The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.591
33(5), 591–601

A new cluster validity index based on connectivity in self-organizing map

Sangmin Kim[a] · Jaejik Kim[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received August 4, 2020; Revised August 19, 2020; Accepted August 19, 2020)

Abstract

The self-organizing map (SOM) is a unsupervised learning method projecting high-dimensional data into low-dimensional nodes. It can visualize data in 2 or 3 dimensional space using the nodes and it is available to explore characteristics of data through the nodes. To understand the structure of data, cluster analysis is often used for nodes obtained from SOM. In cluster analysis, the optimal number of clusters is one of important issues. To help to determine it, various cluster validity indexes have been developed and they can be applied to clustering outcomes for nodes from SOM. However, while SOM has an advantage in that it reflects the topological properties of original data in the low-dimensional space, these indexes do not consider it. Thus, we propose a new cluster validity index for SOM based on connectivity between nodes which considers topological properties of data. The performance of the proposed index is evaluated through simulations and it is compared with various existing cluster validity indexes.

Keywords: cluster validity index, self-organizing map, connectivity, cluster analysis

1. 서론

오늘날 컴퓨터와 데이터베이스(database)의 비약적인 발전으로 인해 방대한 양의 데이터가 쌓이고 있고, 이러한 현상에 대해 우리는 빅데이터(big data)라고 부른다. 이러한 빅데이터의 대표적인 특징 중 하나는 자료가 고차원(high-dimension)이라는 것이다. 일반적으로 고차원은 자료를 통계적으로 분석하고 모형을 세우는데 여러가지 어려움을 야기한다. 이러한 어려움은 자료를 탐색하고 자료의 구조를 파악하는 과정에서도 발생하는데, 이러한 문제를 해결하는 효과적인 방법 중 하나가 자기조직화지도(self-organizing map)이다.

자기조직화지도는 고차원의 데이터를 프로토타입 벡터(prototype vector)를 이용해 저차원의 노드(node)에 투영하는 기법이다 (Kohonen, 1997). 2차원 또는 3차원의 공간에 투영된 노드들을 이용하여 고차원 자료의 시각화가 가능하며 이는 자료의 구조와 특징을 파악하는데 용이하다. 그러나 단순한 시각화로는 자료의 구조를 파악하는데 한계가 있기 때문에 이 노드들에 대한 군집분석을 실시한다면 자료

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2018R1D1A1B07049818).
[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: jaejik@skku.edu

의 구조를 이해하는데 도움이 될 것이다. 자기조직화지도로 부터 구한 노드들에 대한 군집분석 과정을 2단계 접근법(two-level approach)이라고 한다 (Vesanto와 Alhoniemi, 2000).

군집분석의 여러가지 중요한 문제 중 하나는 주어진 자료에 대한 최적의 군집 개수를 결정하는 것이다. 그러나, 주어진 자료의 최적의 군집 개수를 결정하는데는 명확한 성공의 기준이 없기 때문에 쉽지 않은 문제이다. 종종 분석자의 주관적인 판단으로 군집의 수를 정하는 경우도 있지만, 자료를 기반으로 계산하는 군집타당성지수(cluster validity index)를 이용하는 것이 더 일반적이다. 현재까지 많은 군집타당성지수가 개발되어 쓰이고 있으며, 여러 군집타당성지수들의 성능을 비교하고자 모의실험을 한 논문이 발표된 바 있다 (Milligan과 Cooper, 1985). 그러나 데이터의 형태에 따라 지수들의 성능이 다르기에 주어진 상황에 대한 정확한 이해를 바탕으로 군집타당성지수를 쓰는 것이 바람직하다고 할 수 있다. 군집타당성지수는 기본적으로 군집 내의 밀집도와 군집 간의 분리된 정도를 바탕으로 정의되는데, CH 지수 (Calinski와 Harabasz, 1974), 실루엣(silhouette) 지수 (Kaufman과 Rousseeuw, 1990), C 지수 (Hubert와 Levin, 1976), KL 지수 (Krzanowski와 Lai, 1988) 등이 대표적으로 많이 쓰이는 지수들이며, 이 지수들은 기본적으로 데이터의 분포를 구형분포라고 가정하고 군집의 개수를 판단하기 때문에 비대칭 분포에 취약할 수 있다.

이러한 군집타당성지수들은 2단계 접근법에서 자기조직화지도의 노드들에 대한 군집분석에도 군집의 개수를 결정하기 위해 바로 적용될 수 있다. 그러나, 자기조직화지도는 원자료의 위상적 특성을 저차원 공간의 노드를 통해 반영한다는 특성을 갖는데, 이러한 일반적인 군집타당성지수들은 자료의 위상적 특성을 전혀 고려하지 않는 문제가 있고 이는 군집의 개수를 결정하는데 있어 정확성을 떨어뜨리는 결과를 초래할 수 있다. 이에 본 연구에서는 Tasdemir와 Merenyi (2006)가 개발한 원자료의 위상적 특성을 고려한 노드들 사이의 연결강도(connectivity)에 기반한 군집타당성지수를 제안하고자 한다. 연결강도는 자기조직화지도에서 노드들의 유사한 정도를 수치화한 것으로 노드들이 유사할수록 높은 수치의 강도를 제공한다. 연결강도는 노드들 간의 유사도만을 바탕으로 계산되기 때문에 원자료의 분포에 대한 사전적인 가정이 없다. 따라서 2단계 접근법에 대해 일반적인 군집타당성지수보다 높은 정확성을 기대할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 자기조직화지도와 2단계 접근법, 그리고 기존의 군집타당성지수에 대해 간략히 소개하고, 3절에서는 자기조직화지도에서의 연결강도를 소개하고 이를 기반으로 한 새로운 군집타당성지수를 제안한다. 4절에서 모의실험을 통해 본 연구에서 제안된 군집타당성지수의 성능을 검증하고 기존의 지수들과 비교한다.

