International Journal of Hybrid Information Technology
Vol.9, No.2 (2016), pp. 57-66
http://dx.doi.org/10.14257/ijhit.2016.9.2.05

Stock Price Prediction Based on Financial Statements Using SVM

Junyoung Heo and Jin Yong Yang

Department of Computer Engineering, Hansung University
jyheo@hansung.ac.kr
Department of Computer Engineering, Hansung University
jyang0112@gmail.com

Abstract

The support vector machine (SVM) is a fast, and reliable machine learning mechanism. In this paper, we evaluated the stock price predictability of SVM, which is a kind of fundamental analysis that predicts the stock price from corporate intrinsic value. Corporate financial statements are used as input into SVM. Based on the results, we predicted the rise or drop of the stock. In addition, we evaluated how long a given financial statement can be used to predict a stock's price. Compared to the experts forecast, the results of SVM show good predictability. However, as times goes on, the predictability begins to drop. These predictions based on financial statements are excellent, but after a short period, the dissonance between financial statements and stock price can be offset by reasonable investors. These results support the efficient market hypothesis.

Keywords*: SVM, machine learning, stock price prediction, financial statements*

1. Introduction

Many previous studies have been conducted regarding how to predict stock prices in the stock market, and there are innumerable disputes on the reliability of those predictions. As efficient predictions of stock prices can provide a minimal guideline for investors who want to invest in a company or stock index, studies on price predictability will contribute to price discovery functions in the market. Stock price fluctuation seems to be completely random from a long-term perspective, but on the other hand, in the short-term, you can find various patterns in the fluctuations and therefore achieve excessive returns by catching such quickly changing patterns. However, there is no guarantee that studies of stock price prediction in the past will still be effective in the future. That is the reason to continue studies of stock price prediction.

Investors have used fundamental and technical analysis for the prediction of future stock prices. Fundamental analysis is a method to estimate intrinsic value of a stock by analyzing various internal and external variables of a company. It predicts the flow of a stock price based on the belief that the stock price reflects the intrinsic value of the stock. On the other hand, technical analysis is a method to find a pattern in the stock price fluctuations based on which we can predict a future stock price. Unlike fundamental analysis, technical analysis does not take the intrinsic value of a stock price into consideration. It predicts future stock price by using past data such as stock price and trading volumes. A lot of technical analysis methods have been developed so far since the Dow Theory [11] in the early 1990s. As most of them have used charts for analysis, it is also called the *chartist's approach*.

Under the efficient market hypothesis (EMH), however, as the information that may have influence on a stock price is reflected in the stock price accurately and quickly, it is impossible to predict future stock price through fundamental analysis

ISSN: 1738-9968 IJHIT

or technical analysis. [18] In a market where the EMH works, the basic information that determines the intrinsic value of a stock is already reflected in the current stock price, and therefore you cannot generate excessive returns through fundamental analysis. Besides, as the fluctuation of stock price is almost completely random, it is impossible to detect a certain pattern in order to generate excessive returns. [7] For empirical analysis such as EMH, event studies, and other studies of price predictability of returns have been performed. [2, 24] While the former tests to see if the new information is well reflected in the market price, the latter tests to see if the past information is well reflected in the market price.

Event study, which tests for an efficient market through the reaction to market price based on new information, was started by Fama, Fisher, Jensen, and Roll. [9, 16, 22] The results of these studies support the efficient market hypothesis, but other researchers have presented conflicting results. While Studies by Ball and Brown support the EMH, studies by Rendleman, Ibbotson, and others oppose the EMH [21].

Then, how about the study results of price predictability that verify the reaction of market price to the past information? If the stock prices evolve according to a random walk, past stock prices would not be able to provide any information on the current stock price. In that case, does pattern analysis of past stock prices give no help to predicting future stock price? The answer was “yes” in many of the early study results. Earlier studies concluded that technical analysis using past data could not predict future stock price and would not bring excessive returns accordingly. However, studies that refute such conclusions have been conducted continuously. That is, there is a certain correlation between the past and future returns of an investment, and you can partially predict future returns of an investment based on the past returns of that investment.

