Scientia Iranica E (2018) 25(3), 1629–1640

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
http://scientiairanica.sharif.edu

The strategy of investment in the stock market using modified support vector regression model

C.-H. Huang[a], F.-H. Yang[b], and C.-P. Lee[c,*]

a. *Ph.D. Program in Management, Da-Yeh University, No.168, University Rd., Dacun, Changhua 51591, Taiwan (R.O.C.).*
b. *Department of International Business Management, Da-Yeh University, No.168, University Rd., Dacun, Changhua 51591, Taiwan (R.O.C.).*
c. *Department of Maritime Information and Technology, National Kaohsiung Marine University, No.482, Zhongzhou 3rd Rd., Qijin Dist., Kaohsiung City 805, Taiwan.*

Received 21 April 2015; received in revised form 15 December 2016; accepted 21 January 2017

KEYWORDS
Correlation coefficient;
Support vector regression model;
Hybrid model;
Time series data forecasting;
Stock indices.

Abstract. Stock indices forecasting has become a popular research issue in recent years. Although many statistical time series models have been applied to stock indices forecasting, they are limited to certain assumptions. Accordingly, the traditional statistical time series models might not be suitable for forecasting real-life stock indices data. Hence, this paper proposes a novel forecasting model to assist investors in determining a strategy for investments in the stock market. The proposed model is called the modified support vector regression model, which is composed of the correlation coefficient method, sliding window algorithm, and support vector regression model. The results show that the forecasting accuracy of the proposed model is more stable than those of the existing models in terms of average and standard deviation of the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). Accordingly, the proposed model would be used to assist investors in determining a strategy for investing in stocks.

1. Introduction

In recent years, with the development of the economy, many people seek to earn more money and profits by investing in financial markets. Accordingly, there has been much discussion about the best strategy for investing. Moreover, due to the fact that one of the most popular investment objects is the stock market, stock indices forecasting has become more and more important. Recently, many traditional statistical time series models have been successfully applied to the stock indices forecasting issues, such as the autoregressive moving average (ARMA) model and autoregressive integrated moving average (ARIMA) model [1]. However, it is difficult to predict stock indices accurately because stock indices are often affected by many external factors. Furthermore, most of the traditional statistical time series models are limited due to the fact that many of their assumptions might not be fitted to the real life stock indices data. Hence, many researchers have proposed various novel forecasting models to predict financial problems and assist investors in determining an investment strategy without financial knowledge.

Many researchers have utilized popular traditional forecasting models to propose suitable and novel models for forecasting financial issues, such as the

*. *Corresponding author. Tel.: +886 7 8100888 ext. 25322*
E-mail addresses: d0233013@cloud.dyu.edu.tw (C.-H. Huang); leonard@mail.dyu.edu.tw (F.-H. Yang); cplee@nkust.edu.tw (C.-P. Lee)

doi: 10.24200/sci.2017.4440

statistical time series, Fuzzy Time Series (FTS), Artificial Neural Network (ANN), and Support Vector Regression (SVR) models. For example, Tseng et al. [2] proposed a hybrid statistical time series model to forecast the option price. In the model of Tseng et al., a grey-exponential generalized autoregressive conditional heteroscedasticity (Grey-EGARCH) model is developed to decrease the stochastic and nonlinearity of error term sequence, and then the predictability of the option-pricing model is elevated further. Moreover, the FTS model has been gaining popularity in financial forecasting such as the stock price indices forecasting [3,4], foreign exchange rates forecasting [5], option price forecasting [6-8], and futures exchange indices forecasting [9] in recent years. For example, Chen [10] modified the traditional FTS model and proposed a novel FTS model. The novel FTS model was called the high-order FTS model. The high-order FTS model was different from the traditional FTS model. The high-order FTS model used many days to construct the trend (the situation) of time series data for enhancing the prediction performance of the high-order FTS model.

For ANN model, Leu et al. [7] proposed a hybrid ANN model, which is called the fuzzy time series-based neural network model, to predict option prices in 2010. The model by Leu et al. combined the FTS and ANN models to improve the problem in finding the same fuzzy sets from the historical dataset. Liang et al. [11] proposed a simple method based on an ANN model. Liang et al. [11] first used four different linear models to predict option prices. Subsequently, the four predicted option prices are fed into the ANN model to get the final predicted option price. Compared to the traditional time series models, the ANN hybrid models significantly showed better performance under several performance measures.

Furthermore, the SVR model has been widely used to solve many financial forecasting problems in recent years. For example, Lee et al. [6] proposed a hybrid SVR model to predict option prices. In the model by Lee et al. [6], the Least Square Support Vector Regression (LSSVR) model is used to assist the FTS model with option price forecasting, and the bootstrap method is used to enhance the prediction accuracy because the sample size of the option price data might be small. Furthermore, Huang [12] first employed the SVR model to generate surrogates for actual stock returns. Subsequently, the top-ranked stocks could be selected to form a portfolio. Finally, the Genetic Algorithm (GA) was employed for the optimization of model parameters, and feature selection was used to acquire optimal subsets of input variables to the SVR model.

Although those proposed models were successfully applied to time series data forecasting, some problems still existed for dealing with the real-life time series data. Hence, the main aim of this paper is to counter these problems and keep the advantages of the existing proposed model. We, therefore, propose a novel hybrid model, which is called "the Modified Support Vector Regression model" (MSVR model), for stock indices forecasting. The MSVR model is composed of the statistical method and data mining technology. First, the MSVR model shares the advantages of the FTS model, e.g. the MSVR model uses many previous days to construct the trends of time series data and utilizes the trends to forecast the next day's value. Although Leu et al. [5] used the Euclidean distance to search through similar trends for forecasting, the Euclidean distance could not be used to measure the daily change. Hence, the MSVR model uses the correlation coefficient method to search for similar trends in the historical database. Then, the searched similar trends are used as the training samples to construct the training model and predict the stock index on the next day by the trained MSVR model.

The remainder of this paper is organized as follows. Section 2 briefly reviews the definition of time series data, as well as definitions of the FTS model, ANN model, and SVR model. Section 3 introduces the procedure of the MSVR model. Section 4 introduces the dataset and reports the experiment results. Section 5 gives the conclusions.

2. The related works

Due to the fact that the time series data, FTS model, ANN model, and SVR model play important roles in this paper, they will be introduced briefly in this section.

2.1. Time series data

Generally, a time series data is a sequence of data points, namely the tth data point might be affected by the previous data points. Stock indices, stock prices, or other financial investing objects are the well-known time series data. If a method is used to predict the future values (for example, the next day's value) based on the previously observed time series values, the method is called the forecasting method for time series data. For example, the ARMA, ARIMA, and generalized autoregressive conditional heteroskedasticity (GARCH) models are well-known statistical methods for forecasting time series data [1,6,13]. Moreover, data mining models are also widely used to predict time series data, such as the ANN [1,14,15] and SVR models [6,16,17].

