Journal of the Korean Data Analysis Society (August 2018)
Vol. 20, No. 4, pp. 1769-1780

연속형과 범주형 변수가 혼합된 데이터의 군집분석 연구*

한지수[1], 조형준[2]

요 약

연속형 변수와 범주형 변수가 함께 존재하는 혼합형 데이터의 군집분석에서 연속형 변수의 경우에 유클리디안 거리처럼 자연스럽게 거리를 정의할 수 있지만, 범주형 변수, 특히 순서가 없는 명목형 변수의 경우에 개체 간 거리 측정이 모호하여 종종 배제되었다. 개체 간 거리를 기반으로 하는 군집분석 방법에서 개체 간 거리의 정의는 매우 핵심적인 문제이므로 명목형 변수의 합리적 정의는 신뢰할 수 있는 군집분석 위해서는 중요한 요소이다. 따라서 두 가지 형태의 변수가 함께 존재하는 경우에 범주형 변수의 거리 측정을 위해 모형화를 위해 종종 이용되는 가변수 변환 방법, 범주의 일치 여부에 따라 0-1로 거리를 부여하는 Gower의 방법, 그리고 변수의 수준 개수 정보를 활용하는 Eskin의 방법 도입하여 혼합형 데이터에 거리를 측정할 수 있도록 군집 분석 결과를 비교하였다. 거리 정의 방법에 의존하지 않는 공정한 비교를 위해 세 가지 평가 측도를 이용하였다. 그 결과, 군집의 퍼짐 정도 및 군집 별 개체의 불균형한 상황에서 Eskin의 방법의 성능이 우수하였다. 또한, 군집의 응집성, 재현성, 군집의 개수 정확도 측면에서 Eskin의 방법이 더 나은 성능을 보였다.

주요용어 : 명목형 변수, 혼합형 자료, 거리 측정, 계층적 군집분석.

1. 서론

군집분석이란 특성(패턴)을 가진 개체들을 유사한 정도를 기반으로 하여 조직화(군집화)하는 분석이다. 비슷한 특성을 가진 개체들을 동일한 군집으로, 서로 다른 특성을 가진 개체들을 다른 군집에 속하게 하는 것을 목표로 한다. 군집분석의 갈래 중 개체 간 거리를 기반으로 하는 대표적인 군집분석 기법으로 계층적 군집분석과 K-means 군집분석으로 대표되는 분할형 군집분석이 있다. 그런데 이러한 군집분석 방법에서 두 개의 서로 다른 변수로 측정된 개체 간 거리의 정의는 매우 핵심적인 문제임에도 불구하고 종종 군집분석 알고리즘 연구에서 별개의 차원으로 인식되었다. 특히 거리를 기반으로 하는 여러 가지 알고리즘에서 거리 계산 절차를 알고리즘과 종종 독립된 단계로 간주라며, 거리 측정 방법에 대해서는 적은 비중으로 다루고 있다(Park, Cho 2006; Boriah et al., 2008; Song, 2017). 많은 수의 군집분석 방법에 대한 연구는 대부분 연속형 변수로 이루어진 자료를 중심으로 연구되었고, 특히 순서가 없는 범주형 변수인 명목형 변수(nominal variable)가 포함되어 있는 혼합형 데이터 대한 군집분석 방법에 대한 연구는 상대적으로 부족하다(Lee et al., 2012; Chang et al., 2014; Park et al., 2016).

본 연구에서는 연속형 변수와 범주형 변수가 함께 존재하는 혼합형 데이터에 보편적으로 사용되는 거리 측정 방법에 대해 살펴본 후, 범주형 변수가 가지는 특징을 고려하여 개발 된 개체 간

*이 논문은 2015년 한국연구재단(NRF-2015R1D1A1A09058602)지원을 받아 수행되었으며, 제1저자 한지수의 석사학논문(Han, 2018)의 축약본입니다.
[1]02841 서울특별시 성북구 안암로 145, 고려대학교 통계학과 석사졸업. E-mail : hjisu0316@korea.ac.kr
[2](교신저자) 02841 서울특별시 성북구 안암로 145, 고려대학교 통계학과 교수. E-mail : hj4cho@korea.ac.kr
[접수 2018년 7월 20일; 수정 2018년 8월 10일, 2018년 8월 17일; 게재확정 2018년 8월 20일]

거리 측정방법을 이용하여 이를 연속형 변수와 범주형 변수들이 함께 있는 혼합형 자료에 대하여 적절히 도입하고, 군집분석을 실시 및 그 결과를 이용해 거리 측정 방법 간 유용성을 비교하고자 한다.

2. 연구방법

연속형 변수로 이루어진 두 개체 간 거리 측정 방법은 이론 및 수리적으로 이미 잘 정리되어 있으며 유클리디안 거리(Euclidean distance) 측정방법이 보편적으로 사용되고 있다. Gower(1971)는 혼합형 자료에 대하여 거리 측정방법을 제안하였다. 그 중 범주형 변수의 두 값 간의 거리의 계산 방법을 이용하였다.

