Expert Systems with Applications 36 (2009) 3761–3773

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Evolving and clustering fuzzy decision tree for financial time series data forecasting

Robert K. Lai [a], Chin-Yuan Fan [b], Wei-Hsiu Huang [b], Pei-Chann Chang [c,*]

[a] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 32026, Taiwan, ROC
[b] Department of Industries Management, Yuan Ze University, Taoyuan 32026, Taiwan, ROC
[c] Department of Information Management,Yuan Ze University, Taoyuan 32026, Taiwan, ROC

ARTICLE INFO

Keywords:
Fuzzy theory
Decision tree
Step-wise regression
Stock price forecasting
Turning points
Genetic algorithm

ABSTRACT

Stock price predictions have always been a subject of interest for investors and professional analysts. Nevertheless, determining the best time to buy or sell a stock remains very difficult because there are many factors that may influence the stock prices. This paper establishes a novel financial time series-forecasting model by evolving and clustering fuzzy decision tree for stocks in Taiwan Stock Exchange Corporation (TSEC). This forecasting model integrates a data clustering technique, a fuzzy decision tree (FDT), and genetic algorithms (GA) to construct a decision-making system based on historical data and technical indexes. The set of historical data is divided into k sub-clusters by adopting K-means algorithm. GA is then applied to evolve the number of fuzzy terms for each input index in FDT so the forecasting accuracy of the model can be further improved. A different forecasting model will be generated for each sub-cluster. In other words, the number of fuzzy terms in each sub-cluster will be different. Hit rate is applied as a performance measure and the proposed GAFDT model has the best performance of 82% average hit rate when compared with other approaches on various stocks in TSEC.

1. Introduction

Mining stock market trend is a challenging task due to its high volatility and noisy environment. Many factors influence the performance of a stock market including political events, general economic conditions, and traders' expectations. Although stocks and futures traders have relied heavily upon various types of intelligent systems to make trading decisions, the performance have been a disappointment (Abu-Mostafa & Atiya, 1996).

Many attempts have been made to predict the financial markets, ranging from traditional time series approaches to artificial intelligence techniques, such as fuzzy systems and artificial neural network (ANN) methodologies (Abraham, Baikunth, & Mahanti, 2001). However, the main drawback with ANNs, and other black-box techniques, is the tremendous difficulty in interpreting the results. They do not provide an insight into the nature of the interactions between the technical indicators and the stock market fluctuations. Thus, there is a need to develop methodologies that provide an increased understanding of market processes (Chi, Chen, & Cheng, 1999; Zhang, 2007). Another issue to be dealt with is that the dimensionality of financial time series data also creates another challenge in ANN approaches.

The development of a timely and accurate trading decision-making tool is the key for stock traders to make profits. Since the stock price series is affected by a mixture of deterministic and random factors (Chi et al., 1999), new tools and techniques are needed in dealing with noise and nonlinearity in stock price prediction. Decision tree aimed at searching for rules hidden in very large amount of data. This is a new and efficient approach for time series analysis. In addition, decision tree techniques have already been shown to be interpretable, efficient, problem independent and able to treat large-scale applications. However, they are also recognized as highly unstable classifiers with respect to minor perturbations in the training data. Fuzzy logic provides the advantages in handling these variances due to the elasticity of fuzzy sets formalism. In this work, a decision tree tool ID3 (Iterative Dichotomizer 3) (Quinlan, 1986) is combined with the fuzzy theory and genetic algorithms to develop an evolving and clustering fuzzy decision tree for stock trading decision. The proposed model is able to predict the trends of stocks more precisely and to offer speculators a better information platform during the stock trading.

2. Literature review

Conventional research addressing the stock forecasting problem has generally relied on time series analysis techniques, i.e., mixed auto regression moving average (ARMA) as well as multiple regression models. However, the assumptions of these methods become ineffective since a number of missing factors, such as macro economical or political effects can influence stock tendencies greatly.

White (1988) was the first to use neural networks (NNs) for market forecasting. He used a feed-forward NN (FFNN) to study the IBM daily common stock returns. He found that his training results were

* Corresponding author.
E-mail address: iepchang@saturn.yzu.edu.tw (P.-C. Chang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.02.025

over-optimistic, because the result is over-fitting and of irrelevant features. In general, there are two different methodologies for stock price prediction in using ANN as a research tool (Zhang, Akkaladevi, Vachtsevanos, & Lin, 2002). The first method is to consider the stock price variations as a time series and predict the future price based on its past values. In this approach, artificial neural networks (ANNs) have been employed as the predictor, (see, e.g., Aiken & Bsat, 1994; Austin & Looney, 1997; Brownstone, 1996; Chang & Liu, 2006; Chen, Leung, & Daouk, 2003; Chi et al., 1999; Hiemstra, 1994; Hobbs & Bourbakis, 1995; Izumi & Ueda, 1999; Kimoto & Asakawa, 1990; Lee, 2001; Schierholt & Dagli, 1996; Sitte & Sitte, 1999; Yoon & Swales, 1991). These prediction models, however, have their limitations owing to the tremendous noise and high dimensionality of the stock price data. Therefore, the performances of the existing models are not satisfactory (Zadeh, 1965).

The second approach takes the technical indices and qualitative factors, like political effects into account in stock market forecasting and trend analysis. Yao and Poh (1995) use Technical Indicators (%K and %D) along with price information to predict future price values. They achieved good returns, and found their models performed better using daily data rather than weekly data. Hobbs and Bourbakis (1995) predict prices of stocks based on the fluctuations in the rest of the market for the same day. Although the investment is done in a frictionless environment, they show consistently high rates of return. Paying commissions on the large number of trades instigated would certainly take away much of the benefit from the trading strategy proposed. Austin and Looney 5 develop a neural network that predicts the proper time to move money into and out of the stock market. They used two valuation indicators, two monetary policy indicators, and four technical indicators to predict the four week forward excess return on the dividend adjusted S& P 500 stock index. The results significantly outperformed the buy-and-hold strategy. Backpropagation ANNs is applied to predict future elements in the price time series in KOSPI (Kim & Han, 2000).

Mingo López, Díaz, Palencia, Santos, and Jiménez (2002) use time delay connections in enhanced neural networks (that is, the addition of time-dependant information in each weight) to forecast IBEX-35 (Spanish Stock Index) index close prices 1 day-ahead. Stochastic neural networks is applied in forecasting the volatility of index returns for TUNINDEX (Tunisian Stock Index), and finds that the out-of sample neural network results are superior to traditional GARCH models (Slim, 2004). Nenortaite and Simutis (2004) present a trading approach based on one-step ahead profit estimates created by combining neural networks with particle swarm optimization algorithms. The method is profitable given small commission costs, but does not exceed the S& P500 returns when realistic commissions are introduced. Jaruszewicz and Mandziuk (2004) train ANNs using both technical analysis variables and intermarket data, to predict one-day changes in the NIKKEI index. They achieve good results using MACD, Williams, and two averages, along with related market data from the NASDAQ and DAX.

The fuzzy decision tree is similar to the standard decision tree methods (e.g. CART Janikow, 1998; Quinlan, 1986; Weber, 1985) based on a recursive binary partitioning algorithm. At each node during the construction process of a fuzzy decision tree, the most stable splitting region is selected and the boundary uncertainty is estimated based on an iterative resampling algorithm. The boundary uncertainty estimate is used within the region's fuzzy membership function to direct new samples to each resulting partition with a quantified confidence. The fuzzy membership function is used to recover those samples that lie within the uncertainty of the splitting regions. Many attempts (Jaruszewicz & Mandziuk, 2004; Larsen & Yager, 2000; Medasani, Kim, & Krishnapuram, 1998; Mugambi, Hunter, Oatley, & Kennedy, 2004; Sorensen, Miller, & Ooi, 2000; Yuan & Shaw, 1995) have been made in the past to introduce this new technology into stock prediction. Sorensen et al. (2000) use CART to partition assets into outperforming and underperforming assets. Portfolio composed by uniformly weighted outperforming assets.

