KIPS Trans. Softw. and Data Eng.
Vol.10, No.5 pp.169~178
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.5.169

A K-Means-Based Clustering Algorithm for Traffic Prediction in a Bike-Sharing System

Kyoungok Kim† · Chang Hwan Lee††

ABSTRACT

Recently, a bike-sharing system (BSS) has become popular as a convenient "last mile" transportation. Rebalancing of bikes is a critical issue to manage BSS because the rents and returns of bikes are not balanced by stations and periods. For efficient and effective rebalancing, accurate traffic prediction is important. Recently, cluster-based traffic prediction has been utilized to enhance the accuracy of prediction at the station-level and the clustering step is very important in this approach. In this paper, we propose a k-means based clustering algorithm that overcomes the drawbacks of the existing clustering methods for BSS; indeterministic and hardly converged. By employing the centroid initialization and using the temporal proportion of the rents and returns of stations as an input for clustering, the proposed algorithm can be deterministic and fast.

Keywords : Bike Sharing System, Clustering, Demand Prediction, Random Forest

공유자전거 시스템의 이용 예측을 위한 K-Means 기반의 군집 알고리즘

김 경 옥† · 이 창 환††

요 약

최근 들어 공유자전거 시스템은 대중교통 이용이 어렵거나 불가능한 마지막 목적지까지의 거리인 "라스트 마일"을 해소하는 방안으로 주목받고 있다. 공유자전거 시스템에서는 자전거의 대여와 반납의 불균형으로 인해서 사용자가 원하는 시간에 원하는 대여소에서 자전거를 빌리거나 반납할 수 있는 문제가 자주 발생한다. 이에 자전거 재배치는 공유자전거 시스템을 효율적으로 운영하는데 매우 중요한 이슈이다. 자전거 재배치를 효율적이고 효과적으로 진행하기 위해서는 무엇보다 정확한 수요 예측이 이뤄져야 한다. 최근에는 대여소의 수요를 보다 정확하게 예측하기 위해 군집 기반의 수요 예측 모델을 활용하는 방법이 개발되고 있는데, 여기서는 군집 분석 단계가 매우 중요하다. 이 연구에서는 비결정적이고 수렴이 어려운 기존의 공유자전거 수요 예측을 위한 군집 방법의 단점을 극복하는 k-means 기반의 군집 알고리즘을 제안한다. 이 방법은 초기 중심점 방법을 활용하기 때문에 매번 동일한 결과를 얻을 수 있으며, 대여소의 시간별 반납/대여 비중을 이용하여 기존 방법과는 달리 이전 단계의 군집 결과를 필요로 하지 않아 반복해서 군집 분석을 수행할 필요가 없어 빠른 군집 분석이 가능한 장점이 있다.

키워드 : 공유자전거시스템, 군집분석, 수요 예측, 랜덤 포레스트

1. 서 론

공유자전거 시스템은 무료 또는 유료로 단기간 개인에게 자전거를 대여해 주는 서비스를 일컫는다. IT 기술의 발전으로 인해 무인 시스템으로 누구나 손쉽게 자전거의 대여와 반납이 가능해지면서 대도시에서 지속해서 문제시되는 교통 혼잡과 이로 인한 대기 오염의 한 해결책으로 공유자전거 시스템을 도입한 도시가 급속히 늘고 있다[1].

세계적으로 2014년에 비해 2018년 4월에는 공유자전거 시스템 수는 두 배 가까이 늘었으며, 이용 가능한 자전거 대수는 20배 증가했다[2]. 또한 2018년 한 해 동안 미국의 공유자전거 이용 횟수는 5,200만 가량으로 2017년에 비해 45% 이상 증가했다[3]. 국내에도 창원의 '누비자'를 시작으로, 대전 '타슈', 고양 '피프틴', 서울 '따릉이' 등 많은 공유자전거 시스템이 운영 중이다.

대부분의 공유자전거 시스템은 무인으로 대여/반납이 가능한 거치대(dock)가 갖춰진 대여소(station)를 기반으로 운

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1F1A1054496).
† 정 회 원 : 서울과학기술대학교 산업공학과 조교수
†† 비 회 원 : 서울과학기술대학교 데이터사이언스학과 석사
Manuscript Received : October 30, 2020
First Revision : January 26, 2021
Accepted : February 23, 2021
* Corresponding Author : Kyoungok Kim(kyoungok.kim@seoultech.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

영되고 있으며. 사용자는 홈페이지나 모바일 앱을 통해서 특정 대여소의 특정 거치대에 비치된 자전거를 빌려서 같은 시스템 내의 모든 대여소에서 다시 반납할 수 있다.

공유자전거 시스템 운영에 있어 가장 큰 문제는 시간별로 대여소마다 자전거 대여와 반납의 불균형으로 인해 각 정류소에 자전거가 없어 빌리지 못하거나, 거치대가 차 있어 자전거를 반납할 수 없는 현상이다[4]. 이 문제를 해결하기 위해서는 대여소별 비치 자전거의 균형을 맞출 필요가 있는데 많은 공유자전거 시스템은 트럭 등을 이용해서 자전거가 많은 대여소에서 부족한 대여소로 자전거의 재배치를 수행하거나[5-6], 이용자에게 인센티브를 주고 자전거가 부족한 지역으로 자전거를 이동시키는 방법을 활용하고 있다[7-8].

제한된 자원으로 전체 시스템을 효율적으로 운영하고 효과적으로 자전거를 재배치하기 위해서는 대여소의 자전거 수요량을 정확하게 예측해야 한다. 이에 공유자전거 시스템을 위한 자전거 수요 예측의 정확도를 높이기 위한 다양한 연구들이 있었다. 수요 예측 모형은 주로 시계열 모형[9-10]이나 기계 학습의 회귀 알고리즘[11-13]을 이용해서 개발되었다. 예측이 이뤄지는 시간 단위는 10분, 1시간처럼 단시간 내의 수요를 예측하기도 하지만, 24시간과 같이 보다 긴 시간 단위의 수요를 예측하는 연구도 있다.

한편 최근에는 대여소별 수요 예측의 정확도를 높이기 위해 군집 단위의 수요 예측 모델을 함께 활용하는 계층적 방식의 수요 예측 모델에 대한 연구가 시도되고 있다. 공유자전거 시스템은 자전거 불균형 문제로 인해 수요의 변동성이 크므로, 군집 단위로 수요를 예측했을 때 대여소 단위로 수요를 예측했을 때보다 정확한 수요 예측 모형을 구축할 수 있다[14-15]. 이를 이용해 계층적 방식의 수요 예측에서는 군집 수준의 예측 모델을 활용해서 대여소 단위의 수요 예측을 진행한다. 군집 단위의 수요를 군집에 속하는 대여소의 과거 수요에 비례해서 분배하는 방식으로 대여소의 수요를 예측하거나[16], 군집 수준과 대여소 수준의 모델을 통합하여 동일 군집 내 대여소들의 수요의 합과 군집 수준 모델을 통해 예측한 군집 수요의 오차가 최소가 되도록 대여소 수준 모델의 예측값을 보정하는 방식[14] 등이 있다.

