한국빅데이터학회지
제5권 제2호, 2020, pp. 135-143 https://doi.org/10.36498/kbigdt.2020.5.2.135

다차원 데이터의 군집분석을 위한 차원축소 방법: 주성분분석 및 요인분석 비교*

A dimensional reduction method in cluster analysis for multidimensional data: principal component analysis and factor analysis comparison

홍준호[1] · 오민지[1] · 조용빈[2] · 이경희[3] · 조완섭[4†]

충북대학교 대학원 빅데이터학과[1], 농촌진흥청[2], ㈜힐링소프트[3], 충북대학교 경영정보학과[4]

요 약

본 논문은 농식품 소비자패널 데이터에서 소비자의 유형을 나눌 때에 변수간 연관성이 많은 장바구니 분석에서 전처리 방법과 차원축소의 방법을 제안한다. 군집분석은 다변량 자료에서 관측 개체를 몇 개의 군집으로 나눌 때 널리 사용되는 분석기법이다. 하지만 여러 개의 변수가 연관성을 가진 경우에는 차원축소를 통한 군집분석이 더 효과적일 수 있다. 본 논문은 1,987 가구를 대상으로 조사한 식품소비 데이터를 K-means 방법을 사용하여 군집화하였으며, 군집을 나누기 위해 17개의 변수를 선정하였고, 17개의 다중공선성 문제와 군집을 나누기 위한 차원축소의 방법 중 주성분 분석과 요인분석을 비교하였다. 본 연구에서는 주성분분석과 요인분석 모두 2개의 차원으로 축소하였으며 주성분분석에서는 3개의 군집으로 나뉘었지만 분석하고자 하였던 소비 패턴에 대한 군집의 특성이 잘 나타나지 않았으며 요인분석에서는 분석가가 보고자 하는 소비 패턴의 특징이 잘 나타났다.

■ 중심어 : 주성분 분석, 요인 분석, 차원 축소, 군집 분석, 패널데이터

Abstract

This paper proposes a pre-processing method and a dimensional reduction method in the analysis of shopping carts where there are many correlations between variables when dividing the types of consumers in the agri-food consumer panel data. Cluster analysis is a widely used method for dividing observational objects into several clusters in multivariate data. However, cluster analysis through dimensional reduction may be more effective when several variables are related. In this paper, the food consumption data surveyed of 1,987 households was clustered using the K-means method, and 17 variables were re-selected to divide it into the clusters. Principal component analysis and factor analysis were compared as the solution for multicollinearity problems and as the way to reduce dimensions for clustering. In this study, both principal component analysis and factor analysis reduced the dataset into two dimensions. Although the principal component analysis divided the dataset into three clusters, it did not seem that the difference among the characteristics of the cluster appeared well. However, the characteristics of the clusters in the consumption pattern were well distinguished under the factor analysis method.

■ Keyword : PCA, Factor Analysis, Dimensionality Reduction, Clustering, Panel Data

2020년 11월 30일 접수; 2020년 12월 14일 수정본 접수; 2020년 12월 21일 게재 확정.

* 본 연구는 농촌진흥청 연구사업(농식품 소비, 유전체 특성 및 질병의 연관성 분석 (과제번호: PJ01538032020)) 지원에 의해 이루어졌습니다.

† 교신저자 (wscho@chungbuk.ac.kr)

Ⅰ. 서 론

보건복지부에 따르면, 국민건강영양조사(2007~2010년) 분석 결과 30세 이상 성인의 대사증후군 유병률이 28.8%로 나타났다. 특히 남성의 경우, 대사증후군 상대위험도가 사무종사자가 타 직종에 비하여 가장 높은 것으로 나타났으며 상대위험도가 높은 직업군은 신체활동이 부족하고 지방섭취와 스트레스가 많은 등 나쁜 생활습관이 그 원인이 된 것으로 판명되었다[1]. 또한 한국식품커뮤니케이션포럼에 따르면 국민건강영양조사(1998~2012년)와 통계청 자료를 근거로 성인의 비만율을 예측하였다. 국내 성인의 비만율 및 복부 비만율이 해마다 높아져 2030년 남성과 여성의 비만율은 각각 61.5%, 37.0%로 예측되었으며, 남성과 여성의 복부비만율은 46.8%, 35.6%로 예측되었다[2].

