A Master of Science

Importance of Trading Volume-Based Features and Modelling Parameters in Daily Stock Trading Using Neural Networks

School of Electrical Engineering

University of Ulsan

DINH THUY AN

November 2017

Importance of Trading Volume-Based Features and Modelling Parameters in Daily Stock Trading Using Neural Networks

Under the supervision of

Prof. Kwon Yung-Keun

Submitted to

School of Electrical Engineering

University of Ulsan

In partial fulfillment of the requirements for the degree of

Master of Science

울산대학교
UNIVERSITY OF ULSAN

By

DINH THUY AN

November 2017

The Master Thesis presented by

Dinh Thuy An

Approved by an academic committee:

1. Chairman: **Prof. Jun, Heesung**
   University of Ulsan

2. Member: **Dr. Seo, Dongil**
   Google Korea

3. Member: **Prof. Yung-Keun Kwon**, *Advisor*
   University of Ulsan

November 2017

School of Electrical Engineering

University of Ulsan

DINH THUY AN

[ABSTRACT]

Importance of Trading Volume-Based Features and Modelling Parameters in Daily Stock Trading Using Neural Networks

This paper analyses the role of combination of Close Price and Volume Trading value in terms of retrieved profit after short-term stock trading by using machine learning method. Thus, the Volume trading value and Close Price is selected as input of 2 methods Support Vector Machines (SVM)-based predictor and Artificial Neural Network (ANN) because both 2 methods prove itself excellent performance among machine learning methods. To decide the moment trading on simulation, strategy rule based on parametric model and close price is also produced. SVM conducts training and testing procedures to find trading signal. Profitability of these simulations is evaluated against a traditional Buy-and-hold strategy and compared with total average profit among the whole dataset. We tested the proposed model with daily trading data from 2001 to 2015. The empirical result showed satisfactory performance that the accuracy and ratio of trading profit of proposed model trading volume can be a good factor on forecasting profit.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my advisor, Prof. Kwon Yung-Keun. I am appreciated Prof. Kwon for instruction and supported in my studying path. Prof. Kwon inspired me a lot through researching, survey and finding new idea during my master course in Ulsan University. What I learned from 2 years studying such as logical thinking, setting up hypothesis and finding prove is worth for my mindset. His guidance helped me in all the time of research and writing of this thesis. I would be very happy to be his student and continue my PhD course under his supervision. Once again, I would like to be thankful and grateful for all his help and correction for my research.

I would like to say thanks to my friends and lab mate's senior. I thoroughly feel happy since I received many help from them without any hesitation. Because of their sharing, my life became much more easy and familiar with new living and working environment. I would like to say thank to my lab mates for everything talk and sharing in the research path.

I would like to send my exceptional thank you message to my family. I am so lucky to have such a great family in which I always get many love and encouragement. I am always aware of that I have been nurtured and inspired enthusiastic studying over years by my parents. This is big motivation for me to complete this master thesis and try harder for further study. I would like to thank my siblings who are always help me to overcome any troubles I met in my living and studying.

DINH THUY AN

Ulsan, Republic of Korea

October 2017

VITA

Dinh Thuy An was born in Dong Nai province on October 27, 1993 and has been living in Ho Chi Minh, Vietnam. She received Bachelor degree in Computer Science (2015) from University of Information Technology (Vietnam National University, Ho Chi Minh City). In March 2016, she began working full time towards her Master course at University of Ulsan, Korea under the guidance of Prof. Kwon Yung-Keun. Since then, she started to conduct researches in Complex Systems Computing lab, and focused on trading stock with machine learning methods.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1 Motivation

Nowadays developing a stock trading system incorporating artificial intelligence methods is getting more attention of investors. In particular, there were many studies to forecast stock price's trend and therefore determine optimal buying and selling points. However, it is not easy since stock markets are nonlinear and volatile (Abhyankar, Copeland, & Wong, 1997). Recently, researchers have proposed various models and techniques for stock price prediction, which are classified into fundamental analysis and technical analysis. The former analysis usually uses the intrinsic information of a company such as earnings, capital, sakes, share and relationship with other companies to assess the value of the company. On the other hand, the technical analysis uses a historical dataset of the stock prices, the trading volume and many other technical indicators to estimate the stock price in future (Pring, 2015). In this study, we also focus on the second type of analysis.

Many technical variables have been devised using moving average, exponential smoothing and linear regression statistical methods (Kotler, 1971). For example, the weighted average of the past price values was modeled to predict the short-term fluctuation of a price by assuming that the data follow some historical pattern. And the more patterns observed and analyzed, the more relevance and evidence used to predict price for the future (Granger, 1980; Wilson and Keating, 1990). Moreover, machine learning methods are more widely employed in a stock prediction problem. For example, artificial neural networks (ANNs) was applied to predict daily New York Stock Exchange (Gallant, Rossi, & Tauchen, 1992), daily Deutscher Aktienindex stock (DAX), intra-day option contracts on the Financial Times Stock Exchange 100 Index (FTSE100) (Tino, Schittenkopf, & Dorffner, 2001), and S&P 500 future index price (Hamid & Iqbal, 2004). Support vector machines (SVMs) were also used to predict five future contracts collated from the Chicago Mercantile Market (Cao & Tay, 2003). Although these previous studied were successfully applied, we note that there are two issues needed to be discussed: the type of input variables and the modelling parameters for learning. Regarding the first issue, it is significant to construct most informative input variables for a forecasting model. In fact, most input variables in previous stock

prediction methods have been generated from a closing, opening, highest, or lowest price. For example, moving average (Kwon & Moon, 2003), relative strength index (Kwon & Moon, 2007), golden cross and dead cross (Lin, Yang, & Song, 2011), stochastics (Kim, 2003) and many other technical indicators have been devised from a stock price. On the other hand, we note that the trading volume is less emphasized in creating informative technical indicators. Considering that the volume is known to be informative in explaining a status of a stock market (Campbell et al., 1993; (Chen, Hong, & Stein, 2001)), there is room to more intensively use the trading volume as input variables in stock price prediction. The second issue is about the modelling parameters in learning stage. Specially, we focus on three parameters, the number of past days to be considered for input variables, the maximum number of future days for a long position, and the minimum profit rate to sell a stock. These parameters might be critical for accurate prediction because they modify the problem space by modifying input and output variables. We note that they were empirically chosen by trial-and-error in previous studies, though. For example, the number of past days was set to five in KOSPI 200 prediction (Roh, 2007) or even one in National Stock Exchange (NSE) India prediction (Bhat & Kamath, 2013). Values can improve the prediction performance by making the learning problem easier-to-learn.

1.2 Research objectives

In this study, we apply two primary algorithms Multilayer Perceptron (MLP) and SVMs method which is considered as efficient in stock prediction in (Tino et al., 2001) and (Gallant et al., 1992). We define parameters to present the window-size which is considered properly to help the learning model forecast the next day signal trading. The window-size is selected properly through learning process among a list of candidate parameters.

This thesis seeks to find answers to several questions. One is whether trading volume can be important and fundamental factor in terms of forecasting stock. Because most studies pay much attention on technical indicator generated from only closing price rather than the dynamic trading volume daily. Although many previous works proved the efficient and powerful trading indicators to predict the future price as well as decide the trading time, this study investigate the trading volume to make a balance view among the forecasting stock trends technique.

Another concern is how to implement the modelled parameters which modify both closing price and trading volume into MLPs and SVMs learning algorithm. Also, it is also necessary to choose a suitable forecasting horizon since it should be long sufficiently to avoid the missing information (Cao & Tay, 2003). As proposed by (Thomason, 1999), distribution time series data would make the data become more symmetrical advantages. The optimal parameter will maximize the predict accuracy and give the highest profit rather than fix the interval series in advance. This study also examines various forecasting models based on dynamic level of profit which is defined by threshold profit variable. The purpose of proposing these parameters is: (1) from the back-testing stock data perspective, it helps to demonstrate and analyze the impact of period day on the predictability of stock index direction from the back-testing stock data perspective; and (2) from the trading view to develop effective trading strategies which would generate relative performance of these investment schemes.

