Short-Term Time Series Modelling Forecasting Using Genetic Algorithm

Haviluddin and Rayner Alfred

Abstract The prediction analysis of a network traffic time series dataset in order to obtain a reliable forecast is a very important task to any organizations. A time series data can be defined as an ordered sequence of values of a variable at equally spaced time intervals. By analyzing these time series data, one will be able to obtain an understanding of the underlying forces and structure that produced the observed data and apply this knowledge in modelling for forecasting and monitoring. The techniques used to analyze time series data can be categorized into statistical and machine learning techniques. It is easy to apply a statistical technique [e.g., Autoregessive Integrated Moving Average (ARIMA)] in order to analyze time series data. However, applying a genetic algorithm in learning a time series dataset is not an easy and straightforward task. This paper outlines and presents the development of genetic algorithms (GA) that are used for analyzing and predicting short-term network traffic datasets. In this development, the mean squared error (MSE) is taken and computed as the fitness function of the proposed GA based prediction task. The results obtained will be compared with the performance of one of the statistical techniques called ARIMA. This paper is concluded by recommending some future works that can be applied in order to improve the prediction accuracy. AQ1

Keywords Time series · Network traffic · Forecasting · Genetic algorithm · Mean squared error (MSE)

Haviluddin (✉)
Department of Informatics, Faculty of Computer Science and Information Technology, Universitas Mulawarman
e-mail: haviluddin@unmul.ac.id

R. Alfred
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88999 Kota Kinabalu, Sabah, Malaysia
e-mail: ralfred@ums.edu.my

J. H. Abawajy et al. (eds.), *Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015)*, Lecture Notes in Electrical Engineering 520,
https://doi.org/10.1007/978-981-13-1799-6_18

327261_1_En_18_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:24/12/2018 Pages: 10 Layout: T1-Standard

1 Introduction

Time Series Analysis is used for many applications such as Economic Forecasting, Sales Forecasting, Budgetary Analysis, Stock Market Analysis, Yield Projections, Process and Quality Control, Inventory Studies, Workload Projections, Utility Studies, Census Analysis, Network Monitoring and Analysis and many more. Network monitoring is not an easy task and it is a demanding task that is a vital part of a Network Administrators job. Network Managers and Administrators are constantly striving to ensure smooth operation of their networks. In any universities, if a network were to be down even for a small period of time, the teaching and research productivity within these universities would decline and the ability to provide essential learning and teaching services would be compromised. Network Managers need to monitor traffic movement and performance throughout the network in order to maintain smooth operation of their networks. One of the issues that network managers should pay attention to is the bandwidth usage. Network monitoring and analysis on the bandwidth usage can be performed by using a traffic management system tool. This is important in order to avoid any network congestions in the network due to the density of traffic. The traffic management system has the ability to manage the network by setting variables of network elements, so that it presents the optimum use of real-time bandwidth data during the network data communication process [1, 2]. These network traffic datasets are non-linear time series datasets which can be analyzed and predicted to determine the amount of usage on a daily, weekly, monthly and even yearly. There are many related works conducted to perform the analysis and prediction of these type of time series datasets in order to obtain a good forecast accuracy that includes weather, rainfall, temperature, wind speed forecasting [3–5], financial; stock market, stock price [6–8], tourist demand, tourist quantity [9, 10] and engineering, network traffic, internet traffic [1, 2, 11–15].

There is an increasing interests in developing more advanced forecasting techniques in learning time series datasets (e.g., network traffic) as it will provide more information to the University's network manager for better decision making results. A Genetic Algorithm (GA) method is one of the machine learning techniques that is capable in solving the problem of forecasting a non-linear time series dataset [16–18]. As a result, the main objective of the paper is to outline and evaluate a genetic algorithm (GA) based prediction algorithm that is developed to model time series datasets. The ICT Universitas Mulawarman statistical data of the daily inbound outbound network traffic recorded for five days will be used as the main datasets. A step-by-step processes involved in the proposed genetic algorithm will be described clearly and the mean squared error (MSE) is taken and computed as the fitness function of the proposed GA based prediction algorithm. The rest of this paper is structured as follows. Section 2 describes the proposed genetic algorithm approach, including both time series models. The dataset is described in Sect. 3. In Sect. 4, the results of the forecasting are discussed. Finally, this paper is concluded in Sect. 5.

