KIPS Trans. Softw. and Data Eng.
Vol.10, No.10 pp.391~398
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.10.391

Leased Line Traffic Prediction Using a Recurrent Deep Neural Network Model

In-Gyu Lee† · Mi-Hwa Song††

ABSTRACT

Since the leased line is a structure that exclusively uses two connected areas for data transmission, a stable quality level and security are ensured, and despite the rapid increase in the number of switched lines, it is a line method that is continuously used a lot in companies. However, because the cost is relatively high, one of the important roles of the network operator in the enterprise is to maintain the optimal state by properly arranging and utilizing the resources of the network leased line. In other words, in order to properly support business service requirements, it is essential to properly manage bandwidth resources of leased lines from the viewpoint of data transmission, and properly predicting and managing leased line usage becomes a key factor. Therefore, in this study, various prediction models were applied and performance was evaluated based on the actual usage rate data of leased lines used in corporate networks. In general, the performance of each prediction was measured and compared by applying the smoothing model and ARIMA model, which are widely used as statistical methods, and the representative models of deep learning based on artificial neural networks, which are being studied a lot these days. In addition, based on the experimental results, we proposed the items to be considered in order for each model to achieve good performance for prediction from the viewpoint of effective operation of leased line resources.

Keywords : Leased Line, Traffic Modeling, Time Series Analysis, Deep Learning, RNN, LSTM

순환 심층 신경망 모델을 이용한 전용회선 트래픽 예측

이 인 규† · 송 미 화††

요 약

전용회선은 데이터 전송에 있어서 연결된 두 지역을 독점적으로 사용하는 구조이기 때문에 안정된 품질수준과 보안성이 확보되어 교환회선의 급격한 증가에도 불구하고 기업 내부에서는 지속적으로 많이 사용하는 회선 방식이다. 하지만 비용이 상대적으로 고가이기 때문에 기업 내 네트워크 운영자의 중요한 역할 중의 하나는 네트워크 전용회선의 자원을 적절히 배치하고 활용하여 최적의 상태를 유지하는 것이 중요한 요소이다. 즉, 비즈니스 서비스 요구 사항을 적절히 지원하기 위해서는 데이터 전송 관점에서 전용회선의 대역폭 자원에 대한 적절한 관리가 필수적이며 전용회선 사용량을 적절히 예측하고 관리하는 것이 핵심 요소가 된다. 이에 본 연구에서는 기업 네트워크에서 사용하는 전용회선의 실제 사용률 데이터를 기반으로 다양한 예측 모형을 적용하고 성능을 평가하였다. 일반적으로 통계적인 방법으로 많이 사용하는 평활화 기법 및 ARIMA 모형과 요즘 많은 연구가 되고 있는 인공신경망에 기반한 딥러닝의 대표적인 모형들을 적용하여 각각의 예측에 대한 성능을 측정하고 비교하였다. 또한, 실험 결과에 기초하여 전용회선 자원의 효과적인 운영 관점에서 각 모형이 예측에 대하여 좋은 성능을 내기 위하여 고려해야 할 사항을 제안하였다.

키워드 : 전용회선, 트래픽 모델링, 시계열분석, 딥러닝, RNN, LSTM

1. 서 론

데이터 전송을 위한 회선 방식은 전용회선 방식과 교환회선 방식으로 분류된다. 전용회선 방식은 점대점(point-to-point) 방식으로 두 지점을 배타적으로 연결하고 독점적으로 자원을 사용하는 방식으로 회선 사용에 대한 독립성이 확보되며 안정성과 보안성이 우수하다는 장점이 있다. 반면, 교환회선 방식은 다른 트래픽과 통합 및 공유되는 방식으로 운영 효율성이 확보되고 비용이 저렴하다. 네트워크 및 통신의 디자인 관점에서 고려해야 할 사항은 신뢰성(reliability), 응답시간(responsiveness), 용량(capacity)과 비용(cost)의 요소가 있을 수 있으며[2] 기업 IT 비용 관점에서는 가능한 적은 비용으로 많은 용량을 제공할 수 있기 때문에 교환회선 방식이 유리하다[3]. 교환회선은 자원이 공유되는 개념이므로 용량을 효율적으로 사용할 수 있고 전용회선에 비하여 비용 대비 우수함을 보이기 때문이다.

하지만, 전용회선은 신뢰성과 응답시간 및 그리고 현대 네트워크에서 중요한 요소 중의 하나인 보안 관점에서 장점을 가지고 있다. 즉, 두 지점을 독점적으로 연결하는 전용회선은 전송속도(bandwidth), 지연시간(delay), 왜곡(jitter), 패킷 유실(packet loss) 등과 같은 서비스 품질(Quality of Service) 수준이 보장될 뿐 아니라, 보안성이 우수하다는 장점을 지니고 있다. 이러한 장점 때문에 일반 기업 내에서는 일정 수준 이상의 통신 품질이 요구되거나, 일정한 수준 이내의 응답시간이 요구되는 실시간 시스템, 그리고 강력한 보안성이 필요한 곳은

※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회에서 "효율적인 전용회선 자원 사용량 예측을 위한 통계적 기법과 기계학습 모델 비교 연구"의 제목으로 발표된 논문을 확장한 것임.

† 준 회 원 : 세명대학교 정보통신학부 박사과정
†† 정 회 원 : 세명대학교 정보통신학부 조교수

Manuscript Received : June 29, 2021
First Revision : August 2, 2021
Accepted : August 11, 2021

* Corresponding Author : Mi-Hwa Song(mhsong@semyung.ac.kr)

전용회선 방식을 지속적으로 많이 사용하고 있다[4].

