# Application of Genetic Algorithm and Neural Network in Forecasting with Good Data

P. MAKVANDI, J.JASSBI, S.KHANMOHAMMADI
System Department
I.A. University, Science & Research campus
Poonak sq. Ashrafi Esfahani Blv., Hesarak, Tehran,
IRAN

*Abstract:* - Selection of effective input variables on decision making or forecasting problems, is one of the most important dilemmas in forecasting and decision making field. Due to research and problem constraints, we can not use all of known variables for forecasting or decision making in real world applications. Thus, in decision making problems or system simulations, we are trying to select important and effective variables as good data. In this paper we use a hybrid model of Genetic Algorithm (GA) and Artificial Neural Network (ANN) to determine and select effective variables on forecasting and decision making process. In this model we have used genetic algorithm to code the combination of effective variables and neural network as a fitness function of genetic algorithm. The introduced model is applied in a case study to determine effective variables on forecasting future dividend of the firms that are members of Tehran stock exchange. This model can be used in different fields such as financial forecasting, market variables prediction, intelligent robots decision making, DSS structures, etc.

*Key words*: - Good data, Artificial Neural Network, Genetic Algorithm, Forecasting, Fitness function

## 1 Introduction

Decision making can be defined as a selection problem. In such problems the forecasting of decision making parameters, which is usually a very complicated and in some cases impossible process, plays an important role in real world applications. For a long time, forecasting has been an interesting topic in different sciences. In practice, as the number of input variables of a model increases and variables relations become more dynamic, the forecasted results can be deviate form actual output.

There are different tools and methods to analysis such dynamisms and complex relations. Artificial Neural Network (ANN) is a powerful and flexible tool to represent such models and its learning capabilities can be used to recognize the different possible input output relations in the model and to predict different decision making parameters.

The earlier methods for forecasting are usually based on statistical models and historical data. By comparing the statistical models and neural network, Faber and Lapedes (1987) concluded that neural network is more powerful than statistical models. One of the interesting fields for application of ANN is the stock exchanges, where the scientist and researches have different investigations emphasizing on comparing ANN with multivariate regression and ARIMA (Wang and Lee 1996) [9]. Another paper on using ANN for forecasting stock index of Koalalampur is presented by Yang (Yang 1996). Kohzadi and Boyd (1997), applied a combination of ANN and time series model to predict commodity prices [6]. In 1998, Karokawa, Ikeda and Nomura, used combination of GA and ANN in a decision making model on stock purchase or selling [5]. The combination of forecasting methods with the aim of reducing prediction error has been the main topic of many researches. Marcello and Alvaro (2000), combined Neural Network and time series methods as a hybrid linear- neural model for time series forecasting, [2].

However when we use a large umber of data as ANN input variables:

1. Training time increases significantly.
2. Training error increases by using inappropriate variables.
3. We spend extra cost for gathering and processing non effective or inappropriate data.

To solve these draw backs, the effective input data with significant impact (good data) should be considered for training. According to these problems and considering Pareto logic that says, 80% of a system treatment comes from just 20% of its factors, we have applied genetic algorithm (GA), as a heuristic tool, to select appropriate combinations of different variables that have more effect on forecasting decision making parameters. This model is applied in a case study to predict future dividend of the firms that are members of Tehran stock exchange.

## 2 Problem Statement

The basic object of this paper is to find the optimal combination of effective variables on system's behaviors.

Figure (1), shows the conceptual framework of a forecasting system where X is the vector of input variables (independent variables) and Y= f(X) is the vector of output variables (dependent variables) forecasted by f, which is modeled by ANN.

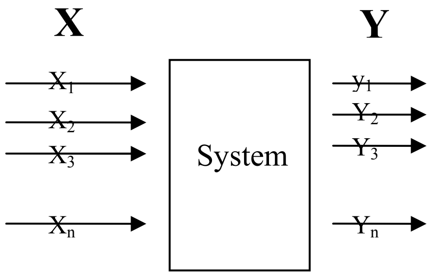


Figure 1. Conceptual framework of a prediction system

The advantage of using GA for selecting the appropriate input variables, instead of using statistical methods, is that in the same time that the number of input variables is decreased the power of forecasting is maintained; because by using the GA the selection of good data is based on considering the correlation between them.

## 3 Methodology

Some of researches in forecasting models have shown that by combination of existing methods, called Hybrid model, the prediction error may be reduced. The model presented in this paper is categorized as a hybrid models consisting artificial neural network (ANN) and genetic algorithm (GA) with necessary modifications.

### 3.1 Model framework

Figure 2 shows the hybrid framework of the model consisting GA and ANN. Input data enter to GA that has neural network as a fitness function inside. This algorithm finds the appropriate combination of input variables. The selected Input variables are used to train ANN.

