날씨에 따른 서울특별시 대중교통 이용량 예측에 관한 연구

김희진*, 오수진**, 김응모***
*성균관대학교 자연과학대학
**성균관대학교 정보통신대학
***성균관대학교 소프트웨어대학
e-mail: thguffkd@skku.edu, bgbanana4@gmail.com, ukim@skku.edu

A Study on the Prediction of Public Transportation Consumption in Seoul by Weather

Hee-Jin Kim*, Sujin OH**, Ung-Mo Kim***
*College of Natural Science, Sungkyunkwan University
**College of Information and Communication Engineering, Sungkyunkwan University
***College of Software, Sungkyunkwan University

요 약

현대 사회에서는 다양한 이동수단 중 지하철, 버스 등의 대중교통에 대한 수요가 높은 편이다. 본 연구의 배경이 되는 서울특별시의 경우에는 출퇴근 시, 과반 수 이상이 대중교통을 이용한다. 대중교통 이용량에는 날씨, 평일-주말, 연착, 도로현황 등 여러 가지에 원인을 둔다. 본 연구에서는 여러 요인 중에서도 날씨 데이터(기온, 강수량, 미세먼지)에 초점을 두어, 날씨에 따른 대중교통 이용량의 변화 양상을 학습하여 예측하는 연구를 진행한다. 서울특별시 25개 자치구마다의 날씨 데이터와 대중교통 이용 데이터를 이용하여 Regression을 통한 데이터 학습을 진행하였으며, 학습된 모델을 통한 날씨에 따른 서울특별시 대중교통 이용량 예측에 따른 평균 오차율은 15.49%로 낮은 오차율을 가진다. 본 연구 결과는 날씨에 따른 버스와 지하철의 배차 간격 조절 등의 대중교통 배치 판단 결정에 기초자료로 사용될 것으로 기대된다.

1. 서론

1.1 연구 배경

현대 사회에서는 다양한 이동 수단 중 지하철, 버스 등의 대중교통에 대한 수요가 높은 편이다. 이를 뒷받침하듯이 서울특별시가 대중교통을 분석한 결과, 2016년도에 총 49억 4천만여 명 즉, 하루 평균 13,491천명의 많은 시민들이 서울 지하철과 버스를 이용했다[1]. 또한 '잡코리아'에서 직장인과 대학생 873명을 대상으로 '출퇴근, 통학 시 이용하는 이동수단은 무엇인가?'라는 질문에 81.5%가 대중교통을 이용한다고 답변했다[2]. 따라서 대중교통의 수요를 예측하고 이에 맞춰 적절하게 대중교통의 배차를 조절하는 것은 매우 중요한 이슈이다.

대중교통 이용량에는 날씨, 평일-주말, 연착, 도로현황 등 여러 가지에 원인을 둔다. 날씨를 제외한 다른 요인들은 많은 연구가 존재 및 진행되고 있으며 대중교통 이용량과의 관계가 쉽게 파악이 가능하다. 반면 날씨와의 관계를 분석한 연구의 수는 적으며, 연구 범위가 넓지 않다. 오직 강우 상태와의 관계만을 분석[3-4], 분석에 사용된 대중교통의 범위가 일정 지역으로 한정[5] 그리고 수집한 분석 자료 기간의 범위가 일주일에서 1년으로 짧았다[3-6]. 또한 대중교통과의 관계 분석으로 끝나며 예측하는 연구는 거의 존재하지 않는다. 따라서 본 연구에서는 여러 요인 중에서도 날씨 데이터(기온, 강수량, 미세먼지)에 초점을 두어, 날씨에 따른 대중교통 이용량의 변화 양상을 학습하여 예측하는 연구를 진행하고자 한다.

본 연구 결과는 날씨에 따른 버스와 지하철의 배차 간격 조절 등의 대중교통 배치 판단 결정에 기초자료로 사용될 것으로 기대된다.

1.2 연구 범위 및 수행 절차

본 연구는 2015년 1월부터 2017년 5월까지의 총 29개월간의 데이터를 사용한다. 그리고 서울특별시 지하철과 버스의 승·하차 인원 데이터와 날씨 데이터(기온, 강수량, 미세먼지)를 이용하여 진행한다. 모든 데이터는 월 단위로 수집되었으며, 서울특별시 25개 자치구를 기준으로 날씨 데이터와 대중교통 평균 승·하차 데이터를 분류하여 데이터를 수집하였다. 수집된 데이터의 수집 기간과 분류 기준이 구체적이며, 데이터 수집의 목적이 분명하기 때문에, 본 연구에서는 기존의 다른 연구에 비해 좀 더 세부적인 분석이 가능하다.

