International Journal of Control and Automation
Vol.6, No.5 (2013), pp.303-314
http://dx.doi.org/10.14257/ijca.2013.6.6.29

A Genetic Programming Model for S&P 500 Stock Market Prediction

Alaa Sheta*, Hossam Faris†, Mouhammd Alkasassbeh‡

Abstract

The stock market is considered one of the most standard investments due to its high revenues. Stock market investment can be risky due to its unpredictable activities. That is why, there is an urgent need to develop intelligent models to predict the for stock market index to help managing the economic activities. In the literature, several models have been proposed to give either short-term or long-term prediction, but what makes these models supersede the others is the accuracy of their prediction. In this paper, a prediction model for the Standards & Poors 500 (S&P500) index is proposed based Genetic Programming (GP). The experiments and analysis conducted in this research show some unique advantages of using GP over other soft computing techniques in stock market modeling. Such advantages include generating mathematical models, which are simple to evaluate and having powerful variable selection mechanism that identifies significant variables.

1 Introduction

Prediction in times series data are clearly beneficial in many business areas like bankruptcy prediction, stock market trends, business failure prediction, stock market indexes and other areas [1–4]. Among all these fields, predicting stock market indices has been a hot topic of research and extensively investigated in the last two decades. Strong motivations for demanding the prediction of stock market prices exists [5–7]. With the growing investments and trading sizes, people urgently searched for an intelligent tool which helps to increase their gains and minimizing their risks [8]. Researchers focused on developing stock market prediction models due to its wide range of financial applications and commercial advantages which lead to maximization in the financial profit.

On the other hand, stock market is dynamic, complex and non-linear [9] because it is highly affected by a number of macro-economic and sometimes uncontrollable factors such as general economic conditions, monetary policies, bank's interest rates and political situations [10]. Therefore, researchers were motivated to adopt different approaches in order to develop accurate prediction models for stock market.

Predicting stock market based time-series models, using past measurements to provide an estimate of future measurements, have been explored in many articles [7, 11, 12]. It is always required

*A. F. Sheta is a Professor with the Computers and Systems Department, Electronics Research Institute (ERI), Cairo, Egypt. (asheta66@gmail.com)

†H. Faris is an Assistant Professor at Business Information Technology Department, The University of Jordan, Amman, Jordan. (hossam.faris@ju.edu.jo)

‡M. Alkasassbeh is an Assistant Professor with the Computer Science Department, Mutah University, Jordan (malkasasbeh@gmail.com)

ISSN: 2005-4297 IJCA

to build a model that has a recurrence relation derived from past measurements. The recurrence relation is then used to provide a near accurate new measurement. These measurements are expected to be good enough compared to the actual measurements.

In general, approaches used by researchers can be classified into two main classes:

- The first one is the econometric models which are statistical based approaches such as Linear Regression, Autoregression and Autoregression Moving Average (ARMA) [13–15]. However, models like ARIMA are developed based on nonrealistic assumptions like that; the financial time-series data are linear and stationary. Such nonrealistic assumptions can degrade the quality of predictions [16].
- The second type of research applies soft computing techniques for forecasting market indices. Soft computing is a term which covers artificial intelligence approaches that resemble biological processes in solving complex and nonlinear problems. Examples of such techniques are Artificial Neural Networks (ANN) [17], Fuzzy logic (FL) [18], Support Vector Machines (SVM) [19] and the particle swarm optimization (PSO) [10].

In this paper, we develop stock market prediction model generated by Genetic programming (GP) and point out some unique advantages of using GP in stock market modeling and compare it with other approaches like fuzzy based models and also Linear models. The paper is organized as follows. In section 2, we introduce the proposed model structure in this study. A brief review for the traditional modeling approach based on linear, fuzzy and GP modeling techniques are presented in sections 3, 4, 5. Section 6 we show the stock market data set adopted in this study, the experimental setup and results for the stock market prediction problem.

