Stock Market Movement Prediction using LDA-Online Learning Model

Tanapon Tantisripreecha
Department of Mathematics
Faculty of Science, Mahidol University
Bangkok, Thailand
tanapon.tan@mahidol.ac.th

Nuanwan Soonthornphisaj*
Department of Computer Science
Faculty of Science, Kasetsart University
Bangkok, Thailand
fscinws@ku.ac.th

***Abstract*—In this paper, an online learning method namely LDA-Online algorithm is proposed to predict the stock movement. The feature set which are the opening price, the closing price, the highest price and the lowest price are applied to fit the Linear Discriminant Analysis (LDA). Experiments on the four well known NASDAQ stocks (APPLE, FACBOOK GOOGLE, and AMAZON) show that our model provide the best performance in stock prediction. We compare LDA-online to ANN, KNN and Decision Tree in both Batch and Online learning scheme. We found that LDA-Online provided the best performance. The highest performances measured on GOOGLE, AMAZON, APPLE FACEBOOK stocks are 97.81%, 97.64%, 95.58% and 95.18% respectively.**

Keywords—LDA; Online learning; Stock movement prediction

I. INTRODUCTION

Stock market movement prediction is a critical task for investors to determine the price direction of the stock before trading. There are two major approaches to estimate the stock prices which are Fundamental analysis and Technical analysis. Fundamental analysis aims to estimate an intrinsic value of the stock price using the performance of the company, the company policy, etc. This background information is required for ideal investors; however, it requires financial and business accounting knowledge that the majority of investors are incapable. Technical analysis is used to estimate the trend or pattern of stock price using historical data. Stock prices are plotted on the stock chart so that the investors can see the movement of the price on the real-time chart and make the decision immediately. Since the traditional stock market data is operated on computer system, many automatic trading systems are developed.

There are two learning schemes using for the stock prediction which are batch learning and online learning. The batch learning is a learning process that a training set and a test set are divided separately. A model obtained from the training set is applied to predict the value of the test set or unseen data. Note that, the value of the unseen data is not included to update the model. On the other hand, the online learning is a method dealing with sequential data as a time series. The model keeps updating its parameters in incrementally style.

In this paper, we propose the LDA-online learning algorithm to predict the stock market movement. We investigate the performance of the proposed algorithm using four IT company stocks traded in NASDAQ stock market which are Apple, Facebook, Google and Amazon.

II. RELATED WORK

Neural Networks, Support Vector Machines, Random Forest and Naïve Bayes were studied by Jigar Patel [1]. Two stock indices namely CNS Nifty and S&P BSE Sensex and two stock prices (Infosys and Reliance Technology) were used as data set. Ten indicators were represented as continuous values. The feature set were 1) 10-day Moving average, 2) Weighted 10-days Simple moving average, 3) Exponential moving average, 4) momentum, 5) Relative Strength Index, 6) K% of stochastic, 7) %D of stochastic, 8) MACD, 9) Larry William's R%, and 10) A/D oscillator. The result showed that Random Forest outperformed other three algorithms. The best performance (accuracy) obtained in their study was 83.59%.

Ensemble techniques (Random Forest, Adaboost and Kernel Factory) were studied by Michel Ballings *et al* [2]. They aimed to predict long term stock price direction (UP/DOWN). They compared Random Forest, Adaboost and Kernel factory with ANN, SVM, logistic regression and K-NN. They collected data from 5,767 publicly listed European companies and used the area under the receiver operating characteristic curve (AUC) as a performance measure. They concluded that Random Forest was the best algorithm followed by Support Vector Machines, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbors and Logistic Regression respectively.

The ability of Artificial Neural Networks in forecasting daily NASDAQ stock index was explored by [3]. They used short-term historical stock value as well as the day of week as inputs. They did experiment using different number of hidden layers. The model applied the first 70 days as the training data and the last 29 days as the test set. To test the model, the performance was compared by using stock index data obtained from four and nine prior working days. The result showed that there was no significant difference between the prediction's ability of the four and nine prior working days.

