효율적인 통신망 관리를 위한 기계학습 알고리즘을 적용한 트래픽 발생량 예측

지세현, Huru Hasanova, 심규석, 김명섭
고려대학교
{sxzer, hhuru, kusuk007, tmskim}@korea.ac.kr

Prediction of Traffic Usage Using Machine Learning Algorithm For Efficient Network Management

Se-Hyun Ji, Huru Hasanova, Kyu-Seok Shim, Myung-Sup Kim
Korea Univ.

요 약

오늘날 인터넷의 발달함에 따라 네트워크 트래픽 사용량이 증가되고 있다. 트래픽 사용량이 증가됨에 따라 망의 혼잡이나 장애가 발생하는 문제점이 있다. 폭주하는 트래픽 문제를 해결하기 위해 네트워크 회선을 늘리는 것은 많은 시간과 비용이 소모되며, 효율적인 망 관리를 할 수 없다. 적합한 QoS(Quality of Service)를 제공하고 안전한 네트워크 환경을 만들기 위해서 효율적인 통신망 관리가 요구되고 있고, 통신망 관리에 있어 트래픽 발생량을 예측하는 것은 필수적인 요소이고, 이에 관한 연구가 활발하게 진행되고 있다. 본 논문에서는 통신망 관리에 효율적으로 사용 될 수 있는 기계학습(Machine Learning) 알고리즘인 순환신경망(Recurrent Neural Network) 알고리즘이 적용된 트래픽 발생량 예측 모델을 제안한다. 제안하는 모델은 학내 망에서 발생하는 트래픽 중 TCP 또는 UDP프로토콜에 관한 트래픽 발생량을 수집하여 이를 바탕으로 3가지 항목에 관하여 다음의 발생량을 예측한다. 예측 모델이 예측한 값과 실제 값의 차이를 평균 제곱근 오차(Root Mean Square Error)값을 통해 정밀도(Precision)를 확인함으로써 모델의 적합성을 검증한다.

Ⅰ. 서 론

오늘날 인터넷이 발달하고 있고, 그에 따라 네트워크 트래픽 사용량이 증가하고 있다. 트래픽 사용량이 증가됨에 따라 트래픽이 과도하게 급증하여 망의 혼잡이나 장애가 발생했거나 우려되는 상황이 발생하고 있다. 이러한 상황이 발생했을 때 적절한 조치를 취하지 않으면 통신망에 심각한 장애가 발생할 수도 있다. 폭주하는 트래픽 문제를 해결할 수 있는 가장 단순한 방법은 네트워크 회선을 늘리는 것이다. 그러나 단순히 네트워크 회선을 늘리는 작업은 많은 시간과 비용이 소모되고, 효율적인 망 관리를 할 수 없다. 따라서 적합한 QoS를 제공하고 안전한 네트워크 환경을 만들기 위해서 효율적인 통신망 관리를 위한 방안이 요구되고 있다. 이러한 통신망 관리에 있어 트래픽 발생량을 예측하는 것은 필수적인 요소이고 예측 결과를 이용하여 동적인 트래픽 관리가 가능하게 된다.

기존의 트래픽 발생량 예측 연구 중 통계학 분야에서 예측 방법으로 널리 쓰이는 시계열 모형인 AR, MA, ARMA, ARIMA 모형을 사용한 트래픽 예측 모델이 있다[1]. 시계열 분석은 과거 시계열 자료의 패턴이 미래에도 지속적으로 유지된다는 가정 하에서 예측을 수행하는 것이므로 가정을 만족할 경우에 한해서 정확한 예측을 수행 할 수 있기 때문에 제한적이다. 이에 본 논문에서는 머신러닝(Machine Learning) 알고리즘인 순환신경망 알고리즘을 설명하고, 트래픽 발생량 예측 모델을 제안한다.

본 논문에서는 서론에 이어 2장에서 Google에서 만든 오픈소스 라이브러리인 Tensorflow에 있는 기계학습 알고리즘인 순환신경망 알고리즘이 적용된 트래픽 발생량 예측 모델을 제안한다. 제안하는 모델은 학내 망에서 발생하는 트래픽 중 TCP 또는 UDP프로토콜에 관한 트래픽 발생량을 수집하여 3가지 항목(Flow, Byte, Pkt)의 사용량을 바탕으로 그 다음 트래픽 발생량을 예측하는 모델에 대해 언급한 뒤 3장에서 만들어진 모델이 예측한 값과 실제 값의 차이를 평균 제곱근 오차(Root Mean Square Error)값을 통해 정밀도를 확인한다. 예측 모델의 정밀도를 확인함으로써 모델의 적합성을 검증한다. 마지막으로 4장에서 결론 및 향후 연구에 대해 언급한 뒤 논문을 마친다.

