Journal of Engineering and Applied Sciences 14 (8): 2752-2763, 2019
ISSN: 1816-949X

Forecasting Stock Index Data Using Hybrid Models

[1]Kumar Vasimalla, [2]C. Narasimham and [3]Deekshitha
[1]Research and Development Center, Bharathiar University, Tamil Nadu, India
[2]Vignan Institute of Engineering and Technology, Department of Computer Science and Engineering, Visakhapatnam, Andhra Pradesh (AP), India
[3]Department of Computer Science, Central University of Kerala, Kasaragod, Kerala, India

Abstract: Forecasting is an important and widely popular topic in the research of system modeling. In this study, we proposed a six 2-stage hybrid prediction models, wherein Laplacian Score (LS), Multi Cluster based Feature Selection (MCFS), Correlation Based feature Selection (CBS) is used to construct Stage-1, followed by invoking Adaptive Network based Fuzzy Inference System (ANFIS) trained by Genetic Algorithm (GA), Particle Swarm Optimization (PSO)(Stage-2). We tested our model with Hang Seng Index (HSI) data and TAIEX stock market transaction data from 1998-2006. The results compared with the existing models in the literature, the comparison shows that the proposed model LS+ANFIS+GA outperformed the listing models in terms of both of Root Mean Squared Error (RMSE) and Theil's U statistic.

Key words: Hybrid model, Laplacian score, multi cluster, co-relation, ANFIS, Genetic algorithm, particle swarm optimization

INTRODUCTION

Time series is a sequence of data gathered based on time. Time series forecasting is a major task in time series analysis and it is a technique for predicting future values based on history and it is the basic technique in real life to take decisions. The purpose of forecasting is to minimize the risk in decision making and reduce unanticipated cost. Forecasting used in several areas of life including finance (e.g., exchange rate prediction, gold price prediction, stock price prediction, demand forecasting and volatility forecasting), medicine (e.g., disease prediction, heart diagnosis prediction, tumor detection) and business (e.g., demand forecasting, monthly sales forecasting).

There are two types of forecasting-qualitative and quantitative. Qualitative techniques are deployed where historical data is not available. These methods generate forecasts based on the judgment of experts. On the other hand, quantitative forecasting methods are used when historical data is available and can be analyzed to get estimates. Time series forecasting is an example for quantitative forecasting technique. In this, data is collected over a period of time to identify trends. Most of the companies want to predict the future of their sales, stock prices to develop their policy based on that. So, financial forecasting is an important challenge for stock dealers, stock investors or stock brokers. It is difficult to make decisions on buying or selling stocks, because many variables are taken into consideration which may affect stock market.

There are several time series models available for forecasting stock market. But accurate predictions of stock markets are important. A good forecast will help the company to exploit all the potential opportunities that the environment can offer or protect them from a disastrous choice. Techniques such as Regression models and ARIMA Models (Laboissiere *et al*., 2015) are also used for stock price forecasting. These methods have been used extensively in past. However, they failed to give accurate results for non-linear time series and have some other limitations. Self-learning techniques like Artificial Neural Networks (ANN) have been used in stock market prediction during the last decade. It is possible to forecast linear and nonlinear data using ANN. ANN are used in several areas like pattern recognition, classification, prediction and process control. It is a computing system which is inspired by the biological neural networks. Fuzzy logic principles are also used for forecasting. It doesn't learn from data. But fuzzy rules are easy to understand. Now a days fuzzy logic with neural networks are used for better prediction.

Challenges: The researchers of financial forecasting facing the following challenges:

Corresponding Author: Kumar Vasimalla, Department of Computer Science, Central University of Kerala, Kerala, India

Forecasting errors: The data collection is a tedious task because data is available in number of ways and contains number of variables, choosing the correct data, number of variables and features selection reflects forecasting accuracy.

Handling uncertainty: Time series data is a complex data and noisy in nature, prediction accuracy depends on the noise removal techniques.

Number of variables: In most of the cases, researchers choose only one input to forecasting model, this yields less prediction accuracy. Because multiple variables are taken into consideration for good accuracy. Rules generated by neural networks are difficult to understand, availability of data.

Literature review: A statistical approach, Box and Jenkins (Engle, 1982) have developed the integrated Autoregressive Moving Average (ARIMA) methodology for linear time series. Statisticians in number of ways have addressed the restriction of linearity in the Box-Jenkins approach. Robust versions of various ARIMA models have been developed. In addition to this, nonlinear models are developed. Most of the researchers used data mining to financial analysis. A methodology proposed by Ricardo (Laboissiere *et al*., 2015), forecasts the maximum and minimum day stock prices of three Brazilian power distribution countries. In this most relevant features are calculated which predict minimum and maximum stock price of company. Artificial neural network is used for prediction.

Jingtao *et al*. (1999) used a back propagation neural network which finds the relationship between the technical indicators and the levels of the index in the market, he made predictions without the use of extensive market data or knowledge. Hybrid model of linear and nonlinear time series models also available for forecasting. A hybrid models with neural network and time series models, proposed by Roh (2007) for forecasting the volatility of stock price index in two viewpoints deviation and direction. Cheng *et al*. (2010) proposed a hybrid model which is based on rough set theory and genetic algorithm for forecasting stock index. In this research, technical indicators are used as feature and which is selected by using correlation matrix, efficient rules are extracted from this model and which would help for forecasting.

Chen *et al*. (2008) proposed a comprehensive fuzzy time-series which factors linear relationships between recent periods of stock prices and fuzzy logical relationships (nonlinear relationships) mined from time-series. Here multi-period adaptation model is used for forecasting stock prices. Wei (2016) proposes a hybrid time-series Adaptive Network based Fuzzy Inference System (ANFIS) Model based on Empirical Mode Decomposition (EMD) to forecast stock price for Taiwan stock exchange capitalization weighted stock index (TAIEX).

