https://doi.org/10.5392/JKCA.2018.18.09.623

익스트림 그라디언트 부스팅을 이용한 지수/주가 이동 방향 예측

Prediction of the Movement Directions of Index and Stock Prices Using Extreme Gradient Boosting

김형도
한양사이버대학교 경영정보학과

HyoungDo Kim(hdkim@hycu.ac.kr)

요약

주가 이동 방향의 정확한 예측이 주식 매매에 관한 전략적 의사결정에 중요한 역할을 할 수 있기 때문에 투자자와 연구자 모두의 관심이 높다. 주가 이동 방향에 관한 기존 연구들을 종합해보면, 주식 시장에 따라서 그리고 예측 기간에 따라서 다양한 변수가 고려되고 있음을 알 수 있다. 이 연구에서는 한국 주식 시장을 대표하는 지수와 주식들을 대상으로 이동 방향 예측 기간에 따라서 어떤 데이터마이닝 기법의 성능이 우수한 것인지를 분석하고자 하였다. 특히, 최근 공개경쟁에서 활발히 사용되며 그 우수성이 입증되고 있는 익스트림 그라디언트 부스팅 기법을 주가 이동 방향 예측 문제에 적용하고자 하였으며, SVM, 랜덤 포리스트, 인공 신경망과 같이 기존 연구에서 우수한 것으로 보고된 데이터마이닝 기법들과 비교하여 분석하였다. 12년간 데이터를 사용하여 1일 후에서 5일 후까지의 이동 방향을 예측하는 실험을 통해서, 예측 기간과 종목에 따라서 선택된 변수들에 차이가 있으며, 1-4일 후 예측에서는 익스트림 그라디언트 부스팅이 다른 기법들과 부분적으로 동등함을 가지면서도 가장 우수함을 확인하였다.

■ 중심어 : | 주가 | 지수 | 이동 방향 | 예측 | 그라디언트 부스팅 | 한국 주식 시장 |

Abstract

Both investors and researchers are attentive to the prediction of stock price movement directions since the accurate prediction plays an important role in strategic decision making on stock trading. According to previous studies, taken together, one can see that different factors are considered depending on stock markets and prediction periods. This paper aims to analyze what data mining techniques show better performance with some representative index and stock price datasets in the Korea stock market. In particular, extreme gradient boosting technique, proving itself to be the fore-runner through recent open competitions, is applied to the prediction problem. Its performance has been analyzed in comparison with other data mining techniques reported good in the prediction of stock price movement directions such as random forests, support vector machines, and artificial neural networks. Through experiments with the index/price datasets of 12 years, it is identified that the gradient boosting technique is the best in predicting the movement directions after 1 to 4 days with a few partial equivalence to the other techniques.

■ keyword : | Stock Price | Index | Movement Direction | Prediction | Gradient Boosting | Korea Stock Market |

접수일자 : 2018년 08월 10일
수정일자 : 2018년 09월 17일

심사완료일 : 2018년 09월 18일
교신저자 : 김형도, e-mail : hdkim@hycu.ac.kr

I. Introduction

With the active industrial applications of artificial intelligence technology, automated stock trading and investment embodied in such as robo-advisors are becoming more attractive, and excellent prediction capability is recognized as a key tool for securing their competitiveness of stock investment. With the potential to generate high benefits and avoid high risks through reasonably accurate predictions, prediction of stock prices has been receiving much attention as one of the important objective of the financial world[1].

Around the efficient market hypothesis[2], however, there have been investors′ beliefs and academic controversies that investors cannot get abnormal returns from the past behavior of stock prices. Despite this debate, it is possible to predict future stock prices at least partially, and some studies show that it is actually feasible by using various data mining algorithms. However, it is still difficult to predict when considering the complicated and dynamic systems that interact with political events, economic conditions, and investors’ expectations[1].

Since accurate stock forecasts are virtually difficult (that is, some errors are inevitable) and investors′ strategies vary widely, direct prediction of stock prices may not be suitable. That is why there have been a number of studies looking at movement directions of financial markets[3]. Traditional forecasting methods such as regression analysis makes it difficult to predict movement directions of stock prices, due to the non-linear complexity, especially that of modeling human behaviors[4]. In order to improve this problem, diverse data mining methods have been developed, showing better forecasting performance[1][3-7].