대사증후군의 구성요소 중 하나인 비만율 및 복부 비만율 증가에 있어 식습관은 중요한 요소이다. 식습관은 한 가구가 구매하는 식품 구매 패턴을 통해 나타나므로 식품 구매 패턴을 파악하여 인구특성정보 및 건강정보와 연계하여 분석하면 의료서비스 및 농식품 마케팅 등 다양한 산업분야에서 유용한 서비스로 이어질 수 있다. 특히, 식품 소비에 따라 고객을 세분화하여 의료서비스 및 마케팅 활동의 대상이 되는 핵심 고객군 선정을 위한 효과적인 타겟팅 활동과 세분화된 고객집단에 맞춤형 서비스 활동을 가능하게 함으로써 고객에게 적절한 대응과 제안을 할 수 있는 등 다양한 가치를 창출할 수 있다(장민석 외, 2018).

고객 세분화 방법은 전통적 방법과 다 기준 스코어를 적용하는 방법으로 나누어질 수 있다(서현지, 2017). 서현지(2017)에 따르면, 고객 세분화의 전통적 방법은 연령, 지역, 구매금액 등 단일 기준에 의한 세분화로서 적용이 쉽다는 장점이 있지만 분류의 지속성과 정확도를 보장할 수 없다는 단점이 있다. 따라서 전통적 고객 세분화의 단점을 보완하여 고객 관련 정보들을 중심으로 유사 고객군을 분류하기 위해 다 기준 스코어를 이용한 고객 세분화 방법인 K-Means 군집분석을 적용할 필요가 있다.

다 기준 스코어를 이용한 고객 세분화 방법으로서 비지도 기계학습 방법은 군집분석은 식품 소비패턴 등 특정 관측 개체를 비슷한 패턴에 따라 분류하는 방법이다. 군집분석 방법 중 비계측정 군집분석인 K-Means 기법은 각 군집까지 유클리디언 거리를 계산하고 군집의 중심까지 거리가 가장 최소가 되는 군집으로 개체를 할당하여 군집내 거리를 최소화하는 군집을 찾는 방법이다. Milligan (1996)에 따르면 변수의 개수가 많은 경우, 군집 분류에 기여하지 않은 변수들이 존재할 가능성이 있으며 이러한 변수들은 군집구조를 찾아내는 것을 방해할 가능성이 있다. 이러한 이유로 다차원 데이터를 저차원으로 축소할 필요성이 있다. 기존 연구에서는 가능한 많은 정보를 포함한 채로 (원본 데이터의 주요 정보를 보존) 차원 축소를 실현하고, 그 결과에 대하여 저차원 군집분석을 실시함으로써 분류의 정확도를 높이는 방법들이 제안되었다.

본 논문의 구성은 다음과 같다. 제 2장에서 관련연구를 설명하고, 제 3장에서는 주성분 분석을 통한 패널데이터 분석과 요인분석을 통한 군집분석을 비교하여 적절한 방법 선택의 기준을 살펴본다. 제 4장에서 연구내용을 요약하고 향후과제를 기술한다.

Ⅱ. 관련연구

식품 소비패턴에 따라 집단의 특성을 파악하려는 다양한 시도가 있었다. 장보영 외(2019)는 에너지 섭취량 분석 및 주요 식사섭취패턴을 조사하기 위해 국민건강영양조사자료를 바탕으로 식품 분류코드를 재분류하여 계층적 군집분석 방

법인 Ward 방법과 비계층적 군집분석인 K-Means 방법을 실시하였다. 데니스 은주 외(2017)는 국민건강영양조사자료를 활용하여 음료군을 재분류한 후 음료군의 섭취량을 기본변수로 요인분석을 실시하고 주성분분석을 통해 설명요인을 추출하였다. 스웨덴 유전자-환경 상호작용 연구 프로그램에서 식습관 패턴과 심혈관 사이의 위험요소를 진단하기 위해 K-Means 방법으로 식습관 패턴을 파악하였다(Christina M Berg et al., 2008).

