논문 22-47-02-22 The Journal of Korean Institute of Communications and Information Sciences '22-02 Vol.47 No.02
https://doi.org/10.7840/kics.2022.47.2.398

이동통신 트래픽 예측을 위한 클러스터링 기법

나 세 현•, 김 영 준*, 유 현 민*, 안 희 준*, 문 정 모**, 홍 인 기°

Clustering Method for Mobile Traffic Prediction

Se-Hyeon Na•, Young-Jun Kim*, Hyeon-Min You*, Hee-Jun Ahn*,
Jung-Mo Moon**, Een-Kee Hong°

요 약

급증하는 모바일 트래픽을 네트워크가 적절히 수용하고, 네트워크 성능을 유지 관리하기 위해서는 미래에 발생할 트래픽을 예측하는 것이 중요하다. 본 논문에서는 시공간 데이터를 학습시키는데 적합한 딥러닝 알고리즘인 ConvLSTM(Convolutional LSTM)을 사용하여 미래 트래픽 데이터를 예측한다. 트래픽 데이터는 시간과 공간에 따라 발생하는 양상이 제각각이기 때문에 서로 다른 양상을 보이는 트래픽 데이터를 한꺼번에 학습 데이터로 사용하여 학습시키는 것은 모델의 성능을 저해할 수 있다. 따라서 본 연구에서는 나름의 트래픽 유사성에 대한 기준을 정하여 유사성에 의한 클러스터링 알고리즘을 통해 클러스터 단위로 트래픽 데이터를 학습시킨다. 본 논문은 연구에 사용한 유사도 기반 클러스터링 방법을 설명하고, 클러스터의 개수를 증가시켜 학습시킬 때의 트래픽 예측 성능의 변화를 분석한다. 연구 결과, 클러스터링 개수를 증가시킬수록 예측 오류가 줄어드는 것을 확인할 수 있었다. 그러나, 너무 많은 클러스터로 나눌 경우 오히려 예측 오류가 증가하였다.

키워드 : 트래픽 예측, 딥러닝, 클러스터링, 컨볼루셔널 LSTM, 평균 제곱근 오차
Key Words : Traffic Prediction, Deep-Learning, Clustering, Convolutional LSTM, RMSE

ABSTRACT

In order to accommodate mobile traffic and maintain network performance, it is important to predict future mobile traffic generation. In this paper, we predict the amount of future traffic data using convolutional LSTM (convLSTM) that is adequate deep learning space-time data modeling. Because traffic data have different attribute over time and space, the learning with all traffic data with different characteristics at once can degrade the performance of the model. Therefore, in this paper, we set the criteria for distinguishing traffic similarity, and group the dataset using the clustering algorithm that reflects the selected criteria. Then, the deep learning model learns traffic data on a cluster-by-cluster basis. This paper describes the similarity-based clustering method and analyzes the traffic prediction performances as the number of clusters is increased. The learning results show that as the number of clustering increases, prediction errors decrease. However, too many clustering results in increase of prediction errors.

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2021-0-02046*)
※ 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2018-0-01659, 5GNR 기반 지능형 오픈 스몰셀 기술 개발)
• First Author : Kyung Hee University, Department of Electronic Information Convergence Engineering, 2015104016@khu.ac.kr, 학생회원
° Corresponding Author : Kyung Hee University, Department of Electronic Information Convergence Engineering, ekhong@khu.ac.kr, 종신회원
* Kyung Hee University, Department of Electronic Information Convergence Engineering, donjomyo@khu.ac.kr, 학생회원; yhm1620@khu.ac.kr, 학생회원; hmk6160@khu.ac.kr, 학생회원
** ETRI(Electronics and Telecommunications Research Institute), jmmoon@etri.re.kr, 정회원
논문번호 : 202110-265-B-RE, Received September 28, 2021; Revised November 19, 2021; Accepted December 2, 2021

Ⅰ. 서 론

오늘날 트래픽이 폭발적으로 증가함에 따라[1] 네트워크의 정체가 빈번하게 발생하고, 이를 수용하기 위한 시스템 용량 증가 및 트래픽 운영 방식에 대한 필요성이 증가하고 있다. 모바일 트래픽 데이터는 특정 장소나 시간대에서 급격하게 증가하는 경향을 보이므로 충분한 네트워크 용량을 미리 확보하지 못한다면 트래픽 폭증과 같은 상황에서 가입자들에게 양질의 서비스를 제공할 수 없다. 따라서 발생할 트래픽을 미리 예측할 수 있다면 장소와 시간에 따라 무선 자원 할당과 네트워크 운용을 효율적으로 할 수 있을 것이다.

트래픽 예측이 네트워크 영역에 적용되면 효율적인 절전과 자원할당이 가능하다. 트래픽 요구가 늘어날 때 프로세서의 수와 복잡성이 증가하여 전력 소비가 많아지는데, 이는 장비의 높은 전력 소비와 그에 따른 냉각 비용 상승으로 이어져 네트워크 운용 비용이 증가한다. 만약 트래픽을 예측할 수 있다면 트래픽이 적은 시간에 추가 프로세서의 전원을 꺼 전력을 절약할 수 있다. 또한 Youtube, Netflix, 비디오 등 새롭게 등장한 응용 프로그램이 차지하는 트래픽에 의해 남용되는 네트워크를 감지하고 방지하기 위한 대책으로 트래픽 예측이 사용될 수 있다. 트래픽 예측을 통한 이상 탐지로 정상적인 트래픽 동작과 구분되는 트래픽을 예측해 자원을 효율적으로 할당하여 양질의 서비스를 제공할 수 있다.