An artificial neural network based ANFIS approach is used for Automatic Generation Control (AGC) of a three unequal area hydrothermal system (Swasti and Khuntia, 1994). ANFIS combines both artificial neural network and fuzzy logic advantages. Adaptive Neuro-Fuzzy Inference System (ANFIS) is used for multi-criteria decision making in supplier evaluation and selection problem (Ozkan and Inal, 2014). The contemporary supply-chain management is looking for both quantitative and qualitative measures other than just getting the lowest price. After evaluating a number of distinct suppliers, determining the reliable suppliers by ANFIS Model with better approximation will support decision makers. Majhi and Anish (2015) proposed a Multi Objective Particle Swarm Optimization (MOPSO) and Non-dominated Sorting Genetic Algorithm Version-II (NSGA-II) have been introduced to effectively train the adaptive stock market prediction models which simultaneously optimize four performance measures. Ghore and Goswami (2015) proposed a method called adaptive neuro-fuzzy inference system which forecast electricity load of Chhattisgarh Grid. There are several hybrid models available for forecasting stock. In past, ARIMA Models are used for forecasting but it is only applicable for linear time series. Some others proposed an artificial neural network which model both linear and nonlinear time series. Aladag *et al*. (2009) proposed a hybrid model called Elman's recurrent neural networks for forecasting stocks. It is a combination of both ARIMA and artificial neural network.

Time series with nonlinear moving average components are not well modeled by feed forward networks or linear models but can be modeled by recurrent networks. Connor *et al*. (1991) proposed a Nonlinear Autoregressive-Moving Average (NARMA) model for electric load forecasting. Azadeh *et al*. (2011) presents an Adaptive Network based Fuzzy Inference System (ANFIS) Auto Regression (AR) Analysis of Variance (ANOVA) algorithm to improve oil consumption estimation and policy making. ANFIS is examined against Auto Regression (AR) in Canada, United Kingdom and South Korea. The algorithm for calculating ANFIS performance is based on its closed and open simulation abilities. This is the first study that introduces an integrated ANFIS-AR-ANOVA algorithm with preprocessing and post processing modules for

improvement of oil consumption estimation in industrialized countries. A new hybrid model for forecasting the electric power load several months ahead is proposed by Lee and Hong (2015). To allow for distinct responses from individual load sectors, this hybrid model, which combines dynamic (i.e., air temperature dependency of power load) and fuzzy time series approaches is applied separately to the household, public, service and industrial sectors. Ravi *et al.* (2017) proposed two 3-stage hybrid prediction models wherein chaos theory is used to construct phase space (stage 1) followed by invoking Multilayer Perceptron (MLP) (stage 2) and multi-objective evolutionary algorithms, includes Multi-Objective Particle Swarm Optimization (MOPSO) and Non-dominated Sorting Genetic Algorithm (NSGA-II), (stage 3) in tandem. In both of these hybrid models, stage 3 improves the prediction yielded by stage 2. Chaos+MLP+NSGA-II yields better forecasting performance compared to other models.

Feature selection is more important in forecasting, because most of the variables mislead the forecasting. Li *et al.* (2017) proposed a novel rolling bearing fault diagnosis strategy which is based on Improved Multiscale Permutation Entropy (IMPE), Laplacian Score (LS) and Least Squares Support Vector Machine-Quantum behaved Particle Swarm Optimization (QPSO-LSSVM). IMPE which was developed to reduce the variability of entropy estimation in time series was used to obtain more precise and reliable values in rolling element bearing vibration signals. The extracted features were then refined by LS approach to form a new set of feature vector which contains main unique information. LS is based on Laplacian Eigen maps and locality preserving projection. By estimating the importance of each feature, LS has the ability to reorder these features according to their locality preserving power. In this, selected feature vector was input to QPSO-LSSVM classifier to distinguish the health status of rolling bearings. The comparative test results indicate that the proposed methodology led to significant improvements in bearing defect identification.

Krzysztof and Halina (2006) proposed a modified pair-wise selection strategy. They suggests that computation time can be significantly lowered while maintaining the quality of the selected feature sets by using mixed univariate and bi-variate feature evaluation based on the correlation between the features. They presents the comparison of the performance of their method with that of the unmodified pairwise selection strategy based on several well-known benchmark sets. Experimental results show that in most cases, it is possible to lower computation time and that with high statistical significance the quality of the selected feature sets is not lower compared with those selected using the unmodified pairwise selection process. In a Handwriting Character Recognition (HCR), the set of features plays as main issues, as procedure in choosing the relevant feature that yields minimum classification error. To overcome these issues and maximize classification performance, many techniques have been proposed for reducing the dimensionality of the feature space in which data have to be processed. Mohamad *et al.* (2015) proposed an overview of some of the methods and approach of feature extraction and selection. Throughout this study, we apply the investigation and analyzation of feature extraction and selection approaches in order to obtain the current trend.

Su and Cheng (2016) proposed an Adaptive Neuro Fuzzy Inference System (ANFIS) time series model based on Integrated Non-linear Feature Selection (INFS) method for forecasting. Firstly, this study proposed an integrated nonlinear feature selection method to select the important technical indicators objectively. Secondly, it used ANFIS to build time series model and test forecast performance, then utilized adaptive expectation model to strengthen the forecasting performance. In order to evaluate the performance of proposed Model TAIEX and HSI stock data are used and methodology is compared with other models. The summary of above studies have four drawbacks:

- Most of the time series models are only suitable for linear datasets
- Previous researches selected important technical indicators dependent on subjective experiences and opinions
- Most of the methods used some assumptions about the variables used in the analysis (Jilani and Burney, 2008), so, it is limited to be applied to all datasets
- Most conventional time series models considered only one variable to forecasts stock price
- The rules generated from ANN are not easy to understand
- Error rate for most of the models is high

Therefore, we proposed a novel method called ANFIS (Adaptive Neuro-Fuzzy Inference System) trained by GA/PSO for decreasing error rate of stock forecasting using multiple variables.

MATERIALS AND METHODS

Technical indicators: Technical indicators are used for forecasting the stock market. In our proposed model, we considered them as inputs to the model. But not all the

technical indicators gives good prediction, so, we need to select the best indicators. Technical indicators used in our proposed model are as follows:

Moving Average Convergence and Divergence (MACD): MACD is a technical analysis indicator created by Gerald Appel, used to spot changes in the strength, direction, momentum and duration of a trend in a stock's price. It is a difference of 2 days exponential moving averages. Generally used days for finding MACD is 12 days Exponential Average (EMA (12)) and 26 days Exponential Average (EMA (26)). Exponential moving average is a type of infinite impulse response filter that applies weighting factors which decrease exponentially. Here, it is used for estimation of the current changes in a trend. MACD is calculated using the following equation:

$$MACD = EMA(12)\text{-}EMA(26)$$

$$EMA = S_t = \alpha * Y_t + (1\text{-}\alpha) * S_{t-1}$$

Where:

α = A degree of weight decrease and $\alpha = 2/N+1$
Y_t = An observation at a time period t
S_t = A value of EMA

Here S_1 is undefined and $S_2 = Y_1$ and in other cases S_2 is an average of the first 4 or 5 observations.

