라지-스케일 GPS 차량 궤적 빅데이터를 위한 딥러닝 연구

이홍석[1] 정희진[2] 배상훈[2]
[1]한국과학기술정보연구원 슈퍼컴퓨팅본부
[2]부경대학교 공간정보시스템공학과
hsyi@kisti.re.kr, hjjung1974@gmail.com, sbae@pknu.ac.kr

Large-scale Traffic Flow Prediction With Big Data : A Deep Learning Approach

영문제목

Hongsuk Yi[1] Heejin Jung[2] Sang-hoon Bae[2]
[1]Korea Institute of Science and Technology Information, Supercomputing Center
[2]Pukyong National University, Department of Spatial Information Engineering

요 약

지난 몇 년 동안 스마트폰과 내비게이션의 GPS 단말기 보급으로 실시간 교통 데이터는 폭발적으로 증가하고 있다. 실제로 교통통신 네트워크는 빅데이터 시대로 접어들었다. 하지만 현존하는 많은 차량흐름 예측 모델은 회귀분석을 통한 모델 중으로, 실제 교통 빅데이터를 분석하기에는 많은 어려움이 있다. 많은 방법론이 개발되고 있지만, 딥러닝 알고리즘은 복잡한 라지-스케일 교통 빅데이터를 효율적으로 다룰 수 있는 기법의 하나다. 본 논문에서는 서울시 강남구 지역 교통 혼잡 흐름 모형을 개발하기 위해 GPS 차량 궤적데이터를 분석하였다. 또한, 딥러닝 이론이 어떻게 교통 빅데이터 모형 분석에 적용될 수 있는지 그 방법론을 제시하고자 한다.

1. 서 론

최근 실시간 교통데이터는 GPS 단말기 보급으로 등으로 더욱 다양해지고 방대해지고 있다. 이에 따라 스마트폰, 내비게이션, 그리고 하이패스 단말기의 GPS를 활용한 연구가 활발히 진행되고 있다[1]. 이처럼 최근 교통 상황에 대한 자료가 풍부해지면서 여러 가지 가정에 기반을 둔 교통 상황 예측 이론기법보다는 데이터 기반 예측 모형이 널리 활용되고 있다[2].

교통 분야에서 인공신경망 학습모형은 비선형 통계적 기법으로 향상된 지도학습 방법을 통해 최적의 결과를 도출한다. 또한, 교통 상황 예측 시 사고 및 정체현상에서도 우수한 예측력을 가지고 있다[3]. 하지만 모형을 개발하는 것이 복잡하며 검정하기 어려운 단점이 있다. 최근에 개발된 딥러닝은 빅데이터 분석에 매우 효율적이라고 알려졌다. 그 이유는 빅데이터를 하나하나 분류해서 구분할 필요가 없고, 또한 전처리(Pre-training) 알고리즘을 통해 정형 학습을 비정형 학습이 가능하도록 해주기 때문이다[4].

슈퍼컴퓨팅 측면에서 그래픽처리장치(GPU)를 이용한 CUDA(Compute Unified Device Architecture) 병렬프로그램이 가능해지면서 딥러닝이론이 급속도로 발전하고 있다. 이러한 딥러닝은 신경망 이론 개발 분야뿐만 아니라, 실제 음성인식, 자연어 처리, 영상처리, 음악, 미술 등 다양한 분야에서 적용되고 있다. 이러한 연구 분야는 CNN(Convolution Neural Network) 알고리즘이 대표적으로 사용되고 있으며, 특히 이미지 분석에서 탁월한 성능을 보여준다. 하지만 딥러닝 이론을 활용한 대도시 빅데이터를 분석 연구는 매우 초기 단계이다[3]. 실제로 교통분야에서 의미 있는 딥러닝 연구 결과는 거의 없는 형편이다. 본 논문에서는 서울시 GPS 차량궤적 빅데이터를 확보하였다. 서울시 전체 자료는 너무 방대해서 강남구 지역을 중심으로 GPS 차량궤적데이터를 통계처리 하였다. 그 이유는 딥러닝을 위한 모형 개발을 위해 교통 자료 수집률이 높은 지역을 대상으로 하기 때문이다. 실제로 교통 혼잡 예측을 효율적으로 하기 위해 지표도 지표(TPI)를 개발하여 교통 혼잡 패턴을 분석하였다. 마지막으로 강남구 교통 빅데이터 패턴 분석과 혼잡 예측 모형 개발을 위하여 딥러닝 시뮬레이션 방법을 제시하고자 한다.