2 Proposed Model Structure

Our goal is to build a GP model structure which has multiple inputs and single output. Consider a dynamical system with $x_1(k)$, $x_2(k)$, ..., $x_n(k)$ are the input variables and $y(k)$ as the output variable, respectively. k is the time sample. The relationship between these variables can be represented as:

$$y(k) \quad = \quad f(x_1(k), x_2(k), \ldots, x_n(k)) \tag{1}$$

Our objective is to find the values of the model output $y(k)$ as a function of past outputs. Models generated by the evolutionary cycle of genetic programming can be used to approximate this relationship function f.

In order to check the performance of the developed stock market predication model, the Variance-Accounted-For (VAF) and the Root Mean Squares Error (RMSE) were measured. These performance criteria are assessed to measure how close the measured values to the values developed using the genetic programming approach. VAF and RMSE are computed as follows:

$$VAF = \left[1 - \frac{var(y - \hat{y})}{var(y)}\right] \times 100\% \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(T(i) - T_r(i))^2}{n}} \tag{3}$$

where y is real actual value, $\hat{y}$ it the estimated target value. n is the total number of measurements.

3 Linear Regression Model

A linear regression model simply has the following mathematical representation.

$$y = a_0 + \sum_{i=0}^{n} a_i x_i \tag{4}$$

where x_i represents the model input variables and y is the model output variable which is the next week's S&P 500 closing price. a_0 and $a_i, i = 1, 2, \ldots, n$ are the model parameters which need to be estimated. To show how the parameter estimation process work, we assume we have a system with four input variables $x_1(t), x_2(t), x_3(t), y(t)$ and single output y as in the case under study. Thus, the model mathematical equation can be represented as:

$$\begin{aligned} y &= f(x) \\ &= a_0 + a_1 x_1(t) + a_2 x_2(t) + a_3 x_3(t) + a_4 y(t) \end{aligned} \tag{5}$$

To find the values of the model parameters a's we need to build what is called the regression matrix ϕ. This matrix is developed based on the experiment collected measurements. Thus, ϕ can be presented as follows given there is a set of measurements m:

$$X = \begin{pmatrix} x_1^1(t) & x_2^1(t) & x_3^1(t) & y^1(t) \\ x_1^2(t) & x_2^2(t) & x_3^2(t) & y^2(t) \\ \vdots & \vdots & \vdots & \vdots \\ x_1^m(t) & x_2^m(t) & x_3^m(t) & y^m(t) \end{pmatrix}$$

The parameter vector θ and the output vector y can be presented as follows:

$$\theta = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}$$

where:

$$y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{pmatrix}$$

The least squares solution of yields the normal equation:

$$\phi^T \theta = y \tag{6}$$

which has a solution:

$$\theta = \phi^{-1} y \tag{7}$$

But since, the regression matrix ϕ is not a symmetric matrix, we have to reformulate the equation such that the solution for the parameter vector θ is as follows:

$$\theta = (\phi^T \phi)^{-1} \phi^T y \tag{8}$$

4 Fuzzy Modeling

Fuzzy logic has been successfully used to solve a variety of complex problems in system identification. A fuzzy model structure can be represented by a small set of fuzzy IF-THEN rules that describe local input-output functions of a nonlinear system [20]. A rule-based fuzzy model requires the identification of the following quantities:

- the antecedent,
- the consequent structure of the membership functions,
- the estimation of the consequent regression parameters and
- in [21], additional parameters was selected, which is the number of rules (clusters) σ. This parameter is specified by the user.

The above quantities have to be defined in various operating regions. The number of rules used to solve the nonlinear modeling problem can be determined automatically. In [21], author used a variation of Fuzzy logic approach called Takagi-Sugeno for predicting the next week S&P 500 for stock market. This approach is a universal approximator of any smooth nonlinear system was proposed by Takagi and Sugeno [22]. Author investigated the use of fuzzy logic based on a set of clusters to see if the prediction capabilities can be improved using number of linear models. Three clusters where used to split the input data into three individual models given by set of rules.

