Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 979~987

A Study of Statistical Approach for Detection of Outliers in Network Traffic

Sahmyeong Kim[1] · Joo-Beom Yun[2] · Eung-ki Park[3]

Abstract

In this research we study conventional and new statistical methods to analyse and detect outliers in network traffic and we apply the nonlinear time series model to make better performance of detecting abnormal traffic rather the linear time series model to compare the performances of the two models.

Keywords : Network Traffic, Outliers, Time Series Models

1. 서 론

통신망을 운용하거나 설계하는 입장에서 보면 최근의 급증하고 있는 네트워크 상에서의 이상 트래픽은 망의 안정성과 신뢰성에 치명적인 문제를 일으키는 대상으로 인식되어 오고 있다. 네트워크에 비정상적으로 입력되는 이상 트래픽은 여러 가지 특성이 있으나 첫째로 이상 과다 트래픽을 탐지했다 할지라도 이를 처리하고 해결하는 상황대처를 위한 시간이 별로 없다는 것이다. 이것은 트래픽을 실시간으로 모니터링하고 이상 트래픽을 거의 실시간에 준하는 시간에 처리해야 하는 것을 의미한다.

현재까지 이상 트래픽 탐지에 대한 연구는 대략 다음과 같다. 첫째로 Silcora(1992)는 통계적 공정관리 기법(Statistical Process Control)을 이용하여 공정에 이상 원인이 발생하면 이를 즉시 탐지하여 이상신호(Signal)을 주는 방식을 택하여 이상트래픽을 탐지하고 관리하였다. 이를 통하여 네트워크 상에 이상과다 트래픽이 발생하면 이를 탐지하여 빠른 시간안에 이를 안겨주도록 하였다. 여기서는 주로 관리도(Control

1) First Author : Associate Professor, Department of Statistics, Chung-Ang University, Seoul, 156-756, Korea
E-mail : sahm@cau.ac.kr
2) Member of Engineering Staff, NSRI, 463-1, Jeonmin-dong, Yuseong-gu Daejeon, 305-811, Korea
E-mail : netair@etri.re.kr
3) Principal Member of Engineering Staff, NSRI, 463-1, Jeonmin-dong, Yuseong-gu Daejeon, 305-811, Korea
EMAIL : ekpark@etri.re.kr

Chart)를 이용하였으며 관리도는 형태에 따라 Shewart 관리도, CUSUM 관리도, EWMA 관리도 등이 있다. 그러나 통계적 관리도 기법은 발생 가능한 이상트래픽의 탐지 및 예측에 문제가 되는 것으로 알려져 있다.

Dietherich et.al(1985)는 데이터 마이닝 기법을 이용하여 네트워크 트래픽의 전반적 패턴을 인식, 분류하여 이상과다 트래픽이 발생하는 모형을 구축하여 이를 탐지할 수 있도록 하였다. 이러한 데이터 마이닝 기법을 이용한 패턴의 분석은 범주형 자료에 대해서는 적합하지만 연속적 자료에는 부적합한 것으로 알려져 있다. 한편 Jiao et.al(1999)는 네트워크 이상 트래픽 탐지를 위해 발생되는 모니터링 비용을 최소화 할 수 있는 방법을 고려한 기법으로 이상트래픽 탐지 기법을 연구 하였고 Kozma et.al(1994)는 신경망(Neural Network)을 이용한 탐지 기법을 연구 하였다. 이 기법은 예측이 주목적이 될 수 없고 예측의 정확도가 떨어지는 단점이 있다고 알려져 있다. 한편 Hellerstein et.al(2001)은 예방적 탐지 기법을 제안하여 이상 트래픽 탐지에 적용하였다. 이 기법도 자기 시계열 모형(AR)을 근간으로 하여 서비스 제공자에게 문제를 진단 할 수 있게 하고 국지적인 문제가 더 튼 문제로 확대되기 전에 조치를 취할 수 있게 하고 나아가 서비스 불능 상태를 방지하기 위한 조치에 따르는 시간을 벌어주는 이점이 있다. 결과적으로 서비스 제공자와 고객관의 관계를 증진시키는 장점이 있는 것이다.