Eskin(2002) 기존의 데이터를 새로운 형상 공간(feature space)에 사상(mapping)하여 비지도학습(unsupervised learning)문제에서 이상치 탐지 아이디어를 제안하면서 범주형 변수를 새로운 형상공간에 사상하였다.

이 2가지 방식과 가변수 방식을 이용한 군집분석을 고려하고 비교 분석을 위해 평가 측도로서 3가지 방법을 이용한다.

2.1. 혼합형 데이터에 대한 군집분석 방법

연속형 변수 N개와 범주형 변수 C개 총 $K(=N+C)$개의 변수로 측정된, M개의 개체로 이루어진 행렬 $X=[x_{ik}]=[x_{in}|x_{ic}]$, $i=1,\cdots,M$, $k=1,\cdots,K$, $n=1,\cdots,N$, $c=1,\cdots,C$가 존재할 때, 임의의 두 개체 $\boldsymbol{x}_a=[x_{a1},\cdots,x_{aK}]$, $\boldsymbol{x}_b=[x_{b1},\cdots,x_{bK}]$ 간 거리의 정의는 다음과 같다.

방법 1. Dummy

Step 1. (연속형 변수)
임의의 연속형 변수 $\boldsymbol{x}_n$, $\boldsymbol{x}_n^T=[x_{1n},\cdots,x_{Mn}]$에 대하여, 최솟값이 0, 최댓값이 1이 되도록 정규화 한다($n=1,\cdots,N$). 이를 통해 정규화된 값을 가지는 연속형 변수를 $\tilde{\boldsymbol{x}}_n^T=[\tilde{x}_{1n},\cdots,\tilde{x}_{Mn}]$라 한다.

Step 2. (범주형 변수)
g_c개의 수준을 갖는 임의의 범주형 변수 $\boldsymbol{x}_c^T=[x_{1c},\cdots,x_{Mc}]$에 대해, 해당 수준 값 발생 시 1 그 외의 값은 0이 되도록 g_c개의 가변수 $Z_{ct}^T=[z_{1ct},\cdots,z_{Mct}]$를 생성한다($c=1,\cdots,C$, $t=1,\cdots,g_c$).

Step 3. (최종거리)
임의의 두 개체 $\boldsymbol{x}_a,\boldsymbol{x}_b$ 간 거리는 다음의 식을 통해 최종 계산된다.

$$D_1(\boldsymbol{x}_a,\boldsymbol{x}_b)=\sqrt{\sum_{n=1}^{N}(\tilde{x}_{an}-\tilde{x}_{bn})^2+\sum_{c=1}^{C}\sum_{t=1}^{g_c}(z_{act}-z_{bct})^2}$$

Dummy 방법은 범주형 변수를 가변수화 하였고, 각 연속형 변수는 최솟값을 0, 최댓값을 1로 정규화하여 거리를 계산하였다. 이를 통해 모든 변수가 가지는 거리의 범위를 동일하게 하여 개체 간 거리계산에 변수들이 균등한 영향을 주도록 하였다. 최종적으로 유클리디안 거리(Euclidean distance) 측정 방식을 취한다.

방법 2. Gower

Step 1. (연속형 변수)
임의의 연속형 변수 $\boldsymbol{x}_n$, $\boldsymbol{x}_n^T = [x_{1n},\cdots,x_{Mn}]$ 대하여, 최솟값이 0, 최댓값이 1이 되도록 정규화 한다. ($n = 1,\cdots,N$) 이를 통해 정규화 된 값을 가지는 연속형 변수를 $\tilde{\boldsymbol{x}}_n^T = [\tilde{x}_{1n},\cdots,\tilde{x}_{Mn}]$라 한다.

Step 2. (범주형 변수)
임의의 범주형 변수 $\boldsymbol{x}_c^T = [x_{1c},\cdots,x_{Mc}]$에 대하여, 두 개체간 거리는 다음과 같이 계산한다.

$$d_g(x_{ac}, x_{bc}) = \begin{cases} 0, & x_{ac} = x_{bc} \\ 1, & x_{ac} \neq x_{bc} \end{cases}$$

Step 3. (최종거리)
임의의 두 개체 $\boldsymbol{x}_a, \boldsymbol{x}_b$ 간 거리는 다음의 식을 통해 최종 계산된다.

$$D_2(\boldsymbol{x}_a, \boldsymbol{x}_b) = \sum_{n=1}^{N} |\tilde{x}_{an} - \tilde{x}_{bn}| + \sum_{c=1}^{C} d_g(x_{ac}, x_{bc})$$

Gower의 방법은 범주형 변수에 대하여 단순 대응(simple matching) 방법을, 각 연속형 변수는 최솟값을 0, 최댓값을 1로 정규화하여 거리를 계산하였다. 이를 통해 모든 변수가 가지는 거리의 범위를 동일하게 하여 개체 간 거리계산에 변수들이 균등한 영향을 주도록 하였다. 최종적으로 맨하탄 거리(Manhattan distance) 측도를 취한다.

