디지털산업정보학회 논문지 제17권 제1호-2021년 3월
http://dx.doi.org/10.17662/ksdim.2021.17.1.075

속성유사도에 따른 사회연결망 서브그룹의 군집유효성*

윤 한 성**

Clustering Validity of Social Network Subgroup Using Attribute Similarity

Yoon, Han-Seong

〈Abstract〉

For analyzing big data, the social network is increasingly being utilized through relational data, which means the connection characteristics between entities such as people and objects. When the relational data does not exist directly, a social network can be configured by calculating relational data such as attribute similarity from attribute data of entities and using it as links. In this paper, the composition method of the social network using the attribute similarity between entities as a connection relationship, and the clustering method using subgroups for the configured social network are suggested, and the clustering effectiveness of the clustering results is evaluated. The analysis results can vary depending on the type and characteristics of the data to be analyzed, the type of attribute similarity selected, and the criterion value. In addition, the clustering effectiveness may not be consistent depending on the its evaluation method. Therefore, selections and experiments are necessary for better analysis results. Since the analysis results may be different depending on the type and characteristics of the analysis target, options for clustering, etc., there is a limitation. In addition, for performance evaluation of clustering, a study is needed to compare the method of this paper with the conventional method such as k-means.

Key Words : Social Network, Subgroup, Govern·Newman Algorithm, Cluster Validity Index

Ⅰ. 서론

빅 데이터의 분석도구로서 활용도가 커지고 있는 사회연결망(social network)의 기본적인 구성 및 분석은 사람, 사물 등의 개체와 개체간의 연결특성을 의미하는 관계형 데이터(relational data)를 통해 이루어진다[1]. 그런데 개체간의 관계형 데이터가 직접 존재하지 않는 경우, 개체들의 속성형 데이터(attribute data)로부터 개체간의 속성유사도(attribute similarity)와 같은 관계형 데이터를 계산한 후 이를 연결선(link)과 연결특성으로 활용하여 사회연결망을 구성할 수 있다[2-5]. 이렇게 구성한 사회연결망에서는 관계형 데이터의 값이 기준치 이하 또는 이상인

* 이 연구는 2020년도 경상대학교 발전기금재단 재원으로 수행되었음.
** 경상대학교 경영대학 교수

경우에 연결선의 존재유무가 결정되기도 한다.

사회연결망에서 연결특성 기반의 군집분석(clustering)이 서브그룹(subgroup)분석을 통해 가능한데[6], 서브그룹분석은 개체간의 속성유사도가 아닌 개체간의 연결밀도를 바탕으로 개체들의 군집을 구한다. 즉, 개체간의 연결관계 여부 또는 개체별로 연결되는 타 개체의 수 등과 같은 연결특성에 따라 군집화가 이루어지므로, k-평균 등과 같은 유사성기반 군집분석[7]과는 차이가 있다. 컴포넌트, n-클릭 등의 방식을 포함하는 서브그룹분석의 여러 방식들은 각기 분석의 과정과 결과에서 특징을 가지며, 목적에 따라 분석방식을 선택할 수 있다. 본 논문에서 활용할 거번·뉴먼 알고리즘은 사회연결망의 연결선이 가지는 구조적 특성인 라인매개성(line betweenness)을 계량적으로 파악하여 사회연결망 개체들의 군집을 구하는 서브그룹 분석방식이다[6].

사회연결망의 서브그룹분석에서 개체간의 속성유사도를 연결선의 관계형 데이터로 활용한 사회연결망에서는 서로 연결된 개체들이 서로 더 유사한 속성을 가질 것으로 추측할 수 있다. 본 논문에서는 이와 같은 관점에서 사회연결망에 대한 서브그룹분석의 여러 방식들이 가지는 군집화 특징들을 정리하고, 유사성기반 군집분석의 목적으로 활용이 용이하다고 판단되는 서브그룹분석 방식에 대해서 속성기반의 군집화 성능에 영향을 미치는 요소, 이에 따른 군집화 과정 및 효과 등을 정리하기로 한다.

Ⅱ. 이론적 배경

2.1 사회연결망과 속성유사도 관계

사회연결망은 노드(node) 및 이들을 연결하는 연결선(link)을 기본 구성요소로 하며, 노드와 연결선의 방향성 및 가중치 등을 통해 사람 또는 사물 간의 관계를 네트워크 관점에서 표현하는 방법이다[8, 9]. 사회연결망에서 노드는 일반적으로 사람 또는 사물 등의 개체를 의미하고, 노드간의 연결선은 개체간의 이전, 의사소통, 권력관계, 친족연결 등과 같은 노드간의 연결특성을 표현한다. 사회연결망을 통한 분석기법에는 중심성(centrality), 파당(clique), 밀도(density), 구조적 틈새(structural hole) 등의 다양한 방식이 있으며, 최근 빅 데이터 분석도구로서 활용도가 커지고 있다[10, 11].