군집분석을 실시함에 있어 데이터의 차원을 축소하는 효과에 대한 연구 또한 활발하다. 양혜리 외(2018)는 메모리와 디스크 사용에 대한 비용을 줄이기 위해 주성분분석을 사용하여 데이터의 차원을 축소한 뒤 K-Means 분석 알고리즘의 성능을 향상시켰다. 이용구 외(2012)는 추천시스템 알고리즘인 협업 필터링의 확장성 문제를 보완하기 위해 요인분석을 이용하여 데이터의 차원을 축소하였다. 식료품 지출에 따른 군집분석에 대한 선행연구로는 박미성 외(2014)는 식품소비 라이프스타일이 가공식품에 미치는 효과를 분석하기 위해 식품 소비에 따른 군집화를 시도하였다. 군집화를 하기 위해 19개의 변수를 요인분석으로 차원축소시켜 5개의 요인으로 나누었고 50%이상의 설명력으로 군집화를 시도하였다.

데이터의 차원을 축소하는 방법을 비교한 연구에 따르면, 주성분분석과 요인분석방법 모두 데이터 축소를 위해 사용된다(Santos, R. D. O. et al., 2019). 주성분분석은 원본데이터를 더 작은 성분집합으로 축소하여 변동성의 일부를 재현하려는 경우에 사용하며 대용량의 데이터 세트를 단순한 차원으로 기술하려는 목적이 있다. 반면, 요인분석은 식생활 패턴을 구축하는데 사용되는 통계모델로서 특정 모집단의 식생활 패턴을 분석하려는 연구에서 식숨의 소비를 설명하는 잠재변수를 타나태는 요인을 생성하기 위해 사용된다. Santos, R. D. O. et al.(2019)에 의하면 두 차원축소 방법은 변수의 공통분산이 낮을 때 다른 추정치가 도출되기 때문에 연구의 목적과 데이터 세트의 특성에 따라 차원축소방법을 선정해야 한다. MDS(Multidimensional scaling), 주성분분석(PCA), ICA(Independent compo-nent analysis), LLE(Locally linear embedding) 등 6가지 차원축소 방법에 따라 데이터 시각화에 소요되는 속도와 정확성을 측정하여 데이터 세트에 따른 효율적인 차원축소방법을 제시한 연구가 진행되었지만(Zubova, J., 2018), 주성분분석 및 요인분석으로 군집분석을 실행하여 이를 비교한 연구는 거의 없다.

Ⅲ. 연구방법

본 장에서는 연구에 사용된 데이터셋을 설명하고, 다변량 자료의 차원축소를 위해 주성분분석과 요인분석한 결과를 기술한 후, 두 기법을 비교한다.

3.1 데이터 셋

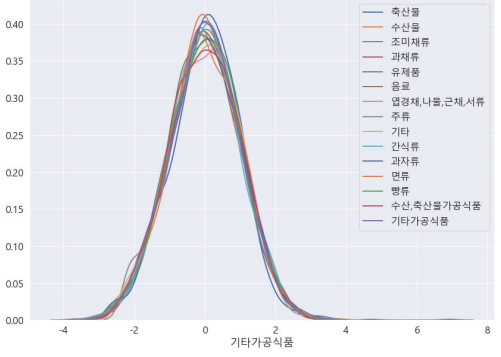
본 연구에 사용된 데이터는 2015년부터 2019년까지 농촌진흥청의 농식품 소비자패널 장바구니 데이터이로 총 패널 수는 1,987이며 5년 동안 5,829,476 건의 구입내역 데이터는 <표 1>과 같다. 장바구니 데이터에 포함된 소비식품은 기준에 따라 대분류-중분류-소분류로 나누어지며, 소비패턴을 분석하기 위해 <표 2>와 같이 재분류하였다. 또한 차원축소를 진행하기 전 <그림 1>과 같이 편향된 데이터를 정규화하였다. 각 변수의 정보량에 차이가 있어 한 객체가 가진 정보의 양이 동일하도록 그룹화하여 17개의 관측개체를 만들었다.

