한국지형공간정보학회지
(Journal of the Korean Society for Geospatial Information System)
Vol.22 No.4 December 2014 pp.175-181

ISSN: 1598-2955 (Print)
ISSN: 2287-6693(Online)
http://dx.doi.org/10.7319/kogsis.2014.22.4.175

연구논문

유전자 알고리즘을 이용한 서울시 군집화 최적 변수 선정

Selection of Optimal Variables for Clustering of Seoul using Genetic Algorithm

김형진* · 정재훈** · 이정빈*** · 김상민**** · 허 준*****

Kim, Hyung Jin · Jung, Jae Hoon · Lee, Jung Bin · Kim, Sang Min · Heo, Joon

要 旨

정부 3.0이라는 새로운 정부운영 계획과 함께 다양한 공공정보를 민간이 활용할 수 있게 되었으며, 특히 서울은 이러한 행정정보 공개 및 활용을 선도하고 있다. 공개된 행정정보를 통해 각 지역을 특징짓는 행정요소를 발견할 경우, 각종 행정정책을 위한 의사결정 수단에 반영할 수 있을 뿐만 아니라 특정 지역의 고객 특성을 파악하여 특화된 서비스나 상품을 판매하는 마케팅 수단으로도 사용할 수 있을 것으로 사료된다. 하지만, 방대한 양의 행정자료로부터 각 군집의 특성을 명확히 구분할 수 있는 최적의 조합을 찾는 과정은 조합최적화 문제로서 상당한 연산량을 요구한다. 본 연구에서는 서울시에서 제공하는 다차원 행정자료로부터 서울시를 대표하는 문화·산업의 중심인 서초구, 강남구, 송파구 등의 강남 3구를 다른 지역과 효과적으로 구분하는 행정요인를 찾고자 하였다. 방대한 양의 행정정보로부터 두 군집간의 차이점을 극대화하는 요인을 선별하기 위한 최적화 방법으로 유전자 알고리즘을 이용하였으며, 군집간 차이를 계산하는 척도로는 Dunn 지수를 이용하였다. 또한 유전자 알고리즘의 연산속도의 향상을 위해 Microsoft Azure에서 제공하는 cloud computing을 이용한 분산처리를 수행하였다. 자료로는 통계청으로부터 취득한 총 718개의 행정자료를 이용하였으며, 그 중 28개가 최적 변수로 선정되었다. 검증을 위해 선정된 28개의 변수를 입력값으로 Ward의 최소분산법 및 K-means 알고리즘을 통한 군집화를 수행한 결과 두 경우 모두 강남 3구가 다른 지역으로부터 효과적으로 분류됨을 확인하였다.

핵심용어 : 군집화, Dunn 지수, Ward의 최소분산법, K-means 알고리즘, 유전자 알고리즘

Abstract

Korean government proposed a new initiative ‘government 3.0’ with which the administration will open its dataset to the public before requests. City of Seoul is the front runner in disclosure of government data. If we know what kind of attributes are governing factors for any given segmentation, these outcomes can be applied to real world problems of marketing and business strategy, and administrative decision makings. However, with respect to city of Seoul, selection of optimal variables from the open dataset up to several thousands of attributes would require a humongous amount of computation time because it might require a combinatorial optimization while maximizing dissimilarity measures between clusters. In this study, we acquired 718 attribute dataset from Statistics Korea and conducted an analysis to select the most suitable variables, which differentiate Gangnam from other districts, using the Genetic algorithm and Dunn’s index. Also, we utilized the Microsoft Azure cloud computing system to speed up the process time. As the result, the optimal 28 variables were finally selected, and the validation result showed that those 28 variables effectively group the Gangnam from other districts using the Ward’s minimum variance and K-means algorithm.