트래픽 예측을위한 방법으로 ES(Exponential Smoothing)[2], ARIMA모델[3]과 같은 고전적인 시계열 예측 기법이 있으나, 환경적 요인(시간적 불규칙성, 공간적 상관성)에 따라 변동이 큰 모바일 데이터를 예측하는 데에는 적합하지 않다. 이러한 모바일 데이터의 특성을 학습시키기에는 3D CNN, convLSTM (convolutional LSTM), STN 등의 인공지능 모델[4~7]을 이용한 예측 방법이 더 적합하다.

본 연구에서는 시공간 데이터 예측에 적절한 convLSTM을 사용하며 클러스터 단위로 트래픽 데이터를 학습시키고 예측하는 알고리즘을 제안한다. 클러스터링을 사용한 인공지능 모델 학습은 기존의 연구 논문[5]에서 수행된 바 있다. 참고 문헌 [5]에서는 SMS, 전화 수/발신 등과 해당 지역의 POI(Point of Interest) 수 등 트래픽과는 성격이 다른 교차 차원 데이터(Cross-Domain Data)를 함께 학습시켰다. 그러나 POI는 해당 지역의 특성을 어느 정도 대변할 수 있겠지만 트래픽 패턴, 트래픽 양과 같은 직접적인 통신 활동의 정보를 대표하지는 못하기 때문에 POI만을 가지고 클러스터링하면 지역을 적절하게 그룹화하지 못한다. 이에 본 연구에서는 셀의 특성이 반영된 클러스터링 파라미터를 선정해 클러스터링을 수행하고 클러스터링 개수에 따른 예측성능을 RMSE(Root Mean Square Error), MASE(Mean Absolute Scaled Error)를 통하여 분석하였다.

Ⅱ. 본 론

2.1 데이터 셋

모델을 학습시키기 위해 사용한 데이터는 2014년 Telecom Italia에서 주최한 'Telecom Italia Big Data Challenge'에서 제공한 시공간 기반 데이터셋이다. 이곳에서 제공하는 데이터는 이탈리아 밀란(Milan) 시를 공간적으로 100x100으로 나눈 그리드($(235\times235)m^2/grid$)에서 일정 기간 동안(2013년 11월 1일 00시 00분부터 2014년 1월 1일 00시 00분까지) 수집되는 시공간 트래픽 데이터를 담고 있다. 이 중, 트래픽이 가장 활발한 밀란의 중심부 20x20 사이즈의 그리드에서 50일 동안 1시간 간격으로 수집되는 트래픽 데이터를 사용하였다. 그림 1의 밀란 시의 지도를 통해 밀란 시가 100x100의 그리드로 공간적으로 나뉘어 있음을 볼 수 있으며 본 연구에서 사용하는 중심부 20x20의 모습 또한 볼 수 있다. 그림 1의 가장 오른쪽 그림은 2013년 11월 1일 09시 00분의 한 시점에서 수집된 트래픽 데이터의 3차원 표현이다. 본 논문에선 전체 50일 분량의 데이터셋 중 43일의 분량의 데이터셋을 훈련 데이터셋으로, 7일 분량의 데이터를 학습 후 예측 평가를 위한 테스트 데이터셋으로 사용한다.

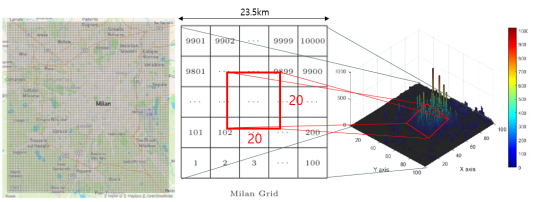

그림 1. 밀란시의 그리드 표현과 트래픽 데이터의 3차원 표현
Fig. 1. The grid expression of Milan and 3D representation of traffic data

2.2 데이터 수집과정

밀란 시에서 발생한 통신 활동에 대한 데이터는 CDR(Call Detail Records) 수에 의해 측정된다.[8] CDR은 기지국을 통해 사용자가 통신을 수행할 때마다 수행 후의 사용 시간과 기지국에서 처리된 정보 등을 기록한 일종의 데이터 레코드이다. CDR은 사용자

가 인터넷 연결을 시작하거나 인터넷을 종료할 때마다 생성되고, 연결이 15분 이상 지속되거나 사용자가 5MB 이상 전송할 때도 CDR이 생성된다. 제공되는 데이터는 (1)과 (2)와 같이 구성된다.

$$S_i(t) = \sum_{v \in C_{map}} R_v(t) \frac{A_{v \cap i}}{A_v} \quad (1)$$

$$S_i^{'}(t) = S_i(t)k \quad (2)$$

$S_i(t)$는 시간 t일 때 i번째 그리드에서 발생한 CDR 수를, $R_v(t)$는 시간 t일 때 기지국의 서비스 영역 안에서 발생한 CDR 수를, A_v는 서비스 영역 v의 면적을, $A_{v \cap i}$는 서비스 영역 v의 면적과 i번째 그리드와 겹치는 면적을 의미한다. 그리고 k는 Telecom Italia에서 정한 임의의 상수로, 기업의 정보 공개와 서비스 이용자의 개인 정보 노출을 막기 위해 실제 인터넷 트래픽 사용 건수를 숨기는 역할을 하며, $S_i^{'}(t)$는 상수 k에 의해 가공되어 제공되는 데이터에 기록되는 CDR 수이다. 하지만 실제 건수가 아니라고 해도, 데이터의 경향성과 패턴은 그대로 남아있기 때문에 트래픽 경향 및 패턴을 학습하여 미래의 트래픽을 예측하는 본 연구에는 영향을 미치지 않는다.