5 Genetic Programming

Briefly, GP was evolved by J. Koza [23–25] at Stanford in 1991. GP is part of the famous evolutionary computation techniques [26] which provide a methodology for the computer to solve wide domain of problems automatically. Moreover, GP can give an insight on the dynamics of the underlying system by generating easy-to-evaluate models in contrast to other soft computing techniques like neural networks and support vector machines.

GP is an evolutionary approach that automatically generates and evolves computer programs in forms of mathematical models [23, 27]. Each of these models can be represented as a tree or as LISP expression. GP evolutionary cycle can be summarized in the following points:

- **Initialization**: GP starts by generating randomly a number of individuals (models) which form the initial population.
- **Fitness evaluation**: each individual is evaluated according to a specific measurement. In this research, squared Pearson's correlation coefficient evaluation is used.
- **Reproduction**: in this process, a new population is created by applying the following three operations:
  - Selection mechanism: the mechanism used for selecting two individuals (called parents) for reproduction. Usually the selection is based on fitness value of the parents.
  - Crossover: It Creates two new individuals by exchanging and recombining randomly chosen subtrees from selected parents as shown in Figure 1.
  - Mutation: It creates new individual by replacing randomly chosen sub tree of an individual by another randomly generated sub tree. An example of the mutation operation is shown in Figure 2.

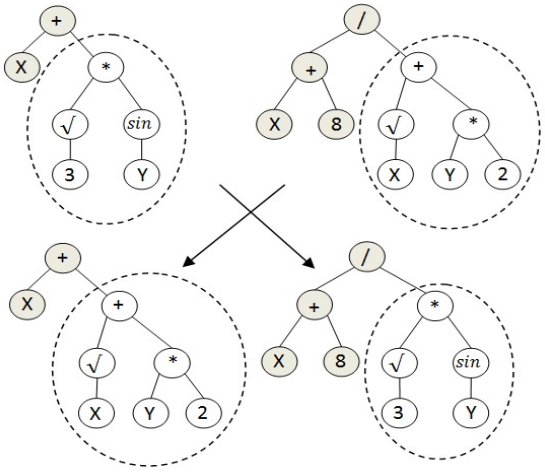

Figure 1. Example of the crossover operator for GP

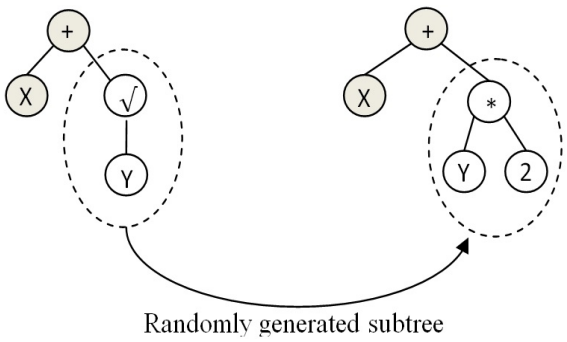

Figure 2. Example of the mutation operator for GP

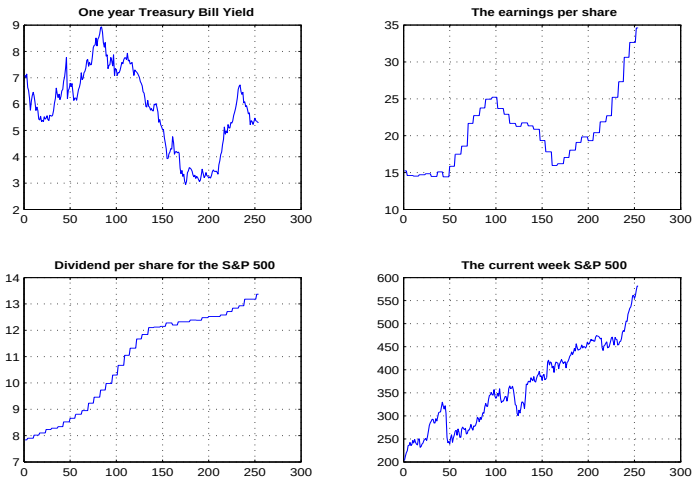

Figure 3. The next week S&P 500 Input Training data set

- **Termination of the cycle**: the evolutionary cycle starting from the fitness evaluation point keeps iterating until an individual with certain fitness value is found or maximum number of iterations is reached.

