한국지형공간정보학회
(Journal of the Korean Society for Geospatial Information Science)
Vol.25 No.4 December 2017 pp.97-106

ISSN: 1598-2955 (Print)
ISSN: 2287-6693(Online)
http://dx.doi.org/10.7319/kogsis.2017.25.4.097

연구논문

개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화 기법

Clustering Public Transit Stops using an Improved DBSCAN Algorithm

이민혁* · 전인우 · 전철민*****

Lee, Minhyuck · Jeon, Inwoo · Jun, Chulmin

要 旨

대중교통 접근성 분석을 위한 O-D 매트릭스 구축 시, 행정구역 단위의 거시적 교통존이 활용되고 있다. 본 연구는 보다 미시적인 교통존을 형성하기 위한 방법으로, 개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화 기법을 제안한다. 기존 DBSCAN 알고리즘은 서로 다른 두 정류장 간의 이웃관계를 판단할 때, 두 정류장 간의 거리만을 고려한다. 본 연구에서 개발한 개선된 DBSCAN 알고리즘은 거리뿐만 아니라 두 정류장 간의 명칭 유사도를 고려하여 이웃관계를 판단한다. 개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화 기법은 2단계로 구성된다. 첫 번째 단계에서는 3개 이상의 정류장을 포함하는 군집을 생성하고 두 번째 단계에서는 군집에 포함되지 않은 나머지 정류장들을 분류한다. 본 연구에서는 서울시 대중교통 정류장에 제안하는 군집화 기법을 적용하여 결과를 분석하였고, 기존 DBSCAN 알고리즘만을 이용한 군집 결과와 비교·분석하였다.

핵심용어 : 대중교통, 미시적 교통존, 정류장 군집화, DBSCAN

Abstract

This study proposes a method to cluster public transit stops using an improved DBSCAN algorithm to build microscopic traffic zones. The classic DBSCAN algorithm considers only the distance between two stops when determining the neighbor relationship. The proposed DBSCAN algorithm determines the neighbor relationship taking into account not only the distance but also the text similarity between two stops. The clustering processes using the proposed DBSCAN algorithm consist of two steps. The first step is a process of creating clusters of three or more stops. And the second step classifies the remaining stops that are not included in the first clustering step. This study applied the proposed clustering method to the transit stops in Seoul City and analyzed the results. And compared it with the clustering results obtained using only the classic DBSCAN algorithm.

Keywords : Public Transit, Microscopic Traffic Zone, Clustering of Transit Stops, DBSCAN

1. 서 론

일반적으로 접근성은 출발 지역으로부터 도착 지역이나 목표 시설에 접근하기 용이한 정도를 의미한다. 대중교통 분야에서 접근성은 운행 노선의 종류, 통행시간, 배차 간격 등의 서비스적 측면과 대중교통 시설까지의 근접성을 종합적으로 고려하여 분석하고 있다(Kim and Park, 2015). 대중교통의 공공재적 성격으로 인해 대중교통 접근성은 지역 간 형평성을 비교하기 위한 지표로 활용되고 있다.

개인의 정류장 승·하차 기록이 저장되는 교통카드 사용이 일반화되면서부터는 정류장 단위의 세밀한 접근성 분석이 가능하게 되었다(Park and Lee, 2015; Choi et al., 2016). 교통카드 데이터베이스에는 승객의 승·하차 정류장과 환승, 이용노선 및 시간 등에 관한 상세한 데이터가 저장되어 있다. 연구자 및 실무자들은 이전보

Received: 2017.11.24, revised: 2017.12.08, accepted: 2017.12.14

* 정회원 · 서울시립대학교 공간정보공학과 박사과정(Member, Ph. D. Student, Dept. of Geoinformatics, University of Seoul, lmhll123@uos.ac.kr)

** 서울시립대학교 공간정보공학과 학·석사과정(Undergraduate student, Dept. of Geoinformatics, University of Seoul, yugo123@uos.ac.kr)

*** 교신저자 · 정회원 · 서울시립대학교 공간정보공학과 교수(Corresponding Author, Member, Professor, Dept. of Geoinformatics, University of Seoul, cmjun@uos.ac.kr)

다 더 정확한 대중교통 수요를 손쉽게 획득할 수 있고 이를 통해 거시적 관점의 장기적인 대중교통 계획 수립뿐만 아니라 개별 정류장 단위의 미시적인 분석도 가능하다.

다만, 대중교통 이용객 측면의 접근성은 이용객 주변에 위치한 다수의 정류장들을 고려해야할 필요가 있다. 이용객들은 목적지에 도달하기 위해 대중교통 정보를 검색할 때, 하나의 정류장만을 이용하지 않는다. 주변에 위치한 다수의 정류장을 이용해 목적지로 갈 수 있는 다양한 방법들 중 최적의 방법을 선택한다. 즉, 대중교통 이용은 단일 정류장이 아닌, 이용객의 위치를 중심으로 근접한 다수의 정류장을 통해 이루어지는 것이다. 따라서 대중교통 이용객 측면의 접근성은 정류장 단위 분석보다는 보행편의성을 고려한 일정 구역 내 정류장 군집을 이용한 분석이 수행되어야 한다.

이에 본 연구에서는 대중교통 접근성 분석을 위해 근접한 다수의 정류장을 군집화하는 기법을 제안한다. 하나의 정류장 군집은 미시적 교통존을 의미하고 군집 내 정류장들은 특정 지역으로의 접근성을 공유한다. 그리고 이러한 정류장 군집은 대중교통 접근성 분석을 위한 origin-destination(O-D) 매트릭스 구축에 활용할 수 있다.

본 연구에서는 정류장 군집을 형성하기 위해 density-based spatial clustering application with noise(DBSCAN) 알고리즘을 개선하여 적용한다. DBSCAN 알고리즘은 임의의 데이터로부터 일정 반경 범위 내에 특정 개수 이상의 데이터가 존재할 경우, 군집을 형성하게 된다(Ester et al., 1996). 본 연구에서 개발한 개선된 DBSCAN 알고리즘은 정류장들 간의 거리뿐만 아니라 명칭 유사도를 추가적으로 고려하여 군집을 형성한다.