Studies of returns on short-term investments have been conducted mainly by verifying if there is any independence on time series in the stock price fluctuation. Studies conducted in the 1960s-1970s including those by Fama showed that there was no significance in stock price fluctuation due to weak autocorrelation. [9] This indicated that prediction of future stock prices was impossible from the analysis of past stock prices. However, later studies showed that the prediction of future returns on an investment was possible, contrary to the previous studies. The representative examples are studies by Lo, by Mackinlay, and by Jegadeesh & Titman [13, 16, 17, 22]. Many studies on the predictability of returns of a long-term investment showed that returns on stocks have negative (-) autocorrelation. The example studies are those by DeBondt and Thaler [3, 16, 22].

Many studies regarding the prediction of stock price have been made with the machine learning mechanism, based on statistics, owing to the brilliant development of computer technology. Machine learning mechanism can perform complicated non-linear analysis in multi-dimensional spaces, which enables various approaches compared with other studies using statistics only. The representative machine learning mechanisms include artificial neural network, genetic algorithm, fuzzy theory, SVM, decision tree, and adaptive boosting (AdaBoost) [10, 13, 19, 10, 25].

This study forecasts stock price fluctuation of a company through fundamental analysis with Support Vector Machine (SVM) mechanism. Inputting financial information of a company with SVM, this study conducts fundamental analysis and forecasts the rise and drop of a stock price based on the analysis result. SVM was selected for this study among the various machine learning mechanisms, because artificial neural network or AdaBoost, which are the representative machine learning mechanisms, have a shortcoming of requiring long times to learn while relatively little improvement in predictability when compared to SVM. In fact, we conducted experiments using artificial neural network, decision tree, and AdaBoost

prior to this study, and they all took longer time and produced less outstanding outcomes. This study utilized information on assets and profits as the financial information for fundamental analysis. In the financial statements, information on assets and profits are the main index to explain the financial status of a corporation [6].

This study evaluated if we can predict stock price fluctuation based on the index with SVM, and how long it can be predicted for, if it is possible. The financial information used in this study was relative valuation indices, which include earning per share (EPS), book value per share (BPS), and net profit growth rate (NPGR) [12, 23]. Actually stock prices are analyzed and predicted not only for academic research purposes, but also done by analysts of security firms. Predictions by analysts are regarded as reliable in the stock market, as their analyses are based on proven analysis methods, as well as their own prediction skills. Therefore, it is required to compare the predictions of experts with other technical prediction methods. This study compared the predictions by analysts and the prediction methods suggested in this thesis and showed that our method is more outstanding than the prediction of experts.

The structure of this thesis is as follows. Chapter 2 explains the issues to be dealt with in this paper. Chapter 3 describes the data used in this thesis. Chapter 4 explains SVM, a machine learning mechanism. Chapter 5 explains the experiment methods and results using SVM. Finally, Chapter 6 contains the conclusion.

2. Description of the Issues

Based on the financial statements of a company reported quarterly, rise and drop of stock price is predicted through a machine learning mechanism. It is conducted in order to figure out how much the quarterly report of a company has influence on the stock price and how the influence will change as time goes by.

Inputting EPS, BPS, and net profit during the term, which are closely related to the stock price of a company, we can predict the rise or drop of the stock price one and two months later. It is known that EPS and BPS are generally significant in predicting a stock price [12, 15]. In order to compare the functions of prediction, this study evaluated the predictability of machine learning utilizing financial information compared to the experts' score for investment intention. Then, through the prediction of rise or drop of a stock price one and two months later, it evaluates the difference in predictability by period according to the financial statements.

While EPS and BPS are relative values depending on the size of a company, net profit during the term shows big variance according to the size of a company. Therefore, we do not use net profit during the term as input data, but use the ratio that compared the net profit during the term with that of the immediate previous term. That is, the increased or decreased ratio of the net profit during the term is used.