기업 내 IT 자원은 비즈니스 목적을 달성하기 위하여 존재한다. 따라서 네트워크 전용회선의 사용률은 비즈니스 운영의 패턴과 깊은 관련성을 가지고 있는 경우가 많다. IT 네트워크 운영자의 역할 중 하나는 전용회선과 같은 IT 자원의 적절한 배치에 있으며, 이러한 비즈니스 요구에 대응하기 위하여 적절한 전용회선 자원을 확보하여 서비스를 제공하여야 한다. 너무 많은 전용회선 자원의 할당은 비즈니스 요구에 충분한 서비스를 제공하지만, 비용 측면에서는 심각한 낭비를 초래한다. 서비스 제공에 대한 효율성을 확보하기 위해서는 비용 대비 최적의 자원 사용률의 확보가 필요하며 이를 위해서는 전용회선 사용량을 사전에 예측(forecasting) 하기 위한 방법이 필요하다[5, 6].

과거의 연구는 NMS(Network Management System)를 통하여 수집된 네트워크 트래픽 데이터에 대하여 미래 사용량을 예측하는 기법으로 주로 통계적인 방법들을 이용하는 경우가 많았으며[7, 8], 요즘에는 인공지능의 여러 방법들을 이용하는 경우가 주를 이루고 있는 추세이다[9]. 하지만 전용회선 사용량에 대하여 통계적인 기법과 인공지능적인 기법을 모두 적용하고 각각의 장단점을 분석하는 통합적인 연구는 많지 않다. 따라서 본 연구에서는 국내 기업에서 신용카드 처리를 위하여 사용하는 전용회선의 실제 데이터를 수집하여 다양한 시계열 분석 모형을 적용하고 성능을 비교 분석하였다.

우선 전용회선 데이터의 수집은 SNMP(Simple Network Management Protocol) 프로토콜을 이용하여 폴링 방식을 이용하였다[10, 11]. 수집되는 정보는 전용회선 사용률을 산정하기 위한 MIB-II를 이용하고, 수집된 데이터는 전처리 과정을 거쳐서 시계열 분석 모형에 적용하였다[12, 13].

전용회선 사용량에 대하여 과거의 트래픽 량을 기준으로 향후의 트래픽 량을 예측하는 시계열 분석 모형에는 전통적으로 통계 기반 모형을 이용하는 방법과 데이터 학습에 기반한 딥러닝 기법으로 분류할 수 있다. 통계적인 방법에서는 과거 데이터의 평균을 이용하여 예측하는 방법인 평활법과 비정상적 요소를 제거하고 자기 상관(autocorrelation)을 이용하는 ARIMA 모형을 적용하였다[14-16]. 또한 최근 기계학습의 발전으로 많이 연구되고 있는 딥러닝 알고리즘에서는 시계열 데이터 처리에 적합한 RNN과 LSTM 알고리즘을 적용하여 분석하였다[17-22]. 각 모형을 대상으로 예측한 결과에 대하여 정확도를 평가하기 많이 사용하는 MAD, MSE, RMSE를 사용하여 각각의 성능을 비교, 평가하였다[23, 24]. 또한 각 기법에 대하여 전용회선 서비스를 제공하는 관점에서 어떠한 장단점이 있는지 분석하였다.

2. 관련 연구

2.1 전용회선 트래픽 데이터의 수집 방법

네트워크 장비에 연동된 전용회선의 트래픽 데이터를 수집하기 위해서는 TCP/IP 기반의 SNMP를 사용하였다[7]. 대부분의 네트워크 장비들이 SNMP를 지원하고 있고 비교적 쉬운 방법으로 네트워크 성능 데이터를 수집할 수 있는 장점이 있다. Fig. 1과 같이 데이터 수집을 담당하는 SNMP 관리자(Manager)는 주기적으로 SNMP 메시지를 이용하여 네트워크 장비인 SNMP 에이전트(Agent)에게 트래픽 데이터를 요청하고 수집한다.

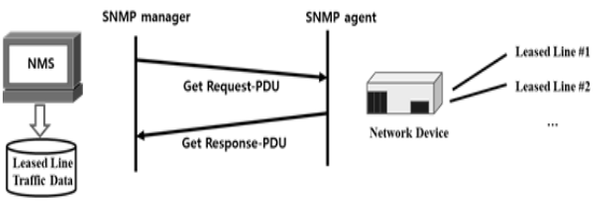

Fig. 1. Collection of Leased Line Traffic Using SNMP

SNMP 관리자와 SNMP 에이전트 사이에 관리 정보는 MIB(Management Information Base)으로 불리는 관리 항목에 의존한다. 본 연구에서는 MIB-II에서 제공하는 관리 정보 중에서 인터페이스의 송수신된 정보를 나타내는 IfInOctet와 IfOutOctet를 이용하였다[12]. MIB-II의 관리 정보 중 IfInOctet와 IfOutOctet 을 이용하여 수신 및 송신에 대한 수집 정보를 이용하여 전용회선 사용률을 측정할 수 있다. 이 정보는 수집하는 시간과 주기의 흐름에 따라 변하는 시계열적인 특징을 가지게 된다.

2.2 시계열 데이터의 예측을 위한 평활화 모형

시계열 데이터의 예측은 시간을 독립변수로 하여 종속변수를 예측한다. 이러한 시계열 분석 기법 중에서 전통적으로 많이 사용되는 평활 기법은 과거 시계열 데이터에 대하여 가중치를 부여하여 미래의 값을 예측하는 방법이다. 대표적인 방법으로는 이동평균법과 지수평활법이 있다.

1) 이동평균법

주식시장의 주가이동평균선 등 다양한 분야에서 많이 이용되는 이동평균법은 과거 일정 기간의 자료의 평균을 이용하여 미래의 값을 예측하는 방법으로 비교적 간단한 시계열 자료 예측 방법이다. 과거 데이터에 동일한 가중치를 부여한 것이 단순 이동평균법이며 수식으로 나타내면 Equation (1)과 같다. (F: 미래의 예측 값, A: 과거 데이터 값, 평균 기간: n)

$$F_{t+1} = \Sigma A_t / n = (A_t + A_{t-1} + \ldots + A_{t-n+1})/n \qquad (1)$$

과거 데이터의 같은 가중치를 적용하는 형태의 평균이 아니라 중요도에 따라 각각 다른 가중치를 부여하는 것을 가중이동평균법이라 한다. 평균 기간을 나타내는 n의 값이 클수록 직선에 가까운 부드러운 형태가 되며 시계열의 변화에 늦게 반응하지만 안정적인 예측이 가능하다.