In 2017, Bin Weng *et al.* [4] proposed the investment tool which integrated news and daily stock movement as the input features. News was collected from Google's news and Wikipedia. Moreover, three classification techniques which are Decision Tree, SVM and ANN were applied in the intelligent trading expert system. They presented a case study based on the

*Corresponding author

978-1-5386-5889-5/18/$31.00 ©2018 IEEE
SNPD 2018, June 27-29, 2018, Busan, Korea

AAPL stock (Apple company traded in NASDAQ) and found that the expert systems got 85% accuracy on predicting the stock movement.

III. Classification Techniques

A. LDA-Online learning method

In this work, we propose the LDA-online learning method that attempt to transform the learning algorithm of batch learning to online learning. The learning framework is shown in Fig. 1 in which T is a training set, N is an unseen data and D is a sequence of days. The concept of LDA-online learning is to reconsider the training set by accumulating the new available data to fit the next LDA model. Note that the proposed method is different from the traditional online learning in stock prediction domain that considers only the fix number of days prior to the forecasting date (see Fig. 2). Since the prediction model is based on the indicators that need to be calculated on the specific day.

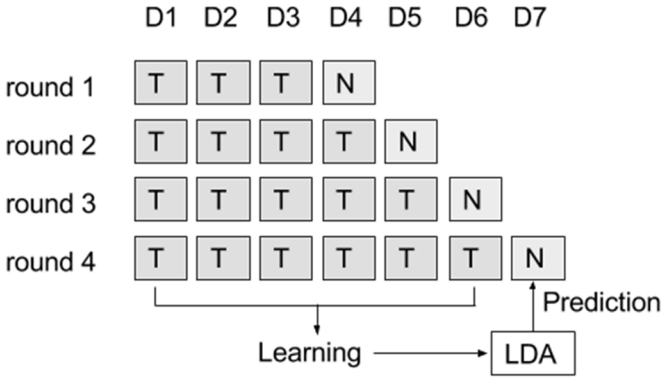

Fig. 1. LDA-Online Learning model

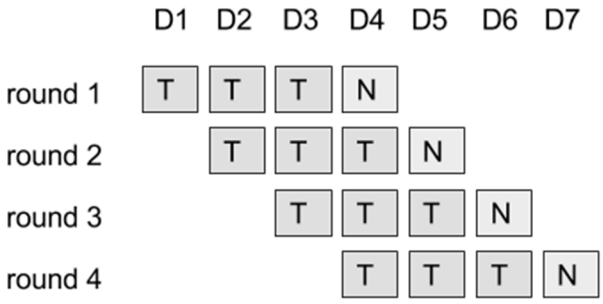

Fig. 2. Traditional online learning for stock prediction

The stock market movement prediction is determined in terms of UP or DOWN. In this work, if the current closing price is higher than the previous price, the algorithm determines the result as UP, vice versa. The performance of LDA-online learning is compared to ANN, KNN and Decision Tree.

B. Linear Discriminant Analysis (LDA)

The main purpose of LDA is to predict the value based on a linear combination of the interval variables. Fisher linear discriminant is the method used to find a linear combination of features [5].

LDA is closely related to the statistical technique namely ANOVA and the regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements [1][2]. The key concept of ANOVA is the use of categorical variables which assume to be independent and a continuous dependent variable, whereas LDA is applied to continuous independent variables and a categorical dependent variable (i.e. the class label) [3]. Logistic regression and logistic regression are more similar to LDA than ANOVA, since they explain the categorical variables by the values of continuous independent variables. These methods are practical for applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

C. Batch Learning model

Batch learning algorithm learns from the training set to create the model and permanently applied its model to predict the new unseen data. Batch learning are not practical for the time series data domain since the concept model is static whereas the value of unseen data depending on a function of time. We found that the value of the current data influences the next value.