Traditionally, the time series data often used the value of the previous data points to predict that of the next data point. To enhance the prediction accuracy, many methods used many values of the

previously observed time series data for forecasting the future values. The value of the previously observed time series data is called a "situation" in this paper. However, it is difficult to determine the number of previously observed time series values. To simplify, this paper defines the situation caused by three days (three previously observed time series values). In our previous study [6], we defined the situation of time series data with a 3-day set, which can be described as 9 types of situations, as shown in Figure 1. Due to the fact that a 3-day set has two continuous fluctuation days in which a fluctuation has three type results (rise, fall, and unchanged), there are $3 \times 3 = 9$ types of situations in a 3-day set. Consequently, the situation with a 3-day set can be represented by (day_1, day_2, day_3) in this paper. Furthermore, for forecasting future value on day t, it can be represented by $(\text{day}_{t-3}, \text{day}_{t-2}, \text{day}_{t-1}) \rightarrow \text{day}_t$, which is called a "trend" in this paper. A trend is composed of Left-Hand (LHS) and Right-Hand Sides (RHS). For example, the LHS of the trend is: day_{t-3}, day_{t-2}, day_{t-1}; the RHS of the trend is the next day of the situation (day_t).

According to the literature [3,5-9], when the situations of two time series data sets are the same, they might have similar values in the future. Accordingly, this characteristic can be used to forecast the value in the future. For instance, one of the important issues for time series data forecasting is to search through the same situations of trends in the historical database for forecasting. However, it is difficult to search through the same situations of trends in the historical database for forecasting, especially for the high-order time series data. When the same situations of trends cannot be searched, many forecasting models might not be trained for forecasting the values in the future. For this reason, this paper uses the correlation coefficient method for searching the similar trends to avoid any inefficiency or incapability on the part of the model in finding the same situations. Subsequently, the searched similar trends are used to train the forecasting model and predict the value in the future.

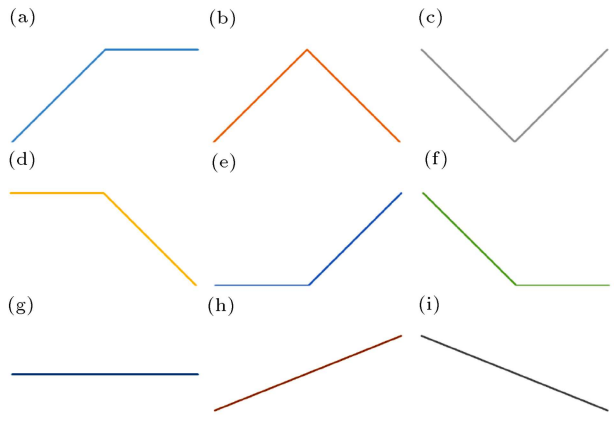

Figure 1. Nine types of situations of time series data with a 3-day set.

2.2. Fuzzy time series model

The FTS model, which is based on the fuzzy logic, is used for forecasting problems [9,13,18]. Song and Chissom first applied it for forecasting enrollments at the University of Alabama [19,20]. Recently, the FTS model has been widely used in financial issues [5,8]. According to the literature [3,5,8,17-21], the following definitions are given to a FTS model.

Definition 1. Suppose that $Y(t)(t = \cdots, 0, 1, 2, \cdots)$, a subset of R, is the universe of discourse, and fuzzy sets $f_i(t)(i = 1, 2\cdots)$ are defined based on the universe of discourse. If $F(t)$ is a collection of $f_i(t)$, $F(t)$ is called a fuzzy time series defined on $Y(t)$.

Definition 2. If $F(t)$ is only caused by $F(t-1)$, then the relationship can be represented as in $F(t) = F(t-1) \circ R(t, t-1)$ where $R(t, t-1)$ and '$\circ$' denote a fuzzy relationship between $F(t)$ and $F(t-1)$ and the max-min composition, respectively. It can be represented by $F(t-1) \rightarrow F(t)$.

Definition 3. If $F(t)$ is caused by $F(t-1)$, $F(t-2), \cdots$, and $F(t-n)$, $F(t)$ is called a 1-factor n-order fuzzy time series, and it can be represented by $F(t-n), \cdots, F(t-2), F(t-1) \rightarrow F(t)$.

2.3. Artificial neural network model

The concept of the ANN model was first introduced in the 1950s [22]. Now, hundreds of different ANN models have been developed. Among them, the feed-forward neural network and radial basis function neural network are widely well known [5,14]. The framework of an ANN model, as shown in Figure 2, contains three layers: the input, hidden, and output layers. Recently, the radial basis function neural network model has been successfully applied to many financial issue applications [8,14,22]. We use the radial basis function neural network model as the main comparison model in this paper.

2.4. Support vector regression model

The Support Vector Machine (SVM) model, as a well-known classification model, has been extended and successfully applied to solve a nonlinear regression estimation problem in 1996. The extended SVM model, proposed by Drucker et al. [23], is called the SVR model. The SVR model has been widely used in many financial problems [6,12,16,17].

Suppose that x_i and y_i are an input variable and a corresponding output variable, respectively, of a dataset (x_i, y_i) $(i = 1, 2, \cdots, l;\ x_i \in R^d;\ y \in R)$. The goal of the SVR model is to search for

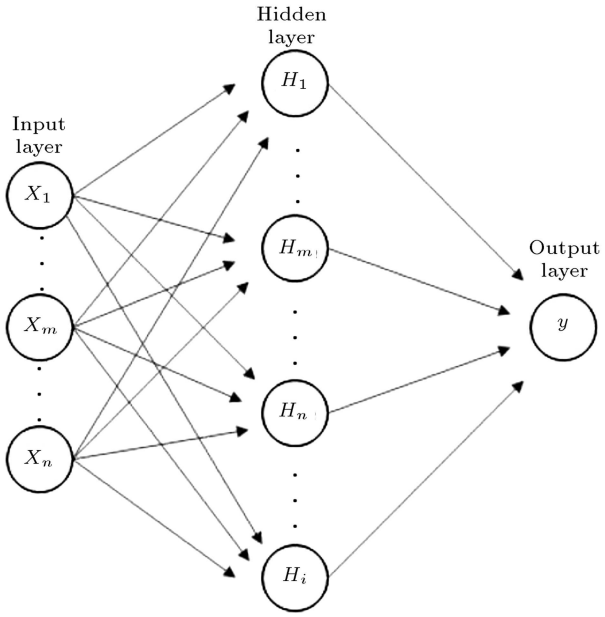

Figure 2. The framework of an artificial neural network.