2. 배경

2.1. 자기조직화지도와 2단계 접근법

자기조직화지도는 비지도학습(unsupervised learning)의 한 방법으로 고차원 데이터를 저차원 상의 노드에 투영하는 기법이며, 경쟁학습(competitive learning)에 의해 지도가 형성된다. p개의 변수를 갖는 i번째 관찰값을 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, $i = 1, \ldots, N$이라고 하고 우리는 N개의 관찰값을 가지고 있다고 가정하자. 이 경우 자기조직화지도의 입력층은 $\mathbf{x}_i$들의 집합이며, 출력층을 노드(node)라고 한다. 그리고 각 노드들의 대푯값을 프로토타입 벡터, $\mathbf{m}_j$, $j = 1, \ldots, M$라고 하고 다음과 같이 나타낼 수 있다

$$\mathbf{m}_j = (m_{j1}, \ldots, m_{jp})^\top, \quad j = 1, \ldots, M. \tag{2.1}$$

자기조직화지도의 학습은 주어진 M개의 프로토타입 벡터를 반복적으로 업데이트(update)하면서 학습과정이 수렴할 때까지 이루어진다.

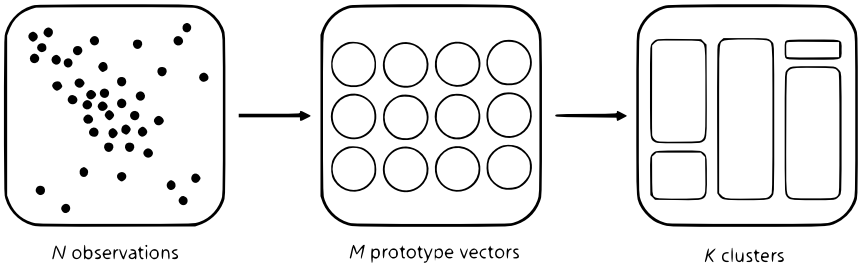


Figure 2.1. Diagram of two-level approach.

각 반복시점 t $(= 1, \ldots, T)$에 대해, 기본적으로 각 관찰값과 프로토타입 벡터들 간의 거리를 계산하며, 거리계산을 위해 일반적으로 유클리디언(Euclidean)거리가 사용된다. 매 반복 t 시점 마다 한 번에 하나의 관찰값에 대해 프로토타입 벡터들이 업데이트된다. 먼저 관찰값 벡터 $\mathbf{x}_i$와 거리가 가장 가까운 프로토타입 벡터를 찾는다. 이를 최적합 유닛(best matching unit)이라고 부른다. 일단 관찰값 $\mathbf{x}_i$에 대해 최적합 유닛 $\mathbf{m}_j$를 찾으면, 그 최적합 유닛에 이웃(neighbor)하는 모든 프로토타입 벡터들 $\mathbf{m}_k$가 $\mathbf{x}_i$를 향해 움직이도록 다음과 같이 업데이트 한다:

$$\mathbf{m}_k \leftarrow \mathbf{m}_k + \alpha h_{kj}(r_k, r_j)(\mathbf{x}_i - \mathbf{m}_k), \quad k = 1, \ldots, M, \ k \neq j, \tag{2.2}$$

여기서 α는 학습률(learning rate)로 프로토타입 벡터가 업데이트되는 정도를 조절해주는 일종의 가중값이고 반복횟수가 늘어날수록 수렴을 위해 단조 감소한다. $h_{kj}(r_k, r_j)$는 근방 함수(neighborhood function)로 최적합 유닛에 가까운 프로토타입 벡터에 더 큰 가중값을 멀리 떨어진 프로토타입 벡터에 작은 가중값을 주는 함수이다. 여기서 r_k와 r_j는 저차원 상에 표시된 노드들의 위치를 나타낸다. 이와 같은 업데이트 과정은 모든 관찰값에 대해 미리 정해진 총 반복횟수 T에 도달할 때까지 또는 α가 충분히 작아질 때까지 반복된다.

위와 같은 자기조직화지도의 학습과정을 통해 프로토타입 벡터들을 얻었다면 이 프로토타입 벡터들을 노드들에 대한 관찰값처럼 취급하여 일반적인 군집분석을 시행함으로써 고차원 원자료의 구조를 파악하는 것이 가능하다. 이러한 방법을 Vesanto와 Alhoniemi (2000)는 2단계 접근법이라고 하였고, Figure 2.1과 같이 도식화할 수 있다. 즉, 1단계에서 자기조직화지도를 통해 M개의 프로토타입 벡터들을 구하고, 이 프로토타입 벡터들을 이용하여 2단계에서 군집분석을 시행하여 주어진 자료의 군집의 개수 및 각 군집의 특징을 파악하는 방식이다. 2단계 접근법은 관찰값의 개수가 매우 클 경우 직접 원자료를 군집화하기 보다 관찰값의 개수보다 훨씬 작은 M개의 프로토타입 벡터들을 군집화하기 때문에 계산 시간을 단축시킬 수 있고, 자료에 잡음(noise)이 있거나 특이값(outlier)이 있을 경우 로버스트(robust)하다는 장점이 있다.