Consider the Stock prediction domain, we found that the batch learning typically calculates the historical stock price using Simple Moving Average (SMA), Exponential Moving Average (EMA) and Relative Strange Index (RSI) (see equation 1-3).

$$SMA_n = \frac{1}{n}\sum_{i=1}^{n} price_i \quad (1)$$

$$EMA_n = price_n\left(\frac{2}{Time+1}\right) + EMA_{n-1}\left(1 - \frac{2}{T+1}\right) \quad (2)$$

Where Time is the time period.

$$\mathrm{RSI} = 100 - \left(\frac{100}{1+RS}\right) \quad (3)$$

$$\mathrm{RS} = \frac{\mathrm{AvgU}}{\mathrm{AvgD}} \quad (4)$$

Note that AvgU is the average of all UP moves in the last *n* price bars, AvgD is the average of all DOWN moves in the last *n* price bars, and *n* is the period of RSI.

SMA, EMA and RSI are well known indicators that are useful for investors. SMA and EMA are average price movement that the investors normally use as the basic estimation tool for considering the trend, and RSI is the Relative Strange

Index that the investors apply to detect the strange of the stock's trend.

Popular batch learning algorithms used in Stock prediction domain are K-nearest neighbor, Decision Tree and Artificial Neural Networks. K-nearest neighbor method is one of the simplest machine learning algorithms used for classifying objects based on the closest training examples in the feature space. An object is classified by a majority class among its k-nearest neighbors. The k-nearest neighbor approach does not rely on prior probabilities like LDA [6].

Decision Tree is a supervised learning method used for classification and regression [7]. Decision Tree builds classification or regression models in the form of a tree structure. The algorithm can handle both numerical and categorical data. The main mechanism is to continuously select the best feature as a node in the tree that reduces the entropy of the predefined class in the training set.

Artificial Neural Networks (ANN) is typically organized in layers. Layers are made up of a number of interconnected nodes which contain an activation function. Training instances are fed to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of connection weights. The hidden layers finally link to an output layer. ANN can be applied for classification or prediction problem.

IV. Feature Set

We collect the NASDAQ daily stock market from Yahoo Finance (https://finance.yahoo.com/). Table I shows the data set of four popular IT company's stocks which have considerably high-volume trading. These stocks include Apple (AAPL, 1999-2017), Facebook (FB, 2012-2017), Google (GOOGL, 2004-2017) and Amazon (AMZN, 1999-2017). The stock information includes date, open price, high price, low price and closing price.

The target classification of LDA-Online aims to predict the stock direction. To consider stock movement, the direction (UP or DOWN) will be determined by comparing the current prediction price with the preceding stock price. If the predicting price is higher than the preceding stock price, the movement is UP. In case that the current stock price is lower than the preceding stock price, then the movement is determined as DOWN. We evaluate the performance of all algorithms by considering the stock direction and measures in term of accuracy.

TABLE I. Dataset

Stock Symbol	Number of days
AAPL	4,256
FB	1,161
GOOGL	3,119
AMZN	4,259

V. Result and Discussion

All experiments are performed using machine learning tool, Scikit-learn, written in Python. In batch learning, the performance of four stocks are compared using four algorithms. The results are shown in Figure 3-6 where axis-x is the number of batches and axis-y is the performance measured in term of accuracy.