사회연결망의 구성을 위해 개체들의 속성형 데이터로부터 계산한 개체간의 속성유사도와 같은 관계형 데이터를 연결선과 연결특성으로 활용한 사례로서, 고객이라는 개체와 이들의 구매내역 데이터로부터 계산되는 고객간의 구매유사도를 들 수 있다. 이 경우에 기준치 이상의 속성유사도에 대해 가상의 연결선으로 구성한 사회연결망을 구성할 수 있다. 속성유사도로써 구한 사회연결망의 사례들이 공통으로 수행하는 작업의 수에 따른 작업자간의 관계[2], 구매상품유사도에 따른 구매자간의 관계[3], 점포간의 판매제품 유사도관계[4], 조직구성원간의 성격특성 유사관계[5] 등의 관련 연구들에서 확인된다. 그런데 이러한 사회연결망 연구들에서도 이루어진 분석들은 대개 중심성(centrality)과 같이 주로 개체간의 연결관계 및 관계성 중심의 분석들이 대부분인 것으로 파악된다.

2.2 사회연결망의 서브그룹분석

한편 사회연결망의 서브그룹분석에는 컴포넌트, k-코어, n-클릭, k-플렉스, 람다집합, 거번·뉴먼알고리즘 등의 여러 방식들이 있으며, 목적에 따라 적절한 방식을 선택할 수 있다. 이러한 방식들에 의한 사회연결망 개체들의 군집화가 가지는 특징을 정리하면 <표 1>과 같다[6, 12]. <표 1>의 서브그룹분석들은 사회연결망의 연결선 밀도에 의해 개체들의 군집화 및

연결구조를 파악하는데 주안점을 두고 있다[13-16]. 또한 개체들의 속성을 고려한 서브그룹분석 또는 이에 의한 군집결과에 대해 개체들의 속성에 의한 군집화 성과는 거의 고려하고 있지 않다.

<표 1> 사회연결망 서브그룹 분석방식

분석방식	차이점
컴포넌트	각 군집별 모든 개체들은 서로 한 개 이상의 연결선으로 구성된 경로로써 연결되고 군집간의 개체들은 서로 연결되지 않는다.
k-코어	코어를 이루는 군집의 각 개체들은 k개 이상의 다른 개체와 연결되며(k=0, 1, n) 1-코어, 2-코어 등의 군집으로 나누어진다. k가 클수록 서로 연결된 개체수가 많은 군집이다.
n-클릭	최단경로가 n개 이하인 경로에 의해 모든 개체가 서로 연결된 형태의 군집을 이룬다.
k-플렉스	서로 경로로써 연결된 군집내의 각 개체는 k개를 제외한 수의 개체들과 서로 직접 연결되는 형태의 군집이다.
람다집합	서로 연결된 개체간의 경로수가 군집외의 개체와 연결된 경로수보다 큰 형태의 군집이다.
거번·뉴먼 알고리즘	높은 라인매개성의 연결선을 제거해가면서 원하는 수의 군집을 탐색한다.

<표 2> 서브그룹 분석방식과 일반적인 유사성기반 군집분석의 차이

분석방식	차이점
컴포넌트	다수의 연결선에 의해 연결된 군집내의 개체들은 개체간의 유사성이 희석될 수 있다.
k-코어	최상위 k값의 코어를 제외하면 이질적 속성의 개체들로 군집을 이룰 가능성이 크다.
n-클릭	작은 값의 n에서는 군집내의 유사성이 높을 수 있지만, n이 커지는 경우 유사성을 유지할 수 있는 적절한 n의 결정이 쉽지 않다.
k-플렉스	k가 작으면 군집의 발견이 어려울 수 있고, k가 커면 군집내의 유사성이 낮아질 수 있다.
람다집합	군집내의 연결경로를 구성하는 연결선 수가 많으면 군집내의 유사성이 낮아질 수 있다.
거번·뉴먼 알고리즘	군집간의 단절효과가 가장 큰 개체간의 연결선을 제거하여 군집내의 유사성을 높인다.

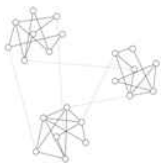
사회연결망 서브그룹 방식들의 특징을 바탕으로, 속성유사도를 연결선으로 구성한 사회연결망에서 여러 서브그룹 방식들과 k-평균을 포함한 일반적인 유사성기반 군집분석의 군집결과에서 보이는 차이를 <표 2>와 같이 정리하였다. 대체로 군집을 이루는 개체들의 유사성을 높이는 효과를 가질 수 있으나, 사회연결망의 연결관계 및 서브그룹분석의 특성으로 인해 k-평균과 같은 유사성기반 군집분석과는 목적과 결과에서 차이가 있다. 그리고 거번·뉴먼알고리즘을 제외하면 군집의 수를 원하는 대로 설정하기에 용이하지 않다.

