Advanced Science Letters
Vol. 21, 3037–3041, 2015

A Performance Comparison of Statistical and Machine Learning Techniques in Learning Time Series Data

Haviluddin[1], Rayner Alfred[2], Joe Henry Obit[2], Mohd Hanafi Ahmad Hijazi[2], and Ag Asri Ag Ibrahim[2]
[1]Dept. Computer Science, Faculty of Mathematics and Natural Science, UNMUL 75119, Indonesia
[2]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88999 Kota Kinabalu, Sabah, Malaysia

The task of analyzing and forecasting time-series data is very crucial task as this Time Series Analysis (TSA) task 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. The techniques used can be classified into two categories, namely statistical and machine learning techniques. As a result, the selection of several prediction methods will continue to be an alternative for researchers to obtain more accurate prediction results. This paper outlines and presents the comparison of predictive performance of statistical and machine learning techniques, namely ARIMA, Back-Propagation Neural Network (BPNN), and genetic algorithms (GA) for analyzing and predicting short-term time series network traffic activity datasets. In other words, this paper examines the forecasting performance of ARIMA, BPNN and GA models for the time series data related to network traffic activity data which is obtained from the ICT Universitas Mulawarman. The performances of these techniques are compared based on the errors measured, namely Mean Squared Error (MSE). Based on the results obtained, BPNN is found to be very efficient in learning a time series data. This paper is concluded by recommending some future works that can be applied in order to improve the prediction accuracy.

Keywords: Forecasting Time Series, ARIMA, Back Propagation Neural Network, Genetic Algorithm, Mean Squared Error.

1. INTRODUCTION

Several research studies on network traffic have been conducted with various solution techniques proposed over the years. The techniques used can be classified into two categories, namely statistical and machine learning techniques. Statistical techniques used for the forecasting time series includes Autoregressive (AR), Moving Average (MA), Exponential Smoothing (ES), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), Autoregressive Integrated Moving Average (ARIMA), and Seasonal Autoregressive Integrated Moving Average (SARIMA) models[1-4].

On the other hand, machine learning (ML) techniques have been also widely used for analyzing and forecasting time series data in the past four decades. Artificial Neural Networks (ANNs) are found to be efficient in solving nonlinear problems and many researchers have been using ANNs widely as a time series analysis method to solve problems in many areas including economic, business, finance, foreign exchange, and stock problems, engineering, energy, internet, and network traffic[5-9]. Besides ANNs, a Genetic Algorithm (GA) method is also one of the ML techniques used to solve the problem of forecasting a non-linear time series dataset[10,11].

*Email Address: haviluddin@unmul.ac.id

1936-6612/2015/4/400/008 doi:10.1166/asl.2015.6490

In this paper, the performances of the Autoregressive Integrated Moving Average (ARIMA), Backpropagation Neural Networks (BPNN), and Genetic Algorithm (GA) models are studied and compared. A time series data related to network traffic activity dataset is used in this work which is obtained from the ICT Universitas Mulawarman. The rest of this paper is organized as follows. Section 2 briefly introduces and describes the time series analysis methods that include ARIMA, BPNN, and GA are introduced. The dataset used in this work will be described in Section 2. In Section 3, based on the results obtained, the performance comparison of the statistical and machine learning techniques in learning time series data is discussed. Finally, this paper is concluded in Section 4.

2. METHODOLOGY

In this section, the principles of ARIMA, BPNN, and GA models for analyzing a time series dataset are briefly described.

2.1. THE PRINCIPLE OF ARIMA

The basic concept of Autoregressive Integrated Moving Average (ARIMA) is developed by George EP Box and Gwilym M. Jenkins in 1976. Thus, this model is often called the Box-Jenkins model. This model consists of Autoregressive (AR), Moving Average (MA), Autoregressive-Moving Average (ARMA) and ARIMA. The Box and Jenkins methodology includes four iterative steps of model identification. The basic tools used to identify model includes

a) The autocorrelation function (ACF) and partial autocorrelation function (PACF)
b) Estimation of the parameter using data differencing and transformation,
c) Diagnostic checking; to help suggest alternative model(s), and
d) Forecasting; the satisfactory model can be used for prediction purposes[4,12].