327261_1_En_18_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:24/12/2018 Pages: 10 Layout: T1-Standard

2 Methodology

2.1 *The Principle of Genetic Algorithm*

The basic concept of GA is found at the University of Michigan, United States of America by John Holland in 1975 as outlined in a book entitled "Adaptation in Natural and Artificial Systems". Then, it was popularized by one of his students, David Goldberg in the 1980s. GA is an algorithm that seeks to apply an understanding of the natural evolution of problem-solving tasks. The approach taken by this algorithm is to randomly combine a wide selection of the best solutions in a set to get the next generation of the best solution based on a condition that maximizes compatibility called fitness. Then, this generation will represent improvements on the initial population [7, 16, 17].

Based on this concept, a GA can be described as a computational abstraction of biological evolution that has worked with a population of possible solutions. A chromosome is normally used to represent the problem-solutions. The initial population that consists of a set of chromosomes is normally generated randomly. Each chromosome will go through an evaluation process using a measure called the fitness function in which this fitness value of a chromosome will show the quality of the chromosomes in the population. Then, the next population, which is also known as offspring, is generated from the process of evolution of chromosomes through iterations called generations. A new chromosome is formed by combining a pair of chromosomes through the crossover and mutation processes [18–21].

2.2 *The Genetic Algorithm Cycle*

In general, the implementation of the GA will go through a simple cycle consisting of four stages that include (1) Constructing a population consisting of several strings of chromosome called initialized population, (2) Evaluation of each string of chromosome value called using predefined fitness function, (3) Performing the selection process to get the best string of chromosome called individual selection, and (4) Genetic manipulation in order to create a new population of chromosomes called reproduction [18, 22]. Figure 1 illustrates the cycle of the GA implementation.

The GA method that will be used to solve the problem of forecasting a non-linear time series dataset is as follows [22];

Step 1 Encoding schemes: Coding genes on *chromosome* using Real Number *Encoding* (RNE) and each chromosome represents a possible solution.

Step 2 Generating Initial Population: Value of genes in each *chromosome* is generated randomly. The size of the population depends on the problem to be solved and the type of genetic operators that will be implemented.

327261_1_En_18_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:24/12/2018 Pages: 10 Layout: T1-Standard

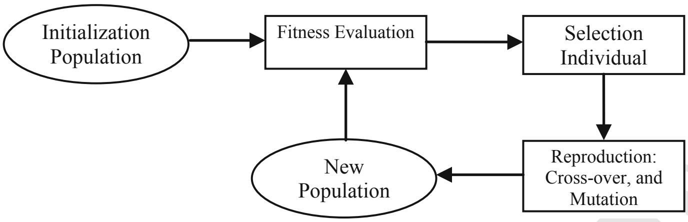


Fig. 1 The Genetic Algorithm Cycle

Step 3 Evaluation function: Individual chromosome is evaluated based on a predefined function because the value of fitness will greatly affect the performance of genetic algorithms.

Step 4 Selection: using the method *roulette-wheel*, *random* and *tournament*.

Step 5 Forming a New Generation: A new generation is formed by using two operators; namely crossover and mutation. The crossover is done by using a *one-point crossover*. Then, the mutation process is carried out by using the *uniform multi point mutation* criteria that is choosing a gene that will be modified based on the probability of mutation.

Step 6 Go to Step 3. This continues until the stopping criteria are met.