방법 3. Eskin

Step 1. (연속형 변수)
임의의 연속형 변수 $\boldsymbol{x}_n$, $\boldsymbol{x}_n^T = [x_{1n},\cdots,x_{Mn}]$ 대하여, 최솟값이 0, 최댓값이 1/2이 되도록 정규화 한다. ($n = 1,\cdots,N$) 이를 통해 정규화 된 값을 가지는 연속형 변수를 $\tilde{\boldsymbol{x}}_n^T = [\tilde{x}_{1n},\cdots,\tilde{x}_{Mn}]$ 라 한다.

Step 2. (범주형 변수)
임의의 범주형 변수 $\boldsymbol{x}_c^T = [x_{1c},\cdots,x_{Mc}]$에 대하여, 두 개체간 거리는 다음과 같이 계산한다.

$$d_e(x_{ac}, x_{bc}) = \begin{cases} 0 & x_{ac} = x_{bc} \\ \dfrac{2}{g_c^2} & x_{ac} \neq x_{bc} \end{cases}$$

Step 3. (최종거리)
임의의 두 개체 $\boldsymbol{x}_a, \boldsymbol{x}_b$ 간 거리는 다음의 식을 통해 최종 계산된다.

$$D_3(\boldsymbol{x}_a, \boldsymbol{x}_b) = \sum_{n=1}^{N} |\tilde{x}_{an} - \tilde{x}_{bn}| + \sum_{c=1}^{C} d_e(x_{ac}, x_{bc})$$

Eskin의 방법은 Eskin 거리 측정 방법이 두 개의 수준을 가질 때의 거리 최댓값을 가지며 그 값이 1/2임을 감안하여, 연속형 변수의 거리의 범위를 최솟값이 0, 최댓값이 1/2이 되도록 정규화하

여 거리를 계산하였다. 이를 통해 모든 변수가 가지는 거리의 범위를 동일하게 하여 개체 간 거리 계산에 변수들이 균등한 영향을 주도록 하였다. 최종적으로 맨하탄 거리(Manhattan distance) 측도를 취한다.

세 가지 방법을 이용하여 4번째 단계(Step 4)로 다음의 $M\times M$행렬의 거리행렬을 계산한다.

$$D_{M\times M}=\begin{bmatrix} D_j(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & D_j(\boldsymbol{x}_1,\boldsymbol{x}_M) \\ \vdots & \ddots & \vdots \\ D_j(\boldsymbol{x}_M,\boldsymbol{x}_1) & \cdots & D_j(\boldsymbol{x}_M,\boldsymbol{x}_M) \end{bmatrix}$$

여기서 j는 위 3가지의 방법 1, 2, 또는 3이다. 마지막 단계(Step 5)로 거리 행렬을 이용하여 계층적 군집분석 (hierarchical cluster analysis) 알고리즘을 수행한다.

2.2. 군집분석 평가 측도

군집 분석 결과의 평가는 군집의 응집 정도에 대한 평가를 통해 군집의 개수 결정에 도움을 주는 내적 유효성 측도(internal validity indices)와 두 군집 예측 결과를 비교하여 군집의 재현성 또는 일치성을 평가하는 외적 유효성 측도(external validity indices)를 이용한다. 여기서 내적 유효성 측도로는 Silhouette 지수, Dunn 지수, 외적 유효성측도로 수정 Rand 지수를 고려하였다.

1) Silhouette 지수

Rousseeuw(1990)은 Silhouette 지수를 제안하고 Silhouette 값을 근거로 최소가 되는 군집 개수를 선택하도록 하였다. 각 개체의 Silhouette 값은 다음과 같이 정의된다.

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}=\begin{cases} 1-\dfrac{a(i)}{b(i)} & , \quad a(i)<b(i) \\ 0 & , \quad a(i)=b(i) \\ \dfrac{b(i)}{a(i)}-1 & , \quad a(i)>b(i) \end{cases}$$

여기서 a_i는 개체 i가 속한 군집의 모든 개체들과 개체 i와의 평균거리이고, b_i는 개체 i가 속하지 않은 군집의 모든 개체들과 개체 i와의 평균 거리이다.

$$Silhouette=\frac{\sum_{i=1}^{n}S(i)}{n},\ Silhouette\in[-1,1]$$

최종적으로 Silhouette 지수는 각 개체의 Silhouette 값의 평균으로 계산되며 Silhouette 값은 개체가 적절한 군집에 배치되었는지 측정하는 신뢰도의 개념이라 할 수 있다. Silhouette 값은 각 개체가 적절한 군집에 배치될수록 1에 가깝고 그렇지 않을수록 -1에 가깝다. 따라서 Silhouette 지수가 최대일 때의 군집의 수를 최적의 군집 개수로 판단한다.

