International Journal of Engineering Research And Management (IJERM)
ISSN: 2349- 2058, Volume-06, Issue-02, February 2019

An Empirical Study of Institutional Investors' Net Buy/Net Sell on Forecasting Stock Prices Using Genetic Programming in Taiwan

Chih-Ming Hsu

***Abstract*— Among various traditional investment tools, stock investment is one of the easily understood targets for ordinary people because the concept of trading stocks is relatively clear and simple. However, for investors, the essential task of accurately forecasting future trends or market prices of stocks is difficult, since the factors that can affect the performance of stocks in the trading market are diverse. Therefore, the problems involved in forecasting stock prices continuously attracts great interest from both researchers and practitioners. Investors making short-term stock investments, especially, prefer to forecast stock prices via technical indicators calculated based on stock-trading information. In addition, an investor pays very close attention to the institutional investors' net buy or net sell of each day , because its value can reflect future expectation about the overall performance of a corporation, as well as judgement regarding the future trends of stock prices. Therefore, this study proposes a stock price forecasting procedure based on genetic programming (GP) and cluster analysis, using technical indicators as predictors in an empirical study on the effects of institutional investors' net buy or net sell in forecasting stock prices in Taiwan. The feasibility and effectiveness of the proposed procedure are illustrated through examining five stocks with the highest trading volumes in the semiconductor section of the TAIEX (Taiwan Capitalization Weighted Stock Index). The implementation results show that the stock price modelling procedure applying the GP method is a robust and practical forecasting technique. Institutional investors' net buy/net sell of stocks can indeed improve forecasting performance. Furthermore, forecasting accuracy can be further refined through classifying the trends in institutional investors' net buy/net sell.**

***Index Terms*— Cluster Analysis, Genetic Programming, Institutional Investors, Stock Price Forecasting.**

I. Introduction

Investing in stocks is one of the traditional and easily understood investment tools for most people, since the concept of trading stocks is clear and simple compared to other investment targets, e.g. options, futures, or swaps. No matter what investing strategies an investor applies, forecasting future trends, even future market prices, is essential work for an investor. Some people believe that the financial condition of a corporation is the most critical to the market prices of stocks issued by the company, thus preferring to employ information in financial reports to project the achievements of a stock, called fundamental analysis. Some investors, especially those making short-term stock investment, however, think that the status in the stock trading market and current events recently are key to future stock prices. This group of investors prefer to forecast stock prices via the indices, i.e. technical indicators, based on the stock trading information, called technical analysis. Whether employing fundamental or technical analysis, forecasting the trends or prices of a stock is always a tough task, since the factors that can affect its performance in a trading market are diverse; thus this topic continuously attracts high interest of researchers coming from both academic and practical worlds. For example, Behravesh [1] applied regression models to forecast the stock prices of the major Iranian petrochemical companies, where the predictors (independent variables) included (1) stock price in the last month, (2) capital of the company, (3) P/E (price-earnings ratio), (4) DPS (dividend per share), and (5) EPS (earnings per share). He used E-Views software to run the regression models, and discussed choosing the best decision among petroleum companies for beneficiary business markets. The author also discovered the application and effectiveness for making stock investment decisions, and assigned a limited budget to each petroleum company. Yeh, Huang and Lee [2] incorporated the sequential minimal optimization and gradient projection methods to develop a two-stage multiple-kernel learning algorithm to resolve stock market forecasting problems. Their proposed algorithm can take several combined advantages from different hyperparameter settings and improve the overall system performance. In addition, the hyperparameter settings need not be specified in advance, and the trial-and-error procedure for determining the optimal values for the hyperparameters can be avoided. The proposed algorithm is demonstrated and compared to single kernel support vector regression (SKSVR) [3], autoregressive integrated moving average (ARIMA)[4], and TSK-type fuzzy neural network (FNN) [5] by carrying out experiments on datasets taken from the TAIEX (Taiwan Capitalization Weighted Stock Index). The experimental results revealed that their approach can provide performance superior to other methods. Zuo and Kita [6] transformed the continuous P/E ratio to a set of digitized values via a clustering algorithm, and forecast the P/E ratio by applying a Bayesian network to the set of digitized values. They took the NIKKEI stock average (NIKKEI225) and Toyota Motor Corporation stock price as examples. The results showed that their approach can attain similar accuracy and a better correlation coefficient compared to time-series forecast algorithms. In addition, their algorithm, using the Ward method, can improve the computational accuracy by

Manuscript received February, 2019.

Chih-Ming Hsu, Department of Business Administration, Minghsin University of Science and Technology, Hsinchu, Taiwan (R.O.C.)