By compiling the existing studies on stock price movements, one can see that different factors are considered depending on the stock market and the forecast period. After all, it suggests that input(predictor) variables should be appropriately chosen depending on them. In the aspect of the techniques of creating forecasting models, there is a lack of consistency in the existing studies about movement directions of stock prices showing that Support Vector Machines(SVM), Artificial Neural Network(ANN), Random Forest(RF), and Decision Tree(DT) models are the best respectively[1][3-5].

This paper aims to analyze what data mining techniques show better performance with some representative index and stock price data of the Korea stock market. Recently, the extreme gradient boosting (XGBoost)[8] has been demonstrating the best performance through many recent open competitions. To the best of our knowledge, there is no research about the application of the technique to the prediction issues of stock prices and their movement directions. This paper intends to compare the prediction techniques for the movement directions of stock prices such as SVM, RF, and ANN with the XGBoost using the Korean stock market data. Prediction periods of 1 to 5 days are experimented with the 12-years data for analyzing its effects on the quality of prediction results. Furthermore, variables are selected before experiments per each stock in each prediction period, compared with the fixed set of variables in the existing studies.

The rest of this paper is organized as follows. Section 2 summarizes related research and section 3 proposes the experimental design for the prediction of movement directions of index and stock prices. Section 4 describes found results from the comparative analysis of the techniques. Finally, section 5 summarizes the paper with concluding remarks.

Ⅱ. Related Research

There are three major approaches to the prediction of stock price behavior[1]: technical analysis with charts and plots, time series forecasting, and data mining. The popularity of the last approach has been growing since the data sets are too big to handle with methods of the other two approaches.

Several algorithms for binary classification have been used in the literature for predicting movement directions of index and stock prices. Simple algorithms such as single decision tree, discriminant analysis, naive Bayes were used in the early days of data mining. As better data-mining algorithms such as SVM, RF, and ANN became popular, they have been also employed for the prediction of index and stock price movement directions.

SVM is a machine learning algorithm that builds a model mapping each case as a single point in a multi-dimensional space. It constructs a hyperplane that has the largest distance to the nearest training case of any class. That is, the hyperplane divides the training examples by a clear gap that is as wide as possible. The model then maps new cases into the same space and predicts which side of the hyperplane they fall. When the dataset to classify is not linearly separable in the original multi-dimensional space, implicit non-linear mapping of the space into much higher dimensional space can be employed for making the separation easier. It is called as kernel-trick[9]. Established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error, SVM is shown to be very resistant to the over-fitting problem, eventually achieving a high generalization performance with always unique and globally optimal solution[5].

ANN is a machine-learning system inspired by biological neural networks that make up animal brains. A set of artificial neurons are connected together and can be structured hierarchically in an ANN. Each connection serves to transmit signals from one neuron to another. Since the first neural network model was created in 1943 based on mathematics, the research on neural network model had been widely spread as multi-layer perceptrons with 1 or more hidden layers in addition to the input and output layers appeared with back-propagation. In recent years, various researches and applications have been actively conducted on deep neural networks such as convolutional and recurrent neural networks[10], where the number of hidden layers is 2 or more.

A RF is a set of decision trees generated from the bootstrap samples of the training data set, created by taking into account a random selection of attributes in each branching process and with no pruning. It is an ensemble learning method of bagging (bootstrap aggregating) which lets the trees vote for the most popular one in classification and calculates the average of their prediction values in regression. RFs give results competitive with boosting and adaptive bagging methods, and the generalization error converges to a limit as the number of trees becomes large without over-fitting[11]. The reason for the better performance is that the bias caused by the noise included in individual training examples can be reduced as the number of trees increases as far as there is no correlation between trees. RFs try to decrease such correlation between trees by bootstrap sampling of training examples and random selection of features in branching.

