비즈니스 데이터 분석을 위한 베이지안 계층 군집분석

Bayesian hierarchical clustering for analyzing business data

저자 (Authors)
류성균, 황범석
Sung Kyun Rhyeu, Beom Seuk Hwang

출처 (Source)
한국데이터정보과학회지 31(1), 2020.1, 159-171 (13 pages)
Journal of the Korean Data And Information Science Society 31(1), 2020.1, 159-171 (13 pages)

발행처 (Publisher)
한국데이터정보과학회
The Korean Data and Information Science Society

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09299959

APA Style
류성균, 황범석 (2020). 비즈니스 데이터 분석을 위한 베이지안 계층 군집분석. 한국데이터정보과학회지, 31(1), 159-171.

Journal of the Korean Data & Information Science Society
2020, **31**(1), 159–171

http://dx.doi.org/10.7465/jkdi.2020.31.1.159
한국데이터정보과학회지

비즈니스 데이터 분석을 위한 베이지안 계층 군집분석[†]

류성균[1] · 황범석[2]

[12]중앙대학교 응용통계학과

접수 2019년 12월 9일, 수정 2020년 1월 7일, 게재확정 2020년 1월 8일

요 약

군집분석은 데이터 마이닝 기법의 일종으로 객체 간의 유사도 혹은 비유사도를 이용하여 비슷한 객체를 군집화하는 방법이다. 흔히 사용되는 군집분석 방법으로는 계층적 군집분석, k-평균 군집분석 등이 있으나 이러한 방법들은 이상치에 민감하고, 군집의 수와 같은 모수들을 사전에 정해야 하는 단점이 있다. 한편, 유전체 분석에서 활용되고 있는 베이지안 계층 군집분석은 가설 검정을 기반으로 군집을 정하기 때문에 앞서 말한 군집분석 방법의 단점을 보완할 수 있다. 본 연구에서는 모의실험을 통해 베이지안 계층 군집분석 방법의 장점과 기존 방법들과의 차이점을 확인하고, 실제 비즈니스 데이터에 이를 적용하여 최적의 군집분석 결과를 얻을 수 있는지 살펴본다.

주요용어: 계층적 군집분석, 고객 데이터, 군집 순도, 베이지안 계층 군집분석, 수정 랜드 지수.

1. 서론

군집분석이란 데이터 상에서는 관측되지 않는 패턴을 포착하기 위해 유사도 혹은 비유사도에 따라 데이터를 그룹화하는 분석방법을 말한다. 군집분석은 데이터를 그룹화한다는 점에서 분류 분석 (classification analysis)과 유사하다. 하지만, 분류 분석의 목적이 데이터에 존재하는 설명변수들을 활용해 결과변수의 범주를 명확하게 맞추는 데에 있다면, 군집분석은 일정한 기준에 따라 데이터에서는 관측되지 않은 숨은 군집을 찾는 데에 그 목적이 있다 (Park과 Yoon, 2017). 군집분석은 크게 거리 기반 군집분석과 모형 기반 군집분석으로 나눌 수 있다. 거리 기반 군집분석은 계층적 군집분석 (hierarchical clustering)이나 k-근접 이웃 군집분석 (k-nearest neighbors clustering)과 같이 데이터 요소 간 거리를 바탕으로 그룹화를 시도하는 반면, 모형 기반 군집분석은 관측치가 정규 혼합 모형 (Gaussian mixture model)과 같은 특정한 모형에 속한다고 가정한 뒤, 데이터를 적합하는 과정을 통해 그룹화를 시도한다. 이 중 계층적 군집분석은 데이터 상에 계층 구조가 있다고 전제한 뒤 이를 탐색해나가는 과정을 취하며, 자연어나 공간 데이터 등 계층 구조가 존재하는 여러 데이터에 적용되어 왔다 (Woo 등, 2014). 최근에는 베이지안 계층 군집분석 방법 (Bayesian hierarchical clustering)이 소개되어 유전체 분야의 데이터에 적용되기도 하였다 (Heller와 Ghahramani, 2005).

본 논문에서는 베이지안 계층 군집분석을 소개하고 모의 실험과 실제 데이터를 활용하여 기존의 계층적 군집분석과 다른 특성을 파악하고자 한다. 2절에서는 계층적 군집분석에 대해 소개하고 기존의 계층

[†] 이 논문은 2018년도 중앙대학교 CAU GRS 지원에 의하여 작성되었고, 2019년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2019R1C1C1011710).

[1] (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 석사과정.
[2] 교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 조교수.
E-mail: bshwang@cau.ac.kr

적 군집분석과 대조되는 베이지안 계층 군집분석만의 특징을 살펴보고자 한다. 3절에서는 다양한 형태의 다변량 정규분포에서 추출한 모의 데이터 상에서 기존의 계층적 군집분석 방법과 베이지안 계층 군집분석의 성능을 비교해보고 관련된 컴퓨팅 이슈도 간략히 소개하고자한다. 4절에서는 실제 도매 유통사의 고객 데이터에 대해 베이지안 계층 군집분석을 적용하여 그 결과를 비교 분석한다. 5절에서는 본 논문을 요약 정리하고 본 논문의 한계와 향후 후속 연구의 방향에 대해 논의한다.