15% and 20% for the NIKKEI stock average and Toyota Motor Corporation stock price, respectively, against the traditional AR (Auto Regressive), MA (Moving Average), ARMA (Auto Regressive Moving Average) and ARCH (AutoRegressive Conditional Heteroskedasticity) methods. Hsu [7] combined the backpropagation (BP) neural network, feature selection, and genetic programming (GP) techniques to develop a hybrid procedure for resolving stock and futures price forecasting problems with the technical indicators as predictors. He first used the BP neural network to construct a preliminary forecasting model, then utilized feature selection through simulation to probe the built neural network, thus selecting the critical technical indicators for forecasting stock and futures prices. Furthermore, the vital technical indicators were also automatically screened out by employing the GP method. Finally, the final forecasting model using the selected technical indicators was established using the BP neural network. The author used TAIEX futures of the spot month to demonstrate the feasibility and effectiveness of the proposed procedure. Based on the experimental results, the forecasting performance was significantly improved through selecting appropriate technical indicators by applying the feature selection method or solely based on the preliminary GP forecasting model. Liu and Hu [8] proposed a feature-weighted support vector machine regression algorithm by providing different weights for different features of the samples, thus improving the performance of traditional SVM. A case study on examining sample stock data sets selected from China was conducted to demonstrate their proposed method, and the result showed that using GCD (grey correlation degree) as the weight value had good generalization capability, and the prediction accuracy improved. Xiong, Bao and Hu [9] applied multi-output support vector regression (MSVR), whose parameters are determined by their proposed approach based on the firefly algorithm (FA), to forecast the interval-valued stock price index series over short and long horizons. Three globally traded broad market indices (the S&P 500 for the US, the FTSE 100 for the UK, and the Nikkei 225 for Japan) were used to illustrate their method. The experimental results showed that their proposed FA-MSVR method outperformed some well-established counterparts based on statistical criteria regarding the forecasting accuracy measure and the accuracy of competing forecasts. Xiao, Xiao, Lu and Wang [10] proposed a three-stage nonlinear ensemble model based on neural networks, improved particle swarm optimization (IPSO), and support vector machines (SVM). In their study, three different types of neural-network-based models, including the Elman network, generalized regression neural network (GRNN) and wavelet neural network (WNN), all further optimized by improved particle swarm optimization (IPSO), were constructed. Later, the support vector machines (SVM) neural network was utilized to generate a neural-network-based nonlinear meta-model. Their proposed approach was able to explore complex nonlinear relationships better, and the built forecasting model was validated by three daily stock indices' time series, including the Shanghai composite index, Shenzhen component index, and Shanghai-Shenzhen 300. The empirical results demonstrated that their proposed ensemble approach significantly improved prediction performance over other individual models and linear combination models. Dan, Guo, Shi, Fang and Zhang [11] presented deterministic echo state network (ESN) models, which construct reservoirs randomly to simplify their structure and applications relative to the standard ESN, for stock price forecasting. They used two benchmark datasets, including the Shanghai Composite Index and S&P 500, to investigate the forecasting performance of their presented method. The experimental results showed that the deterministic ESN was able to outperform the standard ESN by about 20% and 52% in accuracy and stability, respectively, on average. Furthermore, about 23% improvement in efficiency, as well as insignificant improvement in forecasting accuracy, was found for the S&P 500 dataset. Singh and Borah [12] introduced a new type-2 fuzzy time series model, which was then enhanced by applying particle swarm optimization (PSO), in order to utilize more observations while forecasting. The authors' purpose was to tune up the lengths of intervals in the universe of discourse which are used while forecasting, but not to increase the number of intervals. The performance of their proposed model was evaluated by making a study on the daily stock index price data set of SBI (State Bank of India), as well as on the daily stock index price of Google. The experimental results showed that their proposed model is effective and robust compared to the existing fuzzy and conventional time series models. Rounaghi, Abbaszadeh and Arashi [13] utilized the multivariate adaptive regression splines (MARS) model and semiparametric splines technique to predict stock prices. In their study, the MARS model and semi-parametric smoothing splines technique serve as adaptive and nonparametric regression methods, respectively. They utilized 40 variables, including 30 accounting variables and 10 economic variables, to predict stock price by using the MARS model and semi-parametric splines technique. Four accounting variables: (1) book value per share, (2) predicted earnings per share, (3) P/E ratio and (4) risk, were selected by investigating the models to be the influencing variables on stock price forecasting via the MARS model. On the other hand, another combination of four accounting variables, which included dividends, net EPS, EPS forecast and P/E ratio, were chosen as effective variables while forecasting the stock prices. The performance regarding the multi-step ahead forecasting of their proposed approach was evaluated by comparing it to the traditional global linear model through simulation, and the results indicate the nonparametric model can yield superior forecasting performance compared to the global linear model. Furthermore, the intraday data of the Japanese stock price index and time series of heart rates are also analyzed and forecast in their study, and the experimental results revealed that the forecasting performance does not differ significantly in the Japanese stock price index, but the nonparametric model can provide significantly better performance while analyzing heart rates. Guo, Han, Shen and Li [14] applied the support vector machine regression (SVR) technique to tackle the characteristics, including discreteness, non-normality and high noise, which are significantly different in different periods for the same stock, or in the same period for different stocks in high-frequency data. In his study, an adaptive SVR with dynamic optimization of the learning parameters through particle swarm optimization (PSO) is developed to resolve the stock data at three different time scales (daily data, 30-min data, and 5-min data). Compared to the traditional SVR and backpropagation neural networks, his proposed approach can yield better results based on the experimental results. Wang [15] proposed a method based on the big data framework with fuzzy time series to forecast stock prices. She applied fuzzy time series to historical stock big data to predict the fuzzy trend regarding the forecast data, then determined the amount of fluctuation about the forecast data by using an autoregressive model. By integrating trend prediction with fluctuation quantity together, the forecast stock prices are finally obtained. Her proposed forecasting framework was illustrated by forecasting the TAIEX, and produced superior forecasting accuracy compared to existing methods. Chou and Nguyen [16] proposed a sliding-window metaheuristic optimization model by hybridizing the FA (firefly algorithm) and LSSVR (least squares support vector regression), called MetaFA-LSSVR, to predict the stock prices of Taiwanese construction companies one step ahead. Their proposed system is a stand-alone application with a graphical user interface, which is greatly interesting to home brokers that have insufficient investment knowledge. In addition, their proposed model is a favorable predictive technique for dealing with the highly nonlinear time series that traditional models have difficulty capturing.. Their experiments indicated that outstanding prediction performance and improved overall profit were attained by using their developed hybrid system. Cheng and Yang [17] utilized the rough set rule induction to develop a fuzzy time-series model to forecast stock indices. They employed rough sets to generate forecasting rules for replacing fuzzy logical relationship rules according to the lag period, and utilized an adaptive expectation model to improve forecasting performance. Furthermore, they also presented buy and sell rules to provide investment suggestions as references for investors based on three different scenarios. A dataset consisting of the TAIEX, Nikkei, and HSI stock prices from 1998 to 2012 was used to evaluate their proposed model, which was able to outperform existing models with comparing to the listing models under three error indices and profits criteria.