〈표 1〉 연도별 패널 수 및 레코드 수

연도	패널 수	레코드 수
2015년	1,503	1,172,150
2016년	1,411	1,073,598
2017년	1,838	1,367,816
2018년	1,542	1,186,872
2019년	1,469	1,029,040

〈표 2〉 소비 식품 분류 기준

	분류 품목
소비식품 기존분류 (대분류)	가공식품, 축산물, 조미채류, 과일류, 과채류, 엽경채류, 수산물, 나물류, 근채류, 특작류, 서류, 곡물류, 기타채소류, 건과 및 견과류, 과일과채혼합
소비식품 재분류	유제품, 엽경채 및 나물근채서류, 과채류, 기타가공식품, 수산 및 축산가공식품, 축산물, 과자류, 조미채류, 빵류, 음료, 간식류, 면류, 수산물



〈그림 1〉 17개의 변수 정규화

3.2 주성분분석을 통한 차원축소

주성분분석은 1901년에 피어슨(Karl Pearson)에 의해 제안되었으며, 1930년대에 Harold Hotelling에 의해 독자적으로 발전됐다. 이는 고차원의 데이터를 저차원의 데이터로 환원시키는 기법으로서 예측모델을 구축하거나 데이터의 탐색 및 분석의 도구로서 사용되었다. 주성분분석은 서로 연관 가능성이 있는 고차원 표본들을 선형 연관성이 없는 저차원 공간(주성분)의 표본으로 변환하기 위해 직교 변환을 사용한다. 주성분의 차원수는 원래 표본의 차원 수보다 작거나 같다. 데이터를 한 개의 축으로 사상시켰을 때, 그 분산이 가장 커지는 축을 첫 번째 주성분, 두 번째로 커지는 축을 두 번째 주성분으로 놓도록 새로운 좌표계로 데이터를 선형 변환한다.

데이터 집합의 전체 평균이 0이라고 가정하면, 데이터 집합 x의 주성분 w1은 방정식 (1)과 같이 정의된다.

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\mathbf{x}\right)^2\right\} \tag{1}$$

방정식 (2)에서와 같이, k-1개의 주성분이 이미 주어져 있을 때 k번째 주성분은 앞의 k-1개 주성분을 제외함으로써 찾을 수 있다.

$$\hat{\mathbf{x}}_k = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i \mathbf{w}_i^T \mathbf{x} \tag{2}$$

이후 방정식 (3)에서와 같이, 이 값을 데이터 집합에서 제외한 다음 주성분을 새로 찾는다.

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\hat{\mathbf{x}}_k\right)^2\right\} \tag{3}$$

따라서 방정식 (4)에서와 같이 카루엔-뢰브 변환은 데이터 행렬 X의 특잇값 분해를 찾은 다음, 방정식 (5)와 같이 X를 L개의 특이 벡터와 WL로 정의된 부분공간으로 사상시켜 부분 데이터 집합 Y를 찾는 것과 같다. X의 특잇값 벡터 행렬 W는 공분산 C=XXT의 고유 벡터 행렬 W와 동일하다(방정식 (6)). 가장 큰 고윳값을 갖는 고유 벡터는 데이터 집합에서 가장 강한 상관성을 갖는 차원에 대응된다.