Simple Moving Average (SMA): SMA is an un-weighted measure of the previous n data points. In our proposed method, we used simple moving average for 5 and 10 days. For example, 10 days Simple Moving Average (SMA (10)) of open price is the mean of the previous 10 days closing prices. The following equation explain SMA:

$$SMA(10) = \frac{P_t + P_{t-1} + , \cdots, P_{t-9}}{10}$$

P_t, P_{t-1}, ..., P_{t-9} are 10 days opening price.

Relative Strength Indexing (RSI): RSI compares the magnitude of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset. It is used to chart the current and historical strength or weakness of a stock based on the closing prices of a recent trading period. The RSI computes momentum as the ratio of higher closes to lower closes. Stocks which have more or stronger positive changes have a higher RSI and stocks which have more or stronger negative change have a lower RSI:

$$RSI = 100 - \frac{100}{1+RS}$$

RS (Relative Strength) is calculated by following equations:

$$\text{Avg. gain} = \text{Total gain/n}$$

$$\text{Avg. loss} = \text{Total loss/n}$$

RS = Avg. gain/Avg. loss where n is number of days.

BIAS5: BIAS5 is the difference between closing price and moving average for 5 days (SMA5):

$$BIAS5 = SMA5 - \text{Closing price}$$

BIAS10: BIAS10 is the difference between closing price and moving average for 10 days (SMA10):

$$BIAS10 = SMA10 - \text{Closing price}$$

Williams R%: Williams's R% or just R% is a technical analysis oscillator showing the current closing price in relation to the high and low of the past N days. It was developed by Larry Williams. Its purpose is to tell whether a stock or commodity market is trading near the high or the low or somewhere in between, of its recent trading range:

$$R\% = \frac{Close_T - High_N}{High_N - Low_N}$$

Where:

$Close_T$ = Current day closing price
$High_N$ = N days High price
Low_N = N days Low price

Feature selection methods

Laplacian Score (LS): LS is an unsupervised learning method. Unsupervised methods are harder to solve due to the absence of class labels. Previous unsupervised feature selection methods are wrapper techniques which needs a learning algorithm to evaluate the candidate feature subsets. Another category of feature selection method is a filter method which is independent of any learning algorithm. In the case of LS, importance of a feature is evaluated by its power of locality preserving on calculated Laplacian score. For many learning problems the local structure of the data space is more important than the global structure. In order to model the local geometric structure, we construct a nearest neighbor graph. LS is based on Laplacian Eigen maps and locality preserving projection.

Consider m data points each with n features. Let L_r denote the LS of the rth feature value where r = 1, 2, …, L_n. Let r_i denotes the ith element of the r^{th} feature where i = 1, 2, …, L_m. The LS method can be explained as: The first step starts with constructing a nearest neighbor graph G with m nodes. There is a one-to-one relationship between the ith node and the ith data point x_i. Then put an edge between nodes i and j, if x_i and x_j are "close" where "close" is defined as x_i is among k nearest neighbors of x_j or vice versa. Thus, an edge should be constructed between nodes i and j sharing the same label if the label information is available:

$$S_{ij} = \begin{cases} \exp\dfrac{-\left|x_i - x_j\right|^2}{t} & \text{if i and j are connected} \\ 0 & \text{otherwise} \end{cases}$$

where, t is a suitable constant. For the rth feature value, define $A^T f_r = [f_{r1}, f_{r2}, ..., f_{rm}]$. D = diag (S), I= $[1, 1...1]^T$, L = D-S. The matrix L is called graph Laplacian. Remove the mean from the samples as given:

$$F_r = f_r - \frac{f_r^T DI}{I^T DI}$$

LS of the rth feature could be mathematically defined by the following equation:

$$L_r = \frac{\sum_{ij}\left(f_i - f_i\right)^2 S_{ij}}{\text{var } f_r}$$

Multi Cluster based Feature Selection (MCFS): MCFS is an unsupervised learning technique, it also based on finding k-nearest neighbor graph. MCFS consists of three steps. In the first step, it constructs a p-nearest neighbor graph and gets the graph affinity matrix S and the Laplacian matrix L. Then a flat embedding that unfolds the data manifold can be obtained by spectral clustering techniques. In the second step, since, the embedding of data is known, MCFS takes advantage of them to measure the importance of features by a regression model with a L_1-norm regularization. Specifically, given the ith embedding e_i, MCFS regards it as a regression target to minimize:

$$\min_{wi} \left\| X^T w_i - e_i \right\|^2 + \alpha \left\| w_i \right\|$$

where, w_i denotes the feature coefficient vector for the ith embedding. By solving all K sparse regression problems, MCFS obtains K sparse feature coefficient vectors W = $[w_1, ..., w_K]$ and each vector corresponds to one embedding of X. In the third step, for each feature f_j, the MCFS score for that feature can be computed as MCFS (j) = $\max_i |W(j, i)|$. The higher the MCFS score, the more important the feature is.

Correlation based Feature Selection (CFS): Correlation between set of data is a measure of how well they are related. A good feature subset should contain features that are highly correlated with the output variable. Common measure of correlation is Pearson correlation. It shows linear relationship between the data. Correlation (R) between x and y:

$$R = \frac{n\left(\sum xy\right) - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

Forecasting model

Adaptive Neuro Fuzzy Inference System (ANFIS): ANFIS is a kind of artificial neural network which is based on Takagi-Sugeno fuzzy inference system. It integrates both neural network and fuzzy logic principles. It is an inference system with a set of fuzzy if then rules that have learning capability to approximate nonlinear functions. Fuzzy logic itself can't learn from the data but fuzzy based models are easily understood as it utilizes linguistic terms rather than numeric and the structure of if then rules. Linguistic variables are defined as variables whose values are words or sentences in a natural language with associated degrees of membership.