2. 베이지안 계층 군집분석

2.1. 계층적 군집분석

계층적 군집분석 (hierarchical clustering)은 크게 병합 군집화 (agglomerative clustering)와 분할 군집화 (divisive clustering)가 있다. 병합 군집화의 경우 각각의 요소들을 가까운 것부터 합쳐나가는 반면, 분할 군집화의 경우 하나의 그룹에서 시작하여 이를 쪼개가는 방식으로 그룹을 탐색한다. 계층적 군집분석을 시행하려면 두 관측치 사이의 거리 (distance)를 어떻게 정의할 것인지, 관측치가 합쳐졌을 때 다른 관측치들과 새롭게 정의된 군집 사이의 거리는 어떻면 연결할지 (linkage) 등을 사전에 결정해야 한다. 거리를 정의하는 방법에는 유클리드 거리 (Euclidean distance), 맨해튼 거리 (Manhattan distance) 등이 있으며, 연결 방법 (linkage)에는 평균 연결법 (average linkage), 완전 연결법 (complete linkage) 등이 있다. 거리나 연결 방법을 다르게 선택할 경우 군집 분석의 결과가 달라지기도 한다. 뿐만 아니라 계층적 군집분석 결과를 해석하기 위해서는 앞서 선택한 거리와 연결법을 가지고 계산된 덴드로그램 (dendrogram)을 적절한 수준에서 잘라 나타내야 한다. 적절한 수준을 결정할 때에는 보통 해당 분야의 지식 등을 고려하거나 실루엣 지수 (Silhoutte Index)와 같은 측도를 최적화하는 값으로 사용하며 해석의 편의를 위해 가능하면 작은 수의 군집이 나타나도록 한다.

2.2. 베이지안 계층 군집분석

베이지안 계층 군집분석 (Bayesian Hierarchical Clustering; BHC)은 가설검정을 기반으로 수행하는 계층적 군집분석의 한 방법이다 (Heller와 Ghahramani, 2005). 베이지안 계층 군집분석 방법에는 다항 분포와 디리슈레 분포 (Dirichlet distribution) 간의 켤레 관계를 활용한 다항 분포 기반 베이지안 계층 군집분석 방법 (Multinomial Bayesian Hierarchical Clustering; MBHC) (Heller와 Ghahramani, 2005; Savaage 등, 2009)과 정규 분포와 역 감마 분포 (inverse gamma distribution) 간의 켤레 관계를 활용한 정규 분포 기반 베이지안 계층 군집분석 방법 (Gaussian Bayesian Hierarchical Clustering; GBHC) (Sirinukunwattana 등, 2012)이 있다. 본 논문에서는 두 방법 중 다항 분포 가정을 기반으로 한 베이지안 계층 군집분석 (MBHC)에 초점을 맞추고자 한다. 두 방법 모두 연속형 데이터에 대해 사용할 수 있지만, MBHC 방법의 경우 데이터를 이산형 데이터로 변환하여 사용하기 때문에 정보 손실이 있는 반면, 분포의 가정과 이상치에 로버스트한 장점이 있다.

2.2.1. 알고리즘

베이지안 계층 군집분석은 병합 군집화 방법과 마찬가지로 모든 데이터 요소가 하나의 군집에 속한다고 가정하고 이를 합쳐나가는 방식으로 진행된다. 구체적인 베이지안 계층 군집분석의 알고리즘은 다음와 같다 (Heller와 Ghahramani, 2005; Savage 등, 2009).

$\mathcal{D} = \left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}\right\}$을 전체 데이터 집합이라고 할 때, $\mathcal{D}_i$는 전체 데이터를 반영한 덴드로그램 중 가지 T_i에 속하는 데이터 집합을 나타낸다. 베이지안 계층 군집분석은 분석이 시작할 때 모든 관측치 $\mathcal{D}_i = \left\{\mathbf{x}^{(i)}\right\}$가 각각의 가지 $\{T_i : i = 1, \cdots, n\}$에 속한다고 가정한 후, 이를 계속 합쳐나가게 된다. 구

체적으로는 $\mathcal{D}_k = \mathcal{D}_i \cup \mathcal{D}_j$, 즉, T_i와 T_j가 T_k로 합쳐질 수 있는지를 가설검정을 통해 판단하는데, 이때 영가설 H_1^k는 "데이터 집합 $\mathcal{D}_k$에 속하는 모든 관측치가 하나의 확률 모형 $p(x|\theta)$로부터 동일한 독립시행을 통해 생성되었다."로 설정된다. 이때 θ는 데이터가 따르는 분포의 모수를 의미하며 모수에 대한 사전 확률 $p(\theta|\beta)$를 가진다고 정의한다. 또한 β는 모수 θ에 대한 초모수 (hyperparameter)를 나타낸다.

가설검정의 과정은 다음과 같다. 먼저 영가설의 가능도함수와 사전 확률에 따라 영가설 H_1^k를 조건부로 하는 $\mathcal{D}_k$의 확률 $p(D_k|H_1^k)$를 다음과 같이 계산한다.

$$\begin{aligned} p(\mathcal{D}_k|H_1^k) &= \int p(\mathcal{D}_k|\theta)p(\theta|\beta)d\theta \\ &= \int \prod_{\mathbf{x}\in\mathcal{D}_k} \left[p(\mathbf{x}^{(i)}|\theta)\right] p(\theta|\beta)d\theta. \end{aligned}$$

이렇게 계산한 확률은 D_k가 군집 하나에 얼마나 잘 적합하는 지를 나타내며, 가능도함수와 사전 확률은 켤레 관계를 활용하여 효율적으로 계산이 이루어지도록 한다. 또한, 대립 가설인 H_2^k는 "데이터 집합 $\mathcal{D}_k$에 속하는 관측치들은 두 개 혹은 그 이상의 확률분포로부터 나왔다."로 설정하며, T_i와 T_j에 속한 데이터 집합 D_i와 D_j가 각각 다른 확률 분포로부터 생성되었다고 가정하면, $p(\mathcal{D}_k|H_2^k) = p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j)$로 표현할 수 있다. 따라서 이렇게 정의한 영가설과 대립 가설을 지지하는 각각의 확률을 결합하여 T_k에 속한 데이터의 주변 확률 (marginal probability)을 다음과 같이 구할 수 있다.

$$p(\mathcal{D}_k|T_k) = \pi_k p(\mathcal{D}_k|H_1^k) + (1-\pi_k)p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j).$$

이때, 가중치로서 사용되는 π_k는 데이터 집합 $\mathcal{D}_k$에 속하는 모든 관측치가 하나의 군집에 속할 확률, 즉, $p(H_1^k)$를 의미하며 해당 모형의 모수인 α와 해당 서브 트리에 속한 관측값 개수의 함수로써 계산된다.