거번·뉴먼 알고리즘은 사회연결망의 연결선들이 가지는 구조적 특성인 라인매개성(line betweenness)을 계량적으로 파악하여 사회연결망에서 군집을 파악하는 방법이다. 라인매개성은 사회연결망의 모든 노드 쌍 간의 최단경로에 위치하는 횟수를 뜻하며, 거번·뉴먼 알고리즘은 사회연결망에서 중심성지수(centrality index)의 하나인 매개성(betweenness)의 아이디어를 군집의 경계성을 감지하는데 활용한 방법이다[17].

사회연결망에서 거번·뉴먼 알고리즘은 원하는 컴포넌트(서브그룹 또는 군집)의 수가 구해질 때까지 다음 (1)과 (2)의 과정을 반복하는 방식이며, 거번·뉴먼 알고리즘으로 3개의 군집을 나눈 <그림 1>의 사례에서 분석과정 및 결과를 직관적으로 이해할 수 있다.

(1) 존재하는 모든 연결선(link)들에 대해 가장 높은 라인매개성을 가지는 연결선을 제거한다.

(2) 제거한 연결선을 제외한 나머지 연결선들의 경로로써 연결되는 개체들의 그룹(서브그룹)들이 각각 군집이 된다.



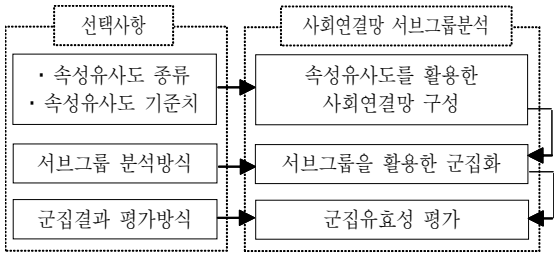
<그림 1> 거번·뉴먼 알고리즘의 군집(3개)화 사례[17]

속성기반 군집화의 결과에 대한 평가는 군집유효성지수(CVI: clustering validity index)를 활용할 수 있는데, CVI는 군집 내 개체들의 응집도(compactness)와 군집 간 개체들의 분리도(separability)를 고려하는 수치이다[18]. 여기서 응집도는 동일한 군집에 존재하는 두 개체간의 유사도가 클수록 높게 평가되는 측정치이고, 분리도는 상이한 각각의 군집에 위치하는 개체간의 유사도가 작을수록 높게 나타나는 수치이다. 속성유사도를 연결선으로 구성한 사회연결망에 대해 <표 2>의 여러 서브그룹 분석방식에 의한 군집화 결과에 대해서도 유사성 기반의 군집유효성을 평가하는 수단으로 CVI를 활용할 수 있다.

개체간의 속성유사도를 활용한 사회연결망에서 직접 서브그룹분석을 통한 군집화의 수행은 사회연결망의 구조를 유지한 채 이루어진다. 이는 중심성과 같은 일반적인 사회연결망분석 결과와 군집화 결과간의 일관성을 가질 수 있을 뿐만 아니라, 사회연결망분석과는 별개로 개체들의 군집분석을 수행하는 번거로움을 줄일 수 있다. 또한 서브그룹에 의한 군집의 수를 설정하기 용이한 경우가 여러 면에서 활용성이 있다고 판단된다. 이러한 필요성에 따라 본 논문에서는 개체간의 속성유사도를 연결관계로 하는 사회연결망의 구성방식, 그리고 구성된 사회연결망에 대해 서브그룹을 활용한 군집화 방식 등을 본 논문에서 정리하고 군집결과에 대한 군집유효성을 평가하기로 한다.

Ⅲ. 연구분석 프로세스

이상의 내용들을 고려하여 본 연구의 범위 및 프로세스는 <그림 2>와 같으며, 사회연결망 서브그룹분석을 위한 선택사항들과 이를 바탕으로 사회연결망의 구성, 서브그룹을 통한 군집화 등을 포함한다. 또한 본 논문에서는 사회연결망의 서브그룹분석에서 군집수의 설정이 용이하고 군집내의 유사성이 적절히 유지되는 것으로 판단되는 거번·뉴먼알고리즘을 중심으로 정리하였다. 덧붙여 서브그룹분석 결과에 대해 군집화 평가방식의 선정과 이를 통한 군집유효성을 평가하는 과정을 실제 데이터와 함께 정리하고자 한다.



<그림 2> 연구 범위 및 구성

Ⅳ. 사회연결망분석을 활용한 군집화

본 장에서는 속성 데이터를 가지는 개체들로부터 관계형 데이터로 활용하기 위한 속성유사도의 선택과 이를 활용한 사회연결망을 구성하고, 구성한 사회연결망에서 거번·뉴먼알고리즘을 통한 군집화와 군집결과에 대한 군집유효성을 평가하는 과정을 정리하기로 한다. 또한, 각 과정에서 조직구성원의 개인별 성격특성을 나타내는 실제 데이터에 활용함으로써 적용가능성을 함께 확인하고자 한다.

4.1 속성유사도를 활용한 사회연결망 구성

수치형 속성 데이터로 이루어진 개체들에 대해 개체간의 속성유사도를 계산하여 사회연결망을 구성하기 위해서는 다음 사항을 먼저 선택하는 것이 필요하다.

