한국지능정보시스템학회논문지 제7권 제2호 2001년 12월(pp. 35~49)

인위적 데이터를 이용한 군집분석 프로그램간의 비교에 대한 연구

김성호
한양대학교 경영대학 경영학부
(kim007@hanyang.ac.kr)
백승익
한양대학교 경영대학 경영학부
(sbaek@hanyang.ac.kr)

인터넷 비즈니스나 전자상거래와 연관되어 고객관계관리 (Customer Relationship Management: CRM)에 대한 관심이 널리 확산됨으로 해서 군집분석에 대한 관심이 한층 높아졌고, 다양한 군집분석 프로그램이 시장에 소개되어 지고 있다. 그러나, 군집분석 프로그램들은 다른 데이터 분석 기법과는 달리 그들의 성능을 측정하기가 매우 힘들다. 본 논문에서는 이미 알려져 있는 군집구조를 지닌 인위적 데이터를 사용하여 다양한 군집분석 프로그램을 평가할 수 있는 하나의 방법론을 제시하고, 그 방법론의 유용성을 보여 주기 위해 현재 많이 사용하고 있는 네 가지의 군집분석 프로그램들을 본 논문에서 제시한 방법론을 사용하여 평가하는데 그 주요 목적을 두고 있다. 본 연구에서 두 가지의 반복적 군집분석 프로그램 (Convergent Cluster Analysis: CCA, SPSS의 Clementine), 전통적인 단순군집 프로그램 (One-Shot Clustering Program: Howard-Harris 프로그램), 그리고 IBM의 데이터 마이닝 기법 중 하나인 데모그래픽 군집분석 프로그램의 성능을 비교한 결과, 군집분석을 위하여 다른 군집분석 방법 보다 좀 더 지능적으로 초기치를 생성한 CCA 방법이 가장 우월한 성능을 보여 주었다.

1. 서론

군집분석은 데이터의 상호 유사성을 기반으로 여러 개의 집단으로 나누는 다변량 분석 방법의 하나이다. 이 분석 방법의 주요 목적은 군집간의 이질성과 군집 내에서의 대상물들의 동질성을 극대화하는 데 있다. 오랫동안 여러 분야에서 이 군집 분석에 대한 다양한 연구가(Joyce & Channon 1966; Green, Frank & Robinson 1967; Frank & Green 1968; Green & Krieger 1991) 되어졌을 뿐만 아니라, 때로는 많은 논쟁을 불러 일으켰던 역사(Neidell, 1970)를 지니고 있다. 이런 과정을 거치면서 군집분석은 마케팅 조사자들에 의해 구매자, 제품, 브랜드, 혹은 구매자의 구매 상황 및 제품 사용상황 등을 기초로 모집단 내에서 동질적인 특성을 지니고 있는 집단을 발굴하기 위한 실증적인 도구로 그 자리를 굳히고 있다. 특히 인터넷 비즈니스나 전자상거래와 연관되어 고객관계관리(Customer Relationship Management: CRM)의 중요성이 널리 확산됨으로 해서 시장 세분화와 표적 마케팅을 실행하기 위한 하나의 방법론으로서 군집분석을 많이 사용하고 있다.

현재 많이 사용되어지고 있는 군집분석 방법으로서는 계층적 군집분석방법(Hierarchical Clustering Models; Blashfield 1976), 계층적 군집분석 방법과 분할적 군집분석방법의 병용(Combination of Hierarchical Clustering and Partitioning Methods; Milligan & Sokol 1980), 중복군집분석 방법(Overlapping Clustering; Arabie, et al., 1981, Srivastava, et al., 1984), 혼합 모형(Mixture Models; Wolfe 1970, McLachlan & Basford 1988), 그리고 K-군집 중앙치을 이용한 중복군집분석(K-Cntroid Overlapping Custering; Chaturvedi, et al., 1997) 등을 들 수 있다.

2. 군집분석의 종류

Berson et al.(2000)과 Jain et al.(1999)은 군집분석 방법을 크게 계층적 군집분석 방법(Hierarchical Clustering)과 비계열적 혹은 분할적 군집분석 방법(Non-Hierarchical 혹은 Partitioning Clustering)으로 분류하고 있다.

2.1 계층적 군집분석(Hierarchical Methods)

계층적 군집분석은 개체들의 분류에 가장 많이 이용되고 있는 군집분석 알고리즘 가운데 하나이다. 이 방법은 군집분석 대상(개체) 각각을 하나의 군집으로 생각하는 데에서부터 출발하며 군집분석을 실행하는 과정에서 우선 가장 유사한 두 개의 대상을 하나의 군집으로 묶는다. 이런 과정을 연속적으로 실시하여 결국에는 모든 응답자들이 하나의 군집으로 묶이게 된다. 그리고 군집분석 결과는 Dendrogram이나 Tree-Structure로 나타나게 된다. 이 때 연구자는 적절한 군집의 수를 결정하게 된다. 계층적 군집분석은 군집을 구성하는 방식에 따라 다시 단일기준결합방식(Single Linkage Method), 완전기준결합방식(Complete Linkage Method), 평균기준결합방식(Average Linkage Method), 그리고 Ward(1963)의 최소분산방식(Ward's Minimum Error Variance Method)으로 분류 할 수 있다.