2.2. THE PRINCIPLE OF BACKPROPAGATION OF NEURAL NETWORK (BPNN)

In general, Backpropagation of Neural Network (BPNN) is a well-known learning method for multi-layer perceptron training with supervised learning method. The BPNN was first proposed by Paul Werbos in 1974, then developed by David Parker in 1982. Afterward, it was popularized by Rumelhart and McCelland in 1986.

The BPNN is a three-layer feed-forward neural network, which includes an input layer, a hidden layer and an output layer with linear neurons. In principles, BPNN involves two steps, a forward propagating step and a backward propagating step. In the forward propagating step, the training data set is presented to the input layer, which will propagate through the hidden layers until it reaches the output layer. In the backward propagating step, the calculated error (the difference between the actual output and the desired output) is propagated back to change the assigned weights[5-7].

2.3. THE PRINCIPLE OF GENETIC ALGORITHM (GA)

The basic concept of Genetic Algorithm (GA) is found at the University of Michigan, United States of America by John Holland in 1975. Then, it was popularized by one of his students, David Goldberg in the 1980s. Later, 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[10,13,14].

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[11,15]. Fig. 1 illustrates the cycle of the GA implementation.

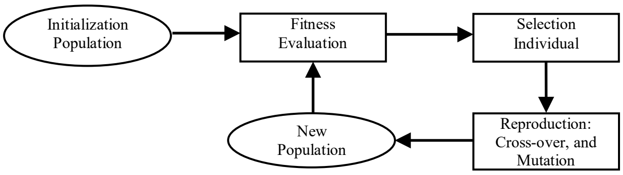
Initialization Population
Fitness Evaluation
Selection Individual
Reproduction: Cross-over, and Mutation
New Population

Fig.1. The Genetic Algorithm Cycle

2.4. 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 [8],[14]. In principle, a time series model is used to predict the current value of data, X_t, based on the data (X_{t-n},…,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 GA based prediction algorithm. Table 1 and Fig.2 show the inbound and outbound of the network traffic real data obtained from the Universitas Mulawarman statistical data.

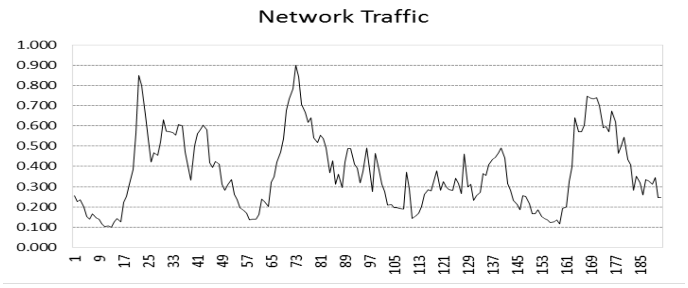

Fig.2. Network Traffic Activity (20 – 24 June 2013)

Table.1. Network traffic real data

Date		Time	Inbound
6/21/2013	1	0:00:00	6293000
	2	0:30:00	5185000
	3	1:00:00	5404000
	…	…	…
	48	23:30:00	11661000
6/22/2013	49	0:00:00	8390000
	50	0:30:00	7307000
	…	…	…
	96	23:30:00	14530000
6/23/2013	97	0:00:00	10517000
	98	0:30:00	6715000
	99	1:00:00	13109000
	…	…	…
	144	23:30:00	5236000
6/24/2013	145	0:00:00	4528000
	146	0:30:00	3603000
	…	…	…
	192	23:30:00	5969000

2.5. DATA AND GA SETTING

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

$$\overline{X} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where, X : actual value of samples, X_{max} : maximum value, X_{min} : minimum value. Table 2 shows the data after the normalization process by applying the function show below in order to learn this time series.