Keywords : Clustering, Dunn’s Index, Ward’s Minimum Variance, K-means Algorithm, Genetic Algorithm

Received: 2014.12.19, accepted: 2014.12.19

* 정회원 · 연세대학교 토목환경공학과 석사과정(Member, Department of Civil and Environmental Engineering, Yonsei University, awg1@yonsei.ac.kr)

** 정회원 · 연세대학교 토목환경공학과 박사후 연구원(Member, Department of Civil and Environmental Engineering, Yonsei University, lionheart_kr@yonsei.ac.kr)

*** 토목환경공학과 박사과정(Department of Civil and Environmental Engineering, Yonsei University, ortolan@yonsei.ac.kr)

**** 정회원 · 토목환경공학과 박사과정(Member, Department of Civil and Environmental Engineering, Yonsei University, netgo82@yonsei.ac.kr)

***** 교신저자 · 정회원 · 연세대학교 토목환경공학과 교수(Corresponding author, Member, Department of Civil and Environmental Engineering, Yonsei University, jheo@yonsei.ac.kr)

1. 서 론

가수 Psy의 대표곡인 '강남 스타일'의 성공은 전 세계에 한국의 명소인 강남 지역을 알리는 계기가 되었다. 일반적으로 우리나라에서 서울 강남 지역이라 함은 서초구, 강남구, 송파구를 가리키고 있다(Fig. 1). 이 지역에는 수많은 기업 사무실과 고급 주택가, 유명한 학군이 몰려있으며 노래 가사에서와 마찬가지로 서울을 대표하는 문화·산업의 중심지역이라 할 수 있다. 그러나 이러한 강남을 비롯하여 서울 내부의 다양한 지역을 구체적으로 특징짓는 속성들이 무엇인지에 대한 군집화 시도는 아직까지 연구가 진행된 바 없다.

우리나라는 최근 이슈가 되고 있는 데이터 주도 경제(data-driven economy)를 활성화시키기 위해 '정부 3.0'이라는 새로운 아젠다를 세우고, 행정기관들이 민간인들을 대상으로 특별한 공개청구 없이 행정정보를 자유롭게 열람할 수 있는 서비스를 제공하고 있다. 특히 서울시는 이미 수많은 행정정보를 대중들에게 공개해 왔으며, 공개되는 대부분의 데이터가 '구' 단위의 공간정보를 포함하고 있다. 이렇게 중요한 정보를 담고 있는 수많은 자료들을 통해 서울시군집화 문제에 대한 분석과 해결방안을 제시 할 수 있을 것으로 판단된다. 이러한 군집화 연구는 첫째, 군집화를 통해 각 지역을 특징짓는 행정요소를 발견, 의사결정 수단에 반영할 수 있으며, 둘째, 비즈니스 관점에서 고객 특성을 파악하여 특화된 서비스나 상품을 판매하는 마케팅 수단으로 이용될 수 있을 것으로 사료된다.

하지만, 서울시는 우리나라에서 가장 큰 도시답게 방대한 양의 행정정보를 포함하고 있다. 이러한 자료로부터 서울시의 특징을 분석하고 군집화하기 위해서는 우선 방대한 종류의 데이터들로부터 군집화 요인을 분류하고, 빠른 시간 내에 처리해야 하는 기술적인 문제가 발생한다. 예를 들어, 강남 3구를 다른 구들로부터 군집화한다고 가정했을 때, 수많은 행정정보로부터 군집화 차이점을 최대화 시킬 수 있는 요인들만을 선별해야 하는 과정은 조합최적화 문제(combinatorial optimization problem)에 해당하며, 일반적으로 자료의 종류가 증가할수록 막대한 양의 연산을 요구한다.

본 연구에서는 서울시를 대상으로 강남 3구가 이미 군집화되어 있다는 가정 하에 다른 구와 차이점을 나타내는 대표적인 요인이 무엇인지를 찾고자 하였다. 자료로는 통계청에서 제공한 718종류의 행정정보를 이용하였으며, 방대한 양의 행정정보로부터 강남 3구를 특징짓는 요인을 효과적으로 선별하기 위해 유전자 알고리즘을 이용한 방법을 제시하였으며, 군집간 차이를 분석