Thus, in the literature, some relatively new techniques, e.g. support vector regression (SVR), neural network (NN), genetic programming (GP), firefly algorithm (FA), and particle swarm optimization (PSO), have been broadly utilized to construct stock price forecasting models to improve the shortcomings or limitations of traditional statistical methods, thus obtaining fairly excellent experimental results. In addition, the trading information regarding institutional investors' net buy/net sell announced every trading day is important reference material when diagnosing the trends of stock prices for an investor. Therefore, this study intends to conduct an empirical study on the effects of institutional investors' net buy/net sell in forecasting stock prices in Taiwan based on genetic programming (GP). The information of institutional investors' net buy/net sell is used to recognize and classify the stock prices' trends, and GP is employed to construct the forecasting model corresponding to each classification (group) of stock prices' trends with technical indicators as predictors. The feasibility and effectiveness of the proposed procedure are demonstrated by examining the five stocks with top trading volumes in the semiconductor section of the TAIEX. The rest of the paper is organized as follows. Section 2 briefly introduces the genetic programming (GP) method, as well as the technical indicators. The proposed stock price forecasting procedure is presented and illustrated by a case study in Sections 3 and 4, respectively. Finally, conclusions and possible research directions are given in Section 5.

II. Methodologies

A. Genetic Programming

Based on the principles of Darwinian natural selection and biologically inspired operations, Koza (1992) invented an evolutionary method for generating programs or functions automatically, called genetic programming (GP), to solve a user-defined problem through evolving a population of chromosomes. The GP utilizes a tree-based structure consisting of terminal and function sets, as shown in Figure 1, to represent an individual program (chromosome). The tree (program) can be interpreted from the left to the right and from the top to the bottom as (5.2-x/15)+(9*sin(y)) . The available terminal elements to the branch in an evolving chromosome such as the constants, variables or zero-argument functions, etc., are defined in the terminal set. The function set consists of the functions of the program, e.g. the square root, minus, logarithm, or sine, etc.

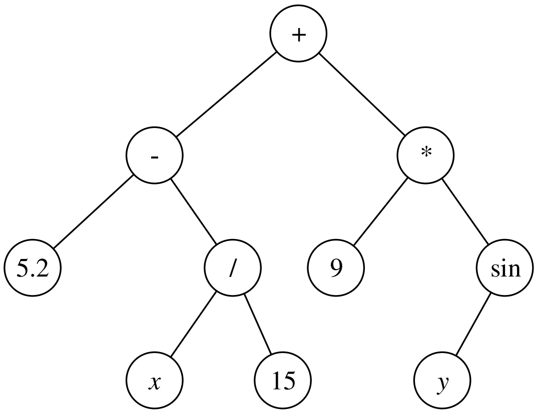

Figure 1. An example of tree structure expression in GP.

The fundamental steps of GP are briefly illustrated as follows [18–20]

Step 1: Create solutions of an initial population

First, create some solutions (chromosomes) of an initial population with a pre-specified population size, called generation 0, which are computer programs consisting of elements from the functional and terminals sets according to the characteristics of the problem.

Step 2: Evaluate the fitness of each chromosome

Execute each program decoded from the corresponding chromosome in the population and measure the degree of how well the program can deal with the problem at hand via a pre-defined fitness function, called fitness value.

Step 3: Select the elite chromosomes

Based on the fitness value of each chromosome, the probability corresponding to the chromosome is first obtained. Some chromosomes (programs) are then picked from the population by using the Russian roulette mechanism according to the obtained probability. These selected chromosomes form a matching pool.

Step 4: Apply genetic operators

To the selected programs in the matching pool, the genetic operators, including reproduction, crossover, mutation, and architecture-altering operator are randomly applied, thus creating an offspring population, called generation g+1, by replacing the elements in the population of the current generation g with the chromosomes in the offspring population based on a certain strategy.

Step 5: Examine the criteria for terminating the GP

If one of the termination criteria is satisfied, the best chromosome, i.e. the chromosome that can provide the best execution result (the highest fitness value) is designated as the final result of the GP run. Otherwise, repeat Steps 2 to 5 iteratively.

The GP widely extends the applications of genetic algorithms to the solution space consisting of computer programs. Hence, a lot of successful practical work using the GP in various application fields has been reported in the literature, e.g. [21–24]

B. Technical Indicators

Fundamental analysis and technical analysis are two major analytical tools for choosing appropriate investment stocks. The fundamental analysis is to analyze a business's financial statements (such as assets, liabilities, and earnings), health and competitors, and markets, as well as consider the overall state of the economy and other influential factors (such as the interest rates, production, earnings, employment, GDP, housing, manufacturing and management). With regard to the technical analysis, it is a method for forecasting stock price trends through studying past market data, primarily prices and volumes. The technical analysts widely use market indicators of many sorts, called technical indicators, which are mathematically calculated based on historic prices, volumes, and other inputs, thus aiming to forecast financial market direction. For example, the commodity channel index (CCI), average convergence/divergence (MACD), relative strength index (RSI), and stochastic oscillator (KD) are some common technical indicators.

III. Proposed Forecasting Procedure

This study intends to study the effects of institutional investors' net buy/net sell upon forecasting stock prices through genetic programming modelling. The research model is designed as depicted in Figure 2 and described as follows:

Step 1: Collect stock trading data

The trading data regarding the investment stocks are first collected.

Path I

Step 1-1: Calculate technical indicators

Some important technical indicators are calculated based on the collected stock trading data.