It has been a new tendency that combining the soft computing (SC) technologies of NNs, fuzzy logic (FL) and genetic algorithms (GAs) may significantly improve an analysis (Abraham et al., 2001; Abraham, Philip, & Saratchandran, 2003; Baba, Inoue, & Asakawa, 2000; Braun & Chandlar, 1987; Chang & Wang, 2006; Chang, Wang, & Yang, 2004; Chang & Warren Liao, 2006; Corani & Guariso, 2005; Golan & Ziarko, 1995; Khokaharm & Sap, 2004; Kosko, 1992; Montana & Davis, 1989; Murata, Ishibuchi, & Gen, 1998; Murata & Ishibuchi, 1996; Su, Liu, & Tsay, 1999; Yu, Wang, & Lai, 2005). In general, NNs are used for learning and curve fitting, FL is used to deal with imprecision and uncertainty, and GAs are used for search and optimization. Zadeh (1965) pointed out, merging these technologies results in a tolerance for imprecision, uncertainty, and partial truth to achieve tractability, robustness, and low solution cost.

This research will follow Zadeh's suggestion by combining several soft computing techniques such as fuzzy decision tree, *K*-means for data clustering, and genetic algorithm to develop a forecasting model for stock trading decision. In addition to fuzzy decision tree, the *k*-means clustering algorithm is applied to cluster the data before the fuzzy decision rules are generated. A set of fuzzy decision rules is generated for each cluster, which enables us to determine the fuzzy terms of each variable. Finally, a GA is applied as an evolving tool to further fine-tune the forecasted result from the FDT model.

3. A forecasting model by fuzzy decision tree

Decision tree induction is free from parametric assumptions and it generates a reasonable tree by progressively selecting attributes to branch the tree. A decision tree is a flow-chart-like structure where each node represents a test on an attribute (such as trading volumes), each branch represents an outcome of the test (such as trading volumes = high) and leaf nodes represent a classification of an instance (such as buy). By combining technical indices, stock price variation, and transaction volumes on stock trading, this research will apply a fuzzy decision tree to develop a forecasting model for generating decision rules in stock trading decisions.

A novel financial time series-forecasting model is developed by clustering and evolving fuzzy decision tree for stocks in TSEC. This forecasting model integrates a data clustering technique, a fuzzy decision tree (FDT), and genetic algorithms (GA) to construct a decision-making system based on historical data and technical indexes. The set of historical data is divided into *k* sub-clusters by adopting a *K*-means algorithm. GA is then applied to evolve the number of fuzzy terms for each input index in FDT. The forecasting accuracy of the model can also be further improved. A different gorecasting model will be generated for each sub-cluster. In other words, the number of fuzzy terms in each sub-cluster will be different.

The detailed model of GAFDT is shown in Fig. 1, and it can be divided into four major steps for stock screening; *K*-means clustering; establishing fuzzy decision tree model; and finally output the forecasting results. The details of each major step are further explained in the following sections.

3.1. The selection of stocks

The source of the data for analysis is selected from stock trading data from 2005/8/10 to 2005/9/30 on TSEC (Taiwan Stock Exchange Corporation). The following analytical indices were applied to select stocks worthy of investment. These indices are listed in Table 1.

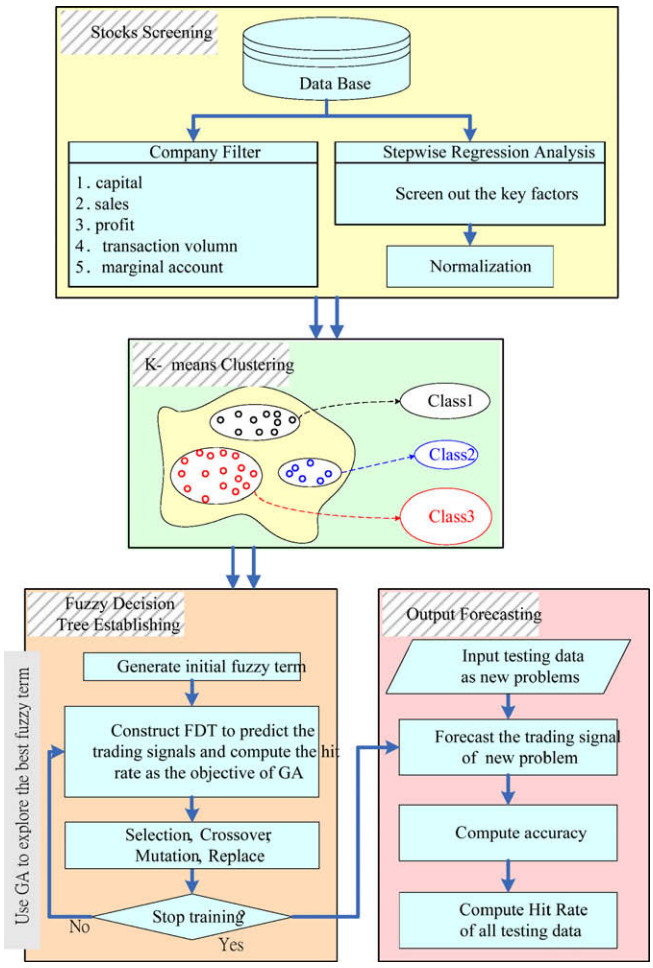

Fig. 1. Detailed model for GAFDT.

3.2. Technical indices

The technical indices (Edwards & Magee, 2001) are calculated by using a specific equation with information on trading prices, volumes and time to reflect the trend of stock price. This is especially useful for judging the trading points and phenomena of overbuying and overselling. Technical indices such as KD, RSI, MACD, MA, BIAS, etc. along with price and volume data will be analyzed using FDT to decide a next day trading decision either sell, buy or hold. These technical indices are listed in Table 2.

3.3. Data pre-processing

3.3.1. Input selection using step-wise regression analysis (SRA)

A set of important factors which will affect the forecasting results are shown in Table 2. These important input factors are further selected through step-wise regression analysis (SRA) method in this research. Step-wise regression is used to sort unaffecting variables out and leave more influential ones in the model. During the process, SRA operates by adding variables onward or removing variables backward in finding the fittest combination of factors for stock prices prediction. The criterion for adding or removing is decided by *F*-test statistical value and decreasing the sum of squared error. After the entrance of first variable to the model, the variable number is increased systematically; once it is removed from this model, it will never enter the model again. Before selecting variables, the critical point, level of significance, and the values of Fe (*F*-to-enter) Fr (*F*-to-remove) have to be determined first. Then the partial *F*-value of each step has to be calculated and compared with Fe and Fr; If $F > Fe$, it is considered as variables to be added to the model; otherwise, if $F < Fr$, variables should be removed from this model. SPSS is used for variable selection and setting up the regression forecasting model.

Table 1
Basic indices for investing in stock

Indices	Descriptions
Capital stock	This index is used to estimate the scale of a company. The higher the amount of capital stock the higher circulating ability. However, The smaller the capital stock the stabler is the stock
Revenue situation	Revenue situation represents operation achievements of a company. The better revenue situation shows the company having ability of earning more profits
Earnings per share (EPS)	EPS = Profit/no. of stock
Turnover number	Turnover no. is an index to be observed the level which investors concern
Net worth and market value ratio (NWMV)	NWMVR = Interest/Market price
Price-earnings ratio, PER	PER = Stock price/profit after taxes the lower ratio represents investors can buy stock with lower price

Table 2
Technical indices used as input variables

Technical index	Descriptions
Six days moving average (MA)	Moving averages are used to emphasize the direction of a trend and smooth out price and volume fluctuations that can confuse interpretation
Six days bias (BIAS)	The difference between the closing value and moving average line, which uses the stock price nature of returning back to average price
Six days relative strength index (RSI)	RSI compares the magnitude of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset
Nine days stochastic line (K,D)	The stochastic line K and line D are used to determine the signals of over-purchasing, over-selling, or deviation
Moving average convergence and divergence (MACD)	MACD shows the difference between a fast and slow exponential moving average (EMA) of closing prices. Fast means a short-period average, and slow means a long period one
13 days psychological Line (PSY)	PSY is the ratio of the number of rising periods over the total number of periods. It reflects the buying power in relation to the selling power
Volume	Volume is a basic yet very important element of market timing strategy. Volume provides clues as to the intensity of a given price move

3.3.2. Normalize the selected variables from SRA

The selected variables from SRA are normalized as follows:

$$X_i' = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

3.3.3. K-means clustering method

After we select these important input factors through step-wise regression analysis, we apply *K*-means clustering to divide the historic data into sub-cluster to generate more homogeneous data. Hopefully, this clustering will produce more accurate forecasting model generated from each sub-cluster and the overall forecasting accuracy can be further improved.