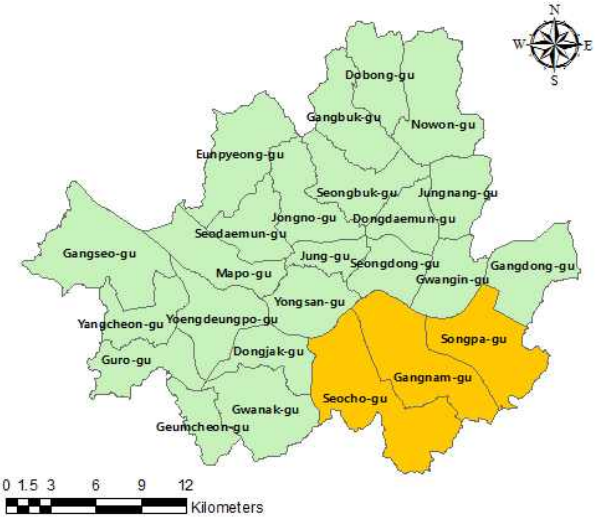

Figure 1. Twenty five districts (Gu) in Seoul and the Gangnam area in orange color

하여 적합성을 평가하기 위한 척도로는 Dunn 지수(Dunn's index)를 사용하였다. 또한 유전자 알고리즘 연산에 소요되는 시간 소요를 효과적으로 줄이기 위해 Microsoft에서 제공하는 Azure cloud computing을 이용한 분산처리를 시도하였다. 아울러, 유전자 알고리즘으로 선별된 요인들이 실제로 강남 3구를 효과적으로 구별하는 지를 검증하기 위해 기존의 군집화 알고리즘인 Ward의 최소분산법(Ward`s Minimum Variance)과 K-means 알고리즘을 이용하였다.

2. 연구내용

2.1 연구 자료

본 연구에서는 강남 3구를 군집화하기 위한 기초 자료로써 통계청에 존재하는 718가지 항목의 자료를 이용하였다. 다음의 Table 1은 통계청으로부터 취득한 자료들을 크게 9가지 항목으로 요약한 자료이며, 해당 자료는 국가통계포털(www.kosis.kr)을 통해 다운받을 수 있다.

2.2 Dunn 지수

다수의 변수에서 강남 3구를 특징짓는 최적화된 변수를 선정하기 위해서는 변수의 종류와 개수를 바꾸어 가면서 군집간의 차이점을 최대화하기 위한 적합성 평가(cluster validity measure) 과정이 필요하다. 군집간 특징을 판단하는 대표적인 척도로 Silhouette 지수와 Dunn 지수 등이 사용되고 있으며 (Kim et al., 2012), 본 연구에서는 그 중에서도 Dunn 지수를 사용하였다.

Dunn 지수는 Euclidean distance를 이용하여 나누어진 군집이 유효한지에 대한 것을 계산하는 index이다. 다음의 수식 (1)은 이러한 Dunn 지수를 나타내고 있다. 여기서 n은 군집의 총 개수를, $dist(x, y)$는 두 군집의 각 원소 간 거리를, $diam(c)$는 군집 내 가장 먼 거리를 이루는 원소 둘이 이루는 반경을, c_i는 i번째 군집을 나타낸다. 분모는 k번째 군집 내의 원소들 중에서 가장 먼 거리의 값이 들어가며, 분자는 i와 j번째 군집이 포함하는 원소들 중에서 최소 거리를 나타낸다. Dunn 지수의 정의에 따르면 Dunn 지수가 크면 클수록 군집화는 잘되었다고 평가할 수 있다(Bezdek and Nikhil, 1995).

Maximize:

$$D = \min_{i=1\ldots n}\left\{\min_{j=1\ldots n}\left(\frac{dist(c_i, c_j)}{\max_{k=1\ldots n}(diam(c_k))}\right)\right\} \quad (1)$$

Where

$$dist(c_i, c_j) = \min_{x\in c_i, y\in c_j}\{dist(x, y)\}$$

$$diam(c_i) = \max_{x, y\in c_j}\{dist(x, y)\}$$

한편 d를 속성 정보의 종류(dimension)라 할 때, 속성 숫자의 증가에 따라 계산량은 $\sum_{k=1}^{d-1} {}_dC_k$로 증가하게 되며, 이는 약 2의 d제곱에 해당한다. 즉, 현 속성 정보의 종류가 718개이므로 전수조사를 위해서는 약 2718번이라는 엄청난 숫자의 Dunn 지수 연산과 각각의 값들을 비교하여 최대값을 찾는 과정이 필요하며, 실질적으로 전수조사는 불가능하다. 이와 같이 변수가 증가할 때 그 연산량이 극단적으로 증가하는 경우를 조합최적화 문제라 하며, traveling salesman problem 등이 대표적인 사례이다 (Rademacher, 2005). 이러한 경우는 전수조사가 아닌 휴리스틱(heuristic) 기법을 통해 문제를 해결하며, 본 연구에서는 유전자 알고리즘을 이용해 문제 해결을 시도하였다.