function $f(x)$ that has ε deviation from actual y_i for all training datasets and is as flat as possible at the same time [12,24,25]. Subsequently, suppose that the function is defined as in $f(x) = wx + b$, where $w \in X$, $b \in R$. We then solve the following problem for $f(x)$, as shown in the following equations:

$$\min \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \right\},$$

$$\text{s.t.} \quad \begin{cases} y_i - wx_i - b \leq \varepsilon \\ wx_i + b - y_i \leq \varepsilon \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{1}$$

$$|\xi|_\varepsilon = \begin{cases} |\xi| - \varepsilon & \text{if } |\xi| > \varepsilon \\ 0 & \text{if } |\xi| \leq \varepsilon \end{cases} \tag{2}$$

In Eqs. (1) and (2), ξ and ξ_i^* are positive slack variables; C determines the trade-off between the flatness of function $f(x)$ and the amount up to which deviations larger than ε are tolerated; $|\xi|_\varepsilon$ is called ε-insensitive loss function. The above-mentioned optimization problem can be translated into a Lagrange dual problem, as shown in Eq. (3), and its solution is given by Eq. (4):

$$\max - \sum_{i,j=1}^{l} \frac{(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j)}{2}$$

$$- \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^*),$$

$$\text{s.t.} \quad \begin{cases} \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0 \\ 0 \leq \alpha_i, \alpha_i^* \leq C \end{cases} \tag{3}$$

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x_j) + b. \tag{4}$$

In Eqs. (3) and (4), α_i and α_i^* are the Lagrange multipliers corresponding to ξ and ξ_i^*; $k(x_i, x_j)$ is the kernel function. for more details of the process, refer to literature [26].

3. The MSVR model

Before using the MSVR model for forecasting stock indices, the daily stock index must be transformed into the trends for constructing a historical-trend database as described in Section 2.1. To enhance the forecasting accuracy, this paper uses the sliding window algorithm [25] to transform the daily stock index. An example of the transformed procedure is shown in Figure 3. There are eight transaction dates shown in Figure 3. According to the sliding window algorithm, the eight transaction dates are transformed into five trends, as shown in Figure 3. Subsequently, the historical trends can be used to construct a historical-trend database, and the prediction day can be forecasted by searching for the similar LHSs between the LHSs of the historical trends and that of the prediction day.

In this paper, the daily stock index must be transformed into the fluctuation of daily stock index as shown in Eq. (5) to enhance the probability in the search for similar historical trends.

$$R_t = (P_t - P_{t-1})/P_{t-1}. \tag{5}$$

In Eq. (5), R_t denotes the fluctuation of the tth day's stock index; P_t and P_{t-1} denote the tth and the $t-1$th days' stock indices, respectively.

Subsequently, the MSVR model uses the correlation coefficient method to calculate the corresponding correlation coefficient between the prediction day and historical trends derived from the historical database.

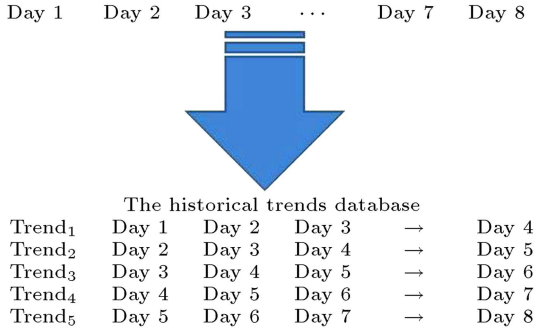

Figure 3. An example of the transformed procedure.

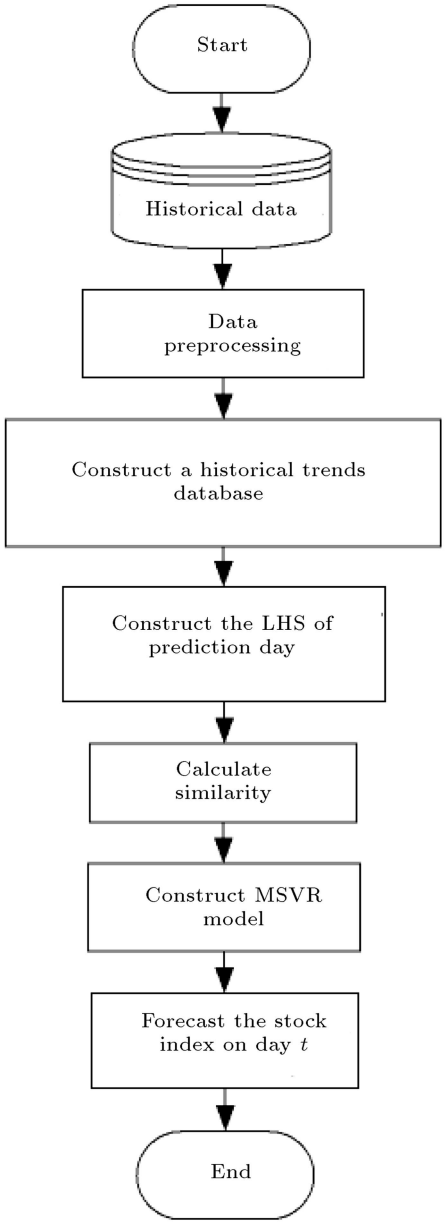

Figure 4. The flowchart of the MSVR model.

Then, the MSVR model searches through the similar historical trends in the historical database according to the results of the correlation coefficient. Finally, the MSVR model uses the searched historical trends as training samples to construct the training model and predict the value on the next day. The flowchart of the MSVR model is shown in Figure 4. The detailed procedure of the MSVR model is described in the following.

Step 1: Data pre-processing. Due to the fact that the fluctuations of the daily stock index might change dramatically according to the international situation, the accuracy of the forecasting model might be affected. To enhance the prediction accuracy of the MSVR model for forecasting stock indices, all of the daily stock indices must be transformed into the fluctuation of the daily data according to Eq. (5);

Table 1. A historical-trend database with a 3-day set.