2.3 convLSTM(convolutional LSTM)

convLSTM의 기본 구조인 RNN은 Hidden Node가 방향을 가지고 연결되어 순환구조를 이루는 인공신경망의 한 종류이다. 그러나 RNN이 갖는 구조적 문제인 Vanishing Gradient Problem[9]을 극복하기 위해 LSTM 방식이 고안되었다. LSTM은 RNN과 마찬가지로 문자, 음성 등 시간상으로 상관관계가 있는 데이터 처리에 유용하게 사용되기 때문에 시계열 데이터인 트래픽 데이터의 분석과 예측에 매우 적합하다. 구조적으로 convLSTM은 데이터의 시간적 관계를 반영하지 못하는 CNN과, 시계열 분석이 가능하지만 1차원 데이터만 학습 가능한 LSTM의 단점이 보완된 모델이다. 이 모델은 CNN처럼 컨볼루션 연산을 하기 때문에 2차원 구조의 데이터 형태를 입력으로 가질 수 있다. 하나의 state에서는 특정 시점의 공간 정보가 담겨있는 2차원 트래픽 데이터를 입력값으로 하고, 각 state마다 입력되는 데이터의 시간대를 순차적으로 변화시킴으로써 시공간적 특징을 학습한다. 그림 2의 (a)는 LSTM, (b)는 convLSTM 구조를 나타내며, 본

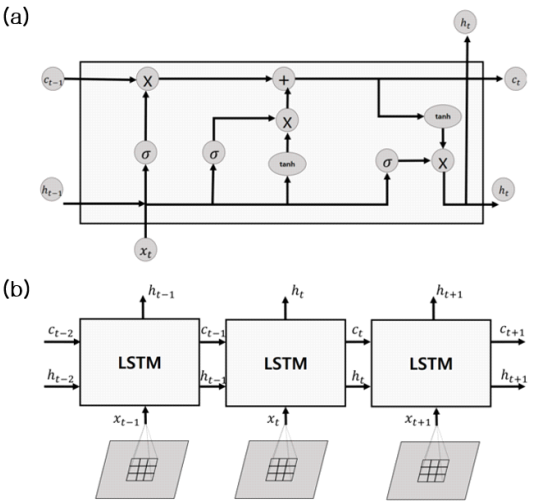

그림 2. (a) LSTM 구조 (b) convLSTM 구조
Fig. 2. (a) LSTM structure (b) convLSTM structure

표 1. 모델 구조와 파라미터
Table 1. Model structure and parameter

Model	Layer	Kernal(filters)	Kernal size	Stride
convLSTM	convLSTM2D	3	(3,3)	(1,1)
	convLSTM2D	1	(1,1)	

Model Parameter	
Learning Rate	1e-3
Epochs	300
Batch size	32

연구에서 사용한 convLSTM 모델 구조와 파라미터는 표 1의 값을 토대로 구현하였다.

2.4 클러스터링 알고리즘

일반적으로 산업지역과 주거지역의 트래픽 피크 시간대나 발생 양이 서로 다른 것처럼[10,11] 통신 활동은 지역마다 서로 다른 동향을 보일 것이다. 따라서 본 연구에서는 20x20 그리드 전체를 한꺼번에 학습시키지 않고, 서로 유사한 트래픽 양상을 보이는 그리드끼리 클러스터링하여 구분하고 클러스터별로 학습을 진행한다.

이때 클러스터링 방식은 흔히 알려진 K-means 클러스터링과 계층적 클러스터링(Hierarchical Clustering)[12] 중 후자를 사용하였다. K-means 클러스터링은 데이터를 K개의 클러스터로 묶는 알고리즘으로 초기에 데이터 중 임의의 위치에 K개의 중심점(센트로이드)을 배치하고 각 데이터들을 가장 가까운 중심점과 연결하는 방식으로 진행되기 때문에 초기에 랜덤하게 설정하는 중심점의 위치에 따라 클러스터링이 잘되지 않

을 수 있다는 단점이 있다. 반면, 계층적 클러스터링은 데이터 포인트를 각각 데이터를 1개씩 가지고 있는 클러스터라고 보고, 모든 데이터 간의 거리를 일일이 계산한 뒤 묶기 때문에 계산량이 많지만 클러스터링이 정교하게 이루어진다는 장점이 있다. 본 논문에서는 트래픽 데이터를 이 두 가지 기법으로 클러스터링 한 후 예측성능을 비교했고, 그림 3과 같이 계층적 클러스터링의 오차가 더 적다는 결론을 얻을 수 있었다. 밀란 데이터에 계층적 클러스터링이 적합함을 확인하였기에 이후 모든 시뮬레이션의 클러스터링 기법으로 계층적 클러스터링을 사용하였다.

비지도학습에서 사용되는 클러스터링의 일반적인 개념은 데이터 간 거리가 가까운 것끼리 묶어 하나의 클러스터로 분류하는 것이다. 그러나 오직 거리만을 가지고 본 연구에서 사용되는 시공간적 데이터를 클러스터링한다면 문제가 발생한다. 예를 들어, 산업지역과 주거지역의 경계 근처에 위치한 그리드들의 경우 그리드끼리 물리적 거리는 가까울지라도 속한 곳이 산업지역이냐 주거지역이냐에 따라 서로 다른 트래픽 양상을 보일 것이다. 따라서, 단순히 거리가 가깝다는 이유로 클러스터링한다면 서로 다른 트래픽 양상을 갖는 그리드끼리 하나의 클러스터로 묶여 학습되는 상황이 발생할 것이고, 모델의 예측 성능은 떨어질 것이다.

효율적인 클러스터링을 위해 그리드 간 거리 이외 클러스터링에 반영하기 위한 파라미터를 추가할 필요가 있다. 본 연구에서는 그리드 간 트래픽 양(Volume), 트래픽의 상관도(Correlation), 거리(Distance) 이 세 가지를 클러스터링 파라미터로 사용한다. 트래픽의 양, 상관도와 같이 직접적인 통신 활동의 정보를 그룹화의 척도로 사용하는 것이다.