2.3 유전자 알고리즘

유전자 알고리즘은 다윈의 적자생존 이론을 기본 개념으로 하여 최적화 문제를 해결하는 하나의 방법이다. 우선 문제의 해를 유전자 형식으로 표현하기 위해 해를 표현하는 유전자들의 집합을 염색체로 정의한다. 각각의 염색체는 적합도 함수를 통하여 평가하며, 선택, 교차, 변이의 과정을 반복하여 최적의 해를 계산하는 방식이다(Ray and Srivastava, 2008; Kim et al., 2010, Kwak et al., 2012).

초기화는 유전자 알고리즘을 수행하기 위한 첫 번째 세대를 생성하는 과정으로 보통 임의로 염색체에 유전자를 채우는 과정이다. 여기서 염색체 내의 유전자 개수와 세대 내의 염색체의 개수는 주어진 문제에 따라 수행자가 값을 선정하게 된다.

Table 1. Some examples of attributes in each category of open dataset(Statistics Korea: www.kosis.kr)

Category	Items
Environment	Time average air pollution level, Daily meteorological information, Day average air pollution information at roadside in period, Hour air pollution information in realtime at roadside, observatory, etc.
Culture	Cultural heritage information, Library for handicapped information, Cultural space In Seoul (Exhibition facilities), Library for handicapped information, etc.
Industry	Market and mart Information, Market status in Seoul (by Gu), City gas usage in Seoul, Financial institutions in Seoul, Distributor status in Seoul, etc.
Safety	Prevention of disasters facility status (shelters), stormwater pumping stations operation information in Seoul, etc..
Welfare	Facilities for disabled, marriages with foreigners, Child care facilities, Population who health insurance coverage status, Elderly welfare facilities, etc.
Public Health	Tubercular status, Pharmaceutical manufacturing and dealership status, Spatial information about public toilet, Smoking status, etc.
Education	Employment rate of vocational high school, Private institutions and reading room, Continuing education person arrangement, Public library, etc.
Administrative	Chartered house price index, Road infrastructure, Unauthorized buildings, Apartment status, Stormwater pumping stations, etc.
Transportation	Persons get on and off by each subway station, Parking capacity in public parking lot, Bike paths, Parking lots, etc.

선택은 부모세대에서 자식세대로 전해지는 염색체를 결정하는 과정이다. 선택방법에는 균등 비례 룰렛 휠 선택, 토너먼트 선택, 순위 기반 선택 등이 있다. 염색체의 선택은 유전자 알고리즘의 성능에 큰 영향을 끼친다. 어떤 선택 방법을 이용하느냐에 따라 나쁜 염색체가 자식 세대에 전해질 수도 있고, 좋은 염색체만 자식세대에 전해질 수 있다. 그렇지만 적합성 함수를 통하여 나쁜 염색체로 결정이 났을 지라도 전역 최적 값에 더 가까울 수 있으므로 실제로는 나쁜 염색체도 선택될 수 있도록 비례균등 룰렛 휠 선택이 많이 이용되고 있다.

교차는 선택을 통해 자식세대에 전해질 염색체들을 교배하여 새로운 유전자 조합의 염색체를 자식세대에 가지게 하는 방법이다. 보통 교차는 부모세대의 염색체 2개를 각각 나누어 서로 위치를 치환하는 방법을 이용한다. 좋은 염색체를 이용하여 교차를 수행한다 하여도 적합도 평가에서는 결과가 나빠질 수 있다. 그렇기 때문에 교차의 경우에도 확률적인 접근을 통하여 수행하는 것이 일반적이다.

마지막으로 변이는 기존의 염색체에서 특정 유전자가 임의로 다른 유전자로 바뀌는 방법이다. 선택과 교차는 부모의 염색체 사이에서 자식세대를 생성하는 과정이지만 변이는 부모의 염색체 이외의 임의의 염색체에서 자식세대를 생성하는 과정으로 지역 최적 값에 빠질 위험을 줄여주며 문제 해의 다양성을 높여준다.