In order to predict the 1-day later movement directions of the daily Korea composite stock price index(KOSPI), Kim(2003) used 12 technical indicators of Stochastic %K, Stochastic %D, Stochastic Slow

%D, Momentum, Rate of Change(ROC), Williams' %R(WPR), Accumulation/Distribution Oscillator(ADO), Disparity 5, Disparity 10, Price Oscillator(OSCP), Commodity Channel Index(CCI), and Relative Strength Index(RSI)[13]. Experiments with the dataset from 1989 to 1998 showed that SVM was the best among SVM, backpropagation ANN, and case-based reasoning. Manish and Thenmozhi(2005) compared SVM, random forest, and ANN using the same technical indicators as those of Kim(2003) to predict the 1-day later movement direction of the S&P CNX NIFTY MARKET index[14]. SVM was also reported as the best. Huang et al.(2005) applied SVM to the NIKKEI 225 index, which combines the 225 stocks traded on the Tokyo stock market, for the weekly moving trends from 1990 to 2002, and performed Linear Discriminant Analysis(LDA), Quadratic Discriminant Analysis(QDA), and Elman Backpropagation Network. S&P500 index and USD/JPY a week ago was used as input variables[5]. They reported that SVM was the best in accuracy.

Kara et al.(2011) estimated the National 100 index of Istanbul stock market in Turkey by applying ANN and SVM for the daily movement directions from 1997 to 2007[4]. They used 10 technical indicators of 10-days Simple Moving Average(SMA), 10-days Weighted Moving Average(WMA), Momentum, Stochastic %K, Stochastic %D, RSI, Moving Average Convergence Divergence(MACD), WPR, ADO, and CCI. Experimental results showed that the average performance of ANN model was found significantly better than that of SVM model.

Imandoust and Bolandraftar(2014) applied the techniques of DT, RF, and Naive Bayes(NB) to the daily data from 2007 to 2012 of Iran Tehran stock market to forecast the movement directions of stock prices for the next day[3]. They used the same 10 technical indicators as those used in the study by Kara et al.(2011) as well as fundamental variables such as oil price, gold price, and USD/IPR. They reported that decision trees outperformed the other ones in accuracy.

Ballings et al.(2015) compared ensemble methods such as RF, AdaBoost, and Kernel Factory with single classifiers such as ANN, Logistic Regression(LR), SVM, and k-Nearest Neighbor(k-NN)[1]. They used 14 variables including price book value ratio(PBR), earnings per share(EPS), book-value per share(BPS), Cash-flow Per Share(CPS), Trade Balance, Unemployment, Inflation, Credit Coverage, and Interest Coverage from the annual dataset collected from the stock markets of several European countries. The best in the area under the receiver operating characteristic curve(AUC) was RF when predicting the movement directions of stock prices for the next year.

Manojlovic and Stajduhar(2015) conducted a comparative experiment on predicting the movement directions after 5 and 10 days for the CROBEX index of the Croatian Zagreb stock market and several stocks of various sectors[6]. They selected 12 technical indicators of 5/10-days WMA, Stochastic %K, Stochastic %D, MACD, CCI, 5/10-days Disparity, Percentage Price Oscillator(PPO), 10-days ROC, 10-dyas Momentum, RSI, 5-days Standard Deviation(SD) for testing SVM, ANN, and RF models. RF was reported as the best in accuracy.

Patel et al.[15] compared ANN, SVM, RF, and NB models in predicting the movement directions of two stocks and two indices in Indian stock market. The technical indicators used are the same as those of Kara et al.(2011). The experimental results suggested that RF outperforms other three prediction models on overall performance with ten technical parameters are represented as continuous values.

Zhong and Enke(2017) used 60 financial and economic factors as potential feature variables for the daily S&P 500 Index ETF(SPY) from 2003 to 2013[7]. These include the SPY return for the current day and three previous days, the relative difference in percentage of the SPY return, exponential moving averages of the SPY return, Treasury bill(T-bill) rates, certificate of deposit rates, financial and economic indicators, the term and default spreads, exchange rates between the USD and four other currencies, the return of seven world major indices(other than the S&P 500), SPY trading volume, and the return of eight large capitalization companies within the S&P 500. Applying three dimensional reduction techniques such as Principal Component Analysis(PCA) to these potential variables, they used the most important and influential factors as inputs to ANN. Experimental results showed that combining ANN with PCA gives slightly higher classification accuracy than the other two combinations.

