기계학습을 활용한 Bay Area지역의 공유자전거 수요 예측

홍성민○ 김현철
고려대학교
sminhong@korea.ac.kr, harrykim@korea.ac.kr

Demand prediction of bicycle-sharing system in the Bay Area using machine learning

Sung Min Hong○ Hyeoncheol Kim
Korea University

요 약

전 세계적으로 공유자전거 시스템의 도입이 확대됨에 따라 이의 수요 예측의 중요성이 높아지고 있다. 공유자전거 시스템이 전산으로 관리되고 있는 만큼, 일자별 운행기록과 정거장 현황, 그 지역의 날씨 등의 속성들을 활용한다면 일일수요량을 예측할 수 있을 것으로 기대됐다. 본 연구는 기계학습을 활용하여 샌프란시스코 만 지역의 일일 공유자전거 이용자를 예측하고자 한다. 일자별 운행기록과 정거장 설치현황, 날씨 정보 등의 데이터를 수집하여 기계학습을 이용해 수요량을 예측하는 방법이다. 의사결정트리방식, 앙상블기법, 기하기반기법, 인공신경망 등의 알고리즘을 활용해 비교를 진행하였다. 그 중 앙상블기법을 통해 94.54%의 예측 정확도를 도출하였고 추후 딥러닝 기법 실험을 통해 성능향상을 도모할 예정이다.

1. 서 론

서울을 포함한 전 세계의 주요 도시들은 공유자전거 시스템을 구축 및 확장하고 있다. 공유자전거 시스템은 공공 소유의 자전거를 일정 기간 동안 대여해주는 것으로 이는 1965년 암스테르담에서 처음 선보였다[1]. 2000년대 이후 현대적인 공유자전거 시스템이 구축되었으며 프랑스 파리가 2007년 대도시 중 최초로 이를 도입, 성공함으로써 이 시스템이 전 세계로 퍼져나가는 계기가 되었다. 공유자전거는 이용자의 건강증진뿐만 아니라, 친환경 대체교통수단으로써 화석연료 사용량을 줄이고, 기존 교통수단의 보완재로써 이동성을 증가시키는 효과가 있다. 이로 인해 2016년 12월 기준으로 1000개 이상의 도시에서 공유자전거 시스템이 도입되어 시행되고 있다[2].

이처럼 단기간에 전 세계로 빠르게 퍼지고 있는 이 시스템이 장기적으로 성공을 거두기 위해서는 이용자의 이용 실태의 분석과 이에 따른 도시 계획이 뒷받침되어야 한다. 본 연구에서는 2013년 처음 도입된 샌프란시스코 만 (the Bay Area) 지역의 공유자전거 시스템의 일일 수요량을 기계학습의 방법을 이용하여 예측한다. 이 연구는 그 자체로 이 지역의 일일 수요량 예측을 가능케 하여, 이를 통해 특정 일자의 자전거 도로 확대 실세 등의 정책으로 이어질 수 있을 것이다. 또한 만들어진 모델을 바탕으로 다른 지역의 공유자전거 데이터에 적용하거나 유사한 특징을 가지고 있는 비즈니스 모델(예: 우버)에도 적용하여 시민들의 전반적인 이동 수단 활용 만족도에 기여할 수 있을 것으로 기대된다.

2. 실험의 진행

2.1 데이터 선정과 처리

본 연구의 데이터는 샌프란시스코 만 지역의 공유 자전거 이용기록과 자전거정거장 현황, 그리고 해당 기간의 그 지역 날씨 정보이다. 샌프란시스코 만 지역의 데이터로 선정한 이유는 공개되어 있지 않거나 데이터의 손실이 많은 다른 도시의 데이터와는 달리 약 2년간의 데이터가 큰 손실 없이 기록되어 있었기 때문이다.

자전거 운행에 관한 데이터는 대여 시점과 반납 시점, 대여 장소 등에 대한 정보를 내포하고 있었는데, 일일 사용량 분석을 위해서 일 단위로 데이터를 분류하여 일일 대여량을 클래스로 설정했다. 그리고 본 데이터에 샌프란시스코 지역 날씨 정보 (최고 및 최저 기온, 습도, 풍량, 강우정보 등 포함)와 날짜정보에 주말 및 공휴일 정보를 덧붙여서 연구데이터를 구성하였다. 그래서 처리된 데이터는 더미변수(dummy variable)로 변환한 항목을 포함해 총 31 개의 속성과 733일에 해당하는 인스턴스들로 구성되었다.