ANFIS, the crisp input signal is converted to fuzzy inputs by the membership function. The membership function used in our proposed model is Gaussian. It is a curve that defines how each point in the input space is mapped to a membership value which is between 0 and 1. The input space is sometimes referred to as universe of discourse. ANFIS method uses fuzzy c means clustering (Suganya and Shanthi, 2012) to partition the universe of discourse for input variables and then generates the fuzzy inference system. The fuzzy input along with membership function fed into the neural network block. The neural network consist of set of rules which is connected to inference engine. Fuzzy inference is a method that interprets the values in the input vectors and based on some set of rules which assign values to the output

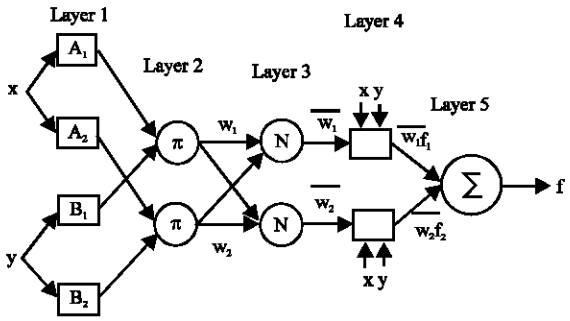

Fig. 1: Layer wise description of ANFIS

vector. For example, if service is poor then tip is cheap. Here, service is interpreted as poor and tip is assigned to be cheap. Inference engine is trained by back propagation method for the proper selection of rule base. In our proposed model, we used Genetic algorithm or particle swarm optimization method for training. After training, proper rules can be generated and fired from the neural network block to yield optimal output. This linguistic output is converted into crisp output by the de-fuzzier unit. These operations are followed by different layers. ANFIS has 5 layers. Layer-wise description is given in Fig. 1.

Layer 1: This layer is called as the fuzzification layer, here the crisp input is fed into the node I which is associated with a linguistic label A_i or B_{i-2}, thus, the membership function determines the membership level of the given input. The output of each node calculated with the following equation:

$$O_{1,i} = \mu_{A_i}(x), \text{ for } i = 1, \cdots, n$$

$$\mu_{A_i(x)} = \frac{1}{1+\left[\left(\frac{x-c_i^2}{a_i}\right)\right]_i^b}$$

where, a_i, b_i, c_i are the parameters and b_i is a positive value and c_i denotes the center of curve.

Layer 2: Every node in this layer is fixed and labeled as $O_{2,1}$, the output of each node is the product of all the incoming values and it is a firing strength of a rule, the following equation used to calculate firing strength:

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_i}(x)$$

Layer 3: This layer is known as a rule layer. Output of this layer is a ratio of the individual rule's firing strength to the sum of all rules firing strengths:

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1+w_2}$$

This weight is called as a normalized weights.

Layer 4; (de-fuzzification layer): This layer calculates the individual output values y from the inferring of rules from the rule base. Individual nodes of this layer are connected to the respective normalization node in layer 3 and also receive the input signal; the calculation performed by using the following equation:

$$O_{4,i} = \overline{w_i}f_i = \overline{w}(p_i+ix+q_iy+r_i)$$

where, p_i q_i, r_i are the parameters.

Layer 5: Last layer is an output layer. It has only one node and it calculates the sum of all the outputs carry from nodes of the de-fuzzification layer to produce the overall ANFIS output:

$$O_{5,i} = \sum_{i=1}^{n}\left(\overline{w_i}f_i\right) = \frac{\sum_{i=1}^{n}\overline{w_i}f_i}{\sum_{i=1}^{n}w_i}$$

Training mechanism of ANFIS Model

Genetic Algorithm (GA): GA is inspired by Darwin's evolutionary theory, it starts with a set of solutions called population. Solutions are selected based on fitness value, the more suitable and they are more chance to reproduce. This is repeated until some condition is satisfied.

Algorithm 1; Genetic algorithm:

1. Start: Generate random set of n populations
2. Fitness: Evaluate fitness of each chromosome x in the population
3. New population: Create a new population by repeating following procedures
   - Selection: Select two best parent chromosomes according to their fitness. The Chromosome with highest fitness value will be selected more times
   - Crossover: With a crossover probability, crossover the parents to form a new offspring. If no crossover is performed, offspring is an exact copy of parents
   - Mutation: With a mutation probability, mutate new offspring at each locus. Accepting: If the generated offspring is good according to fitness, then place new offspring in a new population set
   - Replace: Use new set of population for further run of algorithm
4. Test: If the end condition is satisfied, stop and return the best solution in current population. End condition may be a defined number of iterations or a trial and error
5. Loop: Go to step 2

Particle Swarm Optimization (PSO): PSO is a population based optimization technique, it is similar to GA. In this

method, system is initialized with a set of random solutions, each solution is known as particle. PSO doesn't have crossover and mutation, particles fly through a problem space by following the current optimum particles. PSO stimulates the behavior of bird flocking, suppose a group of birds are randomly searching food in an area, if they doesn't know the position of food but in every iteration, they know how far the food is from the current position. One of the best solution to find food is follow a bird which is nearer to the food. Algorithm starts with generating set of n particle and searches for optimum by updating generation, each particle has a fitness value. In every iteration each particle updates by two best values, gbest and pbest. The best fitness value of an individual is taken as pbest. The best fitness value of all pbest is taken as the gbest. After finding these two values, particle updates it's velocity and position by following equations:

$$v[\ ] = v[\ ] + c1 * mad(\ )(pbest[\ ] - present[\ ]) + c2 * rand(\ )(gbest[\ ] - present[\ ])$$

Position is updated by:

$$Present[\ ] = Present[\ ] + v[\ ]$$

Evaluation

Root Mean Square Error (RMSE): RMSE is one of the performance measure of a model, it is a popularly used evaluation criteria in several models. Equation of RMSE is given as:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - f_t)^2}{n}}$$

Where:

y_t = The real stock index

f_t = The forecasting stock index and n is the size of data

Theil's U statistics: It is another statistic for finding accuracy of a model. Equation for this is:

$$U = \sqrt{\frac{\sum_{t=1}^{n}(y_t - f_t)^2}{\sum_{t=1}^{n} y_t^2 + \sqrt{\sum_{t=1}^{n} f_t^2}}}$$

U value is between 0 and 1, the U value close to 0 indicates the model with good accuracy. Proposed model is explained in a Fig. 2.