Gradient boosting is a term originally used by Friedman(2001), in which a new model is generated that predicts errors or residuals in the existing models[12]. Extreme gradient boosting(XGBoost) is a unique implementation of this kind of gradient boosting. It is very fast and evaluated as very good in performance, proved by the winning solutions of some public competitions in a variety of areas such as sales forecasting, web sentence classification, customer behavior prediction, risk analysis, and so on[8]. It can be a good candidate for the prediction of the movement directions of index and stock prices.

III. Experimental Design for Prediction

In this study, XGBoost will be compared with the existing methods in predicting movement directions of stock prices. As comparison targets, SVM, ANN, and RF, reported to be superior in the previous studies, were adopted. As potential input (predictor) variables, a number of technical indicators suitable for daily forecasting were selected: 5/10/20-days SMA, 5/10/20-days WMA, Stochastic K%, Stochastic D%, MACD, MACD Signal, WPR, Accumulation / Distribution Line(ADL), CCI, 5/10-days Disparity, PPO, 5/10/20-days Momentum, 5/10/20-days ROC, RSI, and 5/10/20-days Psychology Line(PSY). That is, 30 input variables are used in total. The definitions of these variables are shown in the following [Table 1]. Here, L_t, H_t, C_t, and V_t represent the Low Price, High Price, Closing Price, and Volume at day t, respectively. HH_t^{t-n+1} and LL_t^{t-n+1} represent a Highest High Price and a LowestLowPrice from t-n+1 to t, respectively. EMA_t^x represents the value of the x-day geometric moving average obtained at time t.

Table 1. Definitions of Input Variables

Variables	Definition
SMA	$\sum_{i=1}^{n} C_{t-i+1}/n$
WMA	$\sum_{i=1}^{n} C_{t-i+1} * (n-i+1)/\sum_{i=n}^{1} i$
Stochastic %K	$\frac{C_t - LL_t^{t-n+1}}{HH_t^{t-n+1} - LL_t^{t-n+1}} \times 100$
Stochastic %D	$\mathrm{SMA}_t^n(\%\mathrm{K})$
MACD	$EMA_t^{12}(C) - EMA_t^{26}(C)$
MACD Signal	$EMA_t^{9}(MACD)$
WPR	$\frac{HH_t^{t-n+1} - C_t}{HH_t^{t-n+1} - LL_t^{t-n+1}} \times 100$
ADL	$\sum_{i=1}^{t} ((2 \times C_i - L_i - H_i)/(H_i - L_i) \times V_i)$
CCI	$\frac{TP_t - SMA_t^{20}(TP)}{0.015 \times MD_t(TP)}$ $TP_t = \frac{C_t + H_t + L_t}{3}$ $MD_t(TP) = \frac{\sum_{i=1}^{20} \lvert TP_{t-i+1} - SMA_t^{20}(TP) \rvert}{20}$
Disparity	$\frac{C_t}{SMA_t^n(C)} \times 100$

PPO	$\frac{EMA_t^{12}(C) - EMA_t^{26}(C)}{EMA_t^{26}(C)} \times 100$
Momentum	$C_t - C_{t-n}$
ROC	$\frac{C_t - C_{t-n}}{C_{t-n}} \times 100$
SD	$\sqrt{\frac{\sum_{i=1}^{n}(C_{t-i+1} - SMA_t^n(C))^2}{n}}$
RSI	$100 - 100/(1 + \mathrm{Ups}/\mathrm{Downs})$ $\mathrm{Ups} = \sum_{i=1}^{n} \Delta_{t-i+1}\ for\ i\ s.t.\ \Delta_{t-i+1} > 0$ $\mathrm{Downs} = \sum_{i=1}^{n} (-\Delta_{t-i+1})\ for\ i\ s.t.\ \Delta_{t-i+1} <$ $\Delta_{t-i+1} = C_{t-i+1} - C_{t-i}$
PSY	$\frac{UpsNum}{n} \times 100$ $UpsNum = \sum_{i=1}^{n}(C_{t-i+1} - C_{t-i} > 0)$

It is common to carry out all experiments with predetermined input variables in the previous studies. On the other hand, meaningful variables to each experiment are selected among the input variables in this study. An advantage of ensemble techniques such as XGBoost is that it provides the importance of input variables automatically from the predictive model construction process. In this study, input variables are selected additionally by the orders of their importance values provided by XGBoost. Each set of input variables is used for model building to get its performance. Finally the set of input variables in the best performance model is determined to be used in the following experiments.