2.2. 군집타당성지수

군집분석은 N개의 관찰값들을 비슷한 특성을 갖는 K개의 군집으로 묶는 작업이고, 이를 통해 자료의 구조를 파악하는데 용이한 방법이다. 이러한 군집분석에서 최적의 군집의 개수를 결정하는 것은 명확한 성공의 기준이 없기 때문에 쉽지 않은 문제이다. 이 문제 해결을 돕기 위해 지금까지 수많은 군집타당성지수들이 개발되어왔다. 군집내의 분산(또는 거리)이 작고 군집간의 분산(또는 거리)이 클수록 군집화가 잘 이루어졌다고 판단되므로, 대부분의 군집타당성지수는 군집내 분산과 군집간 분산을 이용하여 정의된다. 본 절에서는 몇 개의 대표적인 군집타당성지수들을 소개한다.

먼저 군집분석에서 얻어진 K개의 군집을 C_k, $k = 1, \ldots, K$라고 하자. 각 군집 C_k는 N_k개의 관찰값을 가지고 있고 $\sum_{k=1}^{K} N_k = N$이다. 또한, 군집 C_k내의 N_k개의 관찰값들의 평균벡터(mean vector)를 $\bar{\mathbf{x}}_k$라고 표시하고, 전체 N개의 관찰값들의 평균벡터를 $\bar{\mathbf{x}}$라고 하자.

Calinski와 Harabasz (1974)는 군집내와 군집간 분산들의 비율로 다음과 같은 CH 지수를 제안하였다:

$$\mathrm{CH}(K) = \frac{\mathrm{trace}(\mathbf{B}_K)/(K-1)}{\mathrm{trace}(\mathbf{W}_K)/(N-K)}, \tag{2.3}$$

여기서 $\mathbf{B}_K = \sum_{k=1}^{K} N_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^\top$는 군집간의 산포도를 나타내고, $\mathbf{W}_K = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top$는 군집내의 산포도를 나타낸다. 군집내의 산포도는 최소가 되고 군집간의 산포도가 최대가 되어야 좋은 군집 결과이기 때문에 CH 지수는 그 값이 최대가 되는 K를 최적의 군집 개수로 판단한다.

Kaufman과 Rousseeuw (1990)은 실루엣 지수를 개발하였고, 주어진 군집 $C_1, \ldots, C_k$와 $\mathbf{x}_i \in C_l$에 대해 이는 다음과 같이 정의된다:

$$\mathrm{Silhouette}(K) = \frac{1}{N} \sum_{i=1}^{N} \frac{b_K(i) - a_K(i)}{\max\{a_K(i), b_K(i)\}}, \tag{2.4}$$

여기서 $a_K(i) = (1/(N_l - 1)) \sum_{\mathbf{x}_j \in C_l, i \neq j} d(\mathbf{x}_i, \mathbf{x}_j)$이고, $b_K(i) = \min\{(1/N_k) \sum_{\mathbf{x}_j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j),\ k = 1, \ldots, K,\ k \neq l\}$이다. 여기서 $d(\mathbf{x}_i, \mathbf{x}_j)$는 관찰값 $\mathbf{x}_i$와 $\mathbf{x}_j$간의 거리이다. 결국 $a_K(i)$는 각 관찰값이 속한 군집내의 다른 관찰값들과의 평균 거리이고, $b_K(i)$는 각 관찰값이 속하지 않은 다른 군집의 관찰값들과의 평균거리의 최소값이다. 따라서, $a_K(i)$와 $b_K(i)$ 값의 차가 클수록 군집화가 잘 된 것이므로 실루엣 지수를 최대로 하는 군집의 개수를 최적이라고 판단한다.

Hubert와 Levin (1976)에 의해 개발된 C 지수는 다음과 같다:

$$C(K) = \frac{S_W - S_{\min}}{S_{\max} - S_{\min}}, \tag{2.5}$$

여기서 $S_W = \sum_{k=1}^{K} \sum_{\mathbf{x}_i, \mathbf{x}_j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j)$이고, $S_{\min}$는 관찰값들의 모든 순서쌍들의 거리를 제일 작은 값부터 $P_W = \sum_{k=1}^{K} N_k(N_k - 1)/2$번째 작은 값까지 합한 것이며, $S_{\max}$는 반대로 제일 큰 값부터 P_W번째 큰 값까지 합한 것이다. $S_W = S_{\min}$일 때 가장 가까운 거리를 갖는 관찰값들이 모두 같은 군집으로 묶였다는 뜻이므로 C 지수의 값이 최소일 때 최적이 된다.

KL 지수는 Krzanowski와 Lai (1988)에 의해 제안되었으며 다음과 같이 정의된다:

$$\mathrm{KL}(K) = \left| \frac{(K-1)^{\frac{2}{p}} \mathrm{trace}(\mathbf{W}_{K-1}) - K^{\frac{2}{p}} \mathrm{trace}(\mathbf{W}_K)}{(K)^{\frac{2}{p}} \mathrm{trace}(\mathbf{W}_K) - (K+1)^{\frac{2}{p}} \mathrm{trace}(\mathbf{W}_{K+1})} \right|, \tag{2.6}$$

여기서 p는 변수의 개수이고, $\mathbf{W}_K$는 식 (2.3)의 군집내의 산포도 행렬이다. KL 지수가 최대일 때 군집 개수가 최적으로 결정된다.

Milligan과 Mahajan (1980)에 의해 제안된 Ptbiserial 지수는 다음과 같이 정의된다:

$$\mathrm{Ptb}(K) = \frac{\left(\bar{S}_B - \bar{S}_W\right)\left(P_W P_B / P_T^2\right)^{\frac{1}{2}}}{s_d}, \tag{2.7}$$

여기서 $P_W = \sum_{k=1}^{K} N_k(N_k - 1)/2$, $P_T = N(N-1)/2$, $P_B = P_T - P_W$이다. 또한, $\bar{S}_B = S_B/P_B$이고, $\bar{S}_W = S_W/P_W$이다. 여기서 $S_B = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j)$이고, S_W는 식 (2.5)의 군집내 관찰값들의 거리 합이다. 마지막으로 s_d는 관찰값들의 모든 순서쌍들에 대한 거리들의 표준편차이다. S_B는 크고 S_W는 작을수록 잘된 군집화이므로 Ptbiserial 지수가 최대인 군집의 개수가 최적이 된다.