마지막으로 지금까지 구한 식을 통해 영가설에 대한 사후확률 $r_k \equiv p(H_1^k|\mathcal{D}_k)$를 베이즈 정리를 이용하여 다음과 같이 구할 수 있다.

$$r_k = \frac{\pi_k p(\mathcal{D}_k|H_1^k)}{\pi_k p(\mathcal{D}_k|H_1^k) + (1-\pi_k)p(\mathcal{D}_i|T_i)p(\mathcal{D}_j|T_j)}.$$

이때, 사후확률 r_k가 0.5보다 크면 영가설 H_1^k를 지지하여 T_i와 T_j가 T_k로 합쳐지게 된다. 반면, r_k가 0.5보다 작으면 T_i와 T_j는 병합되지 않는다. 이러한 알고리즘을 모든 요소 쌍에 대해 계산한 후 r_k가 가장 큰 쌍을 병합하게 되며, 가장 큰 r_k가 0.5보다 작아지게 될 때까지 병합을 계속하게 된다. 이런 과정을 요약하면 다음과 같다.

[베이지안 계층 군집분석 알고리즘]

input: data $\mathcal{D} = \left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}\right\}$, model $p(\mathbf{x}|\theta)$, prior $p(\theta|\beta)$

initialize: number of clusters $c = n$, and $\mathcal{D}_i = \left\{\mathbf{x}^{(i)}\right\}$ for $i = 1, \cdots, n$

while $c > 1$ **do**

Find the pair $\mathcal{D}_i$ and $\mathcal{D}_j$ with the highest probability of the merged hypothesis:

$$r_k = \frac{\pi_k p(\mathcal{D}_k|H_1^k)}{p(\mathcal{D}_k|T_k)}$$

Merge $\mathcal{D}_k \rightarrow \mathcal{D}_i \cup \mathcal{D}_j$, $T_k \rightarrow (T_i, T_j)$
Delete $\mathcal{D}_i$ and $\mathcal{D}_j$, $c \rightarrow c-1$
end while
output: Bayesian mixture model where each tree node is a mixture component
The tree can be cut at points where $r_k < 0.5$

2.2.2. 베이지안 계층 군집분석의 특징

베이지안 계층 군집분석은 모든 군집 혹은 관측치 쌍에 대해 사후확률 r_k를 구한 뒤 가장 큰 사후확률이 0.5보다 작아질 때까지, 즉 $\text{logit}(r_k)$가 0보다 작아질 때까지 유사한 군집들을 합쳐나가기 때문에 군집분석을 하는 과정에서 자연스럽게 최적의 군집 수를 구할 수 있다. 또한 베이지안 계층 군집분석은 확률 모형에 기반한 추론 과정이기 때문에 기존의 계층적 군집분석에서 필요로 하는 거리의 정의, 연결법 등이 필요하지 않으며, 새롭게 추가된 관측치에 대해서는 사후 예측 분포 (posterior predictive probability)를 이용하여 별도의 적합 과정 없이 새로운 관측치가 어느 군집에 속할 지를 판단할 수 있다. 베이지안 계층 군집분석 알고리즘은 디리슈레 확률과정 혼합 모형 (Dirichlet process mixture model)의 근사적인 추론방법으로 알려져 있다 (Heller와 Ghahramani, 2005).

또한, 베이지안 계층 군집분석은 마이크로어레이 (microarray) 데이터를 분석하기 위한 방법으로 도입되었기 때문에 다양한 유전체뿐만 아니라 조건들을 군집화하여 하나의 플롯에 함께 시각화하는 양방향 군집분석 (biclustering)이 가능하다. 기존의 계층적 군집분석 방법에서는 변수들을 군집분석 하기 위해 상관계수를 기반으로 한 별도의 군집분석을 수행하여 이를 관측치의 거리를 기반으로 한 군집분석과 결합하여 양방향 군집분석 플롯을 그릴 수 있다. 하지만, 베이지안 계층 군집분석을 활용한다면, 관측치와 다른 별도의 방법으로 변수들을 군집화하지 않고 하나의 알고리즘으로 관측치와 변수 모두 군집화하여 시각화할 수 있다는 장점이 있다. 최근에는 마이크로어레이 데이터뿐만 아니라 세장 세분화 등 다양한 영역에서 양방향 군집분석이 활용되고 있다는 점에서 (Dolnicar 등, 2012; Divina 등, 2019) 베이지안 계층 군집분석의 양방향 군집분석 플롯을 통한 직관적인 시각화가 데이터를 이해하는 데에 도움을 줄 것으로 판단된다.

3. 모의실험

3.1. 모의실험 구성

군집분석은 앞서 언급한 것과 같이 명확하게 정의된 반응변수 (response variable)가 존재하지 않거나 이를 고려하지 않고 이루어진다. 따라서, 명확한 성능측도를 정의하기 어렵고 해당 분야의 지식을 통해 해석되는 것이 일반적이다. 본 연구에서는 실제 소속된 그룹을 알 수 있는 모의실험 데이터를 생성하고 기존의 계층적 군집분석과 베이지안 계층 군집분석에 적합시켜 두 방법론의 성능을 확인하고자 한다. 모의실험은 6차원, 10차원의 다변량 정규 혼합모형을 기반으로 다양한 모수 조합을 시도하였으며, 수정 랜드 지수 (ARI)나 군집 순도 (cluster purity) 등을 성능 측도로써 사용하였다.

3.1.1. 모의실험 1

모의실험 데이터는 3개의 정답 군집이 존재하는 다변량 정규 혼합모형을 기반으로 각 case마다 모수를 바꿔가며 450개의 관측치를 생성했다. 이때, 혼합모형에 사용된 다변량 정규 분포의 공분산은 조작하기 쉽도록 각각의 변수들의 분산 (σ_i^2)과 상관계수 행렬 ρ를 각각 정한 뒤, 이를 합쳐서 구성하였다.

