2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A New Investment Strategy Based on Data Mining and Neural Networks

Chang Liu and Hafiz Malik
Department of Electrical and Computer Engineering, University of Michigan - Dearborn, Dearborn, Michigan 48128
Email: liuchang_01@yahoo.com, hafiz@umich.edu

***Abstract*—In this paper, we present a new investment strategy for optimal gains on investments in the stock market. Neural Network (NN)-based framework is used for trading prediction and forecasting. To this end, statistical measures based on *return* and *volatility* are used to filter out low performing sectors in the stock market. A simple but effective method based on price *Simple Moving Averages (SMAs)* is used to measure volatility for a given stock. The proposed NN-based system uses the strongest performing indices for stock market forecasting. In addition to predicting investment decisions such as *Buy* or *Sell*, the proposed framework also aims at maximizing investment gains (or returns). The proposed NN-based framework rely on historical data and provides investors investing strategies for optimal trading. Training data is extracted extracted from historical weekly data (from the Yahoo Finance). Simulation results indicate that the proposed framework can help investors making investment decisions and increasing their trading profitability.**

I. Introduction

Equity market is a complicated non-linear system driven by many factors. Investors are constantly seeking good strategies to help them making investment decisions. There is a huge amount of data generated by the stock markets around the world. For example, there are about 2788 stocks listed on the NASDAQ, around 1865 stocks listed on the New York Stock Exchange (NYSE), and so on. Each stock prices are updated in ticks and stored on servers for further analysis. The data mining techniques, e.g., association rule mining, correlation analysis, supervised learning based on decision trees, artificial neural networks, support vector machines, fuzzy logic, and so on to classify or predict the price movement are very suitable for handling the enormous amount of data collected from the stock markets.

Data mining techniques have been widely used for the stock analysis. The stock market analysis can be broadly divided in two types of analysis, that is, (i) Fundamental Analysis, and (ii) Technical Analysis. The data mining techniques can be based on either type. Investors generally rely on cluster analysis and decision trees for selecting potential stock candidates for investing strategies. The clustering based techniques rely on features related to the fundamental aspects of the equity, such as earnings per share, business growth rate and profit margin, price-earnings ratio, etc. [1]. One of the limitations of cluster analysis based method is that it requires number of clusters before running the algorithm. Decision tree based methods addresses are interpretable, robustness for a variety of data and measurement levels and easy to implement [2]. Recently, decision tree based methods have proposed for the stock selection, e.g., in [2], five variables, that is, return on equity, return on assets, analyst opinion, growth this year, and price are used to for train decision tree for 'Price Trading' and 'Growth Trading' strategies and identify the top 6 stock candidates. Neural network (NN)-based methods have also been extensively used for stock data [2]–[4]. The power of NN-based modeling lies on the ability to approximate any non-linear function with arbitrary accuracy if appropriate network architecture. For exam, in [2], same-set of five variables is used as the inputs for training NN and probability of increasing price is use to predict the output. The NN-based methods have also been proposed for price forecasting. For example, 30-day historical prices are used to predict the closing price of the current day in [4]. A sliding window-based technique is used to obtain training samples for the underlying neural network. Similarly, in [3] historical data based on past five days High, Low and Closing prices are used as input to the network to predict the current day's closing price; and *Adaptive Neural Fuzzy Inference System* (ANFIS)-based approach is used for stock forecasting.

Mining of sequential patterns can also be used in the stock analysis, for example in [5], a "transaction record" is defined as a sequence of events characterized by *highest price up*, *highest price down*, *volume up*, *volume down*, and so on and the last event is called the "*target*" event, which is produced based on a technical indicator MACD and assumes a "**Buy**" or "**Sell**" value. The discovered association rules can be used to suggest a future Buy or Sell strategy. Correlation analysis is used to measure the interrelationship between the stock prices of two companies. A strong correlation is used determine leading stock which is used to predict the behavior of lagging stocks [6]. Recently, core pattern identification-based framework is proposed for stock market prediction [7]. A core pattern is a representative group of stocks that show coherent behavior specific to their sector [7]. Multiple core patterns may exist in one sector concurrently. The coherence detected within a specific sector may be an indication that the stocks are influenced by the same factors and may continue to do so over a long period.

