



A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting

Ching-Hsue Cheng^a, Tai-Liang Chen^b, Liang-Ying Wei^{c,*}

^a Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan

^b Department of Information Management and Communication, Wenzao Ursuline College of Languages, 900 Mintsu, 1st Road, Kaohsiung 807, Taiwan

^c Department of Information Management, Yuanpei University, 306 Yuanpei Street, Hsin Chu 30015, Taiwan

ARTICLE INFO

Article history:

Received 26 November 2008

Received in revised form 21 December 2009

Accepted 8 January 2010

Keywords:

Rough set theory

Genetic algorithms

Cumulative probability distribution approach

Minimize entropy principle approach

Technical indicators

ABSTRACT

In the stock market, technical analysis is a useful method for predicting stock prices. Although, professional stock analysts and fund managers usually make subjective judgments, based on objective technical indicators, it is difficult for non-professionals to apply this forecasting technique because there are too many complex technical indicators to be considered. Moreover, two drawbacks have been found in many of the past forecasting models: (1) statistical assumptions about variables are required for time series models, such as the autoregressive moving average model (ARMA) and the autoregressive conditional heteroscedasticity (ARCH), to produce forecasting models of mathematical equations, and these are not easily understood by stock investors; and (2) the rules mined from some artificial intelligence (AI) algorithms, such as neural networks (NN), are not easily realized.

In order to overcome these drawbacks, this paper proposes a hybrid forecasting model, using multi-technical indicators to predict stock price trends. Further, it includes four proposed procedures in the hybrid model to provide efficient rules for forecasting, which are evolved from the extracted rules with high support value, by using the toolset based on rough sets theory (RST): (1) select the essential technical indicators, which are highly related to the future stock price, from the popular indicators based on a correlation matrix; (2) use the cumulative probability distribution approach (CDPA) and minimize the entropy principle approach (MEPA) to partition technical indicator value and daily price fluctuation into linguistic values, based on the characteristics of the data distribution; (3) employ a RST algorithm to extract linguistic rules from the linguistic technical indicator dataset; and (4) utilize genetic algorithms (GAs) to refine the extracted rules to get better forecasting accuracy and stock return. The effectiveness of the proposed model is verified with two types of performance evaluations, accuracy and stock return, and by using a six-year period of the TAIEX (Taiwan Stock Exchange Capitalization Weighted Stock Index) as the experiment dataset. The experimental results show that the proposed model is superior to the two listed forecasting models (RST and GAs) in terms of accuracy, and the stock return evaluations have revealed that the profits produced by the proposed model are higher than the three listed models (Buy-and-Hold, RST and GAs).

© 2010 Elsevier Inc. All rights reserved.

* Corresponding author.

E-mail addresses: chcheng@yuntech.edu.tw (C.-H. Cheng), 97007@mail.wtuc.edu.tw (T.-L. Chen), lywei@mail.ypu.edu.tw (L.-Y. Wei).

1. Introduction

In the stock market, it is very difficult to forecast stock trends because of complex factors influencing stock markets and nonlinear relationships, which are contained among different periods of stock prices. Although only a few investors profit from the stock market, millions of them still have not given up trying to make money from the market. Therefore, since the first stock market opened, numerous forecasting methods have been employed in an attempt to predict stock prices.

In the area of stock market forecasting, the technical analysis method is one of the primary analytic approaches used by investors to make investment decisions, and many researchers have been focusing on technical analysis to increase their investment returns [4,13,15]. Furthermore, the technical analysis method has the ability to forecast the future price direction by studying past market data, primarily stock price and volume. The technical analysis method assumes that stock price and volume are the two most relevant factors in determining the future direction and behavior of a particular stock or market, and that the technical indicators, coming from a mathematical formula, based on stock price and volume, can be applied to predict price fluctuations and also provide data for investors, enabling them to determine the timing for the buying or selling of stock [13].

Besides the technical analysis methods, many conventional numeric forecasting models have been proposed by financial researchers, such as Engle's [17] autoregressive conditional heteroscedasticity (ARCH) model, Bollerslev's [6] generalized ARCH (GARCH) model, Box and Jenkins' [7] autoregressive moving average (ARMA) model, and the autoregressive integrated moving average model (ARIMA).

In recent decades, many researchers have employed another approach to financial forecasting: artificial intelligence algorithms. In 1990, Kinoto et al. [29] developed a prediction system for the stock market by using a neural network. Nikolopoulos and Fellrath [32] combined genetic algorithms (GAs) and a neural network to develop a hybrid expert system for investment decisions. Kim and Han [27] proposed a genetic algorithms approach in order to feature discretization and the determination of connection weights for artificial neural networks (ANNs) to predict the stock price index. Huarng and Yu [25] applied a backpropagation neural network to establish fuzzy relationships in fuzzy time series for forecasting stock prices. Roh [42] integrated a neural network and time series model for forecasting the volatility of the stock price index.

From the literature noted above, however, three major drawbacks can be found in their forecasting methods and models: (1) stock market analysts and fund managers apply various technical indicators to forecast stock market trends, based on their personal experience, which could result in erroneous judgments of market signals; (2) for most statistical methods, there are some assumptions about the variables used in the analysis, which can not be applied to those datasets that do not follow the statistical distributions; and (3) the artificial neural network (ANN) is a black-box method, and the rules mined from it are not easily understandable.

To improve upon past forecasting models, a revised model should be able to overcome the drawbacks contained in previous models and should offer a good methodology which could be used more easily by investors. Therefore, this paper proposes a hybrid forecasting model to refine past models in stock price forecasting, and provides four novel methods in the forecasting processes: (1) select essential technical indicators by using a correlation matrix; (2) use CPDA (cumulative probability distribution approach) and MEPA (minimize entropy principle approach) to discretize condition features (technical indicators) and decision features (daily price fluctuation); (3) apply the rough set theory (RST) to produce rules from the linguistic values of technical indicators; and (4) employ genetic algorithms (GAs) to refine the extracted rules to improve forecasting accuracy and stock return.

Empirically, this paper employs two types of stock databases (stock index and individual stock price) as experimental datasets. From the model verification, it is shown that the refined processes are effective in improving forecasting accuracy, and, based on the evidence, a stock analyst or investor can employ the refined processes proposed in this paper to improve their forecasting tools or models.

The rest of this paper is organized, as follows: Section 2 introduces the related works; Section 3 demonstrates the proposed model and algorithm; Section 4 evaluates the performance of the proposed model and describes the findings; and Section 5 draws conclusions and proposes recommendations for future research.

2. Related works

This section reviews related works of technical analysis, cumulative probability distribution approach, minimize entropy principle approach, rough set theory, and genetic algorithms.

2.1. Technical analysis

Technical analysis is an attempt to predict future stock price movements by analyzing a past sequence of stock prices [39]. It relies on charts and looks for particular configurations that are supposed to have predictive value. Analysts focus on investor psychology, which represents common investors' responses to certain price formations and price movements, to analyze the fluctuations of stock market. The price at which investors are willing to buy or sell depends on personal expectation. If investors expect the security price to rise, they will buy it; if investors expect the security price to fall, they will sell it. These simple statements are the cause for a major challenge in setting security prices, because they refer to human expect-

tations and attitudes [39]. It is said that, securities never sell for what they are worth, but for what people think they are worth. It is very important to understand that market participants anticipate future development and take timely action, which compels the price movement. Since stock market processes are highly nonlinear, many researchers have been focusing on technical analysis to improve investment returns [2,4,49].

2.2. Cumulative probability distribution approach (CPDA)

Probability refers to the study of randomness and uncertainty. In any situation, one of a number of possible outcomes may occur. The theory of probability provides methods for quantifying the chances, or likelihoods, associated with the various outcomes. Because a probability distribution on the real line is determined by the probability of being in a half-open interval $p(a, b]$, therefore, $F(b) - F(a)$ if $a < b$. The probability distribution of a real-valued random variable X is completely characterized by its cumulative distribution function (CDF) [1]. For every real number x , the CDF of X is given by

$$x \rightarrow F_X(x) = P(X \leq x) \quad \forall x \in \mathfrak{R} \quad (1)$$

where the right-hand side represents the probability (p) that the random variable X takes on a value less than, or equal to, x . Capital F is used to represent the cumulative distribution function, in contrast to the lower-case f , used to represent probability density functions and probability mass functions. The CDF of X can be defined in terms of the probability density function f , as follows:

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t)dt \quad (2)$$

The inverse of the normal CDF is computed with parameters μ and σ at the corresponding probabilities in P , where μ denotes the mean and σ denotes the standard deviation of the data [1]. The normal inverse function in terms of the normal CDF is defined as

$$x = F^{-1}(p|\mu, \sigma) = \{x : F(x|\mu, \sigma) = p\}, \quad (3)$$

where

$$p = F(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \frac{-(t-\mu)^2}{2\sigma^2} dt \quad (4)$$

The cumulative probability of normal distribution is used to determine the intervals. The steps of the cumulative probability distribution approach are as follows [14]:

Step 1: Test normal distribution.

In this approach, the data must be of normal distribution. This study utilizes CPDA because stock market fluctuations and returns tend to be of normal distribution [3,47].

Step 2: Define the universe of discourse U .

Let $U = [D_{min}^{-\sigma}, D_{max}^{+\sigma}]$, where D_{min} and D_{max} denote the minimum and maximum values in historical data, and σ denotes the standard deviation of the yearly data.

Step 3: Determine the length of intervals and build a membership function.

P_{LB} , as the lower boundary of cumulative probability, and P_{UB} , as the upper boundary of cumulative probability of each linguistic value, are computed by

$$P_{LB} = (2i - 3)/2n \quad (2 \leq i \leq n) \quad (5)$$

$$P_{UB} = i/n \quad (1 \leq i \leq n) \quad (6)$$

where i denotes the order of the linguistic values, and n denotes the number of linguistic values. The lower boundary of the first linguistic value and the upper boundary of the last linguistic value correspond to the lower and upper boundary, respectively. This step computes the inverse of the normal CDF by Eqs. (3) and (4).

