이동통신 트래픽 예측을 위한 클러스터링 기법

나세현, 김영준, 유현민, 안희준, 홍인기
경희대학교 전자정보융합공학과
2015104016, donjomyo, yhm1620, hmk6160, ekhong@khu.ac.kr

Clustering Method for Mobile Traffic Prediction

Se-Hyeon Na, Young-Jun Kim, Hyeon-Min You, Hee-Jun Ahn, Een-Kee Hong
Dept. of Electronics and Information Convergence Engineering, Khung Hee University

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 논문에서는 나름의 트래픽 유사성에 대한 기준을 정하여 유사성에 의한 클러스터링 알고리즘을 통해 클러스터 단위로 트래픽 데이터를 학습시킨다. 본 논문은 연구에 사용한 유사도 기반 클러스터링 방법에 대해 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 나눌경우 오히려 예측 오류가 증가하였다.

1. 서 론

오늘날 트래픽이 폭발적으로 증가함에 따라[1] 네트워크의 효율적인 운영에 대한 필요성이 증가하고 있다. 모바일 데이터 트래픽은 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 예측을 위한 방법으로는 ARIMA모델[2]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[3-4]을 이용한 예측 방법이 더 적합하다. 본 연구에서는 딥러닝 알고리즘 중 convLSTM을 사용하여 트래픽 예측을 수행한다.

2014년 Telecom Italia에서“Telecom Italia Big Data Challenge”을 위해 제공한 데이터를 학습에 사용할 데이터로 사용하였다. 이 데이터는 이탈리아 밀란(Millan) 시를 공간적으로 100x100으로 나눈 그리드에서 일정 기간동안 수집되는 트래픽 데이터를 담고 있다. 이 중, 트래픽이 매우 활발한 밀란의 중심부 20x20 사이즈의 그리드에서 50일 동안 1시간 간격으로 수집되는 트래픽 데이터를 학습과 예측에 사용하였다. 50일 분량의 데이터 중 43일 분량의 데이터를 훈련 데이터로, 7일 분량의 데이터를 테스트 데이터로 구성하였다.

본 논문에서는 클러스터 단위로 트래픽 데이터를 학습시키고 예측하는 알고리즘을 제안한다. 클러스터링을 통한 인공지능 모델 학습은 [3]에서 수행된 바 있으나, [3]에서는 메타 데이터, 교차 차원 데이터와 같이 트래픽 데이터 이외의 데이터를 사용했다. 반면 본 연구에서는 트래픽 데이터와 관련된 클러스터링 파라미터들을 선정해 클러스터링을 수행하였다. 또한, 클러스터링 개수에 따른 예측 성능을 RMSE(Root Mean Square Error), MASE(Mean Absolute Scaled Error)를 통하여 분석하였다.

2. convLSTM (Convolutional LSTM)

시계열 데이터 분석에 적합한 LSTM의 구조가 확장된 convLSTM을 사용한다. convLSTM은 데이터의 시간적 관계를 반영하지 못하는 CNN과, 시계열 분석이 가능하지만 1차원 데이터만 학습 가능한 LSTM의 단점이 보완된 모델이다. ConvLSTM은 CNN처럼 컨볼루션 연산을 하기 때문에 2차원 구조의 시계열 데이터 형태를 입력으로 가질 수 있다. 이는 시공간 데이터를 사용하기에 적합한 기법이며 표 1과 같은 파라미터로 학습에 이용했다.

표 1. 모델 파라미터

Model	Layer	Kerals(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

3. 클러스터링 알고리즘

일반적으로 산업지역과 주거지역의 트래픽 피크 시간대가 다른 것처럼 통신 활동은 지역마다 서로 다른 동향을 보일 것이다. 따라서 20x20 그리드 전체를 한꺼번에 학습시키지 않고, 서로 유사한 트래픽 양상을 보이는 그리드끼리 클러스터링하여 구분하고 클러스터 별로 학습을 진행한다.

효율적인 클러스터링을 위해 그리드 간 트래픽의 상관도(Correlation), 트래픽의 양(Volume), 거리(Distance) 이 세 가지를 파라미터로 사용한다. 그리드 간 거리가 가깝다고해서 유사한 통신 활동 양상을 보인다고 보장을 할 수 없기 때문에 거리 뿐만 아니라 트래픽 상관도, 트래픽 양과 같은 통신 활동의 정보를 반영하여 클러스터링한다.