The main contribution of this paper is Neural Network (NN)-based framework to predict the **Buy long** or **Sell short** trading strategies with their potential profits based on a mechanic trading system. We propose a new comparison metric for evaluating the performance of various Exchange-Traded Funds (ETFs). The best performing sector are used for forecasting strategy. The proposed NN-based framework aims at providing a suggested **Buy** or **Sell** strategy combined with a predicted profit margin. Instead of focusing on a short-term profit based on the prediction of the next closing price, our strategy in general suggests a longer term trade, which could

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

last for several weeks or even months and may result in a higher profit.

II. DETERMINE THE BEST PERFORMING SECTOR OF THE MARKET

A. Exchange-Traded Funds (ETFs)

Although a security can be either bought *long* or *sold short*, most investors prefer to trade on the best performing stocks, because buying a strong steady uptrend increases the chance of profitability and at the same time, exposes to less risks of losing the capital. There are thousands of equities on the markets, selecting the best performing ones to invest is a challenging task. *Exchange-Traded Fund (ETF)* is mostly used to track an index, a commodity, or a basket of assets of a specific sector, but traded like a stock on an exchange. For the sack of simplicity, the *Weekly Price Data* of the following ETFs is used to represent the behavior of their corresponding sectors.

TABLE I. ETFs FOR VARIOUS MARKET SECTORS

Symbol	Description
XLU	- SPDR Utility ETF
XLE	- SPDR Energy ETF
XLB	- SPDR Materials ETF
XLF	- SPDR Financial ETF
XLI	- SPDR Industrial ETF
XLV	- SPDR Healthcare ETF
XLK	- SPDR Technology ETF
XLY	- SPDR Consumer Discrete ETF
XLP	- SPDR Consumer Staples ETF

To evaluate the performances for a specific period the following factors are considered:

- Final investment return at the end of the target period
- Maximum return during the target period
- Minimum return during the target period
- Volatility compared with the Standard and Poor's 500 (S&P500) market index.

B. Volatility Measure

The volatility of an equity is measured in terms of *Standard Deviation* (STD). In our case, a 10-week *Simple Moving Average (SMA)* is constructed for the entire evaluation period except the first 9 weeks. Firstly, we convert the equity's price in US dollar amount to the price change in percentage as illustrated in Eq.(1). It is important to mentioned that after this conversion, value of very first price point of the target period is always 0%.

$$P(i) = \begin{cases} 0 & i = 0 \\ \frac{P_{\$}(i) - P_{\$}(i-1)}{P_{\$}(i-1)} & i = 1, 2, 3, \cdots \end{cases} \quad (1)$$

where i is the index of the week and $P_{\$}(i)$ is the closing price in US dollar of the i^{th} week.

The 10-week SMA is then calculated using Eq.(2) based on the price change values measured in percentage.

$$SMA_{10}(i) = \begin{cases} 0 & i = 0, 1, 2, \cdots, 8 \\ \frac{\sum_{k=0}^{9} P(i-k)}{10} & i = 9, 10, 11, \cdots, (N-1) \end{cases} \quad (2)$$

where N is the total number of weeks included in the target period, $SMA_{10}(i)$ is the 10-period SMA value for the i^{th} week.

The volatility between the true prices and the SMA_{10} values is computed as:

$$Volatility = \sqrt{\frac{\sum_{i=9}^{N} [P(i) - SMA_{10}(i)]^2}{N}} \quad (3)$$

It is important to highlight that the proposed volatility measure is different from the traditional STD-based measure where volatility is computed using STD of each price against the average price value of the entire period. In that case, the underlying assumption is that the expected return can be modeled by a *Normal Distribution* with the *Expected Value* as the average price of the entire period. The investors generally consider the Moving Average as the trend of the market. Therefore the proposed STD computed using SMAs reflects better measure of true volatility seen by the investors.