3. 연결강도에 기반한 군집타당성지수

자기조직화지도는 원자료를 저차원상의 노드에 투영함으로 자료의 시각화에 용이하다는 장점 외에도 저차원의 노드들이 원자료의 위상적인 특성을 잘 보존한다는 장점이 있다. 즉, 고차원 상에서 가까운 관찰값들은 자기조직화지도에서 같은 노드 또는 인접한 노드에 위치한다는 것이다. 2.2절에서 소개된 것과 같은 군집타당성지수들은 이러한 자기조직화지도의 위상적 특성을 고려하지 않기 때문에 2단계 접근법에 의한 군집분석시 잘 작동하지 않을 수 있다. 따라서, 본 절에서는 이러한 위상적 특성을 고려한 연결강도라는 측도를 소개하고 이를 기반으로 하는 새로운 군집타당성지수를 제안한다.

Tasdemir와 Merenyi (2006)가 제안한 연결강도는 각 관찰값에서 가장 가까운 프로토타입 벡터인 최적합 유닛과 두 번째로 가까운 프로트타입 벡터인 차적합 유닛을 이용하여 정의된다. 자기조직화지도에서 원자료의 위상적 특성을 보존하기 위해 각 관측값에 대한 최적합 유닛과 차적합 유닛은 저차원상에서 서로 인접한 곳에 위치해 있어야 한다. 이를 바탕으로 인접행렬(cumulative adjacency matrix) $\mathbf{A}$를 정의할 수 있다. 인접행렬 $\mathbf{A}$는 대각요소는 0인 $M \times M$의 정방행렬이며, 인접행렬의 j번째 행과 j'번째 열의 요소 $A_{jj'}$은 j번째 프로토타입 벡터를 최적합 유닛으로 하고 j'번째 프로토타입 벡터를 차적합 유닛으로 하는 관찰값들의 개수이다. 이 인접행렬을 기반으로 연결강도 행렬(connectivity strength matrix) $\mathbf{\Phi}$의 j번째 행과 j'번째 열의 요소 $\phi_{jj'}$은 다음과 같이 정의할 수 있다:

$$\phi_{jj'} = A_{jj'} + A_{j'j}, \quad j, j' = 1, \ldots, M. \tag{3.1}$$

연결강도 행렬 $\mathbf{\Phi}$는 대각요소는 0인 $M \times M$의 대칭행렬이다. 따라서, 두 프로토타입 벡터가 구성하는 두 노드 사이의 연결강도는 서로를 최적합 유닛과 차적합 유닛으로 하는 관찰값들의 개수들의 합이 된다. 최적합 유닛과 차적합 유닛으로 하는 관찰값들이 많다는 뜻은 두 노드가 매우 인접해있고 연결강도가 강하다는 것을 의미한다.

자기조직화지도에서 얻은 프로토타입 벡터로 군집분석을 실시하는 2단계 접근법을 이용하면, N개의 관찰값 $\mathbf{x}_1, \ldots, \mathbf{x}_N$은 M개의 프로토타입 벡터 $\mathbf{m}_1, \ldots, \mathbf{m}_M$으로 축소가 되고 이는 다시 K개의 군집 $C_1, \ldots, C_K$로 표현될 수 있다 ($N > M > K$). 각 관찰값 $\mathbf{x}_i$가 자기조직화지도에 의해 최적합 유닛인 프로토타입 벡터로 할당되기 때문에 각 군집 C_k는 프로토타입 벡터들의 군집임과 동시에 관찰값들의 군집이 될 수 있다.

두 프로토타입 벡터 $\mathbf{m}_j$와 $\mathbf{m}_{j'}$의 연결강도 $\phi_{jj'}$ 값이 상대적으로 크고 같은 군집으로 묶였다면 군집화가 잘 되었다고 할 수 있다. 반대로, $\phi_{jj'}$ 값이 상대적으로 작다면 두 프로토타입 벡터는 다른 군집에 할당되어야 좋은 군집화라고 할 수 있을 것이다. 본 연구에서는 연결강도의 이러한 특성을 이용하여 새로운 군집타당성지수를 제안한다. 먼저, 연결강도에 기반한 새로운 군집타당성지수를 정의하기 위해서는 기존의 군집타당성지수의 군집내 분산과 군집간 분산의 역할을 하는 측도들이 필요하다. 이에 본 연구에서는 주어진 K개의 군집 $C_1, \ldots, C_K$에 대해 다음과 같은 내부연결강도(internal connectivity) $\phi_I(K)$와 외부연결강도(external connectivity) $\phi_E(K)$를 제안한다:

$$\phi_I(K) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \left[\sum_{\mathbf{m}_j, \mathbf{m}_{j'} \in C_k, j \neq j'} \phi_{jj'} \right], \tag{3.2}$$

여기서 N_k는 군집 C_k에 있는 프로토타입 벡터들을 최적합 유닛으로 하는 관찰값 $\mathbf{x}_i$들의 총개수이다. 식 (3.2)의 내부연결강도는 같은 군집으로 묶인 프로토타입 벡터들의 평균적인 연결강도라고 할 수 있고, 0과 1사이의 값을 갖는다. 모든 관찰값에 대해 각 관찰값의 최적합 유닛과 차적합 유닛에 해당하는 프로토타입 벡터들이 모두 같은 군집에 있다면 내부연결강도는 1의 값을 가질 것이며 반대의 경우는 0이 될 것이다. 따라서 내부연결강도가 1에 가까운 값을 가질수록 군집화가 잘 되었다고 할 수 있다.