정류장 군집화는 3개 이상의 정류장을 포함하는 군집을 생성하는 과정과 군집에 포함되지 않은 나머지 정류장들을 분류하는 과정으로 구성된다. 정류장 군집화에 대한 보다 자세한 사항은 3장에 기술하였으며 4장에는 본 연구에서 제안하는 군집화 기법을 서울시 대중교통 정류장에 적용하여 분석한 결과에 대해 기술하였다.

2. 관련 연구 분석

대중교통 접근성 분석을 위한 O-D 매트릭스 구축 시, 일반적으로는 행정구역 단위의 교통존이 활용되고 있고 격자와 같은 임의의 공간 단위가 활용된 바 있다(Oh, 2017). 교통카드를 이용한 자동요금징수시스템이 국제적으로 확대된 이후에는 미시적 교통존 구성을 위한 정류장 군집화 연구들이 수행되고 있다. 대표적으로 활용되는 군집 알고리즘으로는 k-means(Hartigan and Wong, 1979)와 DBSCAN이 있다.

k-means는 분할 기반 군집 알고리즘 중 하나로, 사용자가 입력한 k개수만큼 군집을 형성한다. 특정 군집에 포함된 데이터들로부터 해당 군집 중심까지의 거리 합을 squared error(SE)라 하고 모든 군집의 SE 합을 sum of squared error(SSE)라 할 때, SSE가 최소화되도록 군집화가 진행된다. Luo et al.(2017)는 k-means 알고리즘을 이용하여 공간적 응집도와 내부 통행 비율이 높은 정류장 군집화 방법론을 제안하였다. 정류장들 간의 높은 근접성을 확보하면서 내부 통행이 활발한 교통존을 구성하기 위함이었다.

DBSCAN은 밀도 기반 군집화 방식으로, 사용자는 임의의 데이터를 중심으로 주변 공간을 탐색하기 위한 반경(eps)과 군집으로 인정하기 위한 반경 내 최소 데이터 개수(minPts)를 입력한다. eps 내에 minPts 이상의 데이터가 존재하면 군집을 형성하고 이웃 데이터를 중심으로도 동일한 검사를 실시하여 군집을 확장해나간다. Kieu et al.(2015)는 정류장 단위의 승·하차 데이터로부터 통행패턴을 군집화하기 위해 기존 DBSCAN의 연산복잡도를 감소시킨 weighted stop DBSCAN(WS-DBSCAN) 알고리즘을 개발하였다.

대중교통 정류장 군집화 연구에서는 연구자의 군집 목적에 따라 서로 다른 알고리즘이 활용되고 동일한 알고리즘이라 하더라도, 데이터셋의 분포에 따라 입력 파라미터가 다양하게 나타난다. Stop aggregation model에서는 정류장 군집화에 고려해야할 요소로 정류장들 간의 거리뿐만 아니라 명칭 유사도, 동일한 대중교통이용권역 여부 등을 언급하였다(Lee et al., 2012).

근접한 정류장들을 이용하여 하나의 군집을 형성한다는 목적만 본다면, Fig. 1과 같이 DBSCAN이 k-means에 비해 군집 성능이 뛰어나다(Tran et al., 2013). 하지만 DBSCAN은 유사한 명칭을 가지는 정류장이 주변에 존재하더라도 그 거리 차이가 eps를 초과하게 되면

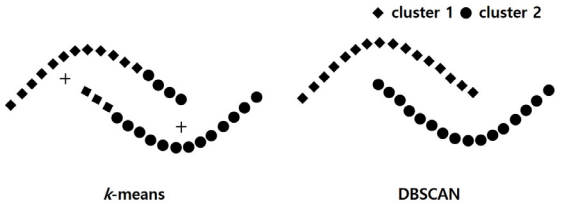

Figure 1. Comparison of result between k-means and DBSCAN

군집에 포함시키지 않는다. 이에 본 연구에서는 정류장들 간의 근접성뿐만 아니라 명칭 유사도도 고려할 수 있도록 개선된 DBSCAN 알고리즘을 개발하였고 이를 이용하여 대중교통 정류장 군집화를 수행하였다.

3. 개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화

3.1 개요

본 연구의 목적은 하나의 정류장처럼 대표할 수 있는 근접한 다수의 정류장을 군집화하는 것이다. 다수의 정류장이 하나의 정류장처럼 고려되기 위해서는 근접성과 더불어 유사한 명칭을 가져야 한다. 또한 지하철역이 포함되어 환승통행이 발생하지 않는 이상, 군집 내부 정류장들 간의 통행은 없는 것이 이상적이다. 그리고 이와 같이 정류장 군집을 형성하는 목적은 대중교통 접근성 분석을 위한 O-D 매트릭스 구축 시, 최소 공간단위의 교통존으로써 역할하기 위함이다. 본 연구에서는 미시적 교통존 형성을 위해 앞서 언급한 바와 같이 명칭 유사도를 고려할 수 있는 개선된 DBSCAN 알고리즘을 개발하였고 이를 이용한 정류장 군집화 방법론을 모색하였다.

3.2 개선된 DBSCAN 알고리즘

3.2.1 DBSCAN 알고리즘

DBSCAN 알고리즘은 eps와 minPts를 입력하면 다음과 같은 단계를 진행하여 군집을 형성한다. Fig. 2는 DBSCAN 알고리즘의 군집 형성 과정을 나타낸 것이다.

Step 1. 데이터셋 중 임의의 데이터를 방문하여 eps 내 모든 이웃 데이터를 검색한다.

Step 2. eps 내 데이터 개수가 minPts 이상인 경우, 군집을 형성하고 eps 내 모든 데이터들을 군집에 포함시킨다.

Step 3. 군집에 포함된 이웃 데이터를 방문하여 eps 범위 내에 minPts 이상의 데이터가 존재하는지 검사한다. 만족할 경우, 해당 데이터들을 현재 군집에 포함시킨다.

Step 4. Step 3를 더 이상 군집에 포함되는 데이터가 나오지 않을 때까지 반복하고 현재 군집의 확장이 종료되었다면, 데이터셋의 방문하지 않은 데이터에 대해서도 Step 2를 진행한다. 만약 minPts 조건을 만족하지 않는다면, 해당 데이터는 노이즈로 판단하며 방문하지 않은 다른 데이터에 대해서 Step 2를 진행한다.

Step 5. 더 이상 방문할 데이터가 없다면 알고리즘을 종료한다.

