Asian Journal of Economics and Banking

ISSN 2588-1396

http://ajeb.buh.edu.vn/Home

A Technique to Predict Short-term Stock Trend Using Bayesian Classifier

Ho Vu[1], T. Vo Van[2], N. Nguyen-Minh[4], and T. Nguyen-Trang[3,4,†]

[1] Faculty of Mathematical Economics, Banking University of Ho Chi Minh City, Vietnam

[2] Department of Mathematics, Can Tho University, Can Tho, Vietnam

[3] Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[4] Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Article Info

Received: 24/02/2019
Accepted: 24/06/2019
Available online: In Press

Keywords

Bayesian Classifier, ROC curve

JEL classification
C11, C15, C3

Abstract

In this paper, an application of Bayesian classifier for short-term stock trend prediction, which is a popular field of study, is presented. In order to use Bayesian classifier effectively, we transform daily stock price time series object into data frame format where the dependent variable is stock trend label and the independent variables are the stock variations with respect to previous days. The numerical example using stock market data of individual firms demonstrates the potential of the proposed method in predicting the short-term stock trend. In addition, to reduce the risk for the investor, a method to adjust the probability threshold using the ROC curve is investigated. Also, it can be implied that the performance of the new technique mainly depends on the skill of investors, such as adjusting the threshold, identifying the suitable stock and the suitable time for trading, combining the proposed technique with other tools of fundamental analysis and technical analysis, etc.

[†]Corresponding author: nguyentrangthao@tdtu.edu.vn

1 INTRODUCTION

Recently, along with the increasing of the number of joint stock companies, the stock market has become more and more vibrant; and therefore, stock investing has been a popular field of study [5,6,16]. In general, there are two major stock investing strategies consisting of technical analysis and fundamental analysis [23]. Fundamental analysis is mainly used for long-term investment by checking a company's financial features, such as average equity, average asset, sales cost, revenues, operating profit, and net income, etc. [10, 19, 28]. Some of the recent fundamental analysis strategies include the mean-variance model [15], the data envelopment analysis [6, 11, 30], and the ordered weighted averaging operator [2, 10]. Long-term investment can create a sustainable business, and therefore it is encouraged for investors, but it takes a long time for investors to generate profit. In addition to fundamental analysis, investors are also interested in technical analysis to get short-term profit [23]. Instead of analyzing the financial statements, technical analysis focuses more on historical price trend and tries to consider some crucial signs for predicting short-term stock trend. There are many simple technical analysis methods, such as chart analysis [7, 20, 24], and complex methods such as: time series, machine learning, neural network, etc. [9,12,14,18,25,29]. In general, although there are plenty of technical analysis algorithms, the main purpose is to identify peaks and troughs so that investors can "buy at the low and sell at the high" [3,8,27].

In short-term investment, predicting the stock trend is more important than predicting the stock values. As shown in Figure 1, the black line represents the actual value of the stock, the red line and blue line represent the predictions of Method 1 and Method 2, respectively. Method 1 results in an error of 2 and Method 2 results in an error of 2.5 compared to the actual value. Based on the error value, investors may follow Method 1, but this can lead to serious mistakes. In fact, Method 1 gives a lower error than Method 2 but it completely mispredicted the trend of the stock. Using Method 1, the investors might still hold on the stock at the time point t and expect further upmove. However, the stock market peak occurred at the time point t and fell at time point t+1, which leads to a loss. For Method 2, although it results in lower performance in terms of predicting the stock value, it is capable of capturing the stock price trend. Therefore, the investors might sell the stock at the peak when using Method 2. Thus, it can be believed that accurately predicting the stock trend is more important than approximating the stock price and can be well applied to the short-term investment.