계층적 방식 모델에서 대여소 단위의 수요 예측의 정확도를 높이기 위해서는 정확한 예측이 가능하고 변동성이 적은 군집을 형성하는 것이 매우 중요하다. 하지만 기존 연구에서는 이용 패턴이 유사한 대여소끼리 군집을 형성하기 위해 특정 대여소에서 대여된 자전거가 반납된 군집의 비중을 활용하는데 이는 군집 결과에 따라 달라지므로 군집 결과가 수렴할 때까지 군집 분석을 여러 단계 반복해야 해서 군집 분석에 오랜 시간이 소요된다. 또한, k-means나 GMM 등 초기값에 따라 군집 결과가 달라질 가능성이 있는 방법을 사용하고 있다[15-16].

이에 본 연구에서는 기존의 군집 방법의 한계를 극복한 대여소 군집을 형성하는 k-means 기반의 알고리즘을 제안한다. 제안하는 군집 방법은 1) 초기 중심점을 데이터 분포에 따라 결정하여 항상 동일한 군집 결과를 얻을 수 있으며, 2) 이전 단계의 군집 결과에 의존하지 않는 대여소의 시간별 대여/반납 비중을 대여소의 이용 패턴을 대표하는 값으로 사용해 군집 결과를 갱신할 필요가 없고 k-means 군집에서 대여소의 위치와 이용 패턴을 동시에 활용하여 군집 시간을 크게 줄일 수 있다.

2. 선행 연구

공유자전거 시스템의 원활한 운영을 위해서는 자전거 대여와 반납의 불균형으로 인해 이용자가 대여소의 거치대가 꽉 차 있어 원하는 대여소에 자전거를 반납하지 못하거나, 대여소에 자전거가 한 대도 비치되어 있지 않아 자전거를 빌리지 못하는 불편함을 해소하는 것이 필요하다. 이 문제를 해결하기 위해서 대다수 시스템은 트럭 등을 이용해 직접 자전거를 재배치하거나 이용자에게 인센티브를 주고 자전거를 옮기게 하는 방법 등을 활용하고 있다. 재배치에 필요한 비용과 시간을 줄이면서 효과적으로 재배치를 수행하기 위해서는 대여소별로 공유자전거의 대여와 반납 수요, 대여소의 비치 자전거 수 등을 정확하게 예측할 필요가 있다. 이에 지금까지 공유자전거 수요 예측의 정확도를 높이기 위한 다양한 연구가 있었다.

공유자전거 수요 예측에는 다른 분야에서의 수요 예측과 마찬가지로 시계열 방법론이나 기계 학습 방법론이 많이 활용되고 있다. [9]에서는 Auto-Regressive Moving Average(ARMA)를 활용해 대여소의 비치 자전거 대수를 예측하는 방법을 제안했다. 이 연구에서는 AR 부분은 개별 대여소의 시계열만 활용했지만, MA 부분에서 주변 대여소의 시계열까지 이용하여 주변 대여소의 정보를 활용하지 않았을 때보다 예측 정확도를 높였다. 하지만 시간에 따라 비치 자전거 대수의 편차는 변동이 커서 시계열이 안정적(stationary)이라는 ARMA의 가정이 위배되는 문제점이 있다. 이에 [10]에서는 ARMA를 대신에 Auto-Regressive Integrated Moving Average(ARIMA) 모델을 기반으로 한 방법을 제안하였다. 이 논문에서도 예측 정확도를 높이기 위해 주변 대여소의 정보를 사용하여 모델링하였다.

기계 학습 기반의 연구에서는 일반화 선형 회귀(generalized linear regression), 서포트 벡터 머신(SVM), 랜덤 포레스트 등 다양한 회귀 알고리즘을 활용하고 있다. [11]은 바르셀로나 공유자전거 데이터를 활용해 대여소에 비치된 자전거의 대수를 10분에서부터 120분까지 여러 시간 단위에 대해 베이지안 네트워크 기반의 예측 모형을 개발하였다. [12]는 선형 혼합 모형(linear mixed model)을 이용해 대여소별로 1시간 간격으로 반납과 대여를 예측하는 모형을 학습하였다. 사람이 외부에 바로 노출되는 자전거의 특성을 고려하여 기온, 습도 등의 날씨 변수와 함께 시간 변수, 대여소 주변의 도로나 다른 공유자전거 대여소 등의 자전거 관련 시설 변수, 대여소가 위치한 장소의 용도나 구축환경(built environment)을 파악할 수 있는 변수, 인구통계학적 변수 등을 설명변수도 활용하였다. [13]에서는 다음 날의 자전거 이용 대수를 예측하는 모형을

ridge 회귀, SVM, 랜덤 포레스트 등 다양한 회귀 방법을 이용해 구축하였다. [17]은 대여소별로 1시간 동안 자전거의 수요를 그래프 합성곱 신경망(graph convolutional neural network) 모형을 활용해 예측하는 방안을 제시하였다.

최근에는 대여소 단위가 아니라 대여소를 여러 개의 군집으로 묶은 다음 군집 단위의 수요 예측 모형을 학습하고 이를 대여소 단위 수요 예측에 활용하는 계층적 방식에 대한 연구가 늘고 있다. 군집 단위로 수요를 예측했을 때의 가장 큰 장점은 대여소 단위의 수요에 비해 변동성이 적어 정확도 높은 예측 모형을 개발할 수 있다는 점이다[15]. 비치 자전거의 불균형 문제로 인해 자전거가 비치되어 있지 않아 놓친 수요가 있으므로, 개별 대여소의 수요는 변동폭이 커 정확한 예측이 어렵다. 그러나 특정 대여소에 자전거가 없는 경우에는 사용자들은 자전거가 비치된 도보로 이동 가능한 가까운 대여소에서 자전거를 대여하기도 하므로 인근 대여소끼리는 서로 수요를 공유한다. 그리고 지리적으로 인접한 대여소들은 인근에서 발생하는 이벤트 등에 동일하게 영향을 받을 수 있다. 그러므로 군집 단위의 수요는 개별 대여소 단위에 비해 더 안정적이다. 자전거의 재배치를 위해서는 군집 단위 모형을 이용해 대여소 단위수요를 추정하는 과정이 추가적으로 필요하지만 대여소의 수요 예측에 군집 수준의 모형을 활용하게 되면 대여소 단위의 예측 모형만으로 수요 예측을 했을 때보다 더 정확한 예측이 가능하다[14].