$$\mathbf{X} = \mathbf{W}\mathbf{\Sigma}\mathbf{V}^T \quad (4)$$

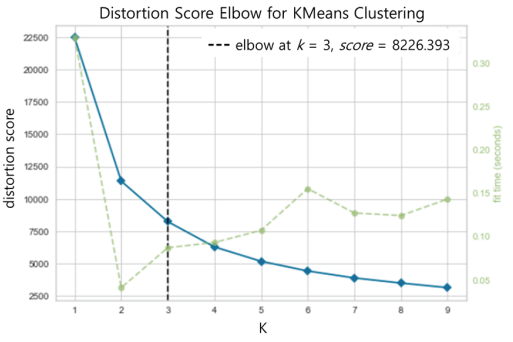
$$\mathbf{Y} = \mathbf{W_L}^T\mathbf{X} = \mathbf{\Sigma_L}\mathbf{V_L}^T \quad (5)$$

$$\mathbf{X}\mathbf{X}^T = \mathbf{W}\mathbf{\Sigma}^2\mathbf{W}^T \quad (6)$$

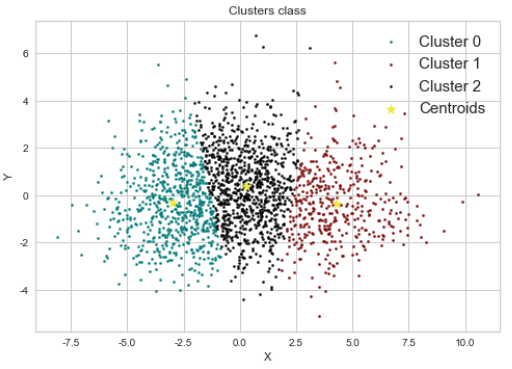
주성분분석을 이용한 차원축소를 통해 17개의 변수를 선형변환하여 PC1(첫 번째 주성분)과 PC2(두번째 주성분)을 도출하였다. 두 개의 주성분이 원본 데이터를 설명하는 설명력은 각각 57%, 18% 이며 첫 번째 주성분의 영향력이 강한 것을 확인하였다.

이후 1,987개의 패널데이터에 대한 K-means 클러스터링을 통해 군집화를 실행하기 위해 군집 k의 값을 설정하였다. 최적의 k를 찾기 위해 변동성을 확인하는 지표인 Eblow 메소드를 활용하였다. <그림 2>에서와 같이 Eblow 메소드의 적합된 3의 값으로 K-Means 군집분석을 실시하였다. 분석결과 <그림 3>과 같이 3개의 군집으로 나뉘었으며, 대체로 X축(PC1)에 의해 군집이 나뉘었음을 확인할 수 있다.

PC1과 PC2와 각 변인들과의 상관관계를 파악한 결과 PC1과 모든 변인들은 0.3이하의 수준에서 음의 상관관계를 보였으며 PC2는 과자류, 음료, 빵류, 간식류와는 0.3 수준의 음의 상관관계를 보였다. 하지만 군집화는 PC1의 수준에서만 분류된 것을 확인할 수 있었으며 주성분 분석 이후 군집화의 결과 주성분 1의 영향이 크면 PC1에 의해 대부분이 군집화 되었음을 알 수 있었다. <그림 3>에서 좌측에 위치한 군집0은 소비를 많이 하는 집단, 중앙에 있는 군집 1은 소비를 적당히 하는 집단, 우측에 있는 군집2는 소비를 적게 하는 집단으로 분류되었다. <표 3>의 소비식품 재분류 항목의 주성분요인 결과, PC1의 특성에서 각 변인들의 특징이 잘 나타나지 않아 집단의 특성을 반영한 군집의 명명을 결정할 수 없었다.



〈그림 2〉 Elbow Plot



〈그림 3〉 주성분분석을 통한 K-Means 군집분석

〈표 3〉 소비식품 재분류 항목의 주성분요인

소비식품 재분류 항목	PC1	PC2
축산물	-0.282	0.131
수산물	-0.217	0.354
조미채류	-0.237	0.366
과채류	-0.255	0.308
유제품	-0.289	-0.092
음료	-0.235	-0.315
엽경채, 나물, 근채, 서류	-0.264	0.327
주류	-0.165	-0.143
기타	-0.224	0.318
간식류	-0.278	-0.247
과자류	-0.242	-0.337
면류	-0.278	-0.187
빵류	-0.25	-0.263
수산, 축산물가공식품	-0.304	-0.055
기타가공식품	-0.306	-0.088