Step 1-2: Prepare training, testing and validation data

For each investment stock, the technical indicators obtained in Step 1-1 and stock trading data collected in Step 1 are first arranged day by day. For each trading day, the input variables, i.e. predictors, are the technical indicators, and the output variable, i.e. response, is the stock closing price in three days. The arranged data are then divided into two parts. The first part is further separated into the training and testing data based on an appropriate proportion, and the second part serves as the validation data.

Step 1-3: Build GP models

The GP technique is applied to construct the stock forecasting model based on the training and testing data prepared in Step 1-2. The GP will be executed several times, and the best GP model is designated by evaluating the weighted forecasting error associated with the training and testing data.

Path II

Step 2-1: Calculate technical indicators along with trends of net buy/net sell

Based on the collected stock trading data, several important technical indicators are calculated. Furthermore, the net buy/net sell by the institutional investors is also confirmed.

Step 2-2: Prepare training, testing and validation data

The obtained technical indicators, information about the trends of net buy/net sell by the institutional investors, and collected stock trading data are organized for each investment target and for each trading day. The data in each row include the predictors and response. The technical indicators serve as the predictors, and the stock closing price in three days is treated as the response. Then, the well-organized data are split into two groups. At the same time, the training and testing data are yielded according to a pre-specified ratio and the second part data form the validation data.

Step 2-3: Build GP models

The GP algorithm is implemented several times to establish the stock forecasting model by using the prepared training and testing data obtained in Step 2-2. The forecasting errors associated with the training and testing data are weighted to determine the best GP stock forecasting model.

Path III

Step 3-1: Calculate technical indicators along with trends of net buy/net sell

According to the stock trading data collected in Step 1, several important technical indicators are figured out. In addition, the trends of net buy/net sell by the institutional investors are also affirmed.

Step 3-2: Group data based on trends of net buy/net sell

First, the trends of net buy/net sell by the institutional investors are classified into three types. The first type's trend illustrates the rising situation, i.e. the net buys of the institutional investors in the previous three trading days are all positive. If all of the net buys of the institutional investors are negative in the previous three days, the trend is determined to be falling and is categorized as the second type. Remaining trading days which do not belong to type one or two are grouped into the trend of the third type.

Step 3-3: Prepare training, testing and validation data for each group of data

For each group of data classified in Step 3-1, the calculated technical indicators, information about the trends of net buy/net sell by the institutional investors, and collected stock trading data are put in order according to the sequences of trading days for each investment stock. Each row of data involves the technical indicators, serving as the predictors, and the corresponding stock closing price in three days, treated as the response. Then two sets are acquired by separating the arranged data. A pre-specified ratio is utilized to split the first part of the data into the training and testing data, while the validation data are the second part of the data.

Step 3-4: Build GP models for each data group

For each data group obtained in Step 3-2, the GP algorithm is executed several times to construct the GP stock forecasting models, where the required data come from the training and testing data corresponding to each data group prepared in Step 3-3. The forecasting performance of each GP model is evaluated by weighting the corresponding training and testing errors, thus settling the optimal GP forecasting model.

Step 2: Compare forecasting performance

The forecasting performance of the GP models resulting from Path I~III are compared by evaluating their RMSE (root-mean-square error), R^2 (R-square), and MAPE (mean absolute percentage error). Then, the effects of institutional investors' net buy/net sell on forecasting stock are concluded.

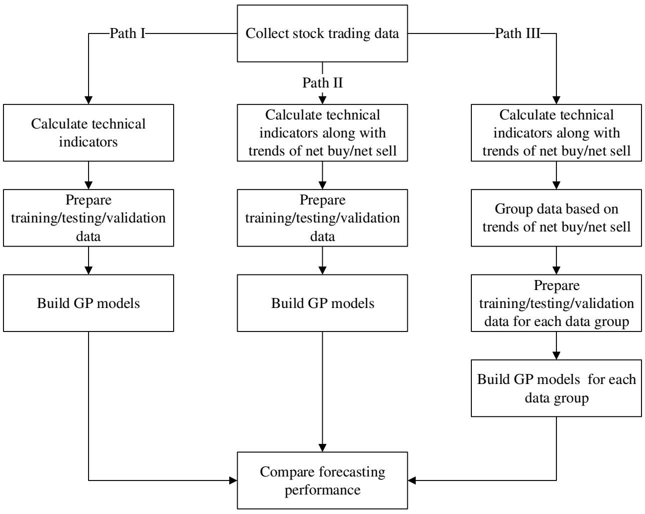
Collect stock trading data
Path I
Path II
Path III
Calculate technical indicators
Calculate technical indicators along with trends of net buy/net sell
Calculate technical indicators along with trends of net buy/net sell
Prepare training/testing/validation data
Prepare training/testing/validation data
Group data based on trends of net buy/net sell
Build GP models
Build GP models
Prepare training/testing/validation data for each data group
Build GP models for each data group
Compare forecasting performance

Figure 2. Proposed forecasting procedure.

IV. Case Study

In this section, we utilize the GP algorithm to construct the stock price forecasting models and compare their forecasting performance for several stocks, thus empirically examining the effects of institutional investors' net buy/net sell on forecasting stock prices.

A. Collection of Stock Trading Data

In this study, the examined stocks include five stocks with the highest yearly trading volumes in the semi-conductor section based on the statistics provided by the TWSE (Taiwan Stock Exchange Corporation). The five investigated stocks include stocks of codes 2303, 2330, 2337, 2344, and 6182. Then, the daily trading data include the open prices, highest prices, lowest prices, closing prices, and trading volumes along with the trading volumes of institutional investors' net buy/net sell for the selected five stocks from 2010/01/03 (the first trading day in 2010) to 2018/09/28 (the last trading day in September 2018).

