Apply Support Vector Regression to Forecast Stock Prices with Feature Selection through Clustering

Chih-Ming Hsu

Abstract— **Accurately forecasting future stock prices is an essential and important issue for an investor to make an expected profit, as well as decrease the investment risk. In the past, the topic that determines the factors crucial to the forecasting of future stock prices is rarely addressed. In this study, the support vector regression and TwoStep cluster analysis are integrated to propose an approach for tackling the stock price forecasting problems with a function of feature selection. The feasibility and effectiveness of the proposed method are validated through making a case study on three different levels' indices that include the TAIEX, FTSE TWSE Taiwan 50 Index, and Taiwan 2303 Stock. The experimental results show that the feature selection mechanism can efficiently screen out the critical technical indicators for forecasting the future stock prices, as well as can remove the superfluous indicators that might interfere the forecasting abilities of other critical technical indicators. Next, the feature selection procedure can greatly reduce the total number of technical indicators to prevent the over-fitting of training the training data. The selected best simplified forecasting model cannot always provide better forecasting performance. However, the investors can only pay attention to less technical indicators, and can obtain a satisfactory forecasting result with a high accuracy. Hence, an investor can save time, cost, and effort while building a forecasting model, thus make more concentration on fewer financial analysis and trading strategies.**

Index Terms—**Stock price forecasting, Support vector regression, TwoStep cluster analysis, Feature selection**

I. INTRODUCTION

Among various financial tools, investing stocks may be the simplest. An investor can make a profit through two main ways: (1) buy a stock at a low price, then sell this stock at a high price; (2) sell a stock at a certain price, and repurchase this stock at a relative low price. In any case, forecasting stock prices accurately is essential for making an expected profit, or controlling the investment risk for an investor. For making a long-term investment plan, the fundamental analysis is usually used to explore a business's financial statements, health, competitors and markets etc. On the other hand, the technical analysis is frequently utilized to forecast the prices (or direction of prices) of a stock by studying the past market data, primarily trading prices and volumes. Therefore, there are a lot of researches on the problems of foresting stock prices through different methods in the past. For example, [1] applied the combined prediction's principle and artificial intelligence's method to propose a hybrid forecasting model. Their study combines the single models in series to develop the principle of hybrid prediction. The simulation results and theoretical proofs reveal that their combined model can provide a superior stock price forecasting accuracy, thus resulting in more profits than other single models. [2] defined the most relevant attributes by calculating the distinct features to be analyzed by means of attribute selection, to propose an approach for forecasting the maximum and minimum day stock prices of three Brazilian power distribution companies. The artificial neural networks are then carried out to yield the actual predictions whose performances are evaluated by the averages of mean absolute error, mean absolute percentage error and root mean square error. The experimental outcomes reveal that the proposed methodology for predicting the maximum and minimum day stock prices is effective, as well as the combination of attribute selection by the correlation analysis and artificial neural networks can achieve their greater results. [3] combined the v-SVR model, principal component analysis (PCA) and brain storm optimization (BSO) to propose a hybrid approach for forecasting the stock price indices. The input variables of v-SVR are first determined from 20 technical indicators by the correlation analysis and PCA. Furthermore, the parameters of v-SVR are optimized by the BSO algorithm. The effectiveness and efficiency of their developed hybrid forecast strategy are evaluated through a case study regarding the China Securities Index 300 (CSI300) and Shenzhen Stock Exchange Component Index (SZSE Component Index). Based on the experimental results, their proposed model is simple, as well as able to sufficiently approximate the actual CSI300 stock price index. [4] combined the variational mode decomposition (VMD) with backpropagation neural network (BPNN) to forecast the intraday stock prices. The VMD, a new multiresolution technique, is applied to decompose series of stock prices into a sum of variational modes (VM). Then, the BPNN model is trained based on the extracted VM, as well as the initial weights of BPNN are optimized by the particle swarm optimization (PSO). A case study of six stocks indicate that his proposed hybrid model can obtain better forecasting results over the baseline predictive model which is a PSO BPNN model trained by the past prices. [5] proposed a trading rule according to the flag pattern recognition. They first develop a dynamic window scheme that can quarterly update the stop loss and take profit. They also use the EMA indicator, calculated both for 15-min and 1-day timeframes for simultaneously considering short and medium terms, to filter trades due to the trend-following pattern of flag pattern. In addition, they also utilize the price range on which they are developed and limit the maximum loss of each trade to 100 points for filtering the flags. Their proposed method is

Manuscript received March 04, 2020.
Chih-Ming Hsu, Department of Business Administration, Minghsin University of Science and Technology, Hsin-Fong, Hsinchu

employed to 91,309 intraday observations of the DJIA (Dow-Jones Industrial Average) index. The execution results reveal that there is a significant improvement while comparing to the results obtained in the previous proposals and those obtained by the buy & hold strategy according to the profitability, risk, and transaction costs. [6] applied several nonparametric machine learning models, including artificial neural network, support vector machines with polynomial and radial basis function kernels to predict the trends of Korea Composite Stock Price Index 200 (KOSPI 200). They also state and test hypotheses about the controversial issues. Their results are significantly inconsistent with those reported in the precedent research, which are usually consider the high prediction performance. Next, they use Google Trends to prove that the factors in the previous studies are not effective while predicting the KOSPI 200 index prices in their approach. In addition, the ensemble methods are also proven that they cannot improve the accuracy of the prediction. [7] applied the fuzzy time series to propose a big data framework to forecast stock prices. This study uses the fuzzy time series to predict the forecasted data's fuzzy trend based on their historical stock big data. The autoregressive model is then utilized to find the fluctuation quantity regarding the forecasted data. Finally, the trend prediction and fluctuation quantity are combined to forecast stock prices. His proposed framework is demonstrated by forecasting the TAIEX, as well as compared to the existing methods for the forecasting accuracy. According to the implementation outcomes, his proposed forecasting framework can provide the superior forecasting performance. [8] utilized the long short-term memory (LSTM) and various generalized autoregressive conditional heteroscedasticity (GARCH)-type models to develop an integrated method to forecast the volatilities of stock prices. They also compare the forecasting performance of their proposed approach with the existing methodologies, e.g. the GARCH, exponential GARCH, exponentially weighted moving average, deep feedforward neural network (DFN), LSTM, and hybrid DFN models combining a DFN with one GARCH-type model. According to the experiments, their proposed hybrid model that combines the LSTM model with three GARCH-type models can provide the lowest prediction errors in terms of mean absolute error (MAE), mean squared error (MSE), heteroscedasticity adjusted MAE (HMAE), and heteroscedasticity adjusted MSE (HMSE). [9] combined the features learned from different representations of the same data, including the stock time series and stock chart images, to develop a model, called the feature fusion long short-term memory-convolutional neural network (LSTM-CNN) model to predict stock prices. The LSTM and CNN can extract temporal features and image features. The SPDR S&P 500 ETF data are used to measure the performance of their proposed model compared to those of single models, i.e. CNN and LSTM. The comparison results reveal that their feature fusion LSTM-CNN model can indeed increase the performance in predicting stock prices. In addition, a candlestick chart is the most appropriate among several stock chart images for forecasting stock prices. [10] hybridized proposes the support vector regression (SVR) with the firefly algorithm (FA) to propose an approach to resolve the stock price forecasting problems. They first use a modified version of the FA, called MFA, that applies the dynamic adjustment strategy and opposition-based chaotic strategy to improve the convergence speed of global optimization. Then, the hybrid SVR model, whose parameters are optimized by the MFA, is proposed to forecast the stock prices. The applicability and superiority of their proposed methodology are demonstrated by conducting comparative experiments. Based on the experimental results, the proposed MFA algorithm can yield superior performance while comparing to the other algorithms, thus the proposed MFA-SVR prediction procedure can be considered as a feasible and effective tool for forecasting stock prices.