세 가지 파라미터는 식 (3)~(5)과 같이 표현되며 각각 트래픽 양, 트래픽 상관도, 그리드 간 거리를 계산할 때 사용된다. 먼저 트래픽 양의 경우 각 그리드에서 발생하는 트래픽 양의 총합의 차를 측정한다. (3) 식의 X_t, Y_t는 그리드 번호 X, Y에서 시간 t일 때 발생한 트래픽이며, T_{XY}는 그리드 번호 X, Y에서 T 시간 동안 트래픽 양의 총합에 대한 차를 의미한다. 그다음으로 트래픽 상관도의 경우, 측정은 그리드에서 발생한 트래픽과 다른 그리드의 트래픽의 상관도를 계산함으로써 얻을 수 있다. 이때 상관도는 피어슨 상관계수(Pearson Correlation Coefficient)[13]를 사용하여 계산한다. (4) 식의 $\overline{X}$, $\overline{Y}$는 각각 그리드 번호 X, Y에서 T시간 동안 발생한 트래픽 양의 평균을, r_{XY}는 그리드 번호 X, Y의 피어슨 상관계수를, C_{XY}는 피어슨 상관거리를 의미한다. 트래픽 양의 유사성을 판단할 때 두 그리드의 트래픽 양의 차이를 사용하는 것과 트래픽 상관도를 판단할 때 피어슨 상관계수(γ_{XY})가 아닌 피어슨 상관 거리($C_{XY}=1-\gamma_{XY}$)를 사용하는 것은 학습 파라미터가 적용될 때 클러스터링 알고리즘이 이들을 거리와 같은 개념으로 인식하기 때문이다. 알고리즘은 그리드 사이 관계 값이 0에 근접할수록 그리드끼리 가깝다고 판단해 같은 클러스터로 묶는 원리이기 때문에, 트래픽 양이 유사하거나 트래픽이 상관관계를 가질 때 같은 클러스터로 묶일 확률을 높여주기 위해서는 그리드 간 관계를 0에 가깝게 만들어 주어야 한다. 마지막으로, 그리드 간 거리는 물리적 거리가 멀어질수록 두 그리드 간의 통신 활동의 유사도가 떨어지는 것이 일반적이라는 가정하에 반영되었다. 계산 과정은 두 그리드 $X(x_X, y_X)$, $Y(x_Y, y_Y)$ 사이 유클리디언 거리(Euclidian Distance)를 측정하는 (5)와 같은 방식으로 이루어진다. 3가지 파라미터들은 각각의 가중치(α, β, γ)를 갖고, 가중치 만큼의 중요도로 반영되어 클러스터링에 사용된다. 예를 들어, $(\alpha=0.25, \beta=0.5, \gamma=0.25)$의 경우 클러스터링에 트래픽 양($\alpha$)이 25%, 트래픽 상관도($\beta$)는 50%, 그리드 간 거리($\gamma$)는 25%의 비중으로 반영된다는 것을 의미한다. 최종적으로 파라미터가 가중치만큼 곱해져 선형결합되는 모습 (6)과 같은 형태가 되어 반영되며, 이때 $\alpha+\beta+\gamma=1$ 을 만족한다.

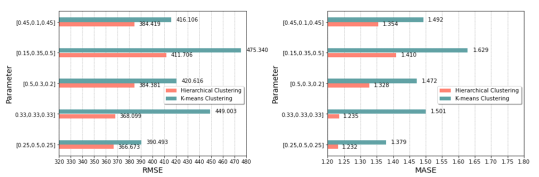

그림 3. K-means 클러스터링과 계층적 클러스터링 성능
Fig. 3. performance of K-means clustering and hierarchical clustering

$$T_{XY}=\left|\sum_{t}^{T}X_t-\sum_{t}^{T}Y_t\right| \tag{3}$$

$$r_{XY}=\frac{\sqrt{\sum_{t}^{T}\frac{(X_t-\overline{X})(Y_t-\overline{Y})}{T-1}}}{\sqrt{\sum_{t}^{T}\frac{(X_t-\overline{X})^2}{T-1}}\sqrt{\sum_{t}^{T}\frac{(Y_t-\overline{Y})^2}{T-1}}},\quad C_{XY}=1-r_{XY} \tag{4}$$

$$D_{XY}=\sqrt{(x_X-x_Y)^2+(y_X-y_Y)^2} \tag{5}$$

$$total = \alpha \cdot \frac{C_{XY}}{\| C_{XY} \|} + \beta \cdot \frac{T_{XY}}{\| T_{XY} \|} + \gamma \cdot \frac{D_{XY}}{\| D_{XY} \|} \quad (6)$$

가중치 (α, β, γ)의 값에 따라 양, 상관도, 거리가 클러스터링에 중요하게 반영되는 정도가 달라지기 때문에 성능 파악을 위해서는 여러 가지 가중치 값을 후보로 정할 필요가 있다. 만약 20x20 그리드를 N개의 클러스터로 클러스터링했을 때 n 번째 클러스터로 함께 묶인 그리드의 개수를 g_n라고 할 때, 본 연구에서 사용한 가중치 조합의 선정 기준은 다음과 같다.

- α, β, γ가 동등하게 33%씩 비중을 갖는 경우를 포함 (조합 2).
- α, β, γ중 하나가 50%의 비중을 갖는 경우를 포함.
- $N=2$의 경우, $g_1 > g_2$라면 $g_1 < 2g_2$를 만족. $g_1 < g_2$라면 $g_2 < 2g_1$을 만족.
- 서로 다른 임의의 값 $l, m, n(\leq N)$에 대해 $g_l < g_m + g_n$를 만족 $(N>2)$.