이때, 6차원 혼합 모형의 상관계수 행렬은 4장에서 사용할 실제 데이터의 상관계수 행렬을 고려하여 구성하였다. 상관계수 행렬과 각각의 모수 조합에 대한 내용은 각각 Table 3.1과 Table 3.2에서 확인할 수 있으며, 각 case마다 100개의 데이터 셋을 생성하였다.

$$\boldsymbol{x} \sim p_1 MVN(\mu_1 \cdot \mathbb{1}_6, \Sigma_6) + p_2 MVN(\mu_2 \cdot \mathbb{1}_6, \Sigma_6) + p_3 MVN(\mu_3 \cdot \mathbb{1}_6, \Sigma_6).$$

이때 다변량 정규분포의 평균에 사용된 $\mathbb{1}_6$는 모든 원소가 1로 이루어진 6차원의 벡터 $\mathbb{1}_6 = (1, \cdots, 1)^T$를 의미한다.

Table 3.1 Correlation matrix of UCI Wholesale data

	Fresh	Milk	Grocery	Frozen	Deter_paper	Delicassen
Fresh	1.0	0.1	0	0.3	-0.1	0.2
Milk	0.1	1	0.7	0.1	0.7	0.4
Grocery	0	0.7	1.0	0	0.9	0.2
Frozen	0.3	0.1	0	1.0	-0.1	0.4
Deter_paper	-0.1	0.7	0.9	-0.1	1.0	0.1
Delicassen	0.2	0.4	0.2	0.4	0.1	1.0

Table 3.2 Parameters in 6 dimensional normal mixture models

case	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	p_1	p_2	p_3
A_6	-10	0	10	100	100	50	50	30	10	0.3	0.4	0.3
B_6	-15	0	15	50	50	50	30	30	10	0.3	0.4	0.3
C_6	-10	0	10	50	50	50	30	30	10	0.3	0.4	0.3
D_6	-20	0	20	50	50	50	30	30	10	0.3	0.4	0.3
E_6	-10	0	10	50	50	50	30	30	10	0.1	0.8	0.1
F_6	-30	0	30	100	100	50	50	30	10	0.3	0.4	0.3

3.1.2. 모의실험 2

10차원 혼합모형 모의실험 또한 6차원 혼합모형 모의실험과 동일한 방식으로 3개의 정답 군집이 존재하는 450개의 관측치를 생성하였다. 이때 사용된 상관계수 행렬 ρ는 QR decomposition을 활용하여 양정치행렬을 생성한 뒤, 생성된 행렬에서 상관계수 행렬을 추출해서 사용했다. 각각의 모수 조합에 대한 내용은 Table 3.3에서 확인할 수 있고, 마찬가지로 각 case마다 100개의 데이터 셋을 생성하였다.

$$\boldsymbol{x} \sim p_1 MVN(\mu_1 \cdot \mathbb{1}_{10}, \Sigma_{10}) + p_2 MVN(\mu_2 \cdot \mathbb{1}_{10}, \Sigma_{10}) + p_3 MVN(\mu_3 \cdot \mathbb{1}_{10}, \Sigma_{10}).$$

이때 다변량 정규분포의 평균을 계산하는 데에 사용된 $\mathbb{1}_{10}$은 모든 원소가 1인 10차원의 벡터 $\mathbb{1}_{10} = (1, \cdots, 1)^T$를 의미한다.

3.1.3. 비교 모형 소개

베이지안 계층 군집분석과 성능을 비교할 방법으로 기존의 계층적 군집분석을 선택하였다. 모든 계층적 군집분석은 유클리드 거리를 기반으로 하며, 연결법은 평균 연결법 (average linkage)과 완전 연결법 (complete linkage)을 사용하여 비교하고자 한다. 기존의 계층적 군집분석은 군집의 수를 정해주어야 군집분석을 수행할 수 있기 때문에 2개에서 10개까지의 군집의 수를 후보로 두고 3.1.4절에서 소개할 실루엣 지수 (Silhouette Index)를 최대로 하는 군집의 경우를 찾아 사용하였다.

Table 3.3 Paramters in 10 dimensional normal mixture models

case	μ_1	μ_2	μ_3	p_1	p_2	p_3
A	-10	0	10	0.3	0.4	0.3
B	-15	0	15	0.3	0.4	0.3
C	-10	0	10	0.3	0.4	0.3
D	-20	0	20	0.1	0.8	0.1

case	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}
A	50	50	50	50	50	30	30	30	10	10
B	100	100	100	50	50	50	30	30	30	10
C	100	100	100	50	50	50	30	30	30	10
D	100	100	100	50	50	50	30	30	30	10

3.1.4. 성능 측도

앞서 설정한 모형들을 비교하기 위해서는 적절한 성능 측도 (performance measure)가 필요하다. BHC 간의 비교를 위해서는 R 패키지에서 계산한 사후확률을 활용할 수 있지만, 다른 모형, 특히 다른 계층적 군집분석 모형과 성능을 비교하기 위해서는 추가적인 성능 측도가 필요하다. 계층적 군집분석 방법의 경우 실루엣 지수 (Silhouette Index) (Rousseeuw, 1987)나 던 지수 (Dunn Index) (Dunn, 1974) 등 거리 기반 측도를 통해 군집분석의 성능을 평가하지만 베이지안 계층 군집분석의 경우에는 관측치 간 거리가 아닌 가설검정을 기반으로 군집을 병합한다는 차이가 있다. 본 연구에서는 실루엣 지수와 같은 거리 기반 측도를 가지고 두 방법을 비교하는 것은 적절하지 않다고 판단하여, 군집 분석의 성능을 평가하는데 빈번히 사용되는 수정 랜드 지수 (ARI)와 군집 순도 (cluster purity)를 성능 측도로써 고려하고자 한다. 측도에 대한 자세한 내용은 아래와 같다.

1) ARI

수정 랜드 지수 (Adjusted Rand Index; ARI)는 두 군집분석의 결과를 비교하여 군집분석의 일치성을 평가하는 측도이다 (Hubert와 Arabie, 1985). ARI는 보통 0과 1의 범위를 가지지만 무작위 배정보다 결과가 더 상이하다면 음수 값을 가지기도 한다. ARI가 0이라면 두 군집분석이 다른 결과를 제시했다는 것을 나타내며, 1이라면 두 군집분석의 결과가 관측치의 모든 순열 조합들에서 일치한다는 것을 의미한다. 또한, ARI 값이 클수록 두 군집분석 결과가 유사하다고 해석할 수 있다.