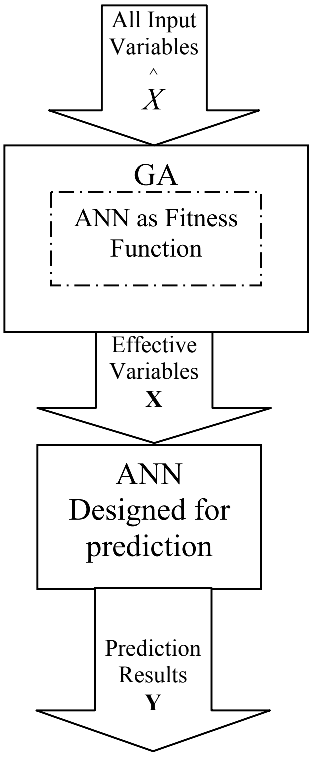


Figure 2. Hybrid Framework of GA and ANN

### 3.2 Genetic Algorithm (GA)

Since reducing the number of inputs to the prediction model may improve the speed of prediction, it is prudent to select only those features that are significant to the particular application. As a by product, reducing the number of input features increase the computation efficiency.

*Genetic algorithm* (GA) is used as a feature selection tool because of its advantages consisting [4]:

- Global optimization
- Suitable for discrete search space
- Efficient search strategy

GA has demonstrated substantial improvement over a variety of random and local search methods. The GA is based on the laws of natural selection in genetics. The principal idea is to search for optimal solution in a large population. It uses a fixed length binary string called a chromosome to represent a possible solution or individual for a given problem domain. Usually a simple GA consists of three operations:

- *Selection*
- *Crossover*
- *Mutation*

The population comprises a group of chromosomes from which candidates can be selected as a solution of a given problem. The initial population (set of possible solutions) consists of member solutions. Each one has a fixed member of randomly selected features from a feature pool. The fitness (a measure of appropriateness of the solution of each chromosome) is evaluated by fitness function.

Here, the fitness function is an artificial neural network (ANN). The selection operator selects chromosomes in the population based on fitness. Individuals with higher fitness have more chance for being selected as parents. A couple of selected chromosomes or parents are then crossovered and their information is exchanged to generate new chromosomes or offsprings.

Mutation fillips the value of all or some bits of randomly selected chromosome. This operation increases exploration of search space and prevents converging to a local optimum point. The process runs cyclically until a stopping criterion is reached. Each cycle is called a generation.

The key assumption of parent selection is to give preference to fitter individuals. Here we use the Roulette Wheel selection scheme [4]:

1. Computing cumulative fitness of individuals.
2. Generating a random number n between 0 and total fitness.
3. Returning the first individual that its cumulative fitness is greater than or equal to n.

In this paper the Uniform crossover scheme is used. That is, each bit position 1 to L, is randomly picked from either of the two parents' strings. This means that each bit is inherited independently from other bit and that there is, in fact, no linkage between bits. After crossover and mutation, there are two groups of candidates: the parents and the offsprings. Each of these groups contributes a fraction of new generation. We have used the elitist selection strategy to transfer a copy of a few fittest individuals in present generation to next generation to ensure crucial convergence of algorithm [8].While any successful application of genetic algorithm for solving a problem is greatly dependent on finding a suitable method for encoding the possible solutions to chromosomes, the creation for the appropriate fitness function is important to have a successful training process.

### 3.3 Artificial Neural Network

In this work the Multilayer Perceptron (MLP) [7] as well as GA fitness function is used for forecasting. The MLP consists of one hidden layer with Tan-Sigmoid activation functions of neurons. Training of the network is carried out by using a standard back propagation (BP) algorithm. In this paper mean square error (MSE) is used as error function [1]:

$$MSE = \sqrt{\frac{\sum_{i=1}^{n}(A_i - F_i)^2}{n}} \qquad (1)$$

Where $A_i$ is the Actual Value, $F_i$ is the Forecasted Value and n is the Number of Validation Data.

### 3.4 Case Study

The selection of effective combination of variables on future dividend (DPS) of 194 firms, that are members of Tehran stock exchange, is considered as a case study. Twenty four initial variables are derived by literature review and experimental suggestion of financial statements as independent variables that affect the future dividend (DPS), as the single dependent variable of the model.

For each variable, 1194 observations have been collected during 6 years. Thousands of observations are used for ANN training and 194 of them are used for validation of training.

Ten of these input variables are selected as effective data by using the introduced algorithm as shown in figure 3.