2.4 Ward의 최소분산법

Ward의 최소분산법(Ward`s Minimum Variance)은 일반적으로 사용되는 계층 군집 방법 중 하나로 잔차제곱합(Error Sum of Squares, ESS)의 증가가 최소화하도록 군집을 만드는 방법이며, 개체간의 거리는 Euclidean distance를 사용한다. i번째 군집내의 잔차제곱합을 ESS_i라 하면, 그 계산은 수식 (2)와 같이 표현된다.

$$ESS_i = \sum_{j=1}^{m_i}\sum_{k=1}^{d}(x_{ijk} - \overline{x_{ik}})^2, \overline{x}_{ik} = \frac{1}{m_i}\sum_{j=1}^{m_i} x_{ijk} \quad (2)$$

여기서 x는 군집 내 포함된 원소 벡터, m은 i번째 군집의 원소 총 개수, d는 전체 속성값의 차원를 나타내며 전체 n개 군집에 대한 잔차제곱합은 다음 수식 (3)과 같이 계산된다.

$$ESS = \sum_{i=1}^{n} ESS_i \quad (3)$$

잔차제곱합은 군집이 병합될 때마다 증가하며 두 군집 c_p와 c_v가 병합될 때에 잔차제곱합의 증분은 다음 수식 (4)와 같이 두 군집사이의 거리를 가리킨다. 즉, 새로운 군집으로 인하여 파생되는 ESS의 증가량을 두 군집 사이의 거리(λ)로 정의하고, 가장 유사성이 큰 군집으로 묶어 나가는 방법이다. 여기서 $\overline{x}$는 군집 원소값의 평균 벡터를 나타낸다(Statistical Research Institute, 2008; Milligan and Cooper, 1985).

$$\lambda = \| \overline{x_p} - \overline{x_q} \|^2 / (\frac{1}{m_p} + \frac{1}{m_q}) \quad (4)$$

2.5 K-Means 알고리즘

K-Means 알고리즘은 군집화 방법 중 대표적인 비계층적 방법의 하나이다. 먼저 K개의 원소는 초기 중심값을 통하여 군집이 이루어지고 각각의 원소는 군집중심으로부터 가까운 거리에 따라 해당 군집으로 할당되게 된다. 초기 군집으로부터 새로운 중심값이 계산되면 각 원소를 다시 재할당하는 연산을 반복하여, 궁극적으로 주어진 데이터를 K개의 군집으로 묶는 과정에서 군집내 제곱합(within sum of squares)이 최소가 되도록 각 원소를 가까운 중심으로 재할당하는 방법이다. 개체들 간의 거리는 Euclidean distance를 사용하게 되며 재할당 기준인 군집 내 제곱합(s)은 다음의 수식 (5)와 같다(Statistical Research Institute, 2008; Hartigan and Wong, 1979).

$$s = \sum_i \sum_j (x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})^T \quad (5)$$

3. 연구결과

강남 3구와 다른 구를 구별짓는 속성정보를 찾기 위하여 미리 강남 3구와 다른 구들이 2가지로 군집화가 되어있다고 가정하고 유전자 알고리즘의 적합도 함수인 Dunn 지수를 최대로 만드는 속성의 종류와 갯수를 찾는 방향으로 연구를 진행하였다. 우선, 유전자 알고리즘을 본 연구에서는 Dunn 지수를 최대로 만드는 속성의 차원을 모르는 상태이기 때문에 이에 대응 되는 염색체의 유전자 수를 1부터 718개로 증가시켜 가며, 각각의 경우에 대하여 유전자 알고리즘을 계산하였다. 만약 유전자 수가 1개일 경우와 718개일 경우는 각각 718번과 1번만의 전수 조사를 수행하고, 그 외의 경우는 사용하는 유전자의 개수(속성의 차원, d)에 따라 조합으로 계산되어($_{718}C_d$) 연산량이 급격히 증가하므로