The last question is about the performance measure. Most studies evaluated the performances based on some statistical measurement such as mean squared/absolute error; false positive/ false positive, and so on (Kwon, 2007). However, this thesis used another measurement, which is profit measurement to evaluate the efficient of trading strategy. In theory, it is profitable to compute the outcome without any transaction cost considered. But in fact, there are some literature supported that it is hard to make profit with stock trading system when transaction cost is considered. For example, Alexander tested several filter rules to advise the trader the buying and the selling points on Down Jones and Standard & Poor's indices. Finally, the trading result showed that the profit is not significant when transaction cost is taken in to account. In this study, I conduct simulation trading with pre-defined input and target variables, then the optimal parameter is chosen in the learning stage. At the final stage, the ratio trading profit is computed for every single year with every stock collected.

1.3 Thesis outline

The remainders of the thesis are organized as follows:

- Chapter 2 discuss about the background knowledge for my work such as introduction of daily trading stock property and machine learning methods (SVMs /MLPs) used in the previous studied.

- Chapter 3 shows detail my proposed method and overall framework used in this trading method and learning method. Our trading strategy which is built up by my hypothesis is also explained in section 2.

- Chapter 4 presents out simulations including the detail data set which is extracted from sites: https://finance.yahoo.com/ and how data is processed among many records data set in this study. Also, implementation set up such as Scikit-learn with Python language and LibSVM library with Java language for these simulations is also described in this chapter. The performance is evaluated by using accuracy of prediction model and profit after processing trading.

- Chapter 5 discuss on my result which is mentioned above about the accuracy and profit measurement. And future works are also offered.

CHAPTER 2. BACKGROUND

2.1 Methods used stock forecasting

2.1.1. Overview

Stock investment is a high-risk financial activity, in which investors are likely to lose their life savings because of poor investment decisions resulting from failure to consider the factors involved in stock price variation or the accumulation of professional investment knowledge and experience. To enhance investor decision-making quality and profitability, analyses conducted for stock investment generally comprise fundamental and technical analysis.

2.1.2. Fundamental analysis

In the former method, the objective is to evaluates the intrinsic value using stock financial condition, industrial development, and macroeconomic environment, while the latter forecasts future stock price trends based on historical stock price variation (Devadoss, 2013). Fundamental analysis is more useful for long-term investors. For example, (Thawornwong, 2004) proposed a novel method that select adaptive financial and economic variables such as reputation, sales, earnings, financial news for stock prediction.

2.1.3. Technical analysis

The technical analysis is characterized by a large of number of formula and trading indicators (mentioned in Section 2.2.1) which committed to explain the trend of dynamic historical prices. Technical analysis is commonly used for deciding the buying and selling points in the security. Candle stick charts is often used in the technical analysis to illustrate the price movement by applying indicators. Technical analysis approach is suitable for short-term trader.

There are various schools of thought in terms of the ability to profit from the equity markets. Some believe that no investor can obtain above average trading advantages based on the historical and present information. The Random Walk Hypothesis (RWH) states that prices on the stock market wander in a purely random and unpredictable manner. As a result, according to this theory, every price change occurs without any

influence from past prices. The Efficient Market Hypothesis (EMH) states that the markets incorporate all available information and prices are adjusted immediately once new information becomes available. If these theories are true, there should not be any advantage in predicting stock performance, as the market would react and compensate for any actions performed due to the predicted information. These theories have been met with a great deal of opposition. The argument against the EMI is that many investors base their expectations on past prices, past earnings, track record as well as other indicators. Since stock prices are largely influenced by investor expectations, many believe it only makes sense that past prices do affect future prices. Compelling evidence has also been if rejects the RWH. It has been illustrated that stock market price movements, of the United States as well as Japan, have conformed only to the weak form of the EMH. There has also been a study of 234 stocks from 8 major European stock markets, which indicated that these stock markets exhibited a slight departure from the RWH. As a result, the above offers encouragement for research into developing market prediction applications.

2.1.4. Artificial Neural Networks (ANNs) in stock market prediction

Recently, investors studied artificial neural network to overcome the difficulty characteristic of stock data. It is well known that is nonlinear and volatility data. As artificial neural network can reduce the noisy data and figure out the input-output relationship of such nonlinear stock data.

Many forms of machine learning are used in the analysis and forecast of the stock market such as Neural Network, Decision Tree, and SVMs. In this study, we focused on daily stock trading based on multi-layer perceptron (MLP) and SVM, respectively.

An ANN is typically composed of layers of nodes. In the popular MLP, all the input nodes are in one input layer, all the output nodes are distributed into one or more hidden layers in between. An MLP is determined by the following variables (Figure 1)

- The number of input nodes
- The number of hidden layers and hidden nodes
- The number of output nodes

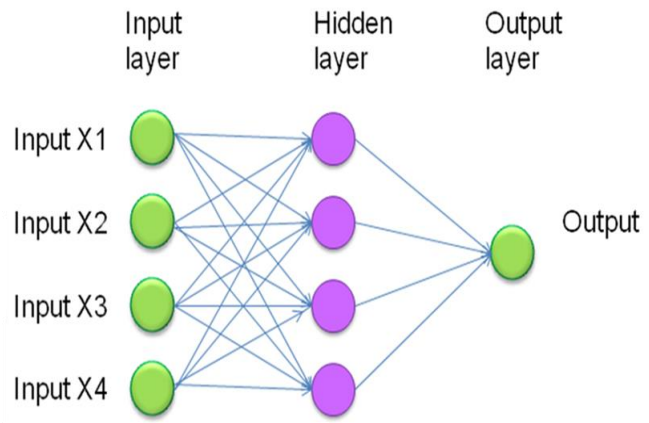

Figure 1 – The structure of Artificial Neural Network

The selection of these parameters is basically problem-dependent. There exist many different approaches such as the pruning algorithm for finding the optimal architecture of an ANN, these methods are usually quite complex in nature and are difficult to implement. Furthermore, none of these methods can guarantee the optimal solution for all real forecasting problems. To date, there is no simple clear-cut method for determination of these parameters. Guidelines are either heuristic or based on simulations derived from limited experiments. Hence the design of an ANN is more of an art than a science.

An artificial neural network is defined as a data processing system consisting of many simple highly interconnected processing artificial neurons in an architecture inspired by the structure of the cerebral cortex of the brain. There are several classes of neural networks. It is classified according to the learning mechanisms. The three broadly classified learning methods are supervised learning, unsupervised learning and reinforced learning. There are three fundamental classes of networks namely, single layer Feedforward network, multilayer Feedforward network and recurrent network. This study used multilayer Feedforward network (a.k.a Multilayer Perceptrons) and SVMs.

SVMs was proposed by Vapnik and his colleagues. SVMs uses linear model to implement nonlinear class boundaries by generating regression function to map the input vectors x into the high-dimensional feature space. Thus, SVM is known as the algorithm solving nonlinear regression estimation problems. Another key characteristic of SVM is that training

SVM is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVM is always unique and globally optimal, unlike other networks' training which requires nonlinear optimization with the danger of getting stuck into local minima. In SVM, the solution to the problem is only dependent on a subset of training data points which are referred to as support vectors. Using only support vectors, the same solution can be obtained as using all the training data points. One disadvantage of SVM is that the training time scales somewhere between quadratic and cubic with respect to the number of training samples. Therefore, a large amount of computation time will be involved when SVM is applied for solving large-size problems.

Feed-forward neural networks are the most commonly used networks for a variety of applications in finance and accounting (Coakley & Brown, 2000). MLP neural network architectures have been used universally in terms of utilizing forecasting, classification and recognizing the upcoming pattern based on givens data.