2) Dunn 지수

Dunn 지수(Dunn, 1974)는 같은 군집에 속해 있는 두 개체 간의 가장 큰 거리에 대한 서로 다른 군집에 속해 있는 두 개체간의 가장 작은 거리의 비(ratio)로써 나타나며 다음과 같이 계산한다.

$$Dunn = \frac{\min\limits_{1 \le i < j \le q} d(C_i, C_j)}{\max\limits_{1 \le k \le q} diam(C_k)}$$

여기서 $d(C_i, C_j)$는 군집 C_i와 C_j내 속해 있는 개체 간 가장 작은 거리를 나타내고, $diam(C_k)$는 군집 C_k에서 가장 큰 두 개체간 거리를 나타낸다. 같은 군집에 속해 있는 두 개체간의 거리가 작을수록, 다른 군집에 속해 있는 두 개체간 거리가 클수록 Dunn 지수는 커지므로, Dunn 지수가 최대일 때 군집의 개수를 최적의 군집 개수로 판단한다.

3) 수정 Rand 지수

Rand 지수(Rand, 1971)는 외적 측도의 하나로 동일한 데이터에 대해 두 군집 방법 결과의 일치성 또는 재현성을 평가한다. Rand 지수는 다음과 같이 계산한다. 총 개체수를 n이라 할 때, 두 군집방법에 의해 각각 k개의 군집결과가 산출된다. 군집방법 1에 의한 결과를 행으로, 군집방법 2에 의한 결과를 열로 하여 교차표를 만들 수 있다. 행 i와 열 j로 분류된 개체 수를 $n_{ij}(i, j = 1, \cdots, k)$라고 할 때, 모든 n_{ij}의 합은 총 개체 수 n이다. 이 중 임의의 두 개체 a, b에 대해 군집방법 U와 또 다른 군집방법 V에 따라 분류하고, 분류결과는 다음의 4가지 경우가 발생된다.

Table 1. Cross classification of a pair of objects

Method U \ Method V	Same Cluster	Different Cluster	Total
Same Cluster	N_1	N_2	$N_1 + N_2$
Different Cluster	N_3	N_4	$N_3 + N_4$
Total	$N_1 + N_3$	$N_2 + N_4$	$\binom{n}{2}$

임의의 두 개체로 만들 수 있는 모든 쌍은 $\binom{n}{2}$개이며, 이 중에서 N_1 - N_4에 해당하는 쌍의 개수는 $N_1 = \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2}$, $N_2 = \sum_{i=1}^{k}\binom{n_{i.}}{2} - \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2}$, $N_3 = \sum_{i=1}^{k}\binom{n_{.j}}{2} - \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2}$, $N_4 = \binom{n}{2} - \sum_{i=1}^{k}\binom{n_{i.}}{2} - \sum_{i=1}^{k}\binom{n_{.j}}{2} + \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2}$로 정의된다. 따라서 두 군집 결과가 동일한 모든 쌍의 수는 $N_1 + N_4$이며, 불일치하는 모든 쌍의 수는 $N_2 + N_3$이다. Rand 지수는 가능한 총 개체의 쌍 중 일치하는 쌍의 비율로 정의된다.

$$Rand = \frac{N_1 + N_4}{\binom{n}{2}}$$

Rand 지수는 0에서 1사이의 값을 갖고, 이 값이 0에 가까울수록 군집결과간의 일치 정도가 없고, 1에 가까울수록 두 결과가 서로 일치한다는 것을 의미한다. 그런데 우연하게 일치하는 개체 쌍이 다수 나올 수 있음을 감안하여 Hubert, Arabie(1985)는 다음과 같이 Rand 지수를 수정하였다.

$$Adjusted\,Rand = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\binom{n_{ij}}{2} - \sum_{i=1}^{K}\binom{n_{i.}}{2}\sum_{i=1}^{K}\binom{n_{.j}}{2}/\binom{n}{2}}{[\sum_{i=1}^{K}\binom{n_{i.}}{2} + \sum_{i=1}^{K}\binom{n_{.j}}{2}]/2 - \sum_{i=1}^{K}\binom{n_{i.}}{2}\sum_{i=1}^{K}\binom{n_{.j}}{2}/\binom{n}{2}}$$

이 같은 수정은 우연히 일치 쌍 개수의 기댓값을 감안한 것이다. 즉 Table 1의 실제 값이 우연적인 일치쌍의 기댓값과 일치한다면, 다시 말해 두 군집방법 결과가 서로 무관하다면 수정 Rand 지수 값이 0이 나오도록 보정한다. 따라서 두 군집결과가 일치하는 경우 1, 무관한 경우는 0을 산출하며, 기댓값보다 작은 경우 음수를 산출한다.

3. 모의실험

연속형 변수와 범주형 변수가 함께 존재할 때 거리 계산 및 군집 성능 평가를 위해 연속형 변수와 범주형 변수를 각각 1개씩 고려하였다. 변수의 생성과정은 다음과 같다. 각 군집 별로 균등분포에서 2개의 변수 X_1, X_2를 독립적으로 추출하여 군집데이터를 생성한다. 추출된 변수 X_1을 구간별로 분할하여 범주형 변수로 변환하여 관측 된 변수 Z_1을 생성하고 X_2를 관측된 연속형 변수 Z_2로 간주한다. X_1의 분할은 고려하는 경우(case)에 따라 다양하게 구성하였다. 그 후 관측된 범주형 변수 Z_1과 연속형 변수 Z_2를 이용하여 계층적 군집분석을 실시한 후 실제 변수 X_1, X_2와 사전 설정한 실제 군집결과를 이용하여 군집결과를 비교하였다.