계층적 방식에서 대여소의 군집을 얻기 위해 기존 연구들은 대여소의 위치 정보와 대여소의 자전거 이용 패턴을 활용하고 있다. 즉, 지리적으로 가까우면서 이용 패턴이 비슷한 대여소를 같은 군집으로 묶는다. [15]는 k-means 군집을 위치 정보와 이용 패턴에 번갈아 가며 적용하는 이분(bipartite) 군집 방법을 제안하였는데, 여기서 이용 패턴은 시간대별로 특정 대여소에서 대여된 자전거가 반납된 군집의 비중으로 규정하였다. [18]은 [15]와 마찬가지로 위치 정보와 시간대별로 특정 대여소에서 대여된 자전거가 반납된 군집의 비중에 affinity propagation을 적용해 군집을 형성하였다. [16]은 [18]과 군집 과정을 같지만 군집의 개수를 지정할 수 없는 affinity propagation 대신 Gaussian mixture model(GMM)을 사용하였다.

이들 방법은 공통으로 시간대별로 특정 대여소에서 대여된 자전거가 반납된 군집의 비중을 대여소별 이용 패턴으로 규정하고 이를 통해서 군집 분석을 수행한다. 그러나 이와 같은 이용 패턴은 군집 결과에 따라 바뀌므로 더는 군집의 구성이 바뀌기 전까지 반복해서 군집 알고리즘을 적용해야 해 시간이 오래 걸리고 군집의 경계에 있는 대여소로 인해 수렴하지 않을 가능성이 있다. 또한, k-means나 GMM을 이용하는 경우 초기값에 따라 군집 결과가 매번 달라지는 문제도 존재한다.

3. 제안 방법론

3.1 초기 중심점 설정

제안하는 군집 알고리즘은 k-means를 기반으로 하지만 매번 같은 결과를 얻기 위해 데이터를 기반으로 초기 중심점을 결정하는 방법을 적용해 항상 같은 군집 결과를 얻을 수 있게 한다. 이 연구에서는 두 가지 초기 중심점 결정 방법을 이용하였다. 첫 번째 방법(CI1로 지칭)은 실제 군집의 중심점과 가까운 초기 중심점을 구하기 위해 서로 떨어져 있으면서 샘플의 밀도가 높은 지역에서 중심을 찾는다[19]. 두 번째 방법(CI2로 지칭)은 PCA를 이용해 원하는 군집의 개수에 도달할 때까지 공간을 나눈 다음, 개별 군집에 속하는 샘플의 평균으로 초기 중심점을 구한다[20].

3.2 k-means 기반 군집 알고리즘

이 연구에서는 시간대별로 특정 대여소에서 대여된 자전거가 반납된 군집의 비중을 대여소의 이용 패턴으로 정의하는 것의 단점을 극복하기 위해 군집 결과가 필요 없는 이용 패턴을 사용하는 방법을 제안한다. 사람들은 출근, 퇴근, 통학, 쇼핑, 여가 등 다양한 목적으로 한 장소에서 다른 장소로 이동하는데, 이동 목적에 따라서 주로 이동하는 시간대가 다르고 이동이 발생하는 지역도 다르므로[21-23], 특정 지역의 시간대별 승/하차의 패턴만 보아도 그 지역이 어떤 특징을 지닌 지역인지 파악할 수 있다[24-25]. 공유자전거도 다른 이동수단처럼 대여소가 설치된 지역의 특성에 따라 시간대별로 대여나 반납, 이용 패턴이 다르게 나타난다. 일반적으로 주거지에 위치한 대여소는 오전에는 반납보다 대여가 많고, 저녁에는 반납이 더 많은 경향을 보인다. 회사나 상업 시설이 밀집된 지역은 주거지와는 반대의 이용 패턴이 나타난다. 즉, 시간에 따른 대여와 반납의 변화 패턴이 대여소가 어떤 특징을 갖는 지역에 위치하는지 간접적으로 보여 준다[12,26].

이에 제안하는 군집 프로세스는 이전 단계의 군집 결과에 영향을 받지 않는 대여소의 시간대별 대여, 반납 패턴을 대여소의 이용 패턴으로 정의한다. 다만, 대여소에 따라 이용객 수가 크게 차이가 날 수 있으므로 대여, 반납 횟수 대신 시간별 대여 비중과 반납 비중을 이용한다. 특정 대여소 s의 이용 패턴 벡터 x_s는 다음과 같이 정의한다.

$$x_s^p = \left(\frac{x_{s,rent,1}}{Z_{rent}}, \ldots, \frac{x_{s,rent,24}}{Z_{rent}}, \frac{x_{s,return,1}}{Z_{return}}, \ldots, \frac{x_{s,return,24}}{Z_{return}} \right) \quad (1)$$

여기서 $x_{s,.,t}^p$는 대여소 s의 t시간대의 대여 또는 반납 횟수를 나타내고, t는 1부터 24까지의 정수로 자정부터 1시간 간격의 시간대를 의미한다. 또한 Z_{rent}와 Z_{return}는 특성 대여소의 전체 대여와 반납 횟수를 나타낸다. 추가로 자전거는 신체가 외부에 노출되기 때문에 이용이 날씨에 영향을 많이 받는데, 특히 겨울에는 낮은 기온으로 인해 전체적으로 이용이 급감하며 이용 시간대도 달라진다[27-28]. 이에 겨울과 다른 계절을 구분하여 x_s^p를 구하고 이를 연결하여 대여소별 이용 패턴 벡터로 사용한다.

여기에 기존 군집 알고리즘과 동일하게 인접한 대여소가 같

Table 1. The Proposed Clustering Algorithm

Algorithm 1: k-means based clustering algorithm for bike-sharing systems

1 Input: number of clusters K, location vectors of stations $\{x_i^l\}_{i=1}^N$, temporal usage vectors of stations $\{x_i^p\}_{i=1}^N$
Output: Cluster assignment for stations, $\{l_i\}_{i=1}^N$
2 Determine K initial spatial centroids $(c_k^l; k \in \{1,2,\ldots,K\})$ using locations vectors of stations
3 Calculate K initial temporal centroids $(c_k^p; k \in \{1,2,\ldots,K\})$ by computing arithmetic mean of temporal vectors of stations within the same A_k obtained during determination of the K initial spatial centroids
4 Apply k-means clustering with K initial spatial and temporal centroids using the distance function between a station s and a cluster k defined as Equation (2)

$$dist(s,k) = d_l(x_s^l, c_k^l) + \alpha \cdot d_p(x_s^p, c_k^p) \quad (2)$$

은 군집에 속할 수 있도록 이용 패턴과 함께 위치 정보도 군집에 사용한다. 기존 방법과의 차이점은 군집 시간을 단축하기 위해 위치와 이용 패턴을 다른 단계로 분리하여 군집 분석을 수행하는 것이 아니라 위치와 이용 패턴을 동시에 이용하는 점이다. 지리적으로 인접한 대여소 간에는 비슷한 이용 패턴을 보이는 경향이 있으므로 위치를 기반으로 군집을 형성하는 것에 더 중점을 두고 이용 패턴을 추가로 고려하는 방식을 제안한다. 이를 통해 지리적으로 인접한 대여소를 묶어 군집을 형성할 때, 서로 다른 이용 패턴을 보이는 대여소 사이를 경계로 하여 군집이 나뉘면서, 더 좋은 성능을 기대할 수 있다. 제안하는 k-means 기반의 군집 알고리즘은 Table 1에 정리하였다.