6 Experiments Results

The $S\&P$ 500 stock market data set used in our case was provided in [28]. The data set consists of 507 weeks of data, which cover an approximately ten-year period. The $S\&P$ 500 data were presented and sampled such that data for weeks 1, 3, 5,.., 505, 507 is used for building a genetic programming model. The data for weeks 2, 4, 6,.., 504, 506 is used for testing the developed model (i.e. validation). Figures 3 and 4 show the training and testing data sets. The proposed model has the following Inputs:

- the 1 year Treasury Bill Yield as $x_1(t)$
- the earnings per share as $x_2(t)$
- dividend per share for the $S\&P$ 500 as $x_3(t)$
- the current week's $S\&P$ 500 as $y(t)$

The proposed model output is:

- the next week's $S\&P$ 500 as $y(t+1)$

The data set was loaded into HeuristicLab framework then a symbolic regression via GP was applied with parameters set as shown in Table 1. HueristicLab framework[1] was used to apply GP in the experiments designed in this research [29].

The cross validation was tuned to 50% for training and 50% for testing. After a run of 1000 generations GP converged to the best model shown in Figure 5. The GP best individual obtained was able to model the S&P 500 Stock Market Data with a VAF value of 99.36%, while it was capable of predicting for the testing part with a VAF value of 99.58%.

[1]HeuristicLab is a framework for heuristic and evolutionary algorithms that is developed by members of the Heuristic and Evolutionary Algorithms Laboratory (HEAL), http://dev.heuristiclab.com

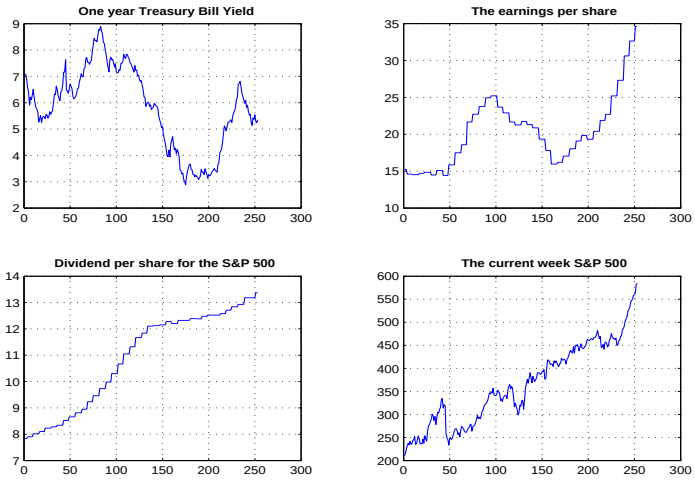

Figure 4. The next week S&P 500 Input Testing data set

Table 1. GP parameters

Parameter	Value
Mutation probability	15%
Population size	1000
Maximum generations	10000
Selection mechanism	Tournament selector
Elites	1
Operators	{+,-,*,/}

$$y(t+1) = \left(c_0 \cdot y(t) + c_1 \cdot x_2(t) + c_2 \cdot x_1(t) + \frac{c_3}{(c_4 \cdot x_2(t) + c_5 \cdot x_1(t))} + \frac{c_6 \cdot x_2(t)}{(c_7 \cdot x_2(t) + x_1(t) \cdot (c_8 \cdot x_1(t) + c_9 \cdot x_2(t)) \cdot c_{10})} + c_{11} \right)$$