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

where X_t is the target output, the sequence of $a_{t-n},\ldots,a_{t-1}$ is a positive real number that represents the weights, $X_{t-n},\ldots,X_{t-1}$ is a sequence of time series data representing the network traffic data. The solution (or the structure of the chromosome) for the problem is defined based on the formula $X_t = a_{t-n}X_{t-n}(k) + \ldots + a_{t-1}X_{t-1}(k)$, and the structure of the chromosome used to model the data shown in Table 2 will be $[a_{t-5},a_{t-4},a_{t-3},a_{t-2},a_{t-1}]$. Individual chromosome is evaluated based on a predefined function: $X_t = a_{t-n}X_{t-n}(k) + \ldots + a_{t-1}X_{t-1}(k)$, 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 $a_{t-n}X_{t-n}(k) + \ldots + a_{t-1}X_{t-1}(k)$.

Table.2. Network traffic data after normalization

Group		Input					Target Output
		Period = $[X_{t-5}, X_{t-4}, X_{t-3}, X_{t-2}, X_{t-1}]$					
		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

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1 A PERFORMANCE COMPARISON OF STATISTICAL AND MACHINE LEARNING TECHNIQUES

This section compares the performance of the Statistical and Machine Learning Techniques, namely ARIMA, BPNN and GA using the time series data related to network activity. In this test, the time series data must be arranged in order of time in one period and the Mean Squared Error (MSE) performance value is used to get the difference between the actual data and predicted data. The lower the MSE computed the better is the model in learning the time series data.

3.2. RESULT OF ARIMA MODEL

Using the ARIMA Model, there are several different parameters of autoregressive *(p)* and moving average *(q)* used in order to determine the best model that provides the lowest MSE as indicated in Table 3. Fig. 3 and 4 shows the time-series forecasting results using the ARIMA $(1,0,1)^{12}$ and it is considered the best model for analyzing and forecasting the network traffic activity time series data in this work.

Table.3. ARIMA results with different parameters

ARIMA	Mean Squared Error (MSE)	ARIMA	Mean Squared Error (MSE)
$(1,0,1)^{12}$	**0.00411**	$(2,0,1)^{12}$	0.00426
$(1,0,2)^{12}$	0.00418	$(2,0,2)^{12}$	0.00959
$(1,1,1)^{12}$	0.00426	$(2,1,1)^{12}$	0.00425
$(1,1,2)^{12}$	0.00428	$(2,1,2)^{12}$	0.00427
$(1,2,1)^{12}$	0.00467	$(2,2,2)^{12}$	0.00799

3.3 RESULT OF BPNN MODEL

In the BPNN Model, there are several different architectures used in order to determine the best model that provides the lowest MSE as indicated in Table 4. The first number indicates the number of neurons in the input layer, the second number represents the neurons in the hidden layers, and the last number represents the neurons in the output layer. Then, *epochs*, *learning rate*, and *momentum* have been set to 1000, 0.1, and 0.8. The BPNN architecture that has been used for the one-hidden layer with the activation function; from input to hidden layer was *tansig*, and from hidden layer to output was *purelin*, in which a gradient descent with momentum (*traingdm*) algorithm is used and the computed MSE values are shown in Fig. 5.

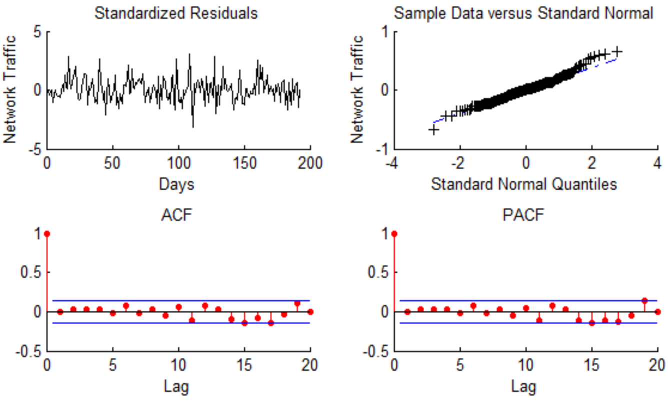

Fig.3. ACF and PACF of ARIMA (1,0,1)[12]

