Research on Modeling and Forecasting of Network Traffic Data with Alpha Distribution Noise Based on FARIMA Model

Qiuting Yan, Hu Sheng*, Hongjia Rong

School of Electrical and Information Engineering, Dalian Jiaotong University, Dalian, China

*Corresponding author e-mail: hu.sheng@djtu.edu.cn

Abstract. Network business traffic data has obvious self-similar characteristics, and its modeling and predictive analysis should also adopt self-similar models. Short correlation models, which cannot accurately describe their self-similarity, is mainly adopted in traditional network traffic data modeling. Therefore, It is necessary to research long correlation models. In addition, network traffic data also has certain pulse characteristics, which leads to inaccurate data modeling. In this research, alpha stable noise is added in FGN time series to simulate network traffic data, and the simulated data is analyzed by improved FARIMA model. Finally, the improved FARIMA model is applied to the BC-p Aug89.TL data modeling and predictive analysis. The analysis shows that the improved FARIMA model has a better prediction effect on network traffic and is of great significance for the application of network traffic prediction.

Keywords: FARIMA model; Alpha stable distribution noise; network traffic; modeling and prediction

1. Introduction

The current research on network traffic flow has fully demonstrated that network traffic flow has a high degree of autocorrelation [1], and most network traffic time series are non-stationary random signals. It is also suggested that the network traffic data has pulse characteristics, and the probability distribution of data with pulse characteristics exhibits significant heavy-tailed distribution characteristics. The parameter $\alpha$ indicates the degree of heavy tail, where $\alpha$ is a parameter indicating the tail thickness of probability distribution. Among them, the heavy tail distribution that best fits the actual sequence is Alpha stable distribution, which is more applicable than the Gaussian distribution. The tailing is attenuated by the square law. The speed of attenuation decreases as the parameter $\alpha$ decreases. It becomes heavy [2].

FARIMA (Fractional Autoregressive Integration Moving Average model) is derived from the ARMA [3] and ARIMA models. It is a model with the characteristics of describing short correlation and long correlation. At the same time, $d$ of the FARIMA model is obtained indirectly using the Hurst parameter and the value of $d$ is not limited to integers. In this research, the value of $d$ is used to estimate the value of $d$, and the FGN sequence contaminated by alpha stable distribution noise is...
simulated and predicted. This method is used to improve the accuracy of the FARIMA prediction model, which is the focus of this study. Many researches on the use of FARIMA models for network traffic prediction are currently in the theoretical stage, but there is little research on the application. How to use the prediction results to guide and improve network management is a current in-depth study.

2. FARIMA Model Introduction
The FARIMA\((p,d,q)\) model is an ARMA model that is stimulated by fractional differential noise FARIMA\((0,d,0)\) [4]. When using the number of differences \(d\) to describe the long correlation structure in the observation sample, the model uses ARMA to characterize the short correlation structure in the sample. FARIMA\((p,d,q)\), \(p\) is the autoregressive order, \(q\) is the moving average order, and \(d\) is the number of differences. The definitions show as below:

\[
\Phi(B)\nabla^d W_t = \Theta(B)a_t
\]  

(1)

Where \((-0.5, 0.5)\{a_t\}\) is a white noise sequence, \(\Phi(B)\) and \(\Theta(B)\) are \(p\)-order and \(q\)-order polynomials, respectively, without common roots, \(B\) is a delay operator, ie \(BW_t = W_{t-1}\).

3. Stochastic sequence modeling and predictive analysis

3.1 Introduction to random data
In this study, a random sequence FGN(fractional Gaussian noise sequence)[5] with long correlation characteristics is added to the FGN, and an alpha stable distribution noise with the characteristic parameter \(\alpha = 1.5\) is added. The selected random sequence contains 10,000 data packets. Through modeling and analyzing, Figure 1 shows the contaminated random FGN data.

![Figure 1. Random FGN data after pollution.](image)
As can be seen from the above figure, after the FGN of the fractional-order Gaussian noise sequence is added with the alpha stable distribution noise, the sharp peak appears. The occurrence of this kind of noise is mainly due to the heavy-tailed distribution of the alpha distribution noise, which causes the time series to appear significantly. The spike part and the heavy-tailed distribution of the time series are also one of the reasons for self-similarity.

3.2 The prediction process of the FARIMA model
The main idea for predicting network traffic through the FARIMA model is using FARIMA(p,d,q) process to separate into fractional difference and ARMA process, which has the following steps.

First, the contaminated random time series is pre-processed and zero-meanized. Second, the characteristic parameter $\alpha$ of the stable distributed noise alpha and the Hurst parameter of the random time series FGN are used to obtain the estimated value of $d$ after performing the fractional difference filtering. The processed data is identified by the ARAM model for prediction of future data. The prediction process is shown in Figure 2.