	LHS		RHS
Trend_1	(R_1, R_2, R_3)	$\rightarrow$	R_4
Trend_2	(R_2, R_3, R_4)	$\rightarrow$	R_5
Trend_3	(R_3, R_4, R_5)	$\rightarrow$	R_6
$\cdots$	$\cdots$	$\rightarrow$	$\cdots$
Trend_n	(R_n, R_{n+1}, R_{n+2})	$\rightarrow$	R_{n+3}

Step 2: Construct a historical-trend database. According to the example of Figure 3, the transformed daily stock indices must be used to construct a historical-trend database using the sliding window algorithm. The example for constructing the historical-trend database is shown in Table 1. Table 1 shows the historical time series' situations and their corresponding values on the next day. For instance, it shows the historical trends' LHS and their corresponding RHS. Each historical trend can be represented by $(R_{i-3}, R_{i-2}, R_{i-1}) \rightarrow R_i$;

Step 3: Construct the LHS of the prediction day. When the historical-trend database has been constructed, we construct the LHS of the prediction day to search through the similar trends for forecasting the stock index on the prediction day. Suppose that the prediction day is day t. The LHS of day t can be represented by $(R_{t-3}, R_{t-2}, R_{t-1})$. Subsequently, the constructed LHS of day t can be used to search through similar trends for forecasting in the next step;

Step 4: Calculate the similarity of the historical trends. According to the literature [4-10,17,18], when two LHSs of trends are the same, their values on the next day might be similar. However, it is difficult to search through the same LHS when the LHS is constructed by many days. To counter this problem, the correlation coefficient method is used to calculate the correlation coefficient between the LHS of the prediction day and each LHS of historical trends from the historical database. In this paper, the threshold of correlation coefficient is 0.9 because it is called a highly positive relation according to the statistics definition when the correlation coefficient is greater than 0.9 [27]. For instance, if the correlation coefficient between the LHS of the prediction day and a LHS of a trend in the historical database is greater than 0.9, that trend will be saved and used in Step 5. In other words, the saved trends will be used as training samples to construct training models and predict stock indices in the following steps.

To demonstrate the method of calculating similarity, we give the following example. Suppose that

the LHS of the prediction day is represented by (0.03, 0.02, –0.005), then we can calculate the similarity between the LHS of the prediction day and each LHS of trends from Table 2 using the correlation coefficient method. According to the example in Table 2, two of five trends' correlation coefficients are greater than 0.9. Accordingly, the 1st and 5th trends will be selected to construct a training model in Step 5;

Step 5: Construct the MSVR model. After having similar trends, the searched similar trends are used to construct the MSVR model. The LHS and RHS of searched trends are input (independent) variables and output (dependent) variables, respectively, of the MSVR model. Hence, the MSVR model is constructed by 3 input variables and an output variable. Note that the parameters setting of the SVR in the MSVR model is set as follows:

(a) The MSVR model uses the radial basis kernel function;
(b) Cost C is set to 1;
(c) The gamma parameter of radial basis kernel function is set to 1/data dimension;
(d) The insensitive-loss function is set to 0.1.

Step 6: Forecast the stock index on day t. When the MSVR model is trained, we perform stock index forecasting on day t by feeding the LHS of the prediction day into the trained MSVR model to get the forecasted value on the prediction day. However, because the forecasted value is a forecasted fluctuation value, the forecasted value must be transformed into the stock index by Eq. (6). In Eq. (6), $\widehat{P}_t$ and $\widehat{R_t}$ denote the forecasted stock indices on day t and forecasted fluctuation value on day t, respectively; P_{t-1} denotes the actual stock index on day $t-1$. Note that an iteration of the above procedure (Steps 1 to 6) forecasts only one forecasting value. For instance, the MSVR model must be rebuilt (procedure Steps 1 to 6) for each new forecasting process (new testing data):

$$\widehat{P}_t = \widehat{R_t} \times P_{t-1} + P_{t-1}. \tag{6}$$

Table 2. The similarity between the LHS of the prediction day and each LHS of trends.

	LHS		RHS	r
Trend_1	(0.01, 0.005, –0.03)	→	0.03	0.9862
Trend_2	(0.04, 0.05, 0.02)	→	0.004	0.8171
Trend_3	(–0.02, –0.01, –0.03)	→	–0.002	0.6934
Trend_4	(0.02, 0.04, 0.01)	→	0.02	0.5447
Trend_5	(0.05, 0.01, –0.05)	→	0.02	0.9919

4. Forecasting stock indices

To demonstrate the proposed model for forecasting stock indices, we offer an example of "Hang Seng Indexes" in 2009 to introduce the procedure of the MSVR model. Table 3 shows the parts of transaction dates' close stock index. In this example, we use the first ten months as a training dataset to forecast the stock index on 2009/11/02.

Step 1: Data pre-processing. Firstly, the stock indices data in Table 3 must be transformed into the fluctuation data by Eq. (5). For example, the fluctuation value on 2009/1/5 equals $(15563.31 - 15042.81)/15563.31 = 0.0346$. In this example, the first ten months include 210 transaction dates; therefore, 209 fluctuation values are calculated. The parts of the transformed results are shown in Table 4;

Step 2: Construct a historical-trend database. After calculating the fluctuation value of the transaction dates, we use the sliding window algorithm

Table 3. The parts of transaction dates' close stock index of "Hang Seng Indexes" in 2009.

Date	2009/1/2	2009/1/5	2009/1/6	2009/1/7	2009/1/8	2009/1/9
Stock index	15042.81	15563.31	15509.51	14987.46	14415.91	14377.44
Date	2009/1/12	2009/1/13	2009/1/14	2009/1/15	2009/1/16	2009/1/19
Stock index	13971	13668.05	13704.61	13242.96	13255.51	13339.99

Table 4. The fluctuation value of the parts of the transaction dates' close stock index of "Hang Seng Indexes" in 2009.

Date	2009/1/5	2009/1/6	2009/1/7	2009/1/8	2009/1/9	2009/1/12
Fluctuation value	0.0346	–0.0346	–0.0337	–0.0381	–0.0267	–0.0283
Date	2009/1/13	2009/1/14	2009/1/15	2009/1/16	2009/1/19	2009/1/20
Fluctuation value	–0.0217	0.0267	–0.0369	0.0948	0.0637	–0.0285

Table 5. A part of the historical-trend database for the fluctuation value of close stock index of "Hang Seng Indexes".

	LHS		RHS
Trend_1	(0.0346, −0.0346, −0.0337)	→	−0.0381
Trend_2	(−0.0346, −0.0337, −0.0381)	→	−0.0267
Trend_3	(−0.0337, −0.0381, −0.0267)	→	−0.0283
Trend_4	(−0.0381, −0.0267, −0.0283)	→	−0.0217
Trend_5	(−0.0267, −0.0283, −0.0217)	→	0.0267
Trend_6	(−0.0283, −0.0217, 0.0267)	→	−0.0369
Trend_7	(−0.0217, 0.0267, −0.0369)	→	0.0948
Trend_8	(0.0267, −0.0369, 0.0948)	→	0.0637

Table 6. A part of the similarity between the LHS of the prediction day and each of the LHS of trends.