Equation (2)에서 $d_l(x_s^l, c_k^l)$는 대여소 간 거리로 km 단위로 계산하며, $d_p(x_s^p, c_k^p)$는 대여소의 이용 패턴 벡터 간의 유클리디안 거리를 의미한다. α는 이용 패턴 거리에 대한 가중치로 α 값을 크게 설정하면 대여소 간의 이용 패턴의 유사도가 군집을 구성하는데 끼치는 영향력이 증가한다.

4. 실험 설계

4.1 사용 데이터 및 데이터 전처리

본 논문에서는 서울의 공유자전거 시스템인 '따릉이' 데이터를 활용하여 실험을 진행하였다. 서울 열린 데이터 광장을 통해 공개된 따릉이의 대여 이력 데이터는 대여 대여소, 대여 시간, 반납 대여소, 반납 시간 등의 정보를 포함하고 있다. 이 연구에서는 2018년 1년 내내 운영되었던 893개의 대여소에 대해서 실험을 진행하였다.

군집의 성능은 군집 단위의 수요 예측의 정확도를 기준으로 평가하였다. 수요 예측을 위한 설명변수로는 기존 문헌을 참고하여 기상 변수, 환경 변수, 인구통계 변수 등을 선정했다[12,29,30]. 먼저, 기상 변수는 기상청에서 제공하는 1시간 단위의 관측 데이터로부터 기온, 풍속, 습도, 운량, 적설량, 강수 여부, 미세먼지 변수를 추출해 사용했다. 대여소 주변 지역에 대한 환경 변수는 카카오 API와 구글 지도 API, 행정안전부(한국지역정보개발원), 관광체육국, 서울도서관, 노동민생정책관, 스마트도시정책관 등에서 제공하는 데이터를 수집했으며, [15]와 동일하게 정류소 250 m 반경 내의 음식점, 카페, 관광 명소, 전통 시장 등의 시설 수를 나타내는 변수와 정류소로부터 가장 가까운 공원, 자전거 도로, 대학교, 운동 시설 등의 시설까지의 거리를 나타내는 변수를 생성했다. 직접 페달을 밟아야 하는 자전거의 특성을 고려해 다른 대여소로 이동할 때의 평균 고도 변화를 계산하여 변수로 활용하였다. 인구통계 변수는 통계지리정보서비스에서 집계구별 인구, 가구, 사업체 통계 데이터를 이용해서 250 m 반경 내의 총인구, 20~40대 인구, 1인 가구 비율, 종사자 수, 사업체수 변수를 생성하였다. 또한, 공유자전거는 대중교통과 연계하여 대중교통 활성화에 기여하므로 250 m 반경 내 버스 정류장과 지하철역을 이용하는 승객수도 변수에 포함시켰다.

종속 변수는 대여소별 시간별 대여횟수로 두고 따릉이 데이터로부터 생성하였다. 2018년 데이터로 예측 모형을 학습하고, 2019년 1월부터 11월까지 11개월간 데이터로 모형을 평가하였다. 이때 따릉이의 이용 패턴이 주중과 주말에 크게 다르게 나타나 주중/주말을 구분하여 실험을 진행했다. 즉, 군집을 위한 대여소별 x_s^p를 주중/주말을 구분해서 구하고 각기 다른 군집 결과를 얻은 다음, 수요 예측 모형도 주중과 주말에 대해 따로 학습하였다.

설명변수와 종속 변수 모두 군집의 구성이 바뀌면 달라지기 때문에 결측치 처리 등과 같은 기본적인 데이터 전처리를 거친 후, 예측 모형 학습을 위한 변수 생성은 군집 분석 후 진행하였다.

4.2 실험 과정

실험은 군집, 설명변수 및 종속 변수 정제, 수요 예측 모형 학습, 평가의 단계로 진행하였다. 제안한 군집 방법은 기존 방법 중에서는 bipartite k-means(Bi-kMeans)[15]와 two-level GMM(TL-GMM)[16]과 비교하였다. 그리고 이용 패턴의 영향력을 알아보기 위해 위치 정보만으로 k-means 군집 알고리즘을 적용한 경우(Loc-kMeans)와도 비교를 진행했다. 기존 군집 방법은 군집 결과가 수렴할 때까지 군집 프로세스를 반복 적용하는데, 최대 반복 횟수는 100회로 설정했다. 그리고 초기값에 따라 매번 최종 군집 결과가 달라지므로 Bi-kMeans, TL-GMM 모두 5회씩 수행하였다. 군집의 개수는 Bi-kMeans와 TL-GMM이 적용된 뉴욕시와 워싱턴 D.C.의 공유자전거 시스템의 대여소가 300개 남짓으로 따릉이의 절반도 되지 않는 것과 서울의 면적이 이 두 도시에 비해 넓은 것을 감안해, 74, 76, 78, 80, 82로 총 5개 값에 대해 실험을 진행했다. Bi-kMeans는 최종 군집의 개수 외에도 중간

단계에서 필요한 군집 개수 파라미터가 하나 더 있는데 이는 기존 연구와 동일하게 5부터 5 간격으로 최종 군집 개수의 60%를 초과하지 않는 선까지 설정하였다[15].

제안하는 방법에서 사용하는 α는 과도하게 크게 설정하면 군집 내 대여소 간의 지리적 인접성이 떨어질 수 있으므로 대여소 간의 거리 분포를 확인한 후, $1,2,...,5$로 값을 바꿔가며 최적의 α를 교차검증으로 결정했다.

군집 단위 수요 예측 모형은 최대 깊이가 20인 의사결정나무 100개로 구성된 랜덤 포레스트를 이용했다. 그리고 군집 수준의 예측 모형을 활용하는 것이 개별 대여소의 수요 예측의 정확도 향상에 도움을 주는지 확인하기 위해 군집 수준의 예측 모형을 이용해 개별 대여소의 수요 예측하고 이를 대여소 수준의 예측 모형의 결과와 비교하였다. 이때 군집 수준의 예측 모형을 활용한 대여소별 수요 예측치는 동일 군집 내에서 개별 대여소의 과거 수요에 비례해서 분배하는 방식을 이용했다. 대여소 수준의 예측 모형은 랜덤 포레스트와 ridge 회귀를 사용하였다.