단일기준결합방식은 가장 가까운 거리에 있는 두 개체 혹은 두 군집을 새로운 군집으로 묶는 방식이며 완전기준결합방식은 반대로 가장 먼 거리에 위치해 있는 두 개체 혹은 두 군집을 새로운 군집으로 묶는 방식이다. 평균기준결합방식에서는 각 군집에 속해 있는 개체간의 거리의 평균을 구한 뒤 가장 가까운 평균거리를 지니고 있는 두 군집이 새로운 군집을 형성하는 방법으로, 일반적으로 단일기준결합방식이나 완전기준결합방식에 비하여 보다 정확하게 군집을 도출하는 방법이다. 그러나 계층적 군집분석에 있어서 가장 군집의 도출능력이 우수한 것은 Ward(1963)의 최소 분산 방식이다. 계층적 군집분석의 특징으로는 첫째, 서로 중복되는 군집이 있을 수 없고 둘째, 일단 서로 같은 군집으로 묶이면 절대 다시 분리 되지 않는다는 것이다.

2.2 분할적 군집분석(Partitioning Methods)

분할적 군집분석 또한 서로 중복되지 않는 군집(Non-Overlapping Clusters)을 도출해 내는 군집분석 방법이다. 이 방법은 종종 비계층적 군집분석(Non-Hierarchical Clustering)이라고도 불리는데 이는 계층적 군집분석이 순차적인 과정을 거쳐 군집을 구성하는 것과는 달리 분할적 군집분석은 단 한번의 분할 과정만을 거치기 때문이

다(Anderberg 1973; Sneath & Sokal 1973). 즉 분할적 군집분석에서는 여러 계층의 군집구조가 형성되지 않고, 단일 계층의 군집이 형성되게 된다. 분할적 군집분석 방법에는 크게 두 가지의 종류의 방법이 있다: 단일 분할적 군집분석 방법(Single Pass Method)과 재분할군집 방법(Reallocation Method) (Berson et al., 2000). 단일 분할적 군집분석 방법은 단 한번 군집대상을 읽어서 분할하도록 하는 반면에, 재분할군집 방법은 하나의 군집대상을 여러 번 읽어 더 좋은 군집에 분류하기 위해서 계속해서 그 군집대상을 다른 군집에 분류해 나간다. 분할적 군집분석방법 중 가장 많이 이용되고 있는 방법은 K-평균 분할적 군집분석(K-Means Partitioning Method)이다.

3. K-평균 군집분석

3.1 K-평균 군집분석 방법의 장점

시장조사와 시장 세분화 연구자들이 다른 군집분석 방법과 비교하여 K-평균 군집분석을 많이 사용하는 이유는 다음과 같이 설명할 수 있다(Arabie et al., 1981; Srivastava et al., 1984). 첫째, 마케팅조사와 시장 세분화 연구자들은 많은 수의 응답자와 변수를 다루고 있기 때문에 대규모의 데이터를 효율적으로 처리할 수 있는 분석기법이 필요할 것이다. 소비자 패널 데이터, 제품 스캐너로부터 수집되는 가구(家口) 단위의 구매 데이터, 혹은 동일한 조사대상에 대하여 장기적으로 수집되는 라이프 스타일(Lifestyle)에 관한 자료 등이 대표적으로 마케팅 조사나 시장 세분화를 위해서 많이 사용하는 데이터일 것이다. K-평균 군집분석은 이와 같이 수많은 응답자와 변수를 지닌 대규모의 자료를 빠른 시간 내에 효율적으로 분석할 수 있는 분석 기법이다(Berson et al., 2000). Chaturvedi et al. (1997)는 실험을 통하여 MAPCLUS(Arabie et al., 1981), INDCLUS (Carroll & Arabie, 1983), SINDCLUS (Chaturvedi & Carroll, 1994), 그리고 GENNCLUS(DeSarbo, 1982)과 같은 계층적 군집분석 프로그램들은 500개 이상의 샘플 크기에서는 사용할 수 없음을 발견하였다. 둘째, 마케팅 관리자들은 세분시장의 수를 결정하고 각 세분시장의 구성원들의 속성을 파악하는 과정에 있어서 중복 군집(Overlapping Cluster) 혹은 계층적 군집(Hierarchical Cluster) 방법보다는 K-평균 군집분석 방법을 선호하는 경향이 있다. 그들은 효과적이고 효율적인 마케팅전략의 수립을 위하여 한 사람의 구매자가 오직 하나의 군집 (세분시장)에만 소속되기를 원한다. 그들은 또한 군집분석을 통하여 나타난 세분시장들이 몇 개의 의미 있고, 요약된 통계(Summary Statistics, 예를 들면 군집평균, Cluster Centroid)에 의하여 선명하게 설명 되어지기를 원한다. 예를 들면, 마케팅 관리자들은 어느 세분시장을 공략할 것인가에 따라 광고 주제를 개발하고 광고 매체를 선택 할 것이다. K-평균 군집분석은 서로 비슷한 크기의 군집을 형성하고 마케팅 관리자들이 쉽게 이용할 수 있는 군집(세분시장)의 특성들을 보여줌으로써 마케팅 관리자들의 이러한 욕구를 충족시켜 주는 방법이라 할 수 있다. 마지막으로, K-평균 군집화 방법을 지원하는 상용화된 제품이 많아 용이하게 마케팅 관리자들이 시장 세분화와 포지션 전략을 수립하기 위해서 많이 사용하고 있다. 현재 시장에서 군집분석 툴로 많이 사용되고 있는 SAS의 E-Miner와 SPSS의 Clementine에도 K-평균 군

집화 방법을 포함하고 있다.