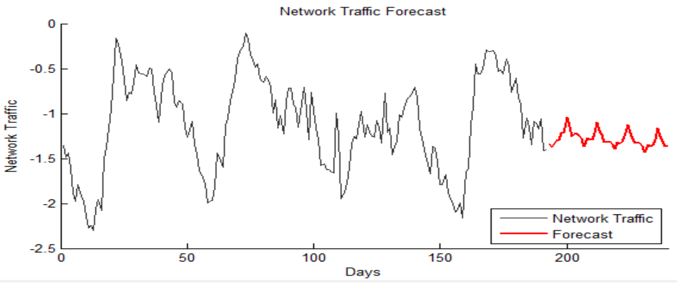

Fig.4. Performance and Plots Forecast of ARIMA (1,0,1)[12]

Table.4. BPNN results with different architectures

Model	Architectures	Epochs	LR	Momentum	MSE
1	**5-10-1**	**10000**	**0.1**	**0.8**	**0.0092089**
2	5-11-1	10000	0.1	0.8	0.0092137
3	5-20-1	10000	0.1	0.8	0.0102578
4	5-30-1	10000	0.1	0.8	0.0101119
5	5-40-1	10000	0.1	0.8	0.0107583

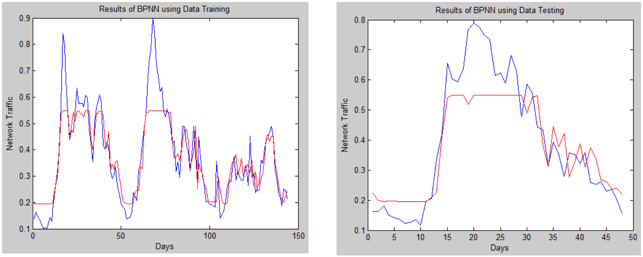

Fig.5. Performance and Plots Forecast of BPNN 5-10-1

3.4 RESULT OF GA MODEL

This section presents the best achieved results for the genetic algorithm based prediction method. The GA parameters that were used includes the *population* size of 200, real number chromosomes, *one-point* crossover method p_c with 0.2 rate and *uniform multi point mutation* p_m with 0.005 rate, 100 *iterations*, *roulette wheel* was used for the selection method. The GA setting has been able to achieve the performance goal, and also has a pretty good MSE value of 0.00497 with the time estimation iteration of 337.815s as shown in Table 5 and Fig. 6.

Table.5. Setting and performance of GA

GA Setting			
Selection Method	Roulette Wheel	p_c (crossover)	0.2
Chromosomes	Real number	p_m (mutation)	0.005
Population	200	Iteration	100

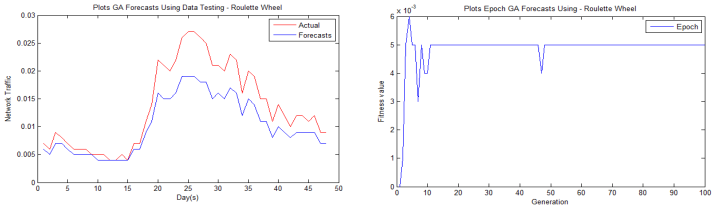

Fig.6. Performance and Plots Forecast of GA

Table.6. Comparison of MSE from three models; ARIMA, BPNN, and GA

Model	**MSE**
ARIMA (1,0,1)[12]	0.00411
BPNN	0.00092
GA	0.00497

4. CONCLUSIONS

This paper has presented the performance comparison of statistical and machine learning techniques, namely ARIMA, BPNN and GA, in learning time series data. The mean squared errors are computed for each model and compared. Based on the results obtained, the BPNN algorithm is found to be more efficient in modelling time series dataset related to network activity. Optimizing the architecture of the BPNN can be considered as one of the future works that can be conducted in order to investigate the best architecture available for the BPNN in learning time series data.