3. Description on Data

The financial data used in this study are those collected from the 1st quarter of 2010 to the 3rd quarter of 2013 from the 200 companies listed in the KOSPI 200 as of 2013. From the financial data, we collected EPS, BPS, net profit, as well as the scores of investment intention (scaled from 5 to 1 points) reported by security firms. Then, we collected the stock price of one and two months later after the end of each quarter. All of the data was retrieved from DataGuide, which was provided by FnGuide.

Although 3,000 data points should have been collected in total for the financial information of the 200 companies for 15 quarters, some were missing and

subsequently a total of 2,913 data points were collected for the study. Table 1 shows the sample information of Samsung Electronics.

Table 1. Financial Information and Stock Price of Samsung Electronics

QUARTER	2010-03-31	2010-06-30	2010-09-30	2010-12-31	2011-03-31
NET PROFIT	3,167,036,000	3,155,614,000	3,550,147,000	3,363,664,000	2,197,639,000
EPS	21,370.00	21,125.00	23,755.00	22,549.00	14,381.00
BPS	445,444.69	460,047.68	478,098.93	501,312.06	509,667.70
STOCK PRICE	814,000	792,000	772,000	945,000	926,000
STOCK PRICE AFTER ONE MONTH LATER	825,000	827,000	764,000	1,010,000	900,000
STOCK PRICE AFTER TWO MONTHS LATER	778,000	776,000	836,000	926,000	884,000
INVESTMENT INTENTION (5-1)	4	4	4	4	4

As EPS and BPS are calculated based on stock price regardless of the size of a company, there is no problem for them to be used in machine learning mechanism. However, net profit during the term shows big variance depending on the size of a company. Therefore, we used net profit growth rate (NPGR) instead of net profit. To obtain NPGR, net profit during the term was collected quarterly and the increase or decrease in ratio was calculated for each quarter.

NPGR(%) = [(Net profit during the term – Net profit of immediate previous quarter) / Net profit of immediate previous quarter] x 100

The stock price after one month and two months later from the quarter was calculated as +1 or -1 depending on the increase or decrease, respectively. For example, if the stock price one month later from the 1st quarter of 2010 (end of April) rises more than the 1st quarter (end of March), it is calculated as +1. If the stock price after two months later from the 1st quarter (end of May) drops when compared with the stock price of the 1st quarter, it is calculated as -1. As for the investment intention scores, point 5 and point 4 were part of the control group to compare the prediction by machine learning mechanism and set as rise (+1). Others (Points 3, 2, 1) were set as drop (-1). Table 2 shows the post processing results of Samsung Electronics in Table 1.

Table 2. Post Processing Results of Samsung Electronics Data

QUARTER	2010-03-31	2010-06-30	2010-09-30	2010-12-31	2011-03-31
EPS	21,370.00	21,125.00	23,755.00	22,549.00	14,381.00
BPS	445,444.69	460,047.68	478,098.93	501,312.06	509,667.70
NPGR	3.71	-0.36	12.50	-5.25	-34.67
STOCK PRICE AFTER ONE MONTH LATER	+1	+1	-1	+1	-1
STOCK PRICE AFTER TWO	-1	-1	+1	-1	-1

MONTHS LATER INVESTMENT INTENTION (5-1)	+1	+1	+1	+1	+1

In sum, we collected financial data and stock prices of 200 companies as we did in Table 3 and used them for learning and testing with SVM.

Table 3. Learning/Test Data

EPS	Earning Per Share
BPS	Book-value Per Share
NPGR	Net Profit Growth Rate
Stock price after one month later (Target_1)	If the stock price after one month rises, it is set as+1. If it drops, it is set as -1
Stock price after two months later (Target_2)	If the stock price after two months rises, it is set as+1. If it drops, it is set as -1
Investment intention	If the investment intention score is 4 or 5, it is set as +1. Others are set as-1.

4. SVM

The SVM (Support Vector Machine) is a statistics-based machine learning mechanism and it is frequently used in pattern recognition. SVM finds a super-plane that minimizes learning data and error rate in a super-space. SVM is summarized as follows.