외부연결강도는 K개의 군집들의 모든 가능한 ${}_K\mathrm{C}_2$개의 군집 쌍 $(C_k, C_{k'})$, $k, k' = 1, \ldots, K$, $k \neq k'$들에 대해 서로 다른 군집에 속하는 프로토타입 벡터들의 평균적인 연결강도로 정의되고 다음과 같이 계산할 수 있다:

$$\phi_E(K) = \frac{1}{{}_K\mathrm{C}_2} \sum_{k=1}^{{}_K\mathrm{C}_2-1} \sum_{k'=k+1}^{{}_K\mathrm{C}_2} \frac{1}{N_k + N_{k'}} \left[\sum_{\mathbf{m}_j \in C_k, \mathbf{m}_{j'} \in C_{k'}} \phi_{jj'} \right], \tag{3.3}$$

여기서 N_k와 $N_{k'}$은 각각 군집 C_k와 $C_{k'}$에 있는 프로토타입 벡터들을 최적합 유닛으로 하는 관찰값들의 개수이다. 내부연결강도와 마찬가지로 식 (3.3)의 외부연결강도는 0과 1사이의 값을 가지고 0에 가까울수록 군집화가 잘 되었다고 할 수 있다.

내부연결강도와 외부연결강도는 모든 군집에 대한 관찰값들의 평균적인 개수에 의존하는 측도이기 때문에 군집의 개수가 늘어날수록 두 값 모두 감소하는 경향을 보인다. 또한, 참된 군집의 개수보다 더 큰 군집의 개수들에 대해서는 내부연결강도와 외부연결강도의 값이 감소는 하나 그 감소폭이 작아지게 된다. 따라서 만일 우리가 내부연결강도에 대한 외부연결강도의 비(ratio)를 고려한다면, 그 비는 참된 군집의 개수와 그 보다 큰 군집의 개수를 비교했을 때 감소폭이 가장 커진다. 이 성질을 이용하여 본 연구에서는 군집의 개수를 K개에서 $K+1$개로 증가시킬 때 외부연결강도와 내부연결강도의 비의 감소폭을 이용한 군집타당성지수를 제안한다. 주어진 K개의 군집과 $K+1$개의 군집들에 대해 본 연구에서 제안하는 연결강도에 기반한 군집타당성지수 $\mathrm{CON}(K)$는 다음과 같다.

$$\mathrm{CON}(K) = \frac{\phi_E(K)/\phi_I(K)}{\phi_E(K+1)/\phi_I(K+1)}. \tag{3.4}$$

식 (3.4)의 분자는 K개의 군집 $C_1, \ldots, C_K$에 대한 외부연결강도와 내부연결강도의 비이고, 분모는 $K+1$개의 군집 $C_1, \ldots, C_{K+1}$에 대한 외부연결강도와 내부연결강도의 비이다. 따라서 식 (3.4)의 비율값이 최대가 되는 K가 최적의 군집개수라고 판단할 수 있다.

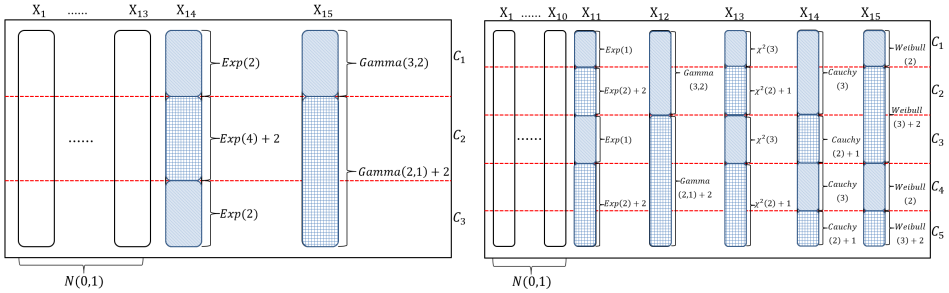
4. 모의실험

본 절에서는 연결강도에 기반한 군집타당성지수의 성능을 다양한 모의실험을 통해 검증하고 기존의 군집타당성지수들과 비교한다. Milligan과 Cooper (1985)는 30개의 군집타당성지수들의 성능을 모의실험을 통해 비교하였다. 그러나 그 모의실험에서는 군집들이 겹쳐있지 않고 잘 나뉘어져있는 경우들을 주로 고려하였으며, 그 경우에는 대부분의 군집타당성지수들이 참된 군집의 개수를 잘 찾아주었다. 또한, 만일 자료에 대한 시각화가 가능한 경우라면 군집들이 잘 나뉘어진 경우 육안으로도 쉽게 군집의 개수를 파악할 수 있을 것이다.