C. Results of the Comparison

Shown in Fig.1 is the S&P500 index and its 10-week SMA at the **2008 market crash**. The corresponding price charts comparing the S&P 500 with some ETFs are given in Fig.2. Finally, the Fig.3 illustrates the same comparison after the 2008 crash. The data used here is collected from the Yahoo Finance [8].

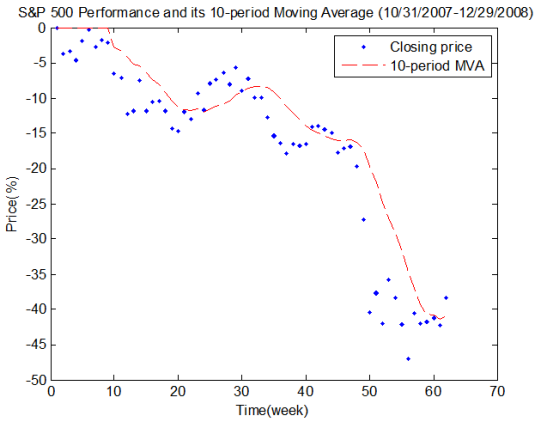

Fig. 1. S&P500 Index price chart and its 10-week Simple Moving Average at 2008 market crash (10/31/2007 to 12/29/2008)

The proposed comparison metric is summarized in Tab.II and Tab.III. It can be observed from Tab.II and Tab.III that that the Consumer Staples (XLP) sector was the most stable sector and outperformed most of the sectors for the periods both before and after the 2008 market crash. On the other hand, the Financial (XLF) sector was the worst performing sector during both periods.

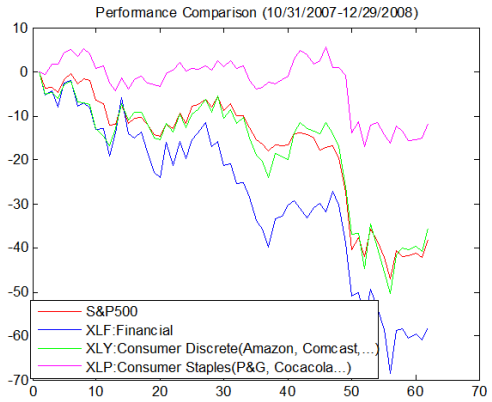

Fig. 2. Comparison between S&P500 Index and some ETFs at the 2008 market crash. (10/31/2007 to 12/29/2008)

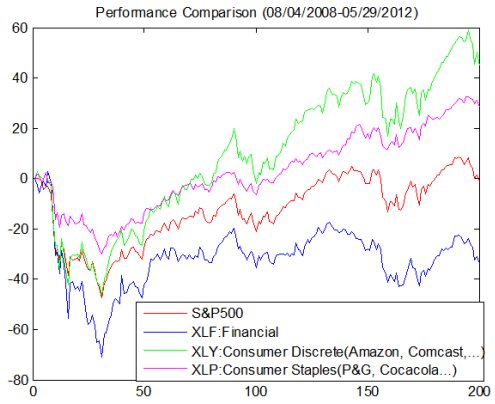

Fig. 3. Comparison between S&P500 Index and some ETFs after the 2008 market crash. (08/04/2008 to 05/29/2012)

TABLE II. PERFORMANCE COMPARISON BETWEEN S&P500 AND ETFS (10/31/2007 TO 12/29/2008)

Symbol	MaxReturn(%)	MinReturn(%)	Volatility(%)	FinalReturn(%)
S&P500	0	-47.01	6.24	-38.28
XLU	5.39	-36.61	5.79	-25.95
XLE	17.90	-42.13	9.36	-33.14
XLB	8.84	-51.85	8.48	-43.55
XLF	0	-68.50	7.74	-58.20
XLI	0.20	-47.80	7.11	-38.45
XLV	3.77	-32.21	5.28	-22.18
XLK	0	-50.1	6.65	-42.55
XLY	0	-50.37	7.35	-35.57
XLP	5.56	-16.90	4.5	-11.75

The best performing ETF (XLP) are selected for performance evaluation of the proposed NN-based trading strategy (discussed in the Section III).