$Where:$

$c_0 = 0.97386,\ c_1 = 0.45609,\ c_2 = -1.3669,\ c_3 = 1.0,\ c_4 = 1.7715$

$c_5 = -5.4218\ c_6 = 0.95483\ c_7 = 1.544\ c_8 = 1.8697\ c_9 = -0.42254$

$c_{10} = -0.34455\ c_{11} = 8.2368$

Figure 5. Developed GP model

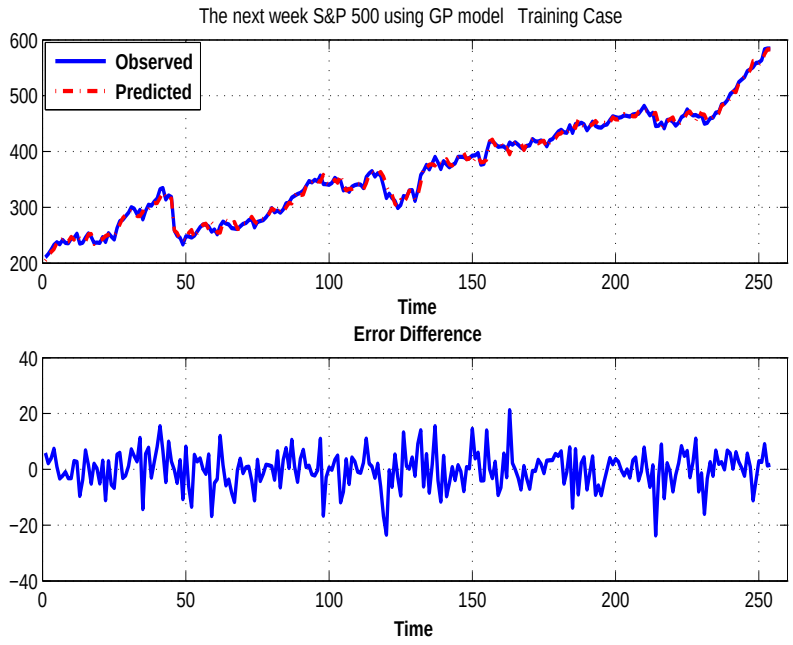

Figure 6. Observed and predicted next week S&P 500 in the Training Case

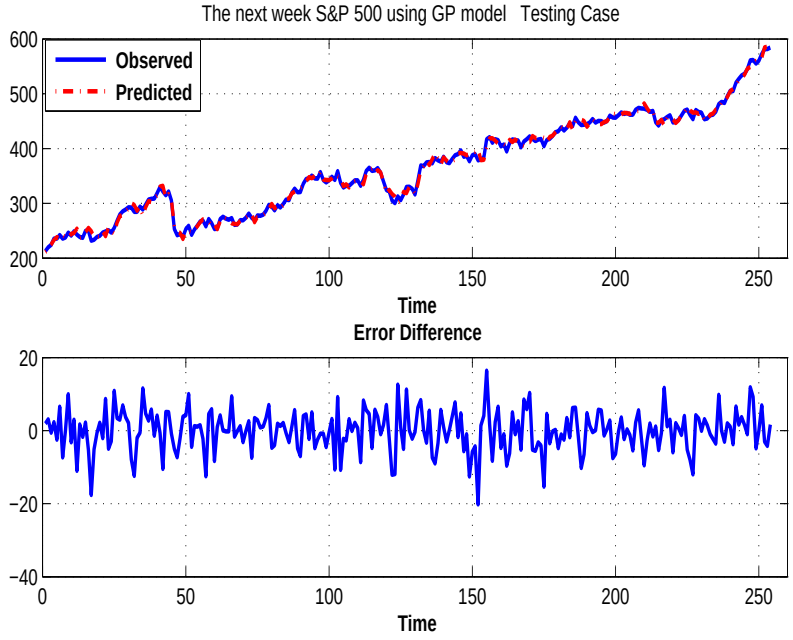

Figure 7. Observed and predicted next week S&P 500 in the Testing Case

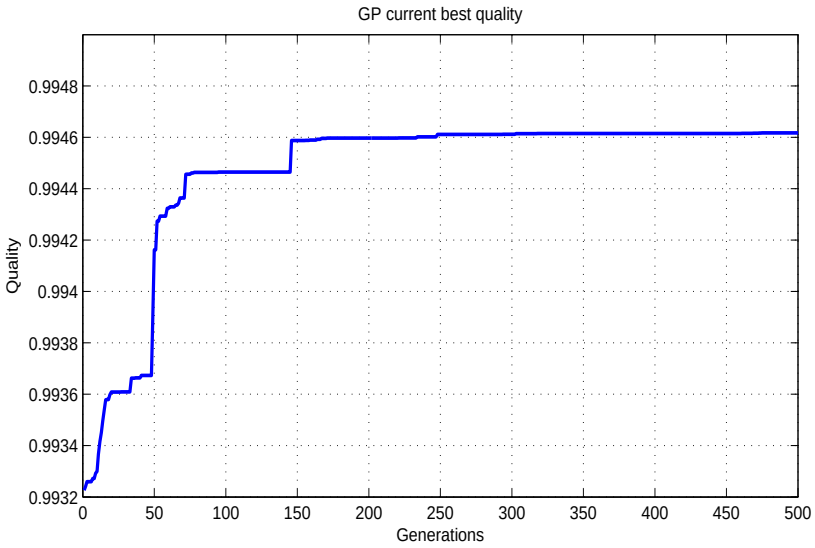

Figure 8. VAF Best so far convergence of the GP evolutionary process

Table 2. A comparison between the three developed model

	GP		Linear Regression		Fuzzy [21]	
	Training	Testing	Training	Testing	Training	Testing
VAF	99.462%	99.605%	99.3620%	99.5865%	99.3977%	99.5682%
RMSE	6.5433	5.6106	7.1280	5.6850	6.9256	5.8096

The predicted values based GP model were compared with results obtained by the Multiple Linear Regression (MLR) model and the Fuzzy model developed in [21]. In Table 2, we show the VAF for the GP model along with Fuzzy model reported in the literature [21]. Obviously, GP model shows better performance for the training case and a very competitive accuracy in the training and testing cases. The predicted values of the next week's S&P 500 closing price along with the actual values are shown in Figure 6 and Figure 7 respectively. The best so far convergence of the GP evolutionary process is shown in Figure 8.

7 Conclusions and Future Work