내적 유효성 척도 2가지(Silhouette 지수, Dunn 지수)는 거리 측정 방식에 따라 그 값이 결정되므로 측정 방법 간 공정한 비교를 위해 실제 변수 X_1, X_2를 유클리디안 거리(Euclidean distance) 측도를 이용하여 군집의 내적 응집성 평가를 하였다. 또한 외적 유효성 척도를 통해 군집 예측 결과와 사전 설정한 실제 군집 결과를 비교하여 군집 결과의 재현성을 평가하였다. 아울러, 방법 간 평균 군집 개수를 비교하여 군집 개수 정확도 평가를 실시하였다. 계층적 군집분석에서 군집 간 거리를 계산하는 방법은 평균 연결측도(average linkage) 방법을 적용하였으며, 2개에서 10개까지를 군집의 개수후보로 설정하였다. 모의실험 결과는 200회의 독립 시행을 통해 산출된 값을 이용하였다.

Case 1: 각 군집이 각 변수 내에서 겹치지 않는 경우

첫 번째 경우에서는 변수 X_1과 X_2의 값이 각 군집을 명확히 설명하도록 설정하였다. 이러한 설정 하에서 Table 2와 같이 분할점을 가진 범주형 변수가 관측되도록 하였으며, Table 3의 결과를 얻었다. X_1의 값이 서로 다른 군집을 측정하는 경우, 방법과 수준 별 분할지점 및 개수에 관계없이 동일한 군집 결과를 보였다. 군집 예측 결과가 세 방법에서 모두 같아 군집 내적 응집성 측도들의 값과 외적 측도의 값이 각 수준의 개수 별로 동일한 값을 산출하였다. 군집 외적 응집성 측도인 수정 Rand 지수 값이 1이므로, 세 방법 모두 정확히 실제 군집의 결과와 정확히 일치하였으며, 군집의 개수도 정확히 3개로 예측하였다. 눈여겨 볼 점은, 범주가 2개인 경우 하나의 수준이 2개의 군집을 설명하고 나머지 하나의 수준이 1개의 군집을 설명하고 있다. 하나의 수준이 두 군집을 설명함에도 불구하고 각 군집을 설명하는 수준의 수가 3개인 경우와 군집 분석 결과가 같았다.

Table 2. Case 1 : Distribution of variables and information of partitions

		Variable		Number of levels				
		X_1	X_2		2	3	4	6
Cluster	C_1	$Unif(1,2)$	$Unif(2,3)$	X_1 Partitions	(1,2) (3,5)	(1,2) (3,4) (5,6)	(1,1.5) (1.5,2) (3,4) (4,5)	(1,1.5) (1.5,2) (3,3.5) (3.5,4) (4,4.5) (4.5,5)
	C_2	$Unif(3,4)$	$Unif(4,5)$					
	C_3	$Unif(4,5)$	$Unif(1,2)$					

Table 3. Case 1: Cluster analysis results

		Number of levels			
		2	3	4	6
Silhouette Index	Dummy	0.8	0.8	0.8	0.8
	Gower	0.8	0.8	0.8	0.8
	Eskin	0.8	0.8	0.8	0.8
Dunn Index	Dummy	1.2	1.2	1.2	1.2
	Gower	1.2	1.2	1.2	1.2
	Eskin	1.2	1.2	1.2	1.2
Adj.Rand.Index	Dummy	1.0	1.0	1.0	1.0
	Gower	1.0	1.0	1.0	1.0
	Eskin	1.0	1.0	1.0	1.0
# of Clusters (Silhouette)	Dummy	3.0	3.0	3.0	3.0
	Gower	3.0	3.0	3.0	3.0
	Eskin	3.0	3.0	3.0	3.0
# of Clusters (Dunn)	Dummy	3.0	3.0	3.0	3.0
	Gower	3.0	3.0	3.0	3.0
	Eskin	3.0	3.0	3.0	3.0

Case 2: **각 군집이 각 변수 내에서 겹치며, 군집 별 개체의 수가 불균형인 경우**

두 번째 경우에서는 세 번째 Case에서 군집 C_1, C_2, C_3의 개체수가 각각 100, 200, 300개로 설정하여 군집 별 개체수가 불균형이 되도록 설정하였다. Table 4와 같이 분할점을 가진 범주형 변수가 관측되도록 하였으며 Table 5의 결과를 얻었다. 군집 별 개체수를 불균형하게 유지하였음에도 방법 간 군집 결과의 전반적인 경향은 비슷하였다. 즉 Eskin의 방법이 나머지 두 방법에 비해 수준의 개수가 3, 5 일 때 군집의 응집성, 재현성, 군집의 개수 측면에서 성능이 매우 좋았다.