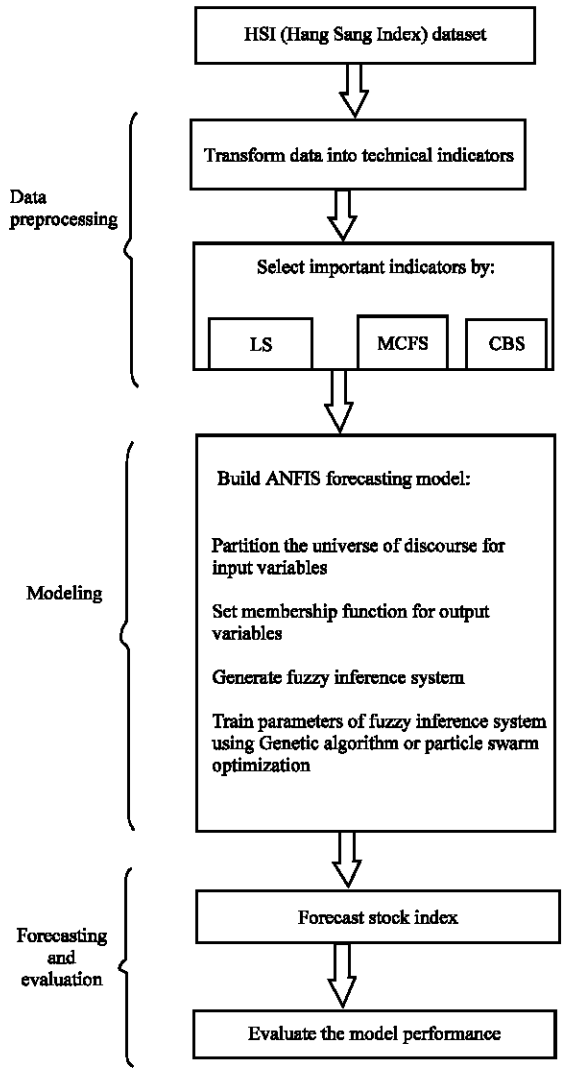

Fig. 2: Proposed model

RESULTS AND DISCUSSION

In order to verify the accuracy of forecasting models, we took HIS and TAIEX datasets. Performance compared with Chen's Model (Chen, 1996), Yu's Model (Yu, 2005), INFS based on ANFIS INFS+ANFIS) and INFS based on Support Vector Regression (INFS+SVR) and adaptive expectation model (INFS+ANFIS) (Su and Cheng, 2016).

HSI dataset: The HSI is a free float-adjusted market capitalization weighted stock market index in Hong Kong, it is used to record and monitor daily changes of the largest companies of the Hong Kong Stock Market and is the main indicator of the overall market performance in Hong Kong. HIS is employed as experimental dataset in our proposed model, daily stock prices of HSI from

January 1998 to December 2006 used as an experimental dataset. The data set contains the open, high, low, close, trading volume and adjusted close prices of HSI Stock on every day throughout these 9 years. Data divided into year wise, for each year from January to October is taken as training data and November and December is taken as a testing data. Seven technical indicators computed and used in the prediction model are simple moving average for 5 days, simple moving average for 10 days, relative strength index, moving average convergence or divergence, Williams's R%, BIAS5 and BIAS10.

Seven technical indicators/features are extracted from the data, best three features selected from the seven technical indicators using LS, MCFS and CBS methods in the data pre-processing stage. In the data modelling stage ANFIS Model is used, first it defines law of discourse for each input, secondly membership function for output is defined then fuzzy c means clustering is used to generate fuzzy inference system. We set number of clusters as 10 model is trained by using evolutionary algorithms, i.e., GA/PSO, we set epoch as 100. To evaluate the model, we used Theil's U statistic and RMSE. Best three features selected by various methods are given in Table 1 and 2. Figure 3-8 representing the actual and predicted closing prices by the six models for the year 1998. The blue line with bubbles indicates actual closing price and red line indicates predicted closing price.

TAIEX dataset: The TAIEX dataset from 1998-2006 downloaded from the Yahoo finance website https://finance.yahoo.com/quote/%5ETWII/history/. For each year the first 10 months data took as training data, and the remaining as test data. For detailed illustration of the proposed model the 2002 year TAIEX data is taken as an