The above literatures reveal that the machine learning techniques, e.g. support vector regression, particle swarm optimization, long short-term memory, and convolutional neural network etc., are broadly used in forecasting stock prices. In addition, the technical indicators are appropriate for resolving the problems of stock prices forecasting. However, the issue for determining the set of technical indicators that are most critical to the stock prices is rarely addressed. Therefore, this study intends to integrate the support vector regression and clustering technique to propose a model for conducting the feature selection thus forecasting stock prices. The remainders of the paper are organized as follows. Section 2 briefly introduces the two main methodologies used in this study. The integrated procedure of stock prices forecasting is represented in Section 3. Then, Section 4 illustrates the usage of our proposed forecasting approach. Section 5 provides the research conclusions and possible directions of future researches.

II. Methodologies

There are two main methodologies used to develop the stock prices forecasting procedure in this study.

A. *Support Vector Regression*

In the field of machine learning, support vector machine (SVM), whose current standard incarnation was published by [11], is a supervised learning model for classification. [12] proposed a version of SVM, called support vector regression (SVR), that extends the original SVM to resolve the regression problems.

Suppose there are N pairs of inputs and their corresponding output (X_k, y_k) where $X_k = (x_{k1}, x_{k2}, \ldots, x_{kn}) \in \mathbf{R}^n$ and $y_k \in \mathbf{R}^n$ are the input vector and output vector in n dimensions, respectively. We intend to build a regression model to represent the functional relationship between X_k and y_k. First, the original inputs X_k ($k = 1,2,\ldots,N$) are mapped into the $\phi_i(X_k)$ in a higher-dimensional space, say, m dimensions, to presumably make the separation easier relative to the original n-dimensional space. Therefore, the linear regression model can be stated as follows:

$$y_k' = f(X_k, W) = \sum_{i}^{m} w_i \phi_i(X_k) + w_0 = W^T \Phi(X_k) + w_0 \quad , \quad k = 1,2,\ldots,N \tag{1}$$

where y_k', W, $\Phi(X_k)$ and w_0 are the predicted output, weight vector, feature vector and bias, respectively. Notably, the weight vector W is consisted of the weight w_i, as well as the feature vector is made up by $\phi_i(X_k)$.

To show that there is no loss arisen when the predicted output is at an acceptable distance from the actual output, [13] introduced an ε-insensitive loss function to evaluate the predicted error as follows:

$$L_\varepsilon(y_k, y_k^{'}) = \begin{cases} 0 & \text{if } |y_k - y_k^{'}| \le \varepsilon \\ |y_k - y_k^{'}| - \varepsilon & \text{otherwise} \end{cases}, \quad k = 1,2,\ldots,N \tag{2}$$

Hence, we re-state the loss as follows:

$$y_k - W^T\Phi(X_k) - w_0 - \varepsilon \le \xi_k,\ k = 1,2,\ldots,N \tag{3}$$

$$W^T\Phi(X_k) + w_0 - y_k - \varepsilon \le \xi_k^{'},\ k = 1,2,\ldots,N \tag{4}$$

$$\xi_k \ge 0,\ k = 1,2,\ldots,N \tag{5}$$

$$W^T\Phi(X_k) + w_0 - y_k - \varepsilon \le \xi_k^{'},\ k = 1,2,\ldots,N \tag{6}$$

In the above equations, the error while the real value y_k is larger and smaller than the predicted value $y_k^{'}$ is measured through the non-negative slack variables, ξ_i and $\xi_i^{'}$, respectively. Therefore, [13, 14] formulated a problem of minimizing the empirical risk as follows:

Minimize

$$\frac{1}{2}\|W\|^2 + C\left(\sum_{k=1}^{N}\xi_k + \sum_{k=1}^{N}\xi_k^{'}\right) \tag{7}$$

that must satisfy the constraints as shown in Equations (3)-(6). Notably, the C is a parameter, that balances the complexity and loss, must be specified by the user in advance. the Lagrangian in primal variables can be formulated as follows:

$$L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')$$
$$= \frac{1}{2}W^TW + C\sum_{k=1}^{N}(\xi_k + \xi_k^{'}) - \sum_{k=1}^{N}\lambda_k\left(W^T\Phi(X_k) + w_0 - y_k + \varepsilon + \xi_k\right) -$$
$$\sum_{k=1}^{N}\lambda_k^{'}\left(y_k - W^T\Phi(X_k) - w_0 + \varepsilon + \xi_k^{'}\right) - \sum_{k=1}^{N}(\gamma_k\xi_k + \gamma_k^{'}\xi_k^{'}) \tag{8}$$

where the $\Xi = (\xi_1,\ldots,\xi_N)^T$ and $\Xi^{'} = (\xi_1^{'},\ldots,\xi_N^{'})^T$ are vectors used to represent slack variables; $\Lambda = (\lambda_1,\ldots,\lambda_N)^T$, $\Lambda^{'} = (\lambda_1^{'},\ldots,\lambda_N^{'})^T$, $\Gamma = (\gamma_1,\ldots,\gamma_N)^T$ and $\Gamma^{'} = (\gamma_1^{'},\ldots,\gamma_N^{'})^T$ are the Lagrangian multiplier vectors regarding Equations (3)-(6), respectively. Through taking the partial derivative of L_P with respect to the primal variables to its saddle point, we can obtain the optimality as follows:

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial W} = 0 \Rightarrow W = \sum_{k=1}^{N}(\lambda_k - \lambda_k^{'})\Phi(X_k) \tag{9}$$

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial w_0} = 0 \Rightarrow \sum_{k=1}^{N}(\lambda_k - \lambda_k^{'}) = 0 \tag{10}$$

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial \xi_k} = 0 \Rightarrow \gamma_k = C - \lambda_k \tag{11}$$

$$\frac{\partial L_P(W, w_0, \Xi, \Xi', \Lambda, \Lambda', \Gamma, \Gamma')}{\partial \xi_k^{'}} = 0 \Rightarrow \gamma_k^{'} = C - \lambda_k^{'} \tag{12}$$

Therefore, the simplified dual form L_D can be constructed by substituting Equation (9), (11) and (12) into Equation (8) as follows:

Maximize

$$L_D(\Lambda, \Lambda') = \sum_{k=1}^{N}d_k(\lambda_k - \lambda_k^{'}) - \varepsilon\sum_{k=1}^{N}(\lambda_k + \lambda_k^{'}) -$$
$$\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\lambda_k - \lambda_k^{'})(\lambda_l - \lambda_l^{'})\Phi(X_k)\Phi(X_l) \tag{13}$$

Notably, the inner product $\Phi(X_k)\Phi(X_l)$ can be re-written in the terms of the kernel function $K(X_k, X_l)$ based on the Mercer's Theorem. Hence, Equation (3) can be re-formulated as follows:

Maximize

$$L_D(\Lambda, \Lambda') = \sum_{k=1}^{N}(\lambda_k(d_k - \varepsilon) - \lambda_k^{'}(d_k + \varepsilon)) -$$
$$\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\lambda_k\lambda_l - \lambda_k\lambda_l^{'} - \lambda_k^{'}\lambda_l + \lambda_k^{'}\lambda_l^{'})K(X_k, X_l) \tag{14}$$

subject to

$$\sum_{k=1}^{N}(\lambda_k - \lambda_k^{'}) = 0 \tag{15}$$

$$0 \le \lambda_k \le C,\ k = 1,2,\ldots,N \tag{16}$$

$$0 \le \lambda_k^{'} \le C,\ k = 1,2,\ldots,N \tag{17}$$

There are some commonly used kernel functions, e.g. linear, polynomial (homogeneous), polynomial (inhomogeneous), radial basis function, hyperbolic tangent, etc. Furthermore, the "support" vector is the data (X_k, y_k) whose corresponding λ_i or $\lambda_i^{'}$ is not zero.