다만, 공개데이터의 특성상 이미 일일 이용량 관련 실험이 Kaggle 웹사이트에 공개되어 있다[3]. 그래서 이미 진행된 연구를 Baseline으로 삼아 비교군을 확대하고 모델을 개선하는 것을 이 연구의 목적으로 한다. 그래서 비교를 위해 교차검증 (Cross Validation)의 겹을 동일하게 15개로 가져가고 평가방식으로 같은 점수체계를 사용하였다. 비록 전처리 과정에서 데이터에 차이가 발생하지만, 전 연구자가 사용한 연구방식을 이번 연구데이터에 적용하여 알고리즘 성능 비교 및 개선에 집중하고자 한다.

2.2 알고리즘의 선정

실험에 활용할 알고리즘의 선정은 이미 여러 연구를 통해 성능과 신뢰성이 확보되는 것을 우선적으로 했다. 그리고 이 기준으로 선정된 알고리즘 중 이미 기존 연구에서 사용된 알고리즘이 있다면 속성값을 수정해 나은 성능을 도출할 수 있도록 개선하고자 했다.

실험에 사용된 알고리즘을 나열하면, 우선 가장 기본적인 의사결정트리방식이 있다. 이 중 수치형 데이터를 회귀적으로 분석하는 Decision Tree Regression이 실험에 사용됐다.

의사결정트리를 포함해 여러 개별적 모델들을 엮어서 좀 더 강력한 성능을 내는 것이 앙상블 학습기법인데, 이 중 대표적인 Random Forest Regression[4]과 Gradient Boosting

Regression[5] 모두 실험에 쓰였다. 우선 전자의 경우, CART (Classification And Regression Tree)에 배깅 (bagging)[6]을 결합해 과대적합(overfitting)의 문제를 해결한 형태이고, 후자는 깊이가 깊지 않은 여러 개의 결정 트리를 묶어 성능을 끌어올린 형태이다.

기하기반 모델의 경우, 모수 없이 인스턴스 간의 거리를 측정하는 게으른 학습 (Lazy Learning)에 일종으로 볼 수 있다. 이 중 k-최근접 이웃 알고리즘 (kNN)[7]의 회귀모형이 실험에 사용되었다.

마지막으로 인공신경망의 경우엔 인간의 신경망을 모방한 형태로 뉴런에 해당하는 노드들을 여러 겹으로 쌓아 오차함수의 기울기를 줄여나가는 방식으로 최적의 모델을 도출한다. 이에 해당하는 다층 퍼셉트론(MLP)[8]을 실험에 사용하였다.

3. 실험결과

3.1 평가기준

각 모델별 평가의 기준은 Median Absolute Error(MAE)의 기법[9]을 적용하였다. n이 검증하는 일자의 개수, m이 교차검증의 횟수, $\widehat{X}$이 예측값일 때, 식은 다음과 같다.

$$MAE = median(median_j(\left|X_i - \widehat{X_i}\right|))$$
$$\text{f or } X_1, X_2, \cdots, X_n, j = 1, 2, 3, \ldots, m$$

즉, 교차검증의 각 겹에서 일자별 예측치 오차의 절대값의 중앙값들의 중앙값을 기준으로 성능을 평가한다. 본 실험에서는 계층 추출된 15개의 데이터집단을 이용하여 교차 검증을 실시하므로 m은 15이다. MAE의 값이 크다는 것은 오차값이 크다는 것으로 값이 작을수록 좋은 성능을 발휘한다고 볼 수 있다.

3.2 Decision Tree Regression의 결과

실험의 Baseline은 Kaggle에 공개되어 있는 실험으로 했다[3]. Decision Tree Regression의 경우 Baseline은 최소 단말노드(Leaf) 3개, 최대 깊이 8의 설정으로 59.8의 결과값을 가졌다. 최소 단말노드 개수와 최대 깊이의 설정변화로 과대적합의 문제를 줄일 수 있는데 최소 단말노드와 최대 깊이를 변수로 여러 차례 실험 후 주요 결과는 다음과 같다.

표 1. Decision Tree Regression의 주요 결과

	최소 단말노드	최대 깊이	MAE
Baseline	3	8	59.8
1	1	9	59.93
2	1	10	58.63
3	2	8	58
4	2	9	59.46
5	6	7	59.68
6	6	8	**57.07**

표에 기입하지 않은 실험 결과는 대개 60 초반에 형성되었고, 비교적 좋은 성능을 낸 모델의 경우에는 대개 최대 깊이가 6개에서 8개 사이에서 형성되었다. (다른 최소 단말노드의 경우에도 위 범위 값 내에서 비교적 좋은 성능을 보여줬다.)