3.2.2 개선된 DBSCAN의 이웃 검색

기존의 DBSCAN 알고리즘은 임의의 데이터로부터 이웃 데이터를 검색할 때, eps 범위, 즉, 거리만을 고려한다(Step 1). 개선된 DBSCAN 알고리즘은 기존 이웃 검색 조건에 최대 허용 반경(maxEps)과 명칭 유사도를 추가적으로 고려한다. Table 1의 명명법을 이용하여 기존 DBSCAN과 개선된 DBSCAN의 이웃 검색 과정을 수도코드로 나타내보면 Table 2와 같다. Fig. 3은 개선된 DBSCAN의 이웃 검색 과정을 시각적으로 나타낸 것이다. eps 범위 내에 존재하는 데이터들을 이웃 데이터셋에 추가하는 것은 동일하고, maxEps 내의 데이터들과 명칭을 비교한 뒤, 유사할 경우 이웃 데이터셋에 추가하는 점이 차별화된다. 단순히 eps의 크기를 늘리게 되면 군집이 과도하게 커질 가능성이 존재한다. 이에 본 연구에서는 maxEps를 두어 해당 범위에서 명칭이 유사한 데이터들만 이웃 데이터셋에 추가하여 군집의 크기가 과도하게 커지는 상황을 방지하고자 하였다.

Table 1. Nomenclature

Variable	Description
D	Entire dataset
$d(i,j)$	Distance of data i and j
N	Neighborhood dataset
$maxEps$	Maximum allowable radius
t_i	Name of data i

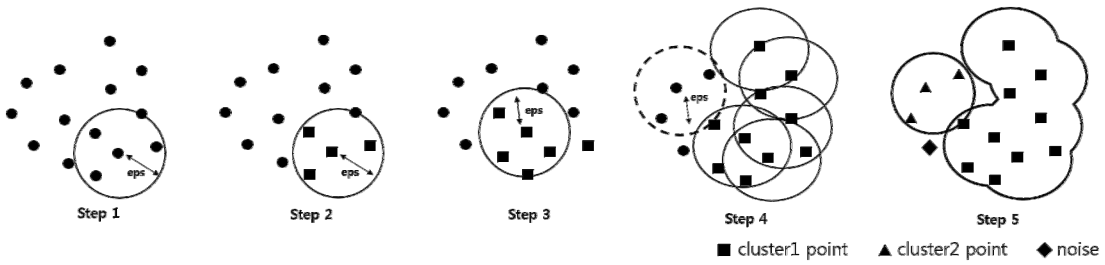

Figure 2. Process of DBSCAN algorithm

Table 2. Difference of retrieving neighborhood between DBSCAN and Improved DBSCAN

Retrieving neighborhood of DBSCAN
Arbitrary select a data $i \in D$ for $\forall j \in D$ If $d(i,j) \le eps$ then add j to N
Retrieving neighborhood of Improved DBSCAN
Arbitrary select a data $i \in D$ for $\forall j \in D$ If $d(i,j) \le eps$ then add j to N If $eps < d(i,j) \le maxEps$ then If t_i and t_j similar? then add j to N

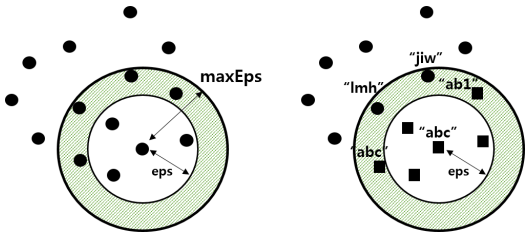

Figure 3. Retrieving neighborhood of Improved DBSCAN

3.2.3 명칭 유사도

두 데이터 간의 명칭 유사도는 편집거리 알고리즘(Yujian and Bo, 2007)을 이용한 부분과 포함관계를 이용한 부분으로 나뉜다. 임의로 선택된 데이터 i와 maxEps 내 데이터 j에 대하여 $t_i \supset t_j$ 혹은 $t_i \subset t_j$, 즉 두 명칭 간의 포함관계가 존재한다면, 명칭이 유사한 것으로 판단하고 이웃 데이터셋(N)에 j를 포함시킨다. “청량리역”과 “청량리역3번출구” 같은 정류장이 명칭 유사도에서 포함관계에 해당한다.

편집거리 알고리즘은 주어진 두 개의 단어에 대하여 두 단어가 동일해지려면 몇 번의 편집(삽입, 삭제, 변경 등)이 필요한지 계산하는 알고리즘이다. 본 연구에서는 음절 단위 비교를 하였고 편집 횟수가 적을수록 두 단어는 유사하다고 볼 수 있다. “청량리역”은 “청량리역3번출구”와 동일해지기 위해서 “3”, “번”, “출”, “구” 4번의 삽입이 필요하다. 따라서 편집거리는 4가 된다.

편집거리 알고리즘은 단어의 길이가 길어질수록 유사도를 판별하기 어려운 한계가 있다. “청량리”와 “서울역”의 편집거리는 3이다. 두 단어 사이에 동일한 음절이 없음에도 불구하고 앞서 살펴본 “청량리역”, “청량리역3번출구”보다 높은 유사도를 보이는 것처럼 나타난다. 이에 본 연구에서는 편집거리를 비교하는 단어의 최대 길이로 나누어 0과 1사이의 값으로 정규화하여 유사도를 비교하였다. 비교하는 두 정류장(i,j) 명칭 간의 편집거리를 $lev(t_i,t_j)$라 하고, 각 명칭의 길이를 $l(t_i), l(t_j)$라 할 때, 명칭 유사도 $s_{i,j}$는 Eq. (1)과 같다. 따라서 “청량리”, “서울역”의 유사도는 0, “청량리역”, “청량리역3번출구”의 유사도는 0.5가 산출된다. 즉, 유사한 명칭일수록 1에 가까운 값이 나오게 된다. 본 연구에서는 명칭 유사도 값이 0.6 이상일 경우 서로 명칭이 유사한 정류장으로 판별하였다.

$$s_{i,j} = 1 - \frac{lev(t_i,t_j)}{max\{l(t_i),l(t_j)\}} \quad (1)$$

3.3 대중교통 정류장 군집화 과정

3.3.1 알고리즘 파라미터 설정

개선된 DBSCAN 알고리즘을 이용한 정류장 군집화를 진행하기 위해서는 알고리즘의 파라미터를 우선적으로 설정해야 한다. 알고리즘의 입력 파라미터로는 군집을 형성하기 위한 최소 데이터 개수 minPts, 이웃 데이터를 검색할 반경 eps, 최대 허용 반경 maxEps 등이 있다.