	LHS		RHS	r
Trend_1	(0.0346, −0.0346, −0.0337)	→	−0.0381	−0.7791
Trend_2	(−0.0346, −0.0337, −0.0381)	→	−0.0267	−0.5269
Trend_3	(−0.0337, −0.0381, −0.0267)	→	−0.0283	0.9996
Trend_4	(−0.0381, −0.0267, −0.0283)	→	−0.0217	−0.3319
Trend_5	(−0.0267, −0.0283, −0.0217)	→	0.0267	−0.1847
Trend_6	(−0.0283, −0.0217, 0.0267)	→	−0.0369	0.9580
Trend_7	(−0.0217, 0.0267, −0.0369)	→	0.0948	−0.8082
Trend_8	(0.0267, −0.0369, 0.0948)	→	0.0637	0.5388

to construct the historical-trend database. Due to the fact that this example includes 209 fluctuation values, 206 historical trends are constructed in the historical-trend database. A part of the historical-trend database includes eight trends, as shown in Table 5;

Step 3: Construct the LHS of the prediction day. When the historical-trend database is constructed, the LHS of the prediction day must be constructed in this step. In this example, the prediction day is 2009/11/2; therefore, the LHS of the prediction day is represented by the previous three days, that is, $(2009/10/28, 2009/10/29, 2009/10/30) = (-0.0184, -0.0228, 0.0229)$;

Step 4: Calculate the similarity of the historical trends. Step 4 uses the LHS of the prediction day to calculate the similarity with each of the historical trends in the historical database. Table 6 shows a part of the similarity between the LHS of the prediction day and each of the LHS of trends. According to the results, 30 of 206 historical trends are similar trends because their correlation coefficients are greater than 0.9. Subsequently, the 30 historical trends are used to construct the MSVR model in Step 5;

Step 5: Construct the MSVR model. The LHS of the 30 selected historical trends and the RHS of the 30 selected historical trends are input (independent) and output (dependent) variables, respectively, of the MSVR model. Furthermore, they are used to construct the MSVR model;

Step 6: Forecast the stock index on 2009/11/2. We perform forecasting stock index on 2009/11/2 by feeding the LHS of the prediction day into the constructed MSVR model to get the forecasted value on the prediction day. In this example, (−0.0184, −0.0228, 0.0229) are fed into the constructed MSVR model, and the forecasted value equals −0.0039. Subsequently, the forecasted value must be transformed into the stock index by Eq. (6). In this example, the forecasted stock index equals 21668.42. Note that the actual stock index on 2009/11/2 equals 21620.19. The forecasted and actual stock indices are much closer.

5. Experiment and results

5.1. Dataset

The dataset of this paper is the daily closed stock index from January 2000 to December 2009. Table 7 shows the descriptive statistics of the dataset. According to Table 7, the fluctuations from 2007 to 2009 are much greater than those in the other years, especially in 2008. Note that the stock indices of the beginning 10 months of each year are training dataset, and the rest are testing dataset. Moreover, in forecasting the stock index of a specific date, the stock indices of the previous transaction dates become the historical data for the MSVR model.

5.2. Performance measures

Two performance measures, i.e. the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), are used to verify predictive accuracy. Ahlburg [28] indicated that the MAPE is helpful in comparing various forecasting models. The MAPE has been widely used for measuring forecasting accuracy [8,29-31]; thus, the MAPE is used as the first alternative accuracy measurement to evaluate the performances of the forecasting methods in this paper. The MAPE is defined in Eq. (7) where t denotes the tth data point ($t = 1, 2, 3, \cdots$, and n), n denotes the number of predicted values, and P_t and $\widehat{P}_t$ denote the tth actual and tth predicted closed stock indices, respectively. A lower MAPE value indicates a more accurate forecasting power:

$$\text{MAPE} = \left\{ \frac{1}{n} \sum_{t=1}^{n} \frac{\left| P_t - \widehat{P}_t \right|}{P_t} \right\} \times 100\%. \tag{7}$$

On the other hand, the RMSE, widely used to verify predictive accuracy [31,32], is the second performance

Table 7. The descriptive statistics of "Hang Seng Indexes".

	Year									
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
n	274	244	247	248	249	247	250	253	251	253
Mean	16058.3	12547.9	10453.8	10291.1	12918.5	14352.6	16885.4	23196.4	20878.1	18106.0
SD	1098.0	1823.2	754.1	1265.6	749.0	615.6	1232.4	3494.6	4250.7	3450.7
Max	18301.7	16164.0	11974.6	12594.4	14266.4	15466.1	20001.9	31638.2	27615.9	22944.0
Min	13722.7	8934.2	8858.7	8409.0	10967.7	13355.2	14944.8	18664.9	11015.8	11344.6

Table 8. The performance of the MSVR model and that of the existing models.

	MSVR		**SVR**		**ANN**		**WFTS**		**FTS**	
Year	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
2000	288.3	1.54	266.2*	1.46+	420.1	2.26	279.0	1.53	275.6	1.52
2001	173.4*	1.11+	202.8	1.47	369.6	2.35	239.1	1.61	282.4	1.99
2002	115.3*	0.87+	1328.9	12.55	283.8	2.40	188.4	1.60	179.5	1.50
2003	132.7*	0.82+	417.9	2.84	274.9	1.92	207.6	1.41	207.7	1.42
2004	115.7*	0.64+	490.3	2.76	509.6	3.25	324.8	2.06	331.9	2.13
2005	114.6	0.56+	107.6*	0.56+	429.5	2.46	205.4	1.16	197.2	1.10
2006	196.1*	0.79+	1564.1	7.61	1205.8	6.09	839.9	4.07	839.4	4.07
2007	704.2*	2.00+	796.0	2.31	1673.5	4.68	901.3	2.45	938.0	2.62
2008	492.4*	2.96+	653.6	3.71	1278.1	7.47	950.6	5.54	945.6	5.64
2009	371.3	1.25+	423.4	1.57	2240.7	9.83	379.8	1.39	361.7*	1.33
Mean	270.4*	1.25+	625.1	3.68	868.6	4.27	451.6	2.28	455.9	2.33
SD	198.4*	0.74+	481.2	3.67	689.5	2.70	314.1	1.42	318.2	1.45

Note: "*" and "+" denote the smallest RMSE and MAPE of all models in the corresponding year, respectively.

measure to test the forecasting accuracies of the MSVR model and that of the existing models. The equation is shown in the following:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(P_t - \widehat{P}_t\right)^2}{n}}, \tag{8}$$

where P_t and $\widehat{P}_t$ denote the actual and forecasted closed stock indices on day t.

5.3. Performance comparison

The performance of the MSVR model is compared with those of the existing models published in the literature [3,10]. Due to the fact that Chen and Yu's models are the well-known time series models, this paper compares the performance of the MSVR model with those of two models. Moreover, this paper also compares the performance of the MSVR model with that of the data mining or big data analysis method, such as the traditional SVR and ANN models. The performance of the MSVR model and those of the existing models are shown in Table 8. In Table 8, FTS denotes the length of intervals of Chen's model equaling 500; similarly, WFTS denotes the length of intervals of Yu's model equaling 500.