In order to accurately predict the stock trend, we need to compute the variations or the first order differences of the stock values rather than the original stock values. As shown in Figure 2, when the current stock price is 1, the stock price in the next time points can rise and fall, arbitrarily. In contrast, if we are interested in the fluc-

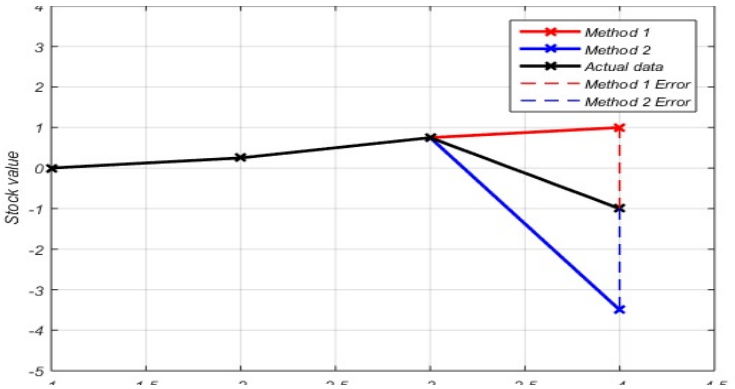

Fig. 1. The prediction of the two methods

tuation of n days before the predicted time, some interesting rules can be discovered. For example, as shown in Figure 2, if the stock price fell in the two previous days (the first order difference < 0), the stock price will rise in the current day; also, if the stock price rose the two previous days, the stock price will fall in the current day. The mentioned rules are also consistent with which we believe that when the stock price has fallen/risen for a few days, it will find the support/resistance and reverse. In fact, the found rules will be more complex and also contains uncertainty.

According to the above discussion, this paper introduces a method to predict the short-term stock trend based on the first order difference of stock price. Specifically, the independent variables are the first order differences of stock prices of n days before the predicted time and the binary dependent variable represents the rise/fall of the stock. For this purpose, the time series collected in the past would be transformed into a data frame and then would be trained by a supervised learning model. In this paper, through a literature survey, we use the Bayesian classifier because it not only can classify the data but also provides the predictive probability of classification, which helps us can evaluate the reliability of the predicted result [1, 4, 17, 22, 26].

The rest of this paper is presented as follow: Section 2 presents the Bayesian classifier. Section 3 presents the proposed method. The experiments are presented in Section 4. Finally is the conclusion.

2 BAYESIAN CLASSIFIER

We consider k classes $w_1, w_2 \ldots, w_k$, with the prior probability $q_i, i = \overline{1,k}$,$X = \{X_1, X_2 \ldots, X_n\}$ is the n-dimensional continuous data with $x = \{x_1, x_2 \ldots, x_n\}$ is a specific sample. Let w_i be the $i-th$ class, according to [17, 21]:

IF $P(w_i|x) > P(w_j|x)$ for $1 \leqslant j \leqslant k, j \neq i$, THEN x belongs to the class w_i. (1)

In the continuous case, $P(w_i|x)$ could be calculated by:

$$P(w_i|x) = \frac{P(w_i)f(x|w_i)}{\sum_{i=1}^{n} P(w_i)f(x|w_i)} = \frac{q_i f_i(x)}{f(x)}$$

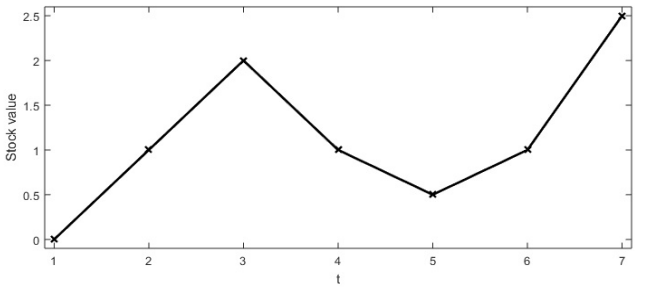

Fig. 2. A time series of stock

Because $f(x)$ is the same for all classes, the classification's rule is:

IF $q_i f_i(x) > q_j f_j(x), \forall j \neq i$, THEN x belongs to the class w_i. (2)

In (2), q_i, and $f_i(x)$ is the prior probability and the probability density function of class i, respectively.