따라서 본 모의실험에서는 관찰값들의 군집들이 서로 잘 나뉘어진 경우와 서로 겹쳐지는 경우 모두에 대해 기존의 군집타당성지수들과 본 연구에서 제안한 군집타당성지수의 성능을 비교해보고자 한다. 모의실험은 Table 4.1에서 보여지듯이 4가지 설정에 대해 시행되었다. 먼저 모든 모의실험에 대해 원자료로서 15개의 변수를 갖는 1,000개의 관찰값들이 생성된다. 군집의 개수는 모든 모의실험에 대해 3개와 5개인 경우를 고려하였다. 군집들이 잘 분리된(separated) 경우 (S1 & S2)에 대해서는 다변량 정규분포로 부터 관찰값들이 생성된다. 즉, 군집 C_k에 대한 관찰값들은 $\mathrm{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$로 부터 무작위로 추출된다. 군집이 3개인 경우 각 군집의 평균벡터는 $\boldsymbol{\mu}_1 = (0, \ldots, 0)^\top$, $\boldsymbol{\mu}_2 = (4, \ldots, 4)^\top$, $\boldsymbol{\mu}_3 = (8, \ldots, 8)^\top$이다. 또한, 군집이 5개인 경우 첫 세 개의 군집에 대한 평균벡터는 앞서 정의된 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$이고, 4번째 군집은 $\boldsymbol{\mu}_4 = (12, \ldots, 12)^\top$, 마지막 군집은 $\boldsymbol{\mu}_5 = (16, \ldots, 16)^\top$이다. 각 군집의 분산-공분산 행렬 $\boldsymbol{\Sigma}_k$는 통계 소프트웨어 R의 패키지 중 하나인 `clusterGeneration`을 이용하여 무

Table 4.1. Original data generation

	Simulation1(S1)	Simulation2(S2)	Simulation3(S3)	Simulation4(S4)
# of obs. (N)	1,000			
# of variables (p)	15			
# of nodes (M)	100, 121, 144, 156, 182 196			
# of clusters (K)	3 & 5			
Type of clusters	Separated		Overlapped	
Sizes of clusters	Equal	Unequal	Equal	Unequal



(a) 3 overlapped clusters (S3 & S4) (b) 5 overlapped clusters (S3 & S4)

Figure 4.1. Simulation settings for S3 and S4; original data are generated from asymmetric distributions.

작위로 생성한다. 군집들이 겹쳐지는 경우 (S3 & S4)에 대해서는 Figure 4.1에서 보여지는 바와 같이 정규분포와 다양한 감마(gamma), 지수, 카이제곱(chi-square), 코시(Cauchy), 와이불(Weibull) 분포와 같은 비대칭 분포들에서 관찰값들이 생성된다. 각 군집의 크기는 모든 군집들이 같은 개수의 관찰값을 갖는 경우 (S1 & S3)와 서로 상이한 개수의 관찰값들을 갖는 경우 (S2 & S4)들 모두를 고려하였다. 군집마다 관찰값의 개수가 상이한 경우에 대해서는 군집이 3개인 인 경우 $N_1 = 500$, $N_2 = 400$, $N_3 = 100$으로 설정하였고, 군집이 5개인 경우는 $N_1 = 300$, $N_2 = 250$, $N_3 = 200$, $N_4 = 150$, $N_5 = 100$으로 하였다.

이렇게 생성된 관찰값들에 대해 먼저 미리 정해진 노드들의 개수(M)에 대해 자기조직화지도를 실행하여 M개의 프로토타입 벡터들을 얻은 후, 그 프로토타입 벡터들에 대해 Ward (1963)가 제안한 Ward 군집방법을 이용하여 군집들을 구하였다. 자기조직화지도에서 노드의 개수를 정하는 명확한 기준은 없다. 그러나 노드의 개수가 군집분석의 결과에 영향을 줄 수 있기 때문에 본 모의실험에서는 6개의 다양한 크기의 노드들을 고려하였다. 이러한 군집분석의 결과에 대해 각 군집타당성지수를 적용하여 각 지수의 기준들에 의해 최적의 군집의 개수를 결정하였다. 이 과정은 총 200번 반복되어 전체 반복횟수 중 참인 군집 개수를 찾는 비율을 성능 비교의 척도로 고려하였고, 그 결과는 Table 4.2에서 보여준다.

Table 4.2에서 군집이 잘 나뉘어지는 경우 (S1 & S2)에는 본 연구에서 제안한 연결강도에 기반한 군집타당성지수보다 다른 기존의 군집타당성지수들이 군집의 개수를 정확하게 찾아주는 비율이 조금 더 높았으나 그 차이가 크지는 않음을 보여주고 있다. 오히려 C 지수, KL 지수, Ptbiserial 지수보다는 더 높은 비율을 보였다. 군집들이 정규분포로 부터 생성되었고 잘 나뉘어진 경우이기 때문에 군집내 분산과 군집간 분산에 기반한 군집타당성지수들이 더 성능이 좋았다. 또한, 각 군집의 관찰값의 개수가 같은 경우 (S1)보다는 서로 상이한 경우 (S2)일 때 모든 군집타당성지수들의 성능이 떨어짐을 알 수 있었다. 그러나, 군집의 관찰값 개수가 상이하고 군집의 개수가 상대적으로 많은 경우 ($K = 5$) 본 연구에서 제안된 군집타당성지수가 전반적으로 참된 군집의 수를 찾는 비율이 기존의 지수들 보다 더 높았다.