3.2 K-평균 군집분석 방법의 단점

K-평균 군집분석이 널리 사용되어 있음에도 불구하고, 군집분석을 시장 세분화에 이용하는 데에는 몇 가지 실무적인 문제가 있다. K-평균 군집분석 방법의 가장 큰 문제점은 군집분석의 정확도가 군집분석을 위해 시장 분석가들이 임의로 정해야 하는 변수(Parameters)들의 값에 따라 민감하게 변화된다는 점이다(Schaffer & Green, 1998). 예를 들어, 군집의 수는 몇 개로 할 것인지, 군집분석의 어떤 알고리즘을 사용할 것인지, 입력변수는 어떤 것을 선택할 것인지, 입력변수 혹은 응답자를 표준화할 것인지 등에 따라 같은 자료를 가지고서 군집분석을 할지라도 정확도에 큰 차이가 있을 것이다. 그 중에서도 특히 군집을 구성하기 위해서 고려해야 할 변수의 선택은 군집분석의 정확도를 가장 크게 좌우하는 요소이다. 그 이유는 선택된 많은 변수 중 비록 한 · 두 개의 비관련 변수(Irrelevant Variables)만 있더라도 실재 존재하는 군집구조(시장세분화)와는 전혀 다른 결과를 도출해 낼 수 있기 때문이다(Milligan & Cooper 1986). 그밖에도 K-평균 군집분석을 위해 사용되는 초기치(Initial Seed)에 따라 분석 결과가 상이할 수가 있을 것이다. 즉, 동일한 자료에 대해서도 초기치를 달리하면 서로 다른 결과가 도출될 수 있을 것이다. 많은 분석가들은 수많은 시도에서 습득한 그들만의 노하우를 기초로 이들 변수들을 선택하고 있다. 이런 단점을 보완하기 위하여 Sawtooth Software社의 Convergent Cluster Analysis(CCA), CONCLUS 프로그램(Helsen & Green 1991) 등은 반복적으로 군집분석을 하여 가장 최적의 군집을 찾아내는 방법을 사용하고 있다. 많은 시장 연구자들은 K-평균 군집방법의 정확도를 높이기 위하여 여러 가지 다양한 초기치 결정방법과 군집 간의 평균 계산방법을 소개하고 있다(Berry & Linoff, 1997).

K-평균 군집분석 방법과 연관되어진 많은 기존의 연구에서는 군집의 수, 군집의 크기, 초기값 설정 방법과 같은 변수에 의하여 군집분석 방법의 정확도가 차이가 있음을 지적하고 있음에도 불구하고, 실증적으로 위의 변수들이 분석방법의 정확도에 어떤 영향을 미치는지에 대해서 보여준 연구는 극소수에 불과하다. 특히, 군집분석 방법의 성과를 측정할 수 있는 방법론이 제시되어 있지 못하다. 본 연구에서는 군집분석 프로그램들의 성과를 비교할 수 있는 하나의 방법론을 제시하고, 이 방법론을 사용함으로써 현재 많이 사용하고 있는 세 가지 종류의 K-평균 군집분석 프로그램(CCA, Howard-Harris, 그리고 SPSS의 Clementine)과 K-평균 군집분석 프로그램이 아니지만 많이 사용하고 있는 IBM의 Intelligent Miner의 성과가 군집의 수, 군집의 크기, 그리고 초기값에 얼마나 민감하게 변하는지를 탐색하는 데 주요 목적이 있다.

4. 연구목적

본 연구의 주요 목적은 시장 세분화에 가장 많이 사용되고 있는 분할적 군집분석 기법인 K-평균 군집분석 방법을 이용한 세 개의 프로그램(Howard-Harris, Convergent Cluster, 그리고 SPSS의 Clementine)과 K-평균 군집분석 방법이 아닌 IBM사의 Intelligent Miner에 제공하는 데

모그래픽 군집분석 프로그램(Demographic Cluster Analysis)의 성과를 인위적으로 생성되어진 데이터를 사용하여 비교하는데 있다. 구체적으로, 본 연구의 목적은 시장 세분화 및 마케팅조사에서 가장 빈번하게 사용되고 있는 세 개의 K-평균 군집분석 프로그램과 Intelligent Miner의 군집분석 프로그램을 군집의 구조가 이미 알려져 있는 인위적자료에 사용하여 그 가치를 비교 평가하는 탐색적 연구를 수행하는데 있다. 본 연구의 구체적인 연구 사항은 다음과 같다.

1. 네 개의 군집분석 프로그램(CCA, Howard-Harris, Clementine, Intelligent Miner) 중에서 어느 프로그램이 이미 알려져 있는 군집의 구조를 가장 잘 도출하는가?
2. 반복적 군집분석 프로그램(CCA)과 전통적인 K-평균 군집분석 프로그램(Howard-Harris), Clementine, 그리고 Intelligent Miner에 알려져 있는 군집구조의 도출에 있어서 유의적인 차이가 있는가?
3. 네 개의 군집분석 프로그램과 군집의 수, 표본의 크기, 군집을 구성하는 변수의 수, 그리고 소음형태 등의 실험요인 중 어느 것이 유의적인가?
4. 본 연구에 사용된 실험 요인간에 유의적인 상호작용이 있는가?

본 연구를 통해서 위의 연구 과제를 수행하기 위한 방법론을 제시하고자 한다.