Another thing to note is about the prediction period in the daily prediction of the movement directions of index and stock prices. Most previous studies have been predicting the movement directions of stock prices after one day later, but it is not uncommon to predict those after 5 days(1 week), 2 weeks, 1 month, 3 months, and even 1 year. In this study, all the experiments on prediction techniques are designed to examine the effect of prediction periods of 1 to 5 days. The movement direction is represented as 1 if the stock price after n days is higher than the current stock price, otherwise 0. As a result, the prediction tasks are treated as a binary classification problem.

There are various measures for evaluating the performance of classifiers such as Accuracy, TPR, TNR, G-mean, Precision, Recall, F-measure and so on. Based on the report that AUC can appropriately and accurately measure the performance of binary classification[16], AUC, which is the area under the ROC(Receiver Operating Characteristic) curve, is adopted as the performance evaluation measure. As the AUC value of a classification system approaches 1, the better it can be evaluated.

A T-tests is applied to the 10 AUC values obtained from the 10-fold cross validation for each combination of techniques, stock, and prediction periods. The composite AUC and accuracy values, obtained from the total predictions combining all the partial predictions in the 10-fold cross validation, are also referred to as a comparison measure.

IV. Experimental Results and Analysis

All the experiments in this paper were conducted using the R environment on a 64-bit Windows PC with an i5-4460 processor and 32G RAM. RF was implemented using the randomForest package, XGBoost the xgboost package, SVM the kernlab package, and ANN the keras/tensorflow package.

1. Datasets

The datasets to be analyzed are about KOSPI, the representative stock market indx of the Korea Stock Exchange, and 3 representative stocks: Samsung Electronics(SSE), Hyundai Motor Company(HMC), and SK Telecom(SKT), which can satisfy the efficient market hypothesis to some extent. They are the daily data(low price, high price, close price, and volume of transactions) of the index/stocks for 12 years(2005~2016), extracted from the Naver finance homepage[17]. Although the total number of trading days for each stock is 2976, only 2923 days are used for training and prediction since 33 days for the first part cannot be used due to the moving average and the data of 20 days later due to future prediction. [Table 2] summarizes the numbers of increasing/decreasing cases of the index/stocks by the prediction periods. The number of increasing(up) cases in each combination is not so different from the number of decreasing(down) cases that no special work is necessary for the class balance in binary classification.

Table 2. The Numbers of Increasing/Decreasing Cases of Index/Stocks by Prediction Periods

Pred. Period	KOSPI	SSE	HMC	SKT
1	1557/1366	1422/1501	1383/1540	1306/1617
2	1597/1326	1470/1453	1428/1495	1354/1569
3	1605/1318	1468/1455	1389/1534	1396/1527
4	1627/1296	1504/1419	1422/1501	1436/1487
5	1657/1266	1520/1403	1436/1487	1442/1481

2. Variable Selection

For the selection of variables, an XGBoost model was constructed by using 80% of the total data as training data and the remaining 20% as validation data to predict the movement direction of stock prices in each prediction period. All the XGBoost parameters are set with default values and the number of decision trees is optimized by stopping additional decision tree generation if the AUC of the validation data does not improve over 100 times.

From the model, the importance of each variable(the gain value of the variable in generating decision trees) is obtained. With an initial set of some best variables, additional variables are added one by one following the importance order of the remainder variables. Each set of the variables created are used for building a model to get their performance. The best performance model is then determined to use the variables in the following experiments. [Table 3] summarizes the selected variables for each prediction period for KOSPI. The size of the best set of variable diminishes as the prediction period increases.