2.3 Time Series Data

A time series data can be described as a period course of action model that illuminates a variable regarding its own past and a spasmodic exacerbation term [9, 18]. In principle, a time series model is used to predict the current value of data, X_t, based on the data ($X_{t-n}, \ldots, X_{t-2}, X_{t-1}$), where *n* is the number of past observation and t is the current time of observation made. Time series models have been widely used for forecasting in the past four decades, with the dominance of Artificial Neural Network models. In this work, the time series data that has been taken by the software CACTI, which is one of the open source software in network management protocol will be fed into the proposed GA based prediction algorithm. Table 1 shows the inbound and outbound of the network traffic real data obtained from the Universitas Mulawarman statistical data.

2.4 Data and Implement Setting

In order to demonstrate the process of forecasting the nonlinear time series, a four days daily network traffic data from 21 to 24 June 2013 (192 samples series data) was taken and the GA based prediction algorithm is applied. The training data was 75%

Table 1 Network traffic real data

Date		Time	Inbound	Date		Time	Inbound
6/21/2013	1	0:00:00	6,293,000	6/23/2013	97	0:00:00	10,517,000
	2	0:30:00	5,185,000		98	0:30:00	6,715,000
	3	1:00:00	5,404,000		99	1:00:00	13,109,000
…	…	…	…	…	…	…	…
	47	23:00:00	12,390,000		143	23:00:00	7,121,000
	48	23:30:00	11,661,000		144	23:30:00	5,236,000
6/22/2013	49	0:00:00	8,390,000	6/24/2013	145	0:00:00	4,528,000
	50	0:30:00	7,307,000		146	0:30:00	3,603,000
	51	1:00:00	7,972,000		147	1:00:00	5,926,000
…	…	…	…	…	…	…	…
	95	23:00:00	10,444,000		191	23:00:00	6,190,000
	96	23:30:00	14,530,000		192	23:30:00	5,969,000

Table 2 Network traffic data after normalization

Group	Input period = $[X_{t-5}, X_{t-4}, X_{t-3}, X_{t-2}, X_{t-1}]$						Target output
	X_{t-5}		X_{t-4}	X_{t-3}	X_{t-2}	X_{t-1}	X_t
Train group	1	0.262	0.231	0.237	0.201	0.154	0.139
	2	0.231	0.237	0.201	0.154	0.139	0.164
	3	0.237	0.201	0.154	0.139	0.164	0.145
	…	…	…	…	…	…	…
	144	0.232	0.213	0.187	0.251	0.246	0.211
Test group	145	0.213	0.187	0.251	0.246	0.211	0.162
	146	0.187	0.251	0.246	0.211	0.162	0.163
	…	…	…	…	…	…	…
	192	0.253	0.262	0.231	0.237	0.201	0.154

(144 samples) and testing data was 25% (48). Before training, the inputs and tests data will be normalized. The aim of the normalization process is to get the data with a smaller size that represents the original data without losing its own characteristics. In this experiment, a MATLAB R2013b was used to perform the process of analyzing and forecasting. The normalization formula form is as follow,

$$\overline{X} = \frac{X - X_{min}}{X_{max} - X_{min}}, \qquad (1)$$

where, X: actual value of samples, X_{max}: maximum value, X_{min}: minimum value. The data after normalization show in Table 2. Based on the data outlined in Table 2, a function can be defined to learn this time series data as shown in Eq. 2,

$$X_t = a_{t-n}X_{t-n}(k) + \cdots + a_{t-1}X_{t-1}(k), \quad (2)$$

where X_t is the target output, the sequence of a_{t-n}, …, a_{t-1} is a positive real number that represents the weights, X_{t-n}, …, X_{t-1} is a sequence of time series data representing the network traffic data.

2.5 *Applying GA in Learning Time-Series Data*

In order to predict the network traffic using the proposed GA based prediction algorithm, the time series data must be arranged in order of time in one period. The purpose of this study is to measure the changes of data by minimizing the value of the difference between the actual and predicted values. The analysis of time series data using the proposed GA has been carried out as follows:

Step 1 Encoding schemes: Each gene in the chromosome is coded using a *real number encoding*. In other words, the chromosome is represented as a sequence of real numbers (describing a sequence of events). Where each chromosome x corresponds to a predefined fitness function f(x).