2절에서는 데이터 수집과 관련하여 각 데이터의 처리 과정을 서술하며, 3절에서는 지도 학습 과정과 그 결과, 마지막으로 4절에서는 결론을 서술한다.

2. 데이터 수집 및 처리

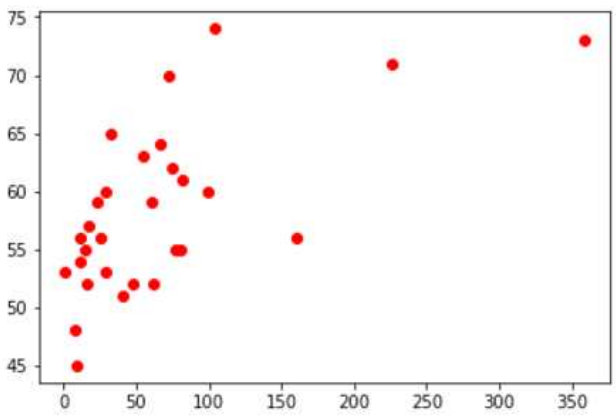
본 연구의 수행 절차는 크게 두 단계로 나눌 수 있다. 첫 번째 단계는, 수집한 데이터에서 필요한 정보만을 생성하는 것이다. 연구 범위인 서울특별시에 포함되지 않는 정보들은 제거하고, 서울특별시 25개 자치구로 데이터를 분류 및 처리를 한다. 두 번째 단계는, 생성한 데이터 학습 모델을 구축하여 학습된 모델을 이용하여 예측하는 것이다. 본 연구에서는 데이터 학습 모델을 구축할 때, Regression을 이용하였으며, 이에 대해서는 3.1에서 서술한다.

2.1 데이터 수집

본 연구에서는 '서울특별시 열린 데이터 광장[7]'에서 제공하는 데이터를 활용하였다. 날씨 데이터는 '서울특별시 월별 평균 기상개황'과 '서울특별시 월별 평균 대기오염도 정보' 두 가지를 이용하였다. 대중교통 데이터 또한 '서울특별시 월별 버스노선별'(이하 버스 데이터)과 '지하철 호선별 역별 승하차 인원 정보'(이하 지하철 데이터) 두 가지를 이용하였다. 모든 데이터는 2015년 1월부터 2017년 5월까지 총 29개월간의 월별 데이터이다.

2.2 날씨 데이터 처리

본 연구에서는 여러 가지 날씨 데이터 중 대중교통에 원인을 주는 요인으로 (평균)기온(℃), 강수량(mm), (평균)상대습도(%), 미세먼지($\mu g/m^3$) 4가지의 날씨 데이터를 선정했다. 선정된 날씨 데이터 중 강수량과 상대습도의 경우 상관관계를 계산해 본 결과 (그림 1)에서도 알 수 있듯이 약 +0.641로 +0.3~+0.7 사이에 존재하여 뚜렷한 양적 선형 관계를 가진다.



(그림 1) x축 : 강수량, y축 : 상대습도

따라서 학습 모델 구축의 간결함을 위해 강수량 하나의 요인만을 이용하였다. 상대습도가 아닌 강수량을 학습 모델 구축의 데이터로 선택한 이유는 3.2에서 언급한다. 따라서 본 논문에서 학습 모델 구축을 위해 사용한 날씨 데이터는 다음 <표 1>과 같다.

<표 1> 날씨 데이터

(평균)기온(℃)	강수량(mm)	미세먼지($\mu g/m^3$)

2.3 대중교통 데이터 처리

2.3.1 지하철 데이터

검색 포탈 네이버 지도를 이용하여 지리적으로 서울특별시에 포함되지 않는 지하철 정보들은 제거하고, 지하철역 주소를 기반으로 역을 25개의 자치구로 분류한다. 이때, 수집된 데이터 기간 동안 변경된 지하철역 이름을 통일하는 과정도 추가한다.