Table 3. Variable Selection by Prediction Periods for KOSPI

Pred. Period	Selected Variables
1	fastD fastK disparity5 CCI PSY10 disparity10 sd5 PSY5 ADL disparity20 MACD ROC20 ROC5 RSI momentum5 sd20 sd10 sma5 PSY20 ROC10 MACDsignal momentum10 sma20 momentum20 PPO sma10 wma20 wma5 wma10 WPR
2	ADL sd5 disparity5 MACDsignal fastD sd20 PSY5 PSY10 MACD ROC10 sma5 RSI CCI fastK sd10 ROC5 PSY20 disparity10 disparity20 ROC20 sma20
3	ADL sd20 disparity5 PSY5 MACDsignal fastD sd5 ROC20 CCI ROC10 MACD sd10 sma5 PSY10 fastK sma20
4	ADL sd20 MACDsignal sd10 fastD MACD PSY5 ROC20 sd5 sma20
5	ADL sd20 MACDsignal sma20 sd10 MACD PSY10

3. Performance Analysis

In order to compare the performance of the models constructed using the techniques, 10-fold cross-validation(10CV) is performed on each index/stock using the selected variables for each prediction period. [Table 4] shows the results of the four models for predicting the movement directions of KOSPI index after 1 to 5 days, respectively. This table presents the mean and standard deviation of the 10CV AUCs and the composite AUC and accuracy obtained by combining all the results of the 10CV.

Table 4. Experimental Results for KOSPI

After n days	Method	Mean AUC	SD AUC	Merged AUC	Merged Accuracy
1	XGBoost	0.54129	0.03131	0.53923	0.53815
	RF	0.51585	0.03648	0.51570	0.51728
	SVM	0.51453	0.02288	0.51209	0.50600
	NN	0.53544	0.03667	0.52894	0.53472
2	XGBoost	0.71357	0.02381	0.70803	0.66028
	RF	0.71640	0.02211	0.71524	0.67362
	SVM	0.65772	0.02636	0.65700	0.62367
	NN	0.56827	0.03276	0.56248	0.56141
3	XGBoost	0.77567	0.03111	0.77209	0.72015
	RF	0.78045	0.02726	0.77966	0.71810
	SVM	0.68829	0.02890	0.68671	0.65036
	NN	0.59043	0.04709	0.58336	0.579200
4	XGBoost	0.82713	0.04080	0.82705	0.75231
	RF	0.83803	0.03506	0.83669	0.75949
	SVM	0.73892	0.02779	0.73716	0.68457
	NN	0.62174	0.04446	0.61585	0.59767
5	XGBoost	0.86004	0.02393	0.85520	0.78857
	RF	0.86691	0.02247	0.86656	0.79747
	SVM	0.77097	0.03089	0.77008	0.70818
	NN	0.63192	0.03619	0.63192	0.60759

[Table 5] summarizes the results of unidirectional t-tests between 10 AUCs obtained from the 10CV for each prediction period of KOSPI. XGBoost can be said to be superior than RF and SVM at 10% of significance in the case of 1-day prediction period. However, it is not better than NN in spite of its AUC is greater than that of NN. In the predictions on the movement directions after 2 to 5 days, XGBoost is not better than RF, but it is better than SVM and NN. At the same time, RF is not better than XGBoost in the significance level. Therefore, XGBoost is the best equally with RF in the prediction periods of 2 to 5 days.

Table 5. p-Values of Right-tailed t-Tests for KOSPI

After n days	T1	T2	p-value	T1	T2	p-value
1	XGBoost	RF	0.05577	RF	SVM	0.46192
	XGBoost	SVM	0.02130	RF	NN	0.12329
	XGBoost	NN	0.35286	SVM	NN	0.07172
2	XGBoost	RF	0.60694	RF	SVM	0.00002
	XGBoost	SVM	0.00005	RF	NN	0.00000
	XGBoost	NN	0.00000	SVM	NN	0.00000
3	XGBoost	RF	0.64048	RF	SVM	0.00000
	XGBoost	SVM	0.00000	RF	NN	0.00000
	XGBoost	NN	0.00000	SVM	NN	0.00001
4	XGBoost	RF	0.73512	RF	SVM	0.00000
	XGBoost	SVM	0.00001	RF	NN	0.00000
	XGBoost	NN	0.00000	SVM	NN	0.00000
5	XGBoost	RF	0.74176	RF	SVM	0.00000
	XGBoost	SVM	0.00000	RF	NN	0.00000
	XGBoost	NN	0.00000	SVM	NN	0.00000