2. 연구내용 및 방법

2.1 서울시 강남구 GPS 차량궤적데이터

본 연구에서 사용된 GPS 차량궤적데이터는 도착시각 기준 링크통행속도이며 차량에 내장된 built-in GPS 단말기를 통해 수집하였다. 각 차량의 GPS 단말기에서 1초마다 위치정보가 생성된 후 단말기에 내장된 링크정보에 GPS 위치데이터를 지도 매칭시켜 차량이 통과한 링크통행시간을 산출한다. 이 링크통행속도는 차량이 링크를 완전히 통과한 시점에 생성되므로 도착시각 기준 링크통행시간이다. 사용된 링크통행속도는 링크정보의 링크 길이를 도착시간 기준 링크통행시간으로 나누어 산출한다.

사용된 자료는 서울 강남구에 있는 간선도로로서 국가교통정보센터를 통해 배포되고 있는 국가표준 노드링크에 등록된 노드와 링크에서 2013년 4월 1일에서 2013년 9월 30일에 수집된 6개월 데이터이다. 강남구에는 1,480여 개 링크가 국가표준 노드링크에 등록되어

있다. 이중 GPS 차량궤적데이터가 수집된 링크는 855개 링크였다. 그림1은 강남구에서 수집된 링크를 보여준다. 공휴일을 제외한 6개월 동안 수집된 자료는 160,084대(45%)의 차량으로부터 수집한 총 46,408,741건의 링크통행시간 데이터이다. 하루 평균 10,168대의 차량이 263,686건의 링크통행시간 자료가 수집되었다.

그림-1 강남구 지역에서 1480개 표준노드링크 중 수집된 GPS 차량궤적 데이터. 파랑색 선은 GPS 데이터가 수집된 855개 링크를 나타내고 있다.

그림 2에서 강남구에서 수집된 GPS 차량궤적데이터를 5분 주기로 수집할 경우 각 시간대에 수집된 자료의 수를 상자 플롯으로 표출하였다. 검은색 영역은 중위 50%의 수집 건수의 분포를 나타내며 굵은 원은 수집 건수의 중앙값을 나타낸다. 5분 주기로 자료를 수집할 경우 새벽 2시에서 7시 사이에는 수집 건수가 매우 적었으며 그 외 시간대는 평균 1,015건 정도의 GPS 차량궤적데이터가 수집되었다. 5분 주기로 수집할 경우 새벽 시간대에 결측이 발생할 확률이 매우 높으므로 연구범위를 오전 7시 이후로 제한하였다.

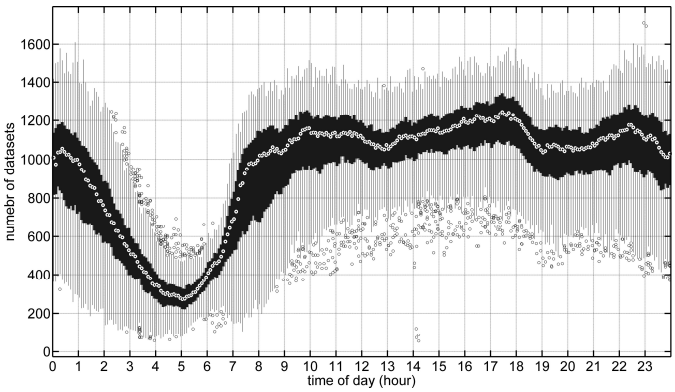

그림 2 강남구의 시간대 별 링크통행시간데이터의 수집건수 분포

2.2 서울시 강남구 교통흐름 패턴

도심부 도로망에서 일반적으로 첨두시간(Peak time)과 비첨두시간(Non-peak time)으로 구분되는 소통상황이 반복된다. 첨두시간이란 교통량이 극도로 증가하여 도로의 용량에 육박하거나 초과하여 통행속도가 극도로 떨어지는 현상이 발생하는 시간대를 의미하며 비첨두시간이란 첨두시간을 제외한 나머지 시간을 의미하며 첨두시간에 비해 원활한 소통상황이 유지되는 시간대를 의미한다. 도심부에서 첨두시간은 출퇴근시간대를 의미하지만, 주변 지역의 특성에 따라 다양한 시간대에서 발생한다. 보편적으로 첨두시간은 출근시간대인 오전 7시에서 10시 사이에 관찰되는 오전첨두시간과 퇴근시간대인 오후 4시에서 8시 사이에 관찰되는 오후첨두 시간으로 구분된다.