Case 3: **각 군집이 변수들 간에 겹치며 경우**

세 번째 Case에서는 군집이 2개이며 두 군집이 두 변수에서 모두 겹치도록 만들었다. 먼저 Table 6과 같이 X_1, X_2를 생성한 후 다음의 회전행렬(rotation matrix) R을 이용하여 새로운 변수 V_1, V_2를 얻었다.

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = R \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \; R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \; \theta = 315^\circ$$

새롭게 생성된 실제 변수 V_1, V_2를 이용하여 분할점을 가진 범주형 변수가 관측되도록 하였고, Table 7의 결과를 얻었다. 수준의 개수가 2 또는 3일 때에 세 방법의 결과는 대체로 비슷하였으나

Table 4. Case 2: Distribution of variables and information of partitions

		Variable		Number of levels		
		X_1	X_2	2	3	5
Cluster	C_1	$Unif(1,5)$	$Unif(1,2)$	X_1 Partitions (1,6) (6,11)	(1,5) (5,7) (7,11)	(1,3) (3,5) (5,7) (7,9) (9,11)
	C_2	$Unif(4,8)$	$Unif(5,6)$			
	C_3	$Unif(7,11)$	$Unif(1,2)$			

Table 5. Case2: Cluster analysis results

		Number of levels		
		2	3	5
Silhouette Index	Dummy	0.352	0.447	0.296
	Gower	0.350	0.446	0.300
	Eskin	0.347	0.688	0.684
Dunn Index	Dummy	0.0069	0.0103	0.0062
	Gower	0.0074	0.0099	0.0059
	Eskin	0.0072	0.5099	0.5116
Adj.Rand.Index	Dummy	0.564	0.629	0.439
	Gower	0.570	0.626	0.448
	Eskin	0.567	1.000	1.000
# of Clusters (Silhouette)	Dummy	2.60	5.00	7.00
	Gower	2.44	5.00	7.00
	Eskin	2.56	3.00	3.00
# of Clusters (Dunn)	Dummy	6.10	5.28	7.96
	Gower	6.14	5.84	7.40
	Eskin	6.56	3.00	3.00

수준의 개수가 5일 때 군집의 내적 측도 그리고 재현성 측면에서 그 외의 방법에 비해 좋은 성능을 보였다.

Table 6. Case 3: Distribution of variables and information of partitions

		Variable		Number of levels			
		X_1	X_2		2	3	5
Cluster	C_1	$Unif(2,9)$	$Unif(7,9)$	V_1 Partitions	$(-\infty, 0)$ $(0, \infty)$	$(-\infty, -1.5)$ $(-1.5, 1.5)$ $(1.5, \infty)$	$(-\infty, -3)$ $(-3, -1)$ $(1, 3)$ $(3, \infty)$
	C_2	$Unif(2,9)$	$Unif(1,3)$				

Table 7. Case 3: Cluster analysis results

		Number of levels		
		2	3	5
Silhouette Index	Dummy	0.503	0.105	0.051
	Gower	0.503	0.105	0.053
	Eskin	0.503	0.100	0.200
Dunn Index	Dummy	0.027	0.013	0.013
	Gower	0.027	0.013	0.013
	Eskin	0.027	0.015	0.020
Adj.Rand.Index	Dummy	0.667	0.019	0.017
	Gower	0.667	0.019	0.018
	Eskin	0.667	0.019	0.111
# of Clusters (Silhouette)	Dummy	2.0	4.2	6.0
	Gower	2.0	4.2	6.0
	Eskin	2.0	4.2	6.3
# of Clusters (Dunn)	Dummy	7.2	7.9	9.2
	Gower	7.2	7.9	9.2
	Eskin	7.2	7.7	8.6

모의실험을 통해 얻은 결론은 다음과 같다. 군집이 실제 각 변수들에 의해 명확히 설명된다면 세 가지 방법 간 군집 결과는 비슷하며, 모두 군집을 정확히 설명한다. 그러나 범주형 변수로 관찰되는 실제 변수가 각 군집을 명확히 설명하지 못하는, 즉 해당 변수가 군집에 대해 부정확한 정보를 제공하는 경우 Eskin의 거리 측정방법이 다른 방법에 비해 더 좋은 성능을 보였다. 이는 상대적으로 해당 변수에 대해 작은 거리를 부여하므로 거리 계산에서 그 영향이 축소되었기 때문이다. 특히 이러한 경우, Eskin의 거리 측정방법의 우수한 결과가 이를 뒷받침한다. 범주형 변수로 변환이 되는 변수가 각 군집을 명확히 설명하지 못하면서 군집이 길쭉한 경우에도 Eskin의 방법이 다른 두 방법에 비해 결과가 좋았으며, 불균형한 경우에도 좋은 성능을 보였다. 아울러, 실제 두 변수가 모두 군집에 대해 부정확한 정보를 가지고 있는 경우에도 Eskin의 방법이 비교적 좋은 결과를 보였다.

4. 사례분석

앞서 비교한 세 가지 거리측정 방법을 이용한 군집분석 방법을 실제 데이터에 적용하여 그 결과를 확인하고자 한다. 데이터는 도매 회사의 고객 데이터(wholesale customers dataset, 2014)를 이용하였다.