When a learning set of which relationship between data and label is set as $(\mathbf{x}_i, y_i)$ is given, SVM can be optimized as follows. Here, $i = 1, \ldots, l$, $\mathbf{x}_i \in R^n$ (that is n-dimensional vector), and $y_i \in \{1, -1\}$. (In this paper, $\mathbf{x}_i = (EPS, BPS, NPGR)$, $y_i = 1 \ or \ -1$). $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{z}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0, \ i = 1, \ldots, l$. It is an equation to get a super-plane $\mathbf{w}$ that minimizes Expression (1). [4, 8]

$$\min_{\mathbf{w},b,\xi} \left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + c\sum_{i=1}^{l}\xi_i\right) \quad (1)$$

Vector $\mathbf{z}_i$ is mapping the learning vector $\mathbf{x}_i$ by using function φ to make it of a higher dimension. That is, $\mathbf{z}_i = \varphi(\mathbf{x}_i)$. In Expression (1), $c > 0$ and c is a penalty parameter on the error term. That is, it is a constant that determines how large the error is. Using Lagrange Multiplier, it can be changed to Expression (2). Herein, $\alpha_i, \beta_i \geq 0$.

$$\min_{\mathbf{w},b,\xi} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + c\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i[y_i(\mathbf{w}^{\mathrm{T}}\mathbf{z}_i + b) - 1 + \xi_i] - \sum_{i=1}^{l}\beta_i\xi_i\right\} \quad (2)$$

Expression (2) can be indicated in the following dual form (Expression (3)).

$$\mathrm{L}(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \quad (3)$$

That is, obtaining $\boldsymbol{\alpha}$ that maximizes $\mathrm{L}(\alpha)$, $\mathbf{w} = \sum_{i=1}^{l}\alpha_i y_i \mathbf{z}_i$. Herein, $0 \leq \alpha_i \leq \mathrm{c}$ and $\sum_{i=1}^{l}\alpha_i y_i = 0$. $k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function. Various kernel functions can be used and this study used radial basis function (RBF) that can provide non-linear classification. RBF kernel function is defined as in Expression (4).

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (4)$$

The SVM is provided in the form of open library format and various commercial calculation programs. Among them, this study used the open library LibSVM, which is

frequently used. [5]

5. Experiments

After mixing 2913 collected samples randomly, we divided them into two for learning and for testing in conduction of the experiment. Then, the two groups were replaced each other and learning and testing were conducted again. This process was repeated 10 times to get the average predictability. By mixing data randomly and repeating mutual verification, this study tried to offset the coincidental SVM prediction results that might be shown incidentally by the learning and testing data [20].

In order to compare which combination had superior predictability among EPS, BPS, and NPGR in the financial data of each quarter, we tested all possible combinations. That is, we made prediction models in the combination of {EPS}, {BPS}, {NPGR}, {EPS, BPS}, {EPS, NPGR}, {BPS, NPGR}, and {EPS, BPS, NPGR}, respectively.

For the non-linear classification of SVM, kernel function was used as radial basis function (RBF). For embodying SVM, this study used a machine learning py (MLPY) that was made for using LibSVM in Python [1].

Figure 1 is a part of a prediction function that uses MLPY among the Python codes used in the experiment. The ml_inputs are an input data that is delivered as a factor to the prediction() function. The ml_targets are a classification data which means increase (+1) or /decrease (-1). These data are randomly mixed by using a random shuffle() function prior to delivering the data to the prediction() function. The prediction() function performs learning with svm.learn() method after initializing LibSVM and performs prediction (testing) with svm.pred() method. The number of predicted successes is saved in the Match. After exchanging the data used for learning and prediction with each other, learning and prediction is repeated again. The number of predicted successes is saved in Match 2. We produced the final predictability by repeating the prediction() functions 10 times in the experiment and getting the average value.

```
def prediction(ml_inputs, ml_targets, ml_experts):
total_count = len(ml_inputs)

if total_count != len(ml_targets):
print 'Fatal Error!'
exit(1)

# libSVM initialization
svm = mlpy.LibSvm(svm_type='c_svc', kernel_type='rbf')

# the front half of input for training, and the last of input for testing
svm.learn(ml_inputs[0:total_count/2,:], ml_targets[0:total_count/2])  # training
preds = svm.pred(ml_inputs[total_count/2:,:])                  # testing
pred_cmp = ml_targets[total_count/2:] - preds
match = (pred_cmp == 0).sum()

# the rear half of input for training, and the last of input for testing
svm.learn(ml_inputs[total_count/2:,:], ml_targets[total_count/2:])    # training
preds = svm.pred(ml_inputs[0:total_count/2,:])                 # testing
pred_cmp = ml_targets[0:total_count/2] - preds
match2 = (pred_cmp == 0).sum()
```