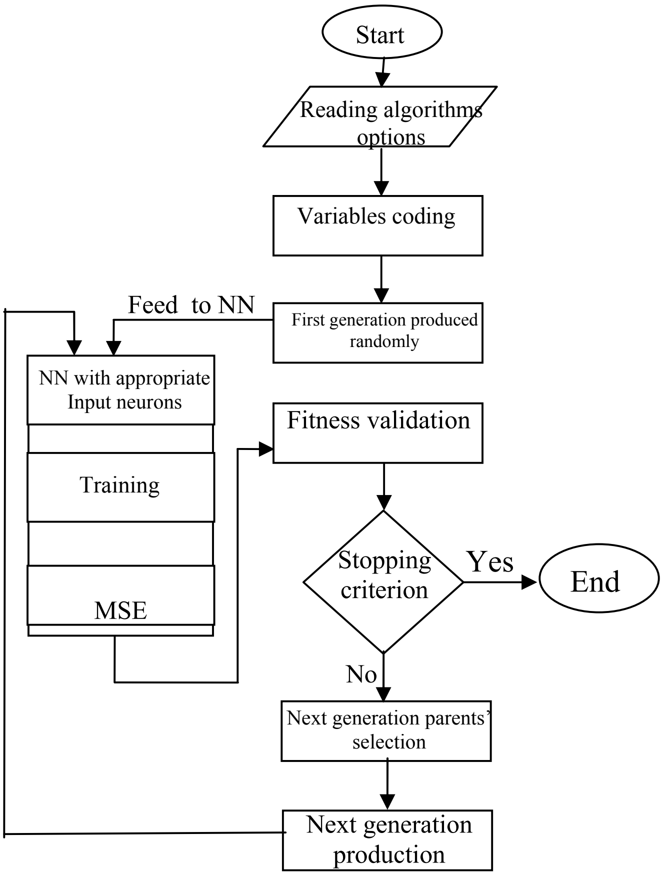

Start
Reading algorithms options
Variables coding
First generation produced randomly
Feed to NN
NN with appropriate Input neurons
Training
MSE
Fitness validation
Stopping criterion
Yes
End
No
Next generation parents' selection
Next generation production

Figure 3. Implementation Flowchart

### 3.5 Implementation and Results

The model variables are coded in a binary string with 29 bits. The first 24 bits of binary string are used to identify the variables and the other 5 bits are used to determine the number of hidden layer neurons of fitness function neural network. For any input variable represented by the 24 first bits, the flag 1 means that corresponding variable must be put in combination and 0 means that it must be removed from the combination.

The data are normalized in the interval [-1, 1] by using:

PN=2*(p-Minp)/(Maxp-Minp)-1; (2)

where P is the matrix of input (column) vectors, PN is the matrix of normalized input vectors, Minp is a vector containing the minimum value for each p and Maxp is a vector containing maximum value for each p [3].

The GA parameters are set as:

Table 1. GA Key Options

| GA Parameters | Rate/Value |
|---|---|
| Number of generations' members | 50 |
| Mating probability | 0.8 |
| Mutation probability | 0.01 |
| Mating kind | Scatterd |
| Selection Function | Reminder, elite count =2 |

Figure 4 shows the best fitness of each generation and figure 5 shows the mean values of the fitness for each generation during the algorithm implementation.

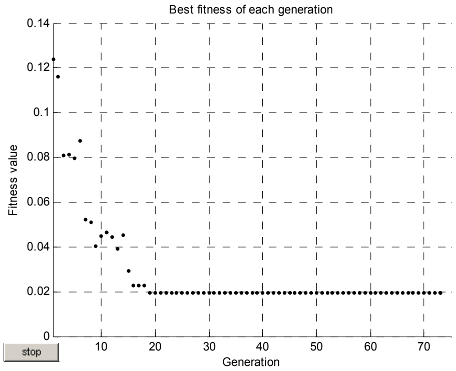


Figure 4. Best Fitness of Each Generation

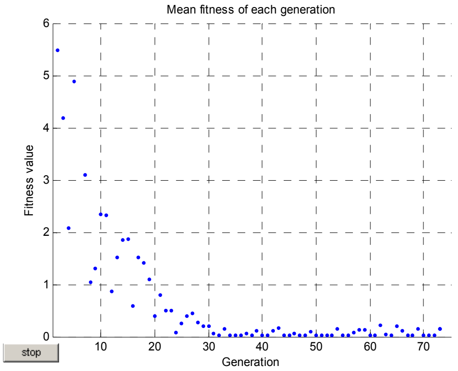


Figure 5. Mean Values of Fitness for Each Generation

The optimal combination is achieved as:

Ans=[0,1,1,0,0,1,0,1,1,0,1,1,0,1,0,0,0,1,0,0,0,1,0,0] ;

It means that 10 variables are selected by GA.

After determination of final optimal combination (the combination of effective variables on DPS that are selected by GA), an ANN proportional to these 10 selected variables (figure 6) is designed and trained by these 10 selected variables.