수요 예측의 정확도는 root mean square error(RMSE), mean absolute error(MAE), symmetric mean absolute relative error (SMARE), 총 3개의 지표를 이용해서 측정했다. 각 지표는 다음과 같이 정의된다.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y_i})^2/n} \tag{3}$$

$$\mathrm{MAE} = \left|y_i - \hat{y_i}\right|/n \tag{4}$$

$$\mathrm{SMARE} = \sum_{i=1}^{n}\left|y_i - \hat{y_i}\right|/\left\{n(y_i + \hat{y_i})\right\} \tag{5}$$

여기서 n은 총관측치의 개수로, 여기서는 데이터에 포함한 날 수, 군집의 개수, 24를 모두 곱한 값이다.

5. 실험 결과

Table 2와 3은 군집 수준에서의 주중과 주말에 대해 평가 결과를 정리한 것이다. 'Proposed'라고 표시된 것이 제안한 방법이고 CI1과 CI2는 사용한 초기 중심점 결정 방법을 뜻한다. 비교 방법에 대한 실험은 5회 수행했기 때문에 평균과 함께 괄호 안에 표준 편차를 같이 표기하였다.

TL-GMM이 가장 우수하게 나타난 80개의 군집을 사용한 주중 모델의 MAE를 제외하고는 주중과 주말, 군집 개수에 관계없이 제안한 방법이 비교 방법보다 더 좋은 결과를 보였다. 특히 RMSE와 SMARE에서는 CI1로 초기 중심점을 구한 모델이 거의 모든 경우에서 가장 우수하게 나타났다. MAE는 CI2를 이용한 모델이 가장 좋은 경우가 많았지만, 이들은 RMSE와 SMARE는 비교 방법보다 성능이 떨어지는 경우가 있었다. Loc-kMeans는 RMSE 측면에서는 가장 안 좋은 결과를 보이지만, MAE와 SMARE에서는 다른 비교 방법과 큰 차이를 보이지 않았다.

RMSE는 MAE에 비해 오차의 크기가 큰 관측치에 더 큰 가중치가 부여되므로 RMSE가 다른 방법에 비해 좋다는 것은 오차의 크기가 큰 관측치가 상대적으로 더 적다고 볼 수 있다. SMARE은 같은 오차라면 실제 값보다 더 크게 예측하는 경우가 실제 값보다 더 작게 예측하는 것에 비해 작은 값을 갖는다. 그러므로 SMARE에서 우위가 있다면 절대적으로 오차의 합이 적을 수도 있지만, 과대 예측할 때의 오차가 과소 예측할 때의 오차보다 작은 경향 때문일 수 있다.

Table 4와 5는 오차의 분포의 특징을 정리해서 보여준다. 이 표에서 Error는 오차의 절대값의 평균이고 Under는 과소 예측한 샘플의 비율을 계산한 것이다. 괄호 안의 S와 L은 실제 관측치의 값이 5 이하인 그룹과 20 이상인 그룹을 뜻한다. 즉, Error(S)는 실제 종속변수가 5 이하인 샘플에 대해

Table 2. Evaluation Results: Weekdays

k	Metric	Proposed(CI1)	Proposed(CI2)	Loc-kMeans	Bi-kMeans	TL-GMM
74	RMSE	**8.3990**	8.9016	9.4927	8.5956(0.1296)	8.8855(0.1993)
	MAE	4.7870	**4.7377**	4.7756	4.7760(0.0245)	4.7599(0.0266)
	SMARE	**0.2233**	0.2430	0.2356	0.2354(0.0093)	0.2445(0.0068)
76	RMSE	**8.1574**	8.6202	9.2974	8.4200(0.1572)	8.7639(0.2829)
	MAE	4.7159	**4.6417**	4.6798	4.6967(0.0265)	4.6778(0.0190)
	SMARE	**0.2254**	0.2421	0.2351	0.2381(0.0070)	0.2440(0.0071)
78	RMSE	**7.9284**	8.4826	8.4934	8.6109(0.5223)	8.5311(0.2028)
	MAE	4.6048	**4.5506**	4.5915	4.6265(0.0220)	4.5771(0.0414)
	SMARE	**0.2272**	0.2454	0.2490	0.2427(0.0056)	0.2532(0.0117)
80	RMSE	**7.8126**	8.1658	9.1799	8.4434(0.3254)	8.7268(0.2335)
	MAE	4.5152	4.4530	4.5334	4.5215(0.0195)	**4.5070(0.0450)**
	SMARE	**0.2317**	0.2465	0.2515	0.2460(0.0054)	0.2588(0.0077)
82	RMSE	**7.6190**	8.0531	8.7500	8.0005(0.1653)	8.6530(0.3769)
	MAE	4.4301	**4.3837**	4.4259	4.4517(0.0216)	4.4090(0.0475)
	SMARE	**0.2334**	0.2494	0.2479	0.2486(0.0033)	0.2537(0.0055)

Table 3. Evaluation Results: Weekends

k	Metric	Proposed(CI1)	Proposed(CI2)	Loc-kMeans	Bi-kMeans	TL-GMM
74	RMSE	**9.2102**	9.6501	10.0187	9.2714(0.1080)	9.7335(0.1325)
	MAE	5.2524	**5.2194**	5.2825	5.2518(0.0158)	5.2571(0.0309)
	SMARE	**0.2630**	0.2816	0.2761	0.2727(0.0053)	0.2830(0.0082)
76	RMSE	**8.7922**	9.3615	9.8375	9.1500(0.1866)	9.6452(0.1224)
	MAE	5.1521	**5.1298**	5.2027	5.1656(0.0284)	5.1555(0.0386)
	SMARE	**0.2631**	0.2812	0.2767	0.2751(0.0037)	0.2909(0.0035)
78	RMSE	**8.5614**	9.2037	9.0966	9.0424(0.3132)	9.7346(0.6913)
	MAE	**5.0258**	5.0431	5.0443	5.0562(0.0439)	5.0863(0.0143)
	SMARE	**0.2650**	0.2848	0.2881	0.2766(0.0040)	0.2910(0.0051)
80	RMSE	**8.4731**	8.7972	9.6116	8.8158(0.1039)	9.1922(0.2497)
	MAE	4.9359	**4.9197**	4.9855	4.9654(0.0076)	4.9407(0.0530)
	SMARE	0.2695	0.2859	0.2908	0.2807(0.0034)	0.2920(0.0047)
82	RMSE	**8.2317**	8.6802	9.3680	8.6850(0.1447)	9.1368(0.3105)
	MAE	**4.8362**	4.8598	4.8962	4.8787(0.0309)	4.8755(0.0234)
	SMARE	**0.2718**	0.2896	0.2891	0.2836(0.0013)	0.2946(0.0017)