B. Calculation of Technical Indicators

For each studied stock, the daily trading data are used to calculate its technical indicators. There are sixteen technical indicators considered in this study in accordance with Kim and Han [25], Kim and Lee [26], Tsang *et al.* [27], Chang and Liu [28], Ince and Trafalis [29], Huang and Tsai [30], and Lai, Fan, Huang and Chang [31]. Notably, the technical indicators are normalized into the range (-1, +1) based on the corresponding technical indicator's maximum and minimum values to avoid the technical indicators with larger indices dictating the forecasting models.

C. Determination of trends regarding net buy/net sell

For each trading day of each examined stock, the trends regarding net buy/net sell are classified into three types. If the trading volumes of institutional investors' net buy for an examined stock are all positive in the previous successive three days (e.g. $t-2$, $t-1$, and t days), the trend on trading day t is determined as rising (type 1). On the other hand, the trend on trading day t is considered falling (type 2) if all of the trading volumes of institutional investors' net buy in the previous successive $t-2$, $t-1$, and t days are negative. The net buy/net sell trends on one trading day are categorized as type 3.

D. Preparation of training, testing and validation data

For each trading day, the corresponding data are arranged in a row in the form (**X**, *y*), where **X** is the vector of input variables and *y* is the output variable. The output variable is the stock closing price three days later. However, there are three different situations for determining the input variables. In the first one, the technical indicators of each trading day are used for the input variables. In the second situation, the input variables are consisted by the technical indicators each trading day along with the trading volumes of institutional investors in the previous three days. Finally, in the last situation, the input variables are the same as those in the second situation, but the paired input/output data are further segmented into three types based on the trends regarding net buy/net sell determined previously. Next, for the first or second situation, the arranged data from 2011/01/03 to 2016/12/27 form the first group, and the arranged data from 2016/12/28 to 2018/09/28 make up the second group. The first data group are then divided into the training and testing data randomly according to a ratio 3:1, while the second data group becomes the validation data. As regarding each type's data in the third situation, the arranged data between 2011/01/03 to 2016/12/27 form the first group, and are split into the training and testing data with a proportion of 3:1 randomly, while the arranged data between 2016/12/28 to 2018/09/28 turn into the validation data.

E. Building forecasting models by GP

For the first and second situations as well as each type in the third situation, in the preparation of training, testing and validation data, the training data are used to construct the stock price forecasting models by applying the GP techniques. Here, Discipulus software is utilized in this study. The main parameters in GP are set as software's default values (population size = 500, mutation rate = 0.95, and crossover rate = 0.5). The fitness value of each chromosome is measured by the RMSE (root-mean-square error), for which smaller is better. In addition, the built models are also evaluated on their forecasting performance via their corresponding testing data to assess the flexibility of the established GP models when applied to unknown data. The implementation results for stock code 2303 are summarized in Table 1. Based on Table 1, the execution result in bold represents the optimal result evaluated according to the weighted training and testing R^2 among each of the ten implementations. All of the coefficients of variation (*CV*) regarding the training RMSE, testing RMSE, weighted training and testing RMSE, training R^2, testing R^2, and weighted training and testing R^2 are considered to be sufficiently small, as shown in Table 1, where the largest *CV* is just 0.064993. Therefore, the stock price modelling procedure via the GP technique can be thought of as a robust and useful tool.

Similarly, the same modelling method is also implemented for the other four investigated stocks. The selected best GP models are given in Table 2.

Table 1. The Implement Results of Stock Code 2303.

(A) Situation 1

GP Run No.	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
1	0.004459	0.004429	0.004444	0.90334	0.90371	0.90353
2	**0.004668**	**0.004633**	**0.004651**	**0.90706**	**0.90747**	**0.90726**
3	0.004937	0.004934	0.004936	0.89696	0.89697	0.89696
4	0.004469	0.004504	0.004486	0.90481	0.90355	0.90418
5	0.005186	0.005366	0.005276	0.89379	0.88943	0.89161
6	0.004679	0.004583	0.004616	0.90274	0.90235	0.90300
7	0.004859	0.004720	0.004790	0.90588	0.90623	0.90606
8	0.004904	0.005105	0.005004	0.90156	0.89636	0.89896
9	0.004457	0.004404	0.004431	0.90562	0.90593	0.90578
10	0.004660	0.004703	0.004682	0.90039	0.89888	0.89963
Mean	0.004758	0.004772	0.004764	0.90209	0.900797	0.901492
Standard deviation	0.000236	0.00031	0.000273	0.004205	0.00563	0.004873
Coefficient of variation	0.049562	0.064993	0.057341	0.004661	0.00625	0.005405

(B) Situation 2

GP Run No.	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
1	0.004107	0.003846	0.003977	0.91105	0.91635	0.91370
2	0.004284	0.004272	0.004278	0.90950	0.90862	0.90906
3	0.004310	0.004166	0.004238	0.90744	0.91124	0.90934
4	0.004279	0.004189	0.004234	0.90763	0.90920	0.90842
5	0.004056	0.003851	0.003953	0.91360	0.91705	0.91533
6	0.004024	0.004000	0.004012	0.91273	0.91360	0.91317
7	0.004316	0.004012	0.004164	0.90787	0.91332	0.91059
8	**0.004049**	**0.003839**	**0.003944**	**0.91211**	**0.91746**	**0.91479**
9	0.003990	0.003931	0.003961	0.91338	0.91507	0.91422
10	0.004014	0.003855	0.003935	0.91297	0.91638	0.91467
Mean	0.004147	0.004013	0.004080	0.910803	0.913549	0.912177
Standard deviation	0.000137	0.000162	0.000141	0.002501	0.003224	0.002680
Coefficient of variation	0.032957	0.040258	0.034557	0.002746	0.003529	0.002938