단속류의 경우 링크를 통과한 차량의 통행시간에는 신호지연시간이 포함하고 있으므로 일정 시간 동안 수집된 차량의 링크통행시간은 신호에 걸린 차량의 속도와 신호에 걸리지 않은 차량의 속도로 나눌 수 있다. 비첨두시간대의 두 속도는 두 개 이상의 그룹으로 분리될 수 있지만 첨두시간대의 경우 그룹의 분류가 어렵거나 신호의 걸린 차량의 데이터가 빠질 수 있다. 본 연구에서 사용된 GPS 차량궤적 데이터는 시속 8km/h 이하의 속도를 이상값으로 처리하여서 첨두시간대 신호에 걸린 차량은 이상값으로 제거된 경우가 있다.

지체도지표(Traffic Performance Index, TPI)는 도로의 자유속도를 기준으로 속도의 오차 비율을 계산하여 지체도를 측정하는 지표이다.

$$TPI = \frac{V_f - V_i}{V_f}$$

여기서, V_f는 자유속도(Free-flow Speed)로서 제한속도의 1.5배를 사용하였다. V_i는 현재의 평균속도이다. TPI는 0에서 1사이의 값으로서 0은 완전자유속도, 1은 완전 정체상황(Traffic Jam)으로 해석할 수 있다.

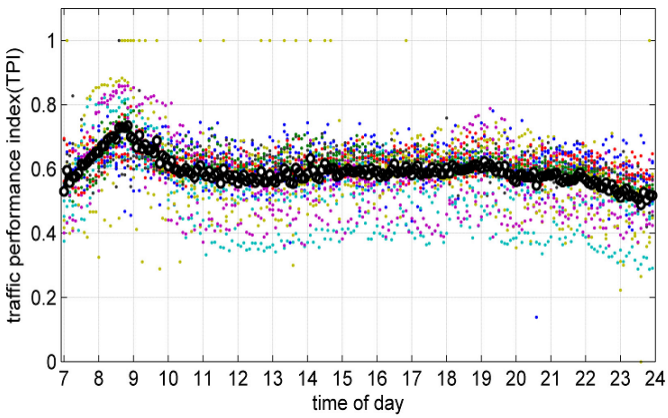

그림-3 소통상황 패턴의 유형으로 오전 첨두시간 (7시 ~ 10시) 패턴

그림 3은 6개월간 수집된 자료를 지체도 지표로 변환하여 패턴화한 결과이다. 사용된 자료는 오전 7시에서 10시 사이 오전첨두 시간에 교통흐름 패턴이 형성하여 데이터 결측 처리에 사용하였다. 일반적으로 TPI 값이 0.4 이하이면 원활, 0.4에서 0.6을 지체, 0.6에서 0.8을 정체, 0.8 이상을 심각한 정체로 분류한다. 따라서 강남구 오전 첨두시간은 TPI 값이 0.6~0.8 사이에 집중된 뚜렷한 정체 패턴을 볼 수 있다.

3 신경망을 이용한 교통데이터 분석 방법

근 신경망 층이 깊어지면서 RBM과 RNN의 장점을 결합하면 좋은 성능을 얻을 것이란 전망으로 딥러닝의 RBM-RNN 아키텍처가 등장했다[2]. RNN-RBM은 시계열 절차의 분포를 예측하기 위한 통계물리의 에너지 함수를 사용한 모델이다. 주어진 시간 스텝 에서 특징은 나타내는 는 고차원 벡터이다. RNN-RBM에서 바이어스 벡터와 는 visible 층과 은닉층의 벡터로 RNN 모델에서 이전 시간 단계의 은닉유닛으로 표현된다. 웨이트 행렬은 RNN와 RBM 모델에서 얻는다. 위의 과정은 아래 식으로 표현되고,

$$b_v^{(t)} = b_v + W_{uu}u^{(t-1)}$$
$$b_h^{(t)} = b_h + W_{uh}u^{(t-1)}$$

은닉층에서 은닉유닛의 활성은 아래와 같이 계산된다.

$$u^{(t)} = \tanh(b_u + W_{uu}u^{(t-1)} + W_{vu}v^{(t)})$$

과거 교통 패턴과 서울시 라지-스케일 교통 네트워크의 링크 수를 고려하면, 교통 흐름 예측은 매우 고차원 시계열 학습문제가 되며, 재귀신경망 (Recurrent Neural Network)를 적용할 수 있다. 각각의 도로 링크에서 교통상황은 신경망 알고리즘에 적용하기 위하여 혼잡과 정상으로 구분할 수 있다. GPS 속도가 20km/h 이하이면 혼잡으로 1이고, 20km보다 크면 정상 흐름으로 0으로 지정하였다.