Table 4. Characteristics in Error Distributions: Weekdays

k	Metric	Proposed(CI1)	Proposed(CI2)	Loc-kMeans	Bi-kMeans	TL-GMM
74	Error(S)	2.2139	2.0065	2.1286	2.1376(0.0830)	1.6573(0.0379)
	Error(L)	9.2287	9.6082	9.6590	9.4182(0.2013)	10.0923(0.2583)
	Under(S)	0.2753	0.2893	0.2915	0.2859(0.0090)	0.4649(0.0077)
	Under(L)	0.6484	0.6618	0.6657	0.6621(0.0073)	0.5229(0.0178)
76	Error(S)	2.1821	1.9878	2.0950	2.1469(0.0705)	1.6696(0.0554)
	Error(L)	8.9996	9.5247	9.7651	9.2969(0.2522)	10.1009(0.2675)
	Under(S)	0.2721	0.2913	0.2981	0.2917(0.0040)	0.4630(0.0088)
	Under(L)	0.6487	0.6548	0.6915	0.6638(0.0113)	0.5100(0.0152)
78	Error(S)	2.1630	1.9571	2.0074	2.0832(0.0480)	1.6129(0.0390)
	Error(L)	8.8652	9.4390	9.6081	9.2225(0.1467)	10.1694(0.2968)
	Under(S)	0.2773	0.2943	0.2979	0.2932(0.0013)	0.4623(0.0077)
	Under(L)	0.6360	0.6659	0.6732	0.6655(0.0182)	0.5262(0.0082)
80	Error(S)	2.0902	1.9174	1.9789	2.0535(0.0332)	1.6063(0.0228)
	Error(L)	8.8430	9.4076	9.6525	9.3592(0.0832)	10.1007(0.1857)
	Under(S)	0.2865	0.2950	0.3014	0.2924(0.0050)	0.4585(0.0098)
	Under(L)	0.6454	0.6700	0.6848	0.6738(0.0053)	0.5230(0.0084)
82	Error(S)	2.0706	1.9008	1.9831	2.0439(0.0346)	1.6219(0.0429)
	Error(L)	8.7836	9.3749	9.5607	9.1017(0.2620)	10.1978(0.3694)
	Under(S)	0.2895	0.2953	0.3089	0.2920(0.0076)	0.4664(0.0063)
	Under(L)	0.6539	0.6541	0.6764	0.6659(0.0112)	0.5186(0.0141)

오차의 절대값의 평균을 구한 것이다. 군집의 개수나 군집 분석 알고리즘에 따라 약간의 차이가 있지만, 5는 관측치의 중앙값보다 작은 값이고 20은 Q3 부근의 값이다.

Loc-kMeans나 Bi-kMeans와 비교했을 때, CI1을 이용한 제안 방법은 S에 대해서는 Error가 더 크지만, L에 대한 Error는 더 작다. 반면에 CI2 이용했을 때는 반대의 경향이 나타난다. TL-GMM은 모든 방법 중에서 S에 대한 Error는 가장 작지만, L에 대한 Error는 가장 크다. 또한, Loc-kMeans나 Bi-kMeans와 비교했을 때, Under의 경우에는 CI1을 이용한 제안 방법이 S와 L 모두에서 작은 경우가 많았다. CI2는 Under 값이 Loc-kMeans보다는 작은 경우가 많았지만, Bi-kMeans보다는 큰 경우가 많았다.

Table 6과 7은 대여소 수준에서의 수요 예측 결과를 정리한 것이다. 이 표에서 RF와 Ridge는 랜덤 포레스트와 ridge 회귀를 이용한 대여소 수준의 예측 모형의 성능을 의미한다. 군집 수준의 수요 예측에서와 마찬가지로 제안한 군집 알고

Table 5. Characteristics in Error Distributions: Weekends

k	Metric	Proposed(CI1)	Proposed(CI2)	Loc-kMeans	Bi-kMeans	TL-GMM
74	Error(S)	2.4094	2.2580	2.3992	2.4276(0.0286)	1.7061(0.0570)
	Error(L)	10.7692	11.4074	11.4371	10.8281(0.0561)	12.1026(0.3861)
	Under(S)	0.2537	0.2632	0.2629	0.2556(0.0061)	0.4347(0.0087)
	Under(L)	0.5991	0.5958	0.6153	0.6033(0.0080)	0.4247(0.0035)
76	Error(S)	2.4054	2.2472	2.3777	2.3872(0.0310)	1.7077(0.0250)
	Error(L)	10.4634	11.1809	11.3717	10.8624(0.1581)	12.1528(0.3464)
	Under(S)	0.2620	0.2640	0.2677	0.2599(0.0058)	0.4374(0.0027)
	Under(L)	0.6092	0.5993	0.6163	0.6074(0.0054)	0.4193(0.0125)
78	Error(S)	2.3828	2.2154	2.2880	2.3496(0.0556)	1.7095(0.0226)
	Error(L)	10.2646	11.1120	11.2277	10.7540(0.1770)	12.0163(0.4903)
	Under(S)	0.2610	0.2649	0.2625	0.2641(0.0019)	0.4370(0.0041)
	Under(L)	0.6056	0.6023	0.6072	0.6117(0.0060)	0.4204(0.0058)
80	Error(S)	2.3362	2.1803	2.2198	2.3450(0.0212)	1.6975(0.0297)
	Error(L)	10.3624	10.8751	11.2705	10.7506(0.3139)	11.7026(0.1488)
	Under(S)	0.2622	0.2672	0.2746	0.2611(0.0015)	0.4345(0.0050)
	Under(L)	0.6058	0.6031	0.6055	0.6119(0.0051)	0.4224(0.0022)
82	Error(S)	2.3089	2.1558	2.2615	2.2876(0.0223)	1.6678(0.0304)
	Error(L)	10.2539	10.8958	11.2121	10.6848(0.2089)	11.5444(0.3154)
	Under(S)	0.2629	0.2675	0.2677	0.2620(0.0029)	0.4329(0.0063)
	Under(L)	0.6097	0.6088	0.6165	0.6097(0.0034)	0.4205(0.0063)

Table 6. Evaluation Results in the Station Level: Weekdays

k	Metric	Proposed(CI1)	Proposed(CI2)	Loc-kMeans	Bi-kMeans	TL-GMM	RF	Ridge
74	RMSE	1.5420	**1.5407**	1.5547	1.5500(0.0121)	1.5602(0.0171)	1.6116	1.9700
	MAE	**0.8436**	0.8438	0.8524	0.8505(0.0061)	0.8548(0.0080)	0.9004	1.0507
	SMARE	**0.3724**	0.3729	0.3788	0.3783(0.0050)	0.3811(0.0077)	0.4067	0.5045
76	RMSE	**1.5369**	1.5391	1.5565	1.5567(0.0196)	1.5578(0.0147)	1.6116	1.9700
	MAE	0.8440	**0.8438**	0.8554	0.8525(0.0073)	0.8541(0.0058)	0.9004	1.0507
	SMARE	**0.3730**	0.3740	0.3827	0.3796(0.0057)	0.3807(0.0045)	0.4067	0.5045
78	RMSE	**1.4776**	1.4796	1.4983	1.4955(0.0189)	1.5037(0.0199)	1.6116	1.9700
	MAE	0.8064	**0.8055**	0.8158	0.8162(0.0081)	0.8180(0.0084)	0.9004	1.0507
	SMARE	0.3523	**0.3522**	0.3573	0.3594(0.0069)	0.3591(0.0073)	0.4067	0.5045
80	RMSE	**1.4798**	1.4855	1.5004	1.4975(0.0193)	1.5005(0.0205)	1.6116	1.9700
	MAE	0.8055	**0.8047**	0.8170	0.8134(0.0061)	0.8146(0.0073)	0.9004	1.0507
	SMARE	**0.3520**	0.3523	0.3612	0.3578(0.0048)	0.3585(0.0060)	0.4067	0.5045
82	RMSE	**1.4822**	1.4836	1.5075	1.5100(0.0245)	1.4956(0.0151)	1.6116	1.9700
	MAE	0.8081	**0.8066**	0.8192	0.8174(0.0059)	0.8163(0.0086)	0.9004	1.0507
	SMARE	**0.3520**	0.3528	0.3591	0.3576(0.0048)	0.3589(0.0075)	0.4067	0.5045