Table 4.2. Separated clusters (S1 & S2)

Simulation settings	# of Clusters	# of Nodes	CON (proposed)	CH	Silhouette	C	KL	Ptbiserial
S1	$K=3$	$M=100$	0.865	**0.890**	0.860	0.640	0.645	0.535
		$M=121$	0.820	**0.875**	0.795	0.595	0.720	0.570
		$M=144$	0.775	**0.825**	**0.825**	0.620	0.605	0.500
		$M=156$	0.770	**0.840**	**0.845**	0.625	0.610	0.515
		$M=182$	0.735	**0.855**	0.850	0.675	0.635	0.510
		$M=196$	0.780	**0.855**	0.820	0.580	0.595	0.490
	$K=5$	$M=100$	0.775	0.820	0.755	0.620	0.605	0.550
		$M=121$	0.745	**0.845**	0.780	0.695	0.640	0.585
		$M=144$	0.720	**0.895**	0.845	0.710	0.720	0.525
		$M=156$	0.755	**0.860**	0.825	0.655	0.685	0.520
		$M=182$	0.720	**0.855**	0.790	0.670	0.560	0.485
		$M=196$	0.775	**0.860**	0.810	0.640	0.580	0.530
S2	$K=3$	$M=100$	0.685	**0.715**	0.645	0.550	0.570	0.470
		$M=121$	0.700	**0.740**	0.640	0.525	0.585	0.490
		$M=144$	0.710	**0.725**	0.700	0.570	0.525	0.530
		$M=156$	0.670	**0.730**	0.635	0.580	0.520	0.500
		$M=182$	**0.680**	0.665	0.605	0.505	0.595	0.550
		$M=196$	0.615	**0.660**	0.610	0.540	0.575	0.525
	$K=5$	$M=100$	**0.625**	**0.625**	0.610	0.510	0.490	0.575
		$M=121$	**0.595**	0.555	0.575	0.470	0.580	0.480
		$M=144$	**0.630**	0.560	0.615	0.550	0.570	0.545
		$M=156$	**0.615**	0.595	0.605	0.500	0.590	0.485
		$M=182$	0.620	**0.645**	0.520	0.525	0.600	0.465
		$M=196$	**0.645**	0.575	0.515	0.455	0.555	0.515
S3	$K=3$	$M=100$	**0.575**	0.145	0.000	0.015	0.010	0.000
		$M=121$	**0.500**	0.140	0.000	0.000	0.005	0.000
		$M=144$	**0.605**	0.160	0.000	0.000	0.020	0.000
		$M=156$	**0.575**	0.170	0.000	0.000	0.040	0.000
		$M=182$	**0.580**	0.155	0.000	0.000	0.015	0.000
		$M=196$	**0.630**	0.220	0.000	0.005	0.040	0.000
	$K=5$	$M=100$	**0.515**	0.170	0.005	0.120	0.000	0.010
		$M=121$	**0.435**	0.180	0.005	0.180	0.005	0.010
		$M=144$	**0.450**	0.105	0.000	0.190	0.000	0.010
		$M=156$	**0.465**	0.120	0.000	0.190	0.000	0.000
		$M=182$	**0.360**	0.075	0.000	0.205	0.005	0.000
		$M=196$	**0.360**	0.105	0.000	0.265	0.005	0.000
S4	$K=3$	$M=100$	**0.505**	0.060	0.000	0.000	0.040	0.000
		$M=121$	**0.600**	0.080	0.000	0.030	0.060	0.000
		$M=144$	**0.535**	0.075	0.000	0.005	0.050	0.000
		$M=156$	**0.540**	0.070	0.000	0.010	0.070	0.000
		$M=182$	**0.575**	0.050	0.000	0.005	0.065	0.000
		$M=196$	**0.515**	0.065	0.000	0.010	0.065	0.000
	$K=5$	$M=100$	**0.660**	0.280	0.000	0.165	0.015	0.000
		$M=121$	**0.585**	0.200	0.000	0.155	0.025	0.010
		$M=144$	**0.440**	0.165	0.000	0.185	0.015	0.000
		$M=156$	**0.405**	0.110	0.000	0.185	0.010	0.000
		$M=182$	**0.310**	0.120	0.000	0.175	0.015	0.000
		$M=196$	**0.250**	0.115	0.000	0.170	0.005	0.000

군집들이 잘 나뉘지 않고 겹쳐져 있고 군집의 분포들이 비대칭인 경우 (S3 & S4)에 대한 Table 4.2의 결과는 본 연구에서 제안한 지수가 군집의 개수와 군집내 관찰값의 개수에 상관없이 기존의 지수들 보다 더 잘 참인 군집의 개수를 찾아주고 있음을 보여준다. 군집들이 겹쳐져있을 때 참인 군집의 구조를 식별

하는 것은 쉽지 않은 문제이다. 비록 아주 높은 비율로 참인 군집의 개수를 찾아주지는 않지만 기존의 지수들 보다는 훨씬 더 좋은 성능을 보여주고 있다. 연결강도에 기반한 지수가 기존의 군집타당성지수들보다 확연히 좋은 성능을 보여준 이유는 대칭인 분포에서 변수를 생성하지 않았기 때문이다. 한쪽으로 치우친 분포를 갖는 경우 각 군집의 중심값이 확연한 차이를 갖지 않는다면 군집들이 상당 부분 겹쳐져 있기 때문에 군집의 산포도에 기반한 기존의 군집타당성지수들로는 군집들의 구조를 식별하는 것이 쉽지 않다. 그러나 연결강도에 기반한 지수는 프로토타입 벡터들의 연결강도를 바탕으로 계산이 되므로 분포의 형태에 구애받지 않는다. 따라서 군집들이 상당히 겹쳐 있는 경우에도 제안된 지수는 군집의 개수를 비교적 잘 찾아준다고 할 수 있다.

마지막으로 자기조직화지도의 노드의 개수가 참된 군집 개수를 식별하는데 어느 정도 영향을 미치는 것으로 보여진다. Table 4.2을 보면 군집내 관찰값이 같거나 상이한 경우, 군집이 잘 나뉘어지거나 겹쳐져 있는 경우 등을 고려한 각 모의실험 설정에서 어떤 특정 노드의 개수에서 가장 좋은 성능을 보여주는 것을 볼 수 있다. 즉 원자료의 군집들의 구조에 따라 그 구조를 가장 잘 반영하는 노드의 개수가 존재하는 것으로 보여진다. 앞서 언급했듯이 자기조직화지도에서 최적의 노드의 개수를 정하는 명확한 방법이 아직 개발되지 않았다. 향후 원자료의 군집 구조를 가장 잘 반영하는 노드의 개수를 구하는 방법이 더 정확한 분석을 위해 필요할 것으로 보여진다.

