응용통계연구 제 **20**권 2호, 2007년, *pp. 219–227*

시계열 모형을 이용한 통신망 트래픽 예측 기법연구*

김삼용[1)]

요 약

시계열 모형은 통신망 트래픽의 예측과 분석에 유용하게 쓰여 왔다. 본 논문에서는 통신망 트래픽의 예측을 위하여 다양한 시계열 모형을 소개하고 성능평가를 하고자 한다. 이를 위하여 실제 통신망 트래픽 자료에 선형 및 비선형 시계열모형을 적합 시키고 비선형 시계열모형이 선형 시계열 모형보다 예측의 정확도가 우수함을 보이고자 한다.

주요용어: 시계열 모형, 통신망 트래픽, 예측(forecasting).

1. 서론

통신망의 운용 및 관리는 통신 사업자에게 있어서 매우 중요한 일 중의 하나이다. 최근에 통신망에 접속하는 트래픽이 급증하고 있다. 이에 상응하여 통신망의 안정적이고 효율적인 운용을 위해 네트워크 트래픽의 통계적 분석을 통한 신뢰성 있는 예측 기법의 개발이 절실히 요구되고 있다. 통계적 예측 기법으로 현재까지 잘 알려져 있는 ARIMA 모형(Box와 Jenkins, 1976)이 있으며 이 기법을 통하여 통신 분야의 많은 시계열 자료들을 분석하고 예측하고 있다. 특히 통신망에서 자주 발생하여 통신망의 안정적 운용에 치명적인 타격을 가하는 인터넷 웜 바이러스, DoS(Denial of Service) 바이러스 등을 조기에 탐지하는 분야에도 시계열 모형이 적용되는데 예를 들어, Hellerstein 등(2001)은 AR모형을 이용하여 이상 트래픽 탐지기법을 개발하였고 Kim 등(2005)는 이를 발전시켜 ARCH모형을 적합하여 이상 트래픽 탐지기법을 개발하고 기존의 AR모형보다 우수함을 시뮬레이션을 통하여 보여주었다. 한편 이동통신 트래픽의 예측을 위하여 Shu 등(2005)은 계절형 ARIMA모형을 이용하여 GSM(Global System for Mobile) 트래픽을 예측 하였다. 본 연구에서는 네트워크 트래픽의 효율적 예측을 위하여 일련의 비선형 시계열 모형을 소개하고 실제 자료를 이용하여 선형 시계열 모형과의 성능비교를 실시하였다. 여기서 비선형 시계열 모형이 선형 시계열 모형보다 예측의 우수성을 보이고자 한다.

2. 시계열 모형

먼저 기존의 선형 시계열 모형에서 AR(p) 모형을 소개하고 모형에 포함된 오차의 분산이 시간에 따라 동일하지 않는 경우인 비선형 시계열 모형을 살펴보기로 한다.

* 이 논문은 2006년도 중앙대학교 학술연구비 지원에 의한 것임.

1) (156-756) 서울시 동작구 흑석동, 중앙대학교 통계학과, 부교수
E-mail: sahm@cau.ac.kr

2.1. AR(Autoregressive) 모형

Box-Jenkins(1976)가 제안한 ARMA 모형의 일환인 AR(p) 모형은 시계열 를 그 이전 시점의 시계열로 회귀시킨 모형이다. AR(p) 모형은 다음과 같이 정의된다.

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t, \quad t = 1, 2, \ldots, n \tag{2.1}$$

여기서 $\epsilon_t \sim iid\,(0, \sigma^2)$ 이고, y_t는 약정상성(weak stationarity) 조건을 만족하는 정상 시계열 이라 가정한다.

2.2. ARCH(Autoregressive Conditional Hetroscadastic) 모형

ARCH 모형은 오차의 분산이 자기 회귀적으로 변하는 이분산성 모형으로 Engel(1982)에 의해 처음 제시되었으며, AR(2) 모형에 근거한 ARCH(p) 모형은 다음과 같다.

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \\ \epsilon_t &= \sqrt{h_t} e_t\ , \quad t = 1, 2, \ldots, n, \quad e_t \sim\ iid\ N(0,1), \\ h_t &= \alpha_0 + \sum_{j=1}^{p} \alpha_j \epsilon_{t-j}^2. \end{aligned} \tag{2.2}$$

여기서 $\alpha_0 > 0,\ \alpha_i > 0,\ \beta_j > 0,\ \sum_{i=1}^{p} \alpha_i < 1$이다.