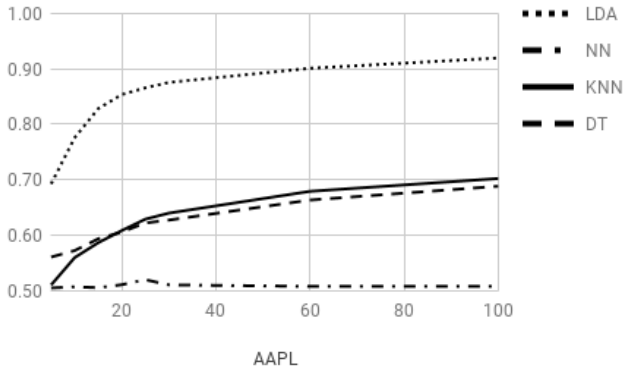

Fig. 3. Accuracy of AAPL in batch learning

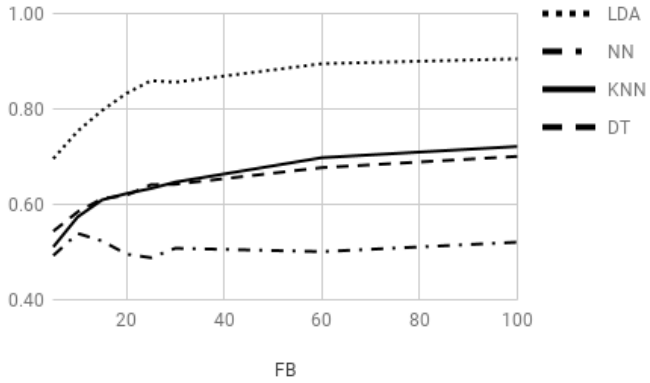

Fig. 4. Accuracy of FB in batch learning

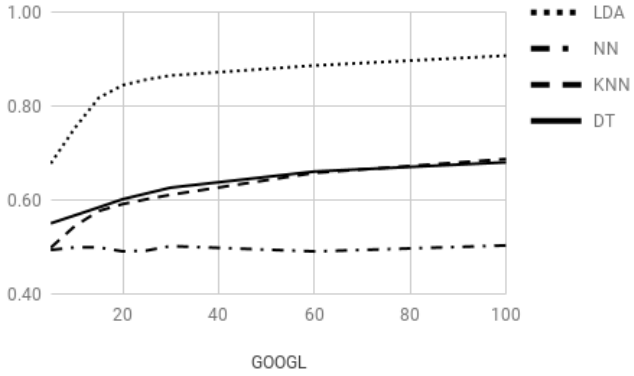

Fig. 5. Accuracy of GOOGL in batch learning

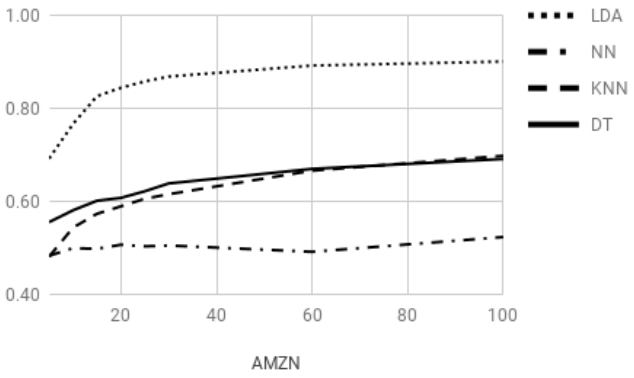

Fig. 6. Accuracy of AMZN in batch learning

Considering the batch learning mode from Figure 3-6, we found that LDA get the highest accuracy in every learning algorithms. Moreover, the accuracy is higher when the number of batches is increased. Therefore the completeness of the historical data plays an important role in stock prediction.

The performance of LDA-Online is compared to ANN, KNN and Decision Tree using batch learning and online learning mode. In table II, the experimental results showed that LDA-Online outperforms other algorithms in both batch learning and online learning.

TABLE II. ACCURACY OF ONLINE AND BATCH LEARNING

Algorithm	AAPL	FB	GOOGL	AMZN	Average
LDA-batch (%)	92.85	91.58	93.22	93.16	
LDA-online (%)	95.58	95.18	97.81	97.64	
% difference	**2.95**	**2.00**	**4.92**	**4.81**	**3.67**
ANN-batch (%)	62.29	54.30	56.91	70.36	
ANN-online (%)	58.68	53.86	50.97	52.74	
% difference	**-5.80**	**-0.81**	**-10.44**	**-25.05**	**-10.52**
KNN-batch (%)	76.91	73.68	76.95	77.96	
KNN-online (%)	83.90	80.26	84.02	85.42	
% difference	**9.09**	**8.93**	**9.19**	**9.56**	**9.19**
DT-batch (%)	76.91	73.68	76.95	77.96	
DT-online (%)	78.11	75.53	78.76	80.13	
% difference	**1.57**	**2.50**	**2.35**	**2.78**	**2.30**

Considering the online learning mode, we found that LDA-Online reaches the highest performance which is quite impressive performance for GOOGL (97.81%) and AMZN (97.64%) stock. The prediction performance of AAPL stock measured in term of accuracy is 95.58%, whereas that of FB stock reaches 95.18%.