III. MARKET FORECASTING USING BP NEURAL NETWORK

A. Neural Network Topology

A *Multilayer Feed-forward Neural Network (NN)* with an input layer of 22 neurons, a hidden layer of 30 neurons, and an output layer of two nodes. Our selection is based on a heuristic approach. By experimenting various topologies with different number of hidden neurons, we observed that this topology yields the best prediction results and there is no improvement by further increasing the number of hidden layer neurons. The input layer of the Neural Network assumes 22 inputs, which include the previous thirteen weekly Closing prices plus the Open, High and Low prices of the last three weeks. Two NN's outputs are labeled as *Action* output and *Profit* output. The *Action* output recommend a *Buy long* or a *Sell short* or *No action* and the *Profit* output provides the predicted profit if the suggested trade action takes place. This resulted in a fixed 22-30-2 network topology for stock market prediction and forecasting.

TABLE III. PERFORMANCE COMPARISON BETWEEN S&P500 AND ETFS (08/04/2008 TO 05/29/2012)

Symbol	MaxReturn(%)	MinReturn(%)	Volatility(%)	FinalReturn(%)
S&P500	16.65	-46.48	4.89	6.37
XLU	11.32	-36.27	3.67	10.47
XLE	18.04	-45.18	6.17	-7.57
XLB	13.13	-51.9	6.2	-6.39
XLF	3.02	-71.11	6.25	-33.77
XLI	14.62	-56.35	5.78	1.71
XLV	20.63	-32.81	4.02	13.89
XLK	35.82	-41.32	5.22	22.44
XLY	59.26	-46.09	6.41	44.92
XLP	32.65	-30.14	3.35	28.21

B. Training Examples Preparation

On the basis of performance comparison of multiple sector ETFs, the Consumer Staples sector ETF (XLP) are selected for the proposed NN-based forecasting. The training data is collected from historical weekly prices from the Yahoo Finance [8]. The weekly *Open*, *High*, *Low* and *Closing* prices are immediately available from the Yahoo Finance data server [8]. These prices are then converted from the US dollars to the changes in percentage using Eq.(1).

Each training example consists of 22-dimensional input vector and 2-dimensional output targets (e.g., *Action* output and *Profit* output). The target values of a training example are correct "Buy long / Sell short / No Action" and the profit generated. In our case, we assume Buy long=1, No action=0, and Sell short=-1. If "No action" is taken, the profit is set to 0. The profit is measured in percentage. For a given week, we construct a training example by collecting the 22 inputs from the historical prices and determining the target values, i.e. the correct trading action and the profit, based on a "*Trailing Stop-Loss*" trading system:

At the closing of a given week, we assumed a *Buy long* trade is taken and set a trading stop-loss of -3% and an initial profit target of 6%. The implication is that if price drops 3% before reaching a positive 6%, this trade will be considered as a failure trade and we will evaluate if a *Sell Short* execution can generate any profit. If the price increases 6% and never dropped more than -3% before that, we then set our stop-loss to +3%. Therefore, from this moment and forward, we are guaranteed to have a minimum of +3% profit. Each time the price increases again, the stop-loss will be also moved

higher to always keep a 3% margin below the price's new high. However, the stop-loss will never move lower if the price drops. This is called a "*Trailing Stop*" strategy. If the price continuously drops lower and hits the stop-loss level, we sell all the positions at the stop-loss level and the trade is closed. The corresponding *Profit* of this trade is recorded as the target profit for this *Buy long* execution. Similarly, we can evaluate if a *Sell Short* execution is profitable. If neither a Buy nor a Sell leads to a profit, the target execution for this given week will be set to *No action* and the target profit will be recorded as zero.