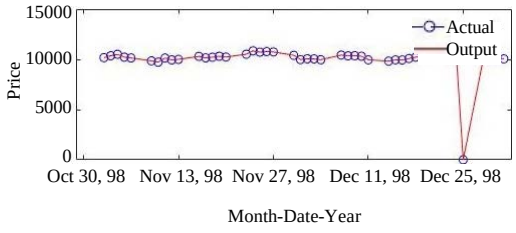

Fig. 3: LS+ANFIS with GA Model for HIS of 1998

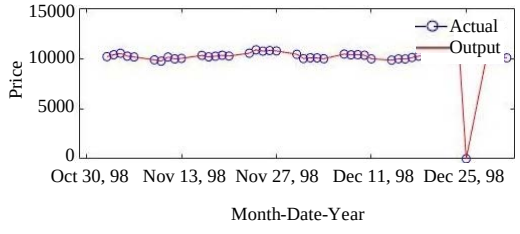

Fig. 4: LS+ANFIS with PSO for HSI of 1998

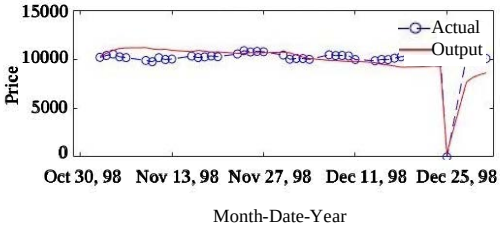

Fig. 5: MCFS+ANFIS with GA for HIS of 1998

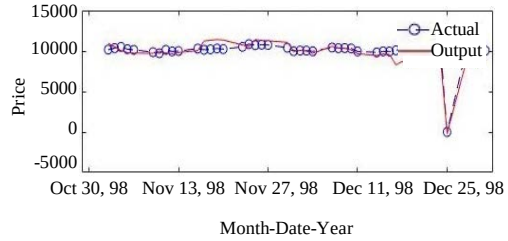

Fig. 6: MCFS+ANFIS with PSO for HIS of 1998

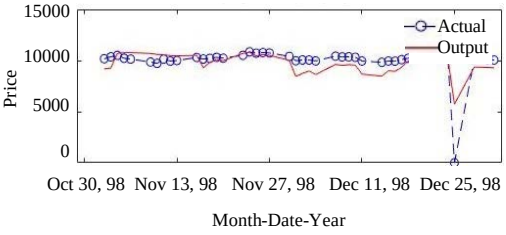

Fig. 7: CBS+ANFIS with GA for HSI of 1998

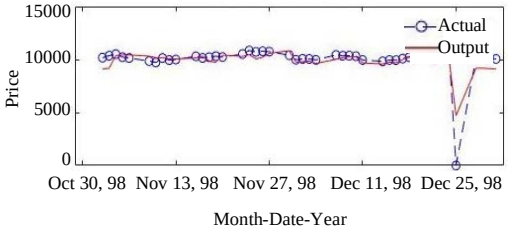

Fig. 8: CBS+ANFIS with PSO for HIS of 1998

Table 1: The technical indicators selected by LS method

Years	Variable 1	Variable 2	Variable 3
1998	MOV_10	MOV_5	BIAS_10
1999	MOV_10	MOV_5	BIAS_10
2000	BIAS_5	MOV_5	BIAS_10
2001	BIAS_10	MOV_10	BIAS_5
2002	BIAS_5	MOV_5	BIAS_10
2003	MOV_10	MOV_5	BIAS_10
2004	BIAS_5	MOV_5	BIAS_10
2005	BIAS_5	MOV_5	BIAS_10
2006	BIAS_5	MOV_5	BIAS_10

example. The three important indicators are selected by LS, MCFS, CBS feature selection methods. The three important indicators selected by LS method shown in Table 2, the three technical indicators used as an input variables for the ANFIS forecasting model. ANFIS Model trained with GA/PSO, training epoch was set as 100. Results of this model evaluated with RMSE and Theil's U statics and shown in Table 3-9.

Comparison: The experimental results comparisons are listed in Table 4, 6, 10 and 11. In order to show the accuracy of proposed model, this study compares the proposed models with other models INFS+ANFIS+AEX (Su and Cheng, 2016) as shown in Table 4. The comparison performed based on RMSE and Thesil's U static. The comparison results shows that the proposed Model LS+ANFIS trained with GA is the best performing model compared to all other models. The comparison tables for HSI and TAIEX datasets shows that the proposed Models LS+ANFIS with GA, LS+ANFIS with

Table 2: The technical indicators selected by LS method for TAIEX dataset

Years	Variable 1	Variable 2	Variable 3
1998	MOV_10	MOV_5	BIAS_10
1999	MOV_10	MOV_5	BIAS_10
2000	MOV_10	MOV_5	BIAS_10
2001	MOV_10	MOV_5	BIAS_10
2002	BIAS_5	MOV_5	BIAS_10
2003	MOV_10	MOV_5	BIAS_10
2004	BIAS_5	MOV_5	BIAS_10
2005	MOV_10	MOV_5	BIAS_10
2006	BIAS_5	MOV_5	BIAS_10

Table 3: The results of various models based on RMSE for HSI dataset

Test set	MCFS+ANFIS+GA	MCFS+ANFIS+PSO	CBS+ANFIS+GA	CBS+ANFIS+PSO	LS+ANFIS+PSO	LS+ANFIS+GA*
1998	881.26	815.26	1.1442×10^{3}	860.96	$1.63*10^{-12}$	$1.29*10^{-12}$
1999	2.920×10^{3}	3.448×10^{3}	1.8661×10^{3}	1.777×10^{3}	$2.52*10^{-12}$	$1.82*10^{-12}$
2000	1.006×10^{3}	2.485×10^{3}	1.722×10^{3}	2.1102×10^{3}	$6.29*10^{-12}$	$3.86*10^{-12}$
2001	1.4176×10^{3}	1.633×10^{3}	1.84×10^{-12}	1.02×10^{-11}	$11.00*10^{-12}$	$0.51*10^{-12}$
2002	937.51	933.09	4.67×10^{-13}	1.52×10^{-12}	$2.14*10^{-12}$	$0.35*10^{-12}$
2003	2.910×10^{3}	2.352×10^{3}	1.4231×10^{3}	1.178×10^{3}	$3.01*10^{-12}$	$3.56*10^{-12}$
2004	1.280×10^{3}	1.223×10^{3}	1.85×10^{-12}	5.89×10^{-12}	$8.03*10^{-12}$	$1.08*10^{-12}$
2005	660.27	682.14	249.25	191.61	$6.63*10^{-12}$	$0.39*10^{-12}$
2006	2.472×10^{3}	2.4866×10^{3}	1.4826×10^{3}	2.4092×10^{3}	$10.40*10^{-12}$	$0.56*10^{-12}$

*Better model

Table 4: The results comparisons for HSI dataset based on RMSE (Su and Cheng, 2016)

Test set	Chen (1996)	Yu (2005)	INFS+ANFIS	INFS+SVR	INFS+ANFIS+AEX (Su and Cheng, 2016)	LS+ANFIS+GA (proposed)
1998	152.14	141.56	147.59	117.28	201.18	$1.29*10^{-12}$
1999	190.11	112.99	115.58	118.91	235.68	$1.82*10^{-12}$
2000	353.00	175.63	180.03	148.52	249.28	$3.86*10^{-12}$
2001	165.31	134.39	133.59	110.76	157.97	$0.51*10^{-12}$
2002	139.64	91.43	81.11	70.38	105.60	$0.35*10^{-12}$
2003	103.96	68.07	77.31	59.22	124.15	$3.56*10^{-12}$
2004	82.32	72.34	56.44	53.07	103.28	$1.08*10^{-12}$
2005	86.12	62.52	55.97	54.72	101.66	$0.39*10^{-12}$
2006	215.64	83.92	77.03	63.22	192.52	$0.56*10^{-12}$

Table 5: The results of various models based on Theil's U statistic for HSI dataset