In the case of two classes like the stock trend prediction, we the following decision rule:

IF $P(w_1|x) > 0.5$ THEN x belongs to the class w_1, ELSE x belongs to the class w_2 (3)

3 THE PROPOSED FRAMEWORK

Normally, we can collect day-by-day stock prices represented by a time series. Let $x(t)$ is the time series data representing stock prices by the time point t, in order to use the Bayesian classifier effectively, pre-processing of the data is very much essential. For predicting the stock trend, we need more information about independent and dependent variables. In that case, the independent variables are the first order differences of stock prices of n days before the predicted time where the first order difference $v(t)$ at the time point t is calculated by $v(t) := x(t) - x(t-1)$, and the dependent variable is binary, that is, $Y(t) = 1$ when the stock prices rise and vice versa. The data representation is carried out using Algorithm 1, which transforms a time series into a tabular representation form so that the data is suitable for supervised learning.

Algorithm 1: Given historical data $X(t), t = 1 : N$, with $x(t)$ is the specific value of $X(t)$ at time t, N is the length of the original time series, Algorithm 1 transforms the time series data to tabular data, which is generally suitable for supervised learning.

INPUT: $X(t)$
FOR $t = 2 : N$
Compute the variation or the first order difference: $v(t) := x(t) - x(t-1)$
ENDFOR
FOR $t = 3 : N$
IF $v(t+1) > 0$
$Y(t) := 1$
ELSE
$Y(t) := 0$
ENDFOR
TrainingData
$= [v(t), v(t-1), \ldots, Y(t)]$,
$t = 3 : N - 1$
OUTPUT: Training Data.

After processing the data, we use the tabular data to build the Bayesian classifier to predict the stock trend. This

process is summarized in Algorithm 2. Algorithm 2: Given training data, this algorithm computes the probability of rise/fall of the stock price at time $t+1$; thereby classifying the stock into one of the two classes.

INPUT: Training data.

Build the Bayesian classifier.

Compute $P(1|X)$ with X is the set of variation before the predicted time point.

IF :$P(1|X) > \Delta$.

The stock price will rise at time $t+1$.

ELSE

The stock price will fall at time $t+1$.

ENDIF

OUTPUT: Class of stock?s rise and fall.

4 NUMERICAL EXAMPLES

4.1 Evaluating the Performance

In this section, a number of examples are presented to evaluate the performance of the proposed framework in predicting the stock trend. The two stocks consisting of NSC (Vietnam National Seed Joint Stock Company) and LPB (Lien Viet Post Joint Stock Commercial Bank) are collected from May 2, 2018 to August 10, 2018. For the test set, we use the stock prices from July 30, 2018 to August 10, 2018. We first have to apply the Algorithm 1 to the training data and build the Bayesian model on the output tabular data. Then, we evaluate the performance of the Bayesian model according to the accuracy on the test set. In this case, the test set plays a role as the actual data because it had not been included when building Bayes classifier until it was predicted. In addition, because the proposed method is applied to predicting in the short-term time, the long-term data may not be suitable in reality. Therefore, when predicting the stock trend at time t, only the variations from time point t-1 to time point t-60 are used to build the training set. In other words, the training set is dynamic by the time. Also it can be noticed that the model can work with arbitrary training sample size, e.g. 50. The problem of training sample size as well as the problem of variable selection (how many days before the predicted time should be used) can be further investigated, however, it is out of the scope of the paper, which focuses on introducing a new technical approach. Therefore, as a case study, we use a training sample size of 60 and two independent variables in this paper. In these examples, the variations of two days before the predicted time points are used as the independent variables, and the binary dependent variable represents the rise or fall of stock with a probability threshold Δ of 0.5. Figure 3 shows the candlestick chart of the LPB stock, where the candle's high and the candle's low represent the highest and lowest prices; the bottom and top of the candle's body represent either the open or close prices; a green candlestick means that the close price is higher than the open price and vice versa for a red candle stick.

For the purpose of data understanding, we need to perform the distribution of data in two classes by scatter plot and compute their probability density functions, as shown in Figure 4 and Figure 5.