5. 연구 대상 군집분석 프로그램

5.1 Howard-Harris 프로그램

Howard-Harris 프로그램은 P개의 변수에 의하여 정의된 N 명의 응답자행렬(N x P matrix)를 Euclidean 거리척도를 사용하여 K개의 집단으로 분할하는 K-평균 군집분석 프로그램이다. 현재 시장 분석가들이 시장 세분화를 위하여 가장 많이 사용하고 있는 프로그램 중 하나이다. 응답자간의 Euclidean거리는 다음과 같이 구한다.

$$d_{ij}^2 = \sum_{t=1}^{P} (x_{it} - x_{jt})^2 \qquad \text{식 (1)}$$

d_{ij} = 응답자 i와 응답자 j 간의 Euclidean 거리척도
x_{it} = 응답자 i의 속성 t의 평가

이들 응답자간의 Euclidean거리의 합은 집단의 평균으로부터의 편차의 합에 직접적으로 비례한다.

$$\sum_{i=1}^{n_k} (x_i - \mu_k)^2 = \frac{1}{2} n_k \left[\sum_{i=1}^{n_k} \sum_{j=1}^{n_k} (x_i - x_j)^2 \right] \qquad \text{식 (2)}$$

μ_k = 집단 k의 평균(group centroid)
n_k = 집단 k의 크기

모든 응답자로부터 구한 총분산은 다음과 같이 나타난다.

$$V_t = \frac{1}{2} N \left[\sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2 \right] \qquad \text{식 (3)}$$

총분산은 집단 내 분산(Vw)과 집단간 분산(Vb)으로 나누어지며 각각의 집단 내 분산은 식(2)에 나타나 있다.

$$V_t = V_b + V_w \qquad \text{식 (4)}$$

따라서 집단 내 분산의 합은 다음과 같다.

$$V_w = \sum_{k=1}^{K} V_k \qquad \text{식 (5)}$$

N명의 응답자를 K개의 집단으로 나누기 위한 기준은 단순히 집단 내 분산(Vw)을 최소화하는 분할 P(N,K)를 구하면 된다. 본 연구에서는 PC용 Howard-Harris 프로그램 (Smith 1990)을 사용하였다. Howard-Harris 군집분석 프로그램은 단일 군집분석 프로그램이다. 단 한 번의 시도로 군집을 구성하는 방법이기 때문에, 이 방법의 성능은 재분할 군집분석 프로그램에 비해서 초기치 설정에 매우 민감하게 반응할 것이다.

5.2 Convergent Cluster Analysis(CCA) 프로그램

CCA는 PC용 군집분석 프로그램으로 사용자가 지정하는 회수(M회)의 군집분석을 실행한 후 각각의 군집분석 결과와 나머지(M-1)개의 군집분석 결과를 군집의 재생성(再生性, Reproducability)을 사용하여 비교한 후, 가장 높은 평균 재생성을 지닌 군집분석결과를 최적(그리고 최종) 군집분석 결과로 선정하는 K-평균 군집분석 프로그램이다. 재생성이란 두 개의 군집분석 결과를 K x K 표(Cross-Tabulation Table; K는 군집의 수를 나타냄)로 나타내어 대각선상에 위치하는 응답자들의 수의 합을 말한다. CCA와 같은 반복적 군집분석방법은 대부분의 K-평균 군집분석 프로그램들이 초기치에 민감하고 따라서 최종적인 군집분석의 결과가 동일한 자료에 대해서도 상이할 수 있다는 단점을 해결할 수 있다. CCA는 다음과 같은 다섯 가지의 초기치 설정 옵션을 가지고 있다.

- Distance-Based: 이 방법은 비교적 멀리 떨어져 있는 응답자들을 초기치로 설정하는 방법으로 가장 빠르게 군집분석을 수행하는 옵션이다. 그러나 이 방법을 사용했을 경우 항상 동일한 군집분석의 결과가 나타나기 때문에 반복적으로 군집분석을 수행할 필요가 없다.
- Hierarchical-Based: 데이터에 있는 응답자의 수가 50명 이상일 경우에는 50명의 응답자를 무작위 추출하여 계층적 군집분석(Complete Linkage)을 수행하여 각 군집의 군집평균(Cluster Centroid)을 초기치로 택하는 방식이다. 데이터의 응답자의 수가 50명 미만일 경우는 전체 데이터를 초기치 설정에 사용한다.
- Density-Based: Hierarchical-Based 방법과 마찬가지로 초기치 설정을 위하여 50명의 응답자를 무작위 추출한다. 이 방법에서는 응답자들이 비교적 밀집된 지역의 중심에 가까이 위치한 응답자를 무작위 하부자료로 추출한다.
- Mixed Strategy: 첫번째 군집분석은 Distance-Based 옵션을 사용하여 실행되고 홀수번째의 반복적 군집분석은 Density-Based 방법을 사용하여, 그리고 짝수 번째의 반복적 군집분석은 Hierarchical-Based 옵션을 사용하여 실행된다. Sawtooth사는 대부분의 K-평균 군집분석에 있어 이 방법을 사용할 것을 권하고 있다.
- 사용자 정의 옵션: 이 방법에서는 프로그램 사

용자가 제공하는 군집소속에 따라 계산된 군집평균을 초기치로 설정한다. 이 방법을 사용하기 위해서는 CCA의 실행 이전에 실행한 군집분석의 결과(군집소속 데이터)가 필요하다.