2) 지수평활법

지수평활법은 가중이동평균법과 유사하지만 과거의 일정 기간이 아닌 전체 데이터를 이용하고, 최근의 값에 중요성을 더 높게 한다. 즉, 가중치를 부여할 때 가장 최근 값에 더 많

은 가중치를 부여하고 과거 데이터로 갈수록 가중치를 지수적으로 줄이는 방법이다[14]. 여기서 평환 계수(α)의 값을 정하는 것이 중요한데, 평환 계수(α)가 1에 가깝게 커질수록 최근 값에 더 비중이 높아지고 예측 값은 시계열에 대해 빠르게 반응하며, 0에 가깝게 작아질수록 이전 시점의 예측 값과 비슷하게 되어 평활 효과가 커진다. 지수평활법을 수식으로 나타내면 Equation (2)와 같다. (F : 미래의 예측 값, A : 과거 데이터 값, 평활 계수: α)

$$F_{t+1} = \alpha A_t + (1-\alpha)F_t \tag{2}$$

2.3 ARIMA/SARIMA 모형

정상적(stationary) 시계열 데이터는 시간의 추이에 관계없이 평균, 분산, 공분산이 일정한 시계열 데이터를 말하며, 예측이 용이한 형태의 데이터이다. 하지만 실생활에 존재하는 대부분의 데이터는 비정상적(non-stationary) 시계열 데이터의 특성을 포함한다. AR, MA, ARMA, ARIMA, SARIMA 등의 모형은 이러한 비정상성 데이터에 대하여 정상성을 확보한 방법이다.

1) ARIMA 모형

Box-Jenkins로 일컬어지는 ARIMA 모형은 자기회귀모형(AR)과 이동평균모형(MA)을 합친 ARMA 모형에서 차분의 단계를 거쳐 비정상성을 제거하고 정상성을 확보한 방법이다[15].

자기회귀모형(AR, Auto-Regression)은 과거와 현재와의 관계를 정리한 것으로 과거의 관측 값이 미래의 관측 값에 영향을 준다는 모형이다. p차의 AR 모형은 AR(p)로 표현하며, 식으로 표현하면 Equation (3)과 같이 나타낼 수 있다. (Z: 시계열 값, t: 현재시점, p: 과거시점, Φ : 모수, α: 오차)

$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \dots + \Phi_p Z_{t-p} + \alpha_t \tag{3}$$

이동평균모형(MA, Moving Average)은 과거와 현재와의 오차를 이용한 것으로 과거의 오차가 미래의 관측 값에 영향을 준다는 모형이다. q차 항의 MA 모형은 MA(q)로 표현하며, Equation (4)로 표현이 가능하다. (Zt: 시계열 값, t : 현재시점, p: 과거시점, θ : 매개변수, α: 오차)

$$Z_t = \Theta_1 \alpha_{t-1} + \Theta_2 \alpha_{t-2} + \dots + \Theta_p \alpha_{t-p} + \alpha_t \tag{4}$$

ARMA 모형은 AR과 MA의 모형을 합친 것으로 간단한 수식만의 결합으로 분석의 정확도를 많이 높일 수 있으며 수식으로 나타내면 Equation (5)와 같다. 여기에 차분을 거쳐 비정상적 시계열을 정상적 시계열로 변환하게 되는데 이때 비정상적 시계열을 ARIMA 모형으로 표현하며 AR모형차수 및 차분과 MA모형차수를 나타내기 위하여 ARIMA(p,d,q) 로 표현한다.

$$\begin{aligned} Z_t = {} & \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \dots + \Phi_p Z_{t-p} + \alpha_t \\ & + \Theta_1 \alpha_{t-1} + \Theta_2 \alpha_{t-2} + \dots + \Theta_p \alpha_{t-p} \end{aligned} \tag{5}$$

2) SARIMA 모형

시계열 자료가 정상성을 확보하였다고 하더라도 시간 주기로 일정한 형태를 반복하고 있을 때 계절시계열라고 부른다. 시계열 자료가 계절성을 가질 경우에는 계절 변동성을 제거해 주어야 한다. 이를 위한 계절 차분이 필요한데 이를 반영한 것이 SARIMA(Seasonal ARIMA) 모형이다. SARIMA 모형은 (p, q, d)(P, D, Q)s 로 나타내며, 소문자는 모형의 비계절성 부분을 대문자는 모형의 계절성 부분을 나타낸다. s는 관측 값의 개수이다.

2.4 인공신경망 기반의 딥러닝 모델

인공신경망 기반의 딥러닝 모델을 이용한 시계열 분석은 데이터의 학습을 통하여 미래의 데이터를 예측하는 방법으로 RNN(Recurrent Neural Networking)과 이를 개선한 LSTM (Long short-term memory) 모델이 있다[17, 18].

1) RNN 모델

딥러닝 기법 중 RNN은 주로 시계열 분석, 자연어 처리 등 순서가 있는 데이터에 효과적이다. RNN은 입력층, 은닉층, 출력층으로 구성된 유닛이 서로 연결되어 있다. 한 방향으로만 동작하는 인공신경망(ANN, Artificial Neural Network)과는 달리 은닉층의 출력이 다시 입력으로 되는 순환구조를 생성하며 데이터를 학습한다(Fig. 2). 이러한 순환구조를 통하여 학습된 과거의 데이터를 이용하여 미래를 데이터를 예측할 수 있게 한다[19].

2) LSTM 모델

RNN의 경우 이전 정보와 지점 사이의 거리가 멀어질 경우 역전파 진행 시 계산되는 양이 점점 많아지고 학습능력이 급격히 떨어지는 문제점(vanishing gradient problem)이 발생한다. 이를 개선하기 위하여 LSTM 에서는 RNN 메모리에 3개의 노드를 추가하여 이 문제를 개선하였다[20, 21]. Fig. 3에서 망각 게이트(Forget gate)는 이전 단계의 정보를 유지할지 또는 버릴지를 결정하고, 입력 게이트(Input gate)는 새로 들어오는 정보 중에서 어떤 것을 저장할 것인지를 결정하며, 출력 게이트(Output gate)는 출력을 내보낼 때 얼마나 값을 반영할지 결정하게 된다[22].