Fig. 3. The candlestick chart of the LPB stock code

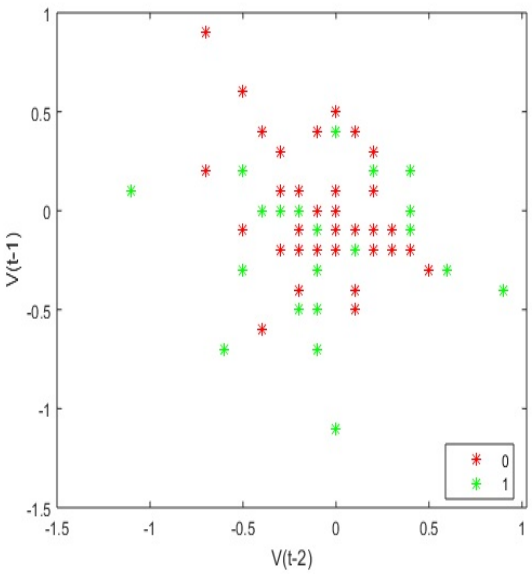

Fig. 4. The scatter plot of data in two classes

Table 1. The classification performance (%) in the case of LPB stock

	True: 0	**True: 1**
Predicted as: 0	77.78	22.22
Predicted as: 1	0.00	0.00
The total accuracy	77.78	

Using the test set for validation, we obtain the classification result. As shown in Table 1, in the case of stock falling, the proposed framework is completely exact. In contrast, in the case of stock rising, the classification result is not correct. The total accuracy of this experimental is 77.78%. Similar to the LPB stock, the classification performance in case of NSC stock is pre-

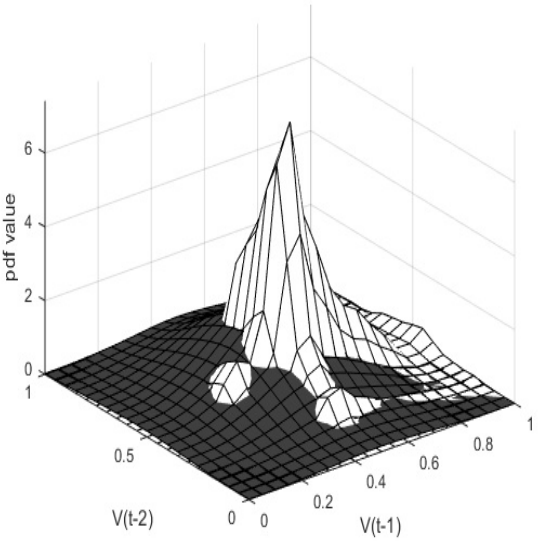

Fig. 5. The probability distribution function of data in two classes

sented in Table 2. According to Table 2, in the case of stock falling, the proposed framework accuracy is 75%, and in case of rising stock prices, the proposed framework accuracy is 100%. The total accuracy of this experimental is 88.89%.

For more detail analysis, it can be observed in Table 1 that the Bayesian algorithm has a high total accuracy, however, the model has no skill at all. In particular, if we said "the stock will fall" every time we predict, we would be right just as often as the sophisticated Bayesian algorithm. For the second stock, if we said "the stock will fall" every time we predict, we would be right only 44.44%, which is lower than that of Bayesian algorithm. Therefore, the proposed algorithm has significant skill here. These are natural comparisons because they emphasize the advantage of Bayesian algorithm compared to what we do in the absence of the algorithm. For more investigation, we perform another experiment on 30 other stocks. Similar to the above experiment, 30 stocks of Vietnam Stock Market are randomly collected from May 2, 2018 to August 10, 2018 and the stock prices from July 30, 2018 to August 10, 2018 are used as the test set. The total accuracy of the proposed technique compared to three other no-skill algorithms consisting of NS1-"the stock will fall" every time we predict, NS2-"the stock will rise" every time we predict, and NS3-a random classification. The comparative result is shown in Table 3.

As shown in Table 3, the proposed technique outperforms NS2 and NS3 and is slightly better than NS1 due to the fact that most stocks in Vietnam stock market have dropped in the test period. This result demonstrates the advantage of the proposed technique compared to what we do in the absence of the algorithm.