Figure 1. Python Code for Testing using MLPY

Table 4 represents the experiment outcome. We can see that the predictability of stock price one month later after reporting financial information (SVM(1M)) is higher than that of the stock price two months later (SVM(M2)), no matter what value is inputted. That is, as time goes by, the predictability of stock price based on financial information decreases.

Table 4. Predictability Results (%)

	EPS,BPS,NPGR	EPS,NPGR	EPS,BPS	BPS,NPGR	EPS	BPS	NPGR
SVM(1M)	57.1	54.6	57.5	56.1	53.7	55.2	56.7
SVM(2M)	50.1	50.6	50.4	49	49	47.7	52.9
EXPERT	45.6	45.6	45.6	45.6	45.6	45.6	45.6

Figure 2 is a graphic form of Table 4. We can easily see that prediction one month later is superior to the prediction of two months later and is higher than the experts' score. However, as the experts' score is only the average values of the experts' scores reported in security firms, it is hardly confirmed that prediction with SVM is more outstanding than prediction by experts from the research results.

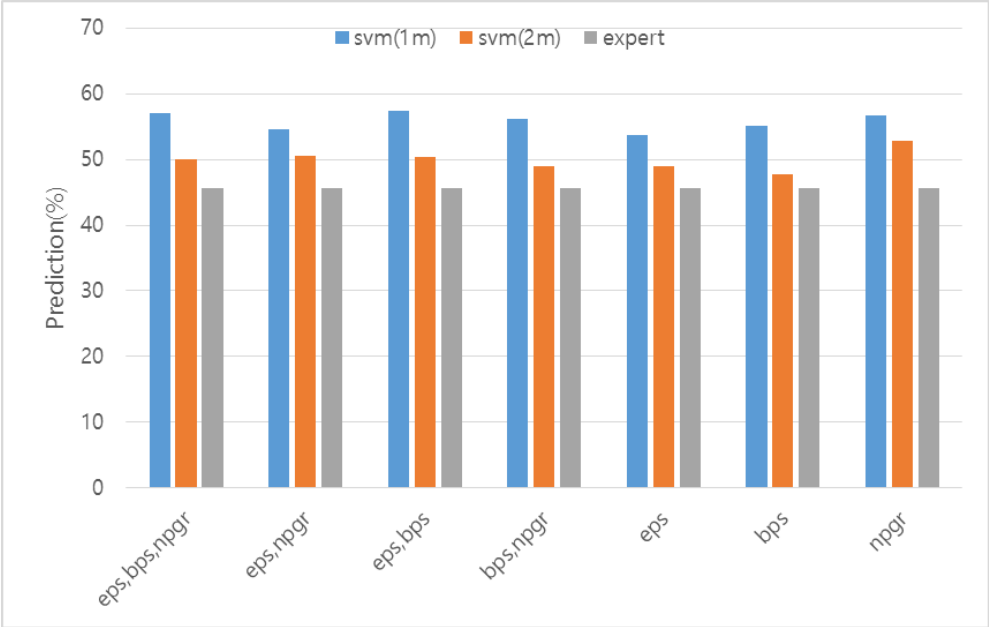

Figure 2. Prediction Outcome (%)

In order to compare predictability easily according to input data type, we selected the results of SVM (2M) in Table 4 and arranged it in the order of superior predictability and made a graph as shown in Figure 3. Although {NPGR} had the best predictability as a single variable, it was found that combination of more than 2 variables helped to improve predictability. As for the data used in this study, combination of {EPS, BPS} showed the best performance. It is superior to the combination of {EPS, BPS, NPGR} that contained NPGR. This means that the greater the number of input datum does not always produce better results.