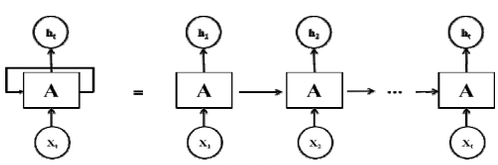

Fig. 2. RNN Structure

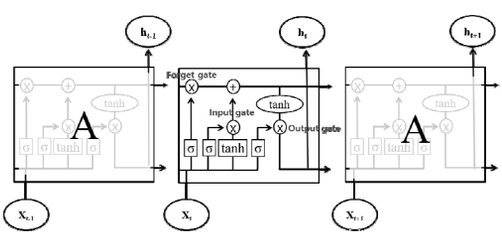

Fig. 3. LSTM Structure

2.5 성능 평가 방법

전용회선 사용률의 예측은 오차율을 통해 성능을 확인할 수 있다. 본 연구에서는 MAE(Mean Absolute Error, 평균 절대 오차), MSE(Mean Squared Error, 평균 제곱 오차), MAPE(Mean Absolute Percentage Error, 평균 절대 백분율 오차)를 사용한다. MAD는 실제 값과 예측 값의 차이의 평균이다. 가장 직관적인 확인이 가능하나 절댓값을 취하기 때문에 과잉성능(overperformance) 인지 저성능(underperformance) 인지 파악은 불가능하다. MSE는 실제 값과 예측 값의 제곱의 평균이며, 특이한 값이 나타날 경우 민감하게 높아지게 된다. MAPE는 MAE와 유사하며 퍼센트로 나타낸다. 이들을 수식으로 나타내면 Equation (6-8)과 같다[23, 24].

$$MAD = \sum_{t=1}^{n} |A_t - F_t| / n \quad (6)$$

$$MSE = \sum_{t=1}^{n} (A_t - F_t)^2 / n \quad (7)$$

$$MAPE = \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| / n * 100 \quad (8)$$

3. 시험 및 평가

3.1 정보의 수집

데이터 수집을 위하여 사용되는 전용회선은 한국 내 한 신용카드사의 승인 처리를 위하여 사용하는 실제 데이터를 대상으로 하였다. SNMP 프로토콜을 이용하고 MIB-II의 IfInOctet와 IfOutOctet 정보를 사용하였다. 폴링 주기는 1시간으로 하여 1개월 동안 720개의 데이터를 수집하였다. 수집된 IfInOctet와 IfOutOctet을 이용하여 회선 사용률을 계산한다. 수신과 송신의 같은 대역폭을 제공하는 일반적인 전용회선의 경우 어느 한쪽만 높더라도 전체 사용률에 영향을 주므로 높은 쪽을 기준으로 선정하여 사용량을 계산하였으며, 계산식은 Equation (9)와 같다.

$$Link\ Util.(\%) = MAX(\Delta IfinOctets, \Delta IfoutOctets) * 8\ bits \\ / Linkspeed * 100 \quad (9)$$

수집된 데이터의 품질을 높이기 위한 전처리 과정을 수행하는데 누락된 데이터를 점검하여 최소화하는 작업이다. SNMP는 신뢰성이 보장되지 않는 UDP 방식을 기반으로 동작되므로 SNMP 메시지의 유실이 있을 수 있고, 시스템 부하 등의 이유로 SNMP 에이전트가 응답을 못 줄 수도 있다. 따라서 누락된 데이터가 없는지 확인하여 보정이 필요하다. 본 연구에서는 결측 값이 있을 경우, 바로 다음의 데이터를 이용하는 방법을 사용하였다.

계산된 회선 사용량에 대하여 70%에 해당되는 504개의 데이터는 학습 데이터로 30%에 해당하는 216개의 데이터는 테스트 데이터로 분리한다. 각각의 모형별로 학습 데이터를 기반으로 알고리즘을 적용한 후 테스트 데이터를 예측하여 실제 값과의 적중률로 성능을 평가하였다.

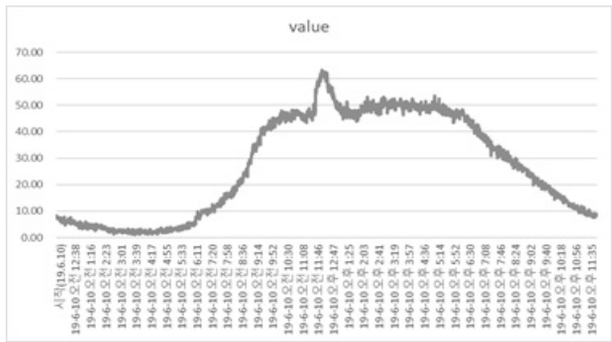

Fig. 4. Daily Pattern of Collected Data

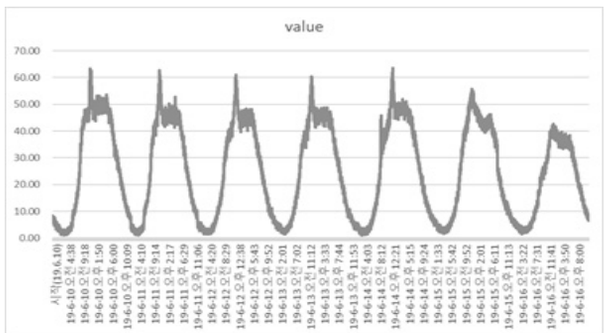

Fig. 5. Weekly Pattern of Collected Data

3.2 예비 분석

실험에서 사용되는 전용회선 트래픽은 신용카드 거래의 승인에 해당하는 데이터이다. 일일과 주별 사용률 패턴은 Fig. 4-5에서 확인되듯이 전반적으로 급격히 변하지 않고 증가 및 감소가 지속적으로 유지되고 완만한 형태를 가진다. 평일의 경우 순차적으로 증가하여 점심시간에 일시적으로 급격히 올라간 후 저녁 9시 이후로는 점점 줄어드는 형태를 가진다. 점심시간에 신용카드 사용이 집중적으로 사용하는 것이 보이며, 토요일의 경우 급격함이 줄고 일요일은 좀 더 평이한 데이터 패턴을 보인다. 이러한 데이터 패턴 때문에 본 데이터의 시계열의 계절주기는 1주일로 나타난다.