Table 2. The classification performance (%) in the case of NSC stock

	True: 0	**True: 1**
Predicted as: 0	33.33	00.00
Predicted as: 1	11.11	55.56
The total accuracy	88.89	

Table 3. The classification performance (%) on 30 stocks

	The proposed method	**NS1**	**NS2**	**NS3**
Total accuracy	62.96	58.14	41.85	50.74

4.2 Probability Threshold Adjustment

In the above experiments, the classification result is calculated with the probability threshold of 0.5, that is, if $P(1|X) > 0.5$ the stock trend is classified to the class "1". In this section, we will discuss a method to adjust the probability threshold so that it can be more suitable for stock investment problem using Receiver Operating Characteristic (ROC) curve. In short-term investment problem, the investors have to make buy and sell orders based on a basic principle? buy at the low and sell at the high? to obtain the highest expected return. We specifically consider the following two scenarios.

Scenario 1: Finding an entry point of investment

Normally, the investors decide to buy the stock after the stock has gone through a period of falling price and can reverse in the future. Specifically, if we believe that the stock price, which closed at time point t, will rise at the time point $t+1$, then t is determined as a suitable entry point of investment. In contrast, t is not suitable time to buy the stock. There are two types of errors that can occur.

Type 1 error: The predicted trend is "rise", but the actual trend is "fall", as shown in Figure 6. This type of error causes serious loss when the investors buy the stock when it is falling continuously.

The Type 2 error: The predicted trend is "fall", but the actual trend is "rise", as shown in Figure 7. This type of error yields loss of investment opportunities, but cannot cause serious loss. Compared to the Type 2 error, the Type 1 error causes a significant risk and needs to be properly controlled.

Scenario 2: Finding an exit point of investment

Normally, the investors decide to sell the stock after the stock has gone through a period of rising price and can reverse in the future. Specifically, if we believe that the stock price, which closed at time point t, will fall at the time point $t+1$, then t is the suitable exit point of investment. In contrast, t is the not suitable time to sell the stock. There are two types of errors that can occur.

Type 1 error: The predicted trend is "rise", but the actual trend is "fall", as shown in Figure 8. This type of error

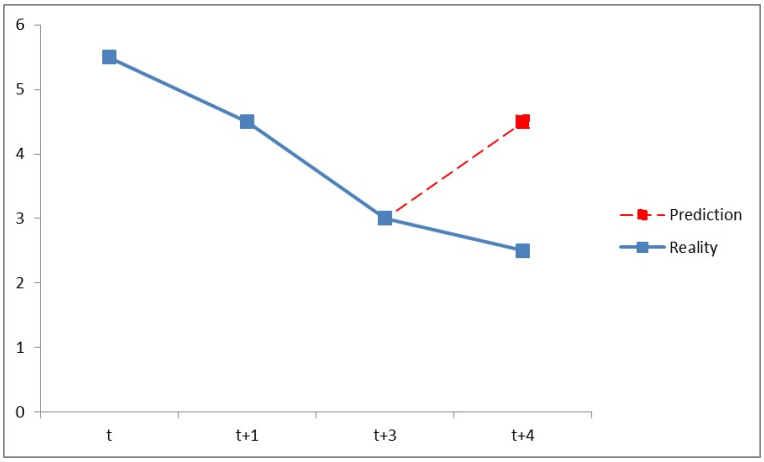

Fig. 6. Type 1 error in Scenario 1

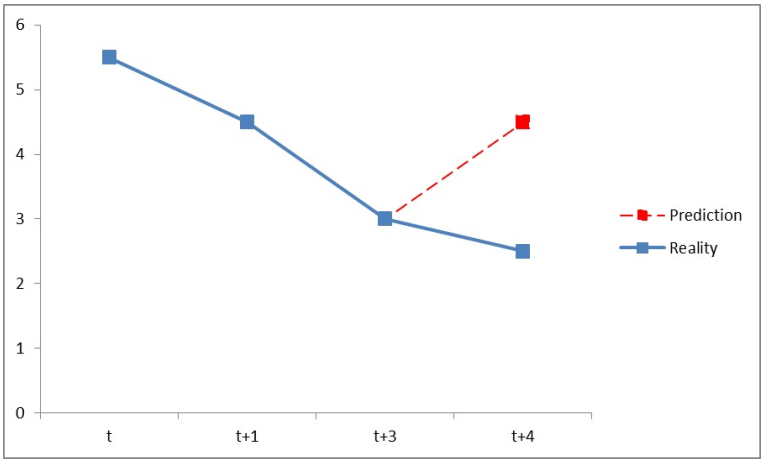

Fig. 7. Type 2 error in Scenario 1

causes serious loss when the investors still hold the stock when it has fallen.