리즘을 이용한 군집 수준의 예측 모형을 이용해 대여소 수준의 수요 예측을 진행했을 때 가장 우수한 결과를 얻을 수 있었다. 대여소 수준의 예측 모형만을 이용했을 때는 계층적 방식으로 대여소의 수요를 예측했을 때보다 정확도가 낮았다.

각 군집 방법의 평균적인 군집 시간을 비교해보면, 제안한 방법은 CI1과 CI2로 초기 중심점을 구하는데 각각 평균 1.69초, 0.25초가 소요되고 k-means를 수행하는데 약 0.07초 정도 소요되는 반면, Bi-kMeans나 TL-GMM는 각각 평균 167.1초, 738.3초가 소요되었다. 게다가 제안한 방법은 초기 중심점을 고정했기 때문에 최종 결과가 항상 일정하므로 여러 번 수행할 필요가 없어 비교 방법에 비해 계산 속도 면에서 매우 우수함을 알 수 있다. 비교 방법 중에서는 Bi-kMeans가 TL-GMM에 비해서는 빠른 편인데 이는 k-means 알고리즘이 GMM에 비해 빠르기 때문이다.

Table 7. Evaluation Results in the Station Level: Weekends

k	Metric	Proposed(CI1)	Proposed(CI2)	Loc-kMeans	Bi-kMeans	TL-GMM	RF	Ridge
74	RMSE	1.6243	**1.6214**	1.6319	1.6306(0.0121)	1.6362(0.0171)	1.7558	2.1278
	MAE	**0.8049**	0.8053	0.8115	0.8113(0.0061)	0.8128(0.0080)	0.8668	1.0022
	SMARE	0.3759	**0.3757**	0.3810	0.3807(0.0050)	0.3820(0.0077)	0.4167	0.5059
76	RMSE	1.6212	**1.6206**	1.6262	1.6274(0.0180)	1.6368(0.0188)	1.7558	2.1278
	MAE	**0.7823**	0.7829	0.7855	0.7862(0.0080)	0.7896(0.0083)	0.8668	1.0022
	SMARE	**0.3616**	0.3618	0.3651	0.3648(0.0068)	0.3675(0.0073)	0.4167	0.5059
78	RMSE	**1.6237**	1.6249	1.6386	1.6369(0.0196)	1.6391(0.0147)	1.7558	2.1278
	MAE	**0.8068**	0.8085	0.8121	0.8149(0.0073)	0.8144(0.0058)	0.8668	1.0022
	SMARE	**0.3775**	0.3788	0.3804	0.3840(0.0057)	0.3835(0.0045)	0.4167	0.5059
80	RMSE	**1.6178**	1.6212	1.6349	1.6313(0.0193)	1.6377(0.0205)	1.7558	2.1278
	MAE	**0.7816**	0.7819	0.7865	0.7893(0.0061)	0.7896(0.0073)	0.8668	1.0022
	SMARE	0.3626	**0.3609**	0.3645	0.3676(0.0048)	0.3678(0.0060)	0.4167	0.5059
82	RMSE	**1.6340**	1.6268	1.6359	1.6408(0.0254)	1.6413(0.0162)	1.7558	2.1278
	MAE	**0.8063**	0.8082	0.8114	0.8142(0.0070)	0.8109(0.0095)	0.8668	1.0022
	SMARE	**0.3784**	0.3798	0.3812	0.3837(0.0054)	0.3816(0.0076)	0.4167	0.5059

6. 결론 및 활용방안

본 연구는 계층적 방식의 수요 예측에서 군집 수준의 모형의 정확도를 높이기 위한 새로운 k-means 기반의 군집 방법을 제안하였다. 제안 방법은 초기 중심점을 설정하는 방법을 이용해 매번 동일한 군집 결과를 얻을 수 있으며, 군집 결과에 기반하지 않는 대여소의 시간별 대여, 반납 패턴을 위치 정보와 함께 활용해 k-means를 단 한 번만 수행하므로 기존 방법보다 군집 속도가 매우 빠르다. 뿐만 아니라, 제안하는 알고리즘을 이용해 서울시 공유자전거인 따릉이 데이터로 군집을 형성하고 군집 수준과 대여소 수준에서의 수요 예측을 진행할 결과, 비교 군집 방법을 사용했을 때보다 더 나은 예측 성능을 보여주었다.

군집 수준에서 Bi-kMeans와 TL-GMM과 비교하였을 때 RMSE, MAE, SMARE 중 RMSE와 SMARE에서는 CI1을 이용한 제안 방법이 모든 경우에서 가장 좋은 결과를 보였다. CI2는 MAE에서는 가장 좋은 결과를 보이는 경우가 다수였으나 RMSE, SMARE는 비교 방법에 비해 뛰어나지 않았다. 그리고 실제값의 크기에 따라 오차의 분포를 확인해 보면 CI1을 이용한 제안 방법이 실제값이 큰 경우에 다른 방법에 비해 더 적은 오차를 보이고 과소 예측하는 비율은 TL-GMM에 비해서는 크지만 다른 방법보다는 낮은 걸 확인할 수 있다.

따릉이는 뉴욕의 City Bike 등 대다수 다른 공유자전거 시스템과 달리 빈 거치대가 없어도 반납이 되므로 원하는 대여소에 반납하지 못하는 일은 발생하지 않는다. 그러므로 자전거 불균형으로 원하는 대여소에서 대여하지 못하는 경우만 문제가 되는데, 이런 현상이 잦은 대여소는 이용량이 많은 경향이 있었다. 이런 경우, 이용이 많은 군집에서 더 정확하게 예측하고 과소 예측하기보다는 과대 예측하는 경향이 있는 것이 불균형으로 인한 이용자의 불편을 최소화하기 위해 선제적으로 자전거 부족에 대응하기 위해서는 더 적절한 군집 방법이라 할 수 있는데 제안 방법이 이런 특징을 보인다.