Step 4: Fuzzify the historical data.

According to the inverse of normal CDF, the lower boundary, midpoint and upper boundary as the triangular fuzzy number of each linguistic value, can be computed. The triangular fuzzy number is applied to build a membership function. The membership degree of each instance is calculated to determine its linguistic value.

2.3. Minimize entropy principle approach (MEPA)

The purpose of entropy minimization analysis is to determine the information content in a given dataset. The entropy of a probability distribution is a measure of the uncertainty of the distribution [51]. To divide the data into membership functions, establishing the point of segmentation between classes of data is needed. A point of segmentation can be determined with an entropy minimization screening method; then, start the segmentation process by first dividing it into two classes. Thereupon, a repeated partitioning with the value of segment point calculations will allow us to partition the dataset into a number of fuzzy sets [43]. In recent years, MEPA has been used in forecasting problems [10].

Assume that the value of the segment point is being sought for a sample in the range between x_1 and x_2 . An entropy equation is written for the regions $[x_1, x]$ and $[x, x_2]$, and denotes the first region p and the second region q . Entropy with each value of x is expressed as: [14]

$$S(x) = p(x)S_p(x) + q(x)S_q(x) \quad (7)$$

where

$$S_p(x) = -[p_1(x) \ln p_1(x) + p_2(x) \ln p_2(x)] \quad (8)$$

$$S_q(x) = -[q_1(x) \ln q_1(x) + q_2(x) \ln q_2(x)]$$

and where $p_k(x)$ and $q_k(x)$ are the conditional probabilities that the class k sample is in region $[x_1, x_1 + x]$ and $[x_1 + x, x_2]$, respectively, and $p(x)$ and $q(x)$ are the probabilities that all samples are in region $[x_1, x_1 + x]$ and $[x_1 + x, x_2]$, respectively.

$$p(x) + q(x) = 1 \quad (9)$$

The value of x that gives the minimum entropy is the optimum value of the segment point. The entropy estimates of $pk(x)$ and $qk(x)$, and $p(x)$ and $q(x)$ are calculated, as follows: [14]

$$pk(x) = \frac{n_k(x) + 1}{n(x) + 1} \quad (10)$$

$$qk(x) = \frac{N_k(x) + 1}{N(x) + 1} \quad (11)$$

$$p(x) = \frac{n(x)}{n} \quad (12)$$

$$q(x) = 1 - p(x) \quad (13)$$

where $n_k(x)$ is the number of class k samples located in $[x_1, x_1 + x]$, $n(x)$ is the total number of samples located in $[x_1, x_1 + x]$, $N_k(x)$ is the number of class k samples located in $[x_1 + x, x_2]$, $N(x)$ is the total number of samples located in $[x_1 + x, x_2]$, and n is the total number of samples in $[x_1, x_2]$.

2.4. Rough set theory

Rough sets theory (RST) was proposed by Pawlak [33–37] in 1982. In recent years, RST has been used in economic and financial prediction. Many researchers have applied RST to discover trading rules [20,48]. The concept of RST is founded on the assumption that with every associated object of the universe of discourse, some information objects characterized by the same information are indiscernible in the view of the available information about them. Any set of all indiscernible objects is called an elementary set and forms a basic granule of knowledge about the universe. Any union of elementary sets is referred to as a precise set; otherwise the set is rough.

With any rough sets, a pair of precise sets, called the *lower* and *upper approximation*, $\underline{BX} = \{x | [x]_B \subseteq X\}$ and $\overline{BX} = \{x | [x]_B \cap X \neq \emptyset\}$ of the rough sets, is associated [33]. The lower approximation consists of all objects that definitely belong to the set, and the upper approximation contains all objects that possibly belong to the set. The difference between the upper and the lower approximation constitutes the *boundary region*, $BN_B(x) = \overline{BX} - \underline{BX}$, of the rough sets. The set X is called “rough” (or “roughly definable”) with respect to the knowledge in B , if the boundary region is non-empty. The basic notions in rough sets are shown in Fig. 1.

The RST is a series of logical reasoning procedures used for analyzing an information system. An information system can be seen as a decision table, denoted by $S = (U, A, C, D)$, where U is the universe of discourse, A is a set of primitive features, and $C, D \subset A$ are two subsets of features, assuming that $A = C \cup D$ and $C \cap D = \emptyset$, where C is called the condition attribute and D is the decision attribute. The measure to describe the inexactness of approximation classifications is called the quality of

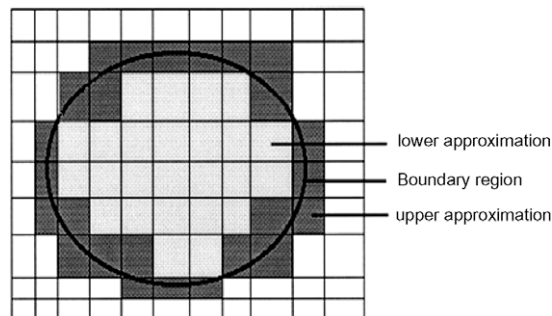


Fig. 1. Basic notions of rough sets.

approximation of X by B . It expresses the percentage of objects that can be correctly classified into class X , employing the attribute B [33]:

$$\gamma_B(A) = \frac{\sum \text{card}(BX_i)}{\text{card}(U)} \quad (14)$$

If $\gamma_B(A) = 1$, then the decision table is consistent; otherwise, it is inconsistent.

An important issue in RST is attribute reduction in which the reduced set of attributes provides the same quality of approximation as the original set. There are two fundamental concepts in connection with this attribute reduction. The B -reduct of A , denoted by $\text{RED}(B)$, is the minimal subset of A , which provides the same quality of approximation of objects into elementary classes of B as the whole attributes of A . The B -core of A , $\text{CORE}(B)$, is the essential part of A , which cannot be eliminated without disturbing the ability to classify objects into the elementary classes of B [33]. It is the intersection of all reducts.

$$\text{CORE}(B) = \bigcap_{R_i \in \text{RED}(B)} R_i, \quad i = 1, 2, \dots \quad (15)$$

Using a reduced algorithm, the rules can be found through determining the decision attributes value, based on the condition attributes values. Therefore, the rules are presented in an “IF condition(s) THEN decision(s)” format. The concept of the decision table is employed in this study to establish rules from fuzzy relationships, which generate rules for better forecasting results.

2.5. Genetic algorithms

Genetic algorithms (GAs) were advanced by Holland [24] in 1975, and expanded by Goldberg [21] in 1989. GAs are search algorithms, inspired by evolution and applied in searching for the global optimum for many applications. Furthermore, GAs have been successfully applied in economic and financial prediction [2,28]. These algorithms encode a potential solution for a specific problem into a simple chromosome-like data structure and apply recombination operators to these structures to preserve critical information. The steps of GAs in the proposed model, based on Goldberg [21], are reorganized for this study, as follows:

Step 1: Initialization.

This step generates the initial population containing N_p chromosomes, which are used to find global optimum initial seeds, where N_p is the number of individuals in each generation. Simultaneously, the probability of crossover P_c , probability of mutation P_m , and the maximum numbers of generations NG are also initialized.

Step 2: Evaluation.

After the initialization step, each chromosome is evaluated using a user-defined fitness function. The fitness value of each string is an index of the problem's design improvement suitability and the probability of survival of reproduction in genetic algorithms.

Step 3: Check termination criteria.

After the previous steps, the processes, from step 2 to 7, are repeated until the termination criteria are satisfied. The proposed algorithm is terminated if either one of the following conditions is satisfied:

1. The maximum number of generations is achieved, or
2. The same solution has not been changed for the present generation.

Step 4: Elitism mechanism.

In order to ensure the propagation of the elite chromosome, GAs use the *Elitism mechanism* [44,50]. This mechanism selects $P\%$ individuals, which have the best fitness values, to be the offspring of the next generation, while the remaining individuals execute the genetic operations (i.e., selection, crossover and mutation).

Step 5: Selection.

Selection is a process in which suitable chromosomes from the parents' populations for the next generation are chosen. In this step, the selection of this model is *tournament selection* [5,22]. It is both effective and computationally efficient. Pairs of chromosomes are selected at random to produce their own fitness values. The chromosomes with the best fitness values will be chosen. This step is repeated until the number of chromosomes selected is equal to the number of the population.

Step 6: Crossover.

The crossover operates by swapping corresponding segments of a string representation of the parents and extends the search for a new solution. Such positional bias [18] implies that the schemas with long-defining lengths suffer biased disruption. In order to reduce positional bias, this model uses *uniform crossover* [45], which can be disruptive especially to the early generations. Uniform crossover operates as shown in Fig. 2.

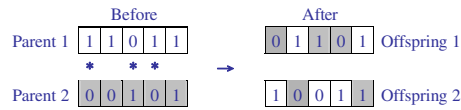


Fig. 2. The crossover operation for the proposed model.



Fig. 3. The mutation operation.

Step 7: Mutation.

The mutation is a GA mechanism. It randomly chooses a member of the population and changes one randomly chosen bit in its bit string representation. The mutation operation is shown in Fig. 3.

3. Proposed concepts and model

3.1. Proposed concepts

As stated in the previous section, past forecasting models exhibited three major drawbacks when used to predict stock market activity: (1) utility technical indicators used to forecast stock prices represent a subjective approach because investors attempt to make knowledgeable judgments about future price trends, based on the values of the technical indicators; (2) some prerequisites, such as statistical probability distribution [9], form the theoretical basis for constructing conventional forecasting models, but usually stock data do not follow a specific data distribution; and (3) rule presentations of stock data, produced from some data-mining algorithms, are difficult for non-professional investors to interpret [46] (i.e., the rules mined from ANN are not easily understandable). We argue that these drawbacks reduce the efficiency and applicability of the forecasting model, and to overcome these drawbacks, four novel methods are proposed in the processes of data-mining for stock price forecasting.