MLP is a feedforward neural network with one or more layers between input and output layer. Feedforward means that data flows in one direction from input to output layer. MLP has three layers; an input layer, one more hidden layers and output layer. The input data are fed to the neurons in the input layer and after processing within the individual neurons of the input layer the output values are forwarded to neurons in the hidden layer and finally to the neurons in the output layer. MLPs are widely used for pattern classification, recognition, prediction and approximation. Connections among the neurons are associated by weights and changing the weights in a specific manner results to learning of the associated network. The procedure by which the weight changes take place in the network is called learning or training algorithm. The backpropagation algorithm is the most commonly used learning technique. The technique consists of a forward pass and a backward pass. In the forward pass, an input vector is applied to the nodes of the network and result of which becomes a set of outputs for the network at the output layer. During this phase the weights are all fixed. In the backward pass, the error term is calculated by finding the difference between actual response of the network and desired response specified to the network and is propagated backward through the network. Here the weights are adjusted so as to make the actual response of the network becomes closer to desired response. The neural network training is an unconstrained nonlinear minimization problem in which synaptic weights of a network are iteratively modified to minimize the overall mean or total

squared error between the desired and actual output values. The most popularly used backpropagation algorithm is used for training which follows the gradient steepest descent method. For the gradient descent algorithm, a step size, called learning rate must be specified. The learning rate is a constant of proportionality which determines the size of the weight changes. The weight change of a neuron is proportional to the impact of the weight from that neuron on the error. A very small learning rate requires more training time. One method to increase the learning rate and thereby speed up training time without leading to oscillation is to include a momentum term in the backpropagation learning rule. The momentum term determines how past weight changes affect current weight changes. Most neural network software programs provide default values for learning rate and momentum that typically work well. Initial learning rates in the literature are found to vary widely from 0.1 to 0.9. Common practice is to start training with a higher learning rate such as 0.7 and decrease as training proceeds. Many neural network programs will automatically decrease the learning rate and increase momentum values as convergence is reached.

2.1.4. Time series forecasting

Time series forecasting is the analysis of the time series data that tries to predict the near future data based on its past data. This is significant in the field of stock market investment, as investors want to make right decisions at right times to maximize their financial profit. Conventional researches used time series analysis techniques like mixed auto regression moving average (ARMA) and multiple regression models. Time series forecasting usually find a trend in the past data to predict future data. The more past data, the easier it is to find a pattern. However, if the history of a stock is short, an accurate analysis and forecast for such little past data is difficult. So, in this case neural networks are described as great tools to use in this scenario.

2.2. Daily stock trading

2.2.1 Introduction

There can be various stock trading schemes such as intraday, daily, weekly, or monthly trading according to the trading time intervals. In this paper, we considered the daily trading scheme among them because it has been most frequently handled in the previous studies ((Kwon & Moon, 2003); (Mittermayer, 2004) and (Bhat & Kamath, 2013)). A

typical formulation for this problem is to approximate an underlying function f for a target variable $y(t)$ and a set of input variables $X(t)$ at day t as follows:

$$y(t) = f\big(X(t)\big).$$

The daily stock dataset includes daily information which consists of losing price, the highest price, the lowest price, and the trading volume. In this thesis, the closing price and trading volume which are named at day t as $\mathrm{x(t)}$ and $\mathrm{v(t)}$, respectively. In many studies, the strategy trading is problem of price series and input variables is based on price-based value. The problem is a kind of time-series data prediction that can be usually first tried with delay coordinates as follows:

2.2.2 Technical trading from Closing Prices

From the financial perspective, there are various statistics technical indicator generated in the market like closing price, volume trading information by financial expert (Kaufman, 2013). I describe the most frequently used in many forecasting study (Kwon, 2007), (Bhat, 2013) in the following:

- Moving Average (MA)
  - The numerical average value of the stock prices over a period.
  - There are 2 forms of (MA), MA_S and MA_L which is short-term and long-term moving average, respectively.
- Golden-cross and dead-cross
  - States that MA_S crosses MA_L upward and downward, respectively
- Moving average convergence and divergence (MACD)
  - Formula for define MACD: $\mathrm{MACD} = MA_S - MA_L$
  - Meaning of MACD is a momentum indicator that shows the relationship between MA_S and MA_L. This indicator gives the trader a signals of trend changes with cross overs.
- Relative Strength index (RSI)
  - An oscillator that indicates the internal strength of a single stock. RSI indicates the overbought and over sold regions. This oscillates between 0 and 100. Above 70 is marked as oversold region and below 30 is marked to be overbought region. RSI can also be used to see the general trend which show the ratio of average gain and average loss.

- Formula: $\text{RSI} = 100 - \frac{100}{1+RS}$

$$RS = \frac{AverageGain}{AverageLoss}$$

- Stochastics
  - An indicator that compares where a stock price closed relative to its price range over a given period.
  - $\%K = \frac{x(t)-L}{H-L}\ x\ 100, H, L$ is the highest and the lowest price in a given time.

In addition, I describe some of indicator formulated used as input variables in previous studies in the as follow:

Variable	Definition	Formula
Rate of change (ROC)	A ratio of price difference between the current price and the price a period time ago.	$\text{ROC} = \frac{x(t) - x(t-K)}{x(t-K)}$
Money flow index (MFI)	A ratio of price difference between the current price and the price a period time ago.	$\text{MFI} = \frac{MFI^{+}}{MFI^{-}}$ $\text{MF} = \text{TP x } \nu$ $\text{TP} = \frac{x_h + x_l + x}{3}$
Relative Strength Index (RSI)	Indicate the internal strength of a single stock	$\text{RSI} = 100 - \frac{100}{1+{}^{U}/_{D}}, U, D$: an average of upward and downward prices changes, respectively

Table 1 - Other technical indicators

2.3 Related works

Many machine learning methods have been used to solve the function approximation problem for an optimal stock trading, and MLPs and SVMs among them, have been most frequently used. As shown in (Kwon & Moon, 2007), the modelled target variable $y(t)$was defined as $\frac{p(t+1)-p(t)}{p(t)}$ where $p(t)$ represents the closing price at day t, i.e. a daily change rate of the next day's closing price over today's. For the input variables, authors used technical indicators being used by financial expert (Kaufman, 2013) or signal. Authors constructed the trading model as follows:

$$\frac{x(t+1)-x(t)}{x(t)} = \mathrm{f}(g_1, g_2, \dots, g_m)$$

Where $g_k(\mathrm{k} = 1, 2, \dots, \mathrm{m})$ is technical indicators or signal. In fact, authors used 75 input variables generated from various technical indicators such as moving average, relative strength index, rate of change, and so on. In particular, 65 and 10 input variables out among 75 ones were created from stock price and trading volume data, respectively. Then MLPs and hybrid genetic was applied over 36 companies in Dow Jones and Nasdaq for 13 years. The result indicated that author's proposed method showed significantly better performance than the "buy-and-hold" strategy because the input node of neural network i.e. stock prices and volume affect the model performance. Various studies also considered Neural network (NN) as a dominant learning method in financial time series prediction. For example, NN approach still can significantly improve the predictability of stock price performance (Panahian, 2011). Feed-forward neural networks are the most commonly used networks for a variety of applications in finance and accounting. The Multi-layer perceptron as well as Radial Basis Function neural network architectures are implemented as classifiers to forecast the closing index price performance (Deo, Srinivasan, & Devanadhen, 2008). The ability of neural networks to discover nonlinear relationships in input data makes them ideal for modeling nonlinear dynamic systems such as the stock market. They have remarkable ability to derive meaning from complicated or imprecise data can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques (Yoon & Swales, 1991), (Patel & Marwala, 2006).

In another study (Chen et al., 2001), the input variables was constructed from moving average indicators such as independent variables consist of $x(t) - EMA_{100}(t), \frac{x(t)-x(t-5)}{x(t-5)}$ and so on and dependent variables $\frac{EMA_3(t+5)-EMA_3(t)}{EMA_3(t)}$, where $EMA_n(t)$ is the n $-$day exponential moving average at day t. And SVMs method was applied over 5 real future contracts collated from the Chicago Mercantile Market and the result showed that SVMs with adaptive parameter can achieve higher performance.

Many other studies also tried to use SVMs to forecast stock value trend price based on price-based predictor and have been proven successfully in (Huang, Nakamori, & Wang, 2005). In (Sapankevych & Sankar, 2009), the authors also examined the forecasting five-time series stock dataset with SVMs to forecast stock value trend price and have been proven successfully.

Most previous studies consider information such as market index, technical indicators, or fundamental factors as inputs variables of neural network. And most technical indicators are derived from closing price, which are used in (Kwon & Moon, 2007). The disadvantage of these previous studies is not to understand adequately contribution of volume in prediction problem performance. It is still limited research studying the volume factor to conclude whether trading volume can contribute to trading performance and there is still room for further analyze this combination. There are a few literatures show that there is relationship between trading volume and price variation. For example, relationship between trading volume and returns and volatility was examined with Pakistan stock market data for the period of July 1998 to October 2008. The result showed that the previous trading volume and can explain on the current market returns (Mubarik & Javid, 2009). In many studies, they show that trading volume can be beneficial factor for long-term forecasting. For example, trading volume and stock closing price can be important determinant of stock prices (Kanas & Yannopoulos, 2001). In (Gallant et al., 1992), it is better to explain the market return by investigating the dynamics of stock price and trading volume than by focusing only on the closing price and technical indicator generated from price. The authors investigated financial time series of daily S&P 500 and total trading volume from 1928 to 1987 including 16127 observations on nonparametric model. The finding suggested that larger price movement associated with higher subsequent volume. Or in (Zhu, Wang, Xu, & Li, 2008), the conclusion is it is possible to modestly improve the network

performance by adding trading volume. And there is one paper, author selected 26 technical indicators and monthly trading volume (Yu, Chen, Wang, & Lai, 2009). This analysis shows that trading volume is strongly related to future stock price movement.