5. 결론 및 토의

방대한 양의 자료를 분석하고 특성을 파악하기 위해 다양한 통계적 기법이 개발되고 있다. 본 연구는 그 중 방대한 양의 고차원 자료를 시각화할 수 있고 원자료의 위상적 특성을 보존할 수 있는 자기조직화지도에 대해 초점을 맞추었다. 자기조직화지도의 프로토타입 벡터들을 이용하여 군집화하는 2단계 접근법은 많은 양의 자료를 군집화하는데 있어 계산 시간을 단축시켜주며, 원자료가 가지고 있는 잡음이나 이상값에 덜 민감하게 반응한다는 장점을 가지고 있다. 그러나 군집들이 서로 겹쳐져 있거나 군집들의 분포가 정규분포가 아닌 경우 기존의 군집타당성지수를 이용하면 군집의 개수를 잘 찾아주지 못한다는 문제를 가지고 있었다. 이에 본 연구에서는 원자료의 위상적 특징을 보존하는 자기조직화지도의 특성을 반영한 연결강도에 기반한 군집타당성지수를 제안하였다. 본 연구에서 제안된 군집타당성지수는 위상적 특징을 이용한 연결강도에 기반하였기 때문에 군집들의 분포의 형태에 상관없이 좋은 성능을 보여주었다.

비록 모의실험에서는 군집분석의 방법으로 Ward 방법만을 고려하였으나, 본 연구에서 제안한 군집타당성지수는 기본적으로 Ward 방법 이외의 다른 계층적인 군집방법과 K-평균 군집분석과 같은 비계층적인 군집분석 방법들에도 적용하는 것이 가능하다. 다만, 군집분석의 방법마다 군집형성의 알고리즘이 다르므로 본 연구에서 제안한 군집타당성지수의 성능이 차이를 보일 수 있기 때문에 향후 이에 대한 깊은 연구가 필요하다.

본 연구에서 모의실험을 통해 자기조직화지도의 노드의 개수가 최적의 군집의 개수를 찾는데 있어 중요한 요인 중에 하나라는 사실을 알 수 있었다. 향후 2단계 접근법에 의한 더 정확한 군집분석을 위해서는 자기조직화지도에서 최적의 노드의 개수를 결정하는 문제가 고려되어야 할 것으로 보인다. 만일 주어진 고차원 자료에 대해 자기조직화지도를 사용했을때 최적의 노드의 개수를 결정할 수 있다면 원자료의 구조를 시각화하고 파악하는데 도움을 줄 것으로 기대된다.

References

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics*,

3, 1–27.

Hubert, L. and Levin J. (1976). A general statistical framework for assessing categorical clustering in free recall, *Psychological Bulletin*, **83**, 1072–1080.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data*, Wiley, New York.

Kohonen, T. (1997). *Self-Organizing Maps* (2nd ed.), Springer-Verlag, Berlin, Germany.

Krzanowski, W. and Lai, Y. (1988). A criterion for determining the number of groups in a dataset using sum of squares clustering, *Biometrics*, **44**, 23–34.

Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, **50**, 159–179.

Milligan, G. and Mahajan, V. (1980). A note on procedures for testing the quality of a clustering of a set of objects, *Decision Science*, **11**, 669–677.

Tasdemir, K. and Merenyi, E. (2006). Data topology visualization for the self-organizing maps. In *Proceeding of the 14th European Symposium on Artificial Neural Networks*, Bruges, Belgium, 277–282.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map, *IEEE Transactions on Neural Networks and Learning Systems*, **11**, 586–600.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, **58**, 236–244.

자기조직화지도에서 연결강도에 기반한 새로운 군집타당성지수

김상민[a] · 김재직[a,1]

[a]성균관대학교 통계학과

(2020년 8월 4일 접수, 2020년 8월 19일 수정, 2020년 8월 19일 채택)

요 약

자기조직화지도는 고차원의 원자료를 노드들로 이루어진 저차원의 공간으로 투영하는 비지도학습 방법이다. 이 방법은 고차원의 자료를 노드들을 사용하여 2 또는 3차원의 공간에서 시각화할 수 있고, 이를 통해 자료의 특성을 탐색하는데 유용하다. 자료의 구조를 파악하기 위해 종종 노드들에 대한 군집분석을 시도하는데, 군집분석의 중요한 문제중 하나는 군집의 개수를 결정하는 것이다. 이 문제를 해결하기 위해 다양한 군집타당성지수들이 지금까지 개발되어 왔고, 이러한 지수들은 자기조직화지도의 노드들의 군집분석에 직접적으로 적용될 수 있다. 그러나, 자기조직화지도가 원자료의 위상적 특성을 저차원 공간에 반영할 수 있다는 특징을 갖는데 반해, 이러한 일반적인 지수들은 이를 고려하지 않는 문제가 있다. 이에 본 연구에서는 원자료의 위상적 특성을 고려한 노드들 사이의 연결강도를 기반으로 하는 군집타당성지수를 제안한다. 이 새로운 군집타당성지수의 성능은 모의실험을 통해 기존의 군집타당성지수들과의 비교되고 검증된다.

주요용어: 군집타당성지수, 자기조직화지도, 연결강도, 군집분석

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2018R1D1A1B07049818).

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대 통계학과. E-mail: jaejik@skku.edu