서울시 대중교통 정류장 군집화에 관한 연구†

A Study on the Clustering of Public Transportation Stations in Seoul

전인우*, 이민혁, 양현재, 전철민
Inwoo Jeon, Minhyuck Lee, Hyunjae Yang, Chulmin Jun
서울시립대학교 공간정보공학과
{yugo123, lmhll123, greenbag1897, cmjun}@uos.ac.kr

요 약

최근 대중교통의 지역별 접근성 분석에서 스마트카드 통행 데이터를 활용한 연구가 활발히 진행되고 있다. 기존 연구에서 이용한 분석 단위는 정류장 단위나 정류장을 행정동 혹은 임의의 교통존으로 군집화한 단위를 이용하고 있다. 하지만 분석 단위가 클수록 서로 다른 접근성 지표를 보이는 정류장들이 하나의 군집에 속할 가능성이 커지게 된다. 따라서 미시적 분석을 위한 작은 크기의 군집 설정이 필요하며 본 연구에서는 기존의 행정동 혹은 교통존보다 세밀한 단위의 군집 생성을 위한 군집화 분석을 수행하였다. 군집내 유사도는 정류장 사이의 물리적 거리를 고려한 공간적 인접성을 이용하였고, *k*-means, DB-SCAN 군집 알고리즘을 서울시 대중교통 정류장에 적용하여 행정동 단위의 군집 결과와 비교·분석하였다.

1. 서론

대중교통을 평가하는 지표로 지역별 접근성이 활용되고 있다. 기존에는 행정동, 교통존 단위의 O/D 기반 접근성 분석이 수행되었고, 스마트카드 통행 데이터의 활용이 가능해지면서 정류장 단위의 분석도 수행되었다. 하지만, 정류장 단위의 접근성 분석은 실제 대중교통 이용객의 이동 패턴의 반영에 한계가 있다. 실제 승객은 정류소-정류소간의 경로 선택이 아니라 출발지점 정류소들-도착지점 정류소들간의 경로들 중 하나의 경로를 선택하기 때문이다. 따라서 정류소들을 군집으로 묶어서 분석하는 연구들이 수행되고 있다. 다만, 기존 군집은 행정동 단위나 임의의 교통존 단위로 설정되는데 이는 세밀한 분석을 수행하기에는 너무 크다. 이에 본 연구에서는 기존 군집보다 크기가 작으며 승객의 경로 선택 패턴을 반영할 수 있는 세밀한 군집 단위 설정을 위한 군집화 분석을 수행하였다. 대표적인 군집 알고리즘인 k-means, DB-SCAN을 서울시 대중교통 정류장 13,525개에 적용하였고, 군집내 유사도는 정류장들 사이의 공간적 인접성을 이용하였다. 또한, 군집 알고리즘의 적용 결과를 행정동 단위의 정류장 군집 결과와 비교·분석하여 군집의 크기와 파라미터에 따른 영향을 살펴보았다.

2. 분석방법론

본 연구에서는 *k*-means 알고리즘과 DB-SCAN 알고리즘을 이용하여 정류장들을 군집화하였다[1,2]. *k*-means 알고리즘은 군집의 개수를 입력값으로 받은 뒤, 군집

† 본 연구는 국토교통부 국토교통기술촉진연구사업의 연구비지원(17CTAP-C133228-01)에 의해 수행되었습니다.

[표 1] 각 군집화 방법별 실험 결과

구분			군집 개수	정류장들 간 평균거리(m)	정류장과 군집 중심과의 평균거리(m)	공간적 응집도
DB-SCAN	eps(m)	150	1379	358.13	205.44	2.28
	minpt	3				
	eps(m)	200	604	1117.26	576.41	6.79
	minpt	5				
k-means			600	373.30	294.97	0.58
			1400	206.47	161.59	1.36
행정동 단위			428 (분할구역 수)	688.20	534.46	-

중심과 군집 내 포인트들 간의 거리 합인 SSE(Sum of Squared Error)를 최소화하는 군집을 형성한다. DB-SCAN 알고리즘은 임의의 포인트로부터 일정 거리(eps) 내 군집을 구성하는 최소 포인트의 개수(minpt)가 있는지 판단하여 점진적으로 군집을 확대해나가는 알고리즘이다.