그뿐만 아니라, 대여소 수준에서도 계층적 방식을 활용한 수요 예측 결과가 대여소 모형만을 이용했을 때보다 우수했으며, 계층적 방식에서는 제안한 군집 알고리즘의 군집 결과를 이용했을 때 가장 좋은 성능을 보였다.

본 연구에서는 대여소별 수요 예측을 위해 군집별 수요를 군집에 속하는 대여소로 분배하는 방식을 사용했다. 추후에는 보다 정확한 대여소 단위의 수요 예측 모형을 개발할 필요가 있다.

References

[1] M. Ricci, "Bike sharing: A review of evidence on impacts and processes of implementation and operation," *Research in Transportation Business & Management*, Vol.15, pp.28-38, 2015.

[2] P. Bhardwaj and S. Gal, "The number of bike-sharing programs has doubled since 2014," [Internet] https://www.businessinsider.com/bike-sharing-programs-doubled-since-2014-public-bikes-charts-2018-7#:~:text=According%20to%20estimates%20and%20data,bikes%20available%20for%20public%20use.

[3] NACTO, "Shared Micromobility in the U.S.: 2018," 2019.

[4] J. C. García-Palomares, J. Gutiérrez, and M. Latorre, "Optimizing the location of stations in bike-sharing programs: A GIS approach," *Applied Geography*, Vol.35, No.1, pp.235-246, 2012.

[5] M. Benchimol et al., "Balancing the stations of a self service 'bike hire' system," *RAIRO Operations Research*, Vol.45, No.1, pp.37-61, 2011.

[6] J. Liu, L. Sun, W. Chen, and H. Xiong, "Rebalancing Bike Sharing Systems," in *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.1005-1014, 2016.

[7] A. Singla, M. Santoni, G. Bartok, P. Mukerji, M. Meenen, and A. Krause, "Incentivizing Users for Balancing Bike Sharing Systems," in *Proceedings of AAAI Conference on Artificial Intelligence*, pp.723-729, 2015.

[8] C. Fricker and N. Gast, "Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity," *EURO Journal on Transportation and Logistics*, Vol.5, No.3, pp.261-291, 2016.

[9] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive and Mobile Computing*, Vol.6, No.4, pp.455-466, 2010.

[10] J. W. Yoon, F. Pinelli, and F. Calabrese, "Cityride: A Predictive Bike Sharing Journey Advisor," in *Proceedings of IEEE International Conference on Mobile Data Management*, pp.306-311, 2012.

[11] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp.1420-1426, 2009.

[12] A. Faghih-Imani, N. Eluru, A. M. El-Geneidy, M. Rabbat, and U. Haq, "How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal," *Journal of Transport Geography*, Vol.41, pp. 306-314, 2014.

[13] R. Giot and R. Cherrier, "Predicting bikeshare system usage up to one day ahead," in *Proceedings of the IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems*, pp.22-29, 2014.

[14] Y. Li and Y. Zheng, "Citywide Bike Usage Prediction in a Bike-Sharing System," *IEEE Transactions on Knowledge and Data Engineering*, Vol.32, No.6 pp.1079-1091, 2019.

[15] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp.1-10, 2015.

[16] W. Jia, Y. Tan, L. Liu, J. Li, H. Zhang, and K. Zhao, "Hierarchical prediction based on two-level Gaussian mixture model clustering for bike-sharing system," Knowledge-Based Systems, Vol.178, pp.84-97, 2019.

[17] L. Lin, Z. He, and S. Peeta, "Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach," *Transportation Research Part C: Emerging Technologies*, Vol.97, pp.258-276, 2018.

[18] W. Jia, Y. Tan, and J. Li, "Hierarchical prediction based on two-level affinity propagation clustering for bike-sharing system," *IEEE Access*, Vol.6, pp.45875-45885, 2018.

[19] Fang Yuan, Zeng-Hui Meng, Hong-Xia Zhang, and Chun-Ru Dong, "A new algorithm to get the initial centroids," in *Proceedings of International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, Vol.2, pp.1191-1193, 2004.

[20] T. Su and J. Dy, "A Deterministic Method for Initializing K-Means Clustering," in *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pp.784-786, 2004.

[21] F. Devillaine, M. Munizaga, and M. Trépanier, "Detection of Activities of Public Transport Users by Analyzing Smart Card Data," *Transportation Research Record: Journal of the Transportation Research Board*, Vol.2276, No.1, pp.48-55, 2012.

[22] L. Gong, X. Liu, L. Wu, and Y. Liu, "Inferring trip purposes and uncovering travel patterns from taxi trajectory data," *Cartography and Geographic Information Science*, Vol.43, No.2 pp.103-114, 2016.

[23] X. Liu, C. Kang, L. Gong, and Y. Liu, "Incorporating spatial interaction patterns in classifying and understanding urban land use," *International Journal of Geographical Information Science*, Vol.30, No.2, pp.334-350, 2016.

[24] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai," *Landscape Urban Planning*, Vol.106, No.1, pp.73-87, 2012.

[25] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," *Journal of Transport Geography*, Vol.43, pp.78-90, 2015.

[26] A. Faghih-Imani and N. Eluru, "Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system," *Journal of Transport Geography*, Vol.54, pp.218-227, 2016.

[27] K. Gebhart and R. B. Noland, "The impact of weather conditions on bikeshare trips in Washington, DC," *Transportation*, Vol.41, No.6, pp.1205-1225, 2014.

[28] P. Lin, J. Weng, Q. Liang, D. Alivanistos, and S. Ma, "Impact of Weather Conditions and Built Environment on Public Bikesharing Trips in Beijing," *Networks and Spatial Economics*, Vol.20, No.1, pp.1-17, 2020.

[29] T. D. , N. Ovtracht, and B. F. d'Arcier, "Modeling Bike Sharing System using Built Environment Factors," *Procedia CIRP*, Vol.30, pp.293-298, 2015.

[30] W. El-Assi, M. Salah Mahmoud, and K. Nurul Habib, "Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto," *Transportation*, Vol.44, No.3, pp.589-613, 2017.

김 경 옥

https://orcid.org/0000-0002-0196-3832
e-mail : kyoungok.kim@seoultech.ac.kr
2008년 POSTECH 신소재공학과(학사)
2013년 POSTECH 산업경영공학과 (석·박사통합)
2013년 ~ 2015년 삼성경제연구소 산업전략1실 연구원
2015년 ~ 현 재 서울과학기술대학교 산업공학과 조교수
관심분야 : Transport Data Analysis, Machine Learning

이 창 환

https://orcid.org/0000-0001-5235-5346
e-mail : lch1181@ds.seoultech.ac.kr
2016년 국민대학교 경영학과(학사)
2020년 서울과학기술대학교 데이터사이언스학과(석사)
관심분야 : Data Mining, Machine Learning