work

This study applies a number of machine learning algorithms on stock price prediction of Dhaka Stock Exchange (DSE). The experimental results with three prominent stocks of DSE show that the outcomes of different algorithms may be used to form an ensemble of predictors, which significantly improves the prediction accuracy. Further experiments reveal that inclusion of technical indicators, such as MACD and RSI, may further improve the prediction performance. Further research may include more ensemble designing techniques, use many other technical indicators and employ more machine learning algorithms, which would provide the researchers with more insights and better algorithms on stock price prediction, especially the DSE listed ones.

References

1. Kim, K.: Financial time series forecasting using support vector machines. Neurocomputing **55**, 307–319 (2003)
2. Lu, C., Chang, C., Chen, C., Chiu, C., Lee, T.: Stock index prediction: a comparison of MARS, BPN and SVR in an emerging market. In: Proceedings of the IEEE IEEM, pp. 2343–2347 (2009)
3. Lucas, K., Lai, C., James, N., Liu, K.: Stock forecasting using support vector machine. In: Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, pp. 1607–1614 (2010)
4. Ince, H., Trafalis, T.B.: Kernel principal component analysis and support vector machines for stock price prediction, pp. 2053–2058 (2004)
5. Kannan, K.S., Sekar, P.S., Sathik, M.M., Arumugam, P.: Financial stock market forecast using data mining techniques. In: Proceedings of the International Multiconference of Engineers and Computer Scientists, pp. 555–559 (2010)
6. Hu, Y., Pang, J.: Financial crisis early warning based on support vector machine. In: International Joint Conference on Neural Networks, pp. 2435–2440 (2008)

7. Chen, K.-Y., Ho, C.-H.: An improved support vector regression modeling for Taiwan Stock Exchange market weighted index forecasting. In: The IEEE International Conference on Neural Networks and Brain, pp. 1633–1638 (2005)
8. Xue-shen, S., Zhong-ying, Q., Da-ren, Y., Qing-hua, H., Hui, Z.: A novel feature selection approach using classification complexity for SVM of stock market trend prediction. In: 14th International Conference on Management Science & Engineering, pp. 1654–1659 (2007)
9. Debasish, B., Srimanta, P., Dipak, C.P.: Support vector regression. Neural Inf. Process. Lett. Rev. **11**(10), 203–224 (2007)
10. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A practical guide to support vector classification. Initial version: 2003, last updated version: 2010
11. Kazema, A., Sharifia, E., Hussainb, F.K., Saberic, M., Hussaind, O.K.: Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Appl. Soft Comput. **13**, 947–958 (2013)
12. Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R.: Evaluating multiple classifiers for stock price direction prediction. Expert Syst. Appl. Int. J. **42**(20), 7046–7056 (2015)
13. Billah, M., Waheed, S., Hanifa, A.: Predicting closing stock price using artificial neural network and adaptive neuro-fuzzy inference system (ANFIS): the case of the Dhaka Stock Exchange. Int. J. Comput. Appl. (0975-8887) **129**(11), 1–5 (2015)
14. Shadman, A.I., Towqir, S.S., Akif, M.A., Imtiaz, M., Rahman, R.M.: Cluster analysis, classification and forecasting tool on DS30 for better investment decision. In: Akagi, M., Nguyen, T.-T., Vu, D.-T., Phung, T.-N., Huynh, V.-N. (eds.) ICTA 2016. AISC, vol. 538, pp. 197–206. Springer, Cham (2017). doi:10.1007/978-3-319-49073-1_22