3.3 Random Forest Regression의 결과

Decision Tree Regression의 약점을 보완한 형태인 Random forest regression은 전 모형보다 더 나은 성능을 기대하게 했다. Baseline에서 54.89으로 실제로 좀 더 나은 성능을 보여줬다. 그리고 속성 값의 변화에 따른 성능 값을 측정하였더니 표 2에 나오는 것처럼 대체로 좋은 성능을 보여줬다.

표 2. Random Forest Regression의 주요 결과

	Estimator 개수	최소 단말노드	MAE
Baseline	55	3	54.89
1	10	1	56
2	20	1	**49.5**
3	30	1	53.93
4	40	1	52.45
5	50	1	52.6
6	20	2	53.28
7	20	3	56.99

Random Forest Regression에서는 Decision Tree Regression과는 다르게 과대적합을 방지하기 위한 최소 단말노드의 수 증가가 오히려 성능저하를 야기하는 것도 목격할 수 있었는데, 이는 앙상블 기법의 적용으로 굳이 최소 단말노드의 숫자를 제한하지 않아도 성능 향상이 가능한 상황에서 오히려 역효과를 나타낸 것으로 추정된다.

3.4 Gradient Boosting Regression의 결과

또 다른 앙상블 기법인 Gradient Boosting Regression에서는 대체로 학습률(Learning rate)=0.1에서 좋은 성능을 보여주었고 Estimator의 값을 Baseline보다 낮췄을 때 좀 더 나은 성적을 보여주었다. 학습률의 최적치는 이전 여러 연구를 통해 다양한 방법이 소개되긴 하였지만 일반화하기 어렵고 실험을 통해 경험적인 최적치를 찾는 것이 일반적이다[10]. 그래서 실험을 통해 가장 좋은 성능의 모델을 탐색하였는데, 0.1의 학습률, 그리고 80개의 Estimator를 구성하였을 때 51.68의 오차값을 보여줬다. 자세한 실험의 결과는 다음 표 3과 같다.

표 3. Gradient Boosting Regression의 성능 비교 (단위: MAE)

	0.13	0.12	0.11	0.1	0.09
150	52.79	*59.23**	52.70	53.17	56.44
100	52.86	58.99	52.95	52.84	55.37
90	52.29	59.04	59.04	52.41	56.53
80	52.01	58.91	52.01	**51.68**	54.80
70	52.02	58.89	52.07	51.92	54.71
60	52.58	58.84	51.98	51.76	55.39

행: 학습률, 열: Estimator의 수, *: *Baseline*

3.5 기하기반 알고리즘과 인공신공망의 실험 결과

마지막으로 기존 연구에서는 진행하지 않은 기하기반 알고리즘과 인공신경망을 이용하여 추가적인 실험을 진행하였다. 기하기반은 K-Neighbours Regressor를 사용하였을 때

57.2의 MAE가 나왔다. 인공신경망의 경우 아래 표 4에 나온 바처럼 16개의 은닉노드를 3겹으로 배치했을 때 80.59로 가장 나은 성능을 보였다. 16개의 은닉노드는 31개의 입력값과 1개의 출력값의 평균을 활용한 값이다. 다른 모델의 경우 파라미터 수가 너무 많아 과대적합 된 것으로 추정된다.

표 4. 다층 퍼셉트론(MLP)의 주요 실험결과

	Hidden Layer	MAE
1	25-25-25-25-25	245.75
2	32-32-32-32	257.41
3	64-64-64-64	225.19
4	16-16	87.76
5	16-16-16	80.59
6	16-16-16-16	139.92

3.6 최종모델의 선택과 분석

앞서 실시한 총 6가지 계열의 알고리즘 중 Random Forest Regression의 20개의 Estimator와 1개의 최소 단말노드를 갖는 모델과 Gradient Boosting Regression의 0.1의 학습률과 80개의 Estimator를 갖는 모델이 가장 좋은 성능을 보여주었다. 각각 모델에서 예측치에 가장 많이 미치는 영향을 분석한 결과, Random Forest Regression 모델의 경우에는 근무일(Business day) 여부가 가장 중요한 요인이었다. 반면, Gradient Boosting Regression 역시 근무일 여부가 가장 중요하였으나 최고기온과 바람세기 등도 상당한 중요도를 보여 영향을 미쳤음을 확인할 수 있었다.