- 개체간의 속성유사도 종류
- 연결선 유무를 판단할 속성유사도 기준치

속성유사도는 유클리디안 거리(Euclidean distance), 피어슨상관계수(Pearson correlation), 코사인유사도(cosine similarity) 등으로 계산된다[19-21]. 본 논문에서는 수치형 속성데이터에 폭넓게 활용되는 다음 식의 피어슨상관계수(PC) 및 유클리디안 거리(UD)를 활용하였으며, X_i 및 Y_i는 각각 개체 x 및 y의 속성(i)별 수치이고 $\overline{X}$와 $\overline{Y}$는 각각 X_i 및 Y_i의 평균값이다.

$$PC = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}}$$

$$UD = \sqrt{\sum_i (X_i - Y_i)^2}$$

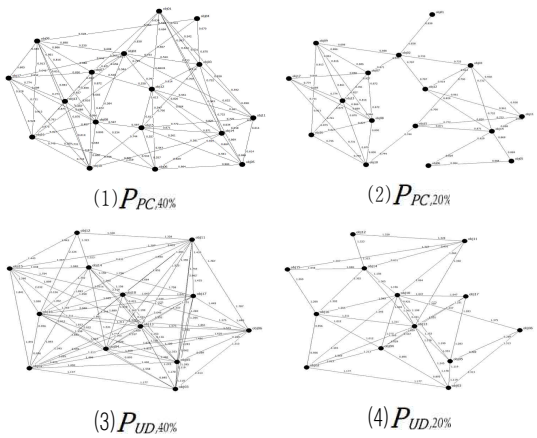
위 식에서 피어슨상관계수(PC)는 -1과 +1사이의 값을 가지는데, +1은 속성이 동일한 경우이며 음(-)의 값은 속성이 상이하다는 것을 의미한다. 유클리디안 거리(UD)는 0 이상의 값을 가지며, 속성이 동일하면 0의 값을 가진다. 유사한 개체들끼리 서로 연결되도록 유사정도가 비교적 큰 값에서 연결선 유무를 결정할 수 있으며, 피어슨상관계수 및 유클리디안 거리의 각각에서는 일정 기준치 이상 또는 이하의 경우에 사회연결망 연결선을 구성할 수 있다.

속성유사도 값이 오름차순으로 나열된 순서에서 연결선 유무를 결정하는 특정 백분위 값 P를 선택할 수 있으며, 예를 들어 피어슨상관계수 또는 유클리디안 거리의 경우에서 각각 '상위 50% 백분위 값 이상($P_{PC,50\%}$)' 또는 '하위 50% 백분위 값 이하($P_{UD,50\%}$)'에 포함되는 속성유사도의 경우 연결선을 구성할 수 있다. 몇 가지 백분위 값에 대해, 조직구성원 개인별 성격특성을 나타내는 <표 3>의 일부 데이터에서 구한 피어슨상관계수 및 유클리디안 거리를 관계형 데이터로 활용하여 구성한 사회연결망의 사례들을 <그림 2>와 같이 구성할 수 있다. <표 3>은 개인별 성격특성을 창의성 등의 6가지 요소에 대해 각 요소별 100점 크기에서 측정한 수치형 속성데이터이다.

<표 3> 개인별 성격특성 데이터 사례(총 80명)

개인	창의성	배려성	주도성	강인성	탐구성	성실성
김원•	43	58	62	56	52	65
서호•	45	53	54	32	58	49
이수•	35	62	45	47	59	55
…	…	…	…	…	…	…

<그림 3>에서 '$P_{PC,40\%} \rightarrow P_{PC,20\%}$' 및 '$P_{UD,40\%} \rightarrow P_{UD,20\%}$'의 순서로 사회연결망에서 연결선의 밀도가 낮아지는 모습을 보인다. 이는 연결망의 존재여부를 판단하는 기준치가 되는 속성유사도가 클수록, 유사성이 큰 경우에만 연결선이 존재하게 되어 사회연결망 전체의 연결선 밀도가 희박해지는 것을 보여준다.



<그림 3> 피어슨상관계수 및 유클리디안 거리를 활용한 사회연결망

4.2 서브그룹을 통한 군집화

앞서 정리한 사회연결망 서브그룹분석들의 특징들

을 고려하여, 본 논문에서는 군집의 수를 비교적 설정하기 용이하고 유사속성의 개체들을 군집으로 분리하는 효과가 크다고 판단되는 거번·뉴먼 알고리즘을 활용하였다.