세 가지 파라미터는 식 (1)~(3)과 같이 표현되며 각각 트래픽 상관도(1), 트래픽 양(2), 그리드 간의 거리(3)를 나타낸다. 파라미터들은 각각의 가중치(α,β,γ)를 갖고, 가중치만큼의 중요도로 반영되어 클러스터링에 사용된다. 파라미터가 가중치만큼 곱해져 선형결합된 식 (4)와 같은형태로 반영되며, 이때 α+β+γ=1 이다.

$$r_{XY}=\frac{\sum_t^T \frac{(X_t-\bar{X})(Y_t-\bar{Y})}{T-1}}{\sqrt{\sum_t^T \frac{(X_t-\bar{X})^2}{T-1}}\sqrt{\sum_t^T \frac{(Y_t-\bar{Y})^2}{T-1}}}\ ,\ C_{XY}=1-r_{XY} \quad (1)$$

$$T_{XY}=\left|\sum_t^T X_t-\sum_t^T Y_t\right| \quad (2)$$

$$D_{XY}=\sqrt{(x_X-x_Y)^2+(y_X-y_Y)^2} \quad (3)$$

$$total = \alpha \cdot \frac{C_{XY}}{\| C_{XY} \|} + \beta \cdot \frac{T_{XY}}{\| T_{XY} \|} + \gamma \cdot \frac{D_{XY}}{\| D_{XY} \|} \quad (4)$$

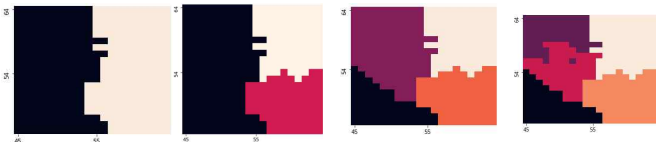
$X(x_X, x_Y)$, $Y(y_X, y_Y)$: 그리드 번호

표 2. 클러스터링 가중치

조합	α (트래픽 양)	β (상관관계)	γ (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

표 2와 같이 가중치(α,β,γ)의 조합으로 5가지를 선정했고, 이를 토대로 클러스터링을 수행한다. 이때, 클러스터의 개수의 증가에 따라 예측 성능이 어떻게 변하는지 보기위해 클러스터 개수를 2개에서 5개까지 증가시키며 시뮬레이션하였다.

4. 시뮬레이션 결과



a. 2 개 b. 3 개 c. 4 개 d. 5 개

그림 1. 클러스터링 결과 (α=0.15, β=0.35, γ=0.5)

그리드마다 수집된 트래픽 데이터를 가지고 앞서 설명한 클러스터링 알고리즘을 토대로 클러스터링한다. 그림 1-a,b,c,d는 20x20의 그리드가 2개, 3개, 4개, 5개의 클러스터로 클러스터링 된 모습이다.

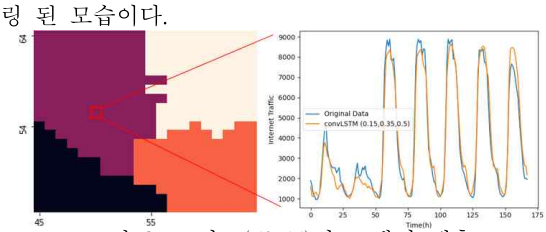


그림 2. 그리드(50,55)의 트래픽 예측

그림 2는 5개의 가중치 조합 중, (α=0.15, β=0.35, γ=0.5) 조합 시나리오에 의해 4개의 클러스터로 나뉜 영역 중, 그리드(50,55)가 테스트 데이터(7일, 168시간)를 예측한 모습이다. 원본의 테스트 데이터(파랑)의 동향을 예측 트래픽(노랑)이 유사하게 예측해낸다.

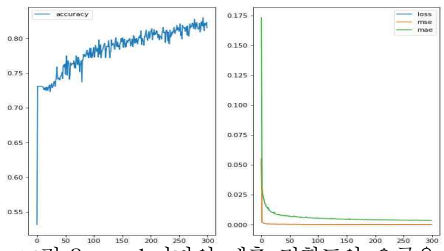


그림 3. epoch마다의 예측 정확도와 오류율

설정한 epoch = 300에 다다를 때까지 모델이 학습되면서 각 epoch마다 학습에 대한 예측 정확도와 오류율을 그림 3에 보였다. 그림 3의 예시는 (α=0.15, β=0.35, γ=0.5) 조합 시나리오로 클러스터를 4개로 나누어 학습시키는 과정이다. epoch이 작은 순간에는 학습 초기여서 예측 정확도가 낮고 오류는 크지만 epoch 300에 다다를수록 예측 정확도는 80% 이상으로 높아지고 오류율은 급감함을 볼 수 있다.