Table 6. Mean AUC Values for Prediction Period of 1 day

Method	KOSPI	SSE	HMC	SKT
XGBoost	0.54129∓ 0.03131	0.55312∓ 0.03714	0.56489∓ 0.0263	0.55439∓ 0.02065
RF	0.51585∓ 0.03648	0.52598∓ 0.0463	0.54195∓ 0.0379	0.53890∓ 0.0317
SVM	0.51453∓ 0.02288	0.54321∓ 0.04618	0.52888∓ 0.01646	0.54705∓ 0.02382
NN	0.53544∓ 0.03667	0.50322∓ 0.04194	0.53502∓ 0.03387	0.53179∓ 0.04139

Table 7. Mean AUC Values for Prediction Period of 2 days

Method	KOSPI	SSE	HMC	SKT
XGBoost	0.71357∓ 0.02381	0.67962∓ 0.03164	0.69009∓ 0.03709	0.69074∓ 0.01494
RF	0.71640∓ 0.02211	0.67993∓ 0.02262	0.70303∓ 0.03847	0.69658∓ 0.02612
SVM	0.65772∓ 0.02636	0.64411∓ 0.0261	0.63261∓ 0.02047	0.6766∓ 0.02325
NN	0.56827∓ 0.03276	0.5588∓ 0.03192	0.56784∓ 0.02783	0.58323∓ 0.02085

Table 8. Mean AUC Values for Prediction Period of 3 days

Method	KOSPI	SSE	HMC	SKT
XGBoost	0.77567∓ 0.03111	0.7421∓ 0.04102	0.76379∓ 0.03005	0.762∓ 0.02233
RF	0.78045∓ 0.02726	0.73216∓ 0.03861	0.77064∓ 0.03167	0.77465∓ 0.02527
SVM	0.68829∓ 0.02890	0.69781∓ 0.03896	0.67746∓ 0.02716	0.71416∓ 0.02471
NN	0.59043∓ 0.04709	0.6136∓ 0.04052	0.60472∓ 0.04138	0.63862∓ 0.03745

Table 9. Mean AUC Values for Prediction Period of 4 days

Method	KOSPI	SSE	HMC	SKT
XGBoost	0.82713∓ 0.04080	0.81305∓ 0.02419	0.79857∓ 0.019	0.79317∓ 0.02418
RF	0.83803∓ 0.03506	0.81036∓ 0.02118	0.79978∓ 0.01999	0.79226∓ 0.02925
SVM	0.73892∓ 0.02779	0.73435∓ 0.02175	0.72531∓ 0.02306	0.73774∓ 0.03084
NN	0.62174∓ 0.04446	0.61567∓ 0.02874	0.62963∓ 0.03317	0.64763∓ 0.03657

Table 10. Mean AUC Values for Prediction Period of 5 days

Method	KOSPI	SSE	HMC	SKT
XGBoost	0.86004∓ 0.02393	0.84774∓ 0.02569	0.83429∓ 0.02224	0.84744∓ 0.02386
RF	0.86691∓ 0.02247	0.86340∓ 0.01533	0.83218∓ 0.02924	0.85401∓ 0.02832
SVM	0.77097∓ 0.03089	0.78563∓ 0.02715	0.75216∓ 0.02366	0.78037∓ 0.02573
NN	0.63192∓ 0.03619	0.65371∓ 0.02535	0.64279∓ 0.03948	0.64392∓ 0.03784

Note that XGBoost is equivalent to SVM in the prediction of the movement directions of SSE after 1 day. With respect to RF, it is equivalent to XGBoost in predicting the movement directions after 2-5 days and is even better than XGBoost in predicting the movement directions of SSE after 5 days.

V. Concluding Remarks

The results of the previous studies on the movement directions of index and stock prices depend on the dataset, the prediction period, and the input variables used. This paper aims to compare the extreme gradient boosting technique with other best ones reported in the previous studies such as SVM, ANN, and Random Forest in the context of the Korean stock market. Some representative index and stock price data of Korea stock market for 12 years are used for the experiments. It is also tested in the experiments about the effect of prediction periods of 1 to 5 days, compared with most other studies of single prediction period, and the effect of variable selection by prediction periods. The difference among the best sets of input variables by prediction periods identifies the importance of variable selection in predicting the movement directions of index and stock prices. As expected, prediction quality of each technique improves as the prediction period expands. XGBoost is mostly the best in predicting the movement directions after 1 to 4 days with a few partial equivalence to the other methods.