In the preprocess of data-mining, two objective approaches, CPDA (cumulative probability distribution approach) and MEPA (minimum entropy principle approach), are suggested to discretize each technical indicator (condition features) and daily price fluctuation (decision features) into linguistic values. There are advantages in using data discretization methods in preprocessing raw data. For example, the data dimension of a database can be reduced and simplified, and use of discrete features is usually more compact and shorter than the use of continuous ones [31]. Additionally, to reduce the amount of stock data and find the effective indicators related to the future stock price, a “correlation matrix” can be used to select, objectively, essential technical indicators from popular indicators, used in the stock market in the preprocess.

In the model-building process, a rough set (RS) algorithm is one proper data-mining algorithm suggested to extract forecasting rules from complex stock data. In recent data-mining techniques, RS methods have provided a basis for a predictive data-mining tool that is especially helpful in dealing with vague, incomplete and uncertain data used for decision-making. Three advantages can be discovered when applying RS methods: (1) RS theory can deal with the original datasets without any additional information or statistical assumptions, unlike the probability of statistics; (2) RS theory can discover important facts hidden in datasets and express them with decision rules of natural language; and (3) the results (rules) from a RS model are easily understood [16,23]. Because of these advantages, RS theory has become an important theory in the fields of artificial intelligence (AI), knowledge discovery in database (KDD), and data-mining (DM). Accordingly, a forecasting model using a RS algorithm to produce rules can overcome the limitations¹ of statistical methods for stock price forecasting, and the produced “if-then” rules can model the qualitative aspects of human knowledge applicable for investors.

Moreover, genetic algorithms (GAs) [2,28] are an effective and robust method for searching for proper solutions to the problems of very large spaces or dimensions in a variety of applications, and are particularly applicable in solving multi-parameter optimization problems. In the stock market, “unexpected events,” which will turbulently influence stock price, will sometimes occur. Although RS methods can extract rules effectively from past stock data, the rules to express the “unexpected events” have very low support value, and they are usually ignored when forecasting the future. Also, the “unexpected events” in the future may not be similar to those of the past. Therefore, to meet the “unexpected events,” we use GAs to produce “mutation rules,” which are evolved from rules extracted from RS methods, to improve classification accuracy and forecasting profit. Consequently, to deal with the unpredictable variations contained in stock markets, GAs are suggested to refine the extracted rules to enhance forecasting performance. The program of the proposed model can be downloaded on the web page [52].

¹ Some statistical distributions are presumed as a basis to construct mathematical models, but usually stock data do not follow a specific data distribution.

3.2. Proposed model

Based on the proposed concepts above, this paper suggests a hybrid model (the overall framework of which is shown as Fig. 4), which incorporates four novel data-mining methods (including CPDA, MEPA, RST and GAs) in the forecasting processes, and provides three major phases (including six processes), as noted below.

3.2.1. Preprocess

There are two works contained in this phase, as follows:

3.2.1.1. Data transformation and selection of essential technical indicators. In the data transformation procedure, one period of stock data is selected as an experimental stock dataset. This dataset contains five daily fundamental stock quantities (maximum price, minimum price, opening price, closing price, and stock trading volume) and the daily price fluctuation, which express the stock price change between any given day and the previous day (daily price fluctuation (t) = stock price (t) – stock price ($t - 1$)). The data of the five daily fundamental stock quantities is converted into the data of several popular technical analysis indicators [26]: moving average (MA), momentum (MTM), stochastic %K (%K), stochastic %D (%D), relative strength index (RSI), psychology line (PSY), Williams' percent range (%R), volume ratio (VR), volume (Volume), and accumulative ratio (AR) [41].

In the procedure of selecting essential technical indicators, the popular indicators are analyzed by a “correlation matrix” to examine their relationship degree to daily price fluctuations, and to select from the popular technical analysis indicators those essential indicators which are highly related to daily price fluctuations. The selection approach employs a statistical method, Pearson correlation with two-tailed tests, to select the essential technical analysis indicators. Pearson's Correlation Coefficient is usually signified by γ . The statistical significance of r is tested using a t -test. The hypothesis for this test is: $H_0 : \gamma = 0$. A low p -value for this test (less than 0.05, for example) means that there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. For example, price fluctuation is related significantly with MA-5, MTM-5, %K-5, %D-5, RSI-5, PSY-5, %R-5, VR-5, volume, and AR-5 (the marker “***” in the last column in Table 19).

3.2.1.2. Data discretization by CPDA and MEPA. In this procedure, two discretization methods, CPDA and MEPA, are utilized to construct membership functions for these selected features and two numeric stock datasets (one dataset consisting of

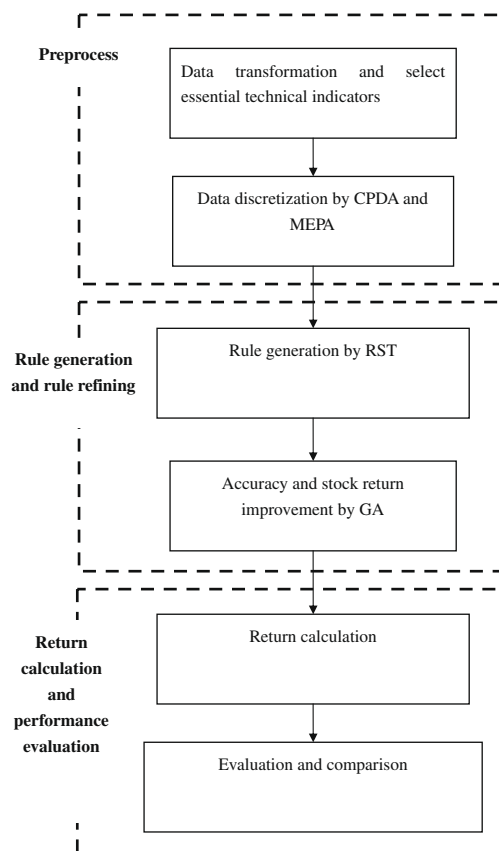


Fig. 4. The framework of proposed model.

essential technical indicators, the other consisting of daily price fluctuations) are fuzzified into linguistic stock datasets (the converted processes of daily price fluctuations, from numeric stock datasets to linguistic stock datasets, are demonstrated in Table 8; the converted processes of a technical indicator, MA-5, from numeric stock datasets to linguistic stock datasets are demonstrated in Table 10). The dataset of essential technical indicators is discretized by MEPA and used as a conditional attribute. The dataset of the daily price fluctuation is discretized by CDPA and employed as a decision attribute.

3.2.2. Rule generation and rule refining

There are two operations contained in this phase, as follows:

3.2.2.1. Rule generation by RST. This procedure applies a toolset, based on RST (LEM2) [38,53], to extract preprocess phase rules from the linguistic stock datasets, and to select the extracted rules with high support value as the initial population for GA operations. To detail this procedure clearly, the computation steps are introduced in Fig. 5.

3.2.2.2. Accuracy and stock return improvement by GA. This procedure employs GA to refine the extracted rules, thus, improving classification accuracy and stock return. To detail this procedure clearly, the computation steps are introduced in Fig. 6.

3.2.3. Return calculation and performance evaluation

There are two processes contained in this phase, as follows:

3.2.3.1. Return calculation. Each selected experimental dataset is split into two subdatasets: training and testing datasets. The refined rules, extracted by the proposed model from the training dataset, are used for forecasting the testing dataset and calculating stock return, based on the trading strategies, as follows: (i) “buy on open and sell on close,” when the selected rule indicates that the forecasted price fluctuation of the next day is “going up;” (ii) “sell on open and buy on close,” when the selected rule indicates that forecasted price fluctuation of the next day is “going down;” and (iii) “no transaction,” when the selected rule indicates that the forecasted price fluctuation of the next day is “staying flat.” Based on the refined rules and the strategies, the returns for the proposed model can be calculated and summarized.

3.2.3.2. Evaluation and comparison. In this procedure, two types of performance evaluations, forecasting accuracy and stock return, are employed. In forecasting accuracy evaluation, RST and GA models, using the same preprocess conditions as with the proposed model, are employed as comparison models. In forecasting accuracy evaluation, the Buy-and-Hold method (defined in Eq. (21)), RST and GA models are employed as comparison models.

To detail the proposed model, each step of the proposed model is described, as follows:

Step 1: Data transformation and selection of essential technical indicators.

In this step, technical indicators are used as condition features, and the next-day price fluctuation ($DPF_{next\ day}$), defined in Eq. (16), is used as a decision feature. The technical indicators are generated from five fundamental quantities (opening price, highest price, lowest price, closing price and trading volume) [28,41]. In order to choose those essential technical indicators

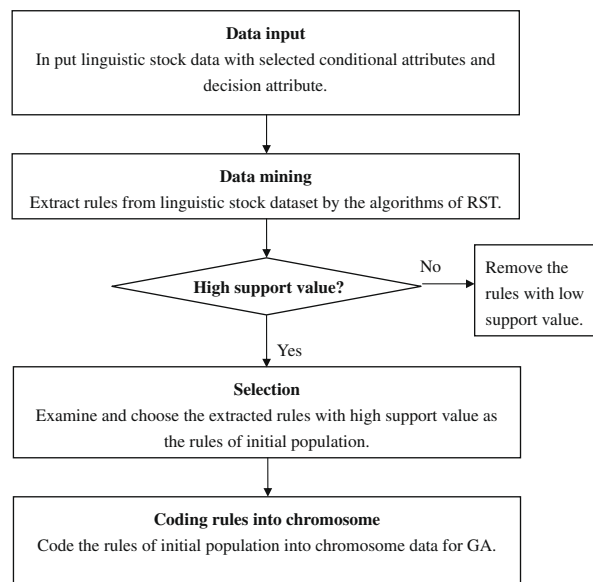


Fig. 5. Rule generation by RST.