Through optimizing the Lagrangian, we can obtain the optimal approximation $\hat{W}$ for the weight vector W can as follows:

$$\hat{W} = \sum_{k=1}^{n_s}(\hat{\lambda}_k - \hat{\lambda}_k^{'})\Phi(X_k) \tag{18}$$

Notably, the index k only runs over the support vector 1 to support vector n_s, where n_s is the number of support vectors.

B. TwoStep Cluster Analysis

Cluster analysis is a grouping methodology that classifies the data into several clusters such that each datum is more similar to the other data in the same cluster than to the data lying in the different cluster. There are lots of approaches that can achieve the cluster analysis, and TwoStep cluster analysis is a famous one among these methodologies. In the TwoStep cluster analysis, a likelihood distance measure is applied while assuming the variables in the cluster model are independent. Furthermore, each continuous variable is assumed to have a normal distribution, as well as the categorical variable has a multinomial distribution. The TwoStep cluster analysis is different from the traditional clustering techniques since it possesses several attractive characteristics. First, it can produce appropriate clusters according to the categorical variables, as well as to the continuous variables. Next, it can automatically determine the optimal number of clusters. Finally, it can analyze data file of a large scale very fast, i.e. efficiently, through constructing a cluster features (CF) tree to summarize the records. The basic execution procedure of TwoStep cluster analysis can be summarized as follows:

Step 1: Build a cluster features (CF) tree

The first case is placed at the root of the CF tree in a leaf node that contains the variable information regarding this case. Each of other cases successively adds to one of the existing nodes or forms a new node where the criterion of its similarity to the existing nodes is utilized to measure the distance. Thus, each node summarizes the variable information regarding the all cases that are clustered in this node. Therefore, the CF tree can provide the summarization about the data file.

Step 2: Group the leaf nodes

The leaf nodes of a CF is then grouped by an agglomerative clustering algorithm to generate various solutions. The optimal number of clusters, as well as the members in each cluster are then determined according to the Schwarz's Bayesian Criterion (BIC) [15] or the Akaike Information Criterion (AIC) [16].

III. Proposed Forecasting Approach

This study proposes an integrated approach for forecasting stock prices with a mechanism of selecting feature variables (critical technical indicators) based on the support vector regression and clustering technique. The procedure of the proposed forecasting approach is conceptually depicted in Figure 1, and is stated in detailed as follows:

Step 1: Collect Stock Trading Data

The daily data of stock trading, including the opening price, highest price, lowest price, closing price, and trade volume, are collected through the stock exchange corporation or off-the-shelf database.

Step 2: Calculate Technical Indicators

The collected stock trading data are used to calculate the technical indicators. Notably, the technical indicators that are beneficial to forecast stock prices are not easy to be determined definitely. The initial technical indicators can be selected according to the previous related researches.

Step 3: Normalize Data

The range of each technical indicator calculated in Step 2 can significantly differ from others. Hence, each technical indicator results in a different degree of impacting the stock prices in the future. Therefore, all technical indicators are normalized into the same range, e.g. 0 to 1, to avoid the situation where certain indicators can affect the future stock prices considerably while others have little influences upon the stock prices in the future. The normalized data are further portioned into two data sets: training data and test data.

Step 4: Build Original Forecasting Model

The support vector regression (SVR) methodology is utilized to build the original forecasting model based on all technical indicators. In other words, all of the technical indicators serve as the input variables, i.e. predictors, as well as the closing price in the future takes as the output variable, i.e. response, for each stock. Notably, the technical indicators used here are chosen in accordance with literatures. Therefore, some technical indicators might interfere with others while forecasting the future stock prices, thus these interfered technical indicators are superfluous. However, we are not easy to judge which technical indicators are critical or superfluous to forecast the future stock prices.

Step 5: Select Feature Variables (Technical Indicators)

In this step, the TwoStep cluster analysis is applied to choose the technical indicators that are crucial for forecasting the future stock prices. The procedures are as follows:

(1) For each technical indicator, the TwoStep cluster analysis is implemented where the two variables, including the technical indicator and future stock closing price, are treated as the clustering variables to segment the data into some groups. The analysis of variance along with the F values of clustering by the TwoStep cluster analysis can then be obtained.

(2) The TwoStep cluster analysis is used again to group all of the acquired F value corresponding to each technical indicator into a certain number of clusters.

(3) The cluster that has the maximum number of members, i.e. F values, is further portioned into a number of classes through the TwoStep cluster analysis.

(4) The procedure (3) is repeated until all of the F values in the class with a largest number of members cannot be further clustered.

(5) Based on the clustering results obtained procedures (2), (3) and (4), the technical indicators can be organized into some categories. The category with a maximum mean of F values is named category 1. The category with second largest mean of F values is then labeled category 2. By the similar methods, each category can then be titled sequentially, such as category 1, category 2, …, category *k*, where category *k* has the smallest average F value among all categories.

Step 6: Build Simplified Forecasting Models

The SVR method is utilized to construct stock price forecasting several simplified models. The output variable (response) is the closing stock price in the future for all simplified models. In the first model, the input variables (predictors) are made up by removing the technical indicators grouped in the last category, i.e. category *k*, from the original full technical indicators. The created forecasting model is named $SVR_{Simp\_all\text{-}Ck}$. Then, discarding the technical indicators grouped in the last two categories, i.e. categories *k*-1 and *k*, from the original full technical indicators. The built SVR forecasting model is called $SVR_{Simp\_all\text{-}Ck\text{-}1}$. Through the similar methods, the other simplified models can be obtained. Finally, forming the input variables by eliminating the technical indicators clustered in groups 2 to *k*, i.e. categories 2, 3, …, *k*-1 and *k*, from all of the technical indicators yields the simplified model $SVR_{Simp\_all\text{-}C2}$.

Step 7: Evaluate Forecasting Performance

Notably, the Steps 4 to 6 are implemented to the training data. To assess the forecasting ability of SVR models to the data that are never encountered before, the models built in Steps 4, 5 and 6 are applied to the test data set up in Step 2. The evaluation measures including the mean squared error (MSE), mean absolute percentage error (MAPE) and coefficient of determination (R^2).

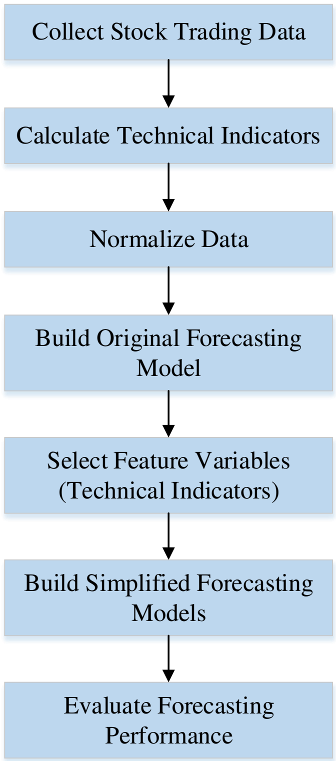


Figure 1. Proposed forecasting approach

IV. CASE STUDY

In this section, three indices of different levels, including the TAIEX (Taiwan Stock Exchange Capitalization Weighted Stock Index), FTSE TWSE (Financial Times Stock Exchange Taiwan Stock Exchange Corporation) Taiwan 50 Index, and Taiwan 2303 Stock (issued by the United Microelectronics Corporation) are used to demonstrate the proposed forecasting approach, as well as validate its feasibility and effectiveness.