DBSCAN 알고리즘에서 minPts는 3 이상이 활용된다. minPts가 1인 경우, eps 범위 내에 항상 데이터 자신을 포함하므로 모든 데이터가 군집이 된다. minPts가 2가 되면, eps 범위 내에 데이터 자신을 제외한 한 개의 이웃 데이터만 존재해도 군집을 형성하므로, 군집의 크기가 과도하게 커질 가능성이 존재한다. 이에 개선된 DBSCAN의 입력 minPts는 3으로 설정하였다.

정류장 군집은 eps와 maxEps에 따라 다양하게 나타난다. 이에 eps와 maxEps의 범위(최소~최대)를 설정하고 일정한 간격으로 두 값을 증가시켜가며 군집의 결과를 살펴보았고 그 중 연구 목적에 가장 합당한 군집 결과를 선정하였다. eps와 maxEps의 범위는 k-최근접 이웃 알고리즘을 통해 결정하였다. k-최근접 이웃 알고리즘은 모든 데이터에 대하여 K번째 가까운 이웃 데이터와의 거리를 계산한다. 본 연구에서는 전체 정류장에 대하여 가장 가까운 첫 번째 이웃 정류장과의 거리를 계산하였다. 산출된 결과를 바탕으로 최근접 정류장과의 평균거리를 eps의 시작점, 최근접 정류장과의 거리가 급격하게 증가하는 변곡점에서의 거리를 maxEps의 시작점으로 설정하였다. eps와 maxEps의 증가 단위는 10m로 하였고 eps는 maxEps의 최소값 전까지, maxEps는 250m까지 증가시켰다.

3.3.2 정류장 군집화

Fig. 4는 개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화의 전체 과정을 나타낸 것이다. 우선, 전체 정류장에 k-최근접 이웃 알고리즘을 적용하여 eps와 maxEps의 범위를 계산한다. 이후, eps와 maxEps의 모든 조합에 대하여 군집 결과를 생성하는데 군집 결과를 생성하는 내부 과정은 2단계로 이루어진다.

특정 eps와 maxEps가 주어졌을 때, 첫 번째 단계에서는 minPts를 3으로 하고 개선된 DBSCAN 알고리즘을 수행한다. 이는 군집 내 정류장 개수가 3개 이상인 제1군집(lv1 cluster) 유형을 도출하기 위함이다. 그리고 제1군집 유형에 속하지 않은 나머지 정류장(lv1 noise)들은 minPts를 2로 한 개선된 DBSCAN 알고리즘의 인풋 데이터가 된다. 2번째 알고리즘이 종료되면, 2개의 정류장 쌍(pair)으로만 구성된 제2군집(lv2 cluster) 유형과 어느 군집에도 포함되지 않는 정류장(noise)이 분류된다.

결과적으로 eps가 최소값부터 최대값까지 총 n개 존재하고, maxEps가 최소값부터 최대값까지 총 n'개 존재한다면, 군집 결과는 $n \times n'$개 산출된다. 그 중 대중교통 접근성 분석을 위한 미시적 교통존으로 가장 적당하다고 판단되는 군집 결과를 최적 군집으로 선정하게 된다. 선정 기준으로는 군집 및 노이즈 개수, 군집의 공간적 응집도를 나타내는 군집 내 정류장 간의 평균/최대 거리, 군집의 크기를 반영하는 군집 내 평균/최대 정류장 개수 등을 고려하였다.

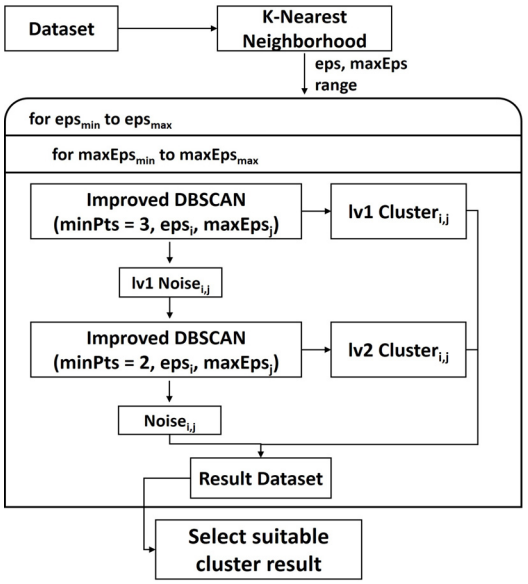

Figure 4. Process of clustering public transit stops using an improved DBSCAN

4. 실험 및 결과 분석

4.1 실험 범위 및 정류장 데이터셋

본 연구에서는 제안하는 군집화 기법을 서울시 대중교통 정류장에 적용해보았다. Fig. 5는 서울 전역에 분포하고 있는 대중교통 정류장을 나타낸 것이다. 서울시에는 만여 개가 넘는 대중교통 정류장이 존재한다. 본 연구에서는 승객의 승·하차가 이루어지지 않는 가상 정류장을 제외하고 총 11,131개의 대중교통 정류장에 대하여 군집화를 수행하였다. 이 중 352개는 도시철도역에 해당하며 나머지 10,779개의 정류장은 버스 정류장에 해당한다.