The type 2 error: The predicted trend is "fall", but the actual trend is "rise", as shown in Figure 9. This type of error makes the investors sell the stock when the stock is still rising, and receive an early profit. Similar to Scenario 1, compared to the Type 2 error, the Type 1 error causes a significant risk and needs to be properly controlled.

In summary, in the above two scenarios, the Type 1 error which can measure by the false positive rate can cause significant risk and needs to be properly controlled. Therefore, our purpose is to reduce the false positive rate but still keep the true positive rate at a permissive value. This purpose can be easily solved by finding out a suitable probability threshold based on the ROC curve. Figure 10 and Table 4 illustrate a ROC curve, the probability thresholds, and the corresponding false positive rates and true positive rates.

It can be seen from Table 4 that the

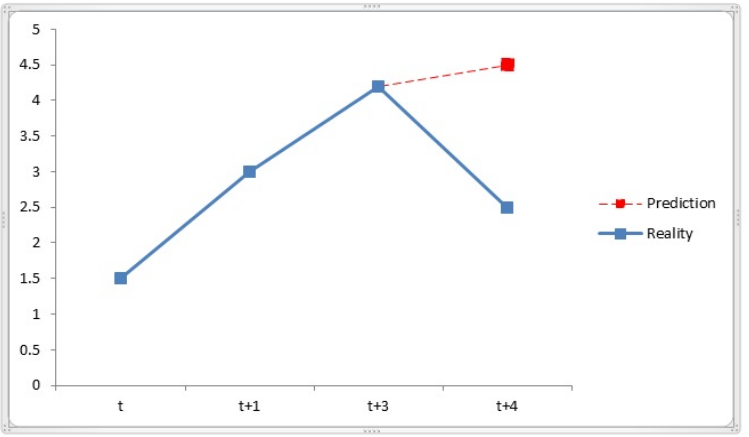

Fig. 8. Type 1 error in Scenario 2

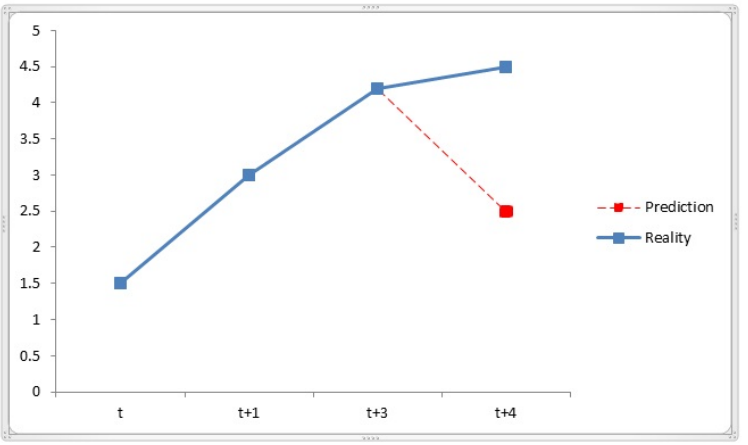

Fig. 9. Type 2 error in Scenario 2

Table 4. Some probability thresholds, and the corresponding false positive rates and true positive rates

Probability Threshold	TPR	FPR
0.8011	0.5000	0.1429
0.7571	1.0000	0.4286
0.5000	1.0000	1.0000

default probability threshold of 0.5 used in the previous experiments results in a true positive rate of 1; however, it also results in a false positive rate of 1, which is too high, and might cause significant risk, as mentioned earlier. In that case, the probability threshold of 0.8 results in a true positive rate of 0.5, which is temporarily accepted, and results in a false positive rate of 0.14, which minimize the risk, can be recommended.