When disagreement (and hence trading volume) is high, it is more likely that bearish investors will wind up at a corner, with their information incompletely revealed in prices. And it is precisely this hiding of information that sets the stage for negative skewness in subsequent rounds of trade, when the arrival of bad news to other, previously more-bullish investors can force the hidden information to come out.

Both theoretical and empirical studies have proven a nonlinear relation between stock return and trading volume. However, whether this nonlinear relationship can be of help to improve the forecasting performance is still an open question. Brooks (1998) explores a number of statistical models (both linear and nonlinear) for predicting the daily stock return volatility of an aggregate of all stocks traded on the New York Stock Exchange market (NYSE). The author finds that lagged volume leads to very modest improvement in forecasting performance. It is inferred that such results are attributed to the transformation method applied in the data, which may lose the important information in trading volume. Brooks (1998) simply focuses on the relationship of trading volume and stock volatility. It is still unclear whether trading volume can improve the forecasting performance of stock return. Kanas and Yannopoulos (2001) introduce stationary transformation of dividends and trading volume as fundamental explanatory variables to neural network models. Their results indicate that inclusion of nonlinear term in the relation between stock returns and fundamentals can improve the out-of-sample forecasting accuracy. However, this study just focuses on long-term (monthly return) forecasting and not short-term, limiting the validation of the conclusion.

In summary, the trading volume was less significantly considered to derive technical indicators. There are still a few studies using volume-based technical indicator. To balance the portion of price-based predictor, this study does not use trading indicators, but closing price and volume trading values are considered ad input variable. We use daily data values, but we also consider dynamic windows size which is needed to be optimized.

CHAPTER 3. PROPOSING METHOD USING TRADING VOLUME AND CLOSING PRICE AS INPUT VARIABLES AND DEFINE TARGET VARIABLE

3.1 Problem formulation

As described in Section 2.1, we also consider the approximation problem in this paper. However, we properly modify it into $\mathbf{y}(t) = \mathbf{f}(X(t))$ where $\mathbf{y}$ is a vector and $\mathbf{f}$ is a vector function considering the output layer can consist of multiple neurons in neural networks. Let $p(t)$ and $v(t)$ be the closing price and the trading volume, respectively, at day t, and we defined $X(t)$ as follows:

$$X(t) = [v'(t-\alpha), v'(t-\alpha+1), \cdots, v'(t-1), v'(t),$$
$$p(t-m), p(t-m+1), \cdots, p(t-1), p(t)]$$

where α is a parameter of the number of past days to be considered for input variables, and $v'(t-i)$ is defined by the ratio of $v(t) - v(t-i)$ over $v(t-i)$. In other words, α determines the time window size of training samples and eventually $X(t)$ includes information on the closing price and the trading volume of the $\alpha + 1$ recent days. In this paper, α ranges from 1 to 10. In addition, we constructed a same number of the volume-specific input variables as the price-specific input variables considering the importance of the trading volume information. With respect to the target variable, we represented $\mathbf{y}(t)$ by a vector of two Boolean values so that they can correspond to two output neurons in our neural networks. Specifically, we defined it as follows:

$$\mathbf{y}(t) = \begin{cases} [1\ 0]^T, & \text{if } p(t+\beta) > p(t) \times (1+\gamma) \\ [0\ 1]^T, & \text{otherwise} \end{cases}.$$

where β is a parameter of the maximum number of future days for a long position and γ is a threshold parameter of the minimum profit rate to sell a stock. Then $\mathbf{y}(t) = [1\ 0]^T$ means that the closing price goes up over $100 \cdot \gamma$ percent within the forthcoming β days. In other words, we could gain a profit of $100 \cdot \alpha$ percent if we bought the stock at the closing price of day t. In this paper, β ranges from 5 to 20 with 5 days interval, and γ ranges from 0.020 to 0.070 with 0.005 interval. As a result, we constructed a set

of 440 parameter combinations $\Sigma = \{(\alpha, \beta, \gamma)\}$ for training, and therefore each specification of parameters represents a different problem formulation.

Once a learning task is completed by MLPs and SVMs, we obtain an approximated function $\tilde{\mathbf{f}}(X(t))$ and interpret it as follows:

$$\hat{\mathbf{y}}(t) = \begin{cases} [1\ 0]^T, & \text{if } \widetilde{\mathbf{f}_1}(X(t)) > \widetilde{\mathbf{f}_2}(X(t)) \\ [0\ 1]^T, & \text{otherwise} \end{cases}$$

where $\widetilde{\mathbf{f}_1}(X(t))$ and $\widetilde{\mathbf{f}_2}(X(t))$ denote the first and the second element of $\tilde{\mathbf{f}}(X(t))$, respectively.

3.2 Overall Framework

Figure 2 shows the overall framework of our approach. To test year Y, two previous consecutive years are used as training and validation sets, respectively. As explained in Section 3.1, the input and the target variables are determined according to a specified parameter combination $(\alpha, \beta, \gamma) \in \Sigma$. Then MLPs and SVMs learn the trading strategy over each of 440 parameter combinations. The best solution over the validation year is assessed with respect to the accuracy and the profit over the test year Y.