3.3 요인분석을 통한 차원축소

요인분석은 요인의 차원을 축소하여 다차원 데이터의 복잡함을 줄이는 것이 특징이다. 요인분석은 복잡하고 추상적인 개념을 간명하게 정리하여 다른 분석기법들과는 달리 독립변인과 종속변인의 개념들이 불필요하다. 또한 요인분석은 질적인 의미해석을 위해 양적인 방법에 의존한다. 주관적 해석을 모두 배제한 분석은 3.2에서 제안한 PCA에 가깝기 때문에 응용학문에 주로 사용되는 경향이 있다.

본 연구에서 요인분석 결과, 2개의 요인을 추출하였으며 factor1은 39.2%, factor2는 32.8%로 총 72%의 분산 설명력을 가진다. <그림 4> 요인회전은 가장 대중적인 기준인 배리맥스(VARIMAX)[4]를 사용하였다. 배리맥스 회전법은 “Variance is maximized”의 축약어로서 분산이 극대화된다는 것을 의미한다.

```
                Factor1 Factor2
SS loadings       5.886   4.918
Proportion Var    0.392   0.328
Cumulative Var    0.392   0.720
```

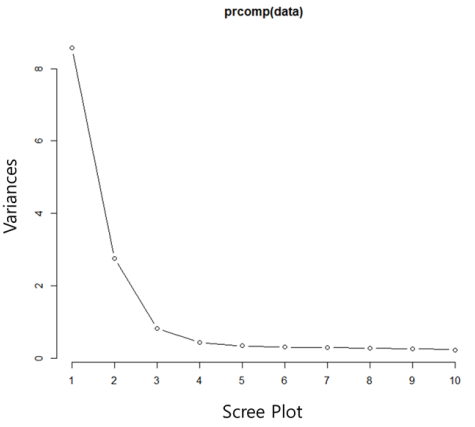
〈그림 4〉 요인분석의 분산설명력

〈표 4〉 소비식품 재분류 항목의 요인분석

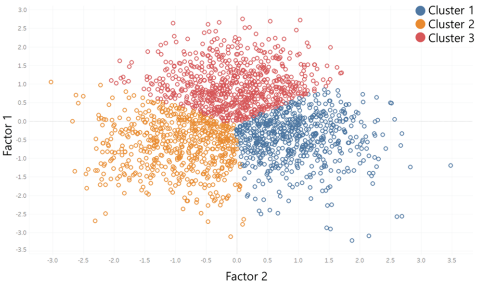
소비식품 재분류 항목	Factor1	Factor2
축산물	0.49	0.68
수산물	0.13	0.80
조미채류	0.14	0.92
과채류	0.26	0.85
유제품	0.74	0.42
음료	0.83	0.06
엽경채, 나물, 근채, 서류	0.24	0.93
주류	0.45	0.15
기타	0.19	0.77
간식류	0.89	0.21
과자류	0.89	0.05
면류	0.80	0.30
빵류	0.82	0.14
수산, 축산물가공식품	0.74	0.49
기타가공식품	0.78	0.46

<표 4>에서와 같이 Factor1은 가공식품 및 음료, 유제품에 영향을 많이 받으며 Factor2는 수산물 및 조미채류, 과채류, 엽경채류, 기타류에 영향을 많이 받는다고 판단할 수 있다.

요인분석으로 데이터의 차원을 축소한 후 <그림 5>과 같이 스크리플롯으로 최적의 군집개수 k를 확인하였다. 군집 k를 3으로 지정한 결과는 <그림 6>과 같다.



〈그림 5〉 요인분석의 Scree Plot



〈그림 6〉 요인분석을 통한 K-Means 군집분석

PCA를 통한 군집분석과는 달리, 요인분석 후의 군집분석 결과는 두 개의 Factor에 의해서 군집화가 되어 Factor1 과 Factor2 가 군집을 형성하는 데 있어 비슷한 영향을 미치는 것을 확인할 수 있었다. 군집의 특성 또한 확연히 차이가 나타나 <표 5>와 같이 통계적 근거에 의한 군집명명을 할 수 있었다.