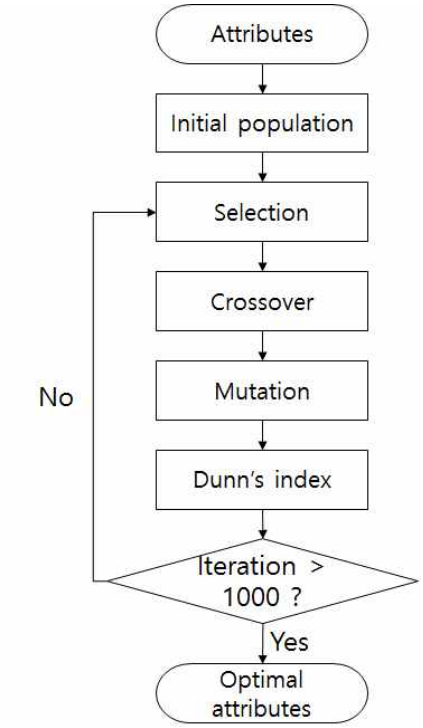

Figure 2. Flow chart

유전자 알고리즘을 통한 방법을 이용하였다. 유전자 알고리즘을 수행하기 위한 초기 염색체의 개수(population number)는 20, 교배 확률은 0.825, 변이 확률은 0.05로 선택하였으며, 각 유전자 알고리즘 시행 시의 반복 횟수(iteration number)는 1,000회로 제한하였다. 다음의 Fig. 2는 본 연구에서 수행한 유전자 알고리즘 연산의 연구 흐름도를 나타내고 있다.

한편, 본 연구에서는 연산 속도의 향상을 위해 Microsoft Azure cloud computing에 기반한 병렬처리 연산을 이용하였다(Microsoft Azure, 2014). 병렬 처리 연산에 이용한 컴퓨팅 환경은 Table 2와 같으며, 병렬 처리 전 연산 시간은 33.7시간이 소요된 데 반해, 병렬 처리 이후의 소요시간은 약 3.8시간 정도로 많은 시간이 단축된 것을 확인할 수 있었다. 최종적으로 유전자 알고리즘 적용 후, Table 3과 같은 28개의 속성을 사용하여 수행하였을 때, Dunn 지수가 0.999 이상으로 가장 높게 나타난 것으로 확인되었다.

Table 2. Test resource

Item		Product specifications
Head node	CPU	2.1 GHz
	Memory	7.0 GB
Worker node	CPU	2.1 GHz
	Memory	14.0 GB
Operating system		Windows Server 2012
Program		MATLAB
Number of cluster		116

또한, 유전자 알고리즘과 Dunn 지수를 통해 선택된 28개의 속성에 대한 검증을 위해 기존에 군집화 방법 중 계층적 방법인 'Ward의 최소분산법'과 비계층적 방법인 'K-means Algorithms'의 결과를 비교하였다. 두 방법은 SAS를 이용하여 계산하였다. 다음의 Fig. 3은 Ward의 최소분산법을 통해 각 군집간의 병합 과정을 보여주는 덴드로그램(dendrogram)이다. 그림에서 x축은 개체('구'), y축은 해당 군집수에 대한 결정계수(R^2)를 나타낸다. 덴드로그램은 군집형성에 대한 시각적 탐색 기능을 제공함으로써 군집의 개수를 정하는 데 좋은

Table 3. The selected attributes by Genetic algorithm and Dunn's index

Number of male volunteers aged 40 to 49	Number of female volunteers aged 20 to 29
Number of oriental medicine wholesale stores	Number of postman
Number of special transports	Number of child welfare institutions
Number of nurses in public health centers	Human resources in public health centers
Value of old public land	Roads in land use
Number of foreign wives	Number of marketplaces
Number of private day care centers	Number of national basic livelihood act recipients
Number of private high school graduates	Number of classrooms in private high schools
Number of parcels from abroad	Land use zoning
Forest areas	Number of rooms per household
Number of unregistered labor unions	Numbers of the Korean Confederation of Trade Union Members
State of land use	Number of vans
Children of families without parents in high school	Female householders among children of families without parents
Number of second grade tourist hotels	Number of guest rooms of tourist hotels

참고가 될 수 있다. Ward의 최소분산법을 통해 28개 속성값이 최후의 2개로 군집화 된 결과를 보면, x축에서 강남 3구를 나타내는 서초구, 강남구, 송파구가 다른 구들과 분리된 것을 쉽게 확인할 수 있었다. K-means의 경우 별도의 그래픽을 제공하고 있진 않지만, K=2일 경우 군집결과에서 역시 강남 3구가 성공적으로 분류된 것을 확인할 수 있었다.