시계열 데이터는 추세와 순환변동, 계절변동을 포함한 불규칙성을 내포한다. 1일 주기로 수집된 데이터를 분해(decomposition) 해보면 변동성을 좀 더 명확히 파악할 수 있으며, 7일 주기로 주중에는 평활한 모습을 보이며 주말에는 낮아지는 경향을 확인할 수 있다. 따라서 본 데이터의 시계열의 계절주기는 1주일이 된다.

3.3 모형의 선정과 피팅

전용회선 사용량 예측을 위한 모형의 선정을 위하여 평활화 방법과 Seasonal ARIMA 및 인공신경망의 딥러닝 세 가지 방법으로 분류하여 비교 분석한다.

1) 평활화 모형

이동평균법은 과거 일정 기간 데이터의 평균을 이용하여

미래의 값을 예측하기 때문에 평균 기간(n)의 선정 값이 예측 결과의 성능에 중요한 영향을 미친다. 본 연구에서는 평균 기간(n) 값을 2부터 지속적으로 늘려가며 가장 최적의 값을 설정하였는데 최솟값인 2로 확인되었다. 이는 평균 기간(n) 값이 클수록 예측 값이 완만해져 예측 오차가 증가하기 때문이다.

지수평활법은 최근 값에 더 많은 가중치를 부여하고 과거 데이터는 지수적으로 가중치를 줄이는 방법으로 효과적인 성능을 추출하기 위해서는 평활 계수(α)의 값이 중요하다. 여기서는 평활 계수(α) 값을 0.1부터 0.9까지 지속적으로 늘려가며 실험을 진행하였으며 계산된 최적의 0.9로 확인되었다. 이는 완만하게 증가 및 감소하는 트래픽 패턴에 따라 가장 최근 값이 가장 효과적인 것을 설명해 준다.

2) ARIMA/SARIMA 모형

ARIMA 및 계절성 ARIMA를 적용하기 위해서는 수집된 전용회선 데이터가 시계열적 추세 및 계절성이 포함된 비정상성 데이터 여부에 대한 파악이 필요하다. ACF(Auto Correlation Function)와 RACF(Partial ACF)를 사용하여 정상성인지 비정상인지를 파악하고 필요한 차분의 값을 찾아서 적용할 수 있는데, 본 연구에서는 계절성 ARIMA 모형에서 각각의 파라미터 조합을 적용하여 가장 최적의 값을 찾는 방법을 이용하였다. 계절성을 있는 경우 주기를 파악한 후 적용하도록 할 필요가 있다. 분해 과정에서 계절성의 주가가 7일인 것을 파악하였으므로 Seasonal ARIMA에 7일 주기인 168을 적용하였다. 해당 실험을 위하여 사용한 파라미터 값은 SARIMA(1, 1, 2)(1, 1, 2, 168)이며 AIC는 1795를 나타냈다.

3) 인공신경망 기반의 딥러닝 기법

인공신경망 기반의 딥러닝 기법은 기본적으로 RNN과 이를 개선한 LSTM이 있다. 수집된 연속적인 과거 데이터를 다음의 값으로 학습시키는 과정을 반복함으로써 실제 테스트 데이터에 대한 예측을 실시하게 하였다. 성능을 높이기 위해서는 학습 레이어 및 파라미터의 구성, 학습 횟수, 연속되는 과거 데이터의 개수, 배치 크기 등의 다양한 파라미터들에 대한 최적화된 값을 찾는 작업이 필요하다.

본 실험에서는 최적화 함수는 adam을 사용하고, 손실 함수는 평균제곱법을 사용하였다. 파라미터 수는 RNN의 경우 180,673이며 LSTM은 707,393이다. RNN과 LSTM의 성능을 결정하는 주요 파라미터는 일반 값과 최적화된 값으로 나눠서 실험을 진행했으며, 수집된 몇 개의 연속적인 과거 데이터를 학습시킬 것인지를 나타내는 변수(Sequence)는 60과 10을 사용하였다. 또한, 몇 번을 훈련할지 결정하는 epoch 값은 10회와 500회를 나눠어서 실시했으며 batch size 도 32과 1로 나눠서 실시하였다.

4. 실험의 결과 및 평가

실험 환경은 구글에서 제공하는 Colab pro를 이용하였고 프로그램은 Python 3.6.9 및 TensorFlow 2.4.0을 사용하였다. 30일 데이터를 기준으로 70%에 해당하는 21일을 학습 데이터로 하고, 30%에 해당하는 9일 치를 예측하여 각 모형별로 실제 데이터와 비교하여 성능을 평가하였다.

평활화 방법 중 이동평균법은 전반적으로 데이터의 예측이 무난하지만 과거 데이터 평균을 적용하기 때문에 트래픽 피크치의 예측이 어렵고 과거의 데이터의 평균이 한 스텝 늦게 적용되는 형태를 보인다. 과거 데이터에 일정한 가중치를 부여하는 지수평활법은 이동평균법보다 전체적으로 양호한 결과를 보인다. Fig. 6-7은 테스트 데이터에 대한 평활화 기법의 예측치 결과이다.

평활화 모형의 단점은 전용회선 트래픽이 급격히 변화할 경우 적절한 예측이 부족하다는 문제가 있다. 평활화는 기본적으로 과거의 데이터를 적절히 평균화하여 반영하기 때문인데, 실제 전산 운영 환경에서 급격한 변화에 대한 최댓값을 적절히 예측하지 못할 경우 전용회선 자원 부족으로 적절한 서비스가 어려운 문제가 생길 수 있다. 이에 반하여 과거 데이터에 대한 회귀 분석적인 방법을 이용하는 ARIMA/SARIMA는 평활화 기법에 대하여 훨씬 개선된 결과를 나타낸다. Fig. 8은 테스트에 대한 ARIMA/SARIMA의 예측치의 결과이다.