2) 군집 순도 (Cluster Purity)

군집 순도는 데이터가 가지고 있는 군집 정보를 알고있다고 가정할 때, 이를 실제 적용한 군집분석 결과와 얼마나 일치하는지를 나타내는 측도로써 정답 군집마다 가장 빈번히 나온 군집분석 결과들을 합친 뒤 이를 총 관측 수로 나누어 구한다 (Manning 등, 2008). 이는 판별 분석의 측도로 활용되는 판별 정확도 (classification rate)와 유사한 개념으로 군집분석의 정확도를 의미한다. 군집 순도는 0에서 1 사이의 범위 값을 가지며, 군집 순도가 1이라는 의미는 모든 군집이 적절하게 배정되었다는 것을 의미한다.

군집 순도는 ARI에 비해 군집의 성능을 더 직관적으로 평가할 수 있다는 장점이 있지만, 추정하는 군집의 수가 늘어날수록 군집 순도는 더 높아지는 단점도 존재한다. 하지만 본 연구의 모의실험에서는 혼합 모형의 그룹이 3개로 고정되어 있으며, 추론된 군집의 수도 2에서 5개 사이로 크게 바뀌지 않기 때문에 이를 연구의 측도로써 활용하고자 한다.

3) 실루엣 지수 (Silhouette Index)

실루엣 지수는 앞선 두 측도와 다르게 같은 군집에 속한 관측치들 사이의 거리와 각기 다른 군집에 속

한 관측치 사이의 거리를 가지고 군집의 성능을 평가한다.

실루엣 지수는 관측치 간 거리를 기반으로 군집분석 결과를 평가하기 때문에 가설검정을 기반으로 군집을 배치하는 베이지안 계층 군집분석의 성능을 비교하는 척도로 활용하기에 적절하지 않다. 실제로 모의실험에서도 실루엣지수를 가지고 군집의 수를 최적화한 계층적 군집분석 방법의 경우 정답 군집보다 더 높은 실루엣 지수를 보인 경우가 존재했다. 이에 본 연구에서 실루엣 지수는 기존의 계층적 군집분석의 군집 수를 정하는 데에만 제한적으로 사용하였다.

3.1.5. R Package

베이지안 계층 군집분석을 수행하기 위해 R (ver. 3.6.1)의 BHC package (ver. 1.36.0)를 사용하였다. 또한 성능 측도를 위해 MixGHD (ver. 2.3.3)와 NMF (ver. 0.21.0) package를 사용하였다.

BHC package의 경우, 다항분포 기반 BHC만 제공하고 있으며, 최대 3개의 범주를 가정하고 모델을 적합하기 때문에 연속형 변수나 4개 이상의 범주를 가지고 있는 경우 군집분석 전에 이산화 과정(discretizing)을 거쳐야 한다. 즉, 시행착오 과정 (trial and error)을 통해 최적의 이산화 분위수를 탐색하고, 여기서 구한 최적의 분위수를 기준으로 이산화를 거친 뒤 군집분석을 진행하도록 하고 있다. 모형의 초모수로서 활용되는 α의 경우 0.001로 고정되며 β 또한 최적의 값을 자동으로 탐색하여 사용하게 된다 (Savage 등, 2009).

3.2. 모의실험 결과

3.2.1. 모의실험 1 결과

6차원 혼합모형 모의실험 결과는 Table 3.4, Table 3.5, Figure 3.1에서 확인할 수 있다. 경우에 따라 베이지안 계층 군집분석 방법 (BHC)과 평균 연결법 (Average)을 사용한 방법이 모수 조합에 따라 가장 좋은 결과를 보인 것으로 나타났다. 하지만 군집 순도 (purity)에 따르면 BHC 방법이 다른 방법들에 비해 안정적인 성능을 보이는 것을 확인할 수 있다. 추가적으로 Table 3.7을 확인했을 때, BHC는 군집의 수를 주로 5개로 추정한 것을 확인할 수 있는데, 이러한 결과가 군집 순도에 비해 ARI의 값에 더 큰 영향을 미친 것으로 보인다.

Table 3.4 Simulation results of 6 dimensional normal mixture models: ARI

ARI	A_6	B_6	C_6	D_6	E_6	F_6
BHC	**0.431**	0.689	**0.453**	0.821	0.031	0.860
Average	0.200	**0.920**	0.319	**0.991**	**0.266**	**0.999**
Complete	0.285	0.706	0.360	0.965	0.261	0.996

Table 3.5 Simulation results of 6 dimensional normal mixture models: purity

purity	A_6	B_6	C_6	D_6	E_6	F_6
BHC	**0.796**	0.937	**0.829**	0.977	0.807	0.988
Average	0.528	**0.972**	0.597	**0.997**	**0.841**	**1.000**
Complete	0.578	0.841	0.619	0.987	0.828	0.999

3.2.2. 모의실험 2 결과

10차원 혼합모형 모의실험 결과는 Table 3.6, Figure 3.2에서 볼 수 있다. 조합에 따라 BHC와 평균 연결법 (Average)을 사용한 방법이 가장 좋은 결과를 보이며, 6차원 혼합모형의 모의실험 결과와 유사