모든 파라미터의 비중을 33%씩 균등하게 한 1번째 조건과 하나의 파라미터가 절반의 비중을 차지하는 2번째 조건에 의해 선정된 조합들의 예측 결과를 비교함으로써 해당 파라미터의 비중을 키웠을 때 결과의 변화를 분석하고자 한다. 3번째, 4번째 조건은 클러스터 내 그리드 개수가 클러스터끼리 비교적 균등한 수를 갖게 하기 위한 것이다. 이는 하나의 클러스터가 과도하게 많은 그리드를 차지하는 경우를 방지하기 위함인데, 예를 들어 2개의 클러스터로 나누었을 때 전체 그리드 400개 중 399개가 하나의 클러스터, 나머지 1개가 다른 하나의 클러스터로 묶였다면 클러스터링하는 의미가 퇴색될 것이다. 따라서 클러스터마다 어느 정도 확보할 수 있는 그리드 개수를 보장하기 위해 3, 4번째 조건을 도입하였다.

위의 선정 조건에 맞추어 (α, β, γ) 값을 바꿔가며 실험해 본 결과, 표 2와 같은 5가지의 경우가 실험해 볼 수 있는 최적의 조합으로 선정되었다. 이후, 본 연구에선 궁극적으로 클러스터의 개수의 증가에 따라 예측 성능이 어떻게 변하는지 보기 위해서 클러스터 개수를 2개에서 5개까지 증가시키며 시뮬레이션을 진행하였다.

표 2. 클러스터링 가중치
Table 2. Clustering weight

조합	α(트래픽 양)	β(상관관계)	r(거리)
1	0.25	0.5	0.25
2	0.33	0.33	0.33
3	0.5	0.3	0.2
4	0.15	0.35	0.5
5	0.45	0.1	0.45

모델이 학습할 데이터셋을 전처리하고, 클러스터링한 뒤 학습시키는 모든 시뮬레이션 과정을 그림 4와 같이 나타내었다. 이 flowchart는 클러스터 개수를 4개로 하였을 때의 과정이다. 20x20의 그리드는 4개의 클러스터로 나뉘고, 각 그리드는 자신이 속한 클러스터의 학습의 순간에만 학습 데이터로 사용되고, 자신이 속하지 않은 클러스터의 학습 때엔 사용되지 않는다.

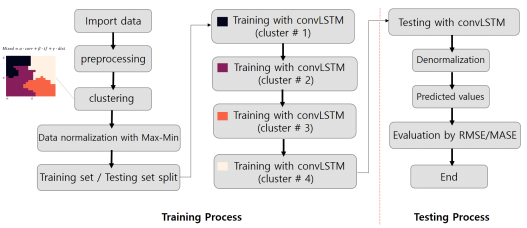

그림 4. 시뮬레이션 흐름도
Fig. 4. simulation flowchart

2.5 모델 평가

회귀 분석 모델의 예측 성능 평가에 많이 사용되는 RMSE(Root Mean Squared Error), MSE(Mean Squared Error), MAE(Mean Absolute Error)는 예측값과 실제값의 절대적 차이를 계산하여 오차가 얼마나 발생하는지를 보여준다. 하지만 이 3가지 평가 지표는 크기 의존적이라는 단점이 존재한다. 이 지표들은 그림 5와 같이 인구가 밀집한 그리드(52, 52)에서 발생한 트래픽과 인구가 많지 않은 그리드(64, 64)에서처럼 지역적 특징에 의해 트래픽 발생의 절대적인

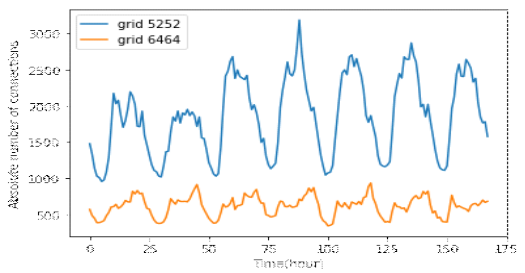

그림 5. 그리드 좌표 (52,52)와 (64,64)에서 7일동안 수집된 트래픽
Fig. 5. Traffic colletcted for 7 days at grid (52,52) and (64,64)

크기의 편차가 심한 경우엔 정확한 성능 평가의 지표로 사용되기 어렵다. 따라서 본 연구에서는 추가적인 평가 지표를 도입해야 한다.

MASE(Mean Absolute Scaled Error)는 그 의미대로 크기가 조정된 에러값이다. 이는 예측값과 실제값의 차이를 평소에 변동되는 평균 변동 폭으로 나눈 값으로, 변동 폭에 비해 얼마나 예측값과 실제값의 차이가 큰지 작은지를 측정하는 것이다. 만약 예측값과 실제값의 차이가 같더라도 변동폭이 큰 스케일의 데이터의 예측이 변동폭이 작은 스케일의 데이터의 예측보다 더 좋은 예측 성능을 지닌다고 판단한다. 결론적으로 본 연구에서 모델 평가를 위해 선정한 지표는 RMSE와 MASE 두 가지이며 식 (7)와 (8)으로 표현된다. Y_i는 실제값, $\hat{Y}_i$는 예측값이며 Y_t는 시간 t에서의 실제값이다.

$$RMSE = \frac{1}{n}\sum\sqrt{(Y_i - \hat{Y}_i)^2} \tag{7}$$

$$MASE = \frac{1}{n}\frac{\sum|Y_i - \hat{Y}_i|}{\sum|Y_t - Y_{t-1}|} \tag{8}$$

Ⅲ. 시뮬레이션

앞서 설명한 클러스터링 알고리즘을 가지고 데이터셋을 서로 다른 클러스터 개수로 클러스터링한 결

과를 그림 6에 나타내었다. 그림 6는 왼쪽 위, 오른쪽 위, 왼쪽 아래, 오른쪽 아래 순서로 전체 20x20의 그리드를 2, 3, 4, 5개의 클러스터로 클러스터링한 모습이다. 이때 클러스터링의 기준으로 사용된 가중치 조합은 $(\alpha = 0.45, \beta = 0.1, \gamma = 0.45)$이다. 학습은 클러스터별로 단독적으로 이루어지기 때문에, 하나의 클러스터 내 그리드들이 함께 묶여 학습될 때 다른 클러스터에 속한 그리드들은 그 학습에 영향을 주지 못한다.