지하철 이용량은 데이터에 존재하는 시간별 승차인원, 하차인원을 이용하여 월 평균 승·하차인원을 계산한다. 그 다음 자치구별 대중교통 이용량에 사용할 k개의 역을 선정한다. k는 지하철과 버스 평균 승·하차인원을 합한 것을 기준으로 상위 k개의 역으로 결정한다. 이 때, k의 값은 최소 3으로[1] 선정한다. 2.3.2에서의 버스 평균 승·하차인원을 고려하기 위해 현 단계에서는 후보 역을 포함하여 자치구별 k^*개의 역을 선정한다.

$$k^* \geq k \tag{1}$$

k^*와 k의 관계는 식 (1)과 같으며, 본 연구에서는 자치구별 지하철역의 수를 고려하여, 아래의 식 (2)과 같이 정하였다.

$$k^* = k+2 \tag{2}$$

2.3.2 버스 데이터

수많은 버스 데이터 중 지하철역 근처에 위치하는 버스 정류장만을 이용한다. 판별하는 기준으로는 역 이름 포함 관계를 사용한다. 예를 들어 역 이름이 '합정'일 경우 '합정역'이 포함된 버스 정류장만을 이용한다. 만약 '합정역'을 포함한 정류장이 하나라도 존재하지 않을시, '합정'이 포함된 정류장으로 선정한다. 불필요한 정류장을 제거하기 위해 2.3.1에서 선정한 자치구별 k^*개의 지하철역을 이용한다. 또한 정류장 정보 중 필요한 정보인 '날짜, 근처 지하철 역 이름' 그리고 시간별 승차인원, 하차인원을 이용한 '월 평균 승·하차인원'만을 이용한다.

2.3.1에서 선정한 자치구별 k^*개의 지하철역과 2.3.2에서 추출한 버스 데이터를 합하여 최종적으로 상위 k개의 역을 선정한다. 마지막으로 날씨 데이터와 구성방식을 통일시키기 위해 자치구별 평균 승·하차인원을 계산 후 모든 데이터를 하나로 통합하여 학습 모델 데이터를 구축한다. 본 연구에서는 자치구별 대중교통 데이터 k의 값으로 4를 사용하였으며, 이에 대해서는 4.2.2에서 설명한다.

1) 자치구에 존재하는 지하철역의 최소 개수가 3개이다.

3. 지도 학습

본 연구에서는 python 언어를 이용한 'scikit-learn' 라이브러리에서 제공하는 Regression 함수를 사용하여 모델을 학습시키고, 이를 이용하여 대중교통 이용량을 예측하였다.

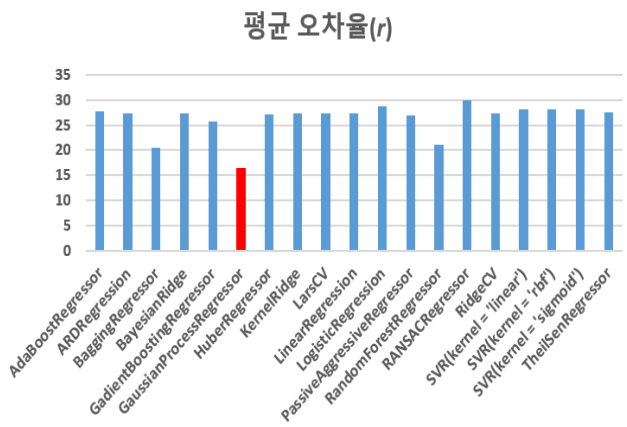
3.1 Regression 함수

모델 학습에 주로 이용하는 기법으로 Classification과 Regression이 있다. Classification의 경우, 학습 데이터가 Category로 분류되어 있어, 이를 기반으로 검증 데이터의 Category를 식별하는 방법이다. 주로 예측하는 결과 값이 이산 값일 때 사용한다. Regression의 경우, 학습 데이터가 연속 형 변수들이며, 이를 기반으로 두 변수 사이의 모형을 구한 후에 검증 데이터의 예상 값을 측정하는 방법이다. 주로 예측하는 결과 값이 연속 값일 때 사용한다. 본 연구는 대중교통 이용량이라는 연속 값을 예측하기에, Classification 함수보다는 Regression 함수가 더 적합하다.