A. Collect Stock Trading Data

The daily trading data of TAIEX are gathered from the Taiwan Stock Exchange Corporation. The collected data including the opening price, highest price, lowest price, closing price, and trade volume form 2014/12/29 to 2019/12/09. The data set includes 1,210 daily trading data.

B. Calculate Technical Indicators

The sixteen technical indicators: (1) 10-day moving average (MA_10), (2) 20-day bias, (BIAS_20), (3) moving average convergence/divergence (MACD), (4) 9-day stochastic indicator K (K_9), (5) 9-day stochastic indicator D (D_9), (6) 9-day Williams overbought/oversold index (WMS%R_9), (7) 10-day rate of change (ROC_10), (8) 5-day relative strength index (RSI_5), (9) 24-day commodity channel index (CCI_24), (10) 26-day volume ratio (VR_26), (11) 13-day psychological line (PSY_13), (12) 14-day plus directional indicator (+DI_14), (13) 14-day minus directional indicator (-DI_14), (14) 26-day buying/selling momentum indicator (AR_26), (15) 26-day buying/selling willingness indicator (BR_26), and (16) 10-day momentum (MTM_10), which are in accordance with [17-23], are calculated based on the formulas shown in Appendix A. The collected stock trading data are used to calculate the technical indicators.

C. Normalize Data

The technical indicators calculated in Section 4.2 have different ranges. For example, the K_9 indicator ranges from about 6 to 100 while the smallest and largest values of DI+ index are about 0 and 0.5, respectively. Therefore, the K_9 and DI+ technical indicators are expected to have different degrees for impacting the future stock prices. To balance the effects of all technical indicators upon forecasting the stock prices in the future, each indicator is normalized into the range of 0 to 1 according to its corresponding smallest and largest values. These normalized data are further portioned into two parts including the training and test data whose information is shown in Table 1.

Table 1. Information of the Training and Test Data

Data	Number of Trading Days	Period
Training Data	987	2014/12/29 to 2018/12/28
Test Data	224	2019/01/02 to 2019/12/04

D. Build Original Forecasting Model

The SVR tool is implemented to the training data shown in Table 1. Notably, for each trading day, all of the normalized technical indicators are treated as the input variables and the closing price in the next third day takes as the output variable, respectively. The program LIBSVM developed by [24], where the parameters (C, γ, ε) in SVR are optimized through the grid-search method, as well as the forecasting performance is evaluated by the cross-validation technique, is applied to build the original SVR forecasting model. The information regarding the obtained SVR model is provided in Table 2.

Table 2. Information Regarding the Original SVR Forecasting Model

Model Name	MSE	C	γ	ε
SVR_{ORI}	0.002992	16.0	0.353553390593	0.011048543456

E. Select Feature Variables

For the training data, the first technical indicator MA_10, along with the response, i.e. the closing price in the next third day, are analyzed by the TwoStep cluster analysis. Thus, the analysis of variance along with the F values of clustering can be obtained. Similarly, the TwoStep cluster analysis is also implemented to the remaining fifteen technical indicators to acquire the analysis of variance regarding each indicator. Table 3 summarizes the executional results.

The F values in Table 3 are analyzed by the TwoStep cluster analysis. These F values are segmented into two groups where the first cluster contains one indicators, as well as the second cluster has the remaining fifteen ones as seen in the first three columns of Table 4. Therefore, the fifteen indicators in the second group are further clustered by the TwoStep cluster analysis, thus yields the results as depicted in the first, second, and fourth columns in Table 4. There are two and thirteen technical indicators in the first and second clusters, respectively. So, the F values in the large group are analyzed one more by the TwoStep clsuster analysis and the results are provided in the first, second and fifth columns of Table 4. The eight F values in the largest, i.e. the second cluster, are fed into the TwoStep cluster analysis to produce the clustering results as exhibited in the first, second and sixth columns in Table 4. Since the eight indicators are all clustered into the same group, the TwoStep cluster analysis is terminated. According the clustering #1 to #4 as shown in Table 4, the technical indicators are finally classified into four categories as depicted in the last column of Table 4. The four categories include one, two, five, and eight technical indicators, respectively.

Table 3. F values of TwoStep Cluster Analysis on Each Indicator

Technical Indicator	F value
MA_10	5063.949
BIAS_20	623.067
MACD	724.530
K_9	1418.047
D_9	859.551
WMS%R_9	2020.253
ROC_10	485.010
RSI_5	810.261
CCI_24	703.374
VR_26	520.263
PSY_13	508.175
+DI_14	461.481
-DI_14	551.744
AR_26	433.844
BR_26	269.970
MTM_10	506.876

Table 4. Clustering Results Regarding the F Values

Technical Indicator	F Value	Clustering #1	Clustering #2	Clustering #3	Clustering #4	Category
MA_10	5063.949	1				1
K_9	1418.047	2	1			2
WMS%R_9	2020.253	2	1			2
BIAS_20	623.067	2	2	1		3
MACD	724.530	2	2	1		3
D_9	859.551	2	2	1		3
RSI_5	810.261	2	2	1		3
CCI_24	703.374	2	2	1		3
ROC_10	485.010	2	2	2	1	4
VR_26	520.263	2	2	2	1	4
PSY_13	508.175	2	2	2	1	4
+DI_14	461.481	2	2	2	1	4
-DI_14	551.744	2	2	2	1	4
AR_26	433.844	2	2	2	1	4
BR_26	269.970	2	2	2	1	4
MTM_10	506.876	2	2	2	1	4

F. Build Simplified Forecasting Models

Let the closing stock price in the next third day be the output variable, as well as make the input variables by removing the technical indicators grouped in the last, i.e. fourth, category from the original full technical indicators. The SVR methodology is implemented again to create the simplified forecasting model, called $SVR_{Simp\_all\text{-}C4}$. Then, discarding the technical indicators grouped in the last two categories, i.e. third and fourth categories, from the original full technical indicators. The obtained forecasting model constructed by the SVR is called $SVR_{Simp\_all\text{-}C3}$. Through the similar method, the final simplified model $SVR_{Simp\_all\text{-}C2}$ can also be obtained.

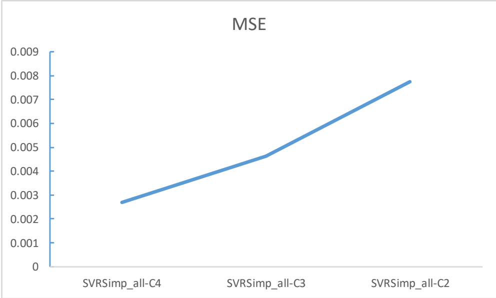
G. Evaluate Forecasting Performance

The SVR models established in Sections 4.4 and 4.6 are applied to the test data made in Section 4.3 for evaluating their forecasting ability regarding the data that are never seen. Table 5 summarizes the evaluation results regarding the training and test data with the measurement including the MSE, MAPE and R^2. Notably, the normalized outputs, the closing stock prices in the next third days, and predictions obtained by the SVR forecasting models must be first de-normalized to their original values in order to calculate the correct MAPEs.