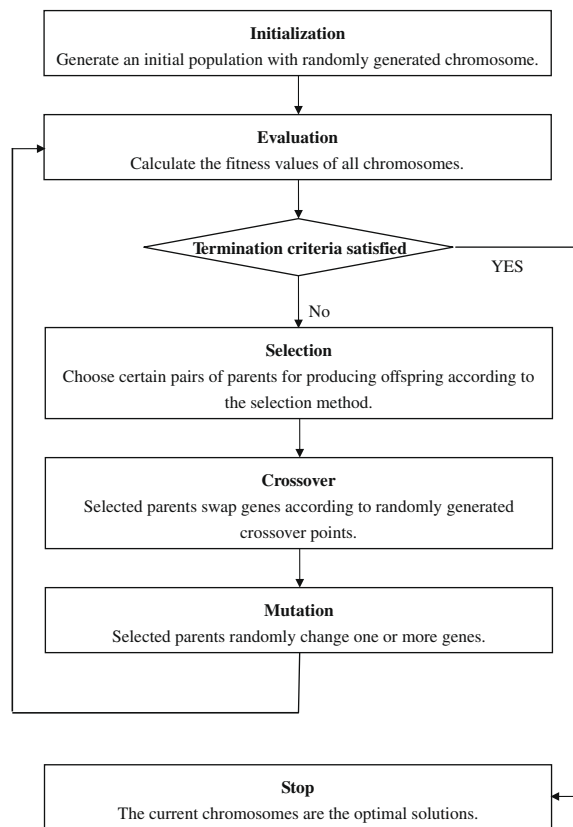


Fig. 6. Accuracy and stock return improvement by GA operations.

as condition features that are highly related to the next-day price fluctuation, a correlation matrix is employed to select the essential indicators from several popular indicators (see Table 19).

$$\text{Daily price fluctuation } (t+1) = \text{closing price } (t+1) - \text{opening price } (t+1) \quad (16)$$

Step 2: Data discretization by CPDA and MEPA

In this step, CPDA is applied to discretize the next-day price fluctuation (decision attribute), and MEPA is used to discretize essential indicators (condition technical attributes). Because price fluctuations in the stock market and stock returns follow normal distribution closely [3,47], CPDA is an appropriate method to discretize stock price data. In this model, the next-day price fluctuation (decision attribute) is defined by three linguistic values: *Up*, *Fair* and *Down*. Additionally, to partition the universe of discourse of each essential technical indicator based on its data characteristic, this step employs MEPA to discretize condition features [11,12]. Granulating features is one important step for data-mining processes, especially for RST, and by this step, the data dimensions of a database can be reduced and simplified.

In this paper, we discretize each condition attribute with 15 linguistic terms, because by using the TSMC as experimental datasets, the accuracy of the proposed model performs best with 15 linguistic terms (see Tables 1 and 2), compared with 2, 3 and 7 linguistic terms.

Step 3: Rule generation by RST.

This step utilizes the algorithms of RST to mine rules from a training stock dataset. The extracted rules are in the form of “if-then” with specific condition attribute values and a decision attribute value, as follows:

Table 1

The accuracy of proposed model with different linguistic terms of condition feature for TSMC dataset (1999).

| Linguistic terms of decision feature | Linguistic terms of condition feature | | | |
|--------------------------------------|---------------------------------------|-------|-----|------|
| | 2 | 3 | 7 | 15 |
| 3 (Up, Fair, and Down) | 0.278 | 0.316 | 0.5 | 0.55 |

Table 2

The accuracy of proposed model with different linguistic terms of condition feature for TAIEX dataset (2000).

| Linguistic terms of decision feature | Linguistic terms of condition attribute | | | |
|--------------------------------------|---|-----|-------|-------|
| | 2 | 3 | 7 | 15 |
| 3 (Up, Fair, and Down) | 0.467 | 0.5 | 0.556 | 0.619 |

If $RSI = L8$ and $Volume$, then $DPF_{next\ day} = Up$

This rule states that if the linguistic value of RSI is $L8$ (medium), and the linguistic value of the $Volume$ is $L9$ (more or less high), then the next day price fluctuation is Up .

Step 4: Accuracy and stock return improvement by GA.

This step uses GAs to refine the extracted rules produced by RST from step 3 to improve forecasting accuracy and stock return. The substeps of performance improvement are described below.

Step 4.1: Generate initial chromosomes.

The rules produced by RST are coded into chromosomes as an initial population, as in Table 3. Each condition feature is coded with a value from “0” to “15,” where “0” signifies that the condition feature is not contained in the extracted rule, and the coded values, from “1” to “15,” represent the linguistic value of this condition attribute, from $L1$ to $L15$, respectively. Additionally, the decision feature is encoded from “1” to “3,” where “1” denotes that the daily price fluctuation of the next day is *Down*, “2” *Flat*, and “3” *Up*.

The proposed model evaluates each chromosome by a user-defined fitness function, as shown in Eq. (17).

$$\text{Fitness function} = \frac{\text{The number of observations classified correctly by the chromosome}}{\text{All of observations in the training data}} \quad (17)$$

Step 4.2: Perform genetic operations.

In this step, genetic operations (selection, crossover and mutation) are employed to refine the rules produced by RST from step 3. In order to ensure the propagation of the elite chromosome, this step implements the genetic operations with the Elitism mechanism [44,50]. This mechanism selects 0.2% of the individuals with the best fitness values to be the offspring of the next generation, while the remaining individuals execute the genetic operations. Besides this, to enhance genetic operations in this substep, the rules with lower rule support value (rule support = 1) are removed. Table 4 demonstrates the parameter settings of GA in this substep.

Step 5: Return calculation.

This step calculates stock return by using the refined rules from step 4 in forecasting the stock price. The inference method based on the refined rule base is described in the following algorithm:

Algorithm 1. Return calculation

The initial forecasting price fluctuation ($t+1$) is set as “Flat”

Do search each rule contained in the refined rule base

If the linguistic values of the actual essential technical indicators in the testing dataset at time t are **equal to** the conditional parts of the rule, **then** the linguistic value of the forecasting price fluctuation ($t+1$) is **equal to** the decision part of the rule

While not end of the refined rule base

There are three trading strategies provided in this step.

(1) “Buy on open and sell on close,” when the selected rule tells that the forecasting price fluctuation ($t+1$) is *Up*; (2) “sell on open and buy on close,” when the forecasting price is *Down*; and (3) “no transaction,” when the forecasting price fluctuation is *Flat*.

Based on the inference method and trading strategies above, stock return is calculated by Eq. (18), when the forecasted price fluctuation ($t+1$) is *Up*, or by Eq. (19), when the forecasted price fluctuation ($t+1$) is *Down*. In Eqs. (18) and (19), *Unit* means the quantity of shares.

Table 3

The structure of the chromosomes for the proposed model.

| Feature name | Feature_1 | Feature_2 | ... | Feature_n – 1 | Feature_n | $DPF_{next\ day}$ |
|--------------|-----------|-----------|-----|---------------|-----------|-------------------|
| Coded value | 0–15 | 0–15 | ... | 0–15 | 0–15 | 1–3 |

Table 4

Parameter settings of GA.

| | |
|-----------------------|-------------------------|
| Population size | 1000 |
| Number of generations | 100 |
| Initialization method | Integer-encoding method |
| Percentage of elite | 0.2 |
| Selection method | Tournament selection |
| Crossover method | Uniform crossover |
| Crossover rate | 0.8 |
| Mutation method | Single point mutation |
| Mutation rate | 0.05 |

$$\text{Return}_{\text{Up}}(t+1) = (\text{closing price}(t+1) - \text{opening price}(t+1)) \times \text{unit} \quad (18)$$

$$\text{Return}_{\text{Down}}(t+1) = (\text{opening price}(t+1) - \text{closing price}(t+1)) \times \text{unit} \quad (19)$$

Step 6: Evaluation and comparison.

In this step, to evaluate forecasting performance carefully, accuracy and stock return are employed as performance indicators. This step calculates the stock return based on the three trading strategies above and sums up the stock return of all transactions. The data-mining model's accuracy is defined in Eq. (20).

$$\text{Accuracy}(T) = \frac{\sum_{i=1}^{|T|} \text{correct classification}(t_i)}{|T|}, \quad t_i \in T \quad (20)$$

$$\text{Correct classification}(t_i) = \begin{cases} 1, & \text{if classify}(t_i) = \text{actual price fluctuation}(t_i) \\ 0, & \text{otherwise} \end{cases}$$

where T is the set of data instances that are classified by the proposed model, $t_i \in T$; actual price fluctuation (t_i) is the classification for each data instance of actual price fluctuation labeled as one of three linguistic values, $Up(L_3)$, $Fair(L_2)$, and $Down(L_1)$; and $\text{classify}(t_i)$ returns the classification of data instance t_i by the proposed model.

In a performance comparison, RST, GAs and the “Buy-and-Hold” approach are used as comparison methods to evaluate the proposed model. To examine the performance difference between the proposed hybrid model and the data-mining models using a single method, RST [20,48] and GAs [2,28] are used as comparison models. Moreover, the “Buy-and-Hold” approach [8,40] is a common strategy for investors in the stock market, and therefore, this approach is used as another comparison model. The stock return of the “Buy-and-Hold” approach is defined in Eq. (21), where *Unit* means the quantity of shares.

$$\text{Return}_{\text{Buy-and-hold}} = (\text{closing price}_{\text{The last day of investing period}} - \text{opening price}_{\text{The first day of investing period}}) \times \text{unit} \quad (21)$$

To detail the proposed model, an empirical case study is introduced in the following section.

3.3. Empirical case study

To demonstrate the proposed model clearly, in this section, a one-year period of Taiwan Semiconductor Manufacturing Company (TSMC) stock data, from 1999/06/23 to 2000/05/11 (see Fig. 7, retrieved from Taiwan Stock Exchange Corporation [54]), is employed as an experimental dataset to introduce the proposed algorithm step by step. Because the stock price on the “ex-dividend day” fluctuates violently [19], this study removes the stock data of the ex-dividend day from the dataset (the dataset selection is from 1999/06/23, the day after the ex-dividend day, to 2000/05/11, the day before the ex-dividend day). The previous 10 months of stock data, from 1999/06 to 2000/03, is used for training, and the rest, from 2000/03 to 2000/05, is used for testing.



Fig. 7. The stock price of TSMC from 1999/06/23 to 2000/05/11.

