Gaussian Process Regression을 이용한 요일 별 지하철 승객 수 추정

주현진 ○ 낭종호

서강대학교 컴퓨터공학과

hjjoo123@sogang.ac.kr, jhnang@sogang.ac.kr

Estimation of Subway Passenger Quantity Using Gaussian Process Regression

Hyunjin Joo○ Jongho Nang

Department of Computer Science and Engineering, Sogang University

요 약

본 논문에서는 기존의 Gaussian Process Regression (GPR)을 이용한 네트워크 내의 유동인구 분석 모델을 확장하여 서울 지하철의 Average Annual Daily Traffic (AADT) 추정 모델을 보인다. 이 추정 모델에서 추정 대상이 되는 역에 인접한 역의 승객 수를 통해 공간적 추정을 위한 데이터를 획득하고, 이 정보를 요일 별로 누적하여 추정의 신뢰도를 높였다.

1. Introduction[1]

유동 인구에 대한 분석은 도시 계획 및 보안/안전 시스템 설계, 마케팅 등 경제 시스템 전반에 응용될 수 있는 정보로서 최근 Data Science 및 Mining 분야에서 Intelligent Environment를 구축하는 하나의 수단으로 부각되어 왔다. 특히 교통 시스템의 전산화는 전례 없이 많은 위치 정보의 축적을 야기하였으며, 좀 더 다양하고 신뢰도 높은 Mobility Mining의 기회를 제공하고 있다.

그 중에서도 현대 도시 환경에서 주요 교통수단으로 자리 잡은 지하철은 유동 인구 분석에 좋은 소스가 되고 있다. 서울시의 경우, 2015년 3월 기준 인구 약 천 만 명을 대상으로 하루 평균 약 7백 만 건 이상의 지하철 승하차가 일어나고 있으며[2], 대중교통지향형 도시개발 정책에 따라 지하철 역세권 분석에 대한 중요성이 더욱 커지고 있다. 이로 인해 컴퓨터공학뿐 아니라 도시공학, 건축, 교통 및 경영 정보 시스템 등 다양한 분야에서 지하철 승객의 분포 및 이동 패턴에 대한 연구가 이루어졌으며, 비슷한 사례는 해외에서도 쉽게 찾아볼 수 있다[4, 9, 10].

유동량 분석은 기준이 되는 시간에 따라 연평균 하루 유동량인 AADT (Average Annual Daily Traffic)와 특정 기간 동안의 일 평균 유동량인 ADT (Average Daily Traffic)로 나눌 수 있다[4]. AADT는 다시 요일 별 또는 하루 중 시간대별 분석으로 나타낼 수도 있으며, Traffic 분석에 있어 중요한 지표이나 장비의 오류 또는 시공간적 한계로 인해 정확히 수집하기 어려운 경우가 많다.

예를 들어 M. May 등은 Street 단위의 Traffic Frequency 예측을 위한 kNN 알고리즘을 제안하였으며, 4일 동안의 비디오 분석을 통해 얻은 보행자 및 차량 데이터를 기반으로 호텔 및 식당의 수, 대중교통 인프라 등의 특성을 Domain Knowledge로 활용하였다[3]. T. Liebig 등은 지하철역이나 공원과 같이 폐쇄된 공간 내에서의 보행자 특성을 이용하여 Gaussian Process Regression (GPR) 으로 특정 공간 내의 보행자 수를 추정하였으며, 일반적인 보행자 움직임의 경향을 표현한 Trajectory Pattern을 Expert Knowledge로 활용하여 정확도를 높였다[4].

그러나 두 연구 모두 ADT 추정을 목표로 하여 시간에 따른 유동인구의 차이는 고려하지 않았으며, 물리적 한계로 인해 실험에 쓰인 데이터가 제한적이었다.

이에 본 논문에서는 서울열린데이터광장[1]에서 제공하는 약 46,000건의 지하철 게이트 집계 자료를 활용하여 GPR 모델을 기반으로 한 서울 지하철의 요일 별 승객 수 추정에 초점을 맞추었다. 구체적으로는 T. Liebig 등이 제시한 폐쇄 공간 내 보행자 추정 문제[4]를 지하철 노선 위의 승객 수 추정 문제로 변환하고 여기에 시간적 정보(요일)에 따른 분석(AADT)을 추가하는 모델을 제안하고자 한다.