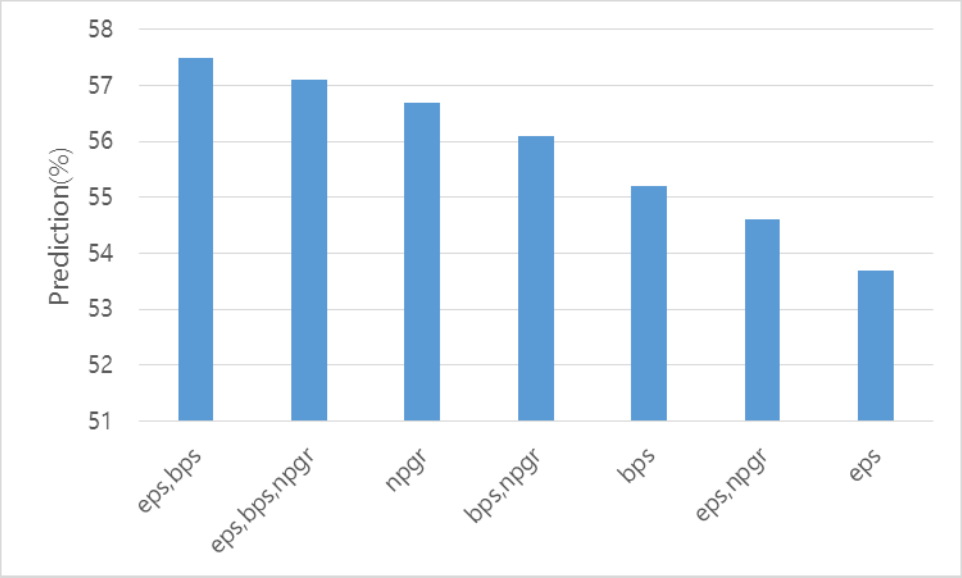

Figure 3. Prediction Result on SVM (1M) (%)

6. Conclusion

This study verified prediction of stock price fluctuation of a company through fundamental analysis with an experimental method using SVM. It performed fundamental analysis based on financial information input with SVM and predicted future fluctuation of stock price based on the experiment results. Information on assets and profits were used as financial information for fundamental analysis. They are the indices to explain the financial status of a company. Based on the indices, this study evaluated if the stock price fluctuation can be predicted and how long the prediction would work, by using SVM.

As a result, it was found that stock price predictability utilizing financial information input with SVM showed superior predictability to expert's predictions, and that predictability decreases as time goes by.

Prediction based on financial information is outstanding in the short-run, but after a certain time passes the mismatch between financial information and stock price is offset by reasonable investors. In other words, the stock market corresponds to efficient market hypothesis in the long run.

References

[1] D. Albanese, R. Visintainer, S. Merler, S. Riccadonna, G. Jurman and C. Furlanello, "Mlpy: Machine Learning Python", **(2012)**.
[2] M. Beechey, D. W. R. Gruen and J. Vickery, "The efficient market hypothesis: a survey", Reserve Bank of Australia, Economic Research Department, **(2000)**.
[3] W. F. Bondt and R. Thaler, "Does the stock market overreact?", The Journal of finance, empirical anomalies, vol. 40, no. 3, **(1985)**, pp. 793-805.
[4] B. E. Boser, I. Guyon and V. Vapnik, "A training algorithm for optimal margin classifier", In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM Press, **(1992)**.
[5] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines", **(2001)**, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[6] C. H. Chung and S. K. Kim, "An Investigation on the Stock Return Predictability of Dividend Yield and Earning-Price Ratio", The Korean Journal of Financial Engineering, vol. 9, no. 3, **(2010)**, pp. 61-87.
[7] J. Conrad and G. Kaul, "Mean reversion in short-horizon expected returns", Review of Financial Studies, vol. 2, no. 2, **(1989)**, pp. 225-240.
[8] C. Cortes and V. Vapnik, "Support-vector network", Machine Learning, **(1995)**, pp. 273-297.
[9] E. F. Fama, "Efficient capital markets: II", The journal of finance, vol. 46, no. 5, **(1991)**, pp. 1575-1617.
[10] E. Hadavandi, H. Shavandi and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting", Knowledge-Based Systems, vol. 23, no. 8, **(2010)**, pp. 800-808.