By using the experimental stock dataset above, each step contained in the three forecasting phases of the proposed model ((I) Preprocess; (II) Rule generation and rule refining; and (III) Return calculation and performance evaluation) is introduced and demonstrated, as follows:

Step 1: Data transformation and selection of essential technical indicators.

In this step, the five fundamental quantities of TSMC stock data (opening price, highest price, lowest price, closing price and trading volume) are demonstrated as Table 5, and are transferred into popular technical indicators by their corresponding equations [28,41]. The eight essential technical indicators (condition features), selected by a correlation matrix for TSMC, are described, as follows: accumulative ratio (AR), moving average (MA), psychology line (PSY), relative strength index (RSI), stochastic %D (%D), stochastic %K (%K), volume, and Williams' percent range (%R) (see Table 6). The daily price fluctuation of the next day (decision features) is produced by Eq. (16) (shown in the last column of Table 6).

Step 2: Data discretization by CPDA and MEPA.

This step utilizes CPDA to discretize decision features and uses MEPA to discretize condition features. There are two sub-steps, as follows:

Use CPDA to discretize decision features

This substep uses CPDA to define the numeric intervals for three linguistic values, *Down* (L_1), *Flat* (L_2), and *Up* (L_3), used for decision features (daily price fluctuation). Table 7 lists the numeric intervals for the linguistic values. Based on these intervals, the membership function for the three linguistic values is established and shown in Fig. 8. According to the membership functions shown in Fig. 8, three membership functions and three membership degrees are produced for each datum (daily price fluctuation on the next day). In the fuzzification procedure, each datum is labeled as a linguistic value among three linguistic values, based on the maximum membership degree. Table 8 demonstrates the daily price fluctuation, the corresponding membership degrees for the three membership functions (μ_{Down} , μ_{Flat} and μ_{Up}), and the labeled linguistic values for the partial TSMC stock data.

Use MEPA to discretize condition features

Table 5

The partial five fundamental quantities of TSMC.

| Date | Opening price | Highest price | Lowest price | Closing price | Volume |
|------------|---------------|---------------|--------------|---------------|--------|
| 1999/06/23 | 126.5 | 127 | 124 | 125 | 44,713 |
| 1999/06/24 | 124 | 130 | 123 | 129 | 69,778 |
| 1999/06/25 | 121 | 127.5 | 121 | 122.5 | 57,865 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2000/05/10 | 194 | 194 | 188 | 188 | 16,876 |
| 2000/05/11 | 182 | 184 | 177 | 179 | 29,310 |
| 2000/05/12 | 182 | 185 | 179 | 185 | 38,309 |

Table 6

The partial instances of forecasting technical indicators data.

| Date | MA-5 | RSI-5 | K-5 | D-5 | R-5 | PSY-5 | AR-5 | Volume | DPF _{next day} |
|------------|-------|-------|-------|-------|--------|-------|------|--------|-------------------------|
| 1999/06/23 | 126 | 72.22 | 16.67 | 5.56 | 50.00 | 60 | 0.89 | 44,713 | 5 |
| 1999/06/24 | 127.6 | 72.22 | 44.44 | 18.52 | 0.00 | 60 | 1.13 | 69,778 | 1.5 |
| 1999/06/25 | 126.3 | 30.30 | 29.63 | 22.22 | 100.00 | 40 | 1.45 | 57,865 | 0 |
| 1999/06/28 | 125.4 | 34.48 | 19.75 | 21.40 | 100.00 | 40 | 2.69 | 29,340 | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2000/05/08 | 190 | 50.00 | 63.97 | 54.75 | 0.00 | 40 | 2.22 | 14,605 | 0 |
| 2000/05/09 | 191 | 72.73 | 75.98 | 61.83 | 0.00 | 40 | 2.11 | 13,523 | −6 |
| 2000/05/10 | 191.2 | 53.33 | 50.65 | 58.10 | 100.00 | 40 | 1.25 | 16,876 | −3 |
| 2000/05/11 | 188.6 | 15.79 | 33.77 | 49.99 | 100.00 | 20 | 0.50 | 29,310 | 3 |

Table 7

The lower/upper boundary cumulative probability and linguistic intervals.

| Linguistic value | Cumulative probability | | Universe of discourse U | | | |
|-----------------------|------------------------|----------|---------------------------|----------|----------------|--------------------|
| | P_{LB} | P_{UB} | Lower boundary | Midpoint | Upper boundary | Length of interval |
| <i>Down</i> (L_1) | 0 | 0.333 | −14.46 | −7.9 | −1.35 | 13.11 |
| <i>Flat</i> (L_2) | 0.167 | 0.667 | −3.2 | −0.79 | 1.63 | 4.83 |
| <i>Up</i> (L_3) | 0.5 | 1 | 0.14 | 10.3 | 20.46 | 20.32 |

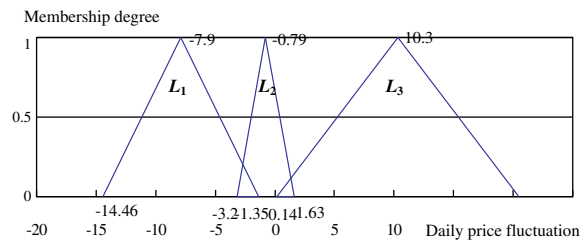


Fig. 8. Membership function of decision feature.

Table 8

The membership degree of decision feature based on CPDA.

| Date | Daily price fluctuation ($DPF_{next\ day}$) | Membership degree for linguistic value | | | Labeled linguistic value |
|------------|--|--|--------------|------------|--------------------------|
| | | μ_{Down} | μ_{Flat} | μ_{Up} | |
| 1999/06/23 | 5 | 0 | 0 | 0.48 | Up |
| 1999/06/24 | 1.5 | 0 | 0.05 | 0.13 | Up |
| 1999/06/25 | 0 | 0 | 0.67 | 0 | Flat |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2000/05/08 | 0 | 0 | 0.67 | 0 | Flat |
| 2000/05/09 | −6 | 0.71 | 0 | 0 | Down |
| 2000/05/10 | −3 | 0.25 | 0.08 | 0 | Down |
| 2000/05/11 | 3 | 0 | 0 | 0.28 | Up |

Table 9

Linguistic intervals for 15 linguistic values produced by MEPA (MA-5).

| Linguistic value | Universe of discourse | | | |
|------------------|-----------------------|----------|----------------|--------------------|
| | Lower boundary | Midpoint | Upper boundary | Length of interval |
| L1 | 104 | 123.1 | 125.7 | 21.7 |
| L2 | 123.1 | 125.7 | 130.4 | 7.3 |
| L3 | 125.7 | 130.4 | 136.05 | 10.35 |
| L4 | 130.4 | 136.05 | 141.85 | 11.45 |
| L5 | 136.05 | 141.85 | 155.3 | 19.25 |
| L6 | 141.85 | 155.3 | 155.9 | 14.05 |
| L7 | 155.3 | 155.9 | 157.4 | 2.1 |
| L8 | 155.9 | 157.4 | 177.8 | 21.9 |
| L9 | 157.4 | 177.8 | 184.7 | 27.3 |
| L10 | 177.8 | 184.7 | 190.2 | 12.4 |
| L11 | 184.7 | 190.2 | 208.6 | 23.9 |
| L12 | 190.2 | 208.6 | 210.1 | 19.9 |
| L13 | 208.6 | 210.1 | 211.2 | 2.6 |
| L14 | 210.1 | 211.2 | 213.3 | 3.2 |
| L15 | 211.2 | 213.3 | 215 | 3.8 |

Notes: L1 is very very very very low, L2 is very very very very low, L3 is very very very low, L4 is very very low, L5 is very low, L6 is low, L7 is more or less low, L8 is medium, L9 is more or less high, L10 is high, L11 is very high, L12 is very very high, L13 is very very very high, L14 is very very very very high, L15 is very very very very high.

In this substep, the interval length of the 15 linguistic values for each condition feature is defined by MEPA (see Table 9), and thus, establishing the membership function of MEPA (see Fig. 9). In Table 10, the degree of membership for each datum is calculated by the membership function of MEPA, and the last column is the linguistic value of the datum, which is fuzzified, based on the maximum membership of the datum.

From the results of two data discretization methods, CPDA and MEPA, the linguistic condition features and the decision feature are demonstrated in Table 11.

Step 3: Rule generation by RST.

This step utilizes RST to construct decision rules from linguistic values in the above step (see Table 11). Table 12 shows the partial rules and the rule support value. “Support” refers to how many records meet the generated decision rules in the stock dataset. For example, “Rule 1: If $AR-5 = L6$ and $PSY-5 = L9$ and $RSI-5 = L2$ and Volume = $L3$ and $K-5 = L7$, then $DPF_{next\ day} = Up$ (support = 3)” indicates that there are three training records that meet the criteria of “Rule 1.”

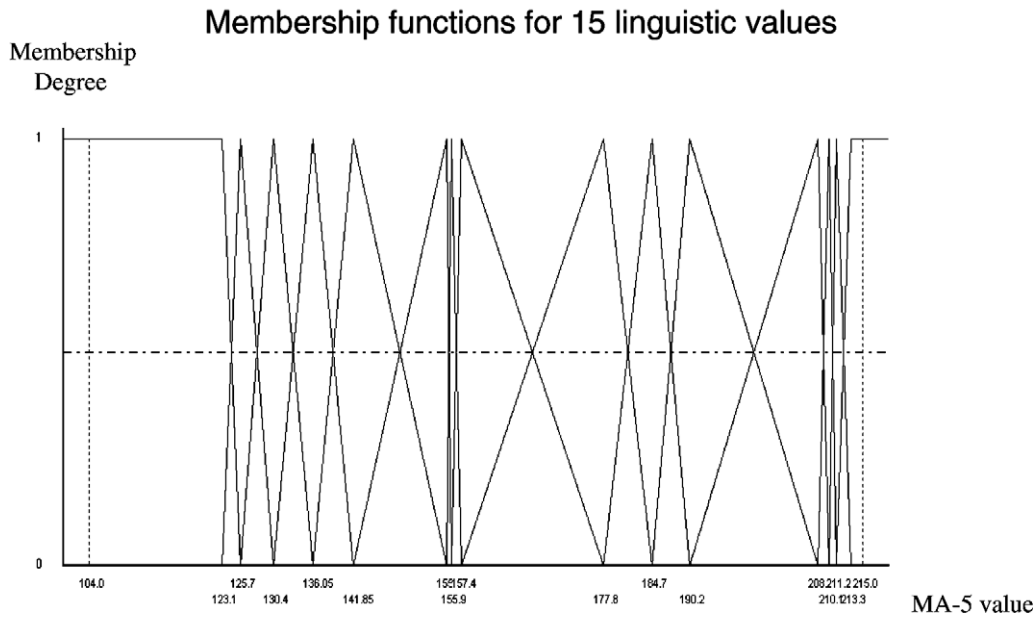


Fig. 9. Membership functions for 15 linguistic values using the attribute of MA-5.