In spite of the successful application of the gradient boosting technique to the prediction problem of the movement directions of index and stock prices, there are some limitations to be noted. This study tested only 4 index and stocks in Korea stock market although they represents the stock market. More extensive research will be needed for making the comparison of the techniques more general. This study used 30 potential input variables, commonly used in the literature, for variable selection. In this respect, the set need to be extended for better quality of predictions and strategic decisions suitable for the Korea stock market. The experiments are also limited in that parameters of each algorithms are set with default values. Another option is to use parameter optimization such as bayesian optimization for parameter tuning[18].

참고문헌

[1] M. Ballings, D. V. Poel, N. Hespeels, and R. Gryp, "Evaluating Multiple Classifiers for Stock Price Direction Prediction," Expert Systems with Applications, Vol.42, pp.7046-7056, 2015.

[2] B. G. Malkiel and E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," The Journal of Finance, Vol.25, No.2, pp.383-417, 1970.

[3] S. B. Imandoust and M. Bolandraftar, "Forecasting the Direction of Stock Market Index Movement Using Three Data Mining Techniques: the Case of Tehran Stock Exchange," Int'l Journal of Engineering Research and Applications, Vol.4, No.6, pp.106-117, June. 2014.

[4] Y. Kara, M. A. Boyacioglu, and O. K. Baykan,

"Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines," Expert Systems with Applications, Vol.38, pp.5311-5319, May 2011.

[5] W. Huang, Y. Nakamori, and S. Y. Wang, "Forecasting Stock Market Movement Direction with Support Vector Machine," Computers and Operations Research, Vol.32, pp.2513-2522, 2005.

[6] T. Manojlovic and I. Stajduhar, "Predicting Stock Market Trends Using Random Forests: A Sample of the Zagreb Stock Exchange," Proc. of MIPRO 2015, Opatija, Croatia, pp.1189-1193, 2015.

[7] X. Zhong and D. Enke, "Forecasting Daily Stock Market Return Using Dimensionality Reduction," Expert Systems with Applications, Vol.67, pp.126-139, 2017.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proc. of the KDD '16 of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, San Francisco, USA, pp.785-794, August 13-17, 2016.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Taining Agorithm for Optimal Margin Classifiers," Proc. of the Fifth Annual Workshop on Computational Learning Theory," Pittsburgh, USA, pp.144-152, July 27-29, 1992.

[10] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," Neural Networks, Vol.61, pp.85-117, 2015.

[11] L. Breiman, "Random forests," Machine Learning, Vol.45, No.1, pp.5-32, 2001.

[12] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Vol.29, No.5 pp.1189-1232, 2001.

[13] K. Kim, "Financial Time Series Forecasting Using Support Vector Machines, Neurocomputing," Vol.55, pp.307-319, 2003.

[14] K. Manish and M. Thenmozhi, "Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest," Proceedings of Ninth Indian Institute of Capital Markets Conference, Mumbai, India, http://ssrn.com/abstract=876544, 2005.

[15] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques," Expert Systems with Applications, Vol.42, No.1, pp.259-268, 2015.

[16] F. Provost, T. Fawcett, and R. Kohavi, "The Case against Accuracy Estimation for Comparing Induction Algorithms," Proc. of the Fifteenth Int'l Conf. on Machine Learning, Madison, USA, pp.445-453, July 24-27, 1998.

[17] http://finance.naver.com/

[18] B. Shahrirai, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," Proceedings of the IEEE, Vol.104, No.1, pp.148-175, 2016.

저자소개

김 형 도(HyoungDo Kim) 정회원

- 1985년 2월 : 서울대학교 산업공학과(학사)
- 1987년 2월 / 1992년 8월 : KAIST 경영과학과(석사 / 박사)
- 1993년 ~ 1999년 : ㈜데이콤 EC 인터넷 기술 팀장
- 2000년 ~ 2002년 : 아주대학교 정보통신전문대학원 교수
- 2003년 ~ 현재 : 한양사이버대학교 경영정보학과 교수

<관심분야> : 전자상거래, 정보보호, 데이터 마이닝, 비즈니스 인텔리전스