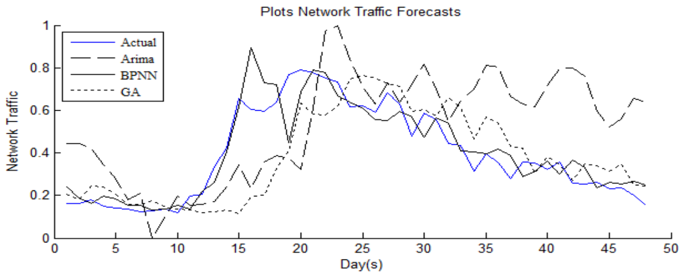

Fig.6. Plots Forecast of ARIMA, BPNN and GA

REFERENCES

[1] X. Zhang, Y. Liu, M. Yang, T. Zhang, A.A. Young, X. Li, Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China, Publisher, City, 2013.

[2] L. Xiao, J. Wang, Y. Dong, J. Wu, Combined forecasting models for wind energy forecasting: A case study in China, Publisher, City, 2015.

[3] A.A. Adebiyi, A.O. Adewumi, C.K. Ayo, Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction, Publisher, City, 2014.

[4] Haviluddin, R. Alfred, Forecasting Network Activities Using ARIMA Method, Publisher, City, 2014.

[5] Haviluddin, R. Alfred, Daily Network Traffic Prediction Based on Backpropagation Neural Network, Publisher, City, 2014.

[6] G. Sermpinis, C. Dunis, J. Laws, C. Stasinakis, Forecasting and trading the EUR/USD exchange rate with stochastic Neural Network combination and time-varying leverage, Publisher, City, 2012.

[7] C.C. Gowda, S.G. Mayya, Comparison of Back Propagation Neural Network and Genetic Algorithm Neural Network for Stream Flow Prediction, Publisher, City, 2014.

[8] O. Claveria, S. Torra, Forecasting tourism demand to Catalonia: Neural networks vs. time series models, Publisher, City, 2014.

[9] S. Agrawal, P.D. Murarka, Stock Price Forecasting : Comparison of Short Term and Long Term Stock Price Forecasting using Various Techniques of Artificial Neural Networks, Publisher, City, 2013.

[10] E.J. Gill, E.B. Singh, E.S. Singh, Training Back Propagation Neural Networks with Genetic Algorithm for Weather Forecasting, in: 8th International Symposium on Intelligent Systems and Informatics, © 2010 IEEE, September 10-11, 2010, Subotica, Serbia, 2010, pp. 465-469.

[11] C.-X. Yang, Y.-F. Zhu, Using Genetic Algorithms for Time Series Prediction, in: 2010 Sixth International Conference on Natural Computation (ICNC 2010), © IEEE, 2010, pp. 4405-4409.

[12] M. Valipour, M.E. Banihabib, S.M.R. Behbahani, Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir, Publisher, City, 2013.

[13] Y. Perwej, A. Perwej, Prediction of the Bombay Stock Exchange (BSE) Market Returns Using Artificial Neural Network and Genetic Algorithm, Publisher, City, 2012.

[14] F. Song, H. Wang, Hybrid Algorithm Based On Levenberg-Marquardt Bayesian Regularization Algorithm and Genetic Algorithm, in: The 2013 International Conference on Advanced Mechatronic Systems, © IEEE, Luoyang, China, 2013, pp. 51-56.

[15] M. Melanie, An Introduction to Genetic Algorithms, in, © 1996 Massachusetts Institute of Technology, 1996, pp. 1-143.

Received: 22 August 2015. Accepted: 10 October 2015