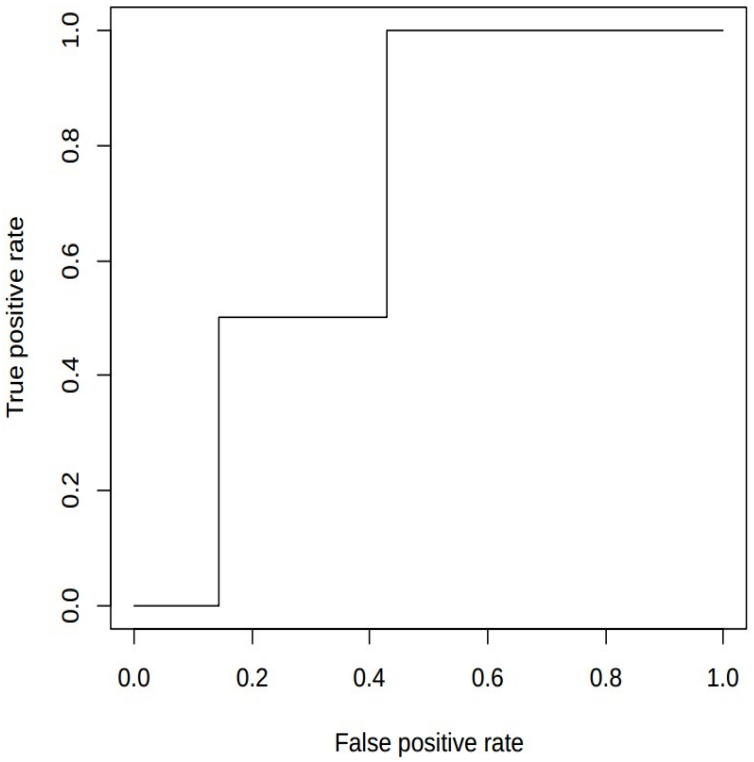

Fig. 10. The ROC curve for stock predict model

5 CONCLUSION

This paper proposes a new framework for stock prediction. In particular, time series data are transformed to tabular data and then predicted using Bayesian classifier. By testing different stocks in Vietnam, the numerical examples indicate that the proposed framework results in reasonable performance and can be considered as a potential method for short-term stock trend prediction. In addition, to reduce the risk for investor, the method to adjust the probability threshold using the ROC curve is investigated. Finally, the proposed framework has been proved to be a potential approach, which can be referred among various technical analysis techniques, and finds its use in a number of specific cases. Also, it can be implied that the performance of the new framework mainly depends on the skill of investors, such as adjusting the threshold, identifying the suitable stock and the suitable time for trading, etc.

References

[1] Addesso P, Capodici F, D'Urso G, Longo M, Maltese A, Montone R, Restaino R, Vivone G (2013). Enhancing TIR image resolution via bayesian smoothing for IRRISAT irrigation management project. In: Remote Sensing for Agriculture, Ecosystems, and Hydrology XV. p 888710

[2] Amin GR, Hajjami M (2016). Application of Optimistic and Pessimistic OWA and DEA Methods in Stock Selection. Int J Intell Syst 31:1220–1233. doi: 10.1002/int.21824

[3] Cartea A, Jaimungal S, Ricci J (2014). Buy low, sell high: A high frequency trading perspective. SIAM J Financ Math 5:415–444

[4] Castellaro M, Rizzo G, Tonietto M, Veronese M, Turkheimer FE, Chappell MA, Bertoldo A (2017). A Variational Bayesian inference method for parametric imaging of PET data. Neuroimage 150:136–149

[5] Chen C, Dongxing W, Chunyan H, Xiaojie Y (2014). Exploiting Social Media for Stock Market Prediction with Factorization Machine. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). pp 142–149

[6] Chen H (2008). Stock selection using data envelopment analysis. Ind Manag Data Syst 108:1255–1268. doi: 10.1108/02635570810914928