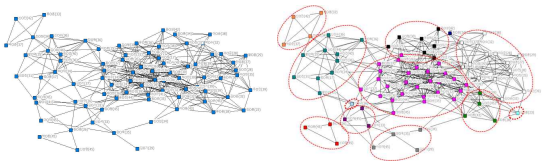
특정 속성유사도를 기준치로 하여 구성한 사회연결망에 대해 임의의 군집수를 설정하여 거번·뉴먼 알고리즘을 통해 서브그룹분석을 할 수 있다. <표 3>의 전체 개체에 대해 '$PC \geq 0.6$'인 경우의 속성유사도를 기준치로 하여 사회연결망을 구성하고, 이에 대해 거번·뉴먼 알고리즘으로 11개까지 서브그룹분석을 시도한 사례를 <그림 4>에서 확인할 수 있다.

사회연결망을 구성할 속성유사도의 기준치와 서브그룹분석을 통한 군집의 수는 분석할 대상의 특성 및 분석목적 등에 따라 적절한 선택이 필요하며, 경험적으로 또는 실험을 통해 선택할 수도 있을 것이다. 본 논문에서는 서브그룹분석의 군집화 과정을 탐색하는 측면과 실험 데이터의 개체수(80)를 고려하여, 중위수 이상의 속성유사도로써 연결되는 사회연결망을 구성할 속성유사도 기준치 및 서브그룹분석을 통한 군집의 수를 <표 4>와 같이 선택하여 분석하기로 한다.

<표 4>에서 선택한 피어슨상관계수 및 유클리디언 거리에서 속성유사도 기준치로 선택한 각 백분위 값에 따라 사회연결망을 작성하여 서브그룹분석을 통해 여러 군집을 구할 수 있다. 피어슨상관계수의 경우에는 유사도 기준치의 상위 백분위수가 작을수록, 그리고 유클리드 거의 경우에는 유사도 기준치의 하위 백분위수가 작을수록 사회연결망의 연결선 밀도가 희박해진다. 피어슨상관계수의 상위 20% 이상 유사도를 의미하는 $P_{PC,20\%}$에서 사회연결망을 구성하고, 거번·뉴먼 알고리즘을 통해 서브그룹분석으로 구한 3~6개의 군집의 경우를 <그림 5>에서 볼 수 있다.

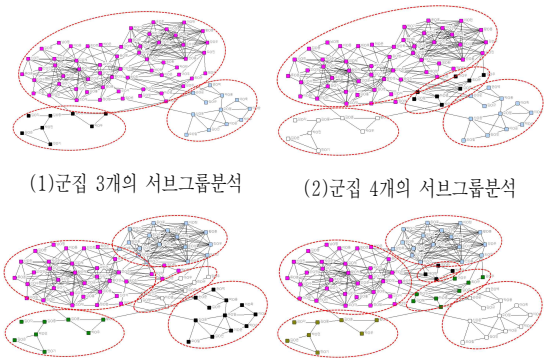
4.3 서브그룹분석 결과의 군집유효성

수치형의 속성데이터를 가지는 개체들을 군집화한



(1)사회연결망의 구성 (2)서브그룹분석 결과

<그림 4> 사회연결망($PC \geq 0.6$) 구성 및 거번·노먼 알고리즘 결과



(1)군집 3개의 서브그룹분석 (2)군집 4개의 서브그룹분석

(3)군집 5개의 서브그룹분석 (4)군집 6개의 서브그룹분석

<그림 5> 피어슨상관계수 $P_{PC,20\%}$ 사회연결망의 서브그룹분석 사례

<표 4> 선택한 속성유사도 기준치 및 군집의 수

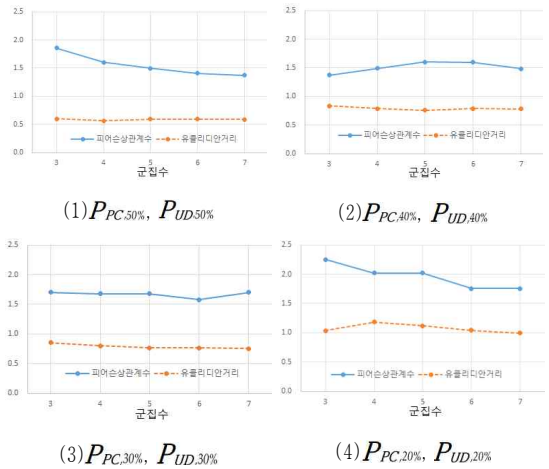
속성유사도 및 기준치		군집의 수
피어슨 상관계수	$P_{PC,50\%}$, $P_{PC,40\%}$, $P_{PC,30\%}$, $P_{PC,20\%}$	각각의 속성유사도 및 기준치로 구성한 사회연결망에서 3~7개의 군집으로 서브그룹분석
유클리디안 거리	$P_{UD,50\%}$, $P_{UD,40\%}$, $P_{UD,30\%}$, $P_{UD,20\%}$	

결과에 대한 군집유효성은 개체간의 응집도와 분리도로써 계산되는 군집유효성지수(clustering validity index: CVI)로써 측정할 수 있다. 여러 CVI 중에서 우수한 것으로 알려진[18] 다음의 DB(Davies-Bouldin) 지수를 활용하기로 한다.