본 연구에서는 *k*-means의 군집 개수를 결정하기 위해 Elbow 기준을 적용하였다[3]. Elbow 기준은 SSE를 군집의 개수에 따라 순차적으로 계산하여 그래프를 그렸을 때, 해당 그래프의 변곡점을 최적의 군집 개수로 결정하는 방식이다. DB-SCAN의 minpt와 eps는 다음과 같은 방식으로 결정하였다. 우선 minpt와 eps를 순차적으로 증가시킨 2차원 행렬을 구성한다. 그리고 행렬 값으로는 SSE를 입력하여 SSE를 최소로 하는 minpt와 eps를 최적의 값으로 결정하였다.

3. 실험 및 분석

k-means는 군집의 개수가 600개일 때 SSE가 최소였고, DB-SCAN은 군집의 개수가 1379개 일 때 eps가 150m, minpt가 3개일 때 SSE가 최소인 것으로 나타났다. 유사한 군집 개수일 때 *k*-means와 DB-SCAN의 군집 결과를 비교하기 위해 본 실험에서는 *k*-means의 군집 개수가 1400개일 경우와 DB-SCAN 군집 개수가 604개일 경우(eps: 200m, minpt: 5)를 추가하였다.

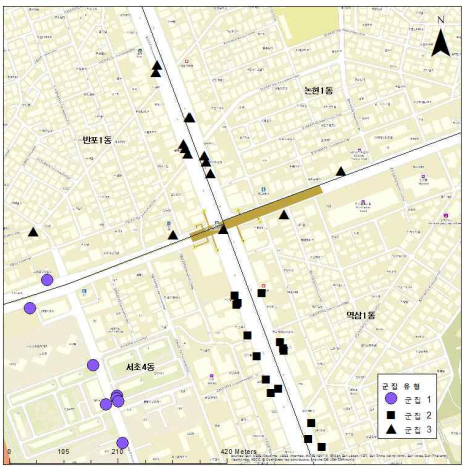
[표 1]은 군집 알고리즘의 적용 결과와 행정동 단위의 정류장 집계 결과를 종합한 것이다. 군집 알고리즘의 적용 결과를 평가하는 지표는 공간적 응집도를 이용하였다. 공간적 응집도란 군집 내부 점들의 인접성과 군집 중심들 간의 인접성을 모두 고려한 지표로 값이 낮을수록 좋은 응집도를 의미한다.

행정동 단위의 집계 결과는 두 가지 군집 알고리즘을 적용한 결과보다 대체로 정류장 사이의 공간적 인접성이 낮은 것을 확인 할 수 있었다. 또한 비슷한 개수의 군집을 가지는 두 알고리즘 적용 결과를 살펴보면, *k*-means가 DB-SCAN에 비해 공간적 응집도가 더 높은 것을 확인할 수 있었고 군집의 개수가 약 600개일때는 그 차이가 매우 컸다.

[그림 1]은 행정동 기반 집계를 포함한 각 군집화 방법을 이용한 결과를 시각화한 모습이다. DB-SCAN은 eps 150m, minpt 3개를 적용하였고 *k*-means는 군집 개수를 1400개로 적용한 경우이다. 대상 지역은 신논현역을 중심으로 반포1동, 논현1동, 서초4동, 역삼1동을 나타내며 신논현역을 남북으로 통과하는 강남대로와 동서로 통과하는 봉은사로를 기준으로 행정동 경계가 구분되어 있다.

(a) 행정동 경계를 이용한 정류장 집계

(b) DB-SCAN을 이용한 정류장 군집 결과



(c) *k*-means를 이용한 정류장 군집 결과

[그림 1] 군집화 결과 시각화

[그림 1]의 (a)는 행정동 구분을 이용하여 정류장들을 집계한 결과이다. 도로에 인접한 정류장들이 공간적 인접성이 뛰어남에도 불구하고 행정경계로 구분하였기 때문에 서로 다른 군집유형에 속하는 것을 확인할 수 있다. 이처럼 도로를 마주하고 있는 정류장들의 경우, 동일한 노선의 하행정류장과 상행정류장의 역할을 하는 경우가 많다. 따라서 두 정류장을 동일한 군집에 포함시켜 대중교통 이용패턴을 분석하는 것이 보다 합리적일 것으로 판단된다.