The *K*-means clustering algorithm is employed for data clustering. *K*-means is a non-hierarchical clustering technique that partitions data into *K*-clusters. During the clustering, the data points are randomly assigned to the clusters to minimize the following squared error (SE):

$$\mathrm{SE} = \sum_{i=1}^{K} \sum_{p \in C_i} |p - m_i|^2 \tag{2}$$

where p are data points in the cluster C_i; m_i is the center of cluster C_i; K is the number of clusters. Once the training data are clustered, we can calculate the parameters of the membership functions for each cluster as follows: S_i is the number of data points in cluster i. In addition, the output training data are also normalized using the mean and standard deviation of those data in each cluster

$$\begin{aligned} a_{ij} &= \frac{1}{s_i} \sum_{i=1}^{s_i} x_j, \\ \sigma_{ij} &= \frac{1}{s_i - 1} (x_j - a_{ij})^2 \end{aligned} \tag{3}$$

3.4. GAFDT forecasting model

This research combines genetic algorithms and fuzzy decision trees to propose a GAFDT forecasting model. The framework of this model is as Fig. 2.

3.4.1. Data fuzzification

This research utilize fuzzy resolution concept in fuzzy set theory to transform data attribute from continuous to discreet. It also applies decision-tree classification method to build a stock forecasting model. Kosko (1992) used Fuzzy Entropy method to revise fuzzy theory data, and Janikow (1998) used fuzzy set's probability to replace clear set probability by calculating fuzzy set data Entropy.

In Fuzzy set theory, membership function is one of the basic concepts, through this concept one will be able to process quantitative fuzzy set data and dispose fuzzy message. How to find an appropriate membership function to approach quantitative fuzzy set data and dispose of fuzzy message becomes very important in fuzzy set theory. Researchers always consider different problems with different membership function; the most used membership function includes Triangles membership functions, trapezoid membership functions, and Gauss membership functions. This research will adopt Triangles membership functions for our primary membership functions. The equation of triangles membership functions include three parameters, i.e., a, b and c. The triangle membership function is determined by the following equation:

$$\mu(x) = \begin{cases} 0, & x \leqslant a \\ \frac{x-a}{b-a}, & a \leqslant x \leqslant b \\ \frac{c-x}{c-b}, & b \leqslant x \leqslant c \\ 0, & c \leqslant x \end{cases} \tag{4}$$

The tri-angle membership function is shown in Fig. 3.

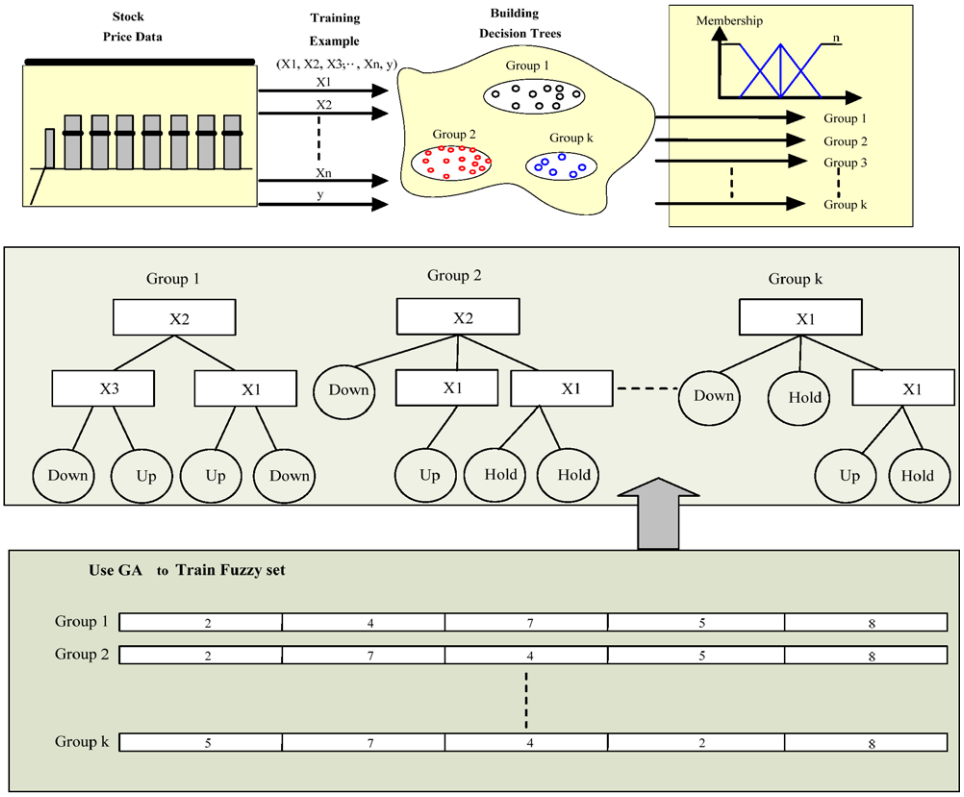

Fig. 2. GA FDT process chart.

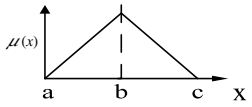

Fig. 3. Tri-angle membership function.

3.4.2. ID3 decision tree

After transforming numerical data into fuzzy set data, next step will be to classify the data by ID3 decision tree for future trading prediction. ID3 decision tree is based on Information Gain, which means chaos of data. The higher the Gain value, the more chaos the data is. As a result, the biggest information gain is selected as an attribute value. Then, the decision tree use this attribute value splitting data for more training sets, those training sets repeatedly attempt to search for the biggest information gain until no more splitting.

First suppose training data set S have m categories, it means C_i, $i = 1,2,3,\ldots,m$, every category number are presented in $freg(C_i,S)$, $|S|$ means all values in data set S. The probability that every category data appears can be calculated as follows:

$$\frac{\text{freg}(C_i,S)}{|S|} \tag{5}$$

Then according to information theory, the function, which measures the amount of information of occurrence of receiving the symbol S, is defined as follows:

$$-\log_2\left(\frac{\text{freg}(C_i,S)}{|S|}\right) \tag{6}$$

Then we multiply the probability that every data appears with the information gain that every categorized data set can get except training data set S. The entropy of data set S having symbols s_i and probability p_i is calculated as follows:

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{7}$$

Through this equation, probability p_i can be described by the following equation:

$$p_i = \frac{\text{freg}(C_i,S)}{|S|} \tag{8}$$

Assume set A can be split into sub-set $s_1,s_2,\ldots,s_m$, then we understand that splitting information equal to sub-set information multiplied by every sub-set proportion amount. The Entropy of set A is calculated follows:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+\cdots+s_{mj}}{s} * I(s_{1j},\ldots,s_{mj}) \tag{9}$$

The detailed procedures of ID3 decision tree is described in Fig. 4.

```
. Function ID3 (I, O, T)
{
/*
I is the set of input attributes
O is the output attribute
T is a set of training data
*/
   If (T is empty)
   {
      Return a single node with the value "Failure";
   }
   If (all records in T have the same value for O)
   {
      Return a single node with that value;
   }
  If (I is empty)
   {
      Return a single node with the value of the most frequent value of O in T;
   }
   Compute the information gain for each attribute in I relative to T;
   Let X be the attribute with largest Gain(X, T) of the attributes in I;
   Let {X_jl j=1, 2, …, m} be the values of X;
   Let {T_jl j=1, 2, …, m} be the subsets of T when T is partitioned according the value of X;
      Return a tree with the root node labelled X and arcs labelled X_1, X_2, …, X_m, where the
      arcs go to the trees ID3(I-{X}, O, T_1), ID3(I-{X}, O, T_2),…, ID3(I-{X}, O, T_m);
                                                    }
```

Fig. 4. The virtual code of the ID3.