(C) Situation 3

GP Run No.	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
Type 1						
1	0.004031	0.003781	0.003906	0.91794	0.92233	0.92013
2	0.004011	0.003675	0.003843	0.91907	0.92536	0.92221
3	0.004130	0.003551	0.003840	0.91542	0.92687	0.92115
4	**0.003858**	**0.003410**	**0.003634**	**0.92180**	**0.92979**	**0.92580**
5	0.004168	0.003547	0.003858	0.91594	0.92682	0.92138
6	0.004065	0.003640	0.003852	0.91798	0.92530	0.92151
7	0.003921	0.003500	0.003710	0.91967	0.92829	0.92398
8	0.004099	0.003518	0.003808	0.91604	0.92763	0.92183
9	0.004178	0.003597	0.003887	0.91643	0.92659	0.92151
10	0.004106	0.003764	0.003935	0.91711	0.92278	0.91995
Mean	0.004060	0.003578	0.003819	0.917718	0.926603	0.922147
Standard deviation	0.000104	0.000118	0.000091	0.001992	0.002321	0.001754
Coefficient of variation	0.025627	0.032924	0.023877	0.002170	0.002504	0.001902
Type 2						
1	0.003188	0.00362	0.003404	0.91181	0.89326	0.90254
2	0.002987	0.003329	0.003158	0.91771	0.90164	0.90968
3	0.002981	0.003355	0.003168	0.91881	0.90026	0.90953
4	0.003137	0.003476	0.003307	0.91392	0.89651	0.90521
5	0.003142	0.003515	0.003329	0.91474	0.89536	0.90505
6	0.002985	0.003408	0.003197	0.91863	0.90065	0.90964
7	0.003068	0.00335	0.003209	0.91639	0.9031	0.90975
8	0.003023	0.003552	0.003287	0.91796	0.89812	0.90804
9	**0.003213**	**0.003201**	**0.003207**	**0.91242**	**0.90623**	**0.90933**
10	0.00334	0.003512	0.003426	0.91021	0.89712	0.90366
Mean	0.003097	0.003411	0.003254	0.915643	0.899888	0.907766
Standard deviation	0.000119	0.000126	0.000096	0.003092	0.003895	0.002830
Coefficient of variation	0.038495	0.036848	0.029580	0.003377	0.004329	0.003118
Type 3						
1	0.004151	0.003904	0.004028	0.90494	0.91121	0.90808
2	0.003786	0.003829	0.003808	0.91333	0.91281	0.91307
3	0.003687	0.003936	0.003811	0.91551	0.91108	0.91330
4	0.004134	0.003882	0.004008	0.90607	0.91239	0.90923
5	**0.003768**	**0.003519**	**0.003644**	**0.91426**	**0.91994**	**0.91710**
6	0.003658	0.003710	0.003684	0.91708	0.91629	0.91668
7	0.003726	0.003805	0.003765	0.91573	0.91482	0.91528
8	0.003877	0.003924	0.003901	0.91320	0.91118	0.91219
9	0.004038	0.003879	0.003958	0.90646	0.91336	0.91141
10	0.003952	0.003574	0.003763	0.91162	0.91957	0.91559
Mean	0.003847	0.003784	0.003816	0.912584	0.914604	0.913761
Standard deviation	0.000182	0.000148	0.000132	0.004421	0.003337	0.003047
Coefficient of variation	0.047430	0.039048	0.034646	0.004844	0.003648	0.003335

Table 2. The Selected Best GP Models.

Stock #	Situation	Type	Training RMSE	Testing RMSE	Weighted Training and Testing RMSE	Training R^2	Testing R^2	Weighted Training and Testing R^2
2303	1	N/A	0.004668	0.004633	0.004651	0.90706	0.90747	0.90726
	2	N/A	0.004049	0.003839	0.003944	0.91211	0.91746	0.91479
	3	1	0.003858	0.003410	0.003634	0.92180	0.92979	0.92580
		2	0.003213	0.003201	0.003207	0.91242	0.90623	0.90933
		3	0.003768	0.003519	0.003644	0.91426	0.91994	0.91710
2330	1	N/A	0.000648	0.000720	0.000684	0.99395	0.99511	0.99453
	2	N/A	0.000645	0.000529	0.000587	0.99468	0.99512	0.99490
	3	1	0.000494	0.000427	0.000460	0.99503	0.99567	0.99535
		2	0.000534	0.000514	0.000524	0.99439	0.99467	0.99453
		3	0.000490	0.000466	0.000478	0.99517	0.99556	0.99537
2337	1	N/A	0.001400	0.001361	0.001381	0.98108	0.98150	0.98129
	2	N/A	0.000816	0.000676	0.000746	0.98752	0.98943	0.98847
	3	1	0.000740	0.000544	0.000642	0.98776	0.99147	0.98962
		2	0.000642	0.000611	0.000627	0.99205	0.99211	0.99208
		3	0.000772	0.000657	0.000715	0.98751	0.98881	0.98816
2344	1	N/A	0.000738	0.000750	0.000744	0.96957	0.96944	0.96951
	2	N/A	0.000644	0.000667	0.000655	0.97237	0.97135	0.97186
	3	1	0.000485	0.000452	0.000469	0.97693	0.97778	0.97735
	3	2	0.000520	0.000547	0.000534	0.97223	0.97170	0.97197
	3	3	0.000611	0.000588	0.000600	0.96907	0.97031	0.96969
6182	1	N/A	0.001468	0.001660	0.001564	0.98289	0.98040	0.90135
	2	N/A	0.001060	0.001164	0.001112	0.98721	0.98583	0.98652
	3	1	0.000932	0.001073	0.001002	0.99069	0.98921	0.98995
		2	0.002421	0.002142	0.002282	0.98121	0.98170	0.98145
		3	0.001152	0.001165	0.001158	0.98561	0.98567	0.98564

F. Comparing the forecasting performance

To realize the institutional investors' net buy/net sell on forecasting stock prices, the forecasting performance for the selected GP models in Table 3 is appraised via MAPE (mean absolute percentage error). Table 3 also reveals the forecasting performance by applying the selected GP models to the never-met validation data while constructing and choosing the optimal models in order to examine the GP models' generalizability. The numbers in bold represent the best result, i.e. the highest R^2 or lowest MAPE, among the three situations for a certain stock. Based on Table 3, we can find that the three GP forecasting models constructed in the third situation for each stock can jointly provide the optimal forecasting for the stock prices in general, except that the GP model built in the second situation yields the smallest testing MAPE. This implies that the information regarding the institutional investors' net buy/net sell is helpful to forecast stock prices. Furthermore, the classification for the trends of net buy/net sell by the institutional investors is also beneficial to improve the forecasting accuracy further. In conclusion, the gathering of institutional investors' net buy/net sell can really assist investors in forecasting future stock prices in addition to collecting common technical indicators.