본 논문에서는 이상 트래픽 탐지를 위한 선형 시계열 모형을 소개하고 이분산성을 설명하기 위하여 ARCH(Autoregressive Conditional Heteroscadastic) 모형을 소개하여 실제자료를 이용하여 Hellersteion(2001)이 사용한 선형 시계열 모형과 ARCH 모형과의 성능 평가를 통하여 ARCH 모형의 우수성을 보이고자 한다.

2. 선형시계열모형을 이용한 탐지기법

Hellerstein(2001)은 각 시점에서 수집된 트래픽 데이터를 다음과 같이 월별, 주별, 일별 효과로 분해하여 그에 따르는 오차 항을 이용한 이상치 탐지를 제안하였다.

$$S_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + y_{ijkl}$$

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0$$

여기서 μ는 총평균, α_i는 일별 효과, β_j는 주별 효과, γ_k는 월별효과, y_{ijkl}는 오차를 나타낸다.

각 관측시점에서의 데이터를 위와 같이 분해하여 오차항 y_{ijkl}을 y_t로 표기하고 y_t에 대한 모형을 구하였다. Hellerstein은 y_t에 대해 시계열모형을 적용하여 아래와 같이 비정상시계열모형 AR(2)을 선택하였다.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t$$

여기서 u_t는 $iid\ N(0, \sigma_u^2)$.

위에서 선택된 모형을 이용하여 모수값(ϕ_1, ϕ_2)을 추정하여 예측모형을 구한다. 즉, 현시점 t에서 한 시점 후의 예측값 $\hat{y}_t(1)$은 $\hat{y}_t(1) = \hat{\phi}_1 y_t + \hat{\phi}_2 y_{t-1}$이 되고 h시점 후의 예측값 $\hat{y}_t(h)$을 다음과 같이 구하고 그 때의 이상치 유무에 대한 확률 또한 구하게 된다.

$$\hat{y}_t(h) = E\ [\ Y_{t+h} | Y_t = y_t\ ,\ Y_{t-1} = y_{t-1}\ ]$$

$$\hat{P}_t(h) = P[\ Y_{t+h}\text{에서의 이상치}\ |\ Y_t = y_t,\ Y_{t-1} = y_{t-1}\ ]$$

따라서 위에서 계산된 $\hat{P}_t(h)$를 이용하여 h시점 후에 이상과다 트래픽이 발생 유무를 판단하는 방법을 제안하였다.

그러나 Hellestein이 제시한 방법은 아직 일반화에 어려움이 있고 또한 모형에 사용된 u_t의 경우 실제로 동일 독립 정규분포를 따른다고 하기 어려움이 있다. 따라서 등분산시계열모형을 적합하여 그 값을 예측하고 이상치 유무에 대한 결정을 내리는 것보다는 조건부이분산모형을 적합하는 것이 타당할 것으로 판단된다.

3. 시계열 모형

이 장에서는 이상 트래픽 탐지를 위해 이용될 수 있는 시계열 모형 중 정상시계열모형인 AR모형과 비선형시계열모형인 ARCH모형에 대해 간략히 설명한다.

3.1 AR모형

AR모형은 정상시계열모형인 ARMA(p, q)의 특수한 경우로 우선 AR(1)의 과정을 살펴보면 다음과 같은 구조식을 갖는 모형이다.

$$z_t = \phi z_{t-1} + e_t \tag{3.1}$$

여기서 e_t는 평균이 0, 분산이 σ^2인 백색잡음과정이고 모든 $k > 0$에 대하여 $Cov(e_t, z_{t-k}) = 0$이 된다.