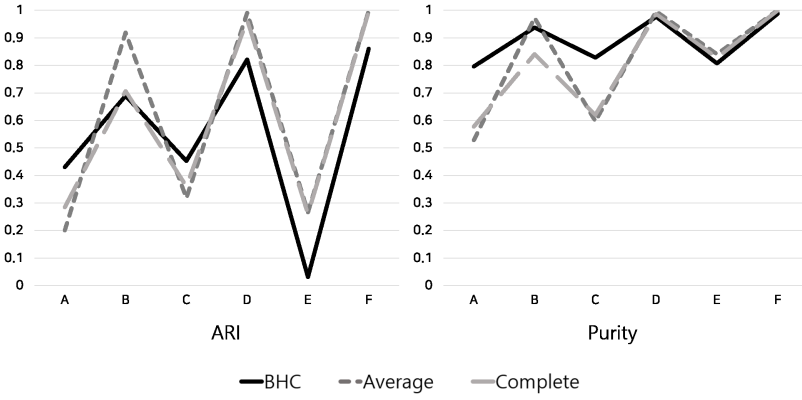

Figure 3.1 Simulation results of 6 dimensional normal mixture models

한 결과가 나타났다. 하지만, BHC는 다양한 모수 조합의 경우에 다른 계층적 군집분석 방법보다 안정적인 성능을 보여주었다. 또한 Table 3.7에서 확인할 수 있듯 BHC 또한 군집의 수를 적절하게 판단했으며 case C의 경우 실험한 방법 중에서 유일하게 군집의 수를 정확하게 추론했다. 전반적으로 Figure 3.1과 3.2를 볼 때 6차원 모의실험보다 10차원 모의실험에서 BHC의 성능이 안정적으로 나타난 것을 확인할 수 있다.

Table 3.6 Simulation results of 10 dimensional normal mixture models: ARI & purity

	ARI				purity			
case	A	B	C	D	A	B	C	D
BHC	**0.926**	0.982	**0.852**	0.873	**0.981**	0.996	**0.956**	0.983
Average	0.861	**0.993**	0.426	**0.984**	0.920	**0.998**	0.645	**0.996**
Complete	0.834	0.968	0.445	0.933	0.912	**0.988**	0.657	0.985

Table 3.7 Estimated number of clusters in simulation studies

	A_6	B_6	C_6	D_6	E_6	F_6	A	B	C	D
BHC	4.67	5.21	5.30	4.81	3.78	4.59	3.19	3.11	3.26	3.23
Average	2.10	3.03	2.15	3.00	2.07	3.00	2.91	3.06	2.20	3.04
Complete	2.01	2.68	2.06	3.00	2.31	3.00	2.81	3.00	2.04	2.95

4. 실제 데이터 분석

4.1. UCI Wholesale Customers Data

UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets`)로부터 이용가능한 wholesale customers 데이터는 포르투갈에 소재하는 한 도매 유통업체의 고객 데이터로서 440개의 고객(사)에 대해 Region (지역)과 Channel (채널), 2개의 범주형 변수와 6개 상품 대분류에 대한 고객의 연간 구매 금액 변수로 구성되어 있다. Table 4.1과 4.2는 분석에 사용된 모든 변수들을 간략히 설명해주고 있다. Table 4.2를 보면 변수에 따라 대표값들에 차이가 존재하는 것을 확인할 수 있다. 따라서 기존의 계층적 군집분석의 경우 연속형 변수를 로그 변환하거나 정규화하는 등 전처리를 해야하며, 어떠한 전처리를 하느냐에 따라 군집 분석의 결과가 달라지기도 한다. 하지만 베이지안 계층 군집분석

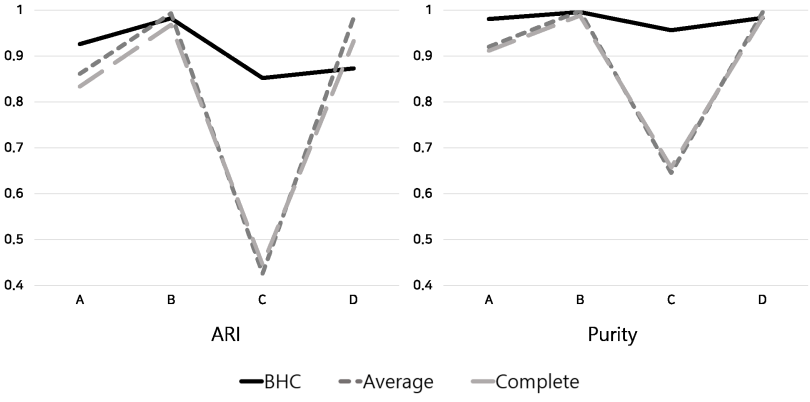

Figure 3.2 Simulation result of 10 dimensional normal mixture models

의 경우 연속형 변수들에 대해 이산화 (discretizing)를 하기 때문에 전처리의 영향력이 크지 않다. 해당 데이터에 한해서는 관측치 기준으로 전처리 여부와 무관하게 동일한 군집분석 결과를 보이는 것을 확인했으며, 범주형 변수의 경우 정규화의 영향을 받아 변수별 군집군석 결과만 달라지는 것을 확인했다.

Table 4.1 UCI Wholesale customers data: categorical variables

Variable	Category	Freq.	Prop.
Channel	1: hotel/restaurant/cafe	298	67.7%
	2: retail	142	32.3%
Region	1: Lisbon	77	17.5%
	2: Porto	47	10.7%
	3: other region	316	71.8%

Table 4.2 UCI Wholesale customer data: continuous variables

Variable	Min	1Q	Median	Mean	3Q	Max
Fresh	3	312	8504	12000	16934	112151
Milk	55	1533	3627	5796	7190	73498
Grocery	3	2153	4756	7951	10656	92780
Frozen	25	742	1526	3071.9	3554.2	60869
Detergents_Paper	3	256.8	816.5	2881.5	3922	40827
Delicatessen	3	408.2	965.5	1524.9	1820.2	47943