Step 2 Generating Initial Population: Initial population process is to determine the value of each gene in the chromosome to generate random numbers. The solution (or the structure of the chromosome) for the problem is defined based on the formula, $X_t = a_{t-n}X_{t-n}(k) + \cdots + a_{t-1}X_{t-1}(k)$, and the structure of the chromosome used to model the data shown in Table 2 will be [a_{t-5}, a_{t-4}, a_{t-3}, a_{t-2}, a_{t-1}]. The initial population size is 200.

Step 3 Evaluation function: Individual chromosome is evaluated based on a predefined function: $X_t = a_{t-n}X_{t-n}(k) + \cdots + a_{t-1}X_{t-1}(k)$, where the values for X_t, X_{t-1}, X_{t-2}, X_{t-3}, X_{t-4} and X_{t-5} are taken from Table 2. In other words, the GA is defined to minimize the Mean Squared Error (MSE) between the X_t and $a_{t-n}X_{t-n}(k) + \cdots + a_{t-1}X_{t-1}(k)$.

Step 4 Selection: The selection process is to establish a set of mating pool in accordance with the number of chromosomes to produce new offspring. In this experiment, three models of selection, namely the *roulette wheel*, *random* and *tournament*. In the Roulette wheel process, individual with the best fitness is not necessarily elected at the next generation but have a better chance of being elected. This process is done by generating random numbers (r), and then be checked against the values of a_1, a_2, a_3, a_4, a_5 to the number of population so that $r \le p_c$. In the Random selection process, the individual with the best fitness randomly selected from the population. In the Tournament process, the individual with the best fitness randomly selected and chosen as a parent with a size parameter value between 2 to N.

Step 5 Forming a New Generation: A new generation is formed by using two operators; namely crossover and mutation. A *one-point* method of crossover p_c with crossover rate of 0.2 and a *uniform multi point mutation* method with

Table 3 Setting and performance of GA

GA setting	Selection method		
	Roulette wheel	Random	Tournament
MSE	**0.004**	0.004	0.005
Time estimation (s)	337.744	337.815	339.632
Population	200	200	200
p_c	0.2	0.2	0.2
p_m	0.005	0.005	0.005
Iteration	100	100	100

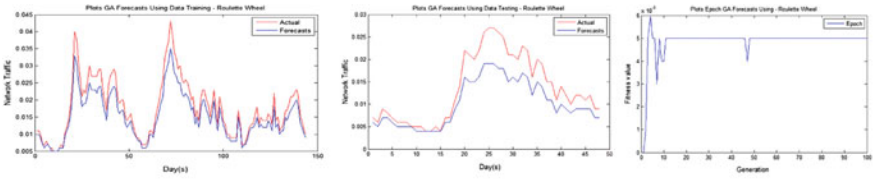
mutation rate of 0.005, and number of *iteration* of 100 times, and finally three selection processes will be used that includes the *roulette wheel*, *random* and *tournament* selections.

3 Results and Discussions

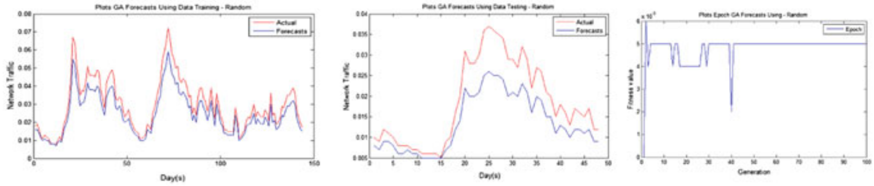
It means that the roulette wheel has better time processed.