3.4.3. Evolving fuzzy decision tree by genetic algorithm

Genetic Algorithm is applied to improve the performance of A FDT (fuzzy decision tree) in financial data forecasting. GA (genetic algorithms) to find the best fuzzy term for each input data. The fitness function of each input data set will then be calculated. In this research, the fitness function is the forecasting accuracy of the proposed model. Then, a set of chromosomes will be selected for cross-over and mutation. The whole process will repeat iteratively until a stopping criteria is satisfied. The process of evolving fuzzy decision tree is described in Fig. 5.

Detailed procedures of evolving fuzzy decision tree are listed as follows:

Step 1. Coding:
Binary code is adopted here, and we assume that there are four technical indices to be considered. Therefore, we randomly generate 12 (i.e., 4 * 3 = 12) genes. The binary code for a feasible gene expression is shown in Fig. 6.

Step 2. De-coding (generate initial solutions):
Since the original chromosome is expressed in binary code, this step will transform gene expression from binary code to decimal code. We can then derive the range of each input technical index and transform this range into a set of fuzzy intervals. The transforming process from binary code to decimal code and fuzzy intervals is illustrated in Fig. 7.

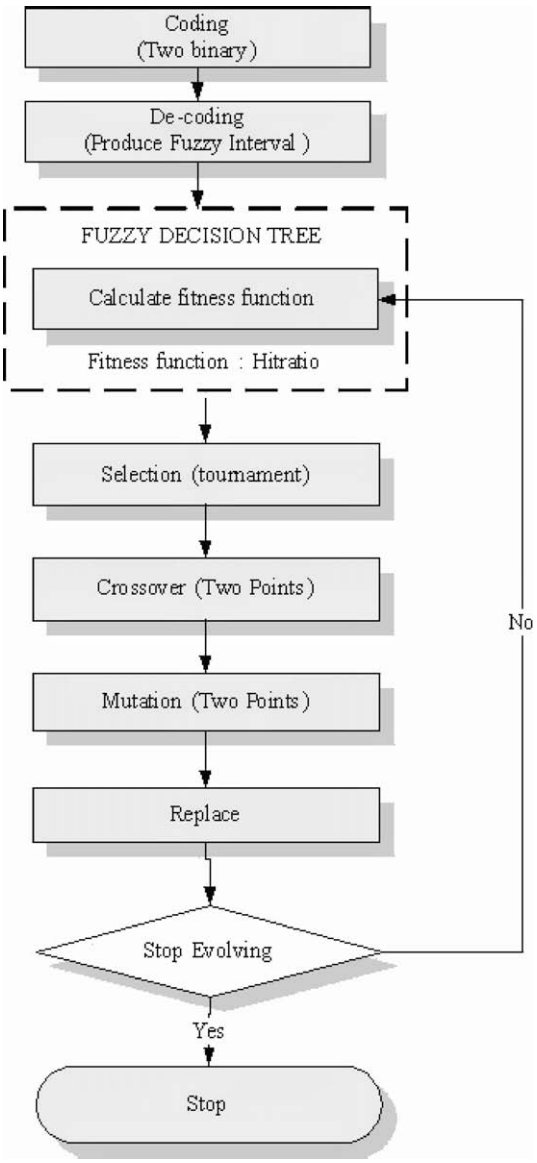

Fig. 5. Process of evolving fuzzy decision tree by genetic algorithm.

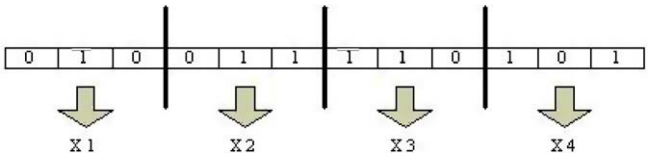

Fig. 6. Binary code for a feasible gene expression.

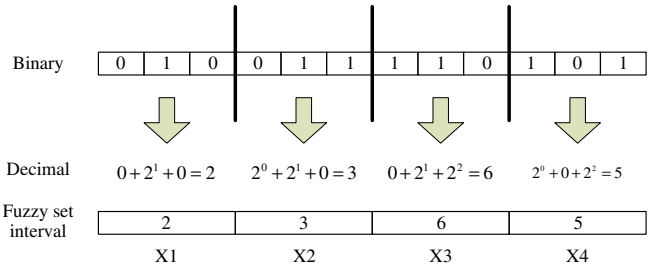

Fig. 7. Generate an initial solution.

Step 3. Calculate the objective value:
The hit ratio of each stock within the studied period will be the objective function of GA and it is defined as follows:

$$\text{Hit ratio} = \frac{\sum_{i=1}^{m} x_i}{n} * 100\% \tag{10}$$

where, m means the number of data clustered; x_i means the number of stocks forecasted correctly, i.e., in up, down or hold decisions, in i cluster, n means total number of data. It represents the accuracy of the prediction model in terms of hit ratio. That is the number of times in correct forecasting for the forecasting model in stock trading decision in terms buy, sell or hold.

Step 4. Representation and selection:
The tournament method is used in this study.

Step 5. Cross-over:
Two points cross-over method is applied and it is shown in Fig. 8.

Step 6. Mutation:
Two points mutation method is adopted in the research as shown in Fig. 9.

Step 7. Replace:
If the fitness function of the current chromosome is better than the best one generate from the previous generation, replace it with the current one.

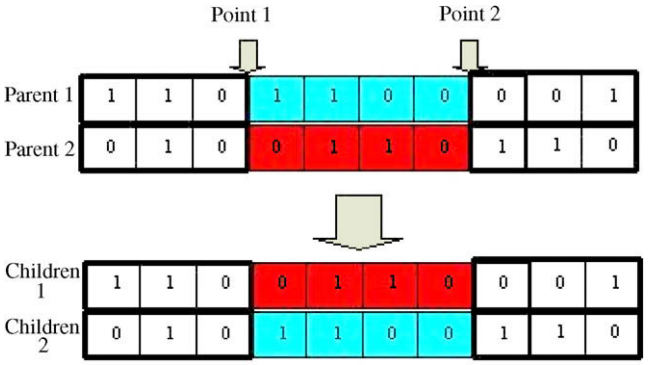

Fig. 8. Two points cross-over.

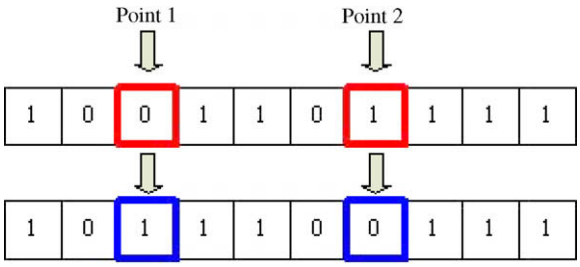

Fig. 9. Two points mutation.

Step 8. Terminate:
If the stopping criteria is satisfied, terminate the GA process and output the best result.

3.5. The decision of an output value

The decision (output) of a stock price movement (trend) is show in Eq. (11). It is in an uptrend (price-hike) when y is greater than +0.5%. On the other hand, it is in a downtrend (price-fall) when y is less than −0.5%. Finally, it is in a steady state (hold) if y is between +0.5% and −0.5%.

$$y = \frac{x_{t-1} - x_t}{x_t} \tag{11}$$

where x_{t-1}: closing price of individual stock in the $(t-1)$th period; x_t: closing price of individual stock in the tth period.