식(3.1)에서 정의된 AR(1) 과정은 z_t를 설명하는데 가장 최근의 값인 z_{t-1}이 가장 많은 정보를 가지고 있고 과거로 갈수록 정보의 양이 줄어드는 시계열자료에 적합한 모형이다.

식(3.1)에서 정의된 AR(1) 과정을 일반화시킨 것이 차수 p를 가지고 있는 자기회귀과정으로 AR(p)로 표기한다. 즉, z_t를 설명하는데 $z_{t-1}, z_{t-2}, \cdots, z_{t-p}$는 어느 정

도 정보를 가지고 있고 z_{t-p}이전의 값은 지수적으로 감소하는 정보를 가지는 시계열에 적합한 모형이다. AR(p)의 수학적 모형은 다음과 같다.

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + e_t \tag{3.2}$$

여기서 e_t는 평균이 0, 분산이 σ^2인 백색잡음이다.

3.2 조건부 이분산 자기회귀모형 - (Autoregressive Conditional Heteroscadestic Model : ARCH)

위에서 살펴본 AR모형은 모두 오차의 분산이 동일하다는 오차등분산 가정에서 생성되고 적합되는 모형이다. 하지만 인터넷 이상 트래픽과 같이 일정 부분에 있어 등분산이 성립되지 않는 모형에 포함된 오차의 분산이 동일하지 않는 경우인 비선형시계열 모형 중 ARCH모형에 대해 살펴보자

ARCH모형은 오차의 분산이 자기회귀적으로 변하는 이분산성 모형으로 1982년 Engel에 의해 처음 제시되었다.

e_t가 ARCH(p) 모형을 따른다고 하면

$$e_t = \sqrt{h_t}\varepsilon_{t,} \quad t = 1, 2, \cdots, n, \quad \varepsilon_t \sim iid\ N(0,1)$$
$$h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j a^2_{j-1}$$

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

4. 시뮬레이션

4.1 시뮬레이션 모형

많은 선행연구의 결과로 인터넷 이상 트래픽 탐색에 AR모형이 많이 이용되었다. 하지만 앞서 말한 것과 같이 이상 트래픽 데이터는 그 특성상 오차의 등분산성을 만족한다고 하기 어려우므로 AR모형 보다는 ARCH모형을 선택하는 것이 타당할 것으로 생각된다.

AR모형과 ARCH모형에서 이상치가 존재하는 경우, 모형의 예측도(정확도)를 확인해보고자 2가지 상황에 대한 시뮬레이션을 실시하였다. 시뮬레이션은 각각의 모형, 즉 AR모형과 ARCH모형에 의해 생성된 자료에 인위적으로 이상치를 생성하여 두 모형의 효율을 비교하였다. 그 모형은 다음과 같다.

(1) AR(2)모형

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + e_t$$

여기서 e_t는 $iid\ N(0,1)$이다.

(2) AR(2)-ARCH(1)모형

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$$

$$\varepsilon_t = \sqrt{h_t} e_t$$

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$$

여기서 e_t는 $iid\ N(0,1)$이다.

위의 두 모형을 이용하여 250개의 관측치를 생성한 후 임의적으로 4지점, 즉 50번째, 100번째, 150번째, 200번째 데이터에 이상치를 부여하였다. 이상치는 각 시점에서 생성데이터에 0.5, 0.1, 0.2, 0.3의 이상치를 부여하였다.

4.2 시뮬레이션 결과

각 모형에 의해 생성된 데이터를 다시 각각의 모형에 적합시켜 예측모형을 구한 후 이상치를 부여한 시점에서의 MSE를 구한 결과가 다음과 같다.