Fig. 6은 서울시 교통정보센터(topis)에서 제공받은 대중교통 정류장 데이터의 일부 속성값을 나타낸 것이다. 정류장ID(NodeID), 정류장 유형(NodeType), 정류장 명칭(NodeName), 정류장 좌표(X,Y) 등이 포함되어 있다. 정류장 유형은 1과 2로 나뉘는데, 1은 도시철도

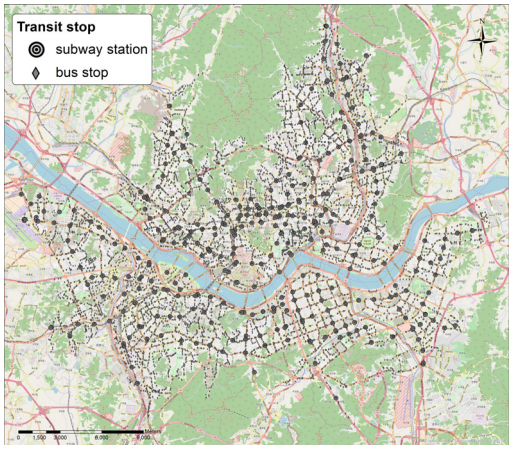

Figure 5. Transit stops in Seoul city

NodeID	NodeType	NodeName	X	Y
9008696	2	강남교육청	960006.834	1945891.678
9009708	2	강남교회.오거리	950936.797	1945814.875
9009809	2	강남교회.오거리	950943.874	1945815.943
9013909	2	강남구민체육관	960554.693	1941956.029
1849	1	강남구청	959459.369	1946529.419
2732	1	강남구청	959460.537	1946531.648
72908	2	강남구청.강남세무서	959994.658	1946667.252
72931	2	강남구청.강남세무서	960027.258	1946647.125
8046	2	강남구청역	959345.226	1946506.201
8049	2	강남구청역	959364.57	1946486.136
70496	2	강남구청역	959409.665	1946672.306
71096	2	강남구청역	959433.552	1946677.737

Figure 6. Structure of transit stop data

역, 2는 버스 정류장을 의미한다. 정류장 명칭은 “강남구청.강남세무서”와 같이 2개 이상의 장소를 나타내는 경우가 있다. 이와 같은 경우, 정류장 간의 명칭 비교 시, “강남구청”, “강남세무서”로 나누어 각각 비교하고 높은 명칭 유사도 값을 두 정류장 간의 명칭 유사도로 결정하였다. “강남구청” 도시철도역은 7호선(NodeID: 2732)과 분당선(NodeID: 1849)이 존재한다. 이처럼 서로 다른 지하철 호선으로 구분되어 있지만 동일한 지하철역인 경우에는 같은 군집에 속하도록 전처리 과정을 수행하였다.

4.2 정류장 군집화 결과

4.2.1 eps 및 maxEps 범위 설정

정류장 군집화를 위해 우선, 서울시 대중교통 정류장들에 대하여 최근접 정류장까지의 거리를 계산하였다. Fig. 7은 각 정류장에서 최근접 정류장까지의 거리를 계산한 뒤, 거리 값을 오름차순으로 정렬하여 나타낸 그래프이다. 평균 거리값은 약 50m로 나타났고 약 180m 지점부터 거리가 급격하게 증가하는 것을 확인할 수 있었다. 이에 평균거리 값인 50m를 eps의 최소값으로 선정하였고 변곡점에서의 거리 값인 180m를 maxEps의 최소값으로 선정하였다. 그리고 앞서 언급한 바와 같이 두 값을 10m 씩 증가시켜 가며 군집 결과를 살펴보았다. eps는 50~170m까지 13단계, maxEps는 170~250m까지 9단계로 구분되어 군집 결과는 총 117개가 산출되었다.

4.2.2 군집 결과 분석 및 최적 군집 선정

Fig. 8은 총 117개 군집 결과에 대하여 제1군집 유형 개수(a), 제2군집 유형 개수(b), 노이즈 개수(c), 평균 정류장 개수(d), 정류장 간 평균 거리(e), 평균 명칭 유사도(f), 최대 정류장 개수(g), 정류장 간 최대 거리(h) 등을 3차원 그래프로 시각화한 것이다. eps의 범위 50~170m는 0~12로 표기하였고 maxEps의 범위 180~250m는 0~7로 표기하였다. 평균 정류장 개수(d)부터 정류장 간 최대 거리(h)까지는 제1군집 유형에 대해 분석한 결과이다. 제2군집 유형의 경우, 2개의 정류장으로만 구성되어 있기 때문에 평균 정류장 개수, 정류장 간 평균 거리, 명칭 유사도 등의 결과를 과도하게 향상시킬 수 있어 배제하였다.

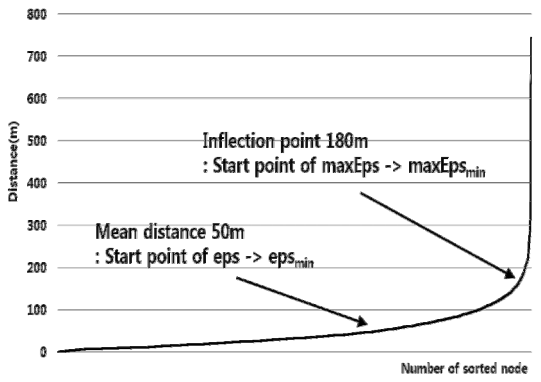

Figure 7. *k*-dist plot (k=1)

Fig. 8(a), 제1군집 유형 개수는 maxEps가 0, eps가 6일 때 1,295개로 최대값이 나타났고 이후부터는 감소하는 패턴을 보였다. Fig. 8(b)부터 Fig. 8(f)까지는 예측 가능한 결과가 나타났다. eps가 커질수록 군집을 형성하기 쉬워지기 때문에 제2군집 유형과 노이즈가 줄어드는 것을 확인할 수 있었다. 또한 군집의 크기가 커질수록 군집을 구성하는 평균 정류장 개수와 정류장들 간의 평균 거리가 증가하는 것을 확인하였고, 명칭 유사도 역시 군집 내 많은 정류장들이 포함되므로 감소하는 패턴이 나타났다.

Fig. 8(g)과 Fig. 8(h)는 eps 및 maxEps의 변화에 따라 결과 값이 민감하게 나타났다. 최대 정류장 개수(g)는 maxEps가 0에서 1로 증가하는 시점, 4에서 5로 증가하는 시점에 급격하게 증가하는 결과가 나타났다. 또한 eps가 5에서 6으로 증가하는 시점과 10 이후부터도 급격하게 증가하는 결과를 확인할 수 있었다.