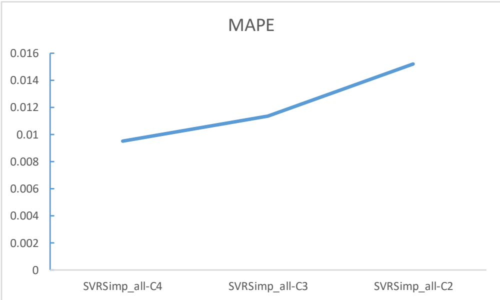
The MSEs, MAPEs and R^2s while implementing the three simplified SVR forecasting models $SVR_{Simp\_all\text{-}C4}$, $SVR_{Simp\_all\text{-}C4}$ and $SVR_{Simp\_all\text{-}C4}$ on the training data are graphically depicted in Figure 2. Compared to the changing from $SVR_{Simp\_all\text{-}C4}$ to $SVR_{Simp\_all\text{-}C3}$, the MSE and MAPE regarding the training data grow more fast when the SVR model changes from $SVR_{Simp\_all\text{-}C3}$ to $SVR_{Simp\_all\text{-}C2}$ based on Figure 2. Furthermore, the R^2 drops with a larger rate if the SVR model switches from $SVR_{Simp\_all\text{-}C3}$ to $SVR_{Simp\_all\text{-}C2}$ while comparing to switching $SVR_{Simp\_all\text{-}C4}$ to $SVR_{Simp\_all\text{-}C3}$. Therefore, the $SVR_{Simp\_all\text{-}C3}$ is thought as the best forecasting model among all simplified models. The forecasting performance regarding $SVR_{Simp\_all\text{-}C3}$ is labeled by the gray base and underline, as well as the evaluation of forecasting for the model SVR_{ORI} is marked by the gray base as shown in Table 5. Although the MSE, MAPE and R^2 of the $SVR_{Simp\_all\text{-}C3}$ are worse than those of the original model SVR_{ORI} for the training data. Especially, the R^2 only decreases 2.2 percentages. However, the simplified model $SVR_{Simp\_all\text{-}C3}$ can improve all of the MSE, MAPE and R^2. Furthermore, the R^2 can be improved drastically from 0.554 to 0.894, i.e. the improvement rate attains 61.3%. Therefore, we can speculate that the feature selection mechanism by using the TwoStep cluster analysis can indeed screen out the technical indicators that are critical to the forecasting of future stock prices, as well as eliminate the unnecessary indicators to avoid the interference to the other important forecasting indicators. Next, the feature selection procedure can significantly reduce the number of technical indicators to prevent the over-fitting regarding the training data thus providing the superior generalization for the unknown test data. In addition, the model $SVR_{Simp\_all\text{-}C3}$ contains only three technical indicators including MA_10, K_9 and WMS%R_9. In other words, the investors can only concentrate on the three indicators, but not pay attention to all sixteen indicators distractingly, and obtain sufficiently good forecasting results. Finally, the forecasting performance regarding the model $SVR_{Simp\_all\text{-}C2}$ drops substantially. The recessive removing of necessary technical indicators that are important to the stock price forecasting might be the reason.

Table 5. Summary of Forecasting Performance

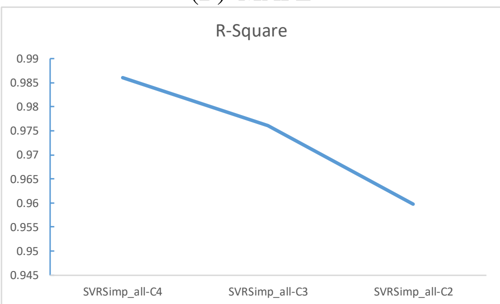
Model	Training data			Test data		
	MSE	MAPE	R^2	MSE	MAPE	R^2
SVR_{ORI}	0.000364	0.003051	0.998098	0.027556	0.022947	0.554403
$SVR_{Simp\_all\text{-}C4}$	0.002693	0.009506	0.986015	0.014514	0.016237	0.748355
$SVR_{Simp\_all\text{-}C3}$	0.004620	0.011377	0.976074	0.005314	0.011273	0.894006
$SVR_{Simp\_all\text{-}C2}$	0.007752	0.015225	0.959775	0.030969	0.022797	0.467561



(A) MSE



(B) MAPE



(C) R^2

Figure 2. Graphical Depictions of MSEs, MAPEs and R^2s of the Training Data

The proposed forecasting approach is also implemented to the FTSE TWSE (Financial Times Stock Exchange Taiwan Stock Exchange Corporation) Taiwan 50 Index, and Taiwan 2303 Stock (issued by the United Microelectronics Corporation). The obtained results by applying the TwoStep cluster analysis to the F values, shown in the analysis of variance for each indicator, are depicted in Tables 6 and 7. The SVR methodology is then utilized to build the simplified forecasting models. The forecasting performance of applying the original and simplified SVR models to the training and test data is summarily shown in Tables 8 and 9, as well as graphed in Figures 3 to 4. For the FTSE TWSE Taiwan 50 Index, the simplified model $SVR_{Simp\_all\text{-}C4}$ is selected as the best forecasting model since the MAPE significantly decrease faster when the model changes from $SVR_{Simp\_all\text{-}C4}$ to $SVR_{Simp\_all\text{-}C3}$ based on Figures 3. The forecasting performance regarding $SVR_{Simp\_all\text{-}C4}$ and SVR_{ORI} is evaluated and shown in Table 8. The evaluation results of $SVR_{Simp\_all\text{-}C4}$ are marked by the gray base and underline, as well as the evaluation of forecasting performance for the model SVR_{ORI} is indicated by the gray base as shown in Table 8. Based on Table 8, the MSE, MAPE and R^2 regarding the $SVR_{Simp\_all\text{-}C4}$ are slightly worse than those corresponding to the original model SVR_{ORI} when evaluating the training data. However, the simplified model $SVR_{Simp\_all\text{-}C3}$ can provide better forecasting performance in terms of the MSE, MAPE and R^2.

As regards the Taiwan 2303 Stock, the original model is also included to diagram the Figure 4 for the purpose of comparison for choosing the best forecasting model since there are only two simplified models. On the basis of Figure 4, the $SVR_{Simp\_all\text{-}C3}$ is considered as the best forecasting model because the forecasting performance descends more obviously when switching the model from $SVR_{Simp\_all\text{-}C3}$ to $SVR_{Simp\_all\text{-}C2}$. The selected $SVR_{Simp\_all\text{-}C3}$ and SVR_{ORI} are evaluated and Table 9 summarizes the assessing results. The gray base and gray base with an underline labels the forecasting performance regarding the $SVR_{Simp\_all\text{-}C3}$ and SVR_{ORI}, respectively. In accordance with Table 9, The simplified forecasting model $SVR_{Simp\_all\text{-}C3}$ cannot significantly improve the forecasting performance in terms of MSE, MAPE and R^2. However, the two models $SVR_{Simp\_all\text{-}C3}$ and SVR_{ORI} can provide near forecasting performance. Furthermore, the simplified model $SVR_{Simp\_all\text{-}C3}$ merely contains nine technical indicators while the original model SVR_{ORI} has sixteen ones. In other words, an investor can forecast the future prices of Taiwan 2303 Stock by using only nine technical indicators to obtain a forecasting accuracy that is almost the same to that yield through utilizing all of the sixteen technical indicators. Hence, an investor can save more time, cost, and effort in constructing a forecasting model.