(1) AR(2) 모형에서 생성된 데이터의 경우

	t=50	t=100	t=150	t=200
AR(2)	1.9428	3.4120	7.5142	14.012
AR(2)-ARCH(1)	1.9363	3.4077	7.4987	14.012
ratio	1.0034	1.0013	1.0021	1.0001

(2) AR(2)-ARCH(1) 생성된 데이터의 경우

	t=50	t=100	t=150	t=200
AR(2)	0.7703	3.7812	5.0122	8.9950
AR(2)-ARCH(1)	0.7090	3.7778	5.0044	8.7818
ratio	1.0865	1.0009	1.0016	1.0243

위의 결과로 볼 때 데이터가 AR(2)모형이나 AR(2)-ARCH(1)모형에서 생성됨에 관계없이 이상치가 존재하는 경우 AR(2)-ARCH(1)모형으로 데이터를 적합하여 예측하

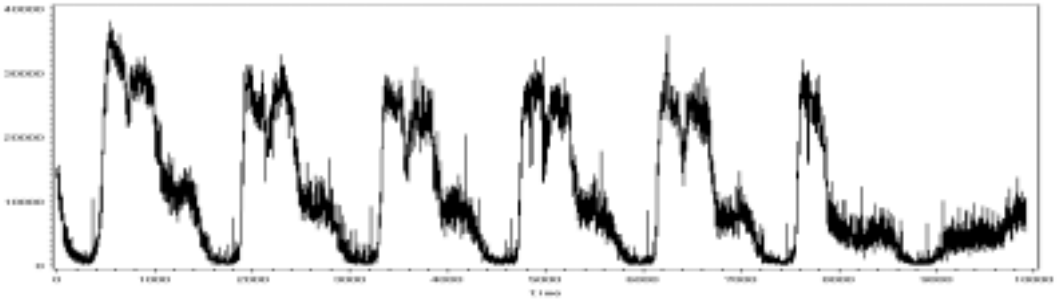
는 것이 더 효과적이라고 할 수 있다.

5. 데이터를 이용한 모형의 선택

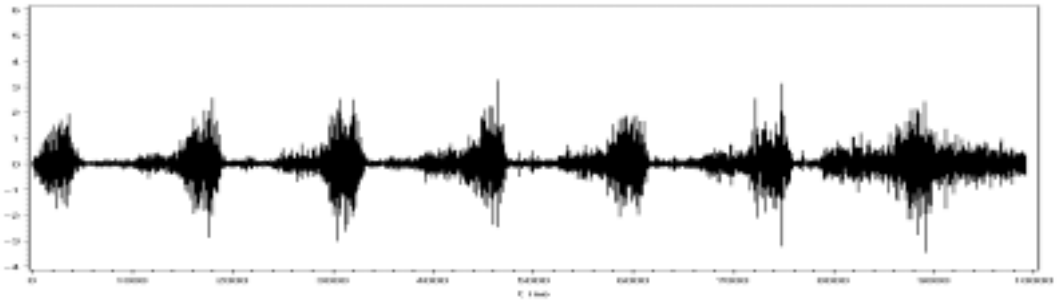
앞서 살펴본 시뮬레이션 결과를 토대로 하여 실제로 인터넷 트래픽 데이터에 존재하는 과다 이상 트래픽의 탐지에 어느 모형이 더 효율적인가를 판단하기 위하여 실제 데이터를 이용하여 두 모형을 적합시켜 그 정확도를 확인해 보았다.

관측시점이 9918개인 10개의 데이터 셋의 원 데이터를 로그(log)변환 후 1차 차분하여 정상시계열 모형으로 변환하여 3000, 6000, 9000 시점에 임의로 이상트래픽을 발생시킨 데이터를 이용하였다.