〈표 5〉 요인분석에 따른 군집분석 결과

구분	Size	특성
군집1	629	수산물, 조미채류, 엽경채류, 기타류 많음
군집2	618	전체적으로 적게 소비 수산물,조미채류,엽경채류,기타류적음
군집3	740	가공식품, 음료, 유제품 많음

이론상 PCA와 요인분석의 결과가 서로 비슷해 지는 경우가 있을 수 있지만 PCA와 요인분석은 기초 논리가 명백히 다르다. 가장 큰 차이점은 PCA에서는 보통 제1,2주성분에서 1주성분을 주로 해석하고 2주성분을 보조로 활용한다. 하지만 요인분석에서는 연구목적에 맞게 요인의 수를 정하여 활용한다. 또한 요인분석에서는 고유요인을 인정하며 분석에 반영하는 것이 가장 큰 차이점이다.

Ⅳ. 결 론

본 연구에서는 1,987개 가구의 5년간 장바구니 데이터로 식품소비 특성에 기반한 가구 군집화를 위한 전처리 방법을 제안하고, 차원축소 방법 두 가지를 비교하였다. 일반적으로 최적의 해를 도출하는 것으로 알려져있는 주성분분석(PCA) 방법과 개인의 주관적 해석이 들어간다고 평가되는 요인분석(FA)을 비교하였다. 주성분 분석은 2개의 주성분이 75%의 설명력을 가졌으며 첫 번째 주성분이 50% 이상으로 군집화의 중심이 되었다. 첫 번째 주성분으로 3개의 군집으로 분류되었지만 각 군집의 소비패턴 특성은 잘 나타나지 않았다. 반면, 요인분석은 2개의 요인이 72%의 설명력을 가지며, 통계적 근거에 의한 군집 명명이 가능도록 군집의 특성이 명확하게 구분되었다. 따라서 본 연구에서 사용한 소비자 패널 데이터는 요인분석을 통한 군집분석이 군집의 특성을 나타내는데 효과적이었음을 제안한다. 주성분분석을 통한 군집화의 결과로 집단의 특징이 뚜렷하게 나타난다면 주성분분석을 이용한 군집화가 효과적일 수 있다. 하지만 본 연구에서는 주성분분석의 결과 첫 번째 주성분의 영향력이 강해 군집분석으로 분류한 군집들이 소비항목의 개수 등 데이터의 크기로 분류될 뿐 소비 항목의 패턴으로 분류되지 않을 때, 선정된 요인이라면 모두 동일하게 취급되는 요인분석을 차원축소의 방법으로 선택하여 더욱 효과적인 군집결과를 도출할 수 있다.

본 연구는 다차원 원본 데이터의 차원을 축소하는 방법으로써 주성분 분석과 요인분석을 비교하고 요인분석을 통한 군집분석이 군집의 특성을 나타낼 수 있는 경우 요인분석이 군집을 도출하는 데에 효과적인 차원축소의 방법이 될 수 있음을 보여주었다. 요인분석을 통한 차원축소 과정과 적용 방안을 설명하였으며, 이러한 요인분석은 데이터에 포함된 고유요인을 인정하여 분석에 반영하는 것으로 마케팅 분석, 심리학적 분석에 적용할 수 있다.

참 고 문 헌

[1] 보건복지부(2012), “0세 이상 성인 대사증후군 유병율 28.8%로 나타나!” 보건복지부 보도자료 (2012년 3월 22일)

[2] 한의신문(2018), “해마다 비만해지는 한국인...2030년 남성 비만율 62%, 여성 37%로 증가”, http://www.akomnews.com/bbs/board.php?bo_table=news&wr_id=15302

[3] 장민석, 김형중. (2018). 빅데이터를 활용한 은행권 고객 세분화 기법 연구. 한국디지털콘텐츠학회 논문지, 19(1), 85-91.