First, we compare the RMSEs of all models. According to Table 8, although the RMSEs of the MSVR model are not the best in 3 of the 10 years (2000, 2005, and 2009), the remaining years and average RMSE of the MSVR model are the best. In contrast, the average RMSE of the FTS model and that of the WFTS model are insignificantly different. For instance, the performances of the FTS and WFTS models are almost the same. The traditional SVR and the ANN

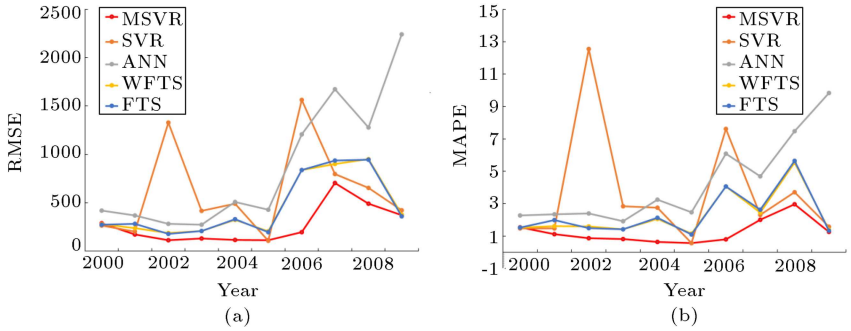

Figure 5. The performance plot of all models in the corresponding year: (a) RMSE and (b) MAPE.

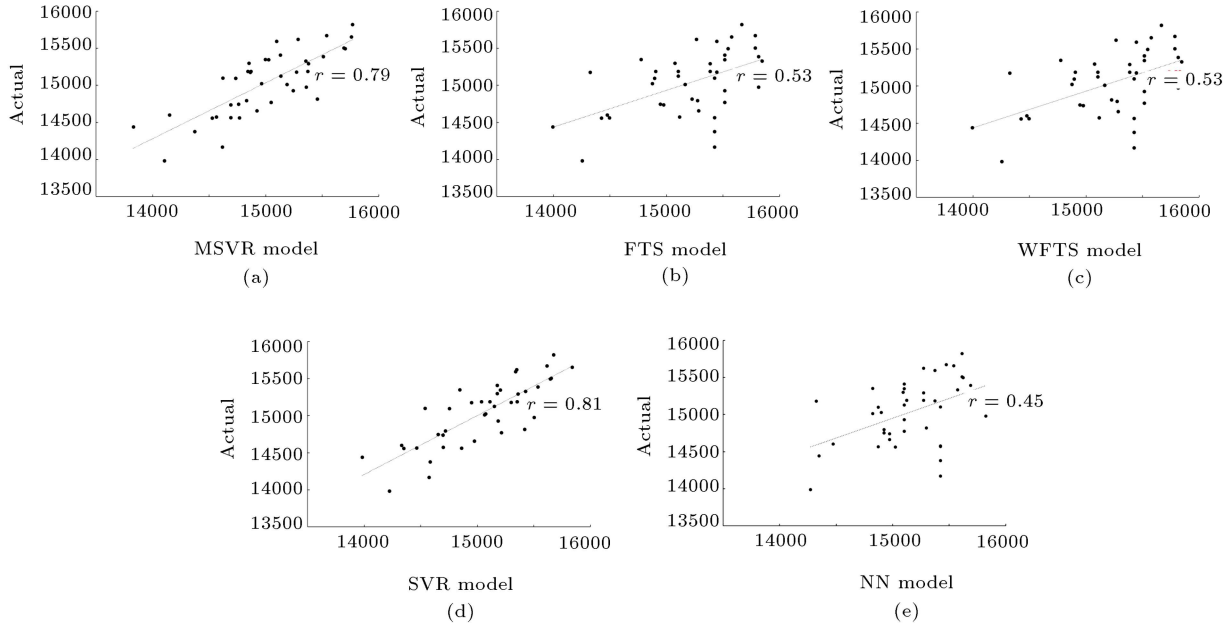

Figure 6. The scatter plot of the actual stock indices and the models' forecasted stock indices in 2000: (a) MSVR model, (b) FTS model, (c) WFTS model, (d) SVR model, and (e) ANN model.

models bring about better performance in the steady years (2000 and 2005). In the extreme fluctuations years, the performances of the existing models are significantly worse than that of the MSVR model in terms of RMSE.

On the other hand, the MAPE of the MSVR model in only one year (2000) is not the best. Obviously, the average and standard deviation of MAPE of the MSVR model are significantly better than those of the other existing models. Furthermore, RMSEs and MAPEs of the MSVR model are almost less than half compared to those of the existing models in 2006-2009. Moreover, Figure 5 shows that the MSVR model provides a stable performance because the plot of the MSVR model does not have extreme fluctuations.

Figure 6 shows the scatter plot of the actual and forecasted stock indices (Figure 6(a) the MSVR model, Figure 6(b) the FTS model, Figure 6(c) the WFTS model, Figure 6(d) the SVR model, and Figure 6(e) the ANN model) in 2000. According to the result of Figure 6, the forecasted stock indices of the MSVR model and those of the SVR model are much closer to the actual stock indices. Obviously, the correlation coefficient of the actual stock indices and the SVR model is the best ($r = 0.81$). The result of Figure 6 is similar to that in Table 8 and Figure 5. In contrast, Figure 6(b), (c), and (e) cannot find any patterns, that is, the forecasted stock indices are not close to the actual stock indices. Moreover, the correlation coefficients of the actual stock indices and the FTS model forecasted stock indices only equal 0.53; those of the actual stock indices and the WFTS model forecasted stock indices also equal 0.53; those of the actual stock indices and the ANN model forecasted stock indices equal 0.45. Obviously, the performance of the MSVR model and that of the SVR model are also better than those of the existing models in terms

of the correlation coefficient in 2000. In short, the forecasting accuracy of the MSVR model is much more stable than that of the existing model according to the experiment results. Furthermore, external factors do not easily affect the forecasting accuracy of the MSVR model.

5.4. Time complexity comparison

Section 5.3 has verified that the performance of the MSVR model is better than those of the existing models in terms of RMSE and MAPE. This section will compare the time complexities of the models. This section uses two results to compare the execution times of the models. First, we use stock indices in 2000 to compare the stability statuses of the execution CPU times for all models in training and forecasting. Figure 7 shows the performance in the stability of the execution CPU time. In Figure 7, the x-axis and y-axis denote the number of executions and CPU time (seconds), respectively. Obviously, the performance in execution CPU time of the SVR, FTS, and WFTS models is better than the others. Although the ANN model needs most of CPU time to train the model and forecast indices, the average CPU time of the ANN model is only 32.618 (seconds). Accordingly, we think only 32.618 (seconds) to train the model (206 training samples) and forecast indices (41 forecasting values) should be used for forecasting stock indices in one year. Hence, all models should be applied for forecasting stock indices in one year according to this result. Second, we use simulation data to compare the training execution CPU times in different numbers of training samples. Figure 8 and Table 9 show the experimental results. In Figure 8, the x-axis and y-axis denote the number of training samples and CPU time, respectively. Obviously, the rates of CPU time of the FTS and WFTS models increase linearly. On the contrary, the increased rates of CPU times of the MSVR, SVR, and ANN models are increasing with the number