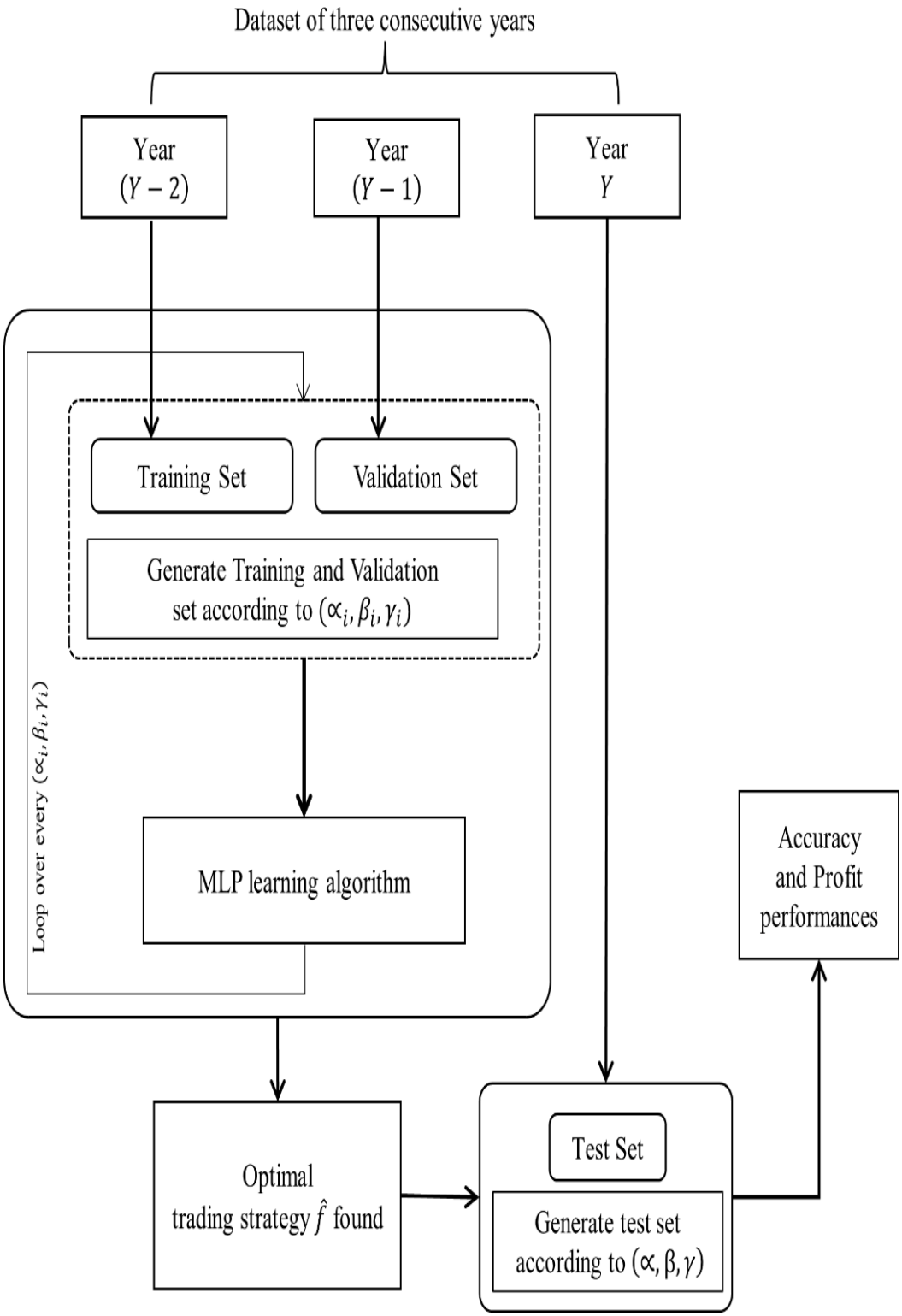

Figure 2 – Overall framework of our method

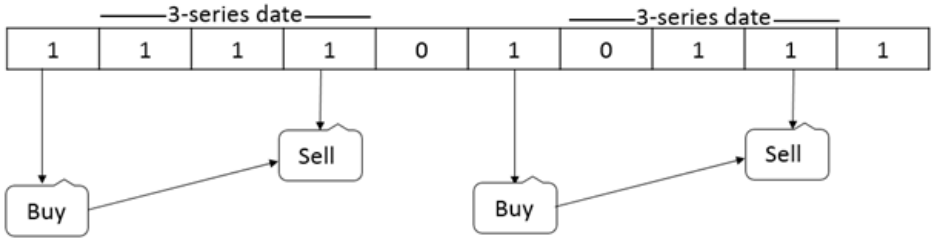

Figure 3 – The buy and sell signals with 3 consecutive

3.3. Trading Strategies

In this study, we compare our trading strategy model with other two trading strategies, which is buy-and-hold strategy. Based on the performance of predicting trading signal, the list of best of parameters that would generate highest accuracy is chosen to apply simulation trading strategy. Different time-lagged would provide different trading signals and the decision selling points based on the period time series considered to maximize to trading profit with optimal parameters,

Figure 3 showed the trading signal and corresponding decision points with 3 consecutive days considered when process the simulation trading. If the predictor system gives the buy signal 1 means, we should take one action buy. The selling point is decided based on the following days parameter which is chosen from the training and that parameter will determine the time-lag series how long investor observer to sell stock.

To compare the profit with this back-testing trading data, we also used very traditional trading strategy buy-and-hold. In the buy-and-hold strategy, the traders hold the stock for long period time, regardless of fluctuation of price. This study considers 1 year holding.

3.4 Performance Evaluation

3.4.1. Overview

The performance is evaluated in testing phase with two primary measurement. One is an accuracy portion of corrected case among total cases in testing phase. The other is ratio trading profit (RTP) after applying our suggested trading strategy with optimal parameter set.

3.4.2. Accuracy measurement

In this study, the objective of applying learning algorithm (SMVs/MLPs) is to predict the trading signal for the whole trading data set.

$$\text{Accuracy} = \frac{\sum_{t=1}^{N} I(\mathbf{y}(t) = \tilde{\mathbf{y}}(t))}{N}$$

where N is the number of days and I is an indicator function which returns 1 if the condition is true, 0 otherwise. The prediction is considered as right prediction if the predicted signal is same as real trading signal output. The total prediction is equal to the total of instances which is examined in data.

3.4.3. Trading Profit ratio

Let $\tilde{\mathbf{y}}(t)$ $(t \in \{1,2,\dots,N\})$ be the predicted results. To evaluate the trading profit of it, we considered a transaction period (b_i, c_i) which is represented by a pair of dates and defined as follows:

$$b_i = \min\{t | c_{i-1} < t \leq N \text{ AND } \tilde{\mathbf{y}}(t) = [1\ 0]^T\} \text{ and}$$

$$c_i = \min\{N, b_i + \beta, \min\{t | p(t) \geq (1+\gamma)p(b_i)\}$$

For convenience, it is assumed that $c_0 = 0$. By these definitions, b_i means the earliest day when the buy signal is generated since the last transaction ends. On the other hand, c_i means the day to sell the stock which was bought at day b_i considering the deadline and the minimum expected profit ratio (γ). Then (b_i, c_i) represents i-th trading transaction where the stock is bought and sold at days b_i and c_i, respectively. The final trading profit of a test year is calculated as follows:

$$\text{RTP} = \sum_{i=1}^{T} \left(\frac{p(c_i)}{p(b_i)} - (1+\eta)\right)$$

where T denotes the number of transactions and η means the transaction cost which was set to 0.025% in this study.

By considering the transaction fee, RTP value represents more realistic trading profit.

CHAPTER 4. EXPERIMENTAL RESULTS

4.1 Overview

We implemented our method by using LibSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) and Sci-kit learn (http://scikit-learn.org/) library. In this study, the MLP consists of one input layer, one hidden layer and one output layer. It is learned by the backpropagation algorithm with the learning rate of 0.025 and the momentum rate of 0.9. The number of hidden neurons was 16. The sigmoid function is chosen as a transfer function. On the other hand, the parameters of SVM are set as follows: C=100, gamma=1, degree=2 of a polynomial kernel. For more stable results, 5-fold cross validation is applied as suggested in (Hsu, Chang, & Lin, 2003). In 5-fold cross validation, the training set is first divided into 5 subsets of equal size, then one subset is tested using the classifier trained on the remaining 4 subsets sequentially. In addition, the transaction cost was set to 0.025% for every buying and selling transaction through the trading experiments.

4.2 Datasets specification

In this study, we tested our approach with six stocks: Hang Seng Index (HSI), NASDAQ Composite Index (NASDAQ), Financial Times Stock Exchange 100 Index (FTSE), Nikkei 225 Index (NIKKEI), Swiss Market Index (SMI) and Google (GOOGLE). We collected the daily closing prices and the trading volumes of them from January 1999 to December 2015. Then the test year varies from 2001 to 2015 as explained in Fig. 1. For each test year and each stock, we considered 440 combinations of three modelling parameters (α, β, γ). In other words, a total of 39600 datasets (=6 stocks × 15 test years × 440 parameter combinations) were learned by our approach.

4.3 Result of accuracy

We first compared the accuracy of our approach and the expected accuracy over all considered parameter combinations (Figure 4 and Figure 5). In other words, the latter value means the average accuracy of neural networks learned over 440 parameter combinations. We need to remind that the best modelling parameter in our approach is chosen based on not the result of the test year but the result of the validation year. Thus,

it is not guaranteed that the accuracy of our approach is always better than the expected accuracy. Figures 4 and 5 show the results of MLPs and SVMs, respectively.

As shown in the figures, our approach showed significantly higher accuracy than the expected average accuracy irrespective of the stock type and the test year. For example, the average accuracy (93.856%) of our approach with MLPs over 15 test years was higher than that (61.455%) of the expected accuracy by 1.5 times in the case of NASDAQ. The accuracy improvements of other stocks by our approach with MLPs were 92.53%, 95.32%, 92.5%, 92.9%, and 94.3% in the case of GOOGLE, HSI, NIKKEI, FTSE, and SMI, respectively. In addition, the number of test years where the accuracy of our approach with MLPs is higher than 0.9 were 12, 9, 13, 10, 10 and 13 in NASDAQ, GOOGLE, HSI, NIKKEI, FTSE, and SMI, respectively. This impressive performance was consistently observed in our approach where SVMs were used as a learning algorithm (Figure 4). These results validate the efficiency of our approach. In particular, it is more interesting that the optimized model parameters dramatically improved the prediction accuracy.