4. Experimental result

The data set applied for test in this research is selected according to the criteria listed in Table 1. There are three particular stocks are selected for test and they are the Epistar Corp. (EPISTAR), Silicon Integrated System Corp. (SiS) and UMC Corp. (UMC) which are in uptrend, downtrend and steady state respectively. The historic data of these companies are derived for observation from 2005/8/10 to 2005/9/30. The major reason for stock selection is that we want to show different patterns of stock price movements to provide different profit making strategies in stock trading by this proposed model. Another reason is to show that the proposed model can have a robust performance even under

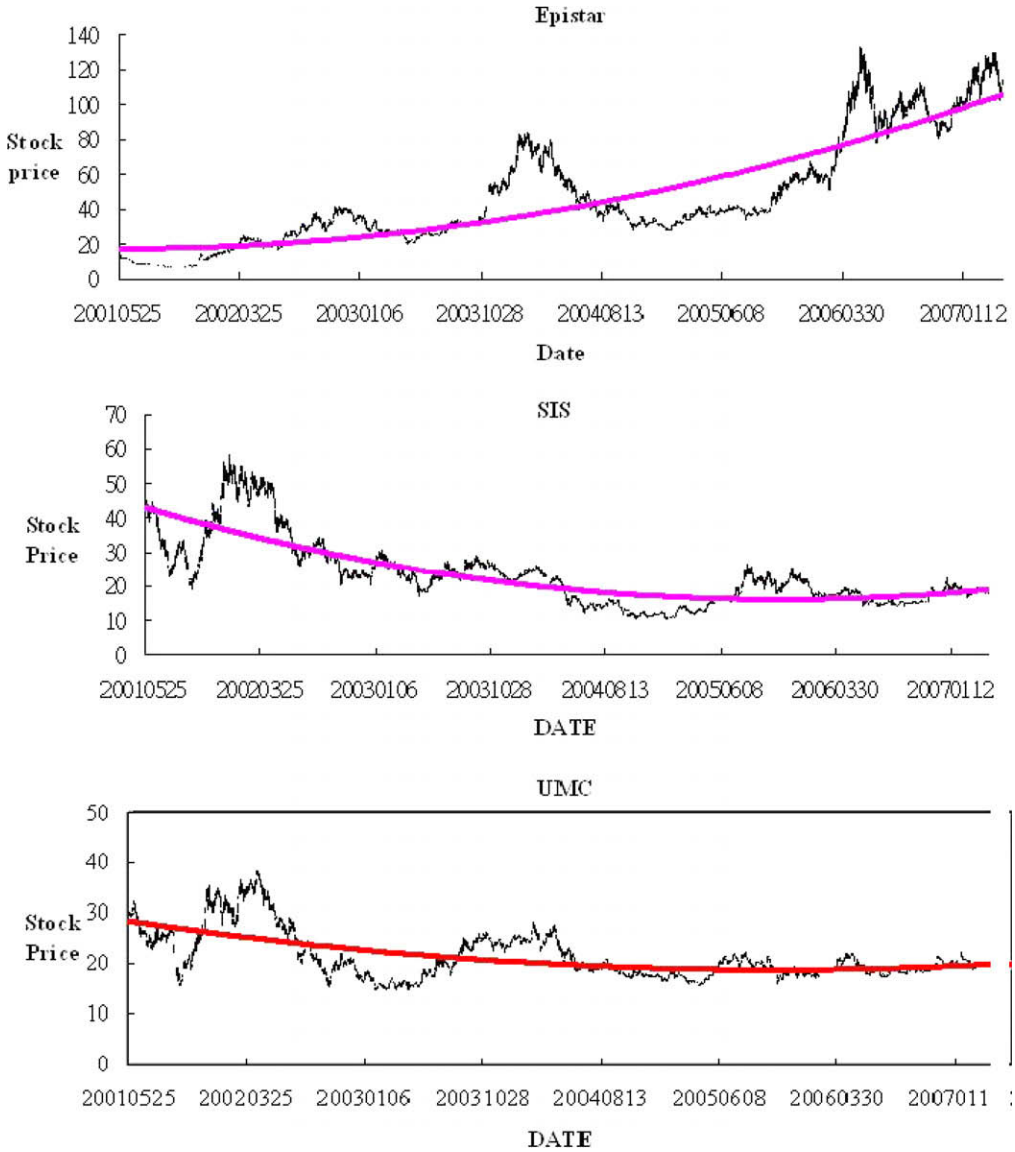

Fig. 10. Three different stocks in uptrend, downtrend and steady state.

Table 3
Step-wise regression analysis choose important input factors

Input factors		
Technical indices	5MA, 5BIAS, 6MA, 10MA, 9K, 9D, 6RSI, 9MACD, 12W%R, 20MA, 12RSI, K_D	
Difference of technical indices	5MA difference, 6MA difference, 10MA difference, 5BIAS difference, 10BIAS difference, 6RSI difference, 9K difference, 9D difference, 12RSI difference, 12W%R difference	
Stock names	Input factors selected	
EPISTAR	Technical indices	12RSI
	Technical indices difference	10BIAS difference, 6RSI difference, 12RSI difference
SiS	Technical indices	12RSI
	Technical indices difference	10BIAS difference, 6RSI difference, 12RSI difference
UMC	Technical indices	12W%R
	Technical indices difference	10BIAS difference, 12RSI difference, 12W%R difference

different types of stock tendency, i.e., uptrend, downtrend and steady state. In the following, the stocks screening process will be introduced first. Then different input factors will be selected according to step-wise regression analysis for different stocks. The details of historical data of each stock are shown in Fig. 10, and the overall comparisons of all instances will be presented in the last section.

4.1. Best parameters setting (SRA)

According to the SRA method described in Section 3.3, significant input factors are chosen to find the best parameter setting. A SPSS software package is applied for the SRA procedure, and significant input factors selected are shown in Table 3.

4.2. K-means clustering and cross-over test

The set of historical data with a total of 1430 records are divided into five sub-set and they are A, B, C, D, and E as shown in Table 4. This is the so called 5-fold cross-over test as shown in Table 5. In addition, *K*-means is applied to cluster the set of sales data from the TSEC stocks and the clustered results are illustrated in Tables 6–8. Although the number of clusters is a key factor to be decided, however, up to now there is not any theory to be applied to select the optimal number of clusters for any data. In the preliminary test, different numbers of clusters are applied to test the effect of data clustering. Therefore, in this research, two to eight clusters are applied for clustering these historical data and the best clustering result of each stock will be selected. Finally, we will compare this result with that of other methods in next section.

Through a series of experimental tests, the best clustering result are derived for these three stocks and for EPISTAR in Table 5, for SiS in Table 6 and for UMC in Table 7. Next, we will use this

Table 4
Cross-over test data distribution

	Data Nunber	I_1	I_2	I_3	I_4	Y	
	1	-0.022633682	-8.5974563	9.41073733	97.8226197	down	
	…	…	…	…	…	…	
	…	…	…	…	…	…	**A**
	286	-0.020190691	0.15110806	28.778217	95.1208427	down	
	287	0.038840974	-0.2257927	-47.2409368	47.8799059	up	
	…	…	…	…	…	…	
	…	…	…	…	…	…	**B**
	572						
	573	…	…	…	…	…	
	…	…	…	…	…	…	
	…	…	…	…	…	…	**C**
	858						
	859	…	…	…	…	…	
	…	…	…	…	…	…	
	…	…	…	…	…	…	**D**
	1144	…	…	…	…	…	
	1145	…	…	…	…	…	
	…	…	…	…	…	…	
	…	…	…	…	…	…	**E**
	1430	0.00195386	-0.0438823	-6.85279188	48.2945944	hold	

Table 5
Cross-over test-situation

Data	Training-data group	Test data group
First cross-over test	BCDE	A
Second cross-over test	ACDE	B
Third cross-over test	ABDE	C
Fourth cross-over test	ABCE	D
Fifth cross-over test	ABCD	E

clustering result to consider the best parameters setup for GA application.

4.3. Best parameters setting (genetic algorithms)

When the best number of clusters of those three stocks is derived, next step will be to consider the factor design for GA evolving applications. Genetic algorithms are applied to evolve the fuzzy terms of each factor in this research. Four important factors are selected in this experimental design. They are population size, number of generation, cross-over rate and mutation rate. After the design of experimental tests, the best factor design for GA in these three stocks is in Table 9. The interaction plots of factor design in these three stocks can be referred to in Appendix 1.

Table 9
The best factor design for GA in these three stocks

Factors	EPISTAR Levels	SiS Levels	UMC Levels
Chromosome	20	20	20
Generation	100	10	100
Cross-over	0.9	0.9	0.9
Mutation rate	0.1	0.1	0.3

4.4. Method comparisons

After setting up the parameters of the experiments, the outputs of GAFDT is compared with those from traditional FDT and these experimental results are shown in Tables 10–12. The experimental results show that GAFDT perform much batter than FDT.