[4] 서현지(2017). 빅 데이터를 이용한 고객 행태 분석에 대한 연구. 국내석사학위논문 가천대학

교, 경기도.

[5] 김청택 (2016). 탐색적 요인분석의 오·남용 문제와 교정. 조사연구, 17(1), 1-29.

[6] Spearman, C. (1904). “General intelligence”: Objectively determined and measured. The American journal of psychology, 15(2), 201-292.

[7] Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In Clustering and classification (pp. 341-375).

[8] 장보영, 부소영. (2019). 군집분석으로 도출한 식사패턴별 에너지 섭취량과 골격근육량의 연관성 분석 : 2008~2010년 국민건강영양조사 자료를 활용하여. Journal of Nutrition and Health, 52(6), 581-592.

[9] 데니스 은주, 강민지, 한성림 (2017). 건강한 한국 성인의 음료섭취패턴과 대사증후군의 연관성 연구. 대한지역사회영양학회지, 22(5), 441-455.

[10] Berg CM, Lappas G, Strandhagen E, Wolk A, Torén K, Rosengren A, Aires N, Thelle DS, Lissner L. Food patterns and cardiovascular disease risk factors: the Swedish INTERGENE research program. Am J Clin Nutr. 2008 Aug; 88(2):289-97.

[11] 양혜리, 윤희용. (2018). 특수 데이터 집합에 대한 K-means clustering 알고리즘을 사용한 PCA 차원 감소. 한국정보기술학회 종합학술발표논문집, 315-318.

[12] 이용구, 양현일, 최정아, 허준. (2012). 화장품 추천 사례에서 요인, 군집분석을 이용한 협업 필터링 추천 모델과 연관성 규칙 기법의 성능 비교 연구, 14(2), 689-705.

[13] 김담희, 안가경. (2018). 머신러닝을 이용한 고객세분화에 관한 연구. 융복합지식학회논문지, 6(2), 115-120.

[14] 박미성, 안병일. (2014). 식품소비 라이프스타일이 가공식품 지출에 미치는 효과 분석: 군집분석과 매칭 기법을 이용하여. 농촌경제, 37(3), 25-58.

[15] Santos, R. D. O., Gorgulho, B. M., Castro, M. A. D., Fisberg, R. M., Marchioni, D. M., & Baltar, V. T. (2019). Principal component analysis and factor analysis: Differences and similarities in nutritional epidemiology application. Revista Brasileira de Epidemiologia, 22, e190041.

[16] Zubova, J., Kurasova, O., & Liutvinavičius, M. (2018). Dimensionality reduction methods: the comparison of speed and accuracy. Information Technology And Control, 47(1), 151-160.

저 자 소 개

홍 준 호 (Jun-Ho Hong)

- 2015년 : 고려대학교 응용통계학과 (학사)
- 2020년~현재: 충북대학교 빅데이터협동과정 석사
- 관심분야: 빅데이터, 머신러닝

오 민 지 (Min-Ji Oh)

- 2017년: 이화여자대학교 교육공학과 (학사)
- 2020년~현재: 충북대학교 빅데이터협동과정 석사
- 관심분야: 빅데이터, 머신러닝

조 용 빈 (Yong-Been Cho)

- 2019년: 충북대학교 빅데이터학과 (박사수료)
- 2015~현재: 농진청 농업빅데이터팀 과장
- 관심분야: 빅데이터, 스마트팜

이 경 희 (Kyung-Hee Lee)
- 2004년 : 충북대 컴퓨터과학과 (박사)
- 2008년~현재 : 충북대 빅데이터학과 교수
- 관심분야: 빅데이터, 알고리즘, 데이터마이닝

조 완 섭 (Wan-Sup Cho)
- 1996년: KAIST 전산학과 (박사)
- 1997년~현재: 충북대학교 경영정보학과 교수
- 관심분야: 데이터베이스, 빅데이터, 블록체인, 인공지능, 데이터 거버넌스