'scikit-learn'에서 제공하는 19가지의 Regression 함수와 2절에서 처리하여 수집된 데이터를 사용하여 학습 모델을 구축하고 각 모델의 평균 오차율(r)을 계산하였다. 평균 오차율(r)은 정확도를 판단하는 기준으로 아래의 식 (3)을 이용하여 계산한다.

$$r = \frac{|\text{예측값} - \text{실제값}|}{\text{제값}} \times 100 \qquad (3)$$

본 연구에서는 19가지의 Regression 함수 중 가장 낮은 평균 오차율(r)을 가지는 GaussianProcessRegressor 함수를 선택하여 학습 모델을 구축하였다(그림 2).



(그림 2) Regression 함수들의 평균 오차율(r)

3.2 데이터 구성

본 연구에서 대중교통 이용량에 영향을 끼치는 요인으로 (평균)기온(℃), 강수량(mm), (평균)상대습도(%), 미세먼지(㎍/㎥) 4가지의 날씨 데이터를 선정하였다. 강수량(mm)과 (평균)상대습도(%)의 경우, 2.2에서와 같이 높은 상관관계를 가진다. <표 2>는 강수량과 상대습도 각각의 요인과 두 요인을 모두 사용했을 때의 평균 오차율(r)을 구한 표이다. <표 2>에 따르면 평균 오차율(r)이 크게 다르지 않기 때문에, 따라서 학습 모델 구축의 간결함을 위해 평균 오차율(r)이 비슷한 강수량 요인만을 이용한다.

<표 2> 학습 데이터에 따른 평균 오차율(r) (단위 : %)

Method Name	강수량	강수량+상대습도	상대습도
GaussianProcess Regressor	15.4944	15.4944	15.4943
RandomForest Regressor	19.6151	19.8069	19.2242
Bagging Regressor	18.8355	18.7584	19.3995

3.3 지도 학습 결과

<표 3>은 GaussianProcessRegressor을 사용하여 '기온, 강수량, 미세먼지, 대중교통 이용량'을 학습 데이터로 가지는 학습 모델 구축한 뒤, 검증 데이터를 이용하여 모델을 검증한 결과이다. <표 3>에 사용된 용어를 간략하게 설명하면, '1_test'는 검증 데이터의 값으로 실제 대중교통 이용량이다(이하 실제 값). '2_pred'는 지도 학습 결과로 예측한 값이다(이하 예측 값). '3_sub'는 |실제 값-예측 값|이다(이하 오차). '4_rate'는 오차/실제 값×100으로 오차율이다.

<표 3> 지도 학습 결과

	1_test	2_pred	3_sub	4_rate
1	1297894	1741135	443240	25.45697
2	1742700	1708277	34424	1.975326
3	1017756	1349691	331935	24.59341
4	922260	948990.1	26730	2.816679
5	1671879	1663782	8098	0.484365
. . .				
214	931660	931660	1	0.000107
215	999405	867250.3	132155	13.22337
216	1698054	1532882	165173	9.727194
217	1164737	1601316	436578	27.26372
218	781559	794623.6	13064	1.64405

본 연구에서 검증 데이터로 사용된 데이터는 총 218개로 15.49%의 낮은 평균 오차율(r)을 가진다.

4. 모델 평가

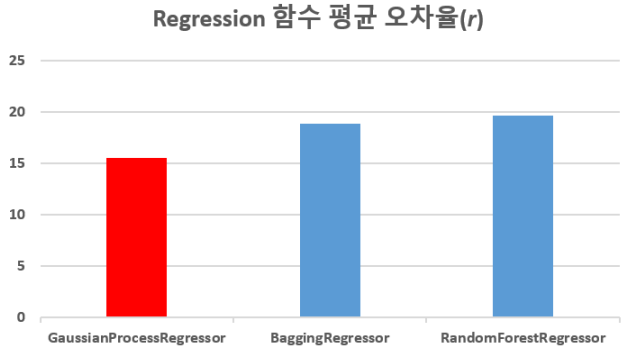
4.1 모델 평가 방법

학습 데이터를 이용하여 학습 모델을 구축하기에 앞서 학습 과정에서 데이터의 과적합(Overfitting)을 막기 위해 전체 학습 데이터를 훈련 데이터와 검증 데이터로 나눈다. train_test_split 함수를 이용하여, 'test_size = 0.3'으로 설정하여 사용한다. 즉 전체 학습 데이터 중 30%를 검증 데이터로 사용한다.