Through a series of experimental tests, we see that the hit ratios of the GAFDT method are batter than those of traditional FDT model in uptrend and downtrend stocks. However, in a steady-trend situation, those two models show almost equal hit ratios. The reason is because the range of variation for the stock price movements is too steady to be distinguished.

Table 6
Forecasting accuracy in different number of clusters in EPISTAR

Cluster	First cross-over	Second cross-over	Third cross-over	Fourth cross-over	Fifth cross-over	Avg
EPISTAR						
No cluster	0.8489	0.7853	0.8689	0.8235	0.8185	0.8290
$c = 2$	0.8626	0.8255	0.8776	0.8297	0.8135	0.8418
$c = 3$	0.8759	0.8909	0.8972	0.8350	0.7824	0.8563
$c = 4$	0.8659	0.8839	0.9088	0.8303	0.8259	0.8630
$c = 5$	0.8639	0.8204	0.8636	0.8014	0.8037	0.8306
$c = 6$	0.9092	0.8871	0.8852	0.8324	0.8369	0.8702
$c = 7$	0.8976	0.8721	0.8737	0.8379	0.8186	0.8600
$c = 8$	0.8934	0.8548	0.8870	0.8310	0.8462	0.8624

Table 7
SiS Corp. forecasting accuracy in different number of clusters

Cluster	First cross-over	Second cross-over	Third cross-over	Fourth cross-over	Fifth cross-over	Avg
SiS						
No cluster	0.8420	0.8001	0.76933	0.8357	0.7804	0.8054
$c = 2$	0.8615	0.8454	0.8411	0.8517	0.7898	0.8379
$c = 3$	0.9129	0.8378	0.8528	0.8618	0.8338	0.8598
$c = 4$	0.8828	0.8326	0.8343	0.8637	0.8206	0.8468
$c = 5$	0.8515	0.8299	0.8364	0.84290	0.8091	0.8339
$c = 6$	0.8546	0.8618	0.8491	0.8750	0.8513	0.8584
$c = 7$	0.8988	0.8755	0.8659	0.88745	0.8536	0.8763
$c = 8$	0.8856	0.8630	0.8638	0.8832	0.8620	0.8715

Table 8
UMC Corp. forecasting accuracy in different number of clusters

Cluster	First cross-over	Second cross-over	Third cross-over	Fourth cross-over	Fifth cross-over	Avg
UMC						
No cluster	0.8318	0.8145	0.7762	0.8143	0.7673	0.8009
$c = 2$	0.8668	0.8304	0.7800	0.8330	0.7885	0.8197
$c = 3$	0.8808	0.8797	0.8578	0.8862	0.8419	0.8693
$c = 4$	0.8682	0.8519	0.8290	0.8962	0.8674	0.8626
$c = 5$	0.9010	0.8703	0.8559	0.9035	0.8802	0.8822
$c = 6$	0.9214	0.8756	0.8493	0.8977	0.8727	0.8833
$c = 7$	0.8969	0.8927	0.8617	0.8920	0.8690	0.8825
$c = 8$	0.9220	0.8857	0.8590	0.9241	0.8905	0.8963

Table 10
The hit ratio of EPISTAR (FDT vs. GA FDT)

EPISTAR	FDT			GAFDT		
	Fuzzy_term numbers			AVG hit ratio	Best hit ratio	Best fuzzy_term
	2	3	4			
First cross-over test	0.7587	0.7517	**0.7727**	0.8489	0.8951	5,8,2,9
Second cross-over test	0.6503	0.6608	**0.6958**	0.7853	0.8496	7,9,3,8
Third cross-over test	0.6818	**0.7202**	0.7167	0.8688	0.8811	7,9,9,8
Fourth cross-over test	0.7027	0.7097	**0.7237**	0.8234	0.8391	5,8,3,9
Fifth cross-over test	0.7097	0.6748	**0.7202**	0.8185	0.8286	7,9,2,8

Table 11
The hit ratio of SiS Corps. (FDT vs. GAFDT)

SiS	FDT			GAFDT		
	Fuzzy_term numbers			AVG hit ratio	Best hit ratio	Best fuzzy_term
	2	3	4			
First cross-over test	0.7972	0.769	**0.8111**	0.8419	0.8846	9,9,7,9
Second cross-over test	0.7202	0.6363	**0.7237**	0.8300	0.8951	9,9,4,9
Third cross-over test	**0.7062**	0.6538	0.6818	0.8342	0.8706	4,9,2,9
Fourth cross-over test	**0.7867**	0.6713	0.7832	0.8556	0.9055	9,9,8,9
Fifth cross-over test	**0.7342**	0.6293	0.6853	0.7804	0.8671	9,9,9,9

Table 12
The hit ratio of UMC (FDT vs. GAFDT)

UMC	FDT			GAFDT		
	Fuzzy_term numbers			AVG hit ratio	Best hit ratio	Best fuzzy_term
	2	3	4			
First cross-over test	0.7867	0.6643	**0.8026**	0.8318	0.8741	8,8,8,8
Second cross-over test	0.7867	0.5594	**0.8146**	0.8145	0.8391	7,8,5,8
Third cross-over test	0.7657	0.5209	**0.7867**	0.7762	0.8041	8,8,6,8
Fourth cross-over test	**0.8091**	0.5489	0.7811	0.8143	0.8951	8,8,8,8
Fifth cross-over test	0.7797	0.5104	**0.8006**	0.7673	0.8286	8,8,5,8

4.5. Discussions

As shown in Tables 10–12, GAFDT is almost performing much better than the FDT method. The evolutionary procedure can further improve the accuracy of the FDT approach. One interesting phenomenon shows that the number of fuzzy terms for each input variables are large when compared with traditional FDT approach. A set of tests is setup to explain this phenomenon. We find out that the number of data and variation of the stock price are two possible reasons in causing a large number of fuzzy terms which can effectively discriminate possible trading decisions. The set of experimental tests for different number of data records are shown

Table 13
Data amounts and fuzzy terms relation

Trail	Data amounts		
	500	100	50
1	9, 4, 8, 9	7, 8, 9, 8	5, 7, 6, 6
2	9, 9, 6, 9	7, 8, 4, 9	6, 7, 2, 4
3	9, 6, 2, 9	7, 8, 9, 8	6, 5, 4, 3
4	9, 6, 6, 9	2, 8, 6, 9	5, 7, 3, 6
5	8, 6, 8, 8	8, 8, 4, 8	4, 7, 7, 8
6	9, 4, 2, 9	2, 8, 3, 9	4, 7, 8, 6
7	9, 6, 4, 9	7, 8, 2, 9	5, 7, 8, 6
8	9, 4, 9, 9	5, 8, 7, 8	2, 7, 6, 6
9	9, 6, 3, 9	8, 8, 8, 8	4, 7, 7, 7
10	9, 8, 5, 9	7, 8, 4, 9	2, 7, 7, 6
7	9, 6, 4, 9	7, 8, 2, 9	5, 7, 8, 6
8	9, 4, 9, 9	5, 8, 7, 8	2, 7, 6, 6
9	9, 6, 3, 9	8, 8, 8, 8	4, 7, 7, 7
10	9, 8, 5, 9	7, 8, 4, 9	2, 7, 7, 6

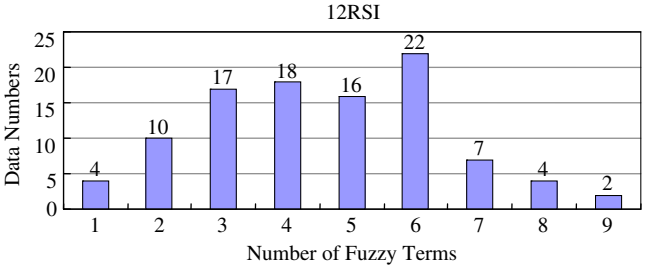

Fig. 11. 9 Fuzzy terms for 12RSI_delta.