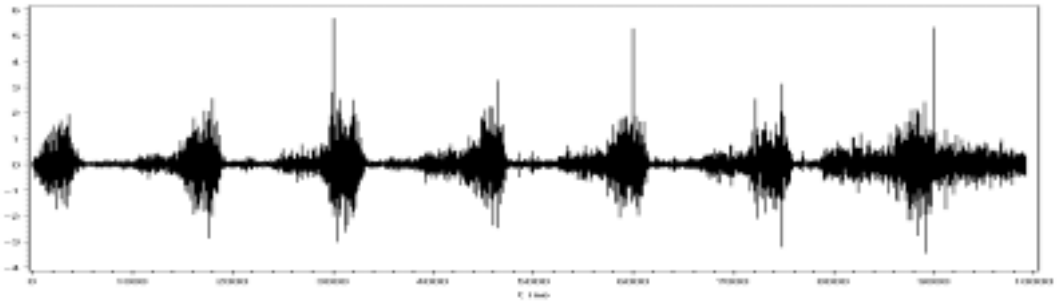
다음의 <그림 1>에서 <그림 5>는 Case2의 데이터의 분석과정에서 나타난 데이터의 형태이다.



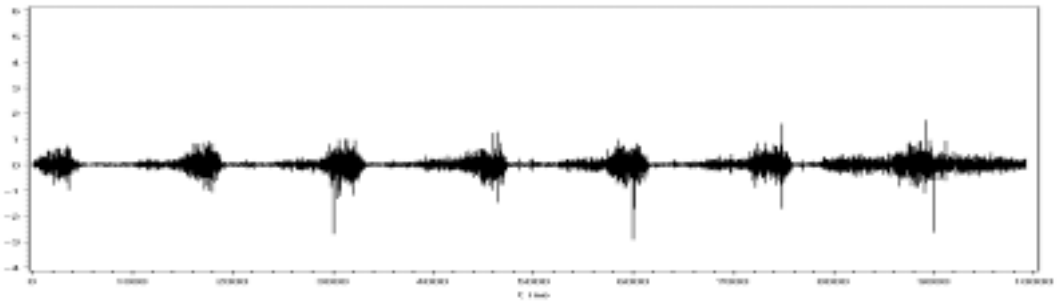
<그림 1> Case 2의 원 데이터



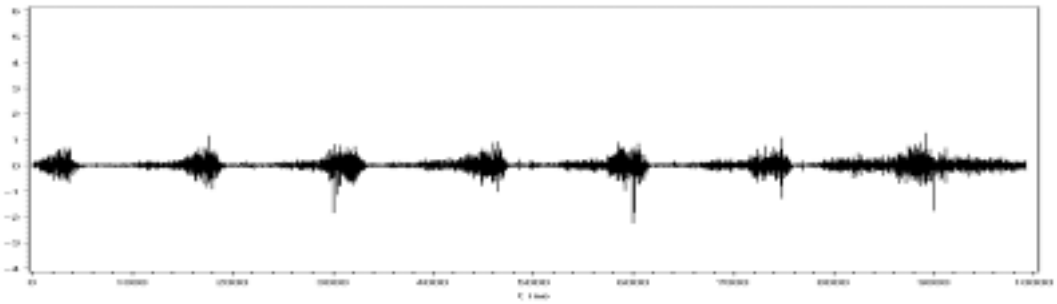
<그림 2> 원 데이터의 로그변환 후 1차 차분 데이터



<그림 3> 차분 데이터에 이상치 부여



<그림 4> AR(2) 모형 적합 후 예측값



<그림 5> AR(2)-ARCH(1) 모형 적합 후 예측값

앞선 설명과 같이 원 데이터를 변형하여 두 가지 모형에 적합시킨 후 임의로 부여한 이상트래픽의 예측력을 살펴보기 위하여 실제값과 모형을 통한 예측값의 차이의 정도를 알 수 있는 MSE와 두 모형으로부터 계산된 MSE의 비율은 다음과 같다.