[7] Elliott N (2007). Ichimoku Charts: An Introduction to Ichimoku Kinko Clouds. Harriman House Limited

[8] Goldman Mb, Sosin Hb, Gatto Mann (2018). Path Dependent Options: "Buy at the Low, Sell at the High." J Finance 34:1111–1127. doi: 10.1111/j.1540-6261.1979.tb00059.x

[9] Gupta S, Wang LP (2010). Stock forecasting with feedforward neural networks and gradual data sub-sampling. Aust J Intell Inf Process Syst 11:14–17

[10] Hajjami M, Amin GR (2018). Modelling stock selection using ordered weighted averaging operator. Int J Intell Syst 0. doi: 10.1002/int.22029

[11] Huang C-Y, Chiou C-C, Wu T-H, Yang S-C (2015). An integrated DEA-MODM methodology for portfolio optimization. Oper Res 15:115–136. doi: 10.1007/s12351-014-0164-7

[12] Huarng K, Yu H-K (2005). A Type 2 fuzzy time series model for stock index forecasting. Phys A Stat Mech its Appl 353:445–462.

[13] Kale A, Khanvilkar O, Jivani H, Kumkar P, Madan I, Sarode T (2018). Forecasting Indian Stock Market Using Artificial Neural Networks. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). pp 1–5

[14] Kohli PPS, Zargar S, Arora S, Gupta P (2019). Stock Prediction Using Machine Learning Algorithms BT - Applications of Artificial Intelligence Techniques in Engineering. In: Malik H, Srivastava S, Sood YR, Ahmad A (eds). Springer Singapore, Singapore, pp 405–414

[15] Markowitz H (1952). Portfolio selection. J Finance 7:77–91

[16] Mladjenovic P (2016) Stock investing for dummies. John Wiley & Sons.

[17] Nguyen-Trang T, Vo-Van T (2017). A new approach for determining the prior probabilities in the classification problem by Bayesian method. Adv Data Anal Classif 11:629–643

[18] Parmar I, Agarwal N, Saxena S, Arora R, Gupta S, Dhiman H, Chouhan L (2018). Stock Market Prediction Using Machine Learning. In: 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). pp 574–576

[19] Patari E, Karell V, Luukka P, Yeomans JS (2018). Comparison of the multicriteria decision-making methods for equity portfolio selection: The U.S. evidence. Eur J Oper Res 265:655–672.

[20] Patel M (2010). Trading with Ichimoku clouds: the essential guide to Ichimoku Kinko Hyo technical analysis. John Wiley & Sons.

[21] Pham-Gia T, Turkkan N, Vovan T (2008). Statistical discrimination analysis using the maximum function. Commun Stat Comput 37:320–336

[22] Pizzo A, Teyssere P, Vu-Hoang L (2018). Boosted Gaussian Bayes Classifier and its application in bank credit scoring. J Adv Eng Comput 2:131–138

[23] Quah T-S (2008). DJIA stock selection assisted by neural network. Expert Syst Appl 35:50–58.

[24] Roscoe P, Howorth C (2009). Identification through technical analysis: A study of charting and UK non-professional investors. Accounting, Organ Soc 34:206–221.

[25] Usmani M, Adil SH, Raza K, Ali SSA (2016). Stock market prediction using machine learning techniques. In: 2016 3rd International Conference on computer and Information Sciences (ICCOINS). IEEE, pp 322–327

[26] Vovan T (2017). Classifying by Bayesian Method and Some Applications. In: Bayesian Inference. InTech, pp 39–61

[27] Zervos M, Johnson TC, Alazemi F (2012). Buy-low and sell-high investment strategies. Math Financ 23:560–578. doi: 10.1111/j.1467-9965.2011.00508.x

[28] Zhai J, Bai M (2018). Mean-risk model for uncertain portfolio selection with background risk. J Comput Appl Math 330:59-69.

[29] Zhang GP (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175

[30] Zhou Z, Jin Q, Xiao H, Wu Q, Liu W (2018). Estimation of cardinality constrained portfolio efficiency via segmented DEA. Omega 76:28–37.