The training dataset is constructed for the period of 10/31/2007 to 12/20/2010. Shown in Fig.4 is the plots of the perfect *Buy Long* or *Sell Short* executions with the *Profits* generated based on the above trading strategy.

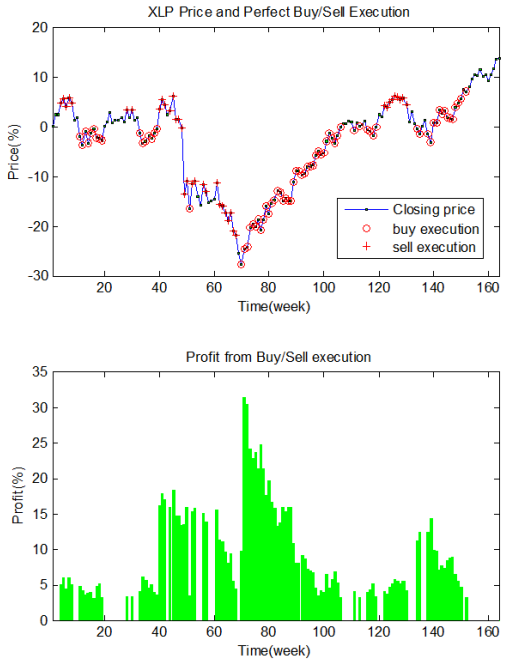

Fig. 4. Perfect Buys and Sells on each week with their generated profits. (10/31/2007 to 12/20/2010)

The predicted outputs of the trained Neural Network on the same period are shown in Fig.5. In this example, we took a *Buy Long* trade if the *Action* output of the Neural Networks is greater or equal than 0.5, and a *Sell Short* trade if the value is less or equal than -0.5.

The two confusion matrices (Tab.IV and Tab.V) were generated for the Buy and Sell actions predicted by our NN. The top left, top right, bottom left and bottom right values represent the counts and percentages of True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN).

The statistical measures for the predicted Buys and Sells is then computed using Eq.(4), which are summarized in Tab.VI.

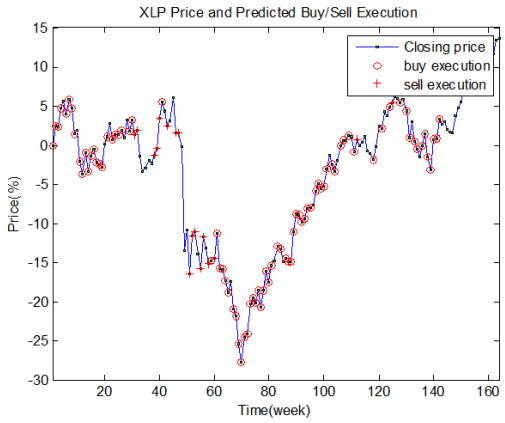

Fig. 5. Buys and Sells predicted by the trained Neural Network on Training Data (10/31/2007 to 12/20/2010)

TABLE IV. BUY CONFUSION MATRIX FOR NN-BASED PREDICTION ON THE TRAINING DATASET (10/31/2007 TO 12/20/2010)

		True	
		Buy	Not Buy
Predicted	Buy	63 (44.4%)	21 (14.8%)
	Not Buy	13 (9.2%)	45 (31.7%)

TABLE V. SELL CONFUSION MATRIX FOR NN PREDICTION ON THE TRAINING DATASET (10/31/2007 TO 12/20/2010)

		True	
		Sell	Not Sell
Predicted	Sell	15 (10.6%)	2 (1.4%)
	Not Sell	18 (12.7%)	107 (75.4%)

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ Accuracy &= \frac{TP+TN}{TP+FP+FN+TN} \end{aligned} \tag{4}$$

TABLE VI. PERFORMANCE MEASURES FOR PREDICTED BUYS AND SELLS (10/31/2007 TO 12/20/2010)

Operation	Precision(%)	Recall(%)	Accuracy(%)
Buy	75.0	82.9	76.1
Sell	88.2	45.5	85.9

The predicted profits are compared against true Buy and Sell events, as shown in Fig.6. The Mean-Square-Error (MSE) is 8.7% and the correlation coefficient is 0.66 of the predicted profits.