$$DB_K = \frac{1}{K}\sum_{i=1}^{K}\max_{i=1,\cdots,K,\, i\neq j}\left\{\left(\sqrt{\frac{1}{n_i}\sum_{x\in c_i} d(x,z_i)^2} + \sqrt{\frac{1}{n_j}\sum_{y\in c_j} d(y,z_j)^2}\right)/d(z_i,z_j)\right\}$$

위 식에서 n_i는 군집 C_i에 포함되는 개체의 수를 뜻하며, DB_K는 군집수를 K개로 군집화한 경우의 '{군집 내 중심과 객체 간 거리(응집도)의 합}÷{군집들의 중심점 간 거리(분리도)}'를 의미한다. 여기서 DB_K의 값이 작을수록 군집유효성을 높아진다고 할 수 있으며, 최소가 되는 K가 최적의 군집 수가 된다. <표 4>에서 나열한 방식들로 구성한 사회연결망과 각 사회연결망에서 구한 군집수에 따라 DB지수를 통해 군집유효성을 본 절에서 평가하기로 한다.

먼저 서로 대응하는 속성유사도 기준치의 상위 또는 하위의 피어슨상관계수 및 유클리디언 거리로 구성한 여러 사회연결망에서 거번·뉴먼 알고리즘으로 3~7개의 서브그룹으로 군집화한 각각의 결과에 대해 DB지수를 구하여 <그림 6>과 같이 비교·정리하였다. <그림 6>에서 피어슨상관계수보다 유클리디안 거리로 구성한 사회연결망에서의 군집유효성지수인 DB_K의 값이 모든 경우에서 작게 나타난다. 이는 주어진 데이터 및 서브그룹 분석방식에서는 DB지수에 있어서 유클리디안 거리가 피어슨상관계수보다 속성기반 서브그룹분석을 통한 군집화의 결과가 우월하다는 것을 나타낸다.

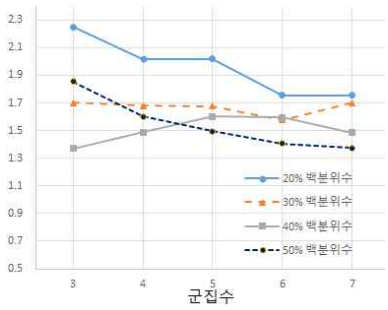


(1)$P_{PC,50\%}$, $P_{UD,50\%}$ (2)$P_{PC,40\%}$, $P_{UD,40\%}$

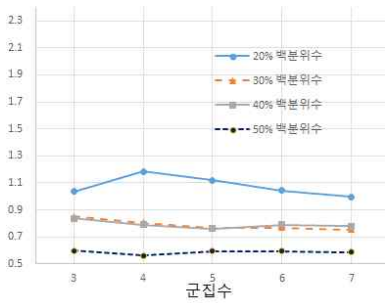
(3)$P_{PC,30\%}$, $P_{UD,30\%}$ (4)$P_{PC,20\%}$, $P_{UD,20\%}$

<그림 6> 각 속성유사도 기준치의 사회연결망에서 거번·뉴먼 알고리즘으로 구한 군집수에 따른 군집유효성지수(DB_K)

다음으로 피어슨상관계수 및 유클리디안 거리를 통해 구성한 각 사회연결망에서 속성유사도의 기준치에 따른 군집유효성지수를 계산하여 <그림 7>과 같이 정리하였다. 충분히 일관적이지는 않으나, 본 논문에서 구성한 사회연결망에서는 대체로 높은 유사도의 연결선이 존재하도록 속성유사도의 기준치를 정한 사회연결망에서 DB_K의 값이 크게 나타난다. 이는 아주 높은 유사도의 연결선으로만 구성되는 사회연결망에서는 오히려 군집유효성이 낮아짐을 의미한다.



(1)피어슨상관계수를 통한 서브그룹분석 결과(DB_K)



(2)유클리디안 거리를 통한 서브그룹분석 결과(DB_K)

<그림 7> 속성유사도 기준치에 따른 사회연결망에서 거번·뉴먼 알고리즘으로 구한 군집수에 따른 군집유효성지수(DB_K)

군집수에 따른 군집유효성을 확인하기 위해 <그림 6>과 <그림 7>에서 각 라인이 나타내는 개별 사회연결망의 군집수에 따른 군집유효성지수를 확인해보면,

다음과 같이 세 가지 경향을 보인다.

(1) 군집수가 많아지면 군집유효성 증가: 피어슨상관계수 $P_{PC,20\%}$/$P_{PC,50\%}$ 등으로 구성한 사회연결망

(2) 군집수에 따라 군집유효성이 오르내림: 피어슨상관계수 $P_{PC,40\%}$ 및 유클리디안 거리 $P_{UD,20\%}$ 등으로 구성한 사회연결망

(3) 군집수에 따른 군집유효성의 변화가 거의 없음: 피어슨상관계수 $P_{PC,30\%}$ 및 유클리디안 거리 $P_{UD,30\%}$/ $P_{UD,40\%}$/ $P_{UD,50\%}$ 등으로 구성한 사회연결망

군집수가 일정 범위에서 증가할수록 군집간의 속성유사도 차이가 커질 가능성이 있고, 이에 따라 군집유효성지수 DB_K가 높아질 가능성은 적다. 본 논문의 경우에서도 군집수의 증가에 따라 군집유효성지수 DB_K가 지속적으로 높아지는 경우는 거의 나타나지 않는다.