Table 3. Forecasting Performance.

Stock #	Situation	Training R^2	Testing R^2	Validation R^2	Training MAPE	Testing MAPE	Validation MAPE
2303	1	0.90706	0.90747	0.92250	0.025637	0.025465	0.027656
	2	0.91211	0.91746	0.93873	0.023123	0.022880	0.031976
	3	**0.92088**	**0.92159**	**0.95002**	**0.022122**	**0.02077**	**0.021480**
2330	1	0.99360	0.99290	0.84177	0.020967	0.021897	0.029619
	2	0.99395	0.99511	0.92424	0.022535	0.020252	0.026072
	3	**0.99497**	**0.99542**	**0.94611**	**0.019202**	**0.019157**	**0.019156**
2337	1	0.98108	0.98150	0.87923	0.009902	0.009959	0.036586
	2	0.98752	0.98943	0.92075	0.007811	0.007509	0.025909
	3	**0.98852**	**0.99025**	**0.94435**	**0.007609**	**0.007273**	**0.021687**
2344	1	0.96957	0.96944	0.81370	0.010749	0.010921	0.046746
	2	0.97237	0.97135	0.87827	0.010264	0.010409	0.035349
	3	**0.97675**	**0.97170**	**0.94157**	**0.009547**	**0.009342**	**0.022642**
6182	1	0.98289	0.98040	0.90135	0.067000	0.068330	0.091958
	2	0.98721	0.98583	0.96717	0.065276	**0.063191**	0.068138
	3	**0.98849**	**0.98804**	**0.97356**	**0.063385**	0.065947	**0.063076**

V. Conclusions

Stock investing is easy to understand for an investor among various investment targets. The ordinary operation for investing in a stock is to buy a stock at a relatively low price and sell it at a high price later. In addition, first selling a stock at a high price and buying back the stock at a low price later, called selling short, is also a possible method. For either, it's an important and critical issue to forecast future stock prices, thus assisting in making proper investment decisions. However, forecasting is a complex and difficult task in itself, with the behavior of a stock's prices being intangible, since the factors that affect the market prices regarding a stock are multitudinous. Traditionally, an investor utilizes fundamental analysis of a business's financial conditions and other influential factors to map the future long-term financial performance of the issued stock. The technical analysis, applying the technical indicators, is more suitable for short-term forecasting. In addition, an investor pays much closer attention to the institutional investors' net buy/net sell of each day in Taiwan, as its value indicates the degree of optimism about the overall performance of a corporation, which assists in judging the future trend of stock prices. Therefore, the present empirical study in Taiwan of the effects of institutional investors' net buy/net sell on forecasting stock prices uses genetic programming (GP), where the technical indicators serve as predictors. An examination procedure is proposed and five stocks with the highest trading volumes in the semiconductor section of the TAIEX are used to illustrate the proposed procedure. The implementation results show that the stock price modelling procedure using the GP method can be considered a steady and practical technique, with low coefficients of variation in the execution results. In addition, the forecasting performance can indeed be improved by gathering information on institutional investors' net buy/net sell of stock. Classification of the trends of net buy/net sell by the institutional investors can further help an investor to raise the forecasting accuracy. In conclusion, the trading volumes and classification of institutional investors' net buy/net sell in Taiwan have significant influence on the behavior and forecasting of stock prices according to this empirical study. Furthermore, genetic programming is shown to be a robust and effective tool for constructing forecasting models for an investor. Researchers can further apply clustering techniques to segment institutional investors' net buy/net sell trends, which may yield superior forecasting performance.

Acknowledgment

The author would like to thank the Minghsin University of Science and Technology, Taiwan, R.O.C. for supporting this research under Contract No. MUST-108BA-1.

References

[1] Behravesh, M., "Forecasting stock price of Iranian major petrochemical companies," African Journal of Business Management, Vol. 5, No. 1, 2011, pp. 7-12.

[2] Yeh, C. Y., Huang, C. W., and Lee, S. J., "A multiple-kernel support vector regression approach for stock market price forecasting," *Expert Systems with Applications*, 2011, Vol. 38, No. 3, pp. 2177-2186.

[3] Tay, F. E. H., and Cao, L., "Application of support vector machines in financial time series forecasting," *Omega: The International Journal of Management Science*, 2001, Vol. 29, No. 4, pp. 309-317.

[4] Box, G. E. P., and Jenkins, G. M., *Time Series Analysis: Forecasting and Control*, 5th ed., New Jersey: John Wiley & Sons, 2016.

[5] Chang, P.-C., and Liu, C.-H., "A TSK type fuzzy rule based system for stock price prediction," *Expert System with Applications*, 2008, Vol. 34, No. 1, pp. 135-144.

[6] Zuo, Y., and Kita, E., "Stock price forecast using Bayesian network," *Expert Systems with Applications*, 2012, Vol. 39, No. 8, pp. 6279-6737.

[7] Hsu, C.-M., "A hybrid procedure with feature selection for resolving stock/futures price forecasting problems prediction," *Expert Systems with Application*, 2013, Vol. 22, No. 3-4, pp. 651-671.