이 자료는 어떤 도매 회사의 440명의 고객(회사)을 대상으로 6가지 제품 별 고객의 연간 지출액과 해당 고객의 정보 2가지 항목, 총 8개의 변수로 구성된다. 6가지 제품은 각각 신선제품(fresh), 유제품(milk), 식료품(grocery), 냉동식품(frozen), 세제 및 휴지(detergent paper), 조리식품(delicatessen)이며 각각 제품에 대한 고객의 연간 지출액이 측정되었다, 고객의 정보로 유통 경로(channel)와 지역(region)을 얻을 수 있으며 유통경로 변수는 각각 외식산업 (hotel/restaurant/cafe)과 소매(retail)로 2개의 수준을, 지역 변수는 리스본(Lisbon), 포르투(Porto), 그 외의 지역(other region)으로 총 3개의 수준을 가지며 두 변수는 명목형 변수이다.

Table 8. Cluster analysis by Gower's method

Channel											
Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
H.R.C	Retail	H.R.C	Retail	H.R.C	Retail	H.R.C	Retail	H.R.C	Retail	H.R.C	Retail
0	105	211	0	59	0	0	18	0	19	28	0

Region																	
Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5			Cluster 6		
L	P	O	L	P	O	L	P	O	L	P	O	L	P	O	L	P	O
0	0	105	0	0	211	59	0	0	18	0	0	0	19	0	0	28	0

Table 9. Cluster analysis by Eskin's method

Channel				Region					
Cluster 1		Cluster 2		Cluster 1			Cluster 2		
H.R.C	Retail	H.R.C	Retail	Lisbon	Porto	Other	Lisbon	Porto	Other
0	142	298	0	18	19	105	59	28	211

6개의 연속형 변수와 2개의 명목형 변수로 구성된 데이터에 대하여 세 가지 거리 측정 방법 별로 각각 계층적 군집분석을 실시하였다. 여기서, 군집 예측 결과의 참값을 알 수 없으므로 수정 Rand 지수를 이용하여 성능의 직접적인 비교는 어렵다. 대신, 군집 예측 결과를 비교하여 좀 더 타당한 방법에 대해 논의하였다. 군집 간의 거리 측정 방법으로 평균연결법(average linkage) 방법을

고려하였고, 군집 개수로 2~10개를 후보로 하였으며, 군집의 개수 예측 측도로 앞서 살펴본 Silhouette 지수 및 Dunn 지수를 이용하였다.

'Dummy' 및 'Gower'의 방법을 이용하여 군집분석 결과 군집의 개수로 6개, 'Eskin'의 방법은 2개가 두 측도에서 동일하게 선택되었다. 여기서 Dummy 방법 및 Gower의 방법은 군집의 예측 결과가 동일하였으므로 Gower의 결과와 Eskin의 군집의 예측결과만을 비교한다. 각 방법에 따라 군집분석을 실시하였고 이에 따른 군집의 특성을 정리하였다.

Gower의 방법의 결과 6개의 군집의 특징을 살펴보았을 때 '지역' 변수를 제외한 모든 변수에서 군집1, 4, 5가 그리고 군집 2, 3, 6이 비슷한 경향을 보였다. 최종적으로 군집이 6개로 선택된 이유는 '지역' 변수의 각 수준에 따라 군집이 분할되었기 때문이다. 반면, Eskin의 방법은 최종적으로 2개의 군집을 선택하였는데 이는 Gower의 방법의 결과에서 비슷한 경향을 보이는 군집을 병합한 결과와 같다. Gower의 방법이 명목형 변수가 가지는 수준의 개수를 곱한 만큼을 최종 군집의 개수로 결정하였음을 살펴볼 때, 군집의 개수를 과도하게 계산할 것이라 추론할 수 있다. 또한 이에 따른 군집 예측 결과도 신뢰하기 어려워진다.

이에 반해, Eskin의 방법은 명목형 변수가 가지는 거리를 조정해주어 적절한 군집의 개수를 계산 및 예측하였다. 실제로, 군집분석이 유사한 특성을 가지는 개체들을 적절한 수로 묶는 방법임을 고려할 때, '지역' 변수의 각 수준 간 거리를 적절히 조정해주어 유사한 특성을 가지는 개체는 하나의 군집이 되도록 한 Eskin의 방법이 성능이 Dummy 및 Gower의 방법에 비해 성능이 좋았다고 결론 내릴 수 있다. Eskin의 방법의 군집분석 결과에서 군집의 개수를 6개로 설정하였을 때 다른 두 방법의 군집 예측 결과와 동일하였다. 즉 적절한 거리 계산을 통해 정확한 군집의 개수 및 이에 따른 타당한 군집 예측 결과를 얻을 수 있다.