본 연구에서는 초기치 설정을 위해서 Mixed Strategy를 사용하였다. CCA는 재분할 군집분석 방법으로 여러 가지 방법으로 초기치를 설정할 수가 있기 때문에, 단일 분할 군집분석 방법인 Howard-Harris 방법보다는 그의 성능이 초기치의 설정에 그리 민감하게 반응하지 않는다.

5.3 SPSS의 Clementine

SPSS의 Clementine은 각 군집대상을 가장 가까운 군집평균을 지닌 군집에 소속시키도록 군집을 하게 된다. 이 과정에서 군집의 평균은 반복적으로 추정되면서 군집대상은 다른 군집에 재할당되어 진다. 다시 말해, Clementine는 군집내의 동질성(Within-Cluster Homogeneity) 및 군집간의 이질성(Between-Cluster Heterogeneity)을 극대화하도록 반복적으로 군집을 만들어 간다(Aldenderfer & Blashfield, 1978). 다른 K-평균 군집분석 프로그램과 같이 군집대상 간의 동질성과 이질성을 판단하기 위한 근접척도로서 Euclidean 거리척도 방법을 사용하였다. Clementine에 있어서는 CCA와 같이 재분할 군집분석 방법을 사용하였으나, 초기치 설정에 있어서는 Howard-Harris에서와 같이 Random하게 초기치를 설정하여 사용하였다.

5.4 IBM의 Intelligent Miner

Intelligent Miner는 모든 데이터 베이스에 있는 레코드들을 한 번에 두 개씩 비교하여 두 데이터 항목의 유사성을 계산하고 그 값을 기초로 하여 군집을 형성하게 된다. 레코드들 사이의 유사성은 그 레코드의 필드 값을 비교하여 결정하게 된다. 그런 다음 동일한 군집내의 모든 레코드 유사성 쌍의 합에서 다른 군집내의 모든 유사성의 합을 뺀 값을 최대화하도록 연속적으로 군집을 만들게 된다. 위에서 설명한 K-평균 군집분석 방법과는 달리, Intelligent Miner는 초기값을 결정할 필요가 없이 군집대상 각각을 비교하여 군집내의 동질성과 군집간의 이질성을 극대화하도록 반복적으로 군집을 만들어 간다.

<표 1>은 4가지 군집분석 방법의 차이점을 요약해 놓았다.

<표 1> 군집분석 방법의 비교

군집분석방법	알고리즘	초기치 설정	차기 초기치 설정	분석 횟수	거리 측정
Howard-Harris	K-Mean	Random	해당 사항 없음	단일	Euclidean
CCA	K-Mean	Mixed Strategy	각 군집의 평균	반복	Euclidean
Clementine	K-Mean	Random	각 군집의 평균	반복	Euclidean
I-Miner	Demographic Algorithm	초기치 설정이 필요하지 않음	각 군집의 평균	반복	Matching

6. 연구방법

본 연구에서는 Monte Carlo 방법을 이용하여 이미 알려져 있는 군집구조(Cluster Structure)를 가진 인위적인 데이터(Synthetic Data)를 네 가지의 다른 군집분석 프로그램을 사용하여 분석함으로써, 어떤 변수가 군집분석 프로그램의 성과에 많은 영향을 미치는지, 어떤 군집분석 프로그램이 가장 정확하게 군집을 도출하는지를 평가해 보았다.

6.1 군집조성 프로그램

본 연구에서는 이미 알려져 있는 군집구조를 지닌 인위적인 데이터를 만들기 위해 Milligan (1985)에 의해 개발된 군집조성 프로그램(Cluster Generation Program)을 사용하였다. 본 연구를 위하여 인위적으로 만들어진 데이터는 다음과 같은 특성을 지니고 있다.

- 군집조성 프로그램은 각 분석마다 2개에서 5개 사이의 군집을 조성한다. - 군집의 수
- 군집들은 4개, 6개, 혹은 8개의 차원(변수)에 의하여 나타내어 진다. - 군집을 구성하는 변수의 수
- 각각의 군집은 100개 와 200개의 개체로 구성된다. - 군집의 크기
- 소음(Noise)을 추가함으로써 원래 군집을 구성하고 있는 변수의 좌표에 오차(Error)를 첨가할 수 있다. - 기본좌표의 혼란
- 한 개, 두 개, 혹은 세 개의 소음차원(Noise Dimension)을 추가함으로써 전체 데이터에 오차(Error)를 첨가할 수 있다. - 추가적인 소음차원

이 프로그램은 다변량 정상분포 (Truncated Multivariate Normal Distribution)로부터 표본을 추출함으로써 군집을 구성하는 점들의 좌표를 도출해 낸다. 이 과정에서 각 군집의 군집평균(Cluster Centroid)의 ±1.5 표준편차 내에 있는 모든 점으로부터 특정한 군집을 구성하는 점들을 도출해 낸다. 그리고 군집을 나누는 기준에 의해 군집간에 서로 겹치는 부분은 없도록 군집을 구성한다 (Non-Overlapping Clusters).

군집을 구성하는 변수에 오차(error)가 첨가된 좌표값(E_{ij})은 다음과 같다.

$$E_{ij} = T_{ij} + \lambda \varepsilon_{ij} \qquad \text{식 (6)}$$

T_{ij} = 군집을 구성하는 점 i의 j번째 오차 없는 기본 좌표(Error-Free Coordiante)
ε_{ij} = 무작위 추출된 오차(Random Error)
λ = 소음의 강도

6.2 실험설계

본 연구에서는 인위적 데이터의 Monte Carlo 실험을 위하여 다음과 같은 실험 변수들을 사용하여 실험을 설계하였다.