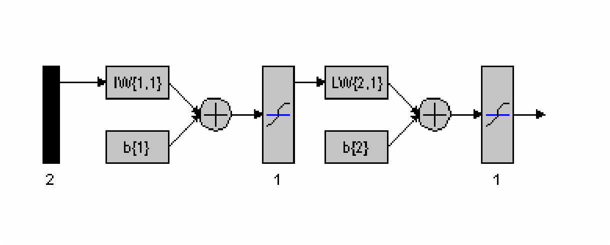


Figure (6) - ANN Proportional to 10 Selected Variables

Figure 7 shows the training result of this network,

In witch MSE = 0.0043.

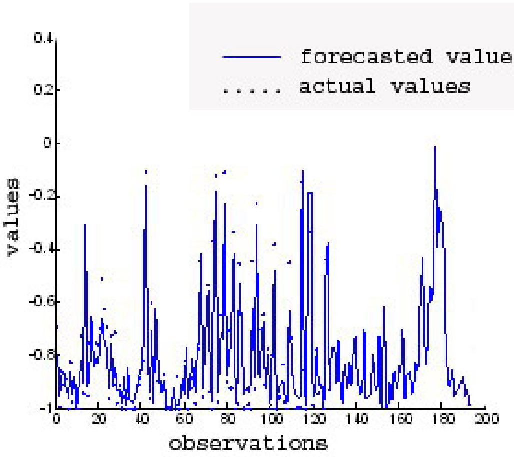
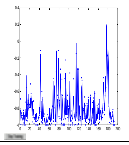
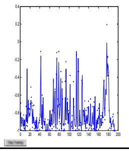
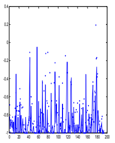


Figure 7. ANN Trained with 10 Selected Variables

In figure7, Horizontal axis belongs to 194 observations that are used to validate the ANN training and the numbers on vertical axis are the values of corresponding observations of output variable. Dots, in this chart, represent actual outputs and the curve is recognized pattern that obtained by ANN. As it shown, the appropriate ANN has detected he output pattern with nearly good estimation. The results obtained by three different methods are presented and compared in table 2.where SSE and MSE represent the Sum of Square Error and Mean Square Error , respectively.

Table 2 –Comparison of different methods

| Methods | SSE | MSE | figures |
|---|---|---|---|
| Training with all 24 variables | 1.0476 | 0.0054 |  |
| Training with 10 variables selected by GA | 0.8342 | 0.0043 |  |
| Training with 8 variables selected by multivariate regression | 9.8552 | 0.0508 |  |

## 4 Conclusions

Forecasting methods in complex systems tie with errors and we always try to reduce these errors. On of the main drawbacks for application of ANN in solving nonlinear problems is it's weakness in learning with large amount of patterns. This drawback is mainly overcome by using this method because of its ability in selecting Good Data. The presented is basically appropriate for the cases where the structure of problems is not predetermined and the use of classic methods that are based on preexisting patterns (theories) is not recommended. Due to stochastic search algorithm of GA and capability of ANN in pattern recognition of complex systems, this model can be used for determination of optimal combination in DSS systems, by reducing the modeling variables in engineering processes, robots decision making problems, etc.

*Reference:*

[1] Alavi H, Ghaffari Saadat M.H., Using Genetic Algorithm in Neural Network Based Feature Selection for Vibration Monitoring of a Gear Train, proceeding of vetomac 03 and ACSIM 2004

[2] Alvaro, V., A hybrid linear- neural model for time series forecasting, IEEE Transactions on Neural Network, V (11), p.p. 1402-1412. 2000

[3] Demuth H., Beale M., Neural Network Toolbox for Use with MATLAB User's Guide, Version 4 , MathWorks, Inc. 2003

[4] Goldberg DE. Genetic algorithm in search, optimization and machine learning. Addison Wesley publishing Co. 1989

[5] Karokawa, T., Ikeda, Y., Nomura, Sh., Hybrid Method of Neural Network and Genetic Algorithm for Stuck Trading, The 3rd. Annual International Conference on Industrial Engineering Theories December 28-31, 1998

[6] Kohzadi, N., Boyd , M., Kermanshahi, B. and Kaastra, I., A Comparison of Artificial Neural Network and Time Series Models for Forecasting Commodity Prices, Neurocomputing, Vol. 15, 1996.

[7] Menhaj MB. Fundamentals of neural networks. Amirkabir Univ. Pub. (in Persian). 2002

[8] Thierens D & Goldberg D. Elitist recombination: An integrated selection recombination GA. IEEE transactions, proceeding of the first IEEE conference on evolutionary computation. 1994

[9] Wang, G.and Leu, J., Stock Market Trend Prediction using ARIMA Based Neural Networks, IEEE, 1996.