5. 결론

본 논문은 연속형 변수와 범주형 변수가 함께 존재하는 혼합형 데이터에 대하여 개체 간 거리의 적절한 측정 및 이를 이용하여 군집분석이 가능토록 하는데 그 목적이 있다. 형태가 다른 변수로 이루어진 데이터의 개체 간 거리를 정확하게 측정하는 것은 쉽지 않다. 특히 순서가 없는 범주형 변수인 명목형 변수가 존재하는 경우 수준 간 대소비교가 쉽지 않아 종종 문제가 된다. 본 논문에서는 이러한 데이터에 대해 세 가지 측정 방법을 고려하였다. 실제 분석에서 빈번하게 활용되는 범주형 변수의 가변수 변환 방법과 Gower의 방법 그리고 범주형 변수의 수준 개수 정보를 활용하는 Eskin의 측정방법 도입하여 혼합형 데이터에 거리를 측정할 수 있도록 응용한 방법이다. Eskin의 방법은 수준의 개수가 증가할수록 범주형 변수의 두 값이 일치하지 않는 경우가 빠르게 증가하여 거리가 과도하게 측정될 가능성이 많아지는 경우를 고려하여 수준의 개수 정보를 이용해 이를 보정한다. 세 가지 거리 측정방법을 계층적 군집분석 알고리즘에 개체 간 거리 측정의 정의하는 단계로써 활용하였다. Eskin의 방법이 다양한 경우에서 수준의 개수에 관계없이 다른 두 방법에 비해 같거나 더 나은 군집 결과를 도출하였다. 특히 실제 변수가 각 군집을 명확히 설명하지 못하는 경우 범주형 변수의 수준의 개수가 3 이상일 때 다른 두 방법에 비해 군집의 응집성, 재현성, 군집의 개수 정확도 측면에서 더 나은 성능을 보였다. 또한 군집의 퍼짐 정도 및 군집 별 개체의 불균형한 상황에서 역시 Eskin의 방법의 성능이 우수하였다. 아울러, 거리 측도들의 타당성을 실제 데이터를 이용하여 비교하였다. 가변수 변환 방법과 Gower의 방법은 개체 간 거리를 과대 측정하여 군집 개수를 증가시킨 반면 Eskin의 방법을 통해 상대적으로 타당한 군집결과를 얻을 수 있었다.

References

Boriah, S., Chandola, V., Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation, *2008 SIAM International Conference on Data Mining,* 243-254.

Chang, H., Kim, K. K., Kang, C. H. (2014). Comparison of clustering methods for categorical data, *Journal of the Korean Data Analysis Society,* 16(5), 689-697. (in Korean).

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions, *Journal Cybernetics,* 4(1), 95-104.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data, D. Barbara and S. Jajodia, editors, *Applications of Data Mining in Computer Security*, 78-100.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, *Biometrics,* 27, 851-871.

Han, J. S. (2018). *A comparative study for cluster analysis with mixed data,* Master´s Thesis, Graduate School, Korea University. (in Korean).

Hubert, L., Arabie, P. (1985). Comparing partitions, *Journal of Classification,* 2, 193-218.

Lee, S. H., Kang, H. C., Choi, H. S., Han, S. T. (2012). Customer segmentation by using two-step cluster analysis, *Journal of the Korean Data Analysis Society*, 14(4), 1849-1860. (in Korean).

Park, H. C., Cho, K. H. (2006). Comparison of clustering algorithms in data mining, *Journal of the Korean Data Analysis Society,* 8(2), 585-596 (in Korean).

Park, J. S., Choi, Y. S., Shin, S. M. (2016). The mixed h-plot for continuous and categorical data, *Journal of the Korean Data Analysis Society*, 18(1), 151-161. (in Korean).

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods, *Journal of American Statistical Association*, 66, 846-850.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics,* 20, 53-65.

Song, J. W. (2017). K-means cluster analysis for missing data, *Journal of the Korean Data Analysis Society,* 19(2), 689-697. (in Korean).

A Study on Cluster Analysis of Mixed Data with Continuous and Categorical Variables*

Jisoo Han[1], *HyungJun Cho*[2]

Abstract

In cluster analysis for mixed data consisting of continuous and categorical variables, the natural definition of distances such as Euclidean distance can be utilized for continuous variables. In contrast, nominal variables have often been eliminated due to their ambiguity in spite many nominal variables exist in real data. Defining distance measures is essential in cluster analysis methodologies based on distances among objects; therefore, it is important to define distance measures reasonably. When there exist both types of variables, we employ and compare the dummy variable transformation method used for modeling, Gower's method assigning distances according to matching of categories, and Eskin's method utilizing their category levels. Three evaluation measures are used for fair comparison regardless of the definitions of distance measures. As a result, Eskin's method performs better than the others in the unbalanced cases of object numbers and spread degrees. In addition, Eskin's method is superior in the cohesion and reproducibility of clusters and finding of cluster numbers.

Keywords : nominal variable, mixed data, distance measure, hierarchical clustering.

*This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A09058602).

[1]M.S, Department of Statistics, 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea.
E-mail : hjisu0316@korea.ac.kr

[2](Corresponding Author) Professor, Department of Statistics, 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Korea. E-mail : hj4cho@korea.ac.kr

[Received 20 July 2018; Revised 10 August 2018, 17 August 2018; Accepted 20 August 2018]