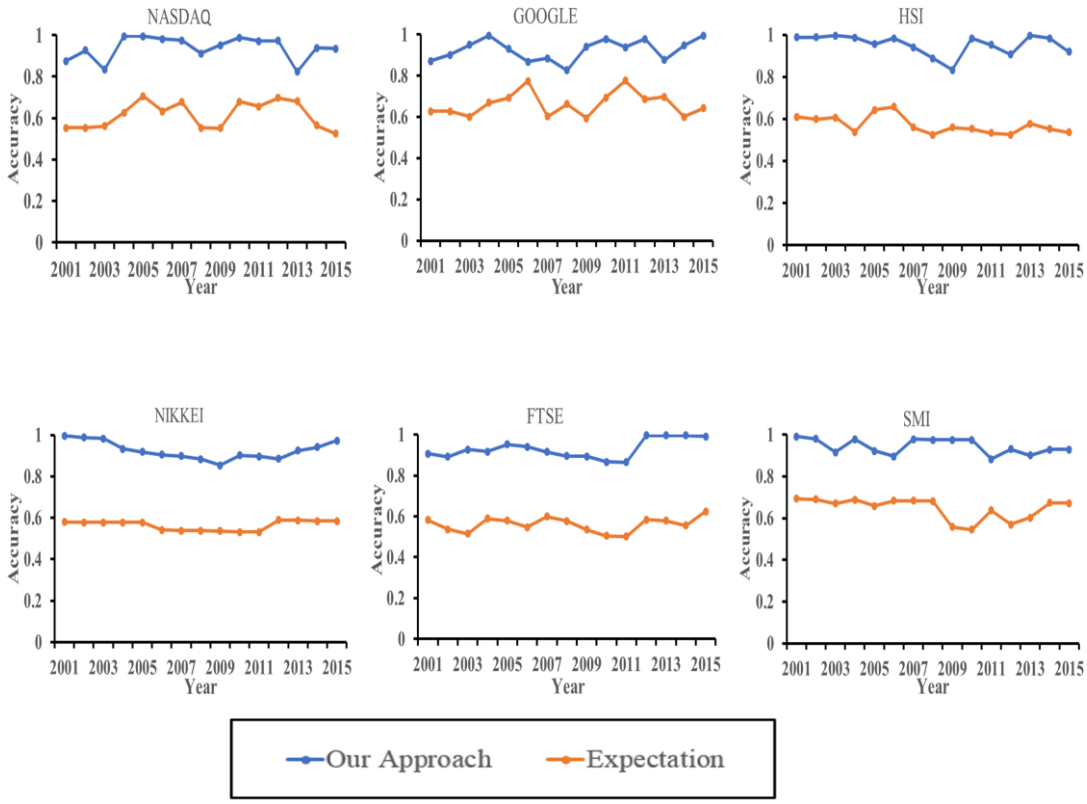

Figure 4 – Accuracy prediction result of MLPs

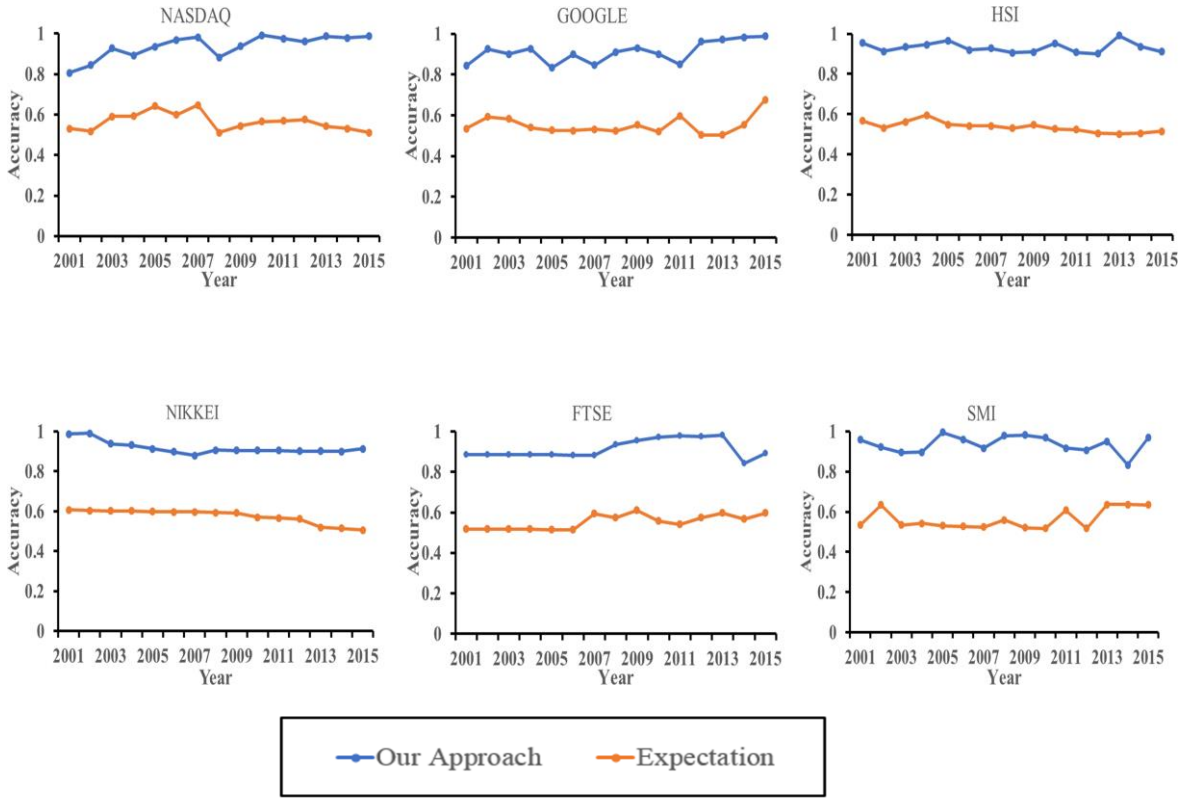

Figure 5 – Accuracy prediction result of SVMs

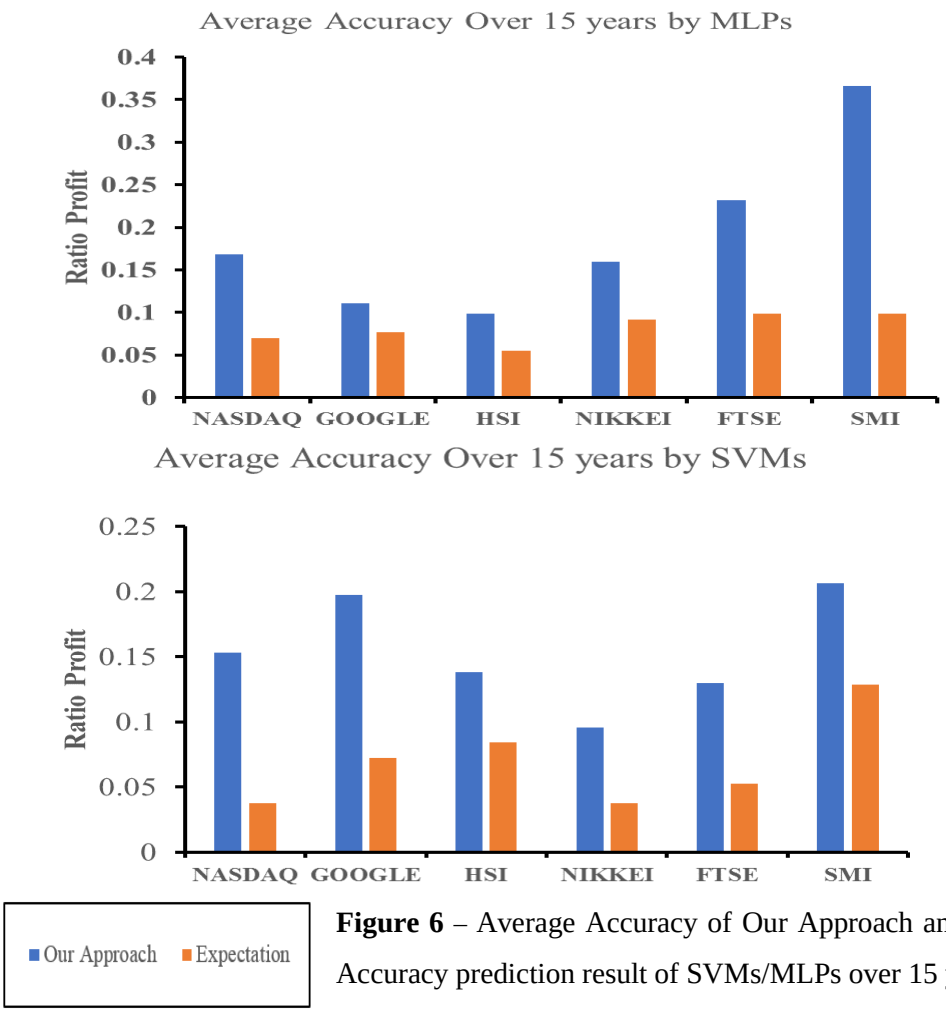

Figure 6 – Average Accuracy of Our Approach and Expected Accuracy prediction result of SVMs/MLPs over 15 years

Figure 6 show the average accuracy over 15 years (from 2000 to 2015) of total 6 stock indices. The average result of our approach fluctuates around 0.92 to 0.96 with both method. For more detail for every single stock index, the exact statistic number is displayed in the table 2. In general, our approach average takes 1.5 times than the other over 15 years.