본 논문에서 속성유사도로써 구성한 사회연결망에 대해 거번·뉴먼 알고리즘의 군집유효성을 분석한 이상의 결과를 다음의 세 가지로 요약하였다.

(1) 피어슨상관계수보다 유클리디안 거리의 유사도로 구성한 사회연결망에서 군집유효성이 우수한 것으로 나타났다. 유클리디안 거리는 속성데이터의 원래 값을 거리로 산정한 값인데 반하여, -1과 1사이의 값을 유도하기 위한 피어슨상관계수의 경우 필요한 계산과정이 영향을 끼쳤을 것으로 판단된다. 그리고 군집유효성지수인 DB지수의 특성에 기인하는 면도 있을 것으로 추측된다.

(2) 연결선의 유무를 판단하는 속성유사도 기준치가 너무 높으면 군집유효성에는 오히려 좋지 않은 결과를 초래하였다. 이는 군집유효성을 높이기 위해서는 적절한 속성유사도 기준치를 정하는 것이 필요함을 뜻한다.

(3) 군집수의 증가에 따른 군집유효성의 변화에서는, 8가지 사회연결망 중에서 유효성이 개선되는 2가지를 제외하고 나머지는 변화가 거의 없거나 또는 약간의 등락을 보였다. 이는 군집수에 따른 군집유효성이 분석데이터의 특성과 사회연결망에서 활용한 속성유사도의 종류에 따라 다양한 결과를 나타내는 것으로 판단된다.

위에서 요약한 세 가지 결과는 분석대상 데이터의 종류와 특성, 선택한 속성유사도의 종류와 기준치에 따라 다르게 나타날 가능성이 크다. 또한 동일한 결과일지라도 군집결과 평가방식에 따라 유효성의 순위가 일정하지 않을 수 있다. 따라서 분석대상의 특성에 따라 <그림 2>의 선택사항에서 적절한 선택이 필요하거나, 또는 선택사항에 대해 실험을 통한 최선의 발견과정이 필요한 것으로 판단된다.

Ⅴ. 결론 및 토의

일반적인 사회연결망은 개체들을 의미하는 노드와 개체간의 연결특성으로 연결선으로 구성되지만, 본 논문에서는 수치형의 속성데이터를 가지는 개체간의 속성유사도로 계산되는 피어슨상관계수 및 유클리디안 거리를 연결특성으로 활용하여 사회연결망을 구성하였다. 그리고 구성한 사회연결망에 대해 거번·뉴먼 알고리즘의 서브그룹분석을 통한 군집화 과정 및 사례를 제시하였다. 또한 분석한 사회연결망의 연결선은 유사속성의 개체들끼리 연결되는 효과가 있으므로, 서브그룹분석에 의한 군집화 결과에 대해 일반적인 속성기반의 군집유효성지수(CVI)를 적용하여 군집유효성을 평가하는 과정의 제시와 함께 응용사례를 분석하였다.

사회연결망의 구성에 있어서 속성유사도로써 활용할 측정치, 연결선의 존재유무를 판단하는 속성유사도의 기준치, 군집화 결과를 평가할 군집유효성지수 등의 선택에 따른 분석의 과정과 결과를 정리하였다. 필요한 요소의 선택에 따라 사회연결망의 구성, 서브그룹분석에 의한 군집의 결과, 군집결과의 군집유효성 등의 성과에 영향을 미친다는 것을 본 논문에서

확인할 수 있었다.

사회연결망은 최근 여러 분야의 데이터 분석도구로 활용되고 있다. 본 논문에서 정리한 사회연결망의 구성 및 서브그룹분석을 통한 군집화는 이러한 사회연결망의 응용범위를 확대할 수 있다. 일반적인 수치형 속성데이터의 군집분석으로 활용되는 k-평균 등과 달리, 서브그룹분석에 의한 군집화는 분석대상 데이터에 대해 사회연결망의 틀에서 이루어지는 타 분석들과 일관된 관점에서 이루어진다는 장점이 있다. 앞서 언급한 기존의 속성유사도로써 구성한 사회연결망들에서 사회연결망구조를 그대로 유지한 채 타 사회연결망분석과 일관된 시각에서 추가적으로 보완할 수 있는 분석대안이 될 수 있다.