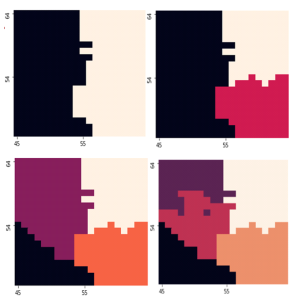

그림 6. 클러스터링 결과 $(\alpha = 0.45, \beta = 0.1, \gamma = 0.45)$
Fig. 6. Clustering results $(\alpha = 0.45, \beta = 0.1, \gamma = 0.45)$

그림 7은 앞선 그림 6의 결과 중 일부인 4개의 클러스터로 나뉜 전체 그리드 중 좌표 (50,55)에서 테스트 데이터(7일, 168시간)를 모델이 예측한 모습을 나타낸 것이다. 원본의 테스트 데이터(파랑선)의 상승과 하강 동향을 예측 트래픽(노랑선)이 유사하게 따라가고 있으며, 그 양 또한 어느 정도 예측함을 확인할 수 있다.

그림 8은 각각 그리드 좌표 (63,49)와 (50,64)에서 테스트 데이터(파랑선)에 대한 예측의 결과이다. 노란색 그래프는 2개의 클러스터로 나누어 학습시킨 모델의 예측 결과이고, 초록색 그래프는 4개의 클러스터로 나누어 학습시킨 모델의 예측 결과이다. 클러스터링을 2개로 한 모델의 예측 결과는 테스트 데이터의 상승과 하강의 동향은 예측하지만 트래픽 양에 대한 정확도는 떨어진다. 하지만 클러스터링을 4개로 한 모델의 예측 결과는 증감의 추세는 물론, 그 양도 매우 유사하게 예측해낸다.

그림 8의 결과를 평가 지표 RMSE와 MASE를 사용해 확인하면 그림 9과 같다. 가로축은 400개의 그리드를, 세로축은 예측 오류 값으로 위 그림은 RMSE, 아래 그림은 MASE를 의미한다. 400개의 그리드 전체적으로 클러스터 개수 2개인 경우(파랑)가 4개인 경우(주황)보다 예측 오류가 더 크게 나타난다.

앞서 시뮬레이션에 사용한 $(\alpha = 0.45, \beta = 0.1, r = 0.45)$조합에 대해 이번에는 클러스터링을 하지 않는 경우, 2개, 3개, 4개, 5개로 클러스터 개수를 하나씩 증가시켜서 모델을 학습시켜

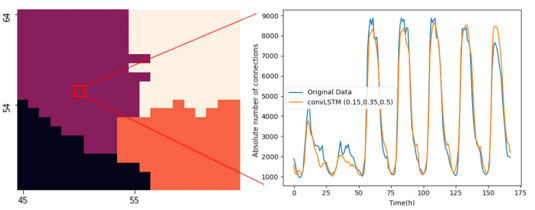

그림 7. 그리드 번호(50,55)의 트래픽 예측 결과
Fig. 7. Traffic prediction result of grid number (50,55)

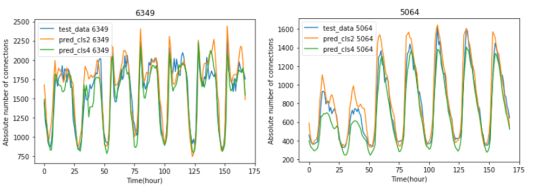

그림 8. 예측 결과 비교(클러스터 개수 2개와 4개)
Fig. 8. Comparision of prediction results(2 clusters and 4 clusters)

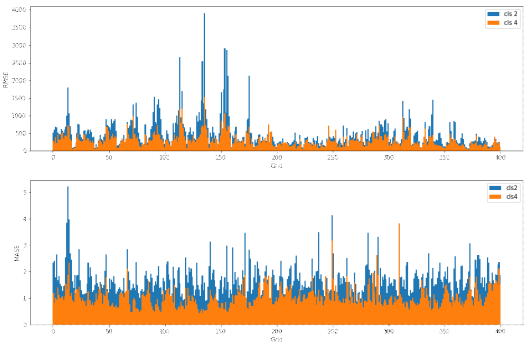

그림 9. RMSE와 MASE로 예측 오류 비교 (클러스터 개수 2개와 4개)
Fig. 9. Comparison of prediction errors with RMSE and MASE(2 clusters and 4 clusters)

보았다. 클러스터 개수가 증가함에 따라 변화하는 예측 오류를 그림 10으로 정리하였다. RMSE를 보면 클러스터링하지 않았을 때에서 2개, 3개, 4개로 점차 증가시켜 갈 때 19%, 10%, 12%씩 꾸준히 감소했다. 그러나 클러스터 개수를 4개에서 5개로 늘렸을 땐 오히려 RMSE가 13% 증가한다. MASE의 경우에도 마찬가지로 클러스터링 하지 않았을 때에서 2개, 3개, 4개로 점차 증가시켜 갈 때 14%, 8%, 17%씩 꾸준히 감소한다. 그러나 클러스터 개수를 4개에서 5개로 늘렸을 땐 오히려 12% MASE가 증가한다.