[8] Liu, J.-N.-K, and Hu, Y.-X., "Application of feature-weighted Support Vector regression using grey correlation degree to stock price forecasting," *Neural Computing and Applications*, 2013, Vol. 22, pp. S143-S152.

[9] Xiong, T., Bao, Y.-K, and Hu, Z.-Y., "Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting," *Knowledge-based Systems*, 2014, Vol. 55, pp. 87-100.

[10] Xiao, Y, Xiao, J., Lu, F.-B., and Wang, S.-Y., "Ensemble ANNs-PSO-GA approach for day-ahead stock e-exchange prices forecasting," *International Journal of Computational Intelligence Systems*, 2014, Vol. 7, No. 2, pp. 272-290.

[11] Dan, J.-P., Guo, W.-B., Shi, W.-R., Fang, B., and Zhang, T.-P., "Deterministic echo state networks based stock price forecasting," *Abstract and Applied Analysis*, 2014, Document No. 137148.

[12] Singh, P., and Borah, B., "Forecasting stock index price based on M-factors fuzzy time series and particle swarm optimization," *International Journal of Approximating Reasoning*, 2014, Vol. 55, No. 3, pp. 812-833.

[13] Rounaghi, M. M., Abbaszadeh, M. R., and Arashi, M., "Stock price forecasting for companies listed on Tehran stock exchange using multivariate adaptive regression splines model and semi-parametric splines technique," *Physics A- Statistical Mechanics and its Applications*, 2015, Vol. 438, 625-633.

[14] Guo, Y.-H., Han, S.-M., Shen, C.-H., and Li, Y., "An adaptive SVR for high-frequency stock price forecasting," *IEEE Access*, 2017, Vol. 6, pp. 11397-11404.

[15] Wang, W.-N., "A big data framework for stock price forecasting using fuzzy time series," *Multimedia Tools and Applications*, 2018, Vol. 77, No. 8, pp. 10123-10134.

[16] Chou, J.-S., and Nguyen, T.-K., "Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression," *IEEE Transactions on Industrial Informatics*, 2018, Vol. 14, No. 7, pp. 3132-3142.

[17] Cheng, C.-H., and Yang, J.-H., "Fuzzy time-series model based on rough set rule induction for forecasting stock price," *Neurocomputing*, 2018, Vol. 302, pp. 33-45.

[18] Koza, J. R., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., and Lanza, G., *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*, New York: Springer, 2005.

[19] Ciglarič, I., and Kidrič, A., "Computer-aided derivation of the optimal mathematical models to study gear-pair dynamic by using genetic programming," *Structural and Multidisciplinary Optimization*, 2006, Vol. 32, No. 2, pp. 153-160.

[20] Koza, J. R., Streeter, M. J., and Keane, M. A., Routine high-return human-competitive automated problem-solving by means of genetic programming," *Information Sciences*, 2008, Vol. 178, No. 23, pp. 4434-4452.

[21] Kazeminia, A., Kaedi, M., and Ganji, B., "Personality-based personalization of online store features using genetic programming: analysis and experiment," *Journal of Theoretical and Applied Electronic Commerce Research*, 2019, Vol. 14, No. 1, pp. 16-29.

[22] Saghafi, H., and Arabloo, M., "Development of genetic programming (GP) models for gas condensate compressibility factor determination below dew point pressure," *Journal of Petroleum Science and Engineering*, 2018, Vol. 171, pp. 890-904.

[23] Fathi, A., and Sadeghi, R., "A genetic programming method for feature mapping to improve prediction of HIV-1 protease cleavage site," *Applied Soft Computing*, 2018, Vol. 72, pp. 56-64.

[24] Shen, J., and Jiménez, R., "Predicting the shear strength parameters of sandstone using genetic programming," *Bulletin of Engineering Geology and the Environment*, 2018, Vol. 77, No. 4, pp. 1647-1662.

[25] Kim, K.-J. and Han, I., "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index," *Expert System with Applications*, 2000, Vol. 19, No. 2, pp. 125-132.

[26] Kim, K.-J., and Lee, W. B., "Stock market prediction using artificial neural networks with optimal feature transformation," *Neural Computing and Applications*, 2004, Vol. 1, No. 3, pp. 255-260.

[27] Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., Tsang, J., Koong, K., and Wong, T.-L., "Design and implementation of NN5 for Hong Kong stock price forecasting," *Engineering Applications of Artificial Intelligence*, 2007, Vol. 20, No. 4, pp. 453-461.

[28] Chang, P.-C., and Liu, C.-H., "A TSK type fuzzy rule based system for stock price prediction," *Expert System with Applications*, 2008, Vol. 34, No. 1, pp. 135-144.

[29] Ince, H. and Trafalis, T. B., "Short term forecasting with support vector machines and application to stock price prediction," *International Journal of General System*, 2008, Vol. 37, No. 6, pp. 677-687.

[30] Huang, Z. W., Li, M. Z., Chousidis, C., Mousavi, A., and Jiang, C. J., "Schema theory-based data engineering in gene expression programming for big data analytics," *IEEE Transactions on Evolutionary Computation*, 2018, Vol. 2, No. 5, pp. 792-804.

[31] Lai, R. K., Fan, C.-Y., Huang, W.-H., and Chang, P.-C., "Evolving and clustering fuzzy decision tree for financial time series data forecasting," *Expert System with Applications*, 2009, Vol. 36, No. 2, pp. 3761-3773.

Chih-Ming Hsu Chih-Ming Hsu is currently a Professor in the Department of Business Administration at Minghsin University of Science and Technology, Taiwan. He holds a PhD in Industrial Engineering and Management from National Chiao Tung University, Taiwan. His present research interests include quality engineering, optimization methods in industrial applications and data mining applications in CRM.