1. 자료의 크기(2개의 변수): 100; 200
2. 군집의 수(4개의 변수): 2개; 3개; 4개; 5개
3. 군집을 구성하는 변수의 수(3개의 변수): 4; 6; 8
3. 오차의 형태(5개의 변수)
   a. 기본좌표의 혼란: λ=0(Error Free Situation); λ=1; λ=2
   b. 추가적인 소음 차원(Additional Noise

Dimensions): 1차원; 2차원

4. 군집분석 프로그램(4개의 변수): Howard-Harris; CCA; Clementine; Intelligent Miner

위의 실험 변수들을 사용한 교차 디자인 (Full Factorial Design)으로부터 Monte Carlo 실험을 위한 144개의 데이터를 구성하였다. 본 연구에서는 그 중 50개의 Data Set을 가지고서 Monte Carlo 실험을 하였다. 각각의 데이터는 네 개의 군집분석 프로그램에 의하여 분석되었다.

6.3 군집구조의 추출(Recovery) 측정척도

본 연구에서는 두 개의 군집(알려져 있는 군집구조와 군집분석 프로그램을 사용하여 도출한 군집구조)간의 일치도를 알아보기 위하여 Adjusted Rand Index(Hubert & Arabie 1985; ARI)를 이용하였다. Rand Index란 두 개의 빈도간의 비율로서 분자(分子)는 동일한 두 명의 응답자 쌍이 동일한 군집에 소속되어 있는가 혹은 상이한 군집에 소속되어 있는가의 빈도수이며 분모(分母)는 전체 응답자 쌍의 수이다. 예를 들면, 전체 응답자의 수가 N일 경우 분모는 N(N-1)/2이다.

<표 2> Rand와 Adjusted Rand Index

True Structure (알려져 있는 군집구조)	Test Structure (군집분석을 통하여 발견한 군집구조)		
	동일군집에 속한 개체의 쌍	상이한 군집에 속한 개체의 쌍	합
동일군집에 속한 개체의 쌍	A	B	A+B
상이한 군집에 속한 개체의 쌍	C	D	C+D
합	A+C	B+D	R

만일 두 개의 군집이 서로 정확하게 일치한다면 Rand Index는 1.0이다. 만일 군집의 구성원들간에 일치가 전혀 이루어지지 않는다면 Rand Index는 0이다. 그러나, 원래 Rand Index에는 상향적 오차가 존재하므로 보통 그 오차를 수정하여 사용하게 된 것이 Adjusted Rand Index (ARI)이다. <표 2>는 Rand와 ARI를 구하는 수식을 보여주고 있다.

Original Rand Index:

$$\frac{(A+D)}{R} = \frac{(A+D)}{\frac{1}{2}N(N-1)} \qquad \text{식 (7)}$$

Adjusted Rand Index

$$\frac{R(A+D)-[(A+B)(A+C)+(C+D)(B+D)]}{N^2-[(A+B)(A+C)+(C+D)(B+D)]} \qquad \text{식 (8)}$$

$$R=\frac{1}{2}N(N-1)$$

7. Monte Carlo 실험 결과

<표 3>은 분산 분석의 결과를 보여 주고 있다. 본 연구의 Monte Carlo 실험결과 오차의 형태, 군집의 수, 그리고 군집 프로그램은 통계적으로 유의한 것으로 나타났다. 그리고, 군집의 크기, 군집을 구성하는 차원 (변수)의 수는 통계학적으로 유의하지 않은 것으로 나타났다. 즉, 오차의 형태, 군집의 수, 그리고 군집 프로그램에 따라 군집 구조의 추출 측정 척도인 ARI가 다르다는 것을 발견하였다. 반면에, 군집의 크기, 군집을 구성하는 차원(변수)의 수는 크게 ARI에 영향을 미치지 않는 것으로 나타났다.

<표 3> 분산분석결과

Source of Variation	DF	Mean Square	F Value	P-Value
Model	13	0.2158	5.43	0.001
자료의 크기	1	0.0489	1.23	0.2685
오차의 형태	4	0.2418	6.65	0.0001***
군집 구성 차원의 수	2	0.1133	2.85	0.0603
군집의 수	3	0.1494	3.76	0.0119**
군집 분석 프로그램	3	0.3714	9.34	0.0001***