인공신경망 기반의 딥러닝 방법은 정상성 데이터로 변환할 필요가 없고 비선형 구조의 데이터도 학습을 통하여 예측하는 방법을 이용한다. 하지만 성능에 영향을 미치는 주요 파라

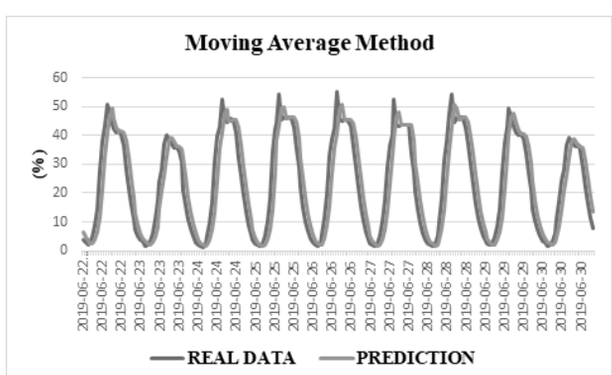

Fig. 6. Permeance Evaluation: Moving Average Method

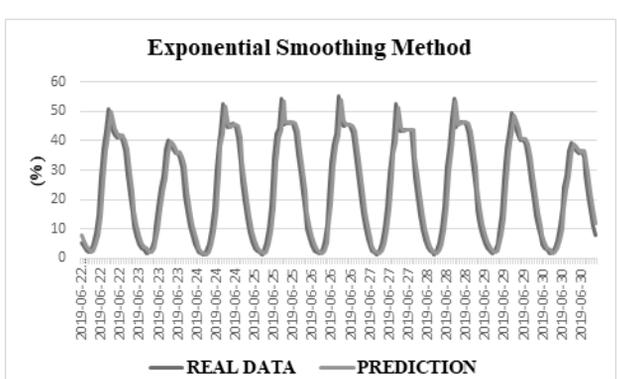

Fig. 7. Permeance Evaluation: Exponential Smoothing Method

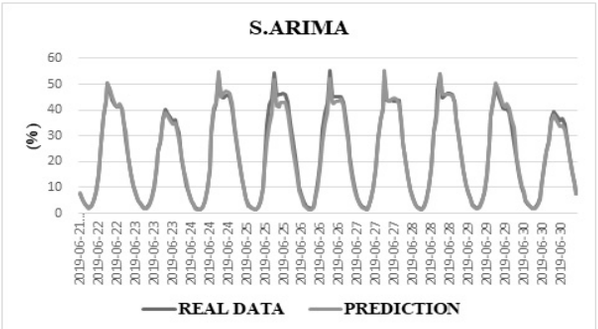

Fig. 8. Permeance Evaluation: Seasonal ARIMA

미터를 최적화하는 작업이 필수적이다. 일반 파라미터(Sequence(60), Epoch(10), Batch size(32))를 적용한 경우와 최적화된 파라미터(Sequence(10), Epoch(500), Batch size(32))에 대해서 학습 횟수에 대한 손실 함수와 성능을 비교하면 Fig. 9-10 및 Table 1과 같다.

최적화된 파라미터를 찾아내면 딥러닝 모형이 시계열 데이터의 처리를 위하여 사용한 통계학적 모형의 성능보다 좋은 결과를 보인다. 또한 RNN의 경우보다는 LSTM 방법이 더 좋은 결과를 나타내는 것으로 확인되어 RNN을 개선한 모형임을 확인 가능하였다. Fig. 11-12은 테스트 데이터에 대한 RNN과 LSTM 모형의 성능 결과이다.

또한 딥러닝 방법은 데이터의 학습을 기반으로 하므로 데이터의 양이 성능에 영향을 미친다. 수집된 전용회선 데이터에 대하여 동일한 기간의 1개월을 그대로 유지하고 polling time 을 1시간에서 1분으로 하여 데이터의 양을 60배 늘린

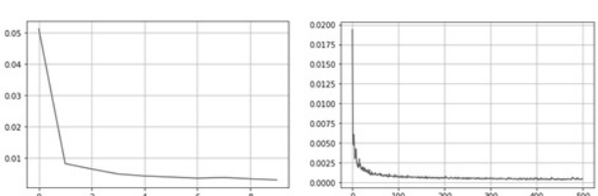

Fig. 9. RNN Loss Func by Parameters (General and Optimized)

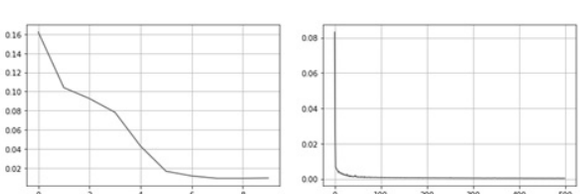

Fig. 10. LSTM Loss Func by Parameters (General and Optimized)

Table 1. Performance Comparison by Parameter Optimization

Model	Parameters	MAD	MSE	MAPE
RNN	general	1.89	8.97	11.02
	optimized	0.88	1.62	6.66
LSTM	general	3.99	30.25	27.61
	optimized	0.76	1.26	6.10