이번에는 클러스터링 개수에 따른 예측 오류를 표 2에서 정하였던 5가지 가중치 조합에 대해 확인해 보았다. 그림 11은 조합 5가지 전체에 대한 실험 결과이며, 클러스터링을 하지 않은 상태에서 5개까지 클러스터를 증가시킬 때 오류의 변화를 보였다. 클러스터 개수마다 가장 예측성능이 좋은 조합이 다르지만, 대체적으로 트래픽 양, 상관도인 α, β의 비중이 거리(γ)의 비중보다 높을 때 오류가 적었다. 이는 단순 거리 기반 클러스터링 방법보다 본 연구에서 추가한 파라미터들이 모델의 예측성능 향상에 기여함을 말한다. 클러스터 개수에 따른 성능을 보면, 5가지의 가중치 조합에 대해 대부분의 경우 클러스터 개수를 증가시킬

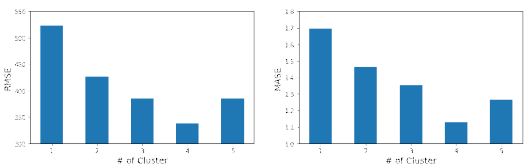

그림 10. 클러스터 개수에 따른 예측 오류 ($\alpha = 0.45, \beta = 0.1, \gamma = 0.45$)
Fig. 10. prediction error according to the number of clusters ($\alpha = 0.45, \beta = 0.1, \gamma = 0.45$)

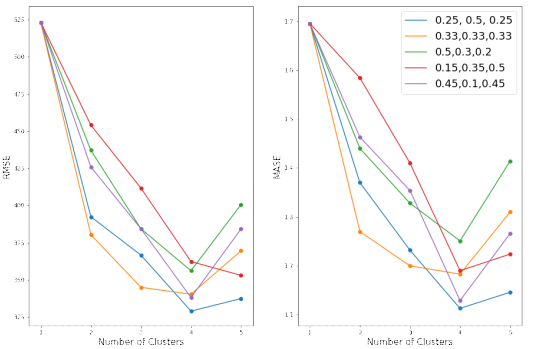

그림 11. 클러스터 개수에 따른 예측 오류
Fig. 11. prediction error according to the number of clusters

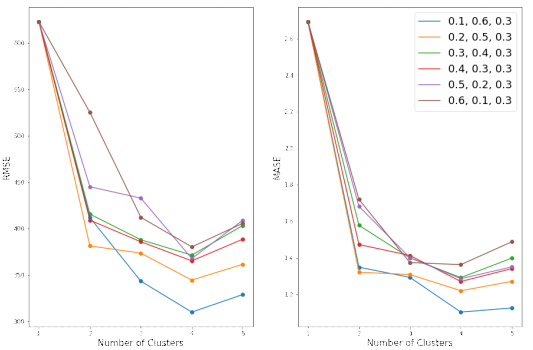

그림 12. 트래픽 양(α)과 상관도(β)의 영향 비교 (거리(γ) 파라미터 비중 고정)
Fig. 12. Comparison of the impact of traffic volume(α) and correlation(β). (Fixed the proportion of distance parameter(γ)

수록 RMSE(좌측)와 MASE(우측)가 감소함을 확인할 수 있다. 이는 유사한 양상의 트래픽 데이터를 갖는 그리드끼리 하나의 클러스터로 묶어 학습했고, 서로 상이한 양상을 갖는 다른 클러스터의 트래픽 데이터는 학습에 간섭으로 작용하지 않았기 때문이다. 하지만 이전 시뮬레이션의 결과와 마찬가지로, 4개에서 5개로 클러스터 개수를 증가시켰을 때는 예측 오류가 커지게 되면서 성능이 오히려 나빠짐을 확인할 수 있다. 이는 20x20개의 그리드를 점차 많은 클러스터로 나눌수록 하나의 클러스터로 같이 묶여 학습하게 되는 시공간 데이터 수가 점점 적어지기 때문에 클러스터가 일정 개수보다 많아지면 클러스터당 학습에 사용할 학습 데이터 수가 충분하지 않아 예측성능이 더 좋아지지 못한다고 추론해 볼 수 있다.

다음으로, 트래픽 양(α)과 상관도(β)가 예측성능에 끼치는 영향을 더 자세히 살펴보기 위해 거리(γ)의 비중을 0.3으로 고정하고 α와 β의 비중을 바꿔가며 실

험해보았다. 이번엔 기존 5가지 조합이 아니라 α와 β가 나머지 0.7의 분량을 0.1:0.6, 0.2:0.5, 0.3:0.4, 0.4:0.3, 0.5:0.2, 0.6:0.1로 나누어 갖는 6가지 조합을 새롭게 도입했고, 그림 12와 같은 결과가 나타났다. 비록 클러스터 개수마다 조합들의 성능 순위가 일관적이진 않지만, 트래픽 상관도(β)의 비중이 높은 파랑선과 노랑선의 성능이 일반적으로 더 좋고 트래픽 양(α)의 비중이 높은 고동선과 보라선의 성능이 더 나쁜 것을 확인할 수 있다. 이는 트래픽 양보단 상관도의 비중을 많이 반영해서 클러스터링하여 학습시킬 때 더 좋은 예측이 가능하다는 것을 말한다.

Ⅳ. 결 론

본 논문에서는 매 시간마다 그리드에서 수집되는 트래픽의 양, 상관도 그리고 그리드 간 거리의 유사관계에 따라 전체 그리드를 클러스터링하고, 클러스터별로 convLSTM 모델을 학습시켜 트래픽 예측을 수행하였다. 학습에 반영된 파라미터 3가지는 비중을 달리할 때마다 예측 성능이 달라졌는데, 트래픽 상관도, 트래픽 양, 그리드 간 거리 순으로 성능의 개선에 영향을 끼쳤다. 4개까지 클러스터 개수를 증가시켜 학습시킬 땐, 개수가 늘어날수록 예측 성능이 좋아졌다. 하지만 클러스터 개수를 5개로 늘렸을 땐 4개일 때보다 성능이 조금 떨어지는 것을 확인할 수 있었다. 결국엔 데이터셋을 클러스터링하여 회귀분석 모델을 클러스터별로 단독 학습시키는 경우, 클러스터 개수가 많아질수록 예측 성능이 좋아지지만 학습에 사용할 수 있는 데이터 수는 점점 적어지기 때문에 클러스터 개수를 아주 많이 늘리더라도 예측성능의 향상에 한계가 있을 것이다.