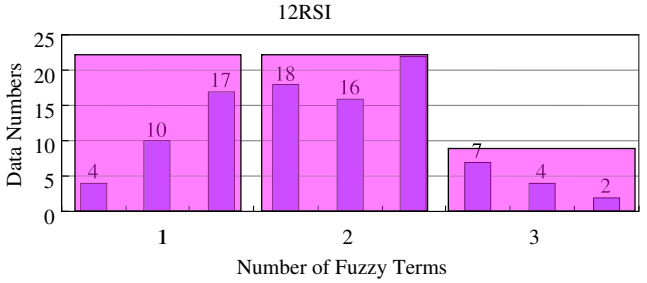

Fig. 12. 3 Fuzzy terms for 12RSI_delta.

in Table 13. The higher the number of data, the higher the number of fuzzy terms is.

Another important reason is the data distribution may cause a large number of fuzzy terms. Fore example, 12RSI_delta is in wide distribution and it is divided into 9 fuzzy terms as shown in Fig. 11. However, if it is divided into 3 fuzzy terms as shown in Fig. 12, a lot of insight information (fuzzy rules) is missing and the trading decision by FDT may not be as accurate as by GAFDT.

5. Conclusions

A considerable amount of researches has been conducted to study the behavior of stock price movement. However, the investor is more interested in making profit by providing simple trading decision such as Buy/Hold/Sell from the system rather than predicting the stock price itself. Therefore, we take a different approach by applying an evolving fuzzy decision tree to decompose the historical data into different segments. Then, a step-wise regression (SRA) method is chosen to select most important factors. Next, we consider applying the *K*-means clustering method to combine homogeneous data in a same group, and use fuzzy decision tree to transform numerical data into abstract syntax data. Finally, a GA is applied to evolve the fuzzy terms of each factor to derive the best forecasting results. Through a series of system training, hit ratio (buy or sell) of the historical stock data can then be detected and help investors to make better decision in trading stocks.

In the future, the proposed system can be further investigated by incorporating other soft computing techniques or better data mining forecasting model other than ID3 decision tree systems. They are listed as follows:

1. *Effective data clustering methods:* Data pre-processing is one of the feature that can be applied in time series data processing. Effective clustering of time series data can further improve the forecasting accuracy of the proposed system. Nevertheless,

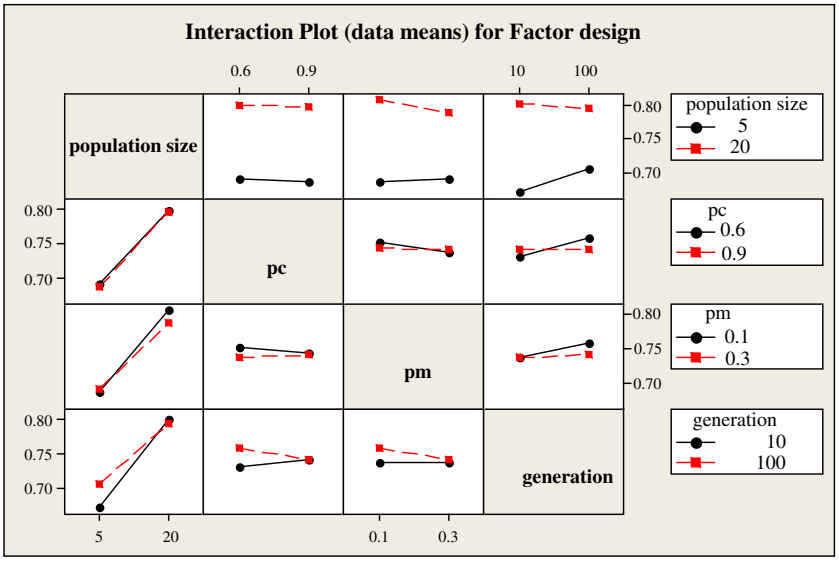

Fig. 13. Interaction plot for factor design in EPISTAR.

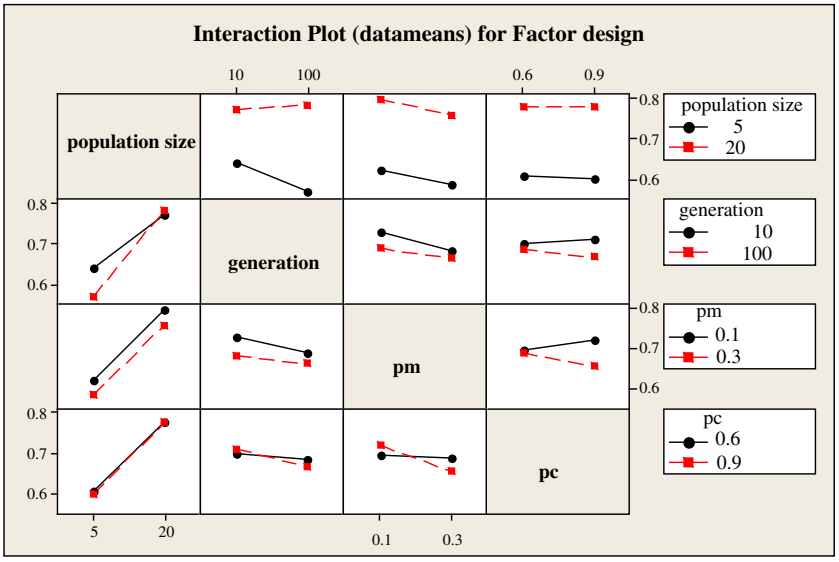

Fig. 14. Interaction plot for factor design in SiS.

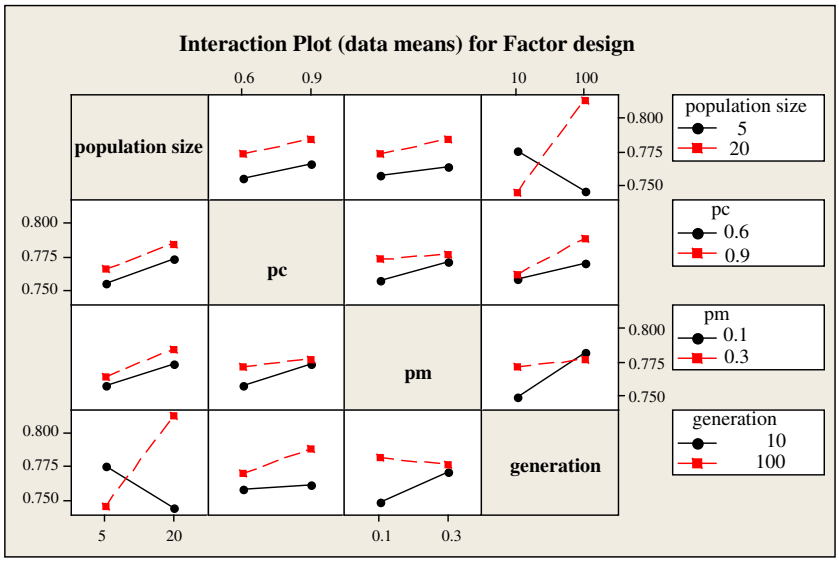

Fig. 15. Interaction plot for factor design in UMC.

how to cluster these data and according to what input factors seem to be interesting issues for future discussion.

2. *A different data mining model:* There are numerous forecasting models other than ID3 decision tree. It is important to study the behavior of these models when applied in prediction of the stock trading point. Different input factors and different forecasting models such as CART, C4.5 are possible candidate models for further study.
3. *Different data fuzzilization methods:* Different data fuzzification functions such as trapezoid membership functions and Gauss membership function can be applied as evolutionary selection by GA to further improve the accuracy of the proposed model.

Appendix 1

MiniTab R14 software is applied to analyze the factors design of genetic algorithm and all interaction plots of different stocks are shown in Figs. 13–15.

References

Abraham, A., Baikunth, N., & Mahanti, P. K. (2001). Hybrid intelligent systems for stock market analysis. *Lecture Notes in Computer Science, 2074*, 337–345.

Abraham, A., Philip, N. S., & Saratchandran, P. (2003). Modeling chaotic behavior of stock indices using intelligent paradigms. *Neural, Parallel and Scientific Computations, 11*, 143–160.