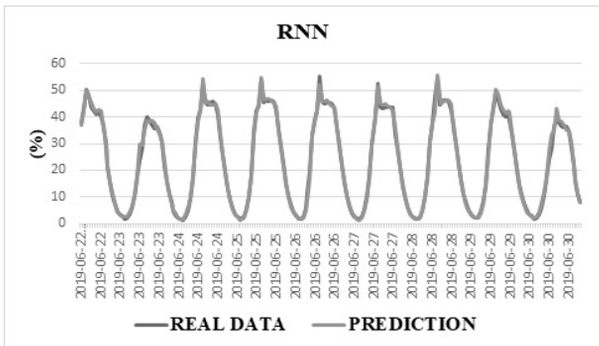

Fig. 11. Permeance Evaluation: RNN

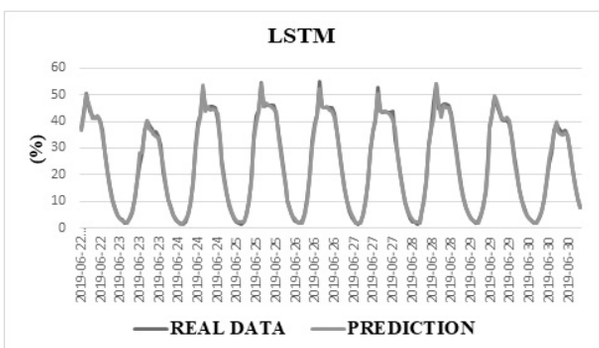

Fig. 12. Permeance Evaluation: LSTM

후 학습을 실시하고 각 성능을 비교하고 Table 2에서 결과를 나타내었다. 파라미터는 일반 Parameters(Sequence(60), Epoch(10), Batch size(32))를 적용하였다. 같은 조건에서 학습 데이터가 많을수록 성능이 훨씬 우수한 것이 확인된다.

Table 3 및 Fig. 13은 지금까지 수행한 각 모형들에 대하여 성능 결과를 나타내는 표와 그림이다. 각 모형들을 동일 조건에서 비교하기 위하여 수집 데이터가 1시간 주기의 데이터(760개)를 기준으로 하였다. 전체적으로 평활화 기법보다

Table 2. Performance Comparison by the Size of Training Data

Model	Parameters	MAD	MSE	MAPE
RNN	1 hour cycle(#720)	1.89	8.97	11.02
	1 min. cycle(#43200)	0.77	1.16	5.51
LSTM	1 hour cycle(#720)	3.99	30.25	27.61
	1 min. cycle(#43200)	0.73	1.04	5.76

Table 3. Performance Evaluation Result

classification	model/method	MAD	MSE	MAPE
statistical method	moving average	5.76	54.32	41.57
	exponential smoothing	4.44	32.80	31.39
	SARIMA	1.08	2.87	6.40
deep learning method	RNN	0.88	1.62	6.66
	LSTM	0.76	1.26	6.10

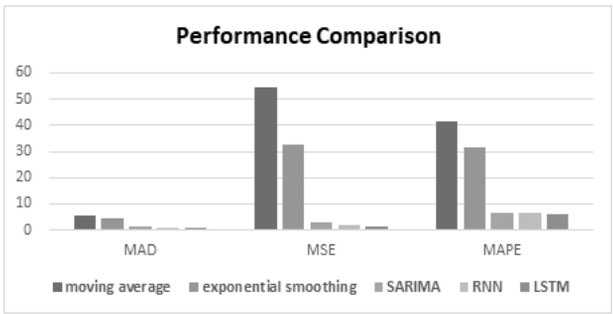

Fig. 13. Performance Evaluation Comparison

는 확률적 방법을 사용하는 ARIMA/SARIMA의 예측 성능이 우수하였으며, 신경망 기반의 딥러닝을 사용하는 방법의 예측 성능이 최고로 우수하였다. 본 실험 결과 딥러닝 기법이 ARIMA/SARIMA 보다 MSE 기준 56%~78% 예측 성능이 향상됨을 확인하였다.

5. 결론 및 향후 연구

본 연구에서는 실제 기업의 운영환경에서 사용 중인 전용회선의 데이터를 이용하여 시계열 데이터의 예측이 가능한 모형에 대해서 각각의 성능을 측정하였다. 전용회선은 특수한 비즈니스 목적을 위하여 설치하는 경우가 많으며 교환회선보다 품질이나 보안성 등 많은 장점을 지니고 있으나 가격이 비싸다는 단점이 있다. 따라서 전용회선을 효율적으로 운영하기 위해서는 정확한 트래픽을 예측하고 그 예측된 데이터에 맞는 전용회선 대역폭 자원을 확보할 필요가 있다.

데이터의 예측을 위한 방법은 시계열 패턴을 분석하고 상관관계, 추세 등의 특징을 파악하는 해석적 측면의 통계학적 방법과, 학습을 통하여 미래 예측데이터의 정확성을 높이는 기계학습 방법으로 나눌 수 있다. 통계학적 방법으로는 전통적인 평활법과 ARIMA 모형의 성능을 검토하였으며 기계학습 방법에서는 신경망 기반의 RNN과 LSTM 모형의 성능을 검토하였다.

평활법에서는 지수평활법이 단순히 과거의 데이터의 균일한 평균을 사용하는 평균이동법 보다 좋은 성능을 나타냈으나 전반전으로 과거 데이터에 대하여 가중치를 둔 평균값이라는 한계성이 있다. 이는 급격히 변화하는 트래픽에 대해서 예측 능력이 저하되고, 실시간 서비스를 제공하는 상황에서 적절한 서비스를 보장하지 못하는 문제가 생길 수 있다. ARIMA/SARIMA 모형은 평활법 보다 좋은 예측 성능을 보였다. 수집된 데이터의 추세 및 계절성을 포함한 데이터이므로 비정상성의 데이터를 차분을 이용하여 정상적인 데이터를 산출하고 일정한 주기로 반복되는 데이터 패턴을 분해 기법을 통하여 확인 후 적절한 파라미터를 적용한 후에 좋은 성능 결과를 확인하였다.