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원(No.2015R1D1A3A01018057) 및 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원(No. 2017-0-00513)을 받아 수행된 연구임.

Ⅱ. 본론

본 장에서는 제안하는 트래픽 발생량을 예측하는 모델에 대해 설명한다. 트래픽 발생량 예측 모델은 그림 1 과 같이 구성된다. 학내 망에서 발생한 트래픽 중 TCP 또는 UDP프로토콜에 관한 트래픽 발생량을 본 연구팀이 개발한 프로그램인 Flow_Twoway checker를 이용해 3가지 항목(Flow, Byte , Pkt)에 대해 10분 단위로 수집한다. 수집 된 트래픽 발생량을 각각의 순환신경망 알고리즘이 적용된 모델에 학습을 시켜 트래픽 발생량 예측 모델을 완성한다.

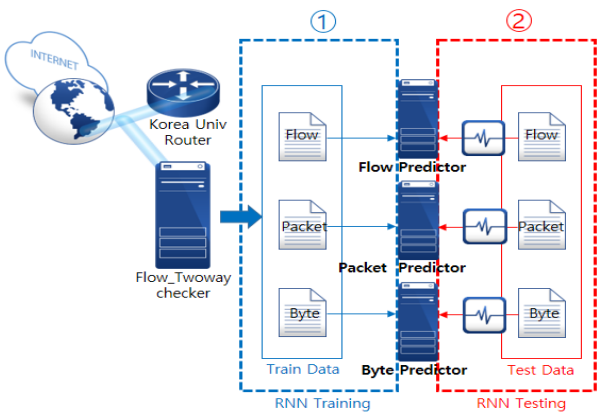


그림 1. 트래픽 발생량 예측 모델

순차적인 데이터를 학습하는데 특화된 순환 신경망의 구조는 그림 2와 같이 구성된다. 시계열(Time Series)로 구성된 트래픽 사용량을 입력 값(Xt), 그리고 임의의 Sequential Length를 설정하면, Sequential Length 이후의 값이 출력 값(ht)으로 나온다[2]. 이 때 순환신경망은 학습에서 나온 과거의 가중치를 기억하여 현재 학습에 반영한다.

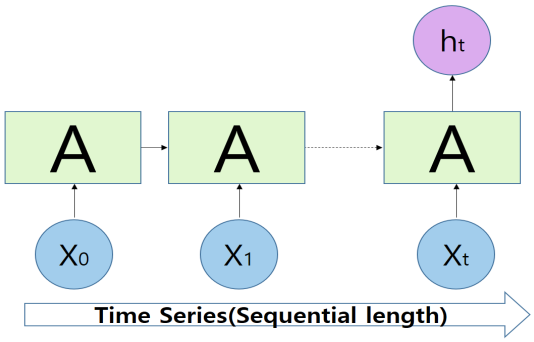


그림 2. 순환신경망 구조

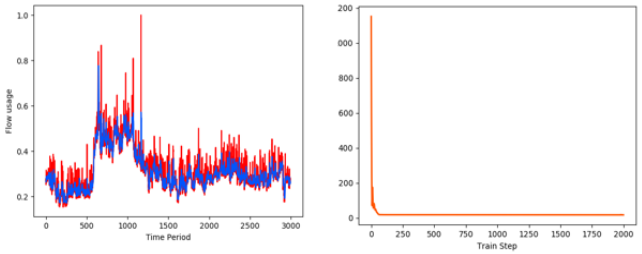
순환신경망으로 이루어진 예측 모델이 예측 한 값과 실제 데이터 값의 차이를 평균 제곱근 오차(Root Mean Square Error)를 통해 정밀도를 표현함으로써 모델의 적합성을 검증한다. 아래의 수식은 평균 제곱근 오차를 계산하는 수식이다[3].

$$RMSE = \sqrt{MSE(\hat{\theta})} = \sqrt{E((\hat{\theta}-\theta)^2} \quad (1)$$

추정량$\hat{\theta}$과 추정모수θ와의 차이는 $\hat{\theta}$이 θ에 얼마나 근접해있는지 나타내 줄 수 있으므로, 이 차이 $\hat{\theta}-\theta$를 밀집성의 정도를 나타내는 지표로 사용할 수 있다. 그러나 $\hat{\theta}-\theta$는 양과 음의 값을 모두 취하게 되므로 $\hat{\theta}-\theta$ 대신 차이의 제곱인 $(\hat{\theta}-\theta)^2$을 지표로 사용해야 한다. 추정량 $\hat{\theta}$은 일종의 확률변수이므로 $(\hat{\theta}-\theta)^2$도 확률 변수가 되기 때문에 이것의 평균(E)을 구할 수 있으며, 그 결과를 평균 제곱 오차(MSE)라 하고 평균 제곱 오차에 제곱근을 취해준 결과가 평균 제곱근 오차(RMSE)이다. 평균 제곱근 오차가 작을수록 정밀도가 높다.