[11] W. P. Hamilton, "The Stock Market Barometer: A Study of its Forecast Value Based on Charles H. Dow's Theory of the Price Movement", Barrons, New York, **(1922)**.
[12] S. Han and R.-C. Chen, "Using SVM with Financial Statement Analysis for Prediction of Stocks", Communications of the IIMA, vol. 7, no. 4, **(2007)**, pp. 63-72.
[13] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency", The Journal of Finance, vol. 48, no. 1, **(1993)**, pp. 65-91.
[14] K. Kim and I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", Expert systems with applications, vol. 19, no. 2, **(2000)**, pp. 125-132.
[15] K. Y. Kim and Y. B. Kim, "Testing the Predictability of Stock Return in the Korean Stock Market", Korean Journal of Industrial Economic, vol. 17, no. 4, **(2004)**, pp. 1255-1271.
[16] K.-P. Lim and R. Brooks, "The evolution of stock market efficiency over time: a survey of the empirical literature", Journal of Economic Surveys, vol. 25, no. 1, **(2011)**, pp. 69-108.
[17] A. W. Lo and A. C. MacKinlay, "Stock market prices do not follow random walks: Evidence from a simple specification test", Review of financial studies, vol. 1.1, **(1988)**, pp. 41-66.
[18] B. G. Malkiel, "Efficient market hypothesis", The new palgrave: A dictionary of economics, vol. 2, **(1987)**, pp. 120-23.
[19] P. F. Pai and C. S. Lin, "A hybrid ARIMA and support vector machines model in stock price forecasting", Omega, vol. 33, no. 6, **(2005)**, pp. 497-505.
[20] A. S. Pandya and R. B. Macy, "Pattern Recognition with Neural Networks in C++", IEEE Press, **(1995)**.
[21] R. J. Rendleman Jr., C. P. Jones and H. A. Latane, "Empirical anomalies based on unexpected earnings and the importance of risk adjustments", Journal of Financial Economics, vol. 10, no. 3, **(1982)**, pp. 269-287.
[22] M. Sewell, "History of the efficient market hypothesis", RN, vol. 11, no. 04, **(2011)**, pp. 04.
[23] D.-S. Song, "A Study on the Relation between the Financial Ratio and Earnings Quality", Korea International Accounting Review, vol. 40, **(2011)**, pp. 135-156.
[24] A. Timmermann and C. W. J. Granger, "Efficient market hypothesis and forecasting", International Journal of Forecasting, vol. 20, no. 1, **(2004)**, pp. 15-27.
[25] M.-C. Wu, S.-Y. Lin and C.-H. Lin, "An effective application of decision tree to stock trading", Expert Systems with Applications, vol. 31, no. 2, **(2006)**, pp. 270-274.

Authors

Junyoung Heo, 2009, He is a PhD, Department of Computer Engineering, Seoul National University. 2009~present, Assistant Professor, Department of Computer Engineering, Hansung University. Research fields: Operating System, Wireless Sensor Networks, Embedded Systems, Machine Learning, Financial Engineering

Jin Yong Yang, 2013, He is a PhD, Department of Economics, Hanyang University. 2014~present, PhD candidate, Department of Computer Engineering, Hansung University. Research fields: Quantitative Finance, Computational Finance, Econophysics.