Test set	LS+ANFIS+PSO	MCFS+ANFIS+GA	MCFS+ANFIS+PSO	CBS+ANFIS+GA	CBS+ANFIS+PSO	LS+ANFIS+GA*
1998	8.0910^{-17}	0.043	0.0400	0.057	0.0431	$6.40*10^{-17}$
1999	8.31910^{-17}	0.105	0.1288	0.063	0.061	6.00810^{-17}
2000	21.410^{-17}	0.033	0.0346	0.0585	0.070	13.110^{-17}
2001	50.310^{-17}	0.060	0.0690	8.46×10^{-17}	4.66×10^{-16}	2.3310^{-17}
2002	11.210^{-17}	0.047	0.0460	2.44×10^{-17}	7.96×10^{-17}	1.8310^{-17}
2003	12.510^{-17}	0.136	0.1080	0.0625	0.0511	14.910^{-17}
2004	29.110^{-17}	0.048	0.0460	6.73×10^{-17}	2.14×10^{-16}	3.9410^{-17}
2005	22.210^{-17}	0.022	0.0233	0.008	0.006	1.3310^{-17}
2006	27.410^{-17}	0.069	0.0690	0.040	0.066	1.4910^{-17}

*Better model

Table 6: The results comparison for HSI dataset based on Theil's U static (Su and Cheng, 2016)

Test set	Chen (1996)	Yu (2005)	INFS+ANFIS	INFS+SVR	INFS+ANFIS+AEX (Su and Cheng, 2016)	LS+ANFIS+GA (proposed)
1998	0.0166	0.0164	0.0152	0.0106	0.00980	$6.40*10^{-17}$
1999	0.0200	0.0123	0.0122	0.0093	0.00770	6.00810^{-17}
2000	0.0125	0.0144	0.0085	0.0084	0.00803	13.110^{-17}
2001	0.0128	0.0155	0.0134	0.0071	0.00710	2.3310^{-17}
2002	0.0107	0.0085	0.0061	0.0056	0.00540	1.8310^{-17}
2003	0.0146	0.0083	0.0054	0.0052	0.00510	14.910^{-17}
2004	0.0101	0.0060	0.0040	0.0041	0.00370	3.9410^{-17}
2005	0.0046	0.0050	0.0052	0.0033	0.00340	1.3310^{-17}
2006	0.0089	0.0071	0.0050	0.0050	0.00510	1.4910^{-17}

PSO are the best performers in 9 testing datasets (1998-2006) and RMSE, Theil's U static values are smaller than other models (The column of table marked with *given good prediction accuracy).

Table 7: The particular action performed on last day of year

Date	Action
31-12-1998	Buy
30-12-1999	Sell
29-12-2000	Sell
31-12-2001	Buy
31-12-2002	Sell
31-12-2003	Sell
31-12-2004	Sell
30-12-2005	Buy
29-12-2006	Buy

Findings: According to Tables 4, 6, 10 and 11 the proposed models under the criteria of RMSE and Theil's U static are better than the fuzzy time series models (Chen,1996; Yu, 2005), since, these models only consider one variable for predicting. According to LS feature selection method, Table 1 and 2 shows that MOV_5, BIAS_5, BIAS_10 are most frequently chosen indicators for HSI dataset, MOV_10, MOV_5, BIAS_10 are most frequently chosen indicators for TAIEX dataset. Thus, we can confirm that MOV_5, BIAS_10 are best indicators for forecasting HIS and TAIEX.

Buying and selling rules: Based on the prediction of our proposed models, we can generate buying and selling rules/decisions as follows:

Table 8: The results of various models based on RMSE for TAIEX dataset

Test set	MCFS+ANFIS+GA	MCFS+ANFIS+PSO	CBS+ANFIS+GA	CBS+ANFIS+PSO	LS+ANFIS+PSO(10^{-12})	LS+ANFIS+GA(10^{-12})*
1998	1.080×10^{3}	1.080×10^{3}	899	1.059×10^{3}	1.820	1.170
1999	255.96	292.98	33.4×0^{-12}	$2.14\ 10^{-12}$	2.520	1.450
2000	2.47×10^{3}	2.45×10^{3}	2.31×10^{3}	2.31×10^{3}	0.848	0.426
2001	694.10	694.10	251.25	251.251	1.520	0.584
2002	1.04×10^{3}	1.04×10^{3}	222.99	87.26	11.90	0.416
2003	1.07×10^{3}	1.07×10^{3}	1.22×10^{3}	1.30×10^{3}	0.500	1.100
2004	453.17	453.15	193.01	193.01	1.820	1.400
2005	213.61	209.56	85.84	266.49	1.330	1.900
2006	673.69	673.69	89.8×10^{-12}	1.33×10^{-12}	5.320	0.888

*Better model

Table 9: The results of various models based on Theil's U statistic for TAIEX

Test set	MCFS+ANFIS+GA	MCFS+ANFIS+PSO	CBS+ANFIS+GA	CBS+ANFIS+PSO	LS+ANFIS+PSO (10^{-17})	LS+ANFIS+GA(10^{-17})
1998	0.074	0.074	0.064	0.0431	13.33	8.60
1999	0.016	0.019	2.16×10^{-17}	13.8×10^{-17}	9.78	9.45
2000	0.190	0.180	0.20	0.20	8.05	4.04
2001	0.070	0,070	0.026	0.026	15.80	6.07
2002	0.107	0.100	0.02	0.0093	12.80	4.47
2003	0.090	0.090	0.116	0.12	4.23	2.12
2004	0.030	0.030	0.01	0.01	15.30	2.15
2005	0.010	0.010	0.006	0.02	10.70	2.19
2006	0.040	0.040	6.03×10^{-17}	8.97×10^{-17}	35.70	5.96

Table 10: The results comparison for the TAIEX dataset based on RMSE (Su and Cheng, 2016)

Test set	Chen (1996)	Yu (2005)	INFS+ANFIS	INFS+SVR	INFS+ANFIS+AEX	LS+ANFIS+PSO(10^{-12})	LS+ANFIS+GA(10^{-12})*
1998	152.14	141.56	147.59	117.28	121.18	1.820	1.170
1999	190.11	112.99	115.58	118.91	112.11	2.520	1.450
2000	353.00	175.63	180.03	148.52	132.19	0.848	0.426
2001	165.31	134.39	133.59	110.76	113.23	1.520	0.584
2002	139.64	91.43	81.11	70.38	65.82	11.900	0.416
2003	103.96	68.07	77.31	59.22	57.62	0.500	1.100
2004	82.32	72.34	56.44	53.07	54.33	1.820	1.400
2005	86.12	62.52	55.97	54.72	54.81	1.330	1.900
2006	215.64	83.92	77.03	63.22	56.31	5.320	0.888

*Better model

Table 11: The results comparison for TAIEX dataset based on Theil's U statistic (Su and Cheng, 2016)