정류장 간 최대 거리(h)도 최대 정류장 개수 변화와 유사한 패턴을 보였다. maxEps가 0에서 1로 증가하는 시점에 1km 가까이 증가하는 결과를 보였고, 4에서 5로 증가하는 시점에는 약 1.5km 증가하는 결과가 나타났다. eps를 기준으로 보면, maxEps가 0일 때는, 5~6, 7~8, 9~10, 11 이후 등에서 정류장 간 최대 거리가 급격하게 증가하는 모습을 확인할 수 있었다. 또한 maxEps가 1~5인 구간에서는 eps가 10인 시점에서 약 1.5km의 거리 증가가 확인되었다.

eps와 maxEps의 증가는 군집 형성을 용이하게 하는 동시에 군집의 확장을 가져온다. 군집의 확장이 과도하게 이루어지면 미시적 교통존으로써의 역할을 기대하기 어렵다. 평균 정류장 개수와 정류장 간 평균 거리는 군집의 크기를 대변하는 지표라 할 수 있다. 평균 정류장 개수 및 정류장 간 평균 거리가 급격하게 증가하기 전 시점인, eps 0~5 구간에서는 모든 maxEps에 대하여 평균 정류장 개수가 약 5개, 정류장 간 평균 거리가 약 110m로 나타난다. 제1군집 유형 개수는 eps가 6인 시점까지 꾸준히 증가하고 노이즈도 해당 시점까지 꾸준히 감소한다.

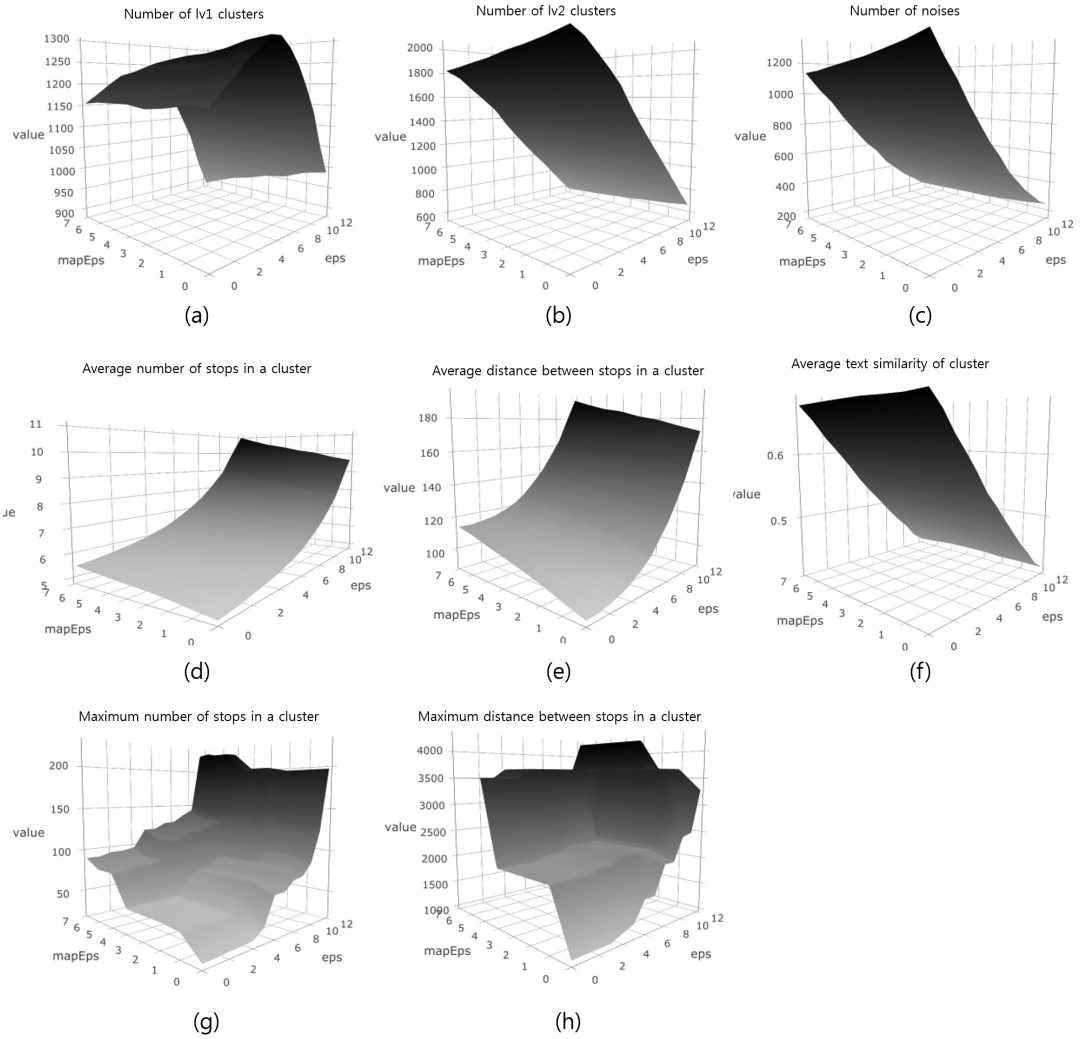

Figure 8. Three-dimensional graph of each indicator for cluster results

종합해보면, maxEps가 특정 값으로 고정되어 있을 때, eps의 0~5 구간은 평균적으로 유사한 군집 크기를 가지고 이 구간에서 제1군집 유형 개수와 노이즈 개수는 eps가 5일 때 최적이다. maxEps의 경우에는 최대 정류장 개수 및 정류장 간 최대 거리를 통해 확인할 수 있듯이, 값이 1만 증가해도 최대 군집의 크기를 급격하게 키우는 것을 알 수 있다. 이에 최적 군집은 eps와 maxEps가 각각 5, 0일 때, 즉, eps는 100m, maxEps는 180m일 때로 선정하였다.