(1) 이상 트래픽의 양이 5인 경우

No. Data Set	MSE		Ratio
	AR(2)	AR(2)-ARCH(1)	
1	21.302018	21.275085	1.0012659
2	33.830584	32.896721	1.0283877
3	21.512406	21.848437	0.9846199
4	23.280837	23.279723	1.0000479
5	29.170132	29.027547	1.0049121
6	25.082708	25.645947	0.9780379
7	17.927047	17.641131	1.0162074
8	19.610271	19.30427	1.0158515
9	24.635742	24.650386	0.999406
10	25.13943	24.851046	1.0116045

(2) 이상 트래픽의 양이 6인 경우

No. Data Set	MSE		Ratio
	AR(2)	AR(2)-ARCH(1)	
1	31.445457	31.416881	1.0009096
2	46.301205	45.35115	1.0209489
3	31.758127	32.092271	0.9895881
4	33.908633	33.85098	1.0017032
5	40.922013	40.808372	1.0027847
6	36.095452	36.925742	0.9775146
7	27.029703	26.671396	1.0134341
8	29.425506	29.034064	1.0134822
9	35.554654	35.584682	0.9991562
10	36.211812	35.72818	1.0135364

위의 결과로 미루어 볼 때, 이상 트래픽이 존재하는 데이터의 경우 AR(2)모형 보다는 AR(2)-ARCH(1)모형의 예측력이 좀 더 높다는 것을 확인할 수 있다. 따라서 ARCH모형을 선택하는 것이 이상 과다 트래픽의 탐지를 위한 시계열모형의 선택에 있어 더 타당한 것임을 알 수 있다.

6. 결 론

시계열자료에서의 이상치(outliers)에 대한 연구는 주로 재무나 경제에 관련된 분야에서 활발히 진행되어 있으며 최근에 통신 트래픽 자료의 이상치를 탐색하고 분석하는 분야로 확장이 되고 있다. 최근에 선형자귀회귀모형을 이용한 이상트래픽을 탐지하는 기법이 소개되었고 이것을 비선형시계열 모형으로 확장하여 이상치를 탐지하는

기법을 본 논문에서 언급하였다. 비선형 시계열 모형을 고려한 주된 이유는 선형 시계열 모형에서 가정하는 오차의 등분산성이 비현실적이며 따라서 오차의 이분산성을 가정하는 일련의 비선형 시계열 모형이 더 효과적으로 이상치 트래픽을 탐지할 수 있다는 가능성을 본 연구에서 예시하고 있다. 향후 보다 정교하고 복잡한 비선형 시계열 모형을 적용하여 다양한 트래픽 자료에 적합 시켜 이상치 탐지의 효율성을 높일 수 있을 것이라 사료된다.

참고문헌

1. Box, G.E.P. G.M.Jenkins, (1976) *Time Series Analysis Forecasting and Control*, Prentice_Hall, Englewood Cliffs, NJ,
2. Dietterich, T. G., Michalski, R. S., (1986) Discovering patterns in sequences of events, *Artificial Intelligence* 25, 187-231.
3. Dijk, D. V., Franses, P. H. and Lucas, A., (1999) Testing for ARCH in the presence of Additive Outliers, *Journal of Applied Econometrics* 14, 539-562.
4. Engle, R. F., (1982) Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U. K. Inflation, *Econometrica* 50, 987-1008.
5. Hellerstein, Josepl, L., Zhang, F., Shahabuddin, P., (2001) A statistical approach to predictive detection. *Computer Networks.* 77-95.
6. Jia, J., et al., (1999) Minimizing the monitoring cost in network management, in: *Integrated Network Management* VIIFIP, 150-170.
7. Kozma, R., Kitamura, M., Sakuma, M., Yokoyama, Y., (1994) Anomaly detection by neural network models and statistical time series analysis. *In Proceedings of the IEEE International Conference on Neural Network*, 2307-3210.
8. Sikora, W. I., (1992) Responce time measurement and SPC. *Computer Measurement Group Transaction*, 35-42
9. Zhang, B., May, G., (1998) Towards real-time fault identification in plasma eteching using neural networks, in: *Proceedings of the 1998 Artificial Networks in Engineering Conference*, vol, 8, 803-810.

[2005년 10월 접수, 2005년 11월 채택]