2. Passenger Quantity Estimation of Unrecorded Area

2.1. Problem Definition and Goals

지하철 승객의 승하차 시간 및 위치에 대한 데이터는 수집 및 공개에 어려움이 있다. 현재 공식적으로 가용한 서울 지하철의 게이트 집계 자료는 해당 라인의 운영기관마다 따로

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(R7718-16-1004).

관리되며, 승차 인원과 하차 인원만 수집 가능하고 환승 구간의 승객 유동은 알 수 없다는 문제점이 있다.

따라서 본 연구는 데이터가 집계되지 않은 구간을 Point of Interest (POI)로 설정하고 다음과 같이 요일 별 승객 수를 추정하는 데 그 목적을 두었다. 이는 지하철 승객 수가 요일에 따른 패턴을 가질 것이라는 가정에 기반을 둔 것이다.

a. 공간적 연관성을 이용한 POI의 승객 수 추정
b. 추정된 값들의 분포를 시간적 정보(요일)에 따라 분석
c. 최종적으로 설정한 POI의 요일 별 평균 승객 수 추정

2.1.1. Experimental Goal

여기서 추정하고자 하는 POI는 실제로 데이터 집계가 이루어지지 않는 구간이므로, 실험적 평가를 위해 데이터가 가용한 구간에서의 Cross Validation을 통한 분석을 실험의 목표로 삼았다.

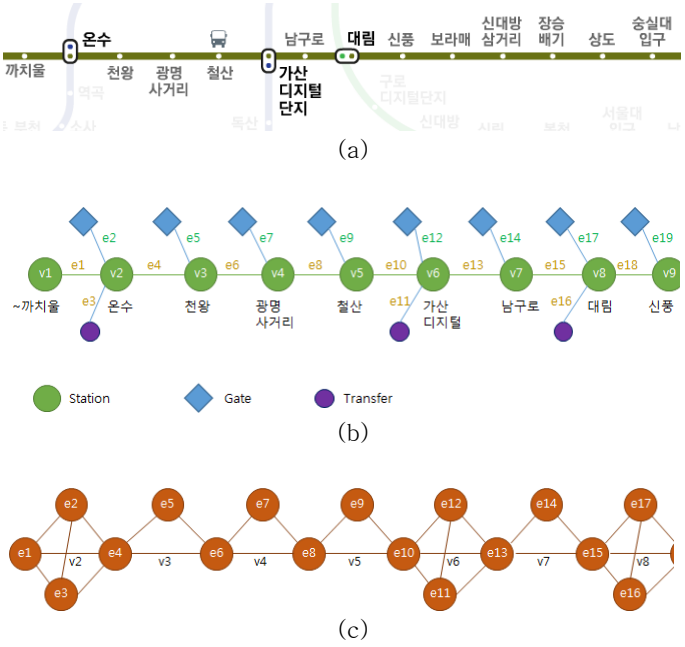

그림 1. (a) 서울 지하철 7호선 일부 (b) (a)의 graph model (부분) (c) (a)의 edge graph model (부분)

2.2. Data

Mining을 위한 데이터로는 서울도시철도공사에서 운영하는 5~8호선의 2009년부터 2014년까지 날짜/시간대별 승하차 승객 수를 사용하였다. 또한 문제를 단순화하기 위해 하루 중 가장 혼잡할 것으로 예상되는 출근 시간대(8시~9시)를 대상으로, [6]의 연구 결과 5~8호선에서 가장 잦은 이동 패턴을 보였던 7호선의 일부 구간을 임의로 설정하였다. Regression 대상인 승객 수는 입력 데이터 전체의 최대/최소값을 기준으로 [0, 1] 사이의 값으로 정규화 하였다. 그림 1. (a)는 실제 모델이 된 14개 역을 보여준다.

2.3. Gaussian Process Regression (GPR) Model

2.3.1. GPR for Spatial Prediction

지하철 노선은 일종의 폐쇄 공간으로 볼 수 있다. 이러한 폐쇄 공간의 특징은 정상적인 이용 환경을 가정했을 때 제한된 시간 동안만 운영되고, 공간 안에 들어온 사람은 일정 시간 후에 다시 빠져 나간다는 것이다. 쇼핑몰, 영화관 등의 건물 내부가 그 대표적인 예이며 공원이나 행사장 등의 제한된 외부 공간도 포함 될 수 있다. 이러한 환경에서는 다수의 사람들에게서 발견되는 특정한 이동 패턴이 있을 수 있는데, 이 이동 패턴은 Prior로 모델에 반영되기도 한다[4]. 지하철 노선에도 다른 곳에 비해 더 많은 사람들이 이용하는 구간이 있을 수 있다. 본 실험에서는 [6]에서 보인 이동 패턴을 Trajectory로서 활용하였다.