Ⅲ. 실험

기계 학습 기반의 트래픽 발생량 예측 모델의 적합성을 검증하기 위해 실험을 진행한다. 2017년 1월 1일부터 2017년 3월 11일까지 학내 망에서 발생한 트래픽을 수집하여, Flow_Twoway Checker를 이용해 10분 단위로 발생한 트래픽 사용량을 3가지(Flow, Packet, Byte)항목별로 구분 하여 연속성이 있는 10000개의 데이터를 구성하였다. 이 중 7000개의 데이터를 순환신경망 모델을 학습시키기 위한 Train Data로 사용하고, 나머지 3000개의 데이터를 예측모델의 검증을 위한 Test Data로 사용하였다. 3000개의 Test Data에 대한 예측 결과와 순환신경망 모델의 학습 결과를 그림 3으로 나타내었다.



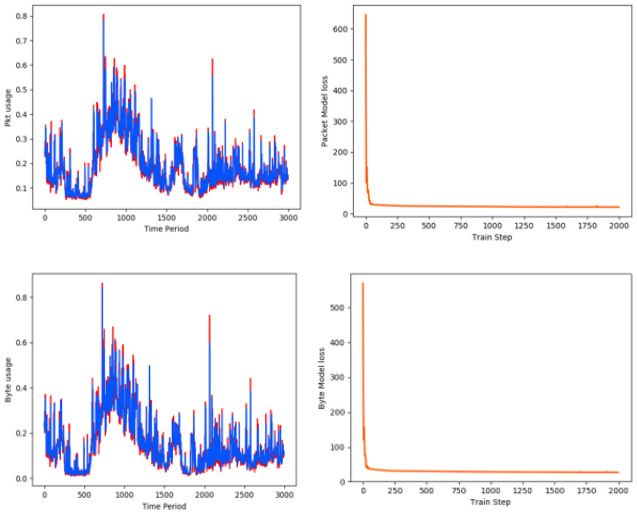


그림 3. 트래픽 발생량 예측 모델 그래프

그림 3의 좌측 그래프에서 적색은 트래픽 발생 실제 값, 청색은 예측 값으로 구분하였다. 우측 그래프에서는 순환신경망 모델의 학습을 횟수에 대한 loss(학습 손실)를 나타내었다. loss는 학습이 얼마나 잘 되었는지 판단 할 수 있는 수치이다. loss가 낮을수록 학습이 잘 이루어 졌다고 판단한다.

순환신경망 트래픽 발생량 예측 모델의 loss와 RMSE값은 표 1과 같다.

표 1. 실험 결과

CATEGORY	LOSS	RMSE
Flow	16.682	0.0405
Packet	21.457	0.0412
Byte	26.917	0.0481

표 1에서 학습이 가장 잘 된 순환신경망 예측 모델은 loss값이 가장 적은 Flow 예측 모델이고 가장 정밀도가 높은 모델도 RMSE의 값이 가장 작은 Flow 예측 모델이라는 결과가 나왔다. 그러나 나머지 두 모델도 높은 정밀도를 나타내고 있음을 확인하였다.

Ⅳ. 결론 및 향후연구

본 논문은 기계 학습 알고리즘을 적용한 트래픽 사용량 예측 모델을 제안 하였다. 실험을 통해 제안한 모델의 성능을 정밀도로 나타냄으로써 확인하였다. 제안한 모델은 동적인 트래픽 관리에 있어 통신망 관리의 효율성을 향상 시킬 수 있다.

향후연구에서는 순환신경망 알고리즘 뿐 만 아니라 다양한 기계학습 알고리즘을 적용하여 보다 정확하고 정교한 검증 실험을 진행할 계획이다.

참 고 문 헌

[1] 정상준, 김동주, 권영헌, 김종근. (2004). 네트워크 트래픽 예측을 위한 시계열 모형의 적합성 검증, 한국통신학회논문지, 29(2B), 217-227.

[2] Christopher Olah. August 27, 2015. "Understanding LSTM Networks", http://colah.github.io/posts/2015-08-Understanding-LSTMs. (2017-12-12).

[3] WIKIPEDIA. December 9, 2015. "Root-mean-square deviation", https://en.wikipedia.org/wiki/Root-mean-square_deviation. (2017-12-12).