Table 6. Clustering Results Regarding the F Values (FTSE TWSE Taiwan 50 Index)

Technical Indicator	F Value	Clustering #1	Clustering #2	Clustering #3	Clustering #4	Category
MA_10	5272.469	1				1
K_9	1109.098	2	2			2
D_9	1073.215	2	2			2
WMS%R_9	1958.714	2	2			2
MACD	649.477	2	1	2	1	3
ROC_10	394.225	2	1	2	1	3
RSI_5	509.276	2	1	2	1	3
CCI_24	460.516	2	1	2	1	3
VR_26	310.341	2	1	2	1	3
PSY_13	563.943	2	1	2	1	3
+DI_9	436.467	2	1	2	1	3
-DI_9	477.122	2	1	2	1	3
AR_26	455.759	2	1	2	1	3
BR_26	702.527	2	1	2	1	3
BIAS_20	116.836	2	1	1		4
MTM_10	1.287	2	1	1		4

Table 7. Clustering Results Regarding the F Values (Taiwan 2303 Stock)

Technical Indicator	F Value	Clustering #1	Clustering #2	Clustering #3	Category
MA_10	3359.059	1			1
MACD	887.85	2	2	1	2
K_9	772.17	2	2	1	2
D_9	979.88	2	2	1	2
WMS%R_9	887.85	2	2	1	2
RSI_5	714.54	2	2	1	2
CCI_24	431.21	2	2	1	2
+DI_9	610.84	2	2	1	2
-DI_9	825.14	2	2	1	2
BIAS_20	36.2	2	1		3
ROC_10	88.9	2	1		3
VR_26	48.6	2	1		3
PSY_13	146.29	2	1		3
AR_26	70.4	2	1		3
BR_26	212.45	2	1		3
MTM_10	505.59	2	1		3

Table 8. Summary of Forecasting Performance (FTSE TWSE Taiwan 50 Index)

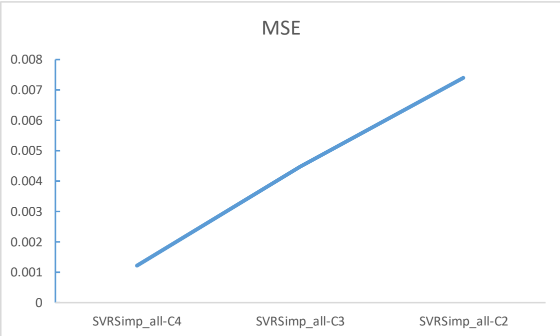
Model	Training data			Test data		
	MSE	MAPE	R^2	MSE	MAPE	R^2
SVR_{ORI}	0.000900	0.005276	0.995283	0.018121	0.021280	0.778113
$SVR_{Simp\_all\text{-}C4}$	0.001201	0.005638	0.993710	0.015819	0.020096	0.816436
$SVR_{Simp\_all\text{-}C3}$	0.004473	0.012852	0.976565	0.008183	0.014665	0.895231
$SVR_{Simp\_all\text{-}C2}$	0.007406	0.016586	0.961580	0.028254	0.024310	0.634948

Note: The gray base indicates the information of the original forecasting model, and the gray based along with an underline represents the information regarding the best selected simplified forecasting model.

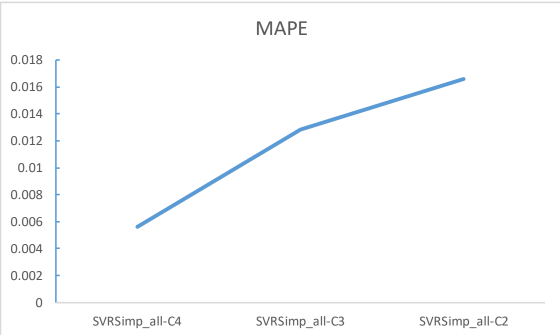
Table 9. Summary of Forecasting Performance (Taiwan 2303 Stock)

Model	Training data			Test data		
	MSE	MAPE	R^2	MSE	MAPE	R^2
SVR_{ORI}	0.001159	0.008160	0.994447	0.008698	0.025365	0.909626
$SVR_{Simp\_all\text{-}C3}$	0.002783	0.012329	0.986744	0.008728	0.023794	0.904069
$SVR_{Simp\_all\text{-}C2}$	0.011117	0.026556	0.946844	0.009456	0.025268	0.904057

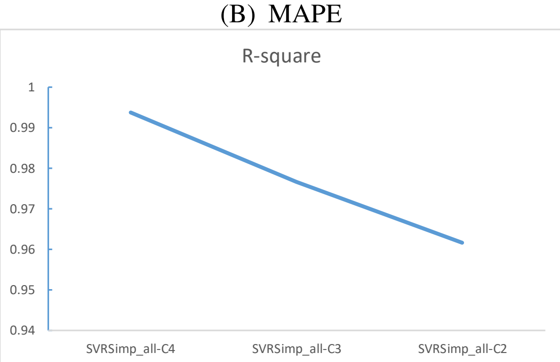
Note: The gray base indicates the information of the original forecasting model, and the gray based along with an underline represents the information regarding the best selected simplified forecasting model.



(A) MSE

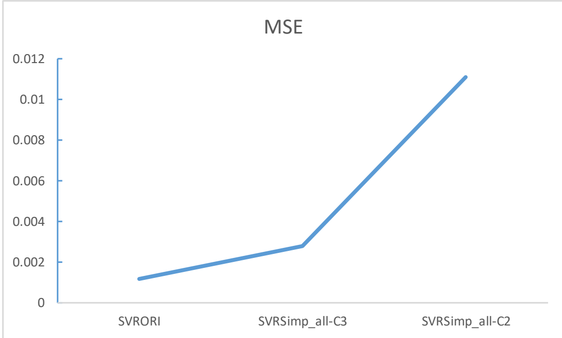


(B) MAPE

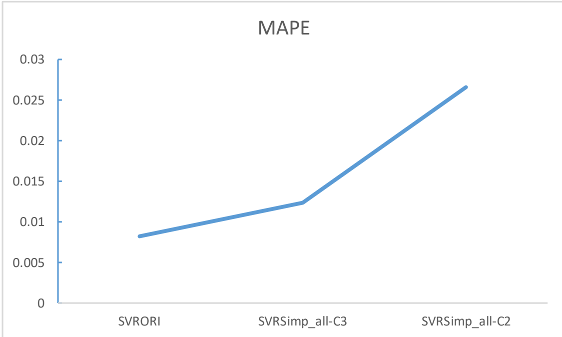


(C) R^2

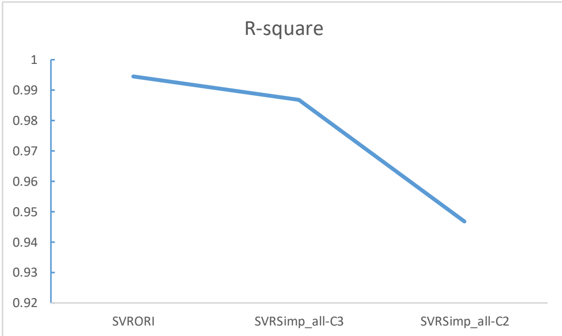
Figure 3. Graphical Depictions of MSEs, MAPEs and R^2s of the Training Data (FTSE TWSE Taiwan 50 Index)



(A) MSE



(B) MAPE



(C) R^2

Figure 4. Graphical Depictions of MSEs, MAPEs and R^2s of the Training Data (Taiwan 2303 Stock)

V. CONCLUSIONS

This study proposes an integrated procedure for resolving the problems of forecasting stock prices based on the support vector regression and TwoStep cluster analysis. The TwoStep cluster analysis is used to screen out the technical indicators that are crucial to the future stock prices, as well as the support vector regression is applied to build the models for forecasting the stock prices in the future. The feasibility and effectiveness of the proposed approach are demonstrated by a case study that aims to resolve the stock price forecasting problems in three different levels' indices that include the TAIEX, FTSE TWSE Taiwan 50 Index, and Taiwan 2303 Stock. According to the experimental results, the designed feature selection mechanism can effectively find out the technical indicators that are helpful to forecast the future stock prices. At the same time, this mechanism can remove the superfluous indicators that will interfere the other critical technical indicators in forecasting the future stock prices. Next, the feature selection procedure can significantly lower the total number of technical indicators thus preventing the situation of over-fitting to the training data. Hence, the well-established model can provide a better ability for generalizing to the unknown test data. In addition, the selected best simplified forecasting model cannot always yield the superior forecasting performance for the training and test data. However, the investors can only concentrate on the less technical indicators, but not pay attention to all of the indicators distractingly, and obtain a forecasting result with a sufficiently high accuracy. In other words, an investor can save more time, cost, and effort while building a forecasting model. Thus investors can spend more concentration on the financial analysis and trading strategies research for the target investment stocks.