4.2. 분석 결과

해당 데이터는 2개의 범주형 변수와 6개의 연속형 변수로 구성되어 있으며 이 중 어느 변수들을 가지고 군집분석을 수행할지가 결과에 영향을 미칠 수 있다. 본 연구에서 고려한 경우의 수는 두 가지이다. 첫째, 범주형 변수는 연속형 변수들에 영향을 미치는 고객정보라고 가정하여 6개의 연속형 변수만을 가지고 베이지안 계층 군집분석을 수행하였다. 두 번째로 Region과 Channel의 범주형 변수까지 고려할 필요성이 있거나 (Han과 Cho, 2018), 최종 소비자의 주거지 및 기호 등이 데이터에는 나타나지 않은 숨은 변수로서 데이터상의 모든 변수에 영향을 미쳤다고 가정하고 이를 분석에 포함하는 경우이다. 분석 결과는 Table 4.3과 Figure 4.1, 4.2에서 확인할 수 있다. 먼저, Table 4.3을 보면, 연속형 변수만 사용할 때의 고객 군집 수는 2개인 반면, 범주형 변수까지 모두 포함할 경우 군집의 수는 3개로 예

측되었다. 즉, 모든 변수를 사용한 case 2에서 군집 1과 군집 2로 분류한 관측치들을 연속형 변수만 가지고 군집분석을 실시할 때 (case 1)는 하나의 군집으로 봤고, case 2에서 군집 3으로 분류된 대부분의 관측치는 case 1에서는 대다수 군집 2로 분류한 것을 확인할 수 있다. 또한 두 경우 모두에서 Grocery, Detergents_Paper, Milk 변수가 같은 군집으로 묶인 것을 확인할 수 있었으며, case 2에서는 Fresh, Frozen 또한 함께 묶인 것을 확인할 수 있었다. 이러한 경향은 case 1에서 두 변수 간의 $\text{logit}(r_k)$가 -3.95로 상대적으로 작게 나타난 것으로도 설명된다.

Table 4.3 Contingency table of BHC in UCI Wholesale customers data

		case 2			
		1	2	3	Subtotal
case 1	1	136	132	20	288
	2	0	0	152	152
	Subtotal	136	132	172	440

Figure 4.1과 4.2는 두 가지 case에 대해 양방향 군집분석을 통해 시각화한 결과를 나타낸 것이다. 가로축과 세로축으로 관측치와 변수들에 대해 군집분석을 한 결과가 나타나 있고, 데이터의 이산화 결과는 색상을 통해 표현되어있다. 양방향 군집분석을 통해 추가로 추론할 수 있는 내용에는 Grocery 변수가 군집을 할당하는 데에 큰 영향을 미친 것으로 보이며, Figure 4.2를 봤을 때, case 2에 추가된 두 명목 변수 중에 Channel은 잘 정렬되어있어 군집의 할당에 큰 영향을 미친 것으로 보이는 반면, Region의 경우 세로축을 기준으로 제대로 정렬되어 있지 않아 군집 구분에 큰 영향이 없었던 것으로 보인다.

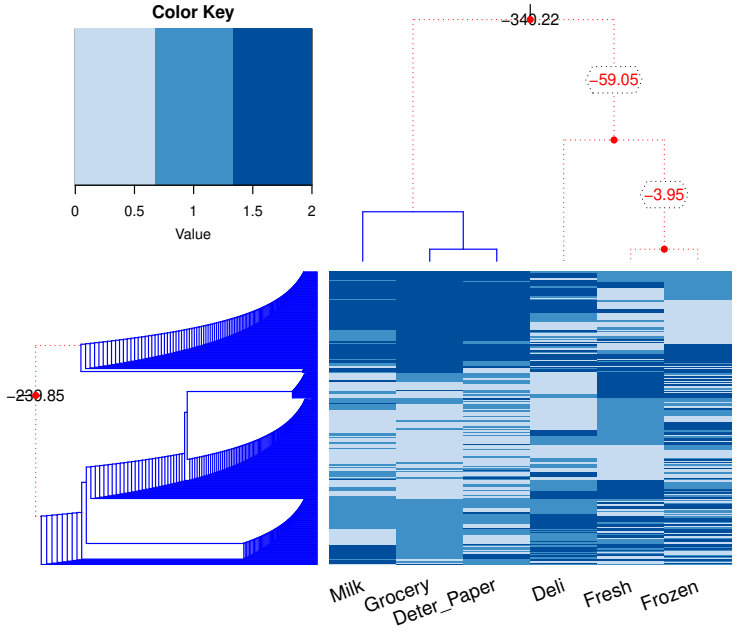

Figure 4.1 Biclustering plot: UCI Wholesale Customers Data - continuous variables only