Table 10

The partial MEPA membership degrees of MA-5 feature for TSMC.

| Date | Technical indicator value (MA-5) | Membership degree for linguistic value | | | | | | | | | | | | | | | Labeled linguistic value |
|------------|----------------------------------|--|------------|------------|------------|------------|------------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------------------|
| | | μ_{L1} | μ_{L2} | μ_{L3} | μ_{L4} | μ_{L5} | μ_{L6} | μ_{L7} | μ_{L8} | μ_{L9} | μ_{L10} | μ_{L11} | μ_{L12} | μ_{L13} | μ_{L14} | μ_{L15} | |
| 1999/06/23 | 126 | 0 | 0.94 | 0.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | L2 |
| 1999/06/24 | 127.6 | 0 | 0.6 | 0.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | L2 |
| 1999/06/25 | 126.3 | 0 | 0.87 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | L2 |
| 1999/06/28 | 125.4 | 0.12 | 0.88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | L2 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2000/05/08 | 190 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 | 0.96 | 0 | 0 | 0 | 0 | L11 |
| 2000/05/09 | 191 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.96 | 0.04 | 0 | 0 | 0 | L11 |
| 2000/05/10 | 191.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.95 | 0.05 | 0 | 0 | 0 | L11 |
| 2000/05/11 | 188.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.29 | 0.71 | 0 | 0 | 0 | 0 | L11 |

Note: LV donates linguistic value.

Step 4: Accuracy and stock return improvement by GAs.

This step uses genetic algorithms (GAs) to improve the accuracy and stock return of rules that are produced by RST. There are two substeps in this step, as follows:

Step 4.1: Generate initial chromosomes.

This substep encodes each rule that is produced by RST in Step 3 as a chromosome in an initial population and evaluates each chromosome in each population using Eq. (17). For example, this substep encodes Rule 1 in Table 12 into a chromosome, as shown in Table 13.

In Table 13, the coded value of *RSI-5* is “2,” which represents that the linguistic value of *RSI-5* is *L2*. In the same way, the coded values of *K-5*, *PSY-5*, *AR-5* and *Volume* are 7, 9, 6 and 3, which show that the linguistic values are *L7*, *L9*, *L6* and *L3*, respectively. The coded values for *MA-5*, *D-5* and *R-5* are “0,” which denote that the rule does not contain the three condition features. Further, the forecasting stock price, DPF_{next_day} , means “Up” because the coded value is “3” (1 for Down, 2 for Fair, 3 for Up). In addition, the proposed model calculates a fitness value to evaluate each chromosome by Eq. (17).

Step 4.2: Perform genetic operations.

This substep refines the initial rules produced by using genetic operators, such as selection, crossover and mutation. Table 14 shows the partial rules refined by GAs and the rule support.

In this paper, the proposed model employs a rule-filter, which selects useful rules for which rule support is greater than 1, and which deletes rules with low rule support (rule support = 1). Moreover, genetic algorithms and the proposed model utilize the same rule-filter condition (i.e., rule support is more than 1), the parameter settings of which are shown in Table 4.

Table 11

The partial linguistic value of TSMC.

| Date | MA-5 | RSI-5 | K-5 | D-5 | R-5 | PSY-5 | AR-5 | Volume | DPF _{next day} |
|------------|------|-------|-----|-----|-----|-------|------|--------|-------------------------|
| 1999/06/23 | L2 | L10 | L3 | L1 | L9 | L11 | L6 | L14 | Up |
| 1999/06/24 | L2 | L10 | L7 | L4 | L1 | L11 | L6 | L15 | Up |
| 1999/06/25 | L2 | L1 | L7 | L4 | L15 | L9 | L7 | L15 | Flat |
| 1999/06/28 | L2 | L2 | L3 | L4 | L15 | L9 | L8 | L6 | Flat |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2000/05/08 | L11 | L2 | L8 | L6 | L1 | L9 | L7 | L2 | Flat |
| 2000/05/09 | L11 | L11 | L12 | L6 | L1 | L9 | L7 | L2 | Down |
| 2000/05/10 | L11 | L2 | L8 | L6 | L15 | L9 | L7 | L2 | Down |
| 2000/05/11 | L11 | L1 | L7 | L6 | L15 | L8 | L4 | L6 | Up |

Table 12

The partial rules generated by rough set theory.

| No. | Rules | Rule support |
|-----|--|--------------|
| 1 | If AR-5 = L6 and PSY-5 = L9 and RSI-5 = L2 and Volume = L3 and K-5 = L7 then DPF _{next day} = Up | 3 |
| 2 | If R-5 = L1 and AR-5 = L7 and RSI-5 = L13 and K-5 = L13 then DPF _{next day} = Up | 3 |
| 3 | If AR-5 = L6 and PSY-5 = L11 and MA-5 = L2 then DPF _{next day} = Up | 3 |
| 4 | If R-5 = L15 and RSI-5 = L1 and K-5 = L2 and MA-5 = L11 then DPF _{next day} = Up | 3 |
| ⋮ | ⋮ | ⋮ |
| 60 | If PSY-5 = L9 and AR-5 = L6 and Volume = L2 and MA-5 = L5 and RSI-5 = L8 then DPF _{next day} = Down | 1 |
| 61 | If AR-5 = L7 and PSY-5 = L11 and MA-5 = L4 and RSI-5 = L10 then DPF _{next day} = Down | 1 |
| 62 | If MA-5 = L12 and Volume = L1 and RSI-5 = L11 then DPF _{next day} = Down | 1 |
| 63 | If RSI-5 = L1 and D-5 = L5 and MA-5 = L12 and K-5 = L6 then DPF _{next day} = Down | 1 |

Table 13

The structure of the chromosomes.

| Feature name | MA-5 | RSI-5 | K-5 | D-5 | R-5 | PSY-5 | AR-5 | Volume | DPF _{next day} |
|--------------|------|-------|-----|-----|-----|-------|------|--------|-------------------------|
| Coded value | 0 | 2 | 7 | 0 | 0 | 9 | 6 | 3 | 3 |

Note: The values of each condition features: 1 represent that the linguistic value is L1, and so on.

The values of DPF_{next day}: 1 is Down, 2 is Flat, 3 is Up.

Table 14

The partial rules refined by genetic algorithms.

| No. | Rules | Rule Support |
|-----|---|--------------|
| 1 | If MA-5 = L8 and K-5 = L1 and R-5 = L15 then DPF _{next day} = Down | 26 |
| 2 | If MA-5 = L8 and K-5 = L1 and D-5 = L7 then DPF _{next day} = Down | 22 |
| 3 | If K-5 = L1 and PSY-5 = L9 and AR-5 = L15 then DPF _{next day} = Up | 20 |
| 4 | If D-5 = L9 and R-5 = L1 and Volume = L2 then DPF _{next day} = Up | 19 |
| ⋮ | ⋮ | ⋮ |
| 281 | If RSI-5 = L15 and D-5 = L15 and AR-5 = L15 then DPF _{next day} = Flat | 1 |
| 282 | If D-5 = L8 and R-5 = L1 and PSY-5 = L8 then DPF _{next day} = Down | 1 |
| 283 | If D-5 = L4 and R-5 = L15 and Volume = L2 then DPF _{next day} = Down | 1 |
| 284 | If RSI-5 = L8 and R-5 = L1 and PSY-5 = L9 and Volume = L2 then DPF _{next day} = Up | 1 |

Step 5: Return calculation.

This step calculates the stock return by using the rules produced from step 4 in forecasting stock prices. The stock return is compared with the listed model in the next step.

Step 6: Evaluation and comparison.

To evaluate the proposed model, the accuracy and stock return of the proposed model are compared with those of the three listed methods: RST, GAs and the “Buy-and-Hold” approach (see Table 15). The accuracy and stock return for the comparison models and the proposed model are listed in Table 16, from which we can see that the proposed model outperforms the listed models.

Table 15

The stock returns of “Buy-and-Hold” in testing period.

| Data period | Opening price | Closing price | Buy-and-Hold return (<i>unit</i>) |
|-----------------------|---------------|---------------|-------------------------------------|
| 2000/03/22–2000/05/11 | 194 | 179 | –15 |

Note: *Unit* means the quantity of share.**Table 16**

The forecasting performance comparison for TSMC.

| Data period | Accuracy | | | Stock return (<i>unit</i>) | | | |
|-----------------------|-----------|--------------------|----------------|------------------------------|-----------|--------------------|----------------|
| | Rough set | Genetic algorithms | Proposed model | Buy-and-Hold | Rough set | Genetic algorithms | Proposed model |
| 2000/03/22–2000/05/11 | 0.545 | 0.42 | 0.55 | –15 | 15 | 10 | 16 |

Note: Some literatures accept the accuracy rate which less than or close to 0.5 in forecasting price fluctuation of stocks [8,30].

4. Model verification

To verify the proposed model, this section provides two types of performance evaluations: (I) *Forecasting accuracy evaluation*: produce the accuracy of the proposed model based on Eq. (20) and provide two other comparison models, which use only one data-mining method, RST or GA, to produce forecasts under the same preprocess conditions as those of the proposed model; and (II) *Stock return evaluation*: produce the stock return based on Eqs. (18) and (19), and provide three other comparison models: Buy-and-Hold (defined in Eq. (21), RST and GAs.