2.3. GARCH(Generalized ARCH) 모형

GARCH 모형은 오차의 분산이 자기 회귀적으로 변하는 ARCH 모형의 일반화된 모형으로써 Bollerslev(1986)에 의해 제시되었으며 AR(2) 모형에 근거한 GARCH(p,q) 모형은 다음과 같다.

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \\ \epsilon_t &= \sqrt{h_t} e_t\ , \quad t = 1, 2, \ldots, n, \quad e_t \sim\ iid\ N(0,1), \\ h_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}. \end{aligned} \tag{2.3}$$

여기서 $\alpha_0 > 0,\ \alpha_i > 0,\ \beta_j > 0,\ \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$이다.

GARCH 모형의 특수한 경우인 GARCH(1,1) 모형은 다음과 같다.

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1}.$$

2.4. IGARCH(Integrated GARCH) 모형

GARCH 모형에서는 자료가 정상성(stationarity)을 만족시켜야 한다. 하지만 이 조건이 만족되지 못하고 식 (2.4)와 같은 경우가 발생할 때 단위 근(unit root)을 가지는 모형과 유사하게 다음과 같은 모형을 제시할 수 있다.

$$\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j = 1. \tag{2.4}$$

AR(2) 모형에 근거한 IGARCH(p,q) 모형은 다음과 같다.

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \\ \epsilon_t &= \sqrt{h_t}e_t\ , \quad t = 1,2,\ldots,n, \quad e_t \sim\ iid\ N(0,1), \\ h_t &= \alpha_0 + \sum_{i=1}^{p}\alpha_i\epsilon_{t-i}^2 + \sum_{j=1}^{q}\beta_j h_{t-j}. \end{aligned} \tag{2.5}$$

여기서

$$\alpha_0 > 0,\ \ \sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j = 1.$$

이다. IGARCH 모형의 특수한 경우인 IGARCH(1,1) 모형은 다음과 같다.

$$h_t = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \beta_1 h_{t-1},\ \ \alpha_1 + \beta_1 = 1$$

3. 자료 분석

본 연구에서는 국가 보안 기술 연구소에서 제공받은 시계열 자료(2004년 11월, 3지역)를 가지고 ARIMA, ARCH, GARCH, IGARCH 모형을 적합 시켰다. 관측시점이 각각이 9918개인 3지역의 자료를 log 변환 후 1차 차분하여 정상시계열 모형으로 변환하였다. 모형의 성능 비교를 위하여 먼저 각각의 모형에 대해 최소 제곱법을 이용하여 모수를 추정하였다. 여기서 9918개중 9762개는 모형 적합에 사용하였고 나머지 156개는 예측의 정확도를 알아 보기 위해 사용하였다. 또한 식 (3.1)과 같이 RMSE(root mean squared error)를 이용하여 모형의 예측 정확도를 비교하였다. 여기서 $\hat{x_t}$는 추정 값이며 $x_t = \log(y_t/y_{t-1})$이다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_t - \hat{x_t})^2}, \quad n = 156. \tag{3.1}$$

표 3.1에서는 각 모형의 모수 추정치를 보여주고 있고, 표 3.2에서 각 모형의 RMSE를 보여 주고 있다. 표 3.2의 결과를 보면 기존의 AR(2) 모형은 통신망 트래픽을 예측하는데