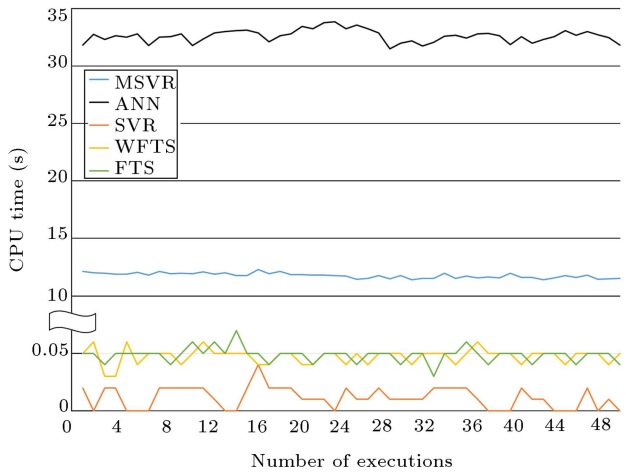

Figure 7. The stability of the execution CPU time.

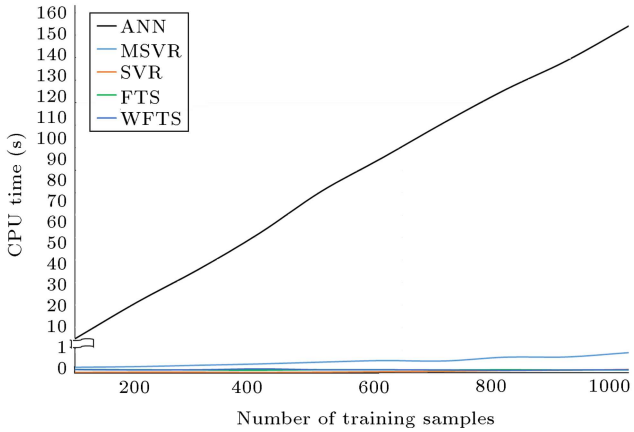

Figure 8. The training execution CPU time in different numbers of training samples.

Table 9. The results of the training execution CPU time (seconds) in different numbers of training samples.

n	MSVR	SVR	ANN	FTS	WFTS
100	0.24	0.01	15.11	0.14	0.14
200	0.27	0.02	31.22	0.13	0.12
300	0.33	0.02	45.81	0.12	0.13
400	0.39	0.01	61.99	0.11	0.17
500	0.47	0.02	80.53	0.13	0.13
600	0.54	0.04	95.56	0.13	0.13
700	0.52	0.06	111.16	0.12	0.12
800	0.69	0.09	125.83	0.13	0.1
900	0.7	0.11	138.77	0.12	0.12
1000	0.89	0.14	154.06	0.12	0.13

of training samples. Although the MSVR and SVR models' increased rates of CPU time are increasing with the number of training samples, the increased rates are still in the acceptable range because their training execution CPU times are both less than 1 second when the numbers of training samples equal 1,000 samples.

According to the above experimental results, the ANN model might be the worst, and the SVR model brings the best performance in comparing time complexities. Although the MSVR model is not the best in comparing time complexities, we still think the MSVR can be applied to the stock indices forecasting problem. The reasons are:

(a) Its execution CPU time is in the acceptable range because the MSVR model can train the model and forecast indices within 15 seconds for one-year stock indices, and the MSVR model can be trained for a large sample size (1,000 daily transactions in 4 years) within 1 second;

(b) The forecasting accuracy of the MSVR model is almost 2 or 3 times better as compared with that of the existing model. Note that the experimental results of Section 5.4 are implemented on a personal computer (OS: MS Windows 10; CPU: Intel Core i7-4660; RAM: 16 GB).

6. Conclusion

6.1. Research contributions

Although many models have been proposed recently for forecasting time series data, this paper proposes a novel hybrid model to enhance forecasting accuracy in stock indices forecasting to assist investors in earning profits. The experiment results show that the MSVR model is more accurate than the other existing methods in terms of RMSE and MAPE. The average and standard deviation of RMSE and MAPE of the MSVR model are much better than those of the existing models, even if in some years, RMSEs or MAPEs of the MSVR model are lower than those of the existing models.

The main contributions of the MSVR model are:

(a) The forecasting accuracy of the MSVR model is much more stable than that of the existing model because it reduces the effect of the extreme fluctuations using the sliding window algorithm;

(b) The external factors do not easily affect the forecasting accuracy of the MSVR model;

(c) The MSVR model can be used to forecast with any assumptions. Hence, the MSVR model offers a useful alternative for stock indices forecasting. The investors can refer to the results of the MSVR model to determine a strategy of investment in the stock market.

6.2. Research limitations and future research

The experiment dataset of this paper involves "Hang Seng Indexes" from 2000 to 2009. The data period includes the financial crises in 2007-2008 that might affect the prediction accuracy of the model. Furthermore, this paper only uses the historical dataset of stock indices for forecasting. However, the stock index may be affected by many external factors. These factors might affect or limit the forecasting performance of the models.

Therefore, we give the following suggestions for future research. First, the data period should be lengthened. The historical data should include more situations, e.g. recent financial crisis in different years. That should reduce the effect of financial crises to enhance the forecasting accuracy of the models. Second, future studies should compare the performances of models using more countries' stock indices to reduce the effect of political events. Third, any future research could use more factors to discuss the influence on forecasting because the stock index could be affected by many external factors.