Index	MLPs		SVMs	
	Our Approach	Expectation	Our Approach	Expectation
NASDAQ	0.938559687	0.614551449	0.937365001	0.564502977
GOOGLE	0.925309155	0.663385934	0.911105427	0.551233522
HSI	0.953190696	0.571892446	0.931703651	0.536101483
NIKKEI	0.92500763	0.563682459	0.918557761	0.575858852
FTSE	0.92992995	0.560678208	0.914930485	0.553998579
SMI	0.943753861	0.646958706	0.936759646	0.564037457

Table 2 – Average Accuracy of Our Approach and Expected Accuracy prediction result of SVMs/MLPs over 15 years

4.4. Results of Profits

We further examine the performance with respect to the trading profit. In addition to the comparison with the expected profit over all combinations of modelling parameters, we further compared the profit of our approach with the trading profit by the Buy-and-Hold strategy which means that the trader buys the stock at the closing price of the first day and sells it at the closing price of the last day of the year. Figure 7 show the results when MLPs and SVMs are used as a learning algorithm, respectively. As shown in these figures, our approach showed significantly higher profit than the expectation over all parameters and the Buy-and-Hold strategy except but the case of HSI by MLP prediction. The best profit case ratio is in 2010 in case of NASDAQ stock index with 1.5 times or more and 1.19 times or more with our method and average method, respectively

Figure 7. Profit comparison. (a) Profit comparison by MLP-based predictor; (b) Profit comparison by SVM-based predictor

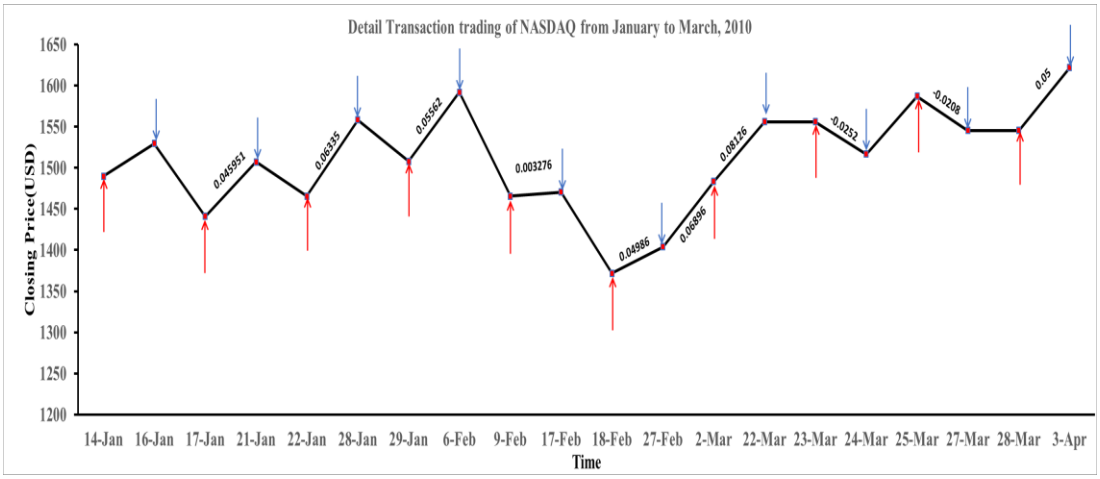

Figure 8 – Detail transaction Buy- Sell of NASDAQ Index from January to March 2010 with parameter set (3,10,0.005). The red Upper and blue Down arrow represents the buy action and corresponding sell action. The weight between two buying and selling point is the gain for one stock in one transaction.

Figure 8 shows the detail profit corresponding to every transaction period which is conducted from January to March 2010 with specific transaction parameter (3,10,0.05). The red Upper and blue Down arrow represents the buy action and corresponding sell action. The weight between two buying and selling point is the gain for one stock in one transaction.

The signal of above simulation trading is based on the output result of MLPs predicator. For more detail, the number of transaction taken in one year is displayed in the table 2.

Table 3 shows a detailed result where our approach with MLP traded NASDAQ in 2010. As shown in the table, a total of 23 transaction occurred and 20 and 3 ones among them were profitable and loss, respectively.

Date Buy	Price Buy	Date Sell	Price Sell	Gain/Loss
14-Jan	1489.640015	16-Jan	1529.33	0.026643981
17-Jan	1440.859985	21-Jan	1507.07	0.045951697
22-Jan	1465.48999	28-Jan	1558.34	0.063357632
29-Jan	1507.839966	6-Feb	1591.71	0.05562261
9-Feb	1465.48999	17-Feb	1470.66	0.00352786
18-Feb	1371.640015	27-Feb	1555.77	0.134240765
2-Mar	1404.02002	22-Mar	1587	0.130325763
23-Mar	1483.47998	24-Mar	1555.77	0.048730041
25-Mar	1557.123	27-Mar	1516.45	-0.026120608
27-Mar	1545.199951	3-Apr	1621.87	0.049618202

6-Apr	1606.709961	13-Apr	1653.31	0.029003429
15-Apr	1626.800049	17-Apr	1673.07	0.028442277
20-Apr	1608.209961	24-Apr	1694.29	0.053525398
27-Apr	1679.410034	4-May	1763.56	0.050106897
5-May	1754.119995	2-Jun	1836.8	0.047134777
5-Jun	1759.099976	8-Jun	1842.4	0.047353788
7-Jul	1746.17004	7-Jul	1746.17	2.29073E-09
24-Jul	1965.959961	27-Jul	1967.89	0.000981736
5-Aug	1993.050049	10-Aug	1992.24	-0.000406442
3-Sep	1983.199951	8-Sep	2037.77	0.027516171
11-Sep	2080.899902	16-Sep	2133.15	0.025109329
21-Sep	2138.040039	24-Sep	2107.61	-0.01423263
20-Oct	2163.469971	29-Sep	2124.04	-0.01822532

Table 3 – Detail transaction of NASDAQ Index in the whole 2010

4.5. Result of comparison without considering volume

In order to analysis the importance of volume, the HSI and NASDAQ is picked out among the whole 6 data sets, we compared performance between our proposed method with both volume and closing price as input variables and without volume as input variables. Our proposed method with the volume trading data and closing price as input show better result in terms of accuracy and profit in the Figure 10 NASDAQ Index. This result is also consistent with HSI Index. The proper reason explained for this result is due to the volume trading information affect to the performance of model.