한편, 두 방법의 군집화 결과의 타당성을 정량적으로 판단하는 데는 일반적으로 PSeudo F(PSF) 통계량이 사용된다(Milligan and Cooper, 1985; Statistical Research Institute, 2008). PSF는 군집의 밀집 정도를 측정하는 통계량으로서, 군집간 분산과 군집 내 분산의 비를 사용한다(Milligan and Cooper, 1983; Statistical Research Institute, 2008). PSF는 다음과 같이 계산된다.

$$PSF = \frac{R^2/(C-1)}{1-R^2/(m-C)} \quad (6)$$

여기서 C는 군집개수, R^2는 결정계수로서, 군집 간 분산을 의미하고, m은 총 원소의 수를 나타낸다. 이때 PSF값이 클수록 좋은 군집결과를 나타낸다.

Table 4는 Ward의 최소분산법과 K-means 방법을 통해 군집 분석을 수행한 결과를 나타내고 있다. SAS에서 통계량이 제공된 최대 5개부터 2개까지의 군집분류 결과를 분석해 본 결과, Ward의 최소분산법 및 K-Means 알고리즘 모두 군집개수가 줄어 듬에 따라 PSF 지수가 지속적으로 증가하여 강남 3구와 다른 22개의 구를 두 군집으로 분류했을 때 최대 수치가 계산됐음을 확인할 수 있었으며, 이를 통해 유전자 알고리즘을 이용하여 선정된 속성값이 강남 3구를 다른 구들과 성공적으로 구별했음을 확인할 수 있었다.

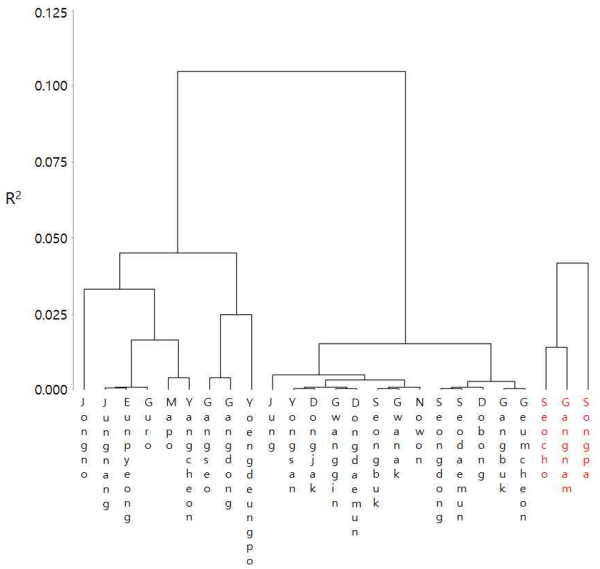

Figure 3. Dendrogram of Ward's minimum variance

Table 4. Clustering results

Methods	Number of clusters	PSF
Ward's minimum variance	5	34.8
	4	34.9
	3	40.9
	2	49.6
K-means	5	25.1
	4	35.9
	3	43.1
	2	49.6

4. 결 론

본 연구는 Dunn 지수를 적합성 평가 함수로 이용한 유전자 알고리즘을 통해 강남 3구(와 나머지 구들을 구별 짓는 속성을 찾아내고자 하였다. 그 결과, 총 718개의 통계자료 중에서 Dunn 지수와 유전자 알고리즘을 통해 강남 3구를 구별 짓는 속성값은 총 28개로 결정되었으며, 그 결과는 Table 3과 같이 나타났다. 또한 선별된 28개 속성을 이용하여 기존의 군집화 방법인 Ward의 최소분산법과 K-means 알고리즘을 이용하여 군집을 형성할 경우 강남 3구와 나머지 구들이 성공적으로 분류되었음을 덴드로그램과 PSF 통계량 계산을 통해 알 수 있었다.