표 5. 주요 항목 번호와 이름

0	Max temperature
9	Max sea level pressure
11	Min sea level preassure
17	Max gust speed
18	Precipitation
20	Wind degree
26	Number of total docks
27	Business day
30	Month
31	Weekday

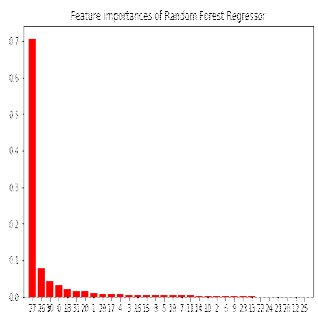

그림 2. Random Forest의 항목중요도

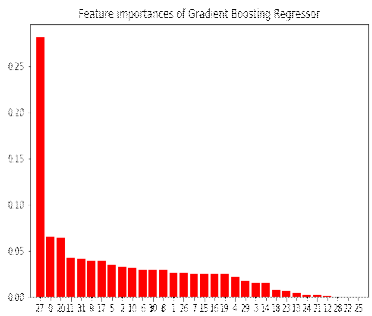

그림 3. Gradient Boosting의 항목중요도

4. 결론

샌프란시스코 만 지역의 공유자전거의 데이터를 활용하여 여러 기계학습 알고리즘의 성능을 비교해 볼 수 있었다. 실험을 통해 평균 906개의 이용량을 가지는 이 시스템을 최소 49.5개의 오차(MAE 기준)로 예측해 94.54%의 정확도를 도출할 수 있었다. 따라서 이 연구를 활용한다면 추후 이 지역의 수요량을 상당한 정확도로 예측할 수 있을 것으로 보인다. 또한, 이번 연구를 바탕으로 향후 다른 지역의 데이터를 분석하거나, 형태가 비슷한 시계열 분석이 가능할 것으로 기대된다.

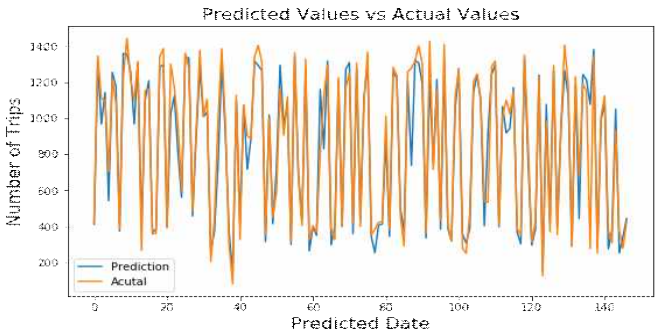

그림 1. 예측값과 실제값의 비교

하지만 아직 정확도 개선의 여지가 있는 만큼, 향후 추가 연구를 통해 Deep Neural Network나 Recurrent Neural Network 등 최근 시계열 분석에 많이 사용되고 있는 딥러닝 기법을 활용해 볼 계획이다. 이를 통해 정확도를 개선할 뿐만 아니라, 모델 자체에 대한 연구도 심층 깊게 진행하여 데이터와 모델의 적합성에 대한 일반적인 규칙에 대해서도 고찰해 볼 것이다.

5. 참고문헌

[1] S. Shaheen et al., Public Bikesharing in North America During a period of rapid expansion: understanding buseinss models, industry trends and user impacts, JTG, 2013

[2] P. Demaio, Bike Sharing: History, Impacts, Models of Provision, and Future, JPT, 2009

[3] Currie32, Bike Sharing in SF and Seattle, Github, 2017

[4] A. Liaw et al., Classification and Regression by randomForest, R news, 2002

[5] L.Mason, et.al., Boosting Algorithms as Gradient Descent in Function Space, RSISE, 1999

[6] L. Breiman, Bagging predictors, Machine Learning, 1996

[7] N. S. Altman, An introduction to Kernel and Nearest-Neighbor Nonparametric Regression, TAS, 1992

[8] D. Rumelhart, Learning internal representations by error propagation, UCSD, 1985

[9] P. Rousseeuw, et al., Alternatives to the median absolute deviation, JASA, 1993

[10] Chinrungrueng, et al., Optimal adaptive k-means algorithm with dynamic adjustment of learning rate, TNN, 1995