A genetic programming model for the S&P500 index was developed in this paper. The developed GP model provided good estimation and prediction capabilities in both training and testing cases. A comparison between prediction models developed by Fuzzy Logic, Linear Regression and the proposed GP model. Future research shall focus on exploring other advantages of GP, which is the capability of identifying the most important variables in such a dynamic and complex problem.

References

[1] W. Dai, J.-Y. Wu, and C.-J. Lu, "Combining nonlinear independent component analysis and neural network for the prediction of asian stock market indexes," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4444 – 4452, 2012.

[2] M.-Y. Chen, "A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering," *Information Sciences*, vol. 220, no. 0, pp. 180 – 195, 2013.

[3] G. S. Atsalakis and K. P. Valavanis, "Forecasting stock market short-term trends using a neuro-fuzzy based methodology," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10696 – 10707, 2009.

[4] H. Etemadi, A. A. A. Rostamy, and H. F. Dehkordi, "A genetic programming model for bankruptcy prediction: Empirical evidence from iran," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3199 – 3207, 2009.

[5] J. J. Murphy, *Technical Analysis of the future Market*. NYIF, New York, 1986.

[6] A. Colin, "Exchange rate forecasting at citibank london," in *Proceedings of the Neural Networks Computing, London*, 1991.