Abu-Mostafa, Y. S., & Atiya, A. F. (1996). Introduction to financial forecasting. *Applied Intelligence, 6*, 205–213.

Aiken, M., & Bsat, M. (1994). Forecasting market trends with neural networks. *Information Systems Management, 6*(4), 42–48.

Austin, M., & Looney, C. (1997). Security market timing using neural network models. *New Review of Applied Expert Systems, 3*, 3–14.

Baba, N., Inoue, N., & Asakawa, H. (2000). Utilization of neural networks and GAs for constructing reliable decision support systems to deal stocks. In *IEEE–INNS–ENNS international joint conference on neural networks (IJCNN'00)* (Vol. 5, pp. 5111–5116).

Braun, H., & Chandlar, S. S. (1987). Predicting stock market behavior through rule induction: An application of the learning-from-example approach. *Decision Science, 18*(3), 415–429.

Brownstone, D. (1996). Using percentage accuracy to measure neural network predictions in stock market movements. *Neurocomputing, 10*, 237–250.

Chang, P. C., & Liu, C. H. (2006). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Applications, 34*(1), 135–144.

Chang, P. C., & Wang, Y. W. (2006). Fuzzy Delphi and back-propagation model for sales forecasting in PCB industry. *Expert Systems with Applications, 30*(4), 715–726.

Chang, P. C., Wang, Y. W., & Yang, W. N. (2004). An investigation of the hybrid forecasting models for stock price variation in Taiwan. *Journal of the Chinese Institute of Industrial Engineering, 21*(4), 358–368.

Chang, P. C., & Warren Liao, T. (2006). Combing SOM and fuzzy rule base for flow time prediction in semiconductor manufacturing factory. *Applied Soft Computing, 6*(2), 198–206.

Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan stock index. *Computers and Operations Research, 30*, 901–923.

Chi, S. C., Chen, H. P., & Cheng, C. H. (1999). A forecasting approach for stock index future using grey theory and neural networks. In *IEEE international joint conference on neural networks* (pp. 3850–3855).

Corani, G., & Guariso, G. (2005). Coupling fuzzy modeling and neural networks for river flood prediction. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 35*(3), 382–390.

Edwards, R. D., & Magee, J. Jr., (2001). *Technical analysis of stock trends* (8th ed.). Amacom.

Golan, R. H., Ziarko, W. (1995). A methodology for stock market analysis utilizing rough set theory. In *Proceedings of the IEEE/IAFE 1996 conference on computational intelligence for financial engineering* (pp. 32–40).

Hiemstra, Y. (1994). A stock market forecasting support system based on fuzzy logic. In *Proceedings of the 27th annual Hawaii international conference on system sciences*.

Hobbs, A., & Bourbakis, N. G. (1995). A neurofuzzy arbitrage simulator for stock investing, In *International conference on computational intelligence for financial engineering (CIFER), New York* (pp. 160–177).

Izumi, K., & Ueda, K. (1999). Analysis of exchange rate scenarios using an artificial market approach. In *Proceeding of the international conference on artificial intelligence* (Vol. 2, pp. 360–366).

Janikow, C. Z. (1998). Fuzzy decision tree: Issues and methods. *IEEE Transactions on System, Man, and Cybernetics – Part B: Cybernetics, 28*(1), 1–14.

Jaruszewicz, M., & Mandziuk, J. (2004). One day prediction of NIKKEI index considering information from other stock markets. In *International conference on artificial intelligence and soft computing ICAISC 2004*. New York: Springer.

Khokaharm, R. H., & Sap, M. N. M. (2004). Fuzzy decision tree for data mining of time series stock market database, In *The fifth international conference for the critical assessment of microarray data analysis, Durham, North Carolina, USA*. (pp. 364–371).

Kim, K. J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications, 19*, 125–132.

Kimoto, T., & Asakawa, K. (1990). Stock market prediction system with modular neural network. In *IEEE international joint conference on neural network* (pp. 1–6).

Kosko, B. (1992). *Neural network and fuzzy systems*. Englewood Cliffs, NJ: Prentice-Hall.

Larsen, H. L., & Yager, R. R. (2000). A framework for fuzzy recognition technology. *IEEE Transaction on Systems, Man, and Cybernetics, Part C, 30*, 65–76.

Lee, J. W. (2001). Stock price prediction using reinforcement learning. In *IEEE international joint conference on neural networks* (pp. 690–695).

Medasani, S., Kim, J., & Krishnapuram, R. (1998). An overview of membership function generation techniques for pattern recognition. *International Journal of Approximate Reasoning, 19*(3–4), 391–417.

Mingo López, L. F., Díaz, M. A., Palencia, V., Santos, E., & Jiménez, P. (2002). IBEX-35 stock market forecasting using time delay connections in enhanced neural

networks. In *World multiconference on systemics, cybernetics and informatics* (pp. 455–460).
Montana, D., & Davis, L. (1989). Training feed forward neural networks using genetic algorithms. In *Proceedings of 11th international joint conference on artificial intelligence* (pp. 762–767). San Mateo, CA: Morgan Kaufmanns.
Mugambi, E. M., Hunter, A., Oatley, G., & Kennedy, L. (2004). Polynomial-fuzzy decision tree structures for classifying medical data. *Knowledge-Based System, 17*(2–4), 81–87.
Murata, T., & Ishibuchi, H. (1996). A genetic-algorithm-based fuzzy partition method for pattern classification problems. In *Proceedings of genetic algorithm and soft computing* (pp. 555–578).
Murata, T., Ishibuchi, H., & Gen, M. (1998). Adjusting fuzzy partitions by genetic algorithms and histograms for pattern classification problems. In *Proceedings of IEEE conference on computational intelligence* (pp. 9–14).
Nenortaite, J., & Simutis, R. (2004). Stocks' trading systems based on the particle swarm optimization algorithm. *Computational Science – ICCS, 3039*(4), 843–850.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*.
Schierholt, K., & Dagli, C. H. (1996). Stock market prediction using different neural network classification architectures. In *Proceedings of the IEEE/IAFE 1996 conference on computational intelligence for financial engineering* (pp. 72–78).
Sitte, R., & Sitte, J. (1999). Analysis of the predictive ability of time delay neural networks applied to the S&P 500 time series. *IEEE Transaction on Systems, Man, and Cybernetics, Part C, 30*, 568–572.
Slim, C. (2004). Forecasting the volatility of stock index returns: A stochastic neural network approach. *Computational Science and Its Applications, 3*, 935–944.
Sorensen, E. H., Miller, K. L., & Ooi, C. K. (2000). The decision tree approach to stock selection. *Journal of Portfolio Management, Fall*, 42–45.
Su, M.-C., Liu, C.-W., & Tsay, S.-S. (1999). Neural-network-based fuzzy model and its application to transient stability prediction in power systems. *IEEE Transaction on Systems, Man, and Cybernetics, Part C, 29*, 149–157.
Weber, R. (1985). Fuzzy-ID3: A class of methods for automatic knowledge acquisition. In *Proceedings of the 2nd international conference on fuzzy logic and neural network* (pp. 265–268).
White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns, In *Proceedings of the 2nd annual IEEE conference on neural networks, II* (pp. 451–458).
Yao, J., & Poh, H. L. (1995). Forecasting the KLSE index using neural networks. In *IEEE international conference on neural networks* (Vol. 2, pp. 1012–1017).
Yoon, Y., & Swales, J. (1991). Prediction stock price performance: A neural network approach. In *Proceeding of 24th annual Hawaii international conference on system science* (pp. 156–162).
Yuan, Y., & Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems, 69*, 125–139.
Yu, L., Wang, S., & Lai, K. K. (2005). Mining stock market tendency using GA-based support vector machines. *Lecture Notes in Computer Science, 3828*, 336–345.
Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338–353.
Zhang, G. P. (2007). Avoiding pitfalls in neural network research. *IEEE Transaction on Systems, Man, and Cybernetics, Part C, 37*, 3–16.
Zhang, Y.-Q., Akkaladevi, S., Vachtsevanos, G., & Lin, T. Y. (2002). Granular neural web agents for stock prediction. *Soft Computing, 6*, 406–413.