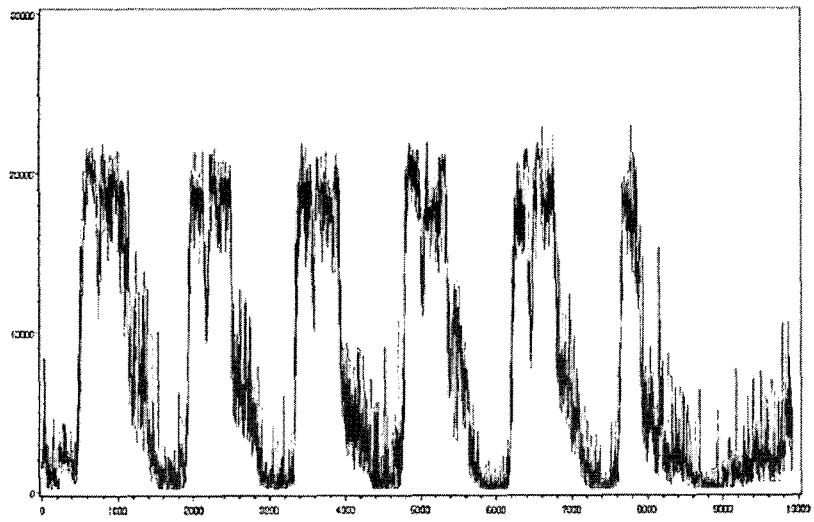

그림 3.1: 자료 1의 원자료

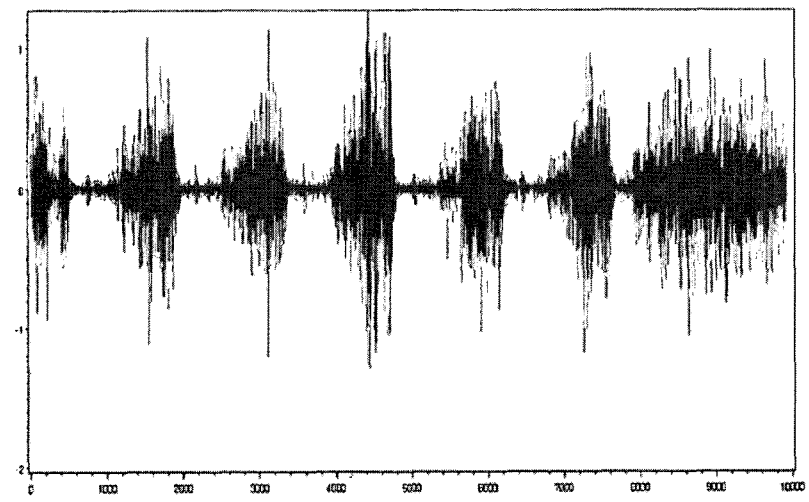

그림 3.2: 자료 1의 로그 변환 후 차분한 자료

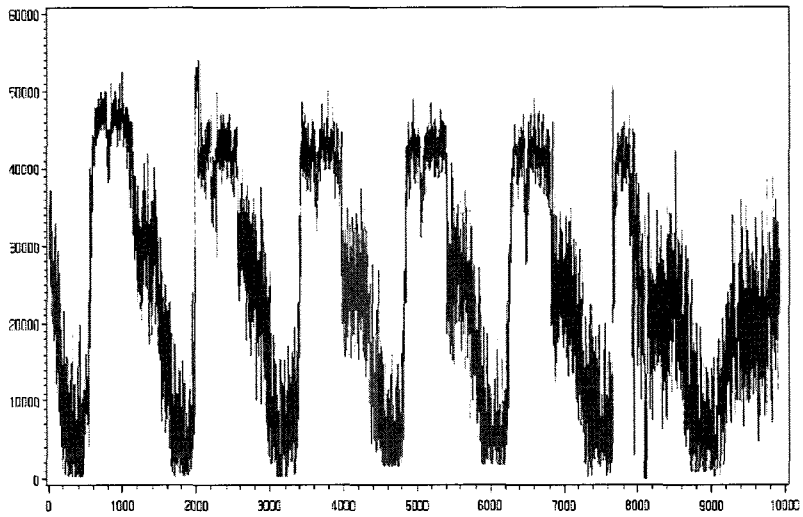

그림 3.3: 자료 2의 원자료

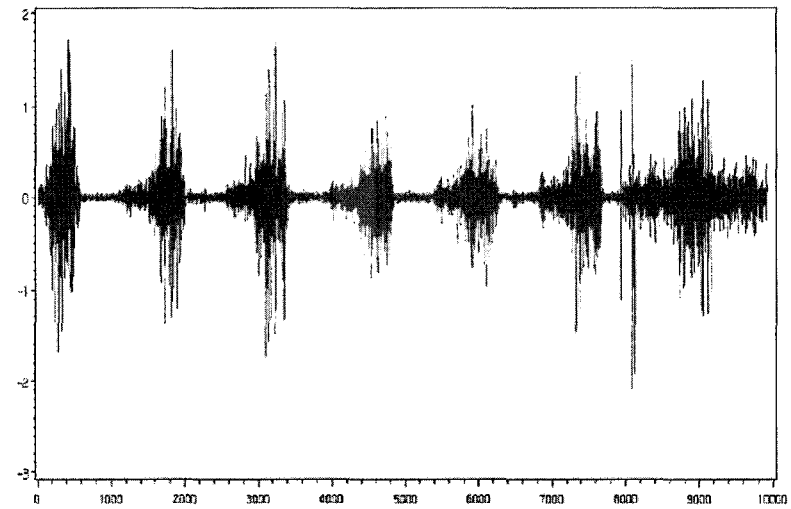

그림 3.4: 자료 2의 로그 변환 후 차분한 자료

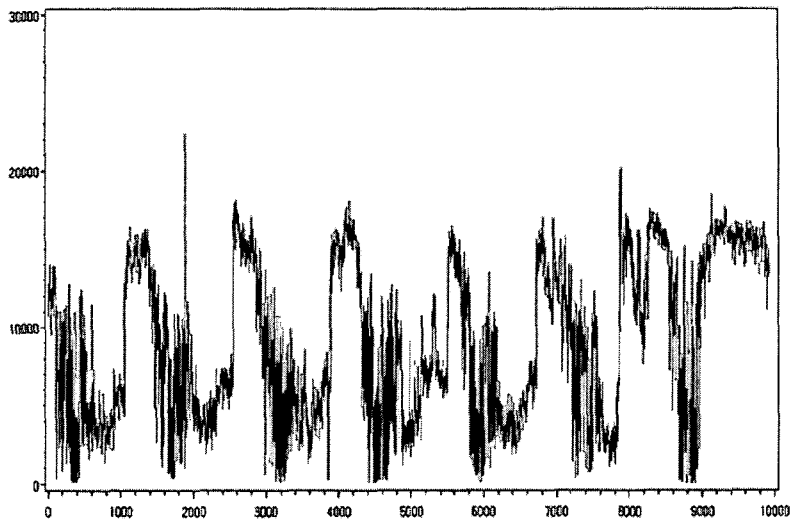

그림 3.5: 자료 3의 원자료

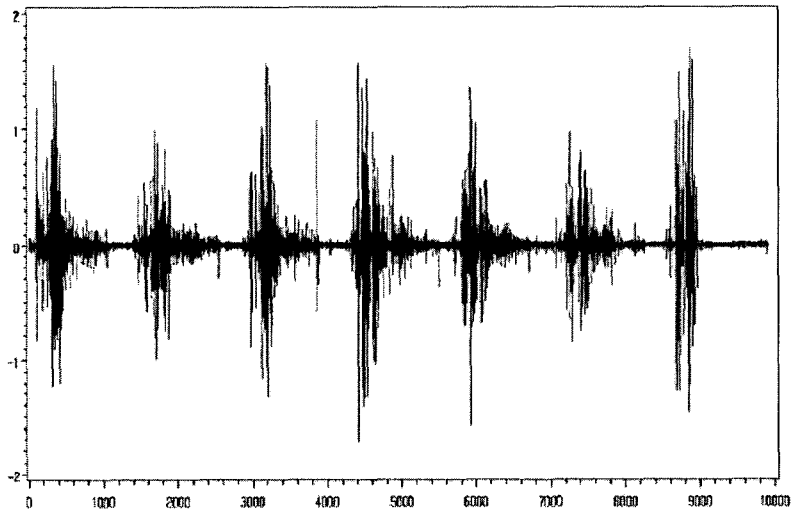

그림 3.6: 자료 3의 로그 변환 후 차분한 자료

표 3.1: 각 모형의 모수의 추정치

	AR(2)					AR(2)-ARCH(1)				
	ϕ_1	ϕ_2	α_0	α_1	β_1	ϕ_1	ϕ_2	α_0	α_1	β_1
자료1	−0.3692	−0.2468				−0.6513	−0.3111	0.0160	0.8908	
자료2	−0.3641	−0.2578				−0.5570	−0.4243	0.0097	1.6476	
자료3	−0.1977	−0.2048				−0.7239	−0.2394	0.0035	3.2108	
	AR(2)-GARCH(1,1)					AR(2)-IGARCH(1,1)				
	ϕ_1	ϕ_2	α_0	α_1	β_1	ϕ_1	ϕ_2	α_0	α_1	β_1
자료1	−0.6713	−0.3063	0.0000	0.1068	0.8625	−0.6727	−0.3042	0.0000	0.0838	0.9162
자료2	−0.6343	−0.3112	0.0000	0.1077	0.8507	−0.6373	−0.2984	0.0000	0.1161	0.8839
자료3	−0.7747	−0.2094	0.0000	0.1175	0.8631	−0.7792	−0.2037	0.0000	0.0789	0.9211

표 3.2: RMSE 값

	ARIMA(2,1,1)	AR(2)-ARCH(1)	AR(2)-GARCH(1,1)	AR(2)-IGARCH(1,1)
자료1	0.3141	0.3189	0.3117	0.3115
자료2	0.1946	0.1499	0.1464	0.1463
자료3	0.0332	0.0042	0.0034	0.0034