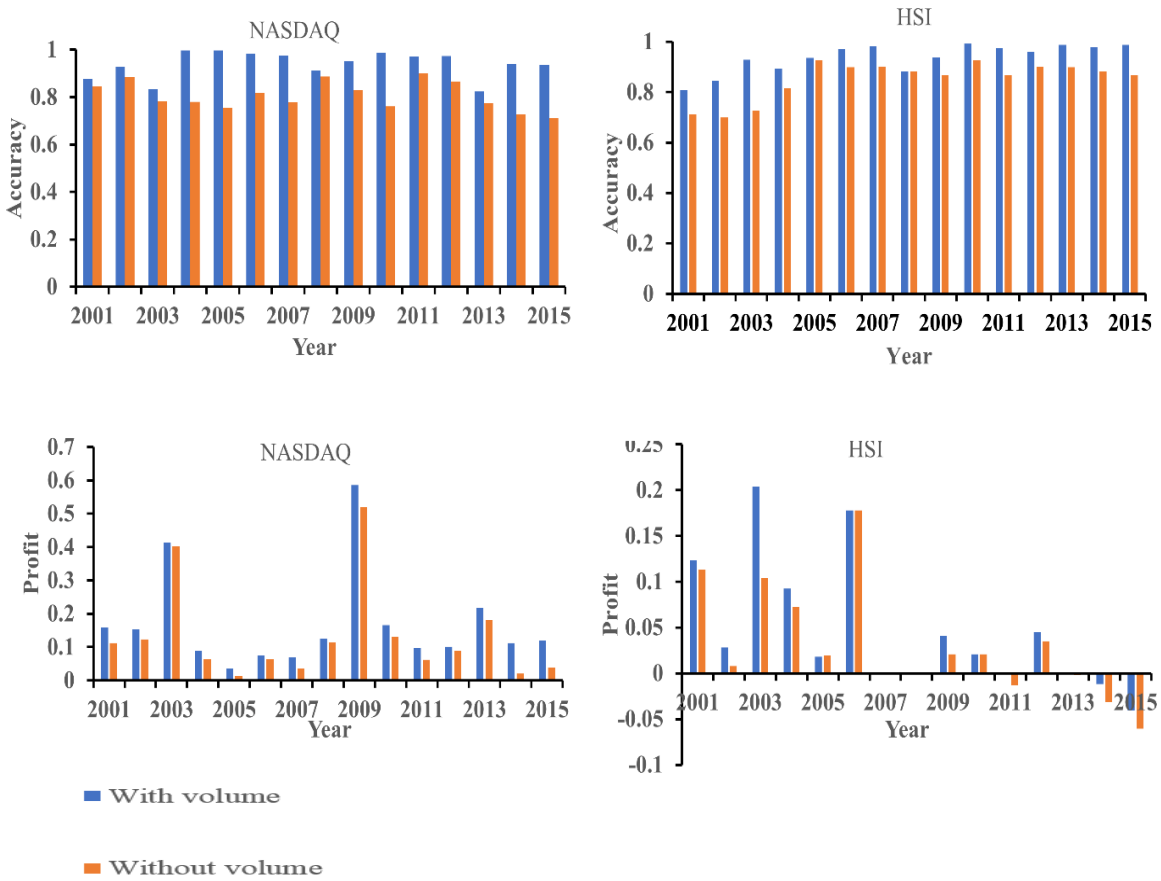

Figure 9 – The comparison performance of HSI & NASDAQG Index result between our model and other model without volume as input data by MLPs learning method

CHAPTER 6. CONTRIBUTION SUMMARY AND FURTHER WORK

6.1 Contribution summary

In this paper, we proposed prediction problem for optimal daily trading stock by using closing price and trading volume value as input variables. For the input nodes of the neural network, we transform volume trading value which is constrained by window size parameters. To define the target variables, other parameters regarded to the period time holding and profit is used. Both learning methods SVMs with and MLs are used independently to evaluate the efficient of proposed model. It was tested with total 6 stock indexes for 15 years from 2001 to 2015 and showed the significantly better performance than the "buy-and-hold" classical trading strategy. To make the trading become much more reasonable, we also applied the transaction cost which is 0.025% per every single transaction. The result of proposed method also outperformed average performance in terms of profit trading and accuracy prediction by choosing the optimal parameters. The result showed the consistent pattern among the whole data set. Therefore, this turned out to conclude that trading volume information can contribute to predict the future price change. And the proposed model also proved to be profitable even when considering transaction costs. It can be explained as the volume trading data and closing price affect to the deciding buying and selling points.

Regarding the modelled parameters, this research deals with set of 3 kinds of parameter, which are parameter of the number of past days to be considered for input variables, parameter of the maximum number of future days for a long position and γ is a threshold parameter of the minimum profit rate. These dynamic parameters contribute to accurate of the SVMs/MLPs predictor because it covers all cases needed to be consider around the problem space. Also, these parameters can help the application for further research will become more flexible in the financial data time series forecasting.

6.2 Future Work

Future study will construct an automatic intelligence trading system that allow user to select optional choices extra; that is trading with only volume values and with price only. And to balance with the price-based indicator in the stock market, volume-based

indicators can be formulated and consider for further research. Since many technical indicators are ever popular used in previous studies, the future work will choose the most powerful technical indicators which is the most efficient prediction when combining with volume values.

Regarding to the volume-based indicator, future study will investigate volume indicators to balance with price-based indicator. Since it can be a useful information for trader to make decision before getting start trading. Also, trading algorithm is explored in a future work to have a significant profit with a good predictor.

REFERENCES

Abhyankar, A., Copeland, L. S., & Wong, W. (1997). Uncovering nonlinear structure in real-time stock market indexes: The S&P 500, the DAX, the Nikkei 225 and the FTSE-100. *Journal of Business & Economic Statistics, 15*, 1-14.

Bhat, A. A., & Kamath, S. S. (2013). *Automated stock price prediction and trading framework for Nifty intraday trading.* Paper presented at the Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on.

Cao, L.-J., & Tay, F. E. H. (2003). Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on Neural Networks, 14*(6), 1506-1518.

Chen, J., Hong, H., & Stein, J. C. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of financial Economics, 61*(3), 345-381.

Deo, M., Srinivasan, K., & Devanadhen, K. (2008). The empirical relationship between stock returns, trading volume and volatility: Evidence from select Asia-Pacific stock market. *European Journal of Economics, Finance and Administrative Sciences, 12*(October), 23-41.

Devadoss, A. V., & Ligori, T. A. A. (2013). Forecasting of stock prices using multi layer perceptron. *International journal of computing algorithm, 2*, 440-449.

Gallant, A. R., Rossi, P. E., & Tauchen, G. (1992). Stock prices and volume. *The Review of Financial Studies, 5*(2), 199-242.

Hamid, S. A., & Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P 500 Index futures prices. *Journal of Business Research, 57*(10), 1116-1125.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification.

Kanas, A., & Yannopoulos, A. (2001). Comparing linear and nonlinear forecasts for stock returns. *International Review of Economics & Finance, 10*(4), 383-398.

Kaufman, P. J. (2013). *Trading systems and methods*: John Wiley & Sons.

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing, 55*(1), 307-319.

Kotler, P. (1971). Marketing decision making: A model building approach.

Kwon, Y.-K., & Moon, B.-R. (2003). *Daily stock prediction using neuro-genetic hybrids.* Paper presented at the Genetic and Evolutionary Computation—GECCO 2003.

Kwon, Y.-K., & Moon, B.-R. (2007). A hybrid neurogenetic approach for stock forecasting. *IEEE Transactions on Neural Networks, 18*(3), 851-864.

Lin, X., Yang, Z., & Song, Y. (2011). Intelligent stock trading system based on improved technical analysis and Echo State Network. *Expert Systems with Applications, 38*(9), 11347-11354.

Mittermayer, M.-A. (2004). *Forecasting intraday stock price trends with text mining techniques.* Paper presented at the system sciences, 2004. proceedings of the 37th annual hawaii international conference on.

Mubarik, F., & Javid, A. Y. (2009). Relationship between stock return, trading volume and volatility: Evidence from Pakistani stock market.

Panahian, H. (2011). Stock market index forecasting by neural networks models and nonlinear multiple regression modeling: study of Iran's capital market. *American Journal of Scientific Research, 18*, 35-51.

Patel, P. B., & Marwala, T. (2006). *Forecasting closing price indices using neural networks.* Paper presented at the Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on.

Pring, M. J. (2015). *Martin Pring's Introduction to Technical Analysis*: McGraw Hill Professional.

Roh, T. H. (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications, 33*(4), 916-922.

Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine, 4*(2).

Tino, P., Schittenkopf, C., & Dorffner, G. (2001). Financial volatility trading using recurrent neural networks. *IEEE Transactions on Neural Networks, 12*(4), 865-874.

Yoon, Y., & Swales, G. (1991). *Predicting stock price performance: A neural network approach.* Paper presented at the System Sciences, 1991. Proceedings of the Twenty-Fourth Annual Hawaii International Conference on.

Yu, L., Chen, H., Wang, S., & Lai, K. K. (2009). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on evolutionary computation, 13*(1), 87-102.

Zhu, X., Wang, H., Xu, L., & Li, H. (2008). Predicting stock index increments by neural networks: The role of trading volume under different horizons. *Expert Systems with Applications, 34*(4), 3043-3054.