지하철 노선을 따라 이동 가능한 모든 구간을 N개의 Vertex와 M개의 Edge를 가진 그래프 G(**V**, **E**)로 표현하면 그림 1. (b)와 같다. 여기서 측정된 데이터는 1시간 동안 Gate를 통과한 승하차 승객 수이다. 본 실험에서는 Gate와 이어진 Edge 위를 통과하는 모든 승객 수(NV, Number of Visits)로 승차 승객과 하차 승객의 합을 사용하였다. NV 는 그래프의 Edge Weight으로 나타낼 수 있으며, Edge Weight은 Gaussian Process에서 각 노드에 주어지는 Latent Variable의 함수이다. 이들 간의 상관관계를 좀 더 간단히 하기 위해 [4]에서는 그래프 모델을 각 Edge를 노드로 한 Edge Graph G'(**E**, **E'**)로 변환하여 실제 승객 수인 $True\_NV$ 를 Latent Variable로 잡았다. Edge Graph의 노드 e_i 에서 관측된 승객 수 y_i 는 일반적인 Gaussian Noise를 고려하였을 때, 참값 $True\_NV_i$ 에 대해 다음과 같이 나타낼 수 있다.

$$y_i = True\_NV_i + \varepsilon_i \ , \ \varepsilon_i \sim N(\Theta, \sigma^2) \tag{1}$$

모든 관측 데이터는 Train Data로서 Unrecorded Area인 Test Node를 예측하기 위해 사용되며, Train Node와 Test Node 사이의 공간적 관계는 그래프의 Adjacent Matrix를 통해 계산된 Kernel Function으로 모델에 반영된다. $True\_NV_i$ 는 노드 e_i 의 위치 x 의 함수로 다음과 같이 나타낼 수 있다.

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big) \tag{2}$$

여기서 $m(\boldsymbol{x})$ 와 $k(\boldsymbol{x}, \boldsymbol{x}')$ 는 각각 Gaussian Process의 Mean Function과 Covariance Function을 의미한다. 정의에 의해, 임의의 관측 값 y 의 집합은 Multivariate Gaussian Distribution에서 뽑힌 하나의 샘플로 가정할 수 있으며, 이는 다시 예측 값과 함께 Joint Gaussian Distribution을 이루는 유한한 Random Variable의 모임이 된다[5]. 이에 따라 예측된 값은 식(3)과 같다.

$$\bar{\mathbf{f}}_* = \mathbf{m}(X_*) + K(X_*, X)K_y^{-1}(\mathbf{y} - \mathbf{m}(X)) \tag{3}$$

여기서 *는 Test Node를 표현하기 위해 사용되었으며, $K_y = K + \sigma_n^2 I$ 이다.

2.3.2. GPR for Temporal Prediction

본 논문에서는 Unrecorded Area의 AADT 추정을 위해 Gaussian Process를 통한 예측을 4년간의 데이터를 활용하여 다시 요일에 대한 분포로 나타내었다. 여기서 요일 x 에 대한 Gaussian Process Model의 Regression값은 실제 관측 값의 요일 별 평균과 비교하여 전체적인 모델의 성능 측정에 사용된다. 자세한 실험에 대한 결과는 다음 장과 같다.

3. Experimental Analysis

3.1. Spatial Analysis

식 (2)과 (3)에서 보듯이 GPR은 어떤 Kernel Function을 사용하느냐에 따라 그 결과가 달라질 수 있다. 본 연구에서는 대상이 되는 지하철 노선을 Edge Graph로 나타내었으므로 기본적인 RBF Kernel과 Graph Kernel의 몇 가지 변형이 비교 대상이 되었다. 표 1은 실험 모델로 설정한 14개의 역 중 Edge Graph에 반영된 12개 역에 대해, 공간적 정보로 추정한 값의 LOOCV (Leave-One-Out Cross Validation) 결과를 보여준다. [4]의 실험 결과에서도 알 수 있듯이 Traffic Network 내의 전체 경로 중 관측된 구간의 수는 예측 정확도에 큰 영향을 미친다. 그러나 본 실험에서는 데이터의 제약으로 모델링 된 28개의 구간 중 12개 구간인 약 43%만을 이용할 수 있었다. 예측 성능 평가로는 Mean Absolute Error를 사용하였으며, 모든 승객 수는 [0, 1] 사이의 값으로 동일한 조건으로 정규화 하였다. 표 1의 4)와 5)에는 [6]의 연구에서 관찰된 이동 패턴이 Adjacency Matrix에 가중치로서 반영되어있다.