그림4는 딥러닝 계산 결과를 검증하기 위해 인공신경망 RNN에서 강남구 GPS 데이터를 시뮬레이션했다. 실제로 강남구 855개 링크에서 5분 단위 6개월 데이터는 매트랩에서 계산하기에는 100GB 이상 큰 메모리를 요구한다. RNN 성능 검증을 위해 100개 링크를 1개월 데이터로 줄이고 계산을 수행하였다.

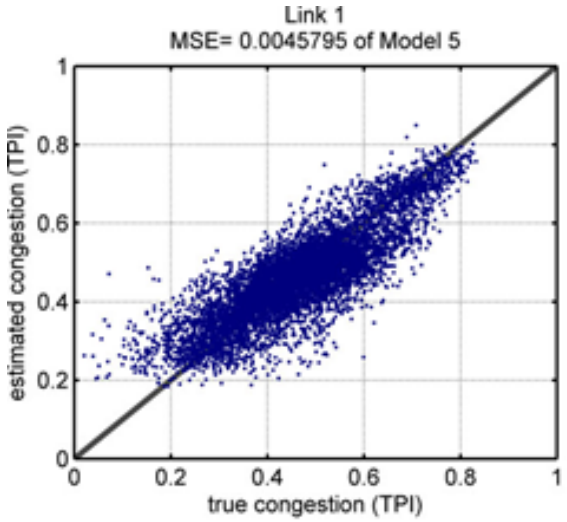

그림-3 서울시 강남구 논현로 정체도 산포도

사용한 시스템은 인텔 Xeon CPU E5-2697 2.6GHz 2개 소갯 28코어를 사용했다. 위의 계산은 매트랩에서 28개 성능확장성을 보여주는 병렬계산 결과이다. 메모리는 32GB이고 GPU는 NVIDIA Geforce GTX 960 2GB memory와 1,024개 스레드를 사용할 수 있다.

4. 결론 및 향후 연구과제

본 논문에서는 서울시 강남구 지역 GPS 궤적차량 빅데이터를 분석하여 교통흐름 패턴을 확인하였다. 강남구에서는 오전첨두 시간에 뚜렷한 교통흐름 정체 패턴을 확인하였다. 최근 빅데이터 기반 딥러닝 이론을 교통에 적용하는 방법을 제시하였다. 딥러닝 RNN-RBM을 이용한 모형은 계속 개발 중이며, 인공신경망 RNN을 이용한 초기 결과를 얻었다. 이 결과를 보면 수렵이 충분하지 않으며, 최적화가 필요하다. 이를 위해 RNN 모델을 최적화하고 사용된 데이터를 충분히 학습할 계획이다. 마지막으로 CUDA 병렬프로그래밍 기반 딥러닝을 위한 알고리즘으로는 THEANO[5]가 있으며, 이 Theano는 파이션 기반으로 RNN-RBM 시뮬레이션을 제공하고 있으며, 초기 시뮬레이션을 진행하고 있다. KERAS[6]는 딥러닝 라이브러리도 사용하여 GPS 기반 교통데이터를 분석할 계획이다. 그리고 딥러닝 계산 결과를 검증하기 위하여 Matlab의 역전파 알고리즘을 계산을 진행하고 있다.

5. 참고문헌

[1] 김태욱, 배상훈, 정희진, 차량궤적데이터를 활용한 도심부 간선도로의 돌발상황 검지, 한국ITS학회논문지, 제13권, 4호, 1, 2014

[2] Deep Learning Tutorials, DeepLearning 0.1 documentation http://deeplearning.net/tutorial/

[3] Xiaolei Ma, Haiyang Yu, Yunpeng Wang, Yinhai Wang, Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory, PLOS ONE, 10(3), 2015

[4] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang, Traffic Flow Prediction With Big Data: A Deep Learning Approach, IEEE transactions on intelligent transportation system, vol 16, no 2, 2015

[5] THEANO, http://deeplearning.net/software/theano/

[6] KERAS, https://github.com/fchollet/keras