Test set	Chen (1996)	Yu (2005)	INFS+ANFIS	INFS+SVR	INFS+ANFIS+AEX	LS+ANFIS+GA(10^{-17})*	LS+ANFIS+PSO (10^{-17})
1998	0.0143	0.0134	0.0160	0.0084	0.0087	8.60	13.33
1999	0.0149	0.0106	0.0074	0.0076	0.0072	9.45	9.78
2000	0.0339	0.0207	0.0170	0.0140	0.0124	4.04	8.05
2001	0.0233	0.0189	0.0139	0.0115	0.0117	6.07	15.80
2002	0.0155	0.0124	0.0087	0.0076	0.0071	4.47	12.80
2003	0.0106	0.0076	0.0065	0.0050	0.0049	2.12	4.23
2004	0.0070	0.0080	0.0048	0.0045	0.0046	2.15	15.30
2005	0.0078	0.0065	0.0045	0.0044	0.0044	2.19	10.70
2006	0.0156	0.0077	0.0052	0.0042	0.0038	5.96	35.70

*Better model

$$\text{Decision} = \begin{cases} \text{Sell} & \text{Forecast}(t+1) - \text{Actual}(t) > 0 \\ \text{Buy} & \text{Forecast}(t+1) - \text{Actual}(t) < 0 \end{cases}$$

Based on the above rules we can decide whether to buy or sell the stocks on a last day of the year for the experimental dataset HSI the buying or selling options are given in Table 7.

CONCLUSION

This study proposed a six two stage hybrid model for forecasting stock market dataset. In this, LS/MCFS/CBS selects important technical indicators and utilized the selected variables as input variables to ANFIS which is trained by evolutionary algorithms GA/PSO to generate initial forecasts. The experimental results shows that the proposed models, LS+ANFIS with GA and LS+ANFIS with PSO effectively increase accuracy of forecasting. These two models has good accuracy compared to listing model and variable selection methods and fuzzy c means clustering decrease computational complexity. The proposed model can produce more reasonable and understandable rules because the "IF-THEN" rules produced by ANFIS can model the qualitative aspects of human knowledge and it 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.

REFERENCES

Aladag, C.H., E. Egrioglu and C. Kadilar, 2009. Forecasting nonlinear time series with a hybrid methodology. Applied Mathe. Lett., 22: 1467-1470.

Azadeh, A., M. Saberi and S.M. Asadzadeh, 2011. An adaptive network based fuzzy inference system-auto regression-analysis of variance algorithm for improvement of oil consumption estimation and policy making: The cases of Canada, United Kingdom and South Korea. Appl. Math. Modell., 35: 581-593.

Chen, S.M., 1996. Forecasting enrollments based on fuzzy time series. Fuzzy Sets Syst., 81: 311-319.

Chen, T.L., C.H. Cheng and H.J. Teoh, 2008. High-order fuzzy time-series based on multi-period adaptation model for forecasting stock markets. Physica A., 387: 876-888.

Cheng, C.H., T.L. Chen and L.Y. Wei, 2010. A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. Inf. Sci., 180: 1610-1629.

Connor, J., L.E. Atlas and D.R. Martin, 1991. Recurrent Networks and NARMA Modeling. Proceedings of the 4th International Conference on Neural Information Processing Systems (NIPS'91), December 2-5, 1991, Morgan Kaufmann Publishers, San Francisco, California, USA., ISBN:1-55860-222-4, pp: 301-308.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica J. Econom. Soc., 50: 987-1007.

Ghore, S. and A. Goswami, 2015. Short term load forecasting of chhattisgarh grid using adaptive neuro fuzzy inference system. Intl. J. Sci. Res., 4: 841-846.

Jilani, T.A. and S.M.A. Burney, 2008. A refined fuzzy time series model for stock market forecasting. Physica A, 387: 2857-2862.

Jingtao, Y., C.L. Tan and H.L. Poh, 1999. Neural networks for technical analysis: A study on KLCI. Int. J. Theor. Applied Finance, 2: 221-241.

Krzysztof, M. and K. Halina, 2006. Correlation-based feature selection strategy in classification problems. Int. J. Applied Math. Comput. Sci., 16: 503-511.

Laboissiere, L.A., R.A. Fernandes and G.G. Lage, 2015. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. Appl. Soft Comput., 35: 66-74.

Lee, W.J. and J. Hong, 2015. A hybrid dynamic and fuzzy time series model for mid-term power load forecasting. Intl. J. Electr. Power Energy Syst., 64: 1057-1062.

Li, Y., W. Zhang, Q. Xiong, D. Luo and G. Mei *et al*., 2017. A rolling bearing fault diagnosis strategy based on improved multiscale permutation entropy and least squares SVM. J. Mech. Sci. Technol., 31: 2711-2722.

Majhi, B. and C.M. Anish, 2015. Multiobjective optimization based adaptive models with fuzzy decision making for stock market forecasting. Neurocomputing, 167: 502-511.

Mohamad, M.A., D. Nasien, H. Hassan and H. Haron, 2015. A review on feature extraction and feature selection for handwritten character recognition. Intl. J. Adv. Comput. Sci. Appl., 6: 204-212.

Ozkan, G. and M. Inal, 2014. Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems. Appl. Soft Comput., 24: 232-238.

Ravi, V., D. Pradeepkumar and K. Deb, 2017. Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms. Swarm Evol. Comput., 36: 136-149.

Roh, T.H., 2007. Forecasting the volatility of stock price index. J. Expert Syst. Appl., 33: 916-922.

Su, C.H. and C.H. Cheng, 2016. A hybrid fuzzy time series model based on ANFIS and integrated nonlinear feature selection method for forecasting stock. Neurocomputing, 205: 264-273.

Suganya, R. and R. Shanthi, 2012. Fuzzy C-means algorithm-a review. Intl. J. Sci. Res. Publ., 2: 1-3.

Swasti, R. and S.P. Khuntia, 1994. Simulation Study for Automatic Generation Control of a Multi-Area Power System by ANFIS Approach. Prentice Hall, Englewood Cliffs, New Jersey, USA.,.

Wei, L.Y., 2016. A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting. Appl. Soft Comput., 42: 368-376.

Yu, H.K., 2005. Weighted fuzzy time series models for TAIEX forecasting. Phys. A. Stat. Mech. Appl., 349: 609-624.