있어서 다른 비선형모형보다 우수하지 못함을 알 수 있다. 다시 말하면 오차의 등분산 가정이 현실적이지 못하며 따라서 이분산성을 가정한 모형이 예측의 정확도가 더 높다고 할 수 있다. 이분산성 모형에서는 ARCH모형보다 GARCH 모형이나 IGARCH 모형이 예측의 정확도가 RMSE 기준으로 볼 때 우수하다고 할 수 있는데 GARCH 모형의 특성상 ARCH 모형보다 더 많은 모수가 필요하고 따라서 ARCH 모형보다 설명력이 높음을 쉽게 짐작 할 수 있다. 아울러 $\alpha_1 + \beta_1$의 값이 모두 1에 가깝게 나오고 있는데 IGARCH 모형이 더 우수함을 보여주고 있다.

4. 결론

본 연구에서는 최근에 많은 관심을 받고 있는 통신망 트래픽 자료의 예측을 위하여 기존의 선형 시계열모형과 더불어 ARCH, GARCH, IGARCH 모형을 소개하고 모형의 예측의 성능을 실제자료를 통하여 비교하여 보았다. Feng 등(2001)은 ARMAX/GARCH 모형이 기존의 multi-fractal wavelet 모형보다 예측의 정확도가 높음을 보였는데 본 연구에서는 기존의 선형시계열 모형보다 GARCH 모형이나 IGARCH 모형이 예측에 있어서 우수함을 보였다. 향후에 threshold ARCH 모형과 같은 더욱 정교한 모형을 가지고 예측의 정확도를 따지는 성능평가가 요구되어진다.

참고문헌

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics*, **31**, 307–327.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control*, Prentice-Hall, Englewood Cliffs, NJ.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50**, 987–1007.

Feng, C., He, D. and Sun, Z. (2004). *IP traffic trace modelling with ARMAX/GARCH, HET-NET* 03, Ilkley, UK, 26–28.

Hellerstein, J. L., Zhang, F. and Shahabuddin, P. (2001). A statistical approach to predictive detection, *Computer Networks*, **35**, 77–95.

Kim, S., Yun, Y. B. and Park, E. G. (2005). A study of statistical approach for detection of outliers in network traffic, *Journal of Korean Data & Information Science Society*, **16**, 979–987.

Shu, Y., Yu, M., Yang, O., Liu, J. and Feng, H. (2005). Wireless traffic modeling and prediction using seasonal ARIMA models, *IEICE Transactions on Communications*, E88-B, 3992–3999.

[2007년 2월 접수, 2007년 3월 채택]

Time Series Models for Performance Evaluation of Network Traffic Forecasting*

S. Kim[1)]

ABSTRACT

The time series models have been used to analyze and predict the network traffic. In this paper, we compare the performance of the time series models for prediction of network traffic. The feasibility study showed that a class of nonlinear time series models can be outperformed than the linear time series models to predict the network traffic.

Keywords: Time series models, network traffic, forecasting.

* This research was supported by the Chung-Ang University Research Grants in 2006.

1) (156-756) Associate Professor, Department of Statistics, Chung-Ang University, Seoul 156-756, Korea
E-mail: sahm@cau.ac.kr