4.2.3 최적 군집 분석

Table 3는 Fig. 8의 각 지표들에 대한 최적 군집의 결과에 군집별 내부통행비율의 평균값을 추가한 것이

Table 3. Result of optimal cluster

Indicator	Value
Number of lv1 clusters	1,294
Number of lv2 clusters	1,536
Number of noises	748
Average number of stops in a cluster	5.65
Average distance between stops in a cluster	101.44m
Average text similarity	0.56
Maximum number of stops in a cluster	35
Maximum distance between stops in a cluster	1,203m
Average intra flow rate	0.003

Table 4. Classification according to the number of stops in a cluster

Number of individual stops in group	Number of groups	%
3-5	848	66
6-10	335	26
11-15	70	5
16-20	23	2
21-25	13	1
26-30	3	0
31-35	2	0
Total	1,294	100

Table 5. Clustering result when classic DBSCAN algorithm applied

Indicator	Value
Number of lv1 clusters	1,317
Number of lv2 clusters	1,878
Number of noises	1,239
Average number of stops in a cluster	4.66
Average distance between stops in a cluster	71.89m
Average text similarity	0.55
Maximum number of stops in a cluster	24
Maximum distance between stops in a cluster	676m

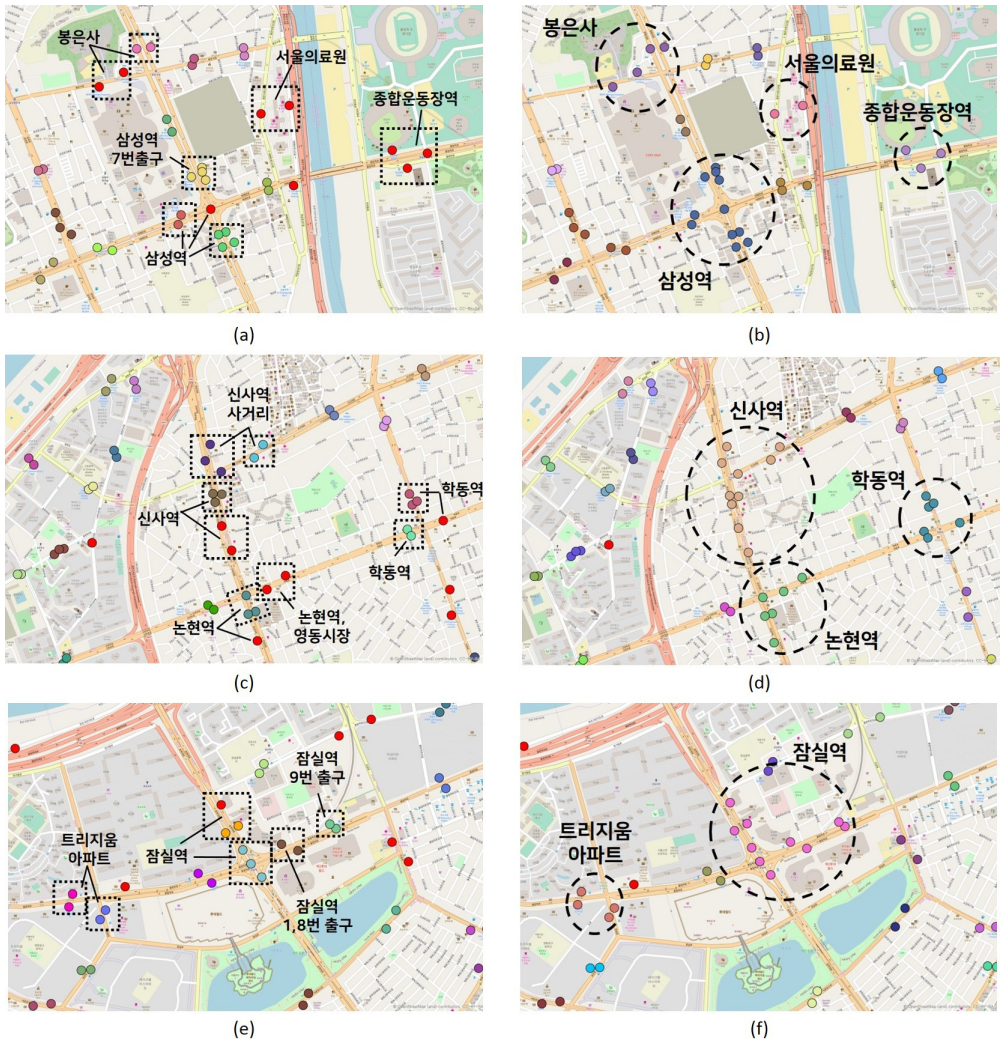

Figure 9. Comparison between case(a, c, e) applying DBSCAN and case(b, d, f) applying improved DBSCAN

다. 통행량은 2015년 10월 12일부터 18일까지 일주일간의 교통카드 데이터를 이용하여 계산하였다. 평균 내부통행비율은 0.003으로 0에 가까운 수치를 보였으며, 군집 내부에서 발생한 환승통행이 포함되어 있었다.

Table 4는 제1군집 유형을 정류장 개수에 따라 분류한 것이다. 제1군집 유형에는 총 1,294개의 군집이 존재한다. 그 중 정류장 개수가 3~5개인 군집은 약 66%, 6~10개인 군집은 약 26%에 해당하였다. 정류장 개수가 20개를 초과하는 군집은 전체 군집의 약 1%로 나타났다. 전반적으로 10개 이하의 정류장을 포함하는 군집이 전체 군집의 약 92%를 차지하였다.

Table 5는 본 연구에서 제안하는 군집화 기법에 개선된 DBSCAN 대신 기존 DBSCAN 알고리즘을 적용했을 때 나타나는 결과이다. 노이즈 개수는 약 2배 가까이 많지만 군집 내 평균/최대 정류장 개수, 정류장들 간의 평균/최대 거리 등의 지표만 보면, 기존 DBSCAN 알고리즘을 적용했을 때가 정류장들 간의 응집도가 높은 것으로 나타난다. Fig. 9는 일부 지역에 대하여 기존 DBSCAN을 적용한 군집 결과(a, c, e)와 개선된 DBSCAN을 적용한 군집 결과(b, d, f)를 시각적으로 비교한 것이다. 일부 정류장들에 대해서는 명칭을 표기하였고 붉은색 점은 노이즈, 하나의 군집은 모두 동일한 색으로 표현하였다.

기존 DBSCAN을 이용한 경우를 살펴보면, 비교적 근접한 거리에 위치해 있으면서 명칭도 유사한 정류장들이 서로 각기 다른 군집을 형성하고 있는 것을 알 수 있다. 특히, 도시철도역 주변으로 과도하게 많은 군집이 형성되었거나 노이즈가 존재하는 것을 파악할 수 있다. 또한 봉은사, 서울의료원, 트리지움 아파트 등 동일한 시설명을 공유하는 정류장들에 대해서도 서로 다른 군집 혹은 노이즈가 나타나고 있다.