This section presents the results obtained as shown in Table 3. The iteration process shows that the *roulette wheel* and *random* selections produced MSE values of 0.00497. But, the *random* selection has longer time estimation iteration than the *roulette wheel* selection which is 337.815 s. Table 3 also shows that the *tournament* selection has MSE value of 0.005 and 339.632 s for longest time estimation iteration. The GA based prediction algorithm has a relative long time estimation iteration process but this process depends on the set of input values. However, the MSE performance of the proposed GA has obtained good results. Figure 2 shows the graphs training and testing of three selections methods and the final MSE performance values which is 75% of the samples. In comparison, the MSE value obtained using the ARIMA (1, 0, 1)[12] is 0.00411 which is comparable with the result obtained using the GA based algorithm. AQ2

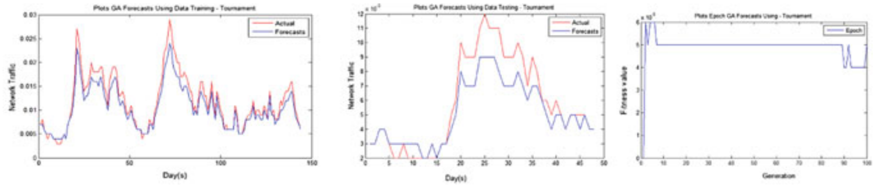
Therefore, the first training which has *population* size was 200, real number chromosomes, p_c with *one-point* method was 0.2 and p_m with *uniform multi point mutation* was 0.005, and *iteration* was used 100 to the output was optimal. The GA setting has been able to achieve the performance goals, and also has a pretty good MSE value.



(a) Roulette Wheel Selection



(b) Random Selection



(c) Tournament Selection

Fig. 2 Plots of results for the GA Modelling; *roulette wheel*, *random*, *tournament* selections

4 Conclusions

This paper examined a time series forecasting with genetic algorithms. The results shown that the proposed genetic algorithm has a pretty good value between training and testing data observed and predicted. Then, this algorithm can be used as an alternative modeling methodology in analyzing and forecasting time series data. Based on the experimental results obtained, it can be concluded that the GA setting with the *population* size of 200, *real number chromosomes*, a *one-point* method of crossover p_c with crossover rate of 0.2 and a *uniform multi point mutation* method with mutation rate of 0.005, with *roulette wheel* selection and number of *iteration* of 100 times, the time that is required to obtain an optimal output is approximately 337.815 s and the obtained MSE is quite encouraging. It means that the GA setting has been able to achieve the performance goals, and comparable to the result obtained using the ARIMA method. Therefore, one of the future works that can be conducted is to combine with neural network method in order to optimize the weights and biases or the structure for generate a higher accuracy of MSE and more efficient in the forecasting of short-term network traffic.