본 논문에서는 속성유사도에 의한 사회연결망의 구성, 거번·뉴먼 알고리즘을 통한 군집분석 과정, 군집결과에 대한 군집유효성 분석 등의 과정을 정리하여 제시하였다. 분석대상 데이터의 종류와 특성, 군집화를 위한 선택사항 등에 따라 분석결과가 상이할 수 있으므로, 이를 고려하여 분석해야 하는 한계가 있다. 그리고 군집화의 성능평가 차원에서 수치형 속성데이터에 대한 본 논문의 방식과 기존의 k-평균 등의 방식을 비교하는 연구가 필요하다.

참고문헌

[1] 박병선·곽기영·김선웅·최흥식, “사회연결망 분석기법을 활용한 기업지배구조와 기업성과 연구,” 경영과학, 제29권, 제2호, 2012, pp.167-183.

[2] 김진광·윤성웅·이상훈, “사회연결망분석 이용 전문가탐색,” 한국컴퓨터정보학회 동계학술대회 논문집, 제27권, 제1호, 2019, pp.171-174.

[3] 조윤호·방정혜, “신상품추천을 위한 사회연결망 분석의 활용,” 지능정보연구, 제15권, 제4호, 2009, pp.183-200.

[4] 딘티퀸·윤한성, “사회연결망분석을 통한 지배적 제품믹스 및 판매점식별: 체인소매점을 대상으로,” 산업혁신연구, 제32권, 제4호, 2016, pp.83-100.

[5] 김창림·윤한성, “사회연결망분석을 활용한 인력배치의 적절성 평가: 성격특성과 직무만족도를 중심으로,” 인터넷전자상거래연구, 제15권, 제6호, 2015, pp.117-128.

[6] Borgatti, S.P., Everett, M.G. and Johnson, J.C., “Analyzing Social Network,” Thousand Oaks, California: Sage Publication Inc., 2013.

[7] Xu R. et al., “Survey of clustering algorithms,” IEEE Trans. on Neural Network, Vol. 16, No. 3, 2005, pp.645-678.

[8] Barnes, J., “Class and Committees in a Norwegian Island Parish,” Human Relations, Vol. 7, 1954, pp.39-58.

[9] Freeman, L.C., “Networks of Innovations: A Synthesis of Research Issues,” Research Policy, Vol. 20, No. 5, 1991, pp.499-514.

[10] Chang, Victor, “A Proposed Social Network Analysis Platform for Big Data Analytics,” Technological Forecasting & Social Change, Vol. 130, 2018, pp.57-68.

[11] 김정숙, “빅 데이터 활용과 관련기술 고찰,” 한국콘텐츠학회지, 제10권, 제1호, 2012, pp.34-40.

[12] 곽기영, 소셜네트워크분석, 청람, 2014.

[13] Freeman, L., “A set of measures of centrality based upon betweenness,” Sociometry, Vol. 40, 1977, pp.35-41.

[14] Rosvall, M. and Bergstrom, C.T., “Maps of random walks on complex networks reveal community structure,” Proc. Natl. Acad. Sci. USA, Vol. 105, No. 4, 2008, pp.1118-1123.

[15] Wu, F. and Huberman, B.A., “Finding

communities in linear time: A physics approach," Eur. Phys. J., Vol. B, No. 38, 2004, pp.331-338.

[16] Zhao, P. and Zhang C.Q., "A new clustering method and its application in social networks," Pattern Recognition Letters, No. 32, 2011, pp.2109-2118

[17] Girvan, M. and Newman, M.E.J., "Community structure in social and biological networks," Proc. Natl. Acad. Sci., Vol. 99, No. 12, 2002. pp.7821-7826.

[18] 이수현·정영선·김재윤, "경영사례를 이용한 군집화 유효성 지수의 성능비교," 한국경영과학회지, 제41권, 제2호, 2016, pp.17-33.

[19] 최슬비·곽기영·안현철, "사용자 간 신뢰관계 네트워크 분석을 활용한 협업 필터링 알고리즘의 예측 정확도 개선," 지능정보연구, 제22권, 제3호, 2016, pp.113-127.

[20] 이현진·지태창, "추천시스템을 위한 연관군집 최적화 기반 협력적 필터링 방법," 디지털산업정보학회논문지, 제6권, 제3호, 2010, pp.19-29.

[21] 김성림·권준희, "상황인식 정보 검색 기법을 이용한 하이브리드 협업 필터링 기법," 디지털산업정보학회논문지, 제6권, 제1호, 2010, pp. 143-149.

논 문 접 수 일 : 2021년 1월 26일
수 정 일(1차) : 2021년 2월 17일
수 정 일(2차) : 2021년 3월 2일
게 재 확 정 일 : 2021년 3월 3일

■ 저자소개 ■

윤 한 성
Yoon, Han Seon

2001년 3월~현재 경상대학교 경영대학 교수
1998년 8월 한국과학기술원 테크노경영대학원(공학박사)
1987년 8월 한국과학기술원 산업공학과(공학석사)
1985년 2월 서울대학교 산업공학과(공학사)

관심분야 : e비즈니스, SCM, 데이터분석 등
E-mail : hsyun@gnu.ac.kr