<표 4>는 각각의 실험요인의 평균 ARI를 보여 주고 있다. 본 연구에 사용된 네 개의 군집 분석 프로그램 가운데서, CCA(0.9548), Clementine (0.9524), Howard-Harris(0.8242), 마지막으로 Intelligent Miner(0.7868)의 순서로 이미 알려진 군집 구조를 정확하게 추출하였다. 반복적 군집분석 프로그램인 CCA와 SPSS의 Clementine는 예상했던 바와 같이 이미 알려져 있는 군집구조의 도출에 있어서 전통적인 단순 군집분석 프로그램인 Howard-Harris보다 우수한 것으로 나타났다. 추가적으로, CCA의 초기치 설정 전략이 다른 K-평균 분석방법의 설정 전략보다 좀 더 체계적 (Random Strategy가 아닌 Mixed Strategy 사용)이어서 네 개의 군집분석 프로그램 중 가장 좋은 결과를 얻었다. 그리고, K-평균 군집분석 프로그램이 아닌 Intelligent Miner는 다른 K-평균 군집분석 프로그램보다 군집구조를 도출하는데 있어서 정확도가 상대적으로 저조하였다. Intelligent Miner는 군집분석을 위한 초기값의 설정이 없이 모든 군집대상을 하나씩 비교하여 반복적으로 군집을 구성하였음에도 불구하고 다른 두 개의 반복적 군집분석 방법보다 열등한 결과가 나온 것은 인위적인 데이터가 가지는 특수한 속성에서 기인되었을 것이다. 본 연구에서 인위적인 데이터 생산을 위하여 사용한 Milligan의 군집 조성 프로그램은 Euclidian 거리척도를 기초로 하여 인위적인 데이터를 생산하였기 때문에 Euclidian 거리척도를 기준으로 군집을 구성하는 K-평균 군집분석 프로그램이 그렇지 않은 Intelligent Miner보다 좋은 결과를 얻은 것은 당연할 것이다.

<표 4>는 예상했던 바와 같이 오차의 강도 (λ)가 증가함에 따라 군집 구조의 도출이 낮아지는 경향을 보이고 있다. 변수의 오차 강도가 증가함에 따라서 ARI는 현격히 감소하였다. 또한 군집의 크기가 증가함에 따라 ARI의 값이 증가함을 보여 주고 있다. 가장 완벽한 군집분석은 아마도 하나의 대상이 하나의 군집을 형성하도록 할 때일 것이다(ARI = 1). 그리고, 가장 나쁜 군집분석은 모든 군집 대상을 하나의 군집에 모두 포함시킬 때일 것이다(ARI=0). 즉, 군집의 수를 증가시키면 자연히 ARI의 값은 증가할 것이다.

<표 4> Monte Carlo 실험 결과

실험 요인	요인 수준	평균 ARI	분산
오차의 형태	λ = 0	0.9604	0.1389
	λ = 1	0.9599	0.1565
	λ = 2	0.8741	0.2253
	추가소음차원 1개	0.8095	0.2603
	추가소음차원 2개	0.7984	0.2737
군집의 수	2개	0.8234	0.2729
	3개	0.9178	0.1857
	4개	0.9083	0.1915
	5개	0.9317	0.1661
군집구성 차원의 수	4	0.8465	0.2420
	6	0.8664	0.2405
	8	0.9206	0.1989
자료의 크기	100	0.8649	0.2531
	200	0.8964	0.1961
군집분석 프로그램	CCA	0.9548	0.1315
	Howard-Harris	0.8242	0.2489
	Intelligent Miner(IBM)	0.7868	0.3136
	Clementine(SPSS)	0.9524	0.1011

<표 5>는 군집분석 프로그램과 다른 실험요인 간의 상호작용을 포함한 분산분석의 결과를 보여주고 있다. 군집분석 프로그램과 오차의 형태, 군집분석 프로그램과 군집의 수간의 상호작용만이 통계적으로 유의한 것(p=0.0001)으로 나타났다. 다른 실험요인 간의 모든 상호작용은 통계적으로 유의하지 않았다.

<표 5> ANOVA 분석 결과

SOURCE	DF	ANOVA SS	MEAN SQUARE	F VALUE	Pr>F
MODEL	34	5.8952	0.1733	6.68	0.001
오차의 형태	4	0.9675	0.2418	9.33	0.0001
군집구성 차원의 수	2	0.2267	0.1139	4.37	0.0141
군집의 수	3	0.4482	0.1494	5.76	0.0009
자료의 크기	1	0.0489	0.0489	1.89	0.1712
군집분석 프로그램	3	1.1143	0.3714	14.32	0.0001
오차의 형태 * 군집분석 프로그램	12	1.0875	0.0906	3.49	0.0001
군집의 수 * 군집분석 프로그램	9	2.0017	0.2224	8.57	0.0001
ERROR		4.2282	0.0259		
TOTAL	197	10.1234			

<표 6>은 군집분석 프로그램과 오차의 형태 간의 유의적인 상호작용에 따른 평균 ARI를 보여주고 있다. 소음이 없는(λ = 0) 대상을 군집화하는데 있어서 반복적 군집분석 방법보다는 Howard-Harris와 같은 단순 군집분석 방법이 이미 알려져 있는 군집구조를 정확하게 도출하였다. 그러나, 소음을 추가함으로써(λ = 1, λ = 2), 반복적 군집분석 방법이 단순 군집분석 방법보다 군집 추출에 있어서 우월하게 되었다. 군집대상의 오차 정도가 증가할수록 군집분석의 정확도는 초기치(Initial Seed)에 매우 민감하게 되고, 이런 문제를 해결하기 위해 반복적으로 군집을 구성하여 최적의 군집구조를 도출해 내는 CCA나 Clementine이 상대적으로 좋은 결과를 얻어냈다.

<표 6> 군집분석 프로그램과 오차 형태의 상호작용

오차의 형태	군집분석 프로그램			
	CCA	Howard-Harris	Intelligent Miner	Clementine
λ=0	0.9874	0.9940	0.8760	0.9843
λ=1	1.0000	0.9970	0.8375	0.9930
λ=2	0.9809	0.9083	0.7040	0.9031
추가소음 차원 1개	0.9738	0.5718	0.7803	0.9122
추가소음 차원 2개	0.8321	0.6500	0.7362	0.9692

군집분석 프로그램과 군집의 수간의 상호작용에 있어서는 가장 적은 군집의 수(2개의 군집)와 가장 큰 군집의 수(5개의 군집)를 도출해낼 때 CCA의 결과가 우수하였다(표 7 참조). 군집의 수가 세 개일 경우에는 SPSS의 Clementine이, 네 개일 경우에는 IBM의 Intelligent Miner가 우수한 결과를 얻었다.