In order to analyze the effect of fractional differential filtering on removing time series typing characteristics, the long-term correlation processing is performed on the data generated by the simulation and the real network traffic. First, the Hurst parameter is estimated, and $\hat{d}$ is estimated through the relationship between $d$ and $H/\alpha$. The data is differentially filtered to detect the degree of self-similarity after filtering. The long correlation characteristics of the data are removed through the fractional difference to meet the short correlation characteristics of the ARMA process, and then used for the prediction of the ARMA process. Through analysis, it can be seen that the degree of self-similarity of the data stream with self-similarity can be reduced to about 0.5 after differential filtering, which indicates that the data after the difference basically does not contain long correlation features.
3.3 FARIMA model parameter estimation

In this study, a FARIMA[6] model was established for the contaminated random FGN data shown in Figure 1. According to the modeling process given in Section 3.2, the Hurst parameter of the contaminated FGN sequence is first estimated and then the value of the model score difference parameter \( d \) is determined, and \( d = 0.1283 \) is obtained according to the formula \( d = H - 1/ \alpha \) [7].

The autocorrelation coefficient(ACF) and partial autocorrelation coefficient(PACF) of the polluted FGN time series are used to estimate the values of the parameters \( p \) and \( q \). It can be seen from Figure 3 that the autocorrelation coefficient ACF is obviously censored. It can be seen from Figure 4 that the PAC is tailing. It can be preliminarily judged that the data can be described by the MA model, and the model parameters are \( p = 0, q = 2 \). Then the AIC[8] minimum criterion is used to test. When the order is 2, the minimum AIC value is 1.6705, so the model order is judged to be 2.

According to the established FARIMA \((0,0.1283,2)\) model, the approximate prediction curve of the
model sample is shown in Figure 5. The red asterisk represents the predicted point. It is can be seen that the predicted curve approaches the random time series FGN, which proves the reliability of the FARIMA model.

![Figure 5. Model sample approximation prediction curve.](image)

4. Modeling and predictive analysis of actual network traffic

4.1 Introduction to network business flow

In this study, BC-Oct89Ext measured flow data provided by Bell Labs was used as the research source[5]. Take some of these data segments to make a traffic prediction, add $\alpha = 1.5$ alpha stable distribution noise to it, the sequence length is 10000, and finally use the FARIMA model for prediction. Figure 6 shows some data segments in the BC-Oct89Ext measured flow data before being contaminated.

![Figure 6. Flow data before pollution.](image)
From the above FARIMA model prediction process, the parameters of the actual network traffic flow are $d = 0.0926, p = 0, q = 2$, respectively. Use FARIMA $(0, 0.0926, 2)$ to predict the network model. The following figure is the prediction result of selecting 48 data points in the actual traffic data. From the figure 7, it can be seen that the prediction result is very close to the actual value, and the average absolute error of the point prediction of the 48 data is 3.3022, which proves the effectiveness of the FARIMA model for actual network prediction.

5. Conclusion
In this paper, alpha stable distribution noise with $\alpha = 1.5$ is added into the BC-Oct89Ext measured flow data provided by Bell Labs, and the FARIMA $(0, 0.091, 2)$ model is established and used for prediction simulation. The experimental results show that the FARIMA model can better describe the actual network traffic flow problem. In the FARIMA model, fractional difference filtering is an important technique for network data flow. In this paper, by injecting alpha stable distribution noise and using the relationship between $H, \alpha$ and $d$ to derive the difference parameters, the analysis of the results shows that the FARIMA model is of high usability value. In the future, the research on network traffic forecasting of the FARIMA model will still be strengthened to realize network service forecasting, access control, and bandwidth allocation in a faster and more convenient way.

Acknowledgments
The study was supported by “The Doctoral Scientific Research Foundation of Liaoning Province, China (Grant No. 20170520215)”, “Natural science research project of Liaoning Provincial Department of Education, China (Grant No. JDL2019014)”

References
[1] M.E. Crovella, Self-similarity in World Wide Web traffic evidence and possible causes, IEEE/ACM Transactions on Networking. 1997 5(6) 835-846.
[2] W. Willinger, M.S. Taqqu, R. Sherman, Self-similarity through high variability: statistical analysis of ethernet lan traffic at the source level, IEEE ACM trans on networking. 1997 1(5) 71-86.
[3] Akaike Hirotugu, Power spectrum estimation through autoregressive model fitting, Annals of the Institute of Statistical Mathematics. 1969 21(1) 407-419.
[4] M.E. Souse, A. Suarez, J.C. Lopez, Application of the whittle estimator to the modeling of traffic based on the M/G/∞ process, IEEE Communications Letters. 2007 11(10) 817-819.
[5] I. Norros, On the Use of Fractional Brownian Motion in the Theory of connectionless Networks, IEEE Journal on Selected Areas in Communications. 1995 13(6) 953-962.
[6] J.P. Montillet, Modeling Geodetic Processes with Levy α-Stable Distribution and FARIMA, Mathematical geosciences. 2014 12(12).
[7] F.C. Harmantzis, Heavy Network Traffic Modeling and Simulation using Stable FARIMA Processes, Research on the Effectiveness of the Next Generation Network (6a). 2005.
[8] G.E.P. Box, Some Recent Advances in Forecasting and Control, Journal of the Royal Statistical Society Series C (6a). 1974 23(2) 158-179.