C. Neural Network Performance Evaluation

To test the forecasting capability of the proposed NN-based framework, the trained neural network is used to analyze

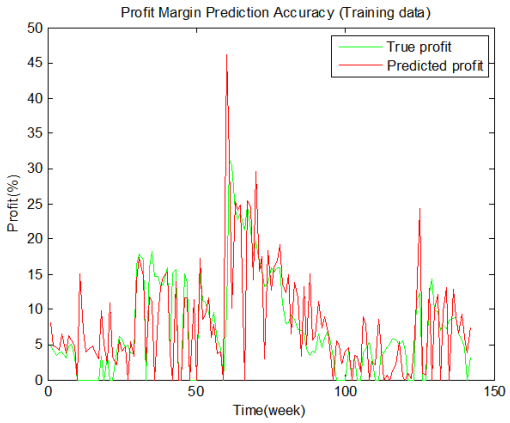

Fig. 6. Profits from the perfect Buys and Sells and profits predicted by the trained Neural Network (10/31/2007 to 12/20/2010). (MSE=8.7%, Corr.Coef.=0.66)

the datased collected from 01/04/2010 to 02/13/2012. The predicted outputs were again compared with the true Buy or Sell events. It is important to noting that the proposed NN used here was trained on the historical data from the period of 10/31/2007 to 12/20/2010. Shown in Fig.7 and 8 illustrate the true Buys/Sells and the predicted trades by the network.

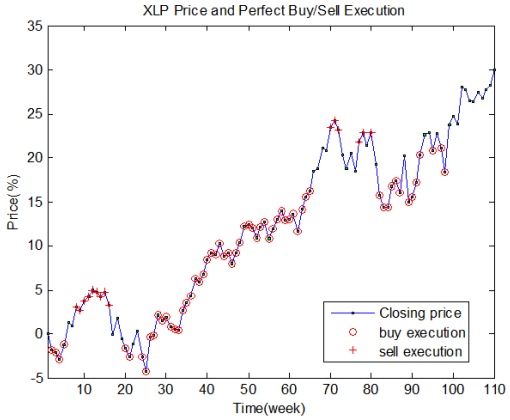

Fig. 7. Perfect Buys and Sells on each week of the period. (01/04/2010 to 02/13/2012)

The confusion matrices and the statistical measures for the Buy and Sell forecasting are summarized in Tab.VII, Tab.VIII, and Tab.IX.

TABLE VII. BUY CONFUSION MATRIX FOR NN PREDICTION ON TESTING DATASET (01/04/2010 TO 02/03/2012)

		True	
		Buy	Not Buy
Predicted	Buy	23 (26.1%)	12 (13.6%)
	Not Buy	34 (38.6%)	19 (21.6%)

Finally, shown in Fig.9 are the predicted profits compared against true Buy/Sell events. Here, the MSE is 11.6% and correlation coefficient is 0.11 between the true and predicted events.

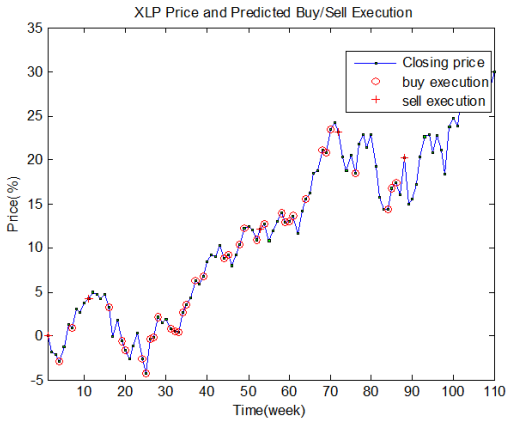

Fig. 8. Buys and Sells predicted by the trained Neural Network (01/04/2010 to 02/03/2012)