[7] D. E. Baestanens and W. M. van den Bergh, *Tracking the Amsterdam stock index using neural networks*. In Neural Networks in the Capital Markets, chapter 10. John Wiley and Sons, 1995.

[8] P. Rausch, A. F. Sheta, and A. Ayesh, *Business Intelligence and Performance Management: Theory, Systems and Industrial Applications*. Springer Publishing Company, Incorporated, 2013.

[9] Y. Zhang and L. Wu, "Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network," *Expert Systems with Applications*, vol. 36, no. 5, pp. 8849 – 8854, 2009.

[10] R. Majhi, G. Panda, G. Sahoo, A. Panda, and A. Choubey, "Prediction of S&P 500 and DJIA stock indices using particle swarm optimization technique," in *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*, pp. 1276–1282, IEEE, 2008.

[11] K. Bergerson and D. Wunsch, *A commodity trading model based on a neural network-expert system hybrid*. In Neural Networks in finance and investing, chapter 23. Probus Publishing Company, 1993.

[12] R. J. Van Eyden, *The Application of Neural Networks in the Forecasting of Share Prices*. Finance and Technology Publishing, 1996.

[13] Y. B. Wijaya, S. Kom, and T. A. Napitupulu, "Stock price prediction: Comparison of ARIMA and artificial neural network methods - an indonesia stock's case," in *Proceedings of the 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, ACT '10, (Washington, DC, USA), pp. 176–179, IEEE Computer Society, 2010.

[14] A. C. Harvey and P. H. J. Todd, "Forecasting Economic Time Series with Structural and Box-Jenkins Models: A Case Study," *Journal of Business & Economic Statistics*, vol. 1, no. 4, pp. 299–307, 1983.

[15] C. Granger and P. Newbold, *Forecasting economic time series*. Economic theory and mathematical economics, New York: Academic Press, 1977.

[16] J. Kamruzzaman and R. A. Sarker, "ANN-based forecasting of foreign currency exchange rates," in *Neural Information Processing*, pp. 49–58, 2004.

[17] C. Chen, "Neural networks for financial market prediction," in *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, vol. 2, pp. 1199–1202, IEEE, 1994.

[18] M. R. Hassan, "A combination of hidden markov model and fuzzy model for stock market forecasting," *Neurocomputing*, vol. 72, no. 1618, pp. 3439 – 3446, 2009.

[19] W. Huang, Y. Nakamori, and S. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, no. 10, pp. 2513–2522, 2005.

[20] B. Kosko, *eural networks and fuzzy systems: a dynamical systems approach to machine intelligence*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1992.

[21] A. Sheta, "Software effort estimation and stock market prediction using takagi-sugeno fuzzy models," in *Proceedings of the 2006 IEEE Fuzzy Logic Conference, Sheraton, Vancouver Wall Centre, Vancouver, BC, Canada, July 16-21*, pp. 579–586, 2006.

[22] S. Jinju, W. Minxiang, and W. Weidong, "Robust takagi-sugeno fuzzy control for a mini aviation engine," *2008 27th Chinese Control Conference*, pp. 775–780, 2008.

[23] J. Koza, "Evolving a computer program to generate random numbers using the genetic programming paradigm," in *Proceedings of the Fourth International Conference on Genetic Algorithms*, Morgan Kaufmann, La Jolla,CA, 1991.

[24] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. The MIT Press, 1992.

[25] J. R. Koza, *Genetic Programming II*. MIT Press, 1994.

[26] K. De Jong, *Evolutionary Computation: A Unified Approach*. MIT Press, 2006.

[27] J. R. Koza, *Genetic Programming*. MIT Press, 1982.

[28] Mathworks, *NeuroSolution for Matlab, http://www.nd.com/nsml*. 1990.

[29] S. Winkler, M. Affenzeller, and S. Wagner, "New methods for the identification of nonlinear model structures based upon genetic programming techniques," in *Proceedings of the 15th International Conference on Systems Science*, pp. 386–393, 2004.