[그림 1]의 (b)는 DB-SCAN 방식을 이용하여 정류장들을 군집화한 결과이다. 강남대로에 인접한 정류장들이 모두 동일한 군집1 유형에 속하는 것으로 나타났고 봉은사로에 인접한 정류장들은 군집1과 군집2 유형으로 구분되었다. 그리고 서초4동의 아파트 단지 인근 정류장들이 군집3 유형에 속하였다. 행정동을 이용한 집계 방식보다는 공간적 인접성을 고려하여 정류장들이 군집화된 것을 확인할 수 있다. 다만, DB-SCAN 방식이 eps를 이용하여 더 이상 minpt 조건에 해당하지 않을 때까지 군집을 확장해나가기 때문에 정류장의 밀도가 높은 지역은 군집이 다소 크게 형성될 수 있다. 따라서 강남대로는 정류장 밀도가 높은 지역이기 때문에 군집의 범위가 넓게 형성된 것으로 판단된다.

[그림 2]의 (c)는 *k*-means를 이용한 군집결과이다. 반포1동과 논현1동에 위치한 정류장들이 군집3 유형을 형성하였고 강남대로에 인접한 역삼1동과 서초4동의 정류장들은 군집2 유형을 형성하였다. 그리고 서초4동의 아파트 단지에 위치한 정류장들이 군집1 유형에 속하는 것으로 나타났다. 행정동을 이용한 집계 방식에 비해서는 DB-SCAN과 마찬가지로 공간적 인접성을 고려한 군집 결과가 나타났으며, DB-SCAN에서는 강남대로에 인접한 정류장들이 모두 하나의 군집에 포함된 반면, *k*-means에서는 두 가지 군집 유형으로

분리되는 것을 확인하였다.

결과적으로 행정동으로 구분하여 정류장들을 집계한 결과보다 DB-SCAN이나 *k*-means를 통해 군집한 결과가 더 공간적인 인접성을 고려하는 것을 확인할 수 있었다. DB-SCAN과 *k*-means를 비교해보면, DB-SCAN과 *k*-means 모두 군집에 속한 평균 정류장 개수가 약 10개로 동일하게 나타났고 군집 안에 속한 정류장들의 최대개수는 DB-SCAN이 89개, *k*-means는 39개로 나타났다. 이는 DB-SCAN 방식이 밀도를 기반으로 군집을 확장해나가기 때문인 것으로 판단된다.

4. 결론

본 연구에서는 두 가지 군집 알고리즘을 이용하여 대중교통 정류장들을 군집화하였다. 군집화 기준은 정류장간의 공간적 인접성을 이용하였고, 각 알고리즘의 파라미터를 조정해가며 SSE가 최소가 되는 결과를 분석하였다. 공간적 응집도를 기준으로 두 가지 알고리즘을 비교해본 결과, *k*-means는 DB-SCAN에 비해 높은 공간적 인접성을 보였다. 다만, *k*-means는 최적의 군집 개수를 추가적으로 설정해야 한다는 한계가 있었다. DB-SCAN은 군집의 개수를 결정하지 않아도 된다는 장점이 있지만 어떠한 군집에도 속하지 않는 이상치에 해당하는 정류장이 산출된다는 한계가 있었다.

본 연구에서는 공간적 인접성만을 군집화 기준으로 설정하였는데, 대중교통 통행량, 토지이용패턴, 명칭유사도 등의 추가적인 변수들을 고려하지 못한 한계가 있다. 따라서 추후에는 이와 같은 변수들을 고려한 군집 분석을 진행할 계획이다. 이를 통해 미시적으로 대중교통의 이용패턴, 접근성 등을 분석하기에 적합한 군집화 알고리즘의 개발이 가능할 것으로 사료된다.

참고문헌

[1] Luo, Ding, Oded Cats, and Hans van Lint. Constructing Transit Origin-Destination Matrices Using Spatial Clustering. No. 17-01552. 2017.

[2] Du, Bowen, Yang Yang, and Weifeng Lv. " Understand group travel behaviors in an urban area using mobility pattern mining." Ubiquitous intelligence and computing, 10th IEEE UIC/ATC, 2013.

[3] 오규협, “시계열 빅데이터 분류 및 클러스터링 기법과 응용”, 경희대학교, 박사학위논문, 2017