TABLE VIII. SELL CONFUSION MATRIX FOR NN PREDICTION ON TESTING DATASET (01/04/2010 TO 02/03/2012)

		True	
		Sell	Not Sell
Predicted	Sell	1 (1.1%)	4 (4.5%)
	Not Sell	11 (12.5%)	72 (81.8%)

TABLE IX. PERFORMANCE MEASURES FOR FORECASTED BUYS AND SELLS (01/04/2010 TO 02/03/2012)

Operation	Precision(%)	Recall(%)	Accuracy(%)
Buy	65.7	40.4	47.7
Sell	20.0	8.3	83.0

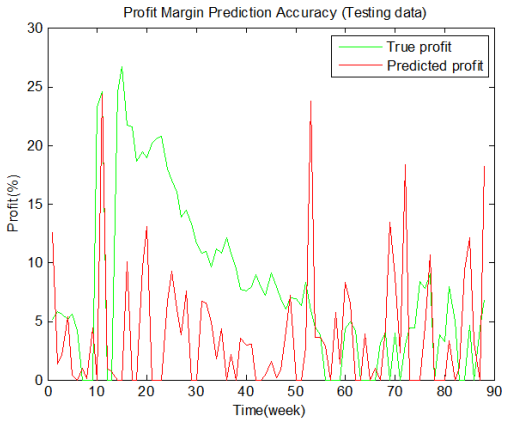

Fig. 9. Profits from the perfect Buys and Sells and profits predicted by the trained Neural Network (01/04/2010 to 02/03/2012) (MSE=11.6%, Corr.Coef.=0.11).

IV. CONCLUSION AND FUTURE WORK

Being profitable on the stock market is a challenging task. The future price forecasting based on the historical data could be considered as a method mainly falling into the technical analysis domain, where the evaluation of the securities is derived from the statistics generated by market activities. The neural network can be used to approximate the inherent non-linear function relating the historical data with future

performance. A new framework based on neural networks is proposed here for investment in the stock market. The proposed framework aims at predicting the Buy/Sell strategy with optimal profit constraints. Our initial investigation has shown the promising outputs. The further continuation of this research work includes experimenting different neural network architectures and topologies to optimize the forecasting performance; and combining the fundamental analysis into the forecasting by using additional information such as local and global economic conditions and events, earning reports, etc.

REFERENCES

[1] R. Wang. "Stock Selection Based on Data Clustering Method". In *2011 7th International Conference on Computational Intelligence and Security*, 2011.

[2] C. Hargreaves and Y. Hao. "Does the Use of Technical and Fundamental Analysis Improve Stock Choice? : A Data Mining Approach applied to the Australian Stock Market". In *Statistics in Science, Business and Engineering (ICSSBE), 2012 International Conference on*, 2012.

[3] F. Zhai, Q. Wen, and Z. Yang etc. "Hybrid Forecasting Model Research on Stock Data Mining". In *New Trends in Information Science and Service Science (NISS), 2010 4th International Conference on*, 2010.

[4] A. Omidi, E. Nourani, and M. Jalili. "Forecasting Stock Prices using Financial Data Mining and Neural Network". In *Computer Research and Development (ICCRD), 2011 3rd International Conference on*, 2011.

[5] A. Galib, M. Alam, and N. Hossain etc. "Stock Trading Rule Discovery Based on Temporal Data Mining". In *Electrical and Computer Engineering (ICECE), 2010 6th International Conference on*, December 2010.

[6] C. Fonseka and L. Liyanage. "A Data Mining Algorithm to Analyse Stock Market Data using Lagged Correlation". In *Information and Automation for Sustainability (ICIAFS), 2008 4th International Conference on*, 2008.

[7] J. Wu, A. Denton, O. Elariss, and D. Xu. "Mining for Core Patterns in Stock Market Data". In *Data Mining Workshops (ICDMW 09), 2009 IEEE International Conference on*, 2009.

[8] Web Source. "Yahoo Finance". http://finance.yahoo.com/. [Online; accessed 19-December-2013].