한편, 본 논문에서 사용한 밀란 데이터의 경우 모델의 학습을 위한 최적의 클러스터 개수가 4개지만, 다른 데이터였다면 그 개수는 달라질 것이다. 최적의 학습 데이터의 양은 문제의 복잡성과 알고리즘에 따라 다르고, 사용하는 데이터의 종류나 성격, 퀄리티에 따라서도 달라지기 때문이다. 따라서 밀란 데이터가 아닌 국내 데이터를 활용해 국내 통신 환경이 반영된 상황 속에서, 모델 학습을 위한 최적의 클러스터의 개수를 예측하는 방법을 최대한 일반화시킴과 동시에 국내 트래픽을 예측하는 것을 향후 목표로 삼아 연구를 진행해 볼 것이다.

References

[1] Cisco, *Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022* (2019), Retrieved Aug. 19, 2020, from https://s3.amazonaws.com/media.mediapost.com/uploads/CiscoForecast.pdf.

[2] D. Tikunov and T. Nishimura, "Traffic prediction for mobile network using Holt-Winter's exponential smoothing," *2007 15th Int. Conf. Softw., Telecommun. and Comput. Netw.*, pp. 1-5, Split, Croatia, Sep. 2007.

[3] H. W. Kim, J. H. Lee, Y. H. Choi, Y. U. Chung, and H. Lee, "Dynamic bandwidth provisioning using ARIMA-based traffic forecasting for Mobile WiMAX," *Comput. Commun.*, vol. 34, no. 1, pp. 99-106, Jan. 2011.

[4] D. Zhang, L. Liu, C. Xie, B. Yang, and Q. Liu, "Citywide cellular traffic prediction based on a hybrid spatiotemporal network," *Algorithms*, vol. 13, no. 1, pp. 1-16, Jan. 2020.

[5] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," in *IEEE J. Sel. Areas in Commun.*, vol. 37, no. 6, pp. 1389-1401, Jun. 2019.

[6] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "An adaptable neural-network model for recursive nonlinear traffic prediction and modeling of MPEG video sources," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 150-166, 2003.

[7] H. Feng and Y. Shu, "Study on network traffic prediction techniques," in *Proc. Int. Conf. Wireless Commun., Netw. and Mob. Comput.*, pp. 1041-1044, Beijing, China, Jun. 2005.

[8] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Scientific Data,* vol. 2, no. 1, pp. 1-15, Oct, 2015.

[9] H. Sepp, "The vanishing gradient problem during learning recurrent neural nets and problems solutions," *Int. J. Uncertainty, Fuzziness and Knowledge Based Syst.*, vol. 6, no. 2, pp. 107-116, Apr. 1998.

[10] L. Chen, D. Yang, D. Zhang, C. Wang, J. Li, and T. Nguyen, *J. Netw. Comput. Appl.*, pp. 59-69, 2018.

[11] Y. H. Yoon, "Analysis of network traffic with urban area characterstics for mobile network traffic model," *The KIPS Trans.: PartC*, vol. 10C, no. 4, pp. 471-478, Aug. 2003.

[12] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd Ed., WILEY, 2014.

[13] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763-1768, May 2018.

나 세 현 (Se-Hyeon Na)

2021년 2월 : 경희대학교 전자공학과 학사 졸업

2021년 3월~현재 : 경희대학교 전자정보융합공학과 석사과정

<관심분야> 이동통신, 5G

[ORCID:0000-0003-1757-1566]

김 영 준 (Young-Jun Kim)

2021년 2월 : 경희대학교 전자공학과 학사 졸업

2021년 3월~현재 : 경희대학교 전자정보융합공학과 석사과정

<관심분야> 5G, 딥러닝

[ORCID:0000-0001-7884-0913]

유 현 민 (Hyeon-Min You)

2021년 2월 : 경희대학교 전자공학과 학사 졸업

2021년 3월~현재 : 경희대학교 전자정보융합공학과 석사과정

<관심분야> 이동통신, 딥러닝

[ORCID:0000-0001-6385-2655]

안 희 준 (Hee-Jun Ahn)

2021년 2월 : 경희대학교 전자공학과 학사 졸업

2021년 3월~현재 : 경희대학교 전자정보융합공학과 석사과정

<관심분야> 5G, Network

[ORCID:0000-0003-4630-6163]

문 정 모 (Jung-Mo Moon)

1992년 2월 : 홍익대학교 전자계산학과 학사 졸업
1994년 2월 : 홍익대학교 전자계산학과 석사 졸업
2004년 2월 : 충남대학교 컴퓨터 과학과 박사 졸업
1994년~현재: 한국전자통신연구원 책임연구원/기술총괄
2021년~현재 : 5G포럼 스몰셀WG 부위원장
<관심분야> 5G, 스몰셀 기지국
[ORCID:0000-0002-0594-1227]

홍 인 기 (Een-Kee Hong)

1989년 2월 : 연세대학교 전기공학과 학사 졸업
1991년 2월 : 연세대학교 전기공학과 석사 졸업
1995년 8월 : 연세대학교 전기공학과 박사 졸업
1995년~1999년 : SKT 선임연구원
1999년~현재 : 경희대학교 전자공학과 교수
2012년~현재 : 미래창조과학부 주파수 정책 자문위원
2013년~현재 : 5G 포럼 주파수 위원회 위원장
2014년~현재 : 국무조정실 주파수 심의위원
2018년~현재 : 한국통신학회 부회장
2018년~현재 : 과기정통부 전파정책 자문위원
2021년~현재 : 위성통신포럼 주파수 위원회 위원장
<관심분야> 5G, 이동통신
[ORCID:0000-0001-6777-7058]