4.2 모델 평가

4.2.1 Regression 함수에 따른 학습 모델 평가

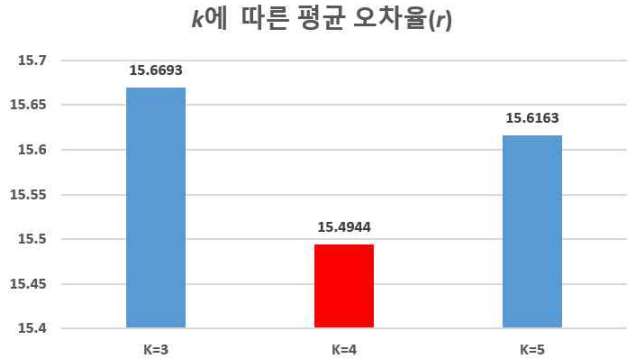
(그림 3)은 3.1의 결과를 바탕으로, 평균 오차율(r)이 25% 이하인 Regression 함수의 평균 오차율(r)을 비교한 결과이다. 본 연구에서 사용한 GaussianProcessRegressor 함수가 가장 낮은 평균 오차율(r)을 가짐을 알 수 있다.



(그림 3) Regression 함수에 따른 평균 오차율(r)

4.2.2 k 선정에 따른 학습 모델 평가

(그림 4)는 k 값에 따라 학습 모델의 평균 오차율(r) 변화를 비교한 결과이다. 본 연구에서는 가장 낮은 평균 오차율(r)을 가진 k=4일 때를 사용한다. k 값이 3에서 4로 증가함에 따라 평균 오차율(r)은 낮아졌고, 반대로 k 값이 4에서 5로 증가할 때는 데이터의 과적합(Overfitting)으로 평균 오차율(r)이 오히려 증가했다.



(그림 4) k에 따른 평균 오차율(r)의 변화

5. 결론

본 연구에서는 '서울특별시 열린 데이터 광장' 데이터를 활용하여 날씨에 따른 서울특별시 대중교통 이용량을 예측했다. 예측에 사용한 데이터는 2015년 1월부터 2017년 5월까지 총 29개월이며 서울특별시 25개의 자치구로 분류하여 진행했다. 70%의 학습 데이터와 30%의 검증 데이터를 GaussianProcessRegressor 학습 모델을 이용하여 구축한 결과 평균 15.49%의 낮은 평균 오차율(r)로 대중교통 이용량을 예측한다. 본 연구 결과는 날씨에 따른 버스와 배차 간격 조절 등의 대중교통 배치 판단 결정에 기초자료로 사용될 것으로 기대된다.

향후 더 효율적인 배치 판단 결정 자료로 이용하기 위해서는 본 연구에 사용된 데이터보다 폭넓은 데이터를 사용하여 연구해야한다. 버스와 지하철로 이용량을 분류하고 또한 시간대 별 이용량도 추가하여 학습 모델을 구축하면 배치가 필요한 대중교통 종류와 시간을 정확히 구할 수 있다. 그리고 월별이 아닌 일별 대중교통 이용량과 일별 날씨 데이터를 이용하여 학습 모델을 구축하면 보다 정확한 하루 이용량을 예측할 수 있다.

참고문헌

[1] 서울특별시청. (2017). "교통카드 데이터로 본 '16년 서울 대중교통 이용현황" http://spp.seoul.go.kr/main/news/news_report.jsp#list/1 (2017-8-31 방문)

[2] 박상우. (2016). "직장인, 대학생 출퇴근·통학 이동수단 1위 '대중교통'" 잡코리아 홈페이지, http://www.jobkorea.co.kr/GoodJob/News (2017-8-31 방문)

[3] 이광섭, 엄진기, 유소영, 민재홍, 양근율 (2014). "강우와 서울시 대중교통 승차인원과의 관계 분석" 한국철도학회 학술발표대회논문집, 252-257

[4] 박근영, 이시복 (2012). "강우 상태에 따른 대중교통 이용패턴 특성연구" 대한토목학회논문집 D, 32(1D), 23-31

[5] 박명근, 고강섭, 곽재덕, 이한열, 박성수 (2010). "과거 패턴자료와 날씨계수를 이용한 버스 통행시간 예측에 관한 연구" 대한교통학회 학술대회지, 63, 253-257

[6] 최상기, 이종호, 오승훈 (2013). "기상조건이 대중교통 수요에 미치는 영향에 관한 연구" 대한토목학회논문집, 33(6), 2447-2453

[7] 서울특별시 열린 데이터 광장. (2017). http://data.seoul.go.kr/ (2017-7-10 방문)