표 1. Regularized passenger quantity estimation by spatial analysis

	1)RBF	2)Lap	3)Diff	4)Diff+ Tr	5)Diff+ Poly + Tr
MAE	0.0896	0.0900	0.0822	0.0787	0.0786

1) RBF kernel
2) Regularized Laplacian kernel (Graph kernel)
3) Diffusion kernel (Graph kernel)
4) Diffusion kernel with trajectory pattern (Graph kernel)
5) Diffusion + polynomial (d=5) kernel with trajectory pattern

3.2. Adding Temporal Analysis

기본적으로 요일 별 Regression은 앞서 수행된 공간적 분석에 따른 추정에 기반을 두므로 Spatial Analysis가 먼저 잘 되어야 할 필요가 있다. 그러나 이러한 요일 별 분석을 통해 특정 날짜가 아닌 일반적인 값에 대한 분포를 예측할 수 있으며, 이로서 Spatial Analysis의 오차를 어느 정도 줄일 수 있었다.

3.3. AADT Estimation

본 논문의 초반에서 목표한 대로 Gaussian Process를 이용하여 공간적, 시간적 Regression을 수행한 최종 결과는 표 2와 같다. 이는 표 1의 실험과 마찬가지로 12개 역에 대하여 LOOCV를 적용한 결과로서, Spatial Analysis만 적용했을 때와 비교하면 전체적으로 오차가 ΔMAE만큼 줄어들었음을 알 수 있다. 또한 본 실험에서는 Diffusion Kernel과 Trajectory를 이용한 모델에 전체적인 경향을 어느 정도 반영하는 Polynomial Kernel을 더하여 약간의 성능 향상을 확인할 수 있었다.

표 2. Passenger quantity estimation (AADT)

	1)NN	2)RBF &RBF	3)Lap &RBF	4)Diff &RBF	5)DPT &R+P
MAE	0.0995	0.0889	0.0897	0.0815	0.0783
Δ_{MAE}	n/a	0.0007	0.0003	0.0007	0.0003

1) Simple Perceptron
2) RBF for space and RBF for time
3) Regularized Laplacian for space and RBF for time
4) Diffusion for space and RBF for time
5) Diffusion + polynomial with trajectory for space and RBF + polynomial for time

4. Conclusions

본 논문에서는 지난 4년 동안의 서울 지하철 승하차 승객 수 데이터를 활용하여 관측되지 않은 구간의 승객 수를 추정하는 것을 목표로, GPR을 통한 공간적 분석과 시간적 분석을 병행한 모델을 구성하였다.

그러나 개별 승객의 승하차 시간 및 위치를 알기 어렵고, Raw Data 자체의 편차가 커 Regression만으로는 실제 값에 가깝게 예측하는 데 한계가 있었으며, 노드의 수에 비례하여 Adjacent Matrix를 만들어줘야 하는 어려움이 있었다. 이에 향후에는 유동 인구에 영향을 줄 수 있는 외적 요인들과 함께 학습의 범위를 넓히는 연구가 필요해 보인다.

참 고 문 헌

[1] Seoul Open Data Plaza, http://data.seoul.go.kr/
[2] Seoul Statistics, http://stat.seoul.go.kr/
[3] M. May, D. Hecker, C. Korner, A Vector-Geometry Based Spatial kNN-Algorithm for Traffic Frequency Predictions
[4] T. Liebig, Pedestrian Quantity Estimation with Trajectory Patterns
[5] Gaussian Processes for Machine Learning, Carl Edward Rasmussen and Christopher K. I. Williams, The MIT Press, 2006. ISBN 0-262-18253-X
[6] K. Kim, K. Oh, Y. Lee, J. Jung, Discovery of Travel Patterns in Seoul Metropolitan Subway Using Big Data of Smart Card Transaction Systems