References

1. Ferrari-Santos, A.C., Simões da-Silva, J.D., de-Sá Silva, L., da-Costa Sene, M.P.: Network traffic characterization based on time series analysis and computational intelligence. J. Comput. Interdisc. Sci. **2**(3), 197–205 (2011)
2. Yu, Y., Wang, J., Song, M., Song, J.: Network traffic prediction and result analysis based on seasonal ARIMA and correlation coefficient. In: Network Traffic Prediction and Result Analysis Based on Seasonal ARIMA and Correlation Coefficient (2010)
3. Meng, X.M.: Weather Forecast based on improved genetic algorithm and neural network. In: Weather Forecast Based on Improved Genetic Algorithm and Neural Network, LNEE 219. Springer-Verlag London (2013)
4. Abhishek, K., Singh, M.P., Ghosh, S., Anand, A.: Weather forecasting model using artificial neural network. Procedia Technol **4**, 311–318 (2012)
5. Upadhyay, K.G., Choudhary, A.K., Tripathi, M.M.: Short-term wind speed forecasting using feed-forward back-propagation neural network. IJEST **3**(5), 107–112 (2011)
6. Vaisla, K.S., Bhatt, A.K.: An analysis of the performance of artificial neural network technique for stock market forecasting. IJCSE **2**(6), 2104–2109 (2010)
7. Perwej, Y., Perwej, A.: Prediction of the Bombay Stock Exchange (BSE) market returns using artificial neural network and genetic algorithm. J. Intell. Learn. Syst. Appl. **4**, 108–119 (2012)
8. Oliveira, F.A.d., Nobre, C.N., Zárate, L.E.: Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index—case study of PETR4, Petrobras, Brazil. Expert Syst. Appl. **40**, 7596–7606 (2013)
9. Claveria, O., Torra, S.: Forecasting tourism demand to Catalonia: Neural networks vs. time series models. Econ. Model. **36**, 220–228 (2014)
10. Zhang, H., Li, J.: Prediction of tourist quantity based on RBF neural network. J. Comput. **7** (2012)
11. Chabaa, S., Zeroual, A., Antari, J.: Identification and prediction of internet traffic using artificial neural networks. J. Intell. Learn. Syst. Appl. **2**, 147–155 (2010)
12. Haviluddin, Alfred, R.: Forecasting network activities using ARIMA method. J. Adv. Comput. Netw. **2**, 173–179 (2014)
13. Haviluddin, Alfred, R.: Daily network traffic prediction based on backpropagation neural network. Aust. J. Basic Appl. Sci. **8**(24), 164–169 (2014)
14. Haviluddin, Alfred, R.: Comparison of ANN back propagation techniques in modelling network traffic activities. In: Comparison of ANN Back Propagation Techniques in Modelling Network Traffic Activities (2014)
15. Purnawansyah, Haviluddin: Comparing performance of Backpropagation and RBF neural network models for predicting daily network traffic. In: Comparing Performance of Backpropagation and RBF Neural Network Models for Predicting Daily Network Traffic (2014)
16. Gill, E.J., Singh, E.B., Singh, E.S.: Training back propagation neural networks with genetic algorithm for weather forecasting. In: Training Back Propagation Neural Networks with Genetic Algorithm for Weather Forecasting (2010)
17. Song, F., Wang, H.: Hybrid Algorithm based on Levenberg-Marquardt Bayesian regularization algorithm and genetic algorithm. In: Hybrid Algorithm Based On Levenberg-Marquardt Bayesian Regularization Algorithm and Genetic Algorithm (2013)
18. Yang, C.-X., Zhu, Y.F.: Using genetic algorithms for time series prediction. In: Using Genetic Algorithms for Time Series Prediction (2010)
19. Alfred, R.: Summarizing relational data using semi-supervised genetic algorithm-based clustering techniques. J. Comput. Sci. **6**(7), 775–784 (2010)
20. Sedki, A., Ouazar, D., El Mazoudi, E.: Evolving neural network using real coded genetic algorithm for daily rainfall–runoff forecasting. Expert Syst. Appl. **36**, 4523–4527 (2009)
21. Alfred, R., Kazakov, D.: A clustering approach to generalized pattern identification based on multi-instanced objects with DARA. In: Local Proceedings of ADBIS, Varna, pp. 38–49 (2007)
22. Melanie, M.: An introduction to genetic algorithms. In: An Introduction to Genetic Algorithms. Massachusetts Institute of Technology (1996)

Author Queries

Chapter 18

Query Refs.	Details Required	Author's response
AQ1	Kindly provide organisation name for the author Haviluddin.	Done
AQ2	Please provide a definition for the significance of bold in Table 3.	Done

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	• • • under matter to remain	
Insert in text the matter indicated in the margin		New matter followed by or
Delete	/ through single character, rule or underline or ⊢—⊣ through all characters to be deleted	or
Substitute character or substitute part of one or more word(s)	/ through letter or ⊢—⊣ through characters	new character / or new characters /
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	= under matter to be changed	=
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	/ through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	/ through character or where required	
Reduce space between characters or words	between characters or words affected	