In this experiment, a six-year period of the TAIEX (Taiwan Stock Exchange Capitalization Weighted Stock Index) stock index, from 2000/01/04 to 2005/12/30, was selected as an experimental database, and was divided into 6 datasets by year. The previous 10-month period of the stock index, from January to October, was used for training, and the rest, from November to December, was used for testing.

The accuracy for the three models (RST, GAs and the proposed model) is listed as Table 17, and the comparisons show that the proposed model outperforms the other two listed models. Further, from the stock return comparisons shown in Table 18,

Table 17

The accuracy comparisons for three models (TAIEX).

| Year | Model | | |
|---------|------------------|--------------------|--------------------|
| | Rough set theory | Genetic algorithms | Proposed model |
| 2000 | 0.602 | 0.574 | ^a 0.619 |
| 2001 | 0.615 | 0.581 | 0.628 ^a |
| 2002 | 0.53 | 0.523 | 0.593 ^a |
| 2003 | 0.512 | 0.535 | 0.582 ^a |
| 2004 | 0.571 | 0.556 | 0.614 ^a |
| 2005 | 0.488 | 0.51 | 0.568 ^a |
| Average | 0.553 | 0.547 | 0.601 ^a |

^a Maximum accuracy among three models.**Table 18**The stock returns comparisons for four models (TAIEX, *unit*).

| Year | Model | | | | |
|---------|---------------------|------------------|--------------------|----------------------|--------------|
| | Buy-and-Hold | Rough set theory | Genetic algorithms | Proposed model | Profit order |
| 2000 | –813.21 | 1590.68 | 1817.29 | ^a 2271.03 | 1 |
| 2001 | 1612.16 | 923.9 | 1272.05 | 1683.51 ^a | 2 |
| 2002 | –144.24 | 304.28 | 353.56 | 421.82 ^a | 4 |
| 2003 | –163.62 | 70.78 | 247 | 336.25 ^a | 5 |
| 2004 | 414.04 | 196.06 | 481.65 | 780.26 ^a | 3 |
| 2005 | 745.09 ^a | 105.85 | 163.8 | 210.83 | 6 |
| Average | 275.04 | 531.93 | 722.56 | 950.62 ^a | |

Note: *Unit* means the quantity of one share.^a Maximum return among four models.

Table 19
Correlations of technical indicators for TAIEX.

| | | MA-5 | MTM-5 | %K-5 | %D-5 | RSI-5 | PSY-5 | %R-5 | VR-5 | Volume | AR-5 | Price fluctuation |
|-------------------|---------------------|----------|----------|----------|----------|----------|----------|-----------|----------|----------|----------|-------------------|
| MA-5 | Pearson correlation | 1 | .741*** | .753*** | .472*** | .797*** | .818*** | -.835*** | .676*** | 1.000*** | .251*** | .169(***) |
| | Sig. (2-tailed) | | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .009 |
| MTM-5 | Pearson correlation | .741*** | 1 | .903*** | .714*** | .880*** | .889*** | -.713*** | .861*** | .741*** | .299*** | .112 |
| | Sig. (2-tailed) | .000 | | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .088 |
| %K-5 | Pearson correlation | .753*** | .903*** | 1 | .860*** | .835*** | .836*** | -.781*** | .768*** | .753*** | .223*** | .133(**) |
| | Sig. (2-tailed) | .000 | .000 | | .000 | .000 | .000 | .000 | .000 | .000 | .001 | .042 |
| %D-5 | Pearson correlation | .472*** | .714*** | .860*** | 1 | .713*** | .705*** | -.424*** | .585*** | .472*** | .074 | .134(**) |
| | Sig. (2-tailed) | .000 | .000 | .000 | | .000 | .000 | .000 | .000 | .000 | .260 | .041 |
| RSI-5 | Pearson correlation | .797*** | .880*** | .835*** | .713*** | 1 | .973*** | -.619*** | .791*** | .797*** | .232*** | .152(**) |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | | .000 | .000 | .000 | .000 | .000 | .020 |
| PSY-5 | Pearson correlation | .818*** | .889*** | .836*** | .705*** | .973*** | 1 | -.627*** | .796*** | .818*** | .269*** | .128(**) |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | | .000 | .000 | .000 | .000 | .047 |
| %R-5 | Pearson correlation | -.835*** | -.713*** | -.781*** | -.424*** | -.619*** | -.627*** | 1 | -.594*** | -.835*** | -.232*** | -.126(**) |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | | .000 | .000 | .000 | .049 |
| VR-5 | Pearson correlation | .676*** | .861*** | .768*** | .585*** | .791*** | .796*** | -.594*** | 1 | .676*** | .367*** | .109 |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | | .000 | .000 | .097 |
| Volume | Pearson correlation | 1.000*** | .741*** | .753*** | .472*** | .797*** | .818*** | -.835*** | .676*** | 1 | .251*** | .169*** |
| | Sig. (2-tailed) | .000 | .000 | .000 | .000 | .000 | .000 | .000 | .000 | | .000 | .009 |
| AR-5 | Pearson correlation | .251*** | .299*** | .223*** | .074 | .232*** | .269*** | -.232*** | .367*** | .251*** | 1 | .137(**) |
| | Sig. (2-tailed) | .000 | .000 | .001 | .260 | .000 | .000 | .000 | .000 | .000 | | .035 |
| Price fluctuation | Pearson correlation | .169*** | .112 | .133(**) | .134(**) | .152(**) | .128(**) | -.126(**) | .109 | .169*** | .137(**) | 1 |
| | Sig. (2-tailed) | .009 | .088 | .042 | .041 | .020 | .047 | .049 | .097 | .009 | .035 | |

Notes: MA-5 denotes that the value of the indicator is calculated using five periods of fundamental stock quantities (maximum price, minimum price, opening price, closing price, and stock trading volume) from present day to previous 4 day; and the values of other indicators (MTM-5, %K-5, %D-5, RSI-5, PSY-5, %R-5, VR-5 and AR-5) are produced in the same way.

** Denotes that correlation is significant at the 0.05 level using 2-tailed test.

*** Denotes that correlation is significant at the 0.01 level.

it is clear that the proposed model surpasses the other three models (Buy-and-Hold, RST and GAs) in each testing period, except 2005. These stock return evaluations demonstrate the outstanding performance of the proposed model.

5. Findings and conclusions

This paper has proposed a new hybrid model, based on four novel methods (CDPA, MEPA, RST and GA), to promote stock market forecasting performance. From the performance evaluation data in the above section, we can conclude that the main objective of this paper has been reached. Furthermore, by examining the performance data carefully, we also ascertain that four important findings for the proposed model emerge, as follows:

- (1) The proposed model produces a positive stock return, whether the market is bullish or bearish. Based on the stock return for the Buy-and-Hold approach (see Table 18), it has been shown that, from 2000 to 2005, there were three bear markets (2000, 2002, and 2003) and three bull markets (2001, 2004, and 2005). From the stock return evaluations shown in Table 18, it is clear that the proposed model produces a positive stock return for each dataset. This evidence points to the exceptional ability of the proposed model to mine correct price patterns in the stock market.
- (2) The proposed model performs outstandingly when the stock market is in a nearly complete bull market (upside trend) or bear market (downside trend).

Figs. 10 and 11 show that the price trends of the TAIEX were mostly downward in 2000 and upward in 2001. From Table 18, it can be observed that the stock return of the proposed model in 2000 is the best (2271.03) among the six datasets and much better than the other three models (−813.21 for Buy-and-Hold, 1590.68 for RST, and 1817.29 for GA). In 2001, the stock return of the proposed model is in second place (1683.51), better than the other three models (1612.16 for Buy-and-Hold, 923.9 for RST, and 1272.05 for GA). Despite the “Buy-and-Hold” approach

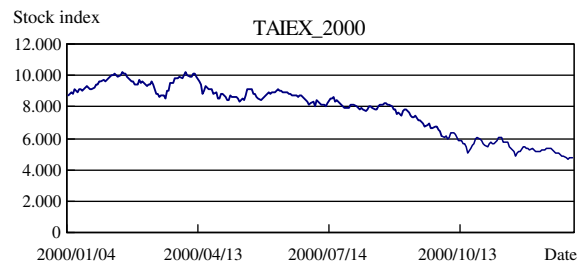


Fig. 10. The actual stock index for the TAIEX in 2000.

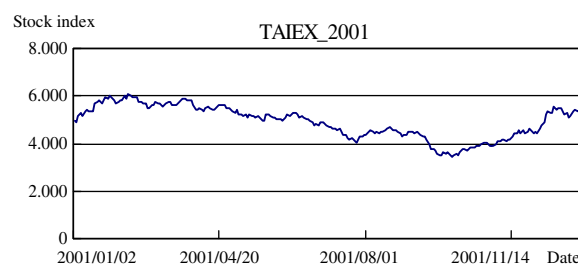


Fig. 11. The actual stock index for the TAIEX in 2001.

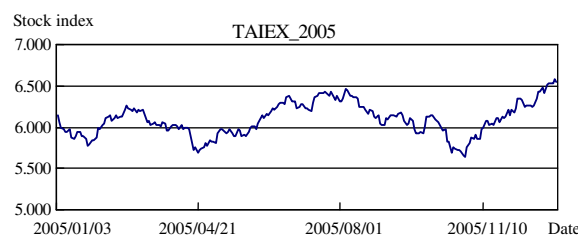


Fig. 12. The actual stock index for the TAIEX in 2005.

making a good stock return when the stock market was in a pure upward trend in 2001, the proposed model still surpasses this approach. These performance evaluations prove that the proposed model can extract tinier fluctuations in the stock price than the other three models when the stock market is in an almost pure bull market or bear market.

- (3) The three listed data-mining models (RST, GAs and the proposed model) perform their worst when there are many violent fluctuations occurring in the stock market.