인공신경망의 딥러닝 방법에서는 통계적 방법의 결과보다는 더 나은 성능 결과를 확인할 수 있었다. 추세 및 계절성 같은 비정상성 데이터를 별도로 계산하지 않아도 스스로 학습을 통하여 정확한 예측치를 찾아내는 편리함을 제공하였다. 또한 RNN보다는 개선된 LSTM이 보다 좋은 성능을 보여줌을 확인하였다. 딥러닝의 이러한 좋은 성능을 만족하기 위해서는 두 가지 요소가 필요하다. 첫째, 최적의 파라미터 조합을 찾는 작업이 필요하다. 딥러닝 기법에서 최적의 파라미터를 찾아주는 여러 기법들이 있으나 본 연구에서는 각 시계열 데이터의 예측 알고리즘의 성능 평가를 목적으로 하므로 대표성을 갖는 몇 가지 파라미터만을 튜닝 하였고 그것만으로도 다른 모형보다는 좋은 결과가 나오는 것을 확인하였다. 둘째, 학습 데이터가 많을수록 좋은 성능을 나타내는 것을 확인할 수 있었다.

본 연구에서는 1개월이라는 비교적 작은 기간의 데이터를 가지고 연구를 진행하였다. 이번 데이터와 같이 패턴과 계절성이 뚜렷하고 큰 변화가 없을 경우에는 1개월 정도의 데이터만 가지고도 각 모델별 성능 평가가 가능하다. 다만 비즈니스 특성에 따라 전용회선 트래픽 패턴이 다양해질 경우에는 좀 더 긴 기간의 데이터가 필요할 것으로 예상된다. 향후 연구에서는 다양한 비즈니스 패턴에 따른 전용회선 트래픽 패턴의 분류와 이에 따라 트래픽 수집 기간을 늘리는 형태로의 연구가 필요할 것으로 판단된다.

References

[1] I. G. Lee and M. H. Song, "A comparative study of statistical techniques and machine learning models for efficient leased line resource usage prediction," *Proceedings of the KIPS*, Vol.28., No.1, pp.474-476, 2021.

[2] L. G. Roberts and B. D. Wessler, "Computer network development to achieve resource sharing," *Proceedings of the May 5-7, 1970, Spring Joint Computer Conference*, 1970.

[3] H. M. Sigurdsson, S. E. Thorsteinsson, and T. K. Stidsen. "Cost optimization methods in the design of next generation networks," *IEEE Communications Magazine*, Vol.42, No.9, pp.118-122, 2004.

[4] Statistical Office, "Business Basic Statistical Survey Report," Each Year (2020).

[5] M. Joshi and T. H. Hadi, "A review of network traffic analysis and prediction techniques," arXiv preprint arXiv: 1507.05722, 2015.

[6] W. Yoo and A. Sim. "Time-series forecast modeling on high-bandwidth network measurements," *Journal of Grid Computing*, Vol.14, No.3, pp.463-476, 2016.

[7] H. W. Taek, A. S. Jin, and C. J. Wook, "Forecasting technique of line utilization based on SNMP MIB-II using time series analysis," *KIPS Journal*, Vol.6, No.9, pp.2470-2478, 1999. DOI: 10.3745/KIPSTE.1999.6.9.2470.

[8] S. J. Jung, D. J. Kim, Y. H. Know, and C. G. Kim, "A fitness verification of time series models for network traffic predictions," *The Journal of Korea Information and Communications Society*, Vol.29, No.2B, pp.217-227, 2004.

[9] S. H. Ji, H. Hasanova, K. S. Shim, and M. S. Kim, "Prediction of traffic usage using machine learning algorithm for efficient network management," *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, pp.824-825, 2018.
[10] W. Stallings, "SNMP and SNMPv2: the infrastructure for network management," *IEEE Communications Magazine*, Vol.36, No.3, pp.37-43, 1998.
[11] J. D. Case, M. Fedor, M. L., Schoffstall, and J. Davin, "RFC1157: Simple network management protocol (snmp)," 1990.
[12] K. McCloghrie and M. T. Rose, "RFC1213: Management information base for network management of TCP/IP-based internets: MIB-II," 1991.
[13] M. T. Rose and K. McCloghrie, "RFC1155: Structure and identification of management information for TCP/IP-based internets," 1990.
[14] R. G. Brown and R. F. Meyer, "The fundamental theorem of exponential smoothing," *Operations Research*, Vol.9, No.5, pp.673-685, 1961.
[15] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, "Time series analysis: Forecasting and control," *John Wiley & Sons*, 2015.
[16] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and practice," OTexts, 2018. [Internet], Available from: https://otexts.com/fpp2
[17] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, Vol.404, pp.132306, 2020.
[18] J. Zhao, et al., "Do rnn and lstm have long memory?," *International Conference on Machine Learning*, PMLR, 2020.
[19] Colah's Blog, Understanding LSTM Networks [Internet], Available from: http://colah.github.io/posts/2015-08-Understanding-LSTMs/(2015)
[20] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Computation*, Vol.9, No.8, pp.1735-1780, 1997.
[21] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, Vol.12, No.10, pp.2451-2471, 2000.
[22] B. Lim and S. Zohren, "Time-series forecasting with deep learning: A survey," *Philosophical Transactions of the Royal Society A*, Vol.379, No.2194, pp.20200209, 2021.
[23] R. Kumar, P. Kumar, and Y. Kumar, "Time series data prediction using iot and machine learning technique," *Procedia Computer Science*, Vol.167, pp.373-381, 2020.
[24] I. Aijaz and P. Agarwal, "A study on time series forecasting using hybridization of time series models and neural networks," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)* Vol.13, No.5, pp.827-832, 2020.

이 인 규

https://orcid.org/0000-0001-9004-3757
e-mail : db999@chol.com
1995년 세명대학교 전자계산학과(학사)
2004년 서강대학교 정보통신학과(석사)
2019년 연세대학교 경영전문대학원(석사)
2019년 ~ 현 재 세명대학교 정보통신학부 박사과정
관심분야 : 인공지능, 데이터 분석, 네트워크, SDN/NFV

송 미 화

https://orcid.org/0000-0001-7047-8032
e-mail : mhsong@semyung.ac.kr
2002년 이화여자대학교 컴퓨터학과(학사)
2005년 서울대학교 전기컴퓨터공학부(석사)
2013년 ~ 현 재 세명대학교 정보통신학부 조교수
관심분야 : Intelligent Information Systems, Natural Language Processing, Human-Computer Interaction, Healthcare Informatics