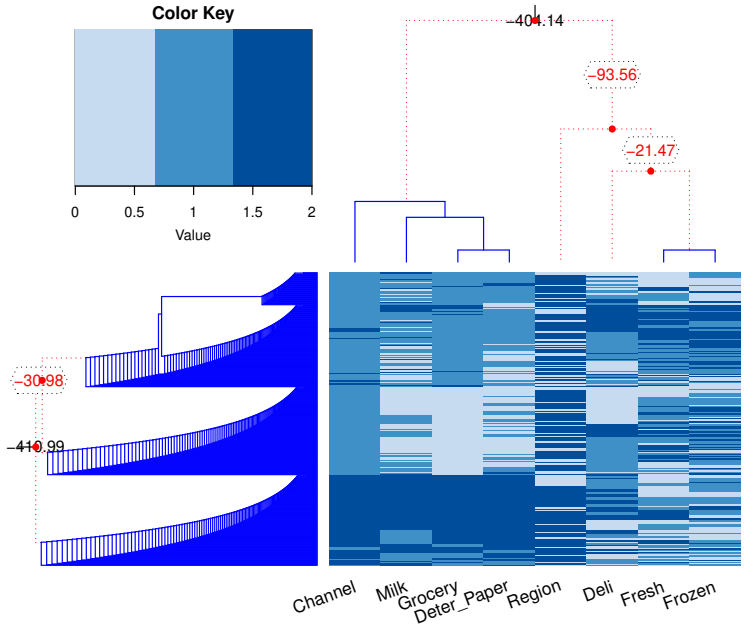

Figure 4.2 Biclustering plot: UCI Wholesale Customers Data - all variables

5. 결론 및 논의

탐색적 데이터 분석 관점에서 실제 데이터를 군집분석할 경우에는 해당 분야의 지식과 경험이 필수적이다. 하지만 그러한 지식과 경험을 가진 전문가가 통계적 군집분석을 이해하기는 쉽지 않을 수 있다. 이에 본 연구에서는 여러 사전 옵션을 고민하지 않고 분석을 진행할 수 있는 베이지안 계층 군집분석에 주목하였다. 베이지안 계층 군집분석의 경우 기존의 군집분석 방법과는 달리 여러 옵션을 사전에 지정해주지 않고도 군집의 수와 배정 결과를 구할 수 있으며, 변수와 관측치를 군집분석할 때에도 하나의 일관된 방법론을 적용하여 군집분석을 수행하고 양방향 군집분석 플롯 또한 그릴 수 있다. 베이지안 계층 군집분석에 관련된 선행연구는 해당 방법론이 출발하였던 마이크로 어레이 데이터나 자연어 처리 등에 적용되었던 반면, 본 연구에서는 일반적인 산업 데이터에서 기존의 계층적 군집분석과 비교해봤을 때 어느 정도의 성능을 보일 것인가를 예측하기 위해 여러 성능 측도를 활용한 모의실험을 고안하여 진행하였으며, 실제 데이터 상에서 베이지안 계층 군집분석이 어떠한 결과를 나타내는지 확인하였다.

모의실험 결과 기존의 계층적 군집분석에 비해 성능 측도가 안정적인 경향이 확인되었으며, 이러한 경향은 6차원 모의실험보다 10차원 모의실험에서 더 강하게 나타났다. UCI의 wholesale customers 데이터에 BHC를 적용해봤을 때는 관측치뿐만 아니라 변수 간의 군집분석을 통해 변수 간의 유사도를 확인할 수 있었으며, 데이터의 이산화 (discretizing) 결과를 함께 시각화하여 주기 때문에 관측치의 군집분석에 어느 변수가 유의미하고, 어느 변수가 그렇지 않았는지 알 수 있었다. 본 연구에서 사용한 다항 분포 기반 베이지안 계층 군집분석 (MBHC)의 경우 기존의 계층적 군집분석이나 정규분포 기반 베이지안 계층 군집분석에 비해 데이터의 전처리와 이상치의 유무, 정규성 등의 여부에 로버스트하며, 변수를 이산화하여 분석하기 때문에 범주형 변수와 연속형 변수가 함께 존재하는 데이터도 쉽게 군집 분석할 수 있다는 장점이 있다. 하지만 이산화 과정과 r_k 계산 등에서 계산 비용이 많이 든다는 단점도 존재한다.

본 연구를 통해 BHC 방법이 우수한 성능을 지니고 있다는 것을 일부 확인하였지만, 유전체 데이터, 비즈니스 데이터 외에 더 다양한 실제 데이터에 적용하여 일반적으로 쉽게 사용할 수 있게 일반화시킬 필요가 있을 것이다. 또한, BHC 방법을 선행 연구에서 언급한 디리슈레 과정에 기초한 비모수 베이지안 군집분석 방법과 비교하거나 GBHC 방법과 MBHC를 비교하는 연구 또한 향후 과제 중 하나가 될 수 있을 것이다.

References

Divina, F., Gomez Vela, F. A. and Garcia Torres, M. (2019). Biclustering of smart building electric energy consumption data. *Applied Sciences*, **9**, 1-15.

Dolnicar, S., Kaiser, S., Lazarevski, K. and Leisch, F. (2012). Biclustering: Overcoming data dimensionality problems in market segmentation. *Journal of Travel Research*, **51**, 41-49.

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

Han, J. and Cho, H. J. (2018). A Study on cluster analysis of mixed data with continuous and categorical variables. *Journal of the Korean Data Analysis Society*, **20**, 1769-1780.

Heller, K. A., and Ghahramani, Z. (2005). Bayesian hierarchical clustering. *In Proceedings of the 22nd International Conference on Machine Learning*, 297-304.

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

Manning, C. D., Raghavan, P. and Schutze, H. (2008). *Introduction to information retrieval*, Cambridge University Press.

Park, D. and Yoon, S. (2017). Clustering and classification to characterize daily electricity demand. *Journal of the Korean Data & Information Science Society,* **28**, 395-406.

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

Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J. and Wild, D. L. (2009). R/BHC: Fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics*, **10**, 1-9.

Sirinukunwattana, K., Savage, R. S., Bari, M. F., Snead, D. R. J. and Rajpoot, N. M. (2013). Bayesian hierarchical clustering for studying cancer gene expression data with unknown statistics. *PLOS ONE*, **8**, e75748.

Woo, S. Y., Lee, J. W. and Jhun, M. (2014). Microarray data analysis using relative hierarchical clustering. *Journal of the Korean Data & Information Science Society*, **25**, 999-1009.

Journal of the Korean Data & Information Science Society
2020, **31**(1), 159–171

http://dx.doi.org/10.7465/jkdi.2020.31.1.159
한국데이터정보과학회지

Bayesian hierarchical clustering for analyzing business data[†]

Sung Kyun Rhyeu[1] · Beom Seuk Hwang[2]

[12]Department of Applied Statistics, Chung-Ang University

Received 9 December 2019, revised 7 January 2020, accepted 8 January 2020

Abstract

Clustering is a kind of data mining methods that groups similar objects by using similarity or nonsimilarity between objects. The hierarchical clustering and k-means clustering are widely exploited, but these methods have some drawbacks in that sensitive to the outliers and require predetermined options such as the number of clusters. Meanwhile, the Bayesian Hierarchical Clustering (BHC) employed in microarray data analysis determines clusters based on the hypothesis testing, and therefore, it does not concern about the problems as mentioned above. In this study, we examine the advantage of BHC and the differences between well-known clustering methods and how this method could be applied to business data to obtain superior clustering result.

Keywords: ARI, Bayesian hierarchical clustering, cluster purity, hierarchical clustering, Wholesale customers data.

[†] This research was supported by the Chung-Ang University Graduate Research Scholarship in 2018, and supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2019R1C1C1011710).

[1] Graduate student, Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea.

[2] Corresponding Author: Assistant professor, Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bshwang@cau.ac.kr