Fig. 12 shows that the TAIEX fluctuated violently in 2005. The performance evaluations (see Tables 17 and 18) show that the worst case of the proposed model occurred for 2005. This can be explained by the fact that there were many conflicting rules extracted from the unstable market, which contained many violent fluctuations, and therefore, the performance value of the proposed model was dramatically reduced. Additionally, this phenomenon also occurred in RST and GA.

- (4) Integrating RST and GA together in forecasting processes can produce a positive effect, enhancing model performance. In the overall performance evaluation of accuracy (see Table 17), RST (0.553) proves marginally better than GAs (0.547), which means that RST can produce more effective rules than GAs. However, in the overall performance evaluation of stock return (see Table 18), GAs (722.56) are much better than RST (531.93), which tells that GAs can deal with the price variations in the stock market better than RST. From Tables 17 and 18, we can see that the proposed hybrid model (0.601, 950.62) performs much better than RST (0.553, 531.93) and GA (0.547, 722.56) in accuracy and stock return. The evidence demonstrates that the proposed hybrid model can acquire the advantages from the two data-mining methods (RST and GA) and therefore, produce superlative results in forecasting the stock market. Besides these findings, by implementing this experiment, two advantages were discovered for the proposed model: (1) the proposed model can produce more reasonable and understandable rules, because the “if-then” rules produced by RST can model the qualitative aspects of human knowledge; and (2) the proposed model can provide stock investors with objective suggestions (forecasts) to make investment decisions in the stock market, because the proposed model produces forecasting rules based on objective stock data rather than subjective human judgments.

For future research, two approaches to refine the proposed model, in order to improve forecasting performance, are suggested: (1) employ other data discretization methods in the preprocessing phase; and (2) use other artificial intelligence algorithms in the forecasting process.

References

- [1] P.J. Acklam, An algorithm for computing the inverse normal cumulative distribution function, (2004), Available from: <http://home.online.no/~pjacklam/notes/invnorm/>.
- [2] F. Allen, R. Karalainen, Using genetic algorithms to find technical trading rules, *Journal of Financial Economics* 51 (1999) 245–271.
- [3] B. Anna, Should normal distribution be normal? The Student's *T* alternative, *Computer Information Systems and Industrial Management Applications* (2007) 3–8.
- [4] M.E. Azo, *Neural Network Time Series Forecasting of Financial Markets*, Wiley, New York, 1994.
- [5] T. Blickle, L. Thiele, A mathematical analysis of tournament selection, in: L.J. Eshelman (Ed.), *Proceedings of the 6th International Conference on Genetic Algorithms*, 1995, pp. 506–511.
- [6] T. Bollerslev, Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics* 31 (1986) 307–327.
- [7] G. Box, G. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, 1976.
- [8] A.P. Chen, Y.H. Chang, Using extended classifier system to forecast S&P futures based on contrary sentiment indicators, *Evolutionary Computation* (2005) 2084–2090.
- [9] A.P. Chen, Y.C. Chen, W.C. Tseng, Applying extending classifier system to develop an option-operation suggestion model of intraday trading – An example of Taiwan index option, *Lecture Notes in AI* (2005) 27–33.
- [10] J.S. Chen, C.H. Cheng, Extracting classification rule of software diagnosis using modified MEPA, *Expert Systems with Applications* 34 (1) (2008) 411–418.
- [11] C.H. Cheng, J.R. Chang, C.A. Yeh, Entropy-based and trapezoid fuzzification-based fuzzy time series approaches for forecasting IT project cost, *Technological Forecasting & Social Change* 73 (2006) 524–542.
- [12] C.H. Cheng, J.S. Chen, Extracting classification rule of software diagnosis using modified MEPA, *Expert Systems with Applications* 34 (2008) 411–418.
- [13] S.C. Chi, W.L. Peng, P.T. Wu, M.W. Yu, The study on the relationship among technical indicators and the development of stock index prediction system, *Fuzzy Information Processing Society* (2003) 291–296.
- [14] R. Christensen, *Entropy minimax sourcebook*, *Fundamentals of Inductive Reasoning*, vol. 1, Entropy Ltd., Lincoln, MA, 1980.
- [15] N. Clarence, W. Tan, A hybrid financial trading system incorporating chaos theory, statistical and artificial intelligence/soft computing methods, in: *Queensland Finance Conference*, School of Information Technology, Bond University, 1999.
- [16] A.I. Dimitras, R. Slowinski, R. Susmaga, C. Zopounidis, Business failure prediction using rough sets, *European Journal of Operational Research* 114 (1999) 263–280.
- [17] R.F. Engle, Autoregressive conditional heteroscedasticity with estimator of the variance of United Kingdom inflation, *Econometrica* 50 (4) (1982) 987–1008.
- [18] L.J. Eshelman, R.A. Caruana, J.D. Schaffer, Biases in the crossover landscape, in: *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, 1989, pp. 10–19.
- [19] E. Faerber, *All About Stocks: From the Inside Out*, Probus, Chicago, 1995.
- [20] E.H. Tay*, Francis, S. Lixiang, Economic and financial prediction using rough sets model, *European Journal of Operational Research* 141 (2002) 641–659.
- [21] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley, Reading, MA, 1989.
- [22] D.E. Goldberg, K. Deb, A comparative analysis of selection schemes used in genetic algorithm, in: G. Rawlins (Ed.), *Foundation of Genetic algorithms*, Morgan Kaufmann, 1991, pp. 69–93.
- [23] S. Greco, B. Matarazzo, R. Slowinski, A new rough set approach to evaluation of bankruptcy risk, in: C. Zopounidis (Ed.), *Operational Tools in the Management of Financial Risks*, Kluwer Academic Publishers, Dordrecht, 1998, pp. 121–136.
- [24] J.H. Holland, *Adaptation in Nature and Artificial Systems*, University of Michigan Press, 1975.
- [25] K. Huang, H.K. Yu, The application of neural networks to forecast fuzzy time series, *Physica A* 363 (2006) 481–491.
- [26] S.H. Irwin, C.H. Park, What do we know about the profitability of technical analysis?, *Journal of Economic Surveys* 21 (4) (2007) 786–826.

- [27] K. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for prediction of stock index, *Expert System with Application* 19 (2000) 125–132.
- [28] M.J. Kim, S.H. Min, I. Han, An evolutionary approach to the combination of multiple classifiers to predict a stock price index, *Expert Systems with Applications* 31 (2006) 241–247.
- [29] T. Kimoto, K. Asakawa, M. Yoda, M. Takeoka, Stock market prediction system with modular neural network, in: *Proceedings of the International Joint Conference on Neural Networks*, San Diego, California, 1990, pp. 1–6.
- [30] J.B. Li, Y.T. Yu, A.P. Chen, Integration of group decisions and XCS in Intelligent financial decision support system – An example of Taiwan index, *Evolutionary Computation, IEEE Congress* (2006) 2389–2396.
- [31] H. Liu, F. Hussain, C. Tan, M. Dash, Discretization: An enabling technique, *Data Mining and Knowledge Discovery* 6 (4) (2002) 393–423.
- [32] C. Nikolopoulos, P. Fellrath, A hybrid expert system for investment advising, *Expert Systems* 11 (4) (1994) 245–250.
- [33] Z. Pawlak, Rough sets, *International Journal of Computational Information Science* (1982) 341–356.
- [34] Z. Pawlak, Rough sets and intelligent data analysis, *Information Sciences* 147 (2002) 1–12.
- [35] Z. Pawlak, A. Skowron, Rudiments of rough sets, *Information Sciences* 177 (2007) 3–27.
- [36] Z. Pawlak, A. Skowron, Rough sets: Some extensions, *Information Sciences* 177 (2007) 28–40.
- [37] Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, *Information Sciences* 177 (2007) 41–73.
- [38] B. Predki, R. Slowinski, J. Stefanowski, R. Susmaga, Sz. Wilk, ROSE – Software implementation of the rough set theory, in: L. Polkowski, A. Skowron (Eds.), *Rough sets and current trends in computing, Lecture Notes in Artificial Intelligence*, vol. 1424, Springer-Verlag, Berlin, 1998, pp. 605–608.
- [39] M.J. Pring, *Technical Analysis*, New York, 1991.
- [40] F.K. Reilly, *Investment Analysis and Portfolio Management*, Dryden Press, Chicago, 1989.
- [41] J.B. Richard, R.D. Julie, *Technical Market Indicators: Analysis & Performance*, John Wiley, New York, 1999.
- [42] T.H. Roh, Forecasting the volatility of stock price index, *Expert Systems with Applications* 33 (2007) 916–922.
- [43] T.J. Ross, *Fuzzy Logic with Engineering Applications*, McGraw-Hill, USA, 2000.
- [44] H. Shimodaira, A new genetic algorithm using large mutation rates and population-elitist selection (GALME), in: *Proceedings of 8th IEEE Conference on Tools with Artificial Intelligence*, 1996, pp. 25–32.
- [45] G. Syswerda, Uniform crossover in genetic algorithms, in: *Proceeding of the 3rd International Conference on Genetic Algorithms*, 1989, pp. 2–9.
- [46] R.R. Trippi, D.D. Sieno, Trading equity index futures with a neural network, *Journal of Portfolio Management* (1992) 27–33.
- [47] C. Tuncay, D. Stauffer, Power laws and Gaussians for stock market fluctuations, *Physica A* (2007) 325–330.
- [48] Y.F. Wang, Mining stock price using fuzzy rough set system, *Expert Systems with Applications* 24 (2003) 13–23.
- [49] L. William, P. Russell, M.R. James, Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support, *Decision Support Systems* 32 (2002) 361–377.
- [50] C. Xipin, J. Min, X. Bin, C. Jie, Application of elitist preserved genetic algorithms on fuzzy controller, in: *Proceedings of the 3th World Congress on Intelligent Control and Automation*, Hefei, PR China, 2000.
- [51] R. Yager, D. Filev, Template-based fuzzy system modeling, *Intelligent and Fuzzy System* 2 (1994) 39–54.
- [52] <<http://140.125.83.36/fuzzy/ga/>>.
- [53] <<http://alfa.mimuw.edu.pl/~rses/>>.
- [54] <<http://www.tse.com.tw>>.