References

1. Donate, J.P., Sanchez, G.G., and De Miguel, A.S. "Time series forecasting. a comparative study between an evolving artificial neural networks system and statistical methods", *Int. J. Artif. Intell. Tools*, **21**(01), 1250010 (2012).
2. Tseng, C.-H., Cheng, S.-T., Wang, Y.-H., and Peng, J.-T. "Artificial neural network model of the hybrid EGARCH volatility of the Taiwan stock index option prices", *Physica A*, **387**(13), pp. 3192-3200 (2008).
3. Yu, H.-K. "Weighted fuzzy time series models for TAIEX forecasting", *Physica A*, **349**(3-4), pp. 609-624 (2005).
4. Cheng, C.-H., Chen, T.-L., Teoh, H.J., and Chiang, C.-H. "Fuzzy time-series based on adaptive expectation model for TAIEX forecasting", *Expert Syst. Appl.*, **34**(2), pp. 1126-1132 (2008).
5. Leu, Y., Lee, C.-P., and Jou, Y.-Z. "A distance-based fuzzy time series model for exchange rates forecasting", *Expert Syst. Appl.*, **36**(4), pp. 8107-8114 (2009).
6. Lee, C.-P., Lin, W.-C., and Yang, C.-C. "A strategy for forecasting option price using fuzzy time series and least square support vector regression with a bootstrap model", *Sci. Iran.*, **21**(3), pp. 815-825 (2014).
7. Leu, Y., Lee, C.-P., and Hung, C.-C. "A fuzzy time series-based neural network approach to option price forecasting", In *Intelligent Information and Database Systems*, N. Nguyen, M. Le and J. Świątek, pp. 360-369, Springer Berlin Heidelberg (2010).
8. Yang, C.-C., Leu, Y., and Lee, C.-P. "A dynamic weighted distanced-based fuzzy time series neural network with bootstrap model for option price forecasting", *Rom. J. Econ. Forecast*, **2014**(2), pp. 115-129 (2014).
9. Lee, L.-W., Wang, L.-H., Chen, S.-M., and Leu, Y.-H. "Handling forecasting problems based on two-factors high-order fuzzy time series", *IEEE Trans. Fuzzy Syst.*, **14**(3), pp. 468-477 (2006).
10. Chen, S.-M. "Forecasting enrollments based on high-order fuzzy time series", *Cybern. Syst.*, **33**(1), pp. 1-16 (2002).
11. Liang, X., Zhang, H., and Li, X. "A simple method of forecasting option prices based on neural networks", In *Next-Generation Applied Intelligence*, B.-C. Chien, T.-P. Hong, S.-M. Chen, and M. Ali, pp. 586-593, Springer Berlin Heidelberg (2009).
12. Huang, C.-F. "A hybrid stock selection model using genetic algorithms and support vector regression", *Appl. Soft. Comput.*, **12**(2), pp. 807-818 (2012).
13. Wang, Y.-H. "Nonlinear neural network forecasting model for stock index option price: hybrid GJR-GARCH approach", *Expert Syst. Appl.*, **36**(1), pp. 564-570 (2009).

14. Panda, C. and Narasimhan, V. "Forecasting exchange rate better with artificial neural network", *J. Policy Model.*, **29**(2), pp. 227-236 (2007).

15. Wei, L.-Y., Cheng, C.-H., and Wu, H.-H. "A hybrid ANFIS based on n-period moving average model to forecast TAIEX stock", *Appl. Soft. Comput.*, **19**, pp. 86-92 (2014).

16. Wang, Y., Wang, B., and Zhang, X. "A new application of the support vector regression on the construction of financial conditions index to CPI prediction", *Procedia Comput. Sci.*, **9**, pp. 1263-1272 (2012).

17. Chen, S.-M. and Kao, P.-Y. "TAIEX forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines", *Inf. Sci.*, **247**, pp. 62-71 (2013).

18. Chen, S.-M. "Forecasting enrollments based on fuzzy time series", *Fuzzy Sets Syst.*, **81**(3), pp. 311-319 (1996).

19. Song, Q. and Chissom, B.S. "Forecasting enrollments with fuzzy time series - Part I", *Fuzzy Sets Syst.*, **54**(1), pp. 1-9 (1993).

20. Song, Q. and Chissom, B.S. "Forecasting enrollments with fuzzy time series-part II", *Fuzzy Sets Syst.*, **62**(1), pp. 1-8 (1994).

21. McCulloch, W.S. and Pitts, W. "A logical calculus of the ideas immanent in nervous activity", *Bull. Math. Biol.*, **5**(4), pp. 115-133 (1943).

22. Lee, C.-P., Shieh, G.-J., Yiu, T.-J., and Kuo, B.-J. "The strategy to simulate the cross-pollination rate for the co-existence of genetically modified (GM) and non-GM crops by using FPSOSVR", *Chemometrics Intell. Lab. Syst.*, **122**, pp. 50-57 (2013).

23. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., and Vapnik, V. "Support vector regression machines", *NIPS'1996*, pp. 155-161 (1996).

24. Basak, D., Pal, S., and Patranabis, D.C. "Support vector regression", *Neural Inf. Process Lett. Rev.*, **11**, pp. 203-224 (2007).

25. Kapoor, P. and Bedi, S.S. "Weather forecasting using sliding window algorithm", *ISRN Signal Processing*, **2013**, p. 5 (2013).

26. Smola, A.J. and Schölkopf, B. "A tutorial on support vector regression", *Stat. Comput.*, **14**(3), pp. 199-222 (2004).

27. Kleinpeter, M.A. "Multivariable analysis: A practical guide for clinicians", 2nd Edn., *J. Natl. Med. Assoc.*, **99**(6), pp. 684-684 (2007).

28. Ahlburg, D.A. "How accurate are the U.S. bureau of the census projections of total live births?", *J. Forecast.*, **1**(4), pp. 365-374 (1982).

29. Ou, S.-L. "Forecasting agricultural output with an improved grey forecasting model based on the genetic algorithm", *Comput. Electron. Agric.*, **85**, pp. 33-39 (2012).

30. Coshall, J.T. "Combining volatility and smoothing forecasts of UK demand for international tourism", *Tourism Manage.*, **30**(4), pp. 495-511 (2009).

31. Gani, A., Mohammadi, K., Shamshirband, S., Altameem, T.A., Petković, D., and Ch, S. "A combined method to estimate wind speed distribution based on integrating the support vector machine with firefly algorithm", *Environ. Prog. Sustain. Energy*, **35**(3), pp. 867-875 (2016).

32. Hadavandi, E., Shahrabi, J., and Shamshirband, S. "A novel Boosted-neural network ensemble for modeling multi-target regression problems", *Eng. Appl. Artif. Intell.*, **45**, pp. 204-219 (2015).

Biographies

Chih-Hua Huang has a master's degree in TESOL from the Department of Applied Foreign Languages at National Yunlin University of Science and Technology in Taiwan and is currently a lecturer at several universities. In addition to teaching English, she is also a student in the PhD Program in Management at Da-Yeh University. Her research interests are in organizational behavior, financial forecasting, and human resource management. Topics include emotions, work engagement, psychological safety, work-family conflict, job satisfaction, etc.

Feng-Hua Yang received his PhD in International Business Administration from Chinese Culture University, Taiwan. He is an Associate Professor at the Department of International Business Management at Da-Yeh University, Taiwan. His research interests include behavioral finance, consumer behavior, organizational behavior, and human resource management.

Chien-Pang Lee received the BS degree in Applied Statistics from Ming Chuan University, Taiwan in 2003, the MS degree in Biostatistics from Nation Chung Hsing University, Taiwan in 2006, and PhD degree in Information Management from Nation Taiwan University of Science and Technology, Taiwan, in 2010. He is currently an Associate Professor in the Department of Maritime Information and Technology at National Kaohsiung University of Science and Technology. His research includes big data analysis, bioinformatics, statistics, and financial forecasting.