APPENDIX A

Notations:

i: the day i

HP_i: the highest price of day i

LP_i: the lowest price of day i

OP_i: the opening price of day i

CP_i: the closing price of day i

TV_i: the trade value of day i

1. MA_10: 10-day moving average

The 10-day moving average is the mean price of a security over the most recent 10 days, and is calculated by:

$$\mathrm{MA\_10}_i = \frac{\sum_{j=i-9}^{i} CP_j}{10} \tag{19}$$

2. BIAS_20: 20-day bias

The 20-day bias is the deviation between the closing price and the 20-day moving average (MA_20), and is calculated by:

$$\mathrm{BIAS\_20}_i = \frac{CP_i - \mathrm{MA\_20}_i}{\mathrm{MA\_20}_i} \tag{20}$$

3. MACD: moving average convergence/divergence

The moving average convergence/divergence is a momentum indicator that shows the relationship between two moving averages. First, define the demand index (DI) as:

$$\mathrm{DI}_i = (HP_i + LP_i + 2 \times CP_i)/4 \tag{21}$$

Next, define the 12-day exponential moving average (EMA_12) and 26-day exponential moving average (EMA_26) as:

$$\mathrm{EMA\_12}_i = \frac{11}{13} \times \mathrm{EMA\_12}_{i-1} + \frac{2}{13} \times \mathrm{DI}_i \tag{22}$$

and

$$\mathrm{EMA\_26}_i = \frac{25}{27} \times \mathrm{EMA\_26}_{i-1} + \frac{2}{27} \times \mathrm{DI}_i \tag{23}$$

, respectively. Then, the difference between EMA_12 and EMA_26 can be calculated by:

$$\mathrm{DIF}_i = \mathrm{EMA\_12}_i - \mathrm{EMA\_26}_i \tag{24}$$

Hence, the moving average convergence/divergence can be defined by:

$$\mathrm{MACD}_i = \frac{8}{10} \times \mathrm{MACD}_{i-1} + \frac{2}{10} \times \mathrm{DIF}_i \tag{25}$$

4. K_9: 9-day stochastic indicator K

The 9-day stochastic indicator K is defined as:

$$\mathrm{K\_9}_i = \frac{2}{3} \times \mathrm{K\_9}_{i-1} + \frac{1}{3} \times \frac{CP_i - LP\_9_i}{HP\_9_i - LP\_9_i} \times 100 \tag{26}$$

where, LP_9_i and HP_9_i are the lowest and highest prices of the previous 9 days, i.e. days i, i -1, …, i -7 and i-8, respectively.

5. D_9: 9-day stochastic indicator D

The 9-day stochastic indicator D is defined as:

$$\mathrm{D\_9}_i = \frac{2}{3} \times \mathrm{D\_9}_{i-1} + \frac{1}{3} \times \mathrm{K\_9}_i \tag{27}$$

where, $\mathrm{K\_9}_i$ is the 9-day stochastic indicator K of day i, as previously defined.

6. WMS%R_9: 9-day Williams overbought/oversold index

The 9-day Williams overbought/oversold index is a momentum indicator that measures overbought and oversold levels, and is calculated by:

$$\mathrm{WMS\%R\_9}_i = \frac{HP\_9_i - CP_i}{HP\_9_i - LP\_9_i} \tag{28}$$

where, LP_9_i and HP_9_i are the lowest and highest prices of the previous 9 days, i.e. days i, i -1, …, i -7 and i-8, respectively.

7. ROC_10: 10-day rate of change

The 10-day rate of change measures the percent changes of the current price relative to the price of 10 days ago, and is calculated by:

$$\mathrm{ROC\_10}_i = \frac{CP_i - CP_{i-10}}{CP_{i-10}} \times 100 \tag{29}$$

8. RSI_5: 5-day relative strength index

The relative strength index is a momentum oscillator that compares the magnitude of recent gains to the magnitude of recent losses. First, define the gain of day i as:

$$\mathrm{G}_i = \begin{cases} CP_i - CP_{i-1} & \textit{if } CP_i > CP_{i-1} \\ 0 & \textit{otherwise} \end{cases} \tag{30}$$

Similarly, the loss of day i is calculated by:

$$\mathrm{L}_i = \begin{cases} CP_i - CP_{i-1} & \textit{if } CP_i < CP_{i-1} \\ 0 & \textit{otherwise} \end{cases} \tag{31}$$

Next, the 5-day average gain (AG_5) and 5-day average loss (AL_5), which can be calculated by:

$$\mathrm{AG\_5}_i = \frac{4}{5} \times \mathrm{AG\_5}_{i-1} + \frac{1}{5} \times \mathrm{G}_i \tag{32}$$

and

$$\mathrm{AL\_5}_i = \frac{4}{5} \times \mathrm{AL\_5}_{i-1} + \frac{1}{5} \times \mathrm{L}_i \tag{33}$$

, respectively. Hence, the 5-day relative strength index can be defined by:

$$\mathrm{RSI\_5}_i = \frac{\mathrm{AG\_5}_i}{\mathrm{AG\_5}_i + \mathrm{AL\_5}_i} \times 100 \tag{34}$$

9. CCI_24: 24-day commodity channel index

The commodity channel index is used to identify cyclical turns in commodities. First, define the typical price (TP) as:

$$\mathrm{TP}_i = \frac{HP_i + LP_i + CP_i}{3} \tag{35}$$

Next, calculate the 24-day simple moving average of the typical price (SMATP_24) by:

$$\mathrm{SMATP\_24}_i = \frac{\sum_{j=i-23}^{i} \mathrm{TP}_j}{24} \tag{36}$$

Then, the 24-day mean deviation (MD_24) can be calculated by:

$$\mathrm{MD\_24}_i = \frac{\sum_{j=i-23}^{i} |\mathrm{TP}_j - \mathrm{SMATP\_24}_i|}{24} \tag{37}$$

Hence, the 24-day commodity channel index can be defined as:

$$\mathrm{CCI\_24}_i = \frac{\mathrm{TP}_i - \mathrm{SMATP\_24}_i}{0.015 \times \mathrm{MD\_24}_i} \tag{38}$$

10. CCI_24: 24-day commodity channel index

The 26-day volume ratio is defined by:

$$\mathrm{VR\_26}_i = \frac{\mathrm{TVU\_26}_i - \mathrm{TVF\_26}_i/2}{\mathrm{TVD\_26}_i - \mathrm{TVF\_26}_i/2} \times 100\% \tag{39}$$

where, $\mathrm{TVU\_26}_i$, $\mathrm{TVD\_26}_i$, and $\mathrm{TVF\_26}_i$ represent the total trade volumes of stock prices rising, falling, and holding, respectively, from the previous 26 days, i.e. days *i*, *i* -1, …, *i* -24 and *i*-25.

11. PSY_13: 13-day psychological line

The psychological line is a volatility indicator based on the number of time intervals that the market was up during the preceding period. The 13-day psychological line is defined by:

$$\mathrm{PSY\_13}_i = \frac{\mathrm{TDU\_13}_i}{13} \times 100\% \tag{40}$$

where $\mathrm{TDU\_13}_i$ is the total number of days regarding stock price rises of the previous 13 days, i.e. days *i*, *i* -1, …, *i* -11 and *i*-12.