반면에 개선된 DBSCAN을 이용한 경우를 살펴보면, 도시철도역을 중심으로 유사한 명칭을 가지는 정류장들이 하나의 군집으로 형성되어 있는 것을 알 수 있다. 동일한 시설명을 공유하는 정류장들 역시 하나의 군집을 형성하고 있다. 정류장 간의 거리가 eps를 초과하여 서로 다른 군집 혹은 노이즈로 분류될 정류장들이 maxEps 범위 안에서 유사한 명칭을 가졌기 때문에 동일한 군집을 이루게 된 것이다.

4.3 실험 결론

도시철도역을 중심으로 연계되는 대중교통 노선들을 고려할 때, 도시철도역이 포함된 교통존은 버스 정류장들로만 구성된 교통존에 비해 넓은 범위를 포괄해야 할 필요가 있다. 개선된 DBSCAN 알고리즘을 이용한 대중교통 정류장 군집화 기법은 정류장들 간의 거리뿐만 아니라 명칭 유사도를 추가적으로 고려하여 군집을 형성하기 때문에 위의 특성을 반영한, 보다 유연한 교통존을 구성할 수 있다. 이는 도시철도역에 국한되는 것이 아니라, point of interest(POI)를 중심으로 위치한 다수의 정류장들을 하나의 미시적 교통존으로 군집화할 수 있음을 의미한다. 따라서 대중교통 접근성 분석을 위한 미시적 교통존 구성 문제에 있어, 기존 DBSCAN 방식보다는 본 연구에서 제안하는 군집화 기법이 보다 적합할 것으로 판단된다.

5. 결 론

본 연구에서는 대중교통 접근성 분석을 위한 미시적 교통존을 형성하기 위해 근접한 다수의 정류장을 군집화하는 기법을 제안하였다. 군집화 기법은 3개 이상의 정류장을 포함하는 군집을 생성하는 과정과 군집에 포함되지 않은 나머지 정류장들을 분류하는 과정으로 구성되어 있다. 각각의 과정에는 본 연구에서 개발한 개선된 DBSCAN 알고리즘을 활용하였다.

기존의 DBSCAN 알고리즘은 사용자가 입력한 검색 반경만을 이용하여 이웃 데이터를 검색하고 해당되는 데이터들은 이웃 데이터셋에 추가한다. 개선된 DBSCAN 알고리즘은 기본 검색 반경을 초과하더라도 최대 허용 반경 이내에 존재하는 데이터들 중 명칭이 유사한 데이터들은 이웃 데이터셋에 추가한다. 즉, 데이터들 간의 거리 뿐만 아니라 명칭의 유사성도 고려하여 군집을 형성하는 것이다.

개선된 DBSCAN 알고리즘을 서울시 대중교통 정류장에 적용한 결과에서는 평균적으로 6개의 정류장이 약 100m 거리를 두고 군집화 되었으며, 총 1,294개의 제1군집이 나타났다. 기존 DBSCAN 알고리즘을 적용한 경우에는 평균적으로 5개의 정류장이 약 72m 거리를 두고 군집을 형성하였으며, 1,317개의 제1군집이 나타났다. 기존 알고리즘을 적용하는 것이 보다 응집력이 높은 군집 결과를 도출하는 것처럼 보이나, 개선된 DBSCAN 알고리즘을 적용할 때 노이즈 개수가 약 2배 감소하며, 유사 명칭을 가지며 근거리에 위치한 정류장들이 군집을 형성하는 보다 합리적인 결과가 도출되는 것을 확인할 수 있었다.

개선된 DBSCAN 알고리즘은 다양한 지표들을 살펴보긴 하였지만, 다소 정성적인 방법으로 파라미터를 결정한 한계가 있다. 따라서 향후연구에는 보다 정량적인 방법을 통해 선정한 파라미터를 바탕으로 정류장 군집화가 수행되어야 할 것으로 판단된다. 또한 군집 결과

를 이용하여 대중교통 접근성을 분석해보고 정책적 의의를 이끌어낼 수 있는 연구가 수행되어야 할 것이다. 이 외에 지오태깅된 SNS의 군집화와 같이 공간적 요소를 고려한 텍스트 마이닝 분야에도 알고리즘을 활용할 수 있도록 발전시킬 필요가 있다.

감사의 글

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

References

1. Choi, S. U., Jun, C. M. and Cho, S. K., 2016, Micro-scale Public Transport Accessibility by Stations – KTX Seoul Station Case Study -, Journal of the Korean Society for Geospatial Information Science, Vol. 24, No. 1, pp. 9-16.
2. Ester, M., Kriegel, H. P., Sander, J. and Xu, X., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd, Vol. 96, No. 34, pp. 226-231.
3. Hartigan, J. A. and Wong, M. A., 1979, Algorithm AS 136: A k-means clustering algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1, pp. 100-108.
4. Kieu, L. M., Bhaskar, A. and Chung, E., 2015, A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data, Transportation Research Part C: Emerging Technologies, Vol. 58, pp. 193-207.
5. Kim, D. H. and Park, D. J., 2015, A study on Customer-Oriented Measure Methodology of Public Transport Accessibility Under Time and Space Constraints, The 73rd Conference of Korean Society of Transportation, pp. 545-550.
6. Lee, S., Hickman, M. and Tong, D., 2012, Stop Aggregation Model: Development and Application, Transportation Research Record: Journal of the Transportation Research Board, No. 2276, pp. 38-47.
7. Luo, D., Cats, O. and van Lint, H., 2017, Constructing Transit Origin-Destination Matrices with Spatial Clustering, Transportation Research Record: Journal of the Transportation Research Board, No. 2652, pp. 39-49.
8. Oh, G. H., 2017, Techniques and Applications of Classification and Clustering of Big Time-Series Data, Ph. D. dissertation, Kyunghee University, pp. 16-18.
9. Park, J. S. and Lee, K. S., 2015, Time-distance Accessibility Computation of Seoul Bus System based on the T-card Transaction Big Databases, Journal of the Economic Geographical Society of Korea, Vol. 18, No. 4, pp. 539-555.
10. Tran, T. N., Drab, K. and Daszykowski, M., 2013, Revised DBSCAN algorithm to cluster data with dense adjacent clusters, Chemometrics and Intelligent Laboratory Systems, Vol. 120, pp. 92-96.
11. Yujian, L. and Bo, L., 2007, A normalized Levenshtein distance metric, IEEE transactions on pattern analysis and machine intelligence, Vol. 29, No. 6, pp. 1091-1095.