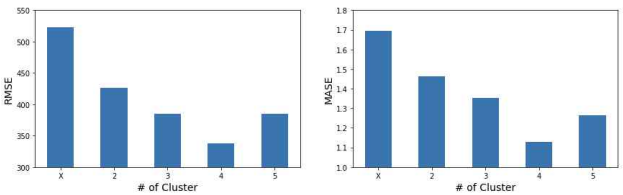
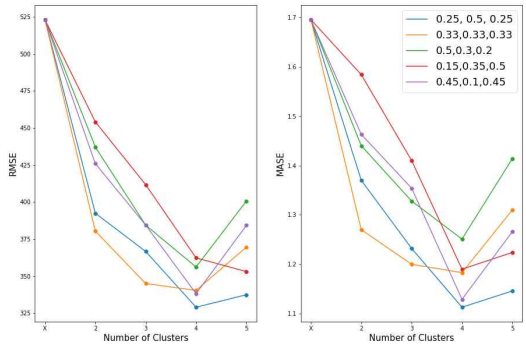


그림 4. 클러스터 개수에 따른 예측 오류(α=0.45, β=0.1, γ=0.45)

그림 4는 클러스터 개수가 증가함에 따라 변화하는 예측 오류를 클러스터링 가중치 조합 중 (α=0.45, β=0.1, γ=0.45)인 하나의 조합에 대해 보인 것이다. 클러스터링 개수가 늘어날수록 예측 오류(RMSE, MASE)가 줄어들다가, 클러스터링을 5개로 할 경우 4개의 경우보다 예측 오류가 증가한다.



a. b.

그림 5. 클러스터 개수에 따른 예측 오류

그림 5은 클러스터링 가중치 조합 5가지 전체에 대한 시뮬레이션 결과로, 클러스터 개수를 2개에서 5개까지 증가시킬 때 예측 오류의 변화를 보였다. 5가지의 가중치 조합에 대해 대부분의 경우 클러스터 개수를 증가시킬수록 RMSE(5-a)와 MASE(5-b)가 감소한다. 하지만 그림 4에서와 마찬가지로, 4개에서 5개로 클러스터 개수를 증가시켰을 때는 예측 성능이 떨어져, 오류가 좀 더 발생하는 경향성을 보인다.

5. 결론

본 논문에서는 그리드 간 트래픽의 상관도, 양, 거리의 유사관계에 따라 20x20 전체 그리드를 클러스터링하고, 클러스터별로 convLSTM 모델을 학습시켜 트래픽 예측을 수행하였다. 2개에서 4개까지 클러스터 개수를 증가시켜 학습시킬 땐, 클러스터 개수가 늘어나면 늘어날수록 예측 성능은 좋아졌다. 이는 유사한 양상의 트래픽 데이터를 갖는 그리드끼리 하나의 클러스터로 묶여 학습했기 때문에 서로 상이한 양상을 갖는 다른 클러스터의 트래픽 데이터가 학습에 간섭으로 작용하지 않기 때문이다. 하지만 클러스터 개수가 5개일 경우엔 4개일 때보다 예측 성능이 조금 떨어지는 것을 볼 수 있었는데, 이는 20x20개의 그리드를 점차 다수의 클러스터로 클러스터링 할수록 하나의 클러스터로 같이 묶여 학습하게 되는 시공간 데이터 수가 점점 적어지기 때문에 클러스터가 일정 개수보다 많아지면 클러스터 당 학습에 사용할 학습 데이터 수가 충분하지 않아 예측 성능이 더 좋아지지는 못한다고 볼 수 있다.

6. Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음 (IITP-2021-2016-0-00291)

7. 참고 문헌

[1] Forecast, G. M. D. T. (2019). Cisco visual networking index: global mobile data traffic fore-cast update, 2017–2022. Update, 2017, 2022.
[2] Kim, H. W., Lee, J. H., Choi, Y. H., Chung, Y. U. & Lee, H. (2011). Dynamic bandwidth provi Sioning for Mobile WiMAX. Computer Communi cations, 34(1), 99-106.
[3] Zhang, D., Liu, L., Xie, C., Yang, B., & Liu, Q. (2020). Citywide Cellular Traffic Prediction Based on a Hybrid Spatiotemporal Network. Algorithms, 13(1), 20.
[4] Zhang, C., Zhang, H., Qiao, J., Yuan, D., & Zhang, M. (2019). Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. IEEE Journal on Selected Areas in Communications, 37(6), 1389-1401.