12. +DI_14: 14-day plus directional indicator

First, define plus directional movement (+DM) and minus directional movement (-DM) as:

$$+\mathrm{DM}_i = HP_i - HP_{i-1} \tag{41}$$

and

$$-\mathrm{DM}_i = LP_{i-1} - LP_i \tag{42}$$

, respectively. The plus true directional movement (+TDM) can be calculated by:

$$+\mathrm{TDM}_i = \begin{cases} +\mathrm{DM}_i & \text{if } +\mathrm{DM}_i > -\mathrm{DM}_i \text{ and } +\mathrm{DM}_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{43}$$

Similarly, the minus true directional movement (-TDM) can be calculated by:

$$-\mathrm{TDM}_i = \begin{cases} -\mathrm{DM}_i & \text{if } +\mathrm{DM}_i < -\mathrm{DM}_i \text{ and } -\mathrm{DM}_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

Hence, the 14-day plus directional movement (+DM_14) can be calculated by:

$$+\mathrm{DM\_14}_i = \frac{13}{14} \times (+\mathrm{DM\_14}_{i-1}) + \frac{1}{14} \times (+\mathrm{TDM}_i) \tag{45}$$

Similarly, the 14-day minus directional movement (-DM_14) can be calculated by:

$$-\mathrm{DM\_14}_i = \frac{13}{14} \times (-\mathrm{DM\_14}_{i-1}) + \frac{1}{14} \times (-\mathrm{TDM}_i) \tag{46}$$

Next, define the true range (TR) as:

$$\mathrm{TR}_i = Max\{HP_i - LP_i, |HP_i - CP_{i-1}|, |LP_i - CP_{i-1}|\} \tag{47}$$

The 14-day true range (TR_14) can be calculated by:

$$\mathrm{TR\_14}_i = \frac{13}{14} \times \mathrm{TR\_14}_{i-1} + \frac{1}{14} \times \mathrm{TR}_i \tag{48}$$

Therefore, the 14-day plus directional indicator can be defined as:

$$+\mathrm{DI\_14}_i = \frac{+\mathrm{DM\_14}_i}{\mathrm{TR\_14}_i} \tag{49}$$

13. -DI_14: 14-day minus directional indicator

The 14-day minus directional indicator is defined as:

$$-\mathrm{DI\_14}_i = \frac{-\mathrm{DM\_14}_i}{\mathrm{TR\_14}_i} \tag{50}$$

where, $-\mathrm{DM\_14}_i$ and $\mathrm{TR\_14}_i$ are the 14-day minus directional movement and 14-day true range of day *i*, respectively, as previously defined.

14. AR_26: 26-day buying/selling momentum indicator

The 26-day buying/selling momentum indicator is defined as:

$$\mathrm{AR\_26}_i = \frac{\sum_{j=i-25}^{i} (HP_j - OP_j)}{\sum_{j=i-25}^{i} (OP_j - LP_j)} \tag{51}$$

15. BR_26: 26-day buying/selling willingness indicator

The 26-day buying/selling willingness indicator is defined as:

$$\mathrm{BR\_26}_i = \frac{\sum_{j=i-25}^{i} (HP_j - CP_{j-1})}{\sum_{j=i-25}^{i} (CP_{j-1} - LP_j)} \tag{52}$$

16. MTM_10: 10-day momentum

The 10-day momentum measures the price changes of a security during a period of 10 days, and is calculated by:

$$\mathrm{MTM\_10}_i = CP_i - CP_{i-10} \tag{53}$$

ACKNOWLEDGMENT

The author would like to thank the Minghsin University of Science and Technology, Taiwan, ROC for partially supporting this research under Contract No. MUST-109BA-01.

REFERENCES

[1] Wu, C., Luo, P., Li, Y. L., Wang, L., & Chen, K. (2015). Stock price forecasting: hybrid model of artificial intelligent methods. *Inzinerine Ekonomika-Engineering Ecomics,* *26*(1), 40-48. doi: 10.5755/j01.ee.26.1.3836

[2] Laboissiere, L. A., Fernandes, R. A. S., & Lage, G. G. (2015). Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. *Applied Soft Computing, 35,* 66-74. doi: 10.1016/j.asoc.2015.06.005

[3] Wang, J. Z., Hou, R., Wang, C., & Shen, L. (2016). Improved v -Support vector regression model based on variable selection and brain storm optimization for stock price forecasting. *Applied Soft Computing, 49,* 164-178. doi: 10.1016/j.asoc.2016.07.024

[4] Lahmiri, S. (2016). Intraday stock price forecasting based on variational mode decomposition. *Journal of Computational Science, 12,* 23-27. doi: 10.1016/j.jocs.2015.11.011

[5] Arevalo, R., Garcia, J., Guijarro, F., & Peris, A. (2017). A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting. *Expert Systems with Applications, 81,* 177-192. doi: 0.1016/j.eswa.2017.03.028

[6] Pyo, S., Lee, J., Cha, M., & Jang, H. (2017). Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets. *Plos One*, *12*(11), e0188107. doi: 10.1371/journal.pone.0188107

[7] Wang, W. N. (2018). A big data framework for stock price forecasting using fuzzy time series. *Multimedia Tools and Applications, 77*(8), 10123-10134. doi: 0.1007/s11042-017-5144-5
[8] Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications, 103,* e0212320. doi: 10.1016/j.eswa.2018.03.002
[9] Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *Plos One, 14*(2), e0212320. doi: 10.1371/journal.pone.0212320
[10] Zhang, J., Teng, Y. F.,& Chen, W. (2015). Support vector regression with modified firefly algorithm for stock price forecasting. *Applied Artificial Intelligence, 49*(5), 1658-1674. doi: 10.1007/s10489-018-1351-7
[11] Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. doi: 10.1007/BF00994018
[12] Drucker, H., Burges, C. C., Kaufman, L., Smola, A. J., & Vapnik, V. N. (1997). Support Vector Regression Machines. In M. C. Mozer, M. I. Jordan & T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9* (pp. 155-161). Cambridge, MA: MIT Press.
[13] Vapnik, V. N. (1998). *Statistical Learning Theory.* New York: Wiley.
[14] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* New York: Springer-Verlag.
[15] Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag.
[16] Akaike, H (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21*(1), 243-247. doi: 10.1007/BF02532251
[17] Kim, K.-J., & Han, I. (2015). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications, 19*(2), 125-132. doi: 10.1016/S0957-4174(00)00027-0
[18] Kim, K.-J., & Lee, W. B. (2004). Stock market prediction using artificial neural networks with optimal feature transformation. *Neural Computing & Applications, 13*(3), 255-260. doi: 10.1007/s00521-004-0428-x
[19] Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., Tsang, J., Koong, K., & Wong, T.-L. (2007). Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence, 20*(4), 453-461. doi: 10.1016/j.engappai.2006.10.002
[20] Chang, P.-C., & Liu, C.-H. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Applications, 34*(1), 135-144. doi: 10.1016/j.eswa.2006.08.020
[21] Ince, H., & Trafalis, T. B. (2008). Short term forecasting with support vector machines and application to stock price prediction. *International Journal of General Systems, 37*(6), 677-687. doi: 10.1080/03081070601068595
[22] Huang, C.-L., & Tsai, C.-Y. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications, 32*(2), 1529-1539. doi: 10.1016/j.eswa.2007.11.062
[23] Lai, R. K., Fan, C.-Y., Huang, W.-H., & Chang, P.-C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications, 36*(2), 3671-3773. doi: 10.1016/j.eswa.2008.02.025
[24] Chang, C.-C., & Lin, C.-J. (2011). LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology,* 2:27:1-27:27.

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.