For batch learning, we found that LDA-Batch performs better than other batch learning methods as well. The overall experimental result shows that the Online-learning scheme was preferable for all learning algorithm.

Considering other classification techniques, in figure 7, we found that KNN-online and DT-online provide higher accuracy compared to KNN-batch and DT-batch. However, ANN-online gets lower accuracy than ANN-batch learning.

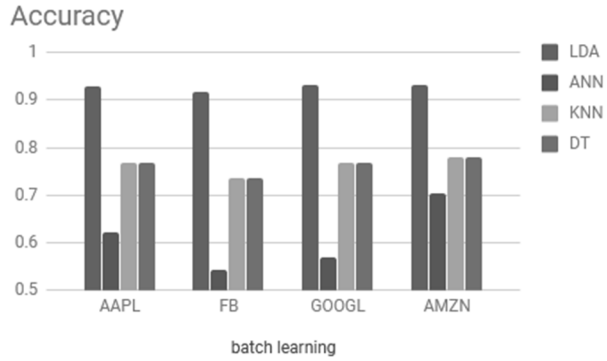

Fig. 7. Performance of Batch learning

We investigate our online learning scheme using different algorithms. As shown in the bar chart of Fig. 8, we found that LDA-Online outperforms other algorithms in every dataset.

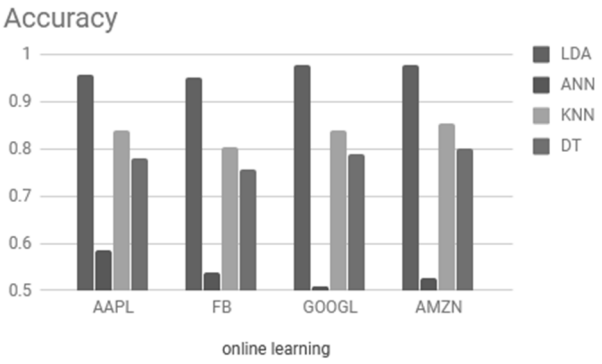

Fig. 8. Performance of Online learning

VI. CONCLUSION AND FUTURE WORK

This paper proposes the LDA-Online learning method that incrementally applies the historical stock as the training set for LDA to fit the learning model. Our method differs from traditional online learning since LDA-Online does not fix the number of prior dates in the learning phase. The advantage of LDA-Online is the completeness of the historical pattern. We found that the proposed online scheme provides promising result for KNN, ANN and Decision Tree. It showed that the completeness of historical data played an important role in stock market movement prediction.

We conclude that LDA is a method of choice since the computational time is not a problem. The total running time is shorter compared to ANN and Decision Tree.

We plan to study on feature extraction to develop new feature set for stock price prediction and apply various indicators to empirically investigate their potentials on LDA-Online in the near future.

ACKNOWLEDGEMENT

We would like to thank Department of Computer Science, Faculty of Science, Kasetsart University for providing the partial support.

REFERENCES

[1] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques", Expert Systems with Applications, vol. 42, p. 259-268, 2015.

[2] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction", Expert Systems with Applications, vol. 42, no.20, p. 7046-7056, 2015.

[3] M. Amin Hedayati, M. Moein Hedayati, E.Morteza, "Stock market index prediction using artificial neural network", Journal of Economics, Finance and Administrative Science, p. 89-93, 2016.

[4] W. Bin, A. Mohamed, M. Fadel,"Stock market one day ahead movement prediction using disparate data sources", Expert system with application, p. 153-163, 2017.

[5] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques", Expert Systems with Applications, vol. 42, no 1, p. 259-268, 2015.

[6] O. Phichhang, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques", Modern Applied Science, vol. 3, no. 12, p28-42, 2009.

[7] J. R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, vol. 4, p. 77-90, 1996.