강남 3구를 특징짓는 요소는 호텔 등록 현황, 토지평가액, 사립고등학교 졸업자 수 등 일반적으로 생각할 수 있을 만한 강남의 특징을 대변하는 속성 정보가 포함되어 있는 반면, 민주노총 조합원수, 보건소 인력현황, 한약도매상 판매업소 등의 예상치 못한 속성 정보 역시 포함되어 있다. 후속 연구에서는 분류된 속성 정보에 대한 심도 깊은 분석을 통해 앞서 설명한 마케팅, 사업 전략 수립 및 의사결정 지원 등 실직적인 활용 방안을 모색해 볼 것이다. 한편, 향후 '동' 단위 혹은 인구조사 단위까지 세밀한 공간구분의 적용이 가능할 경우, 새로운 형태의 군집 간 특성을 찾아 낼 수 있을 것으로 사료된다.

한편, 유전자 알고리즘은 휴리스틱 기법 중 하나이므로 매 수행마다 결과가 달라질 수 있다. 따라서 국소 최저치를 피하고 최적화된 결과를 얻기 위해서는 파라미터를 변경하면서 연산을 반복할 필요가 있다. 또한 본 연구에서 적합도 평가를 위해 사용한 Dunn 지수와 군

집화 알고리즘인 Ward의 최소분산법 및 K-means 알고리즘 외에 보다 다양한 알고리즘 간의 비교·분석을 수행할 예정이다. 아울러 본 연구에서 효과적인 유전자 알고리즘 연산을 위해 사용한 클라우드 분산 컴퓨팅 방식은 MATLAB 기반의 for문 처리에만 사용하였으며, 사용한 cluster의 숫자에 비해 효율성은 떨어지는 결과를 나타내었다. 따라서 향후 프로그래밍 최적화를 통한 연산 속도 향상에 대한 연구를 병행할 예정이다.

감사의 글

본 연구는 '국토교통부 국토공간정보연구사업 국토공간정보의 빅데이터 관리, 분석 및 서비스 플랫폼 기술개발(14NSIP-B081011-01)과제'의 연구비 지원에 의해 연구되었음.

References

1. Bezdek, J. C. and Nikhil R. P., 1995, Cluster validation with generalized dunn's indices, Proc. of the 2nd New Zealand Conference, pp. 190-193.
2. Hartigan, J. A. and Wong, M. A., 1979, Algorithm as 136: a k-means clustering algorithm. Journal of the Royal Statistical Society, Vol. 28, No. 1, pp. 100-108.
3. Hinneburg, A. and Kein, D. A., 1998, An efficient approach to clustering in large multimedia databases with noise, Proc. of the 4th International Conference on Knowledge Discovery and Data Mmining, pp. 58-65.
4. Kwak, S. Y., Nam, H. W. and Jun, C. M., 2012, An optimal model for indoor pedestrian evacuation considering the entire distribution of building pedestrians, Korea Society for Geospatial Information System, Vol. 20, No. 2, pp. 23-29.
5. Kim, S. W. and Ahn, H. C., 2010, Development of an intelligent trading system using support vector machines and genetic algorithms, Korea Intelligent Information System Society, Vol. 16, No. 1, pp. 71-92.
6. Kim, U. G., Ahn, W. S., Lee, C. Y. and Um, M. J., 2012, The optimal analysis of data preprocessing method for clustering the region of precipitation, Journal of Korean Society of Hazard Mitigation, Vol. 12, No. 5, pp. 233-240.
7. Microsoft, 2014, Microsoft azure, http://azure.microsoft.com/ko-kr/
8. Milligan, G. W. and Cooper, M. C., 1985, Anexamination of procedures for determining the number of clusters in a data set, Psychometrika, Vol. 50, pp. 159-179.
9. Rademacher, L., 2005, Combinatorial optimization, http://www-math.mit.edu/ ~goemans/18433-FALL05.html.
10. Ray, A. and Srivastava, D. C., 2008, Non-linear least squares ellipse fitting using the genetic algorithm with applications to strain analysis, Journal of Structural Geology, Vol. 30, pp. 1593-1602.
11. Statistical Research Institute, 2008, Segmentation of rural areas based on the attributes of agricultural and fishing villages, Technical report, p. 40.