<표 7> 군집분석 프로그램과 군집의 수의 상호작용

군집의 수	군집분석 프로그램			
	CCA	Howard-Harris	Intelligent Miner	Clementine
2개	0.9756	0.8592	0.5289	0.9300
3개	0.9469	0.7938	0.9353	0.9970
4개	0.8905	0.7859	0.9797	0.9770
5개	0.9855	0.8230	0.9971	0.9279

8. 요약

본 논문에서는 군집분석 프로그램의 성능을 평가하기 위한 하나의 방법론을 제시하고, 그 방법론을 사용하여 현재 많이 사용하고 있는 네 개의 군집분석 프로그램의 성과를 평가하는데 그 주요 목적을 두고 있다. 본 논문에서는 이미 알려져 있는 군집구조를 지닌 인위적 데이터를 사용하여 반복적 K-평균 군집분석 프로그램 (Convergent Cluster Analysis: CCA), 단순 K-평균 군집분석 프로그램 (Howard-Harris 프로그램), 그리고 두 개의 상업적인 프로그램 (SPSS의 Clementine과 IBM의 Intelligent Miner)의 성과를 비교하였다. 전반적으로 반복적 군집분석 프로그램인 CCA나 SPSS의 Clementine이 다른 군집분석 프로그램에 비해 이미 알려진 군집구조를 잘 발견하는 것으로 나타났다. 그 중에서도 Random한 방법이 아닌, 좀 더 체계적인 방법으로 초기치를 설정한 CCA는 본 논문에서 평가한 네 가지의 군집분석 프로그램 중에 가장 우수한 결과를 보여 주었다. 이와 같은 결과는 K-평균 군집분석 방법의 정확성이 초기치에 매우 민감하다는 기존의 연구 결과를 보조하고 있다 (Hair et al., 1992). 그리고, 한번에 군집을 나누는 것보다 여러 번 반복하여 군집을 나누었을 때 성과가 좋음을 발견하였다. 추가적으로, 본 연구에서는 군집분석 프로그램의 종류에 관계없이 오차의 강도 (λ)가 증가함에 따라 군집 구조의 도출이 낮아지는 경향이 있다는 것을 발견하였고, 또한 군집의 크기가 증가함에 따라 ARI의 값이 증가함을 발견하였다. 그러나, ARI 값이 변화하는 추세가 K-평균 군집분석 방법과 IBM의 Intelligent Miner(비 K-평균 군집분석 방법)간에 차이가 있다는 것을 발견하였다(표 6과 표 7 참조).

본 연구에서는 외생 변수의 값에 따라 군집분석 프로그램의 성과가 어떻게 달라지고, 네 가지의 군집분석 프로그램들간의 성과에 어떤 차이가 있는지를 탐색적으로 연구하였다. 본 연구에서는 기존의 연구에서 주장하였던 내용을 실증적으로 검증하였다는 데서 그 의의를 찾아 볼 수가 있다. 많은 기존의 연구에서 군집분석 프로그램의 성과는 여러 가지 외생 변수(자료의 크기, 오차의 형태, 군집의 수 등)에 의하여 영향을 받을 것이고, 그 외생 변수를 결정하는 것은 매우 어려운 일이라고 지적하고 있다. 그러나, 그 변수들이 어떻게 군집분석 프로그램 성과에 영향을 미치는 지를 탐색한 연구는 찾아 보기가 매우 힘들다. 특히, 군집분석 프로그램 사이에는 성과가 다를 것이라고 예측은 하고 있지만 얼마나, 어떻게 다른지에 대하여 실증적으로 탐색한 연구는 극히 드물다. 본 연구에서는 군집분석 프로그램에 따라 그 성과가 달라질 것이라는 사실을 실증적으로 검증하였으나, 프로그램 간의 성과 차이가 프로그램의 어떠한 속성의 차이에서 기

인한 것인지는 파악하지 못하였다. 단지, 연구 대상 프로그램의 속성을 요약한 표 1을 근간으로 하여 초기치 설정 방법과 군집분석 횟수가 단일이냐, 반복이냐에 따라 프로그램의 성과가 다를 것이라는 사실을 추론하였을 뿐이다. Helsen & Green(1991)에서는 Random하게 초기치를 설정하는 방법을 포함한 다양한 초기치 설정 방법의 성과를 비교하였다. 본 연구에서와 같이 Random하게 초기치를 설정하는 것보다 체계적으로 초기치를 설정하는 것이 더 좋은 성과가 나타남을 실증적으로 보여 주었다. 본 연구에서는 프로그램 간의 성과에 영향을 미칠 수 있는 또 하나의 다른 속성을 발견하였다. 군집분석의 횟수 또한 초기치 설정 방법과 같이 프로그램 간의 성과에 영향을 미칠 가능성 있는 변수임을 본 연구를 통하여 보여 주었다. 차후의 연구에서는 프로그램의 어떤 속성에서 기인한 것인지를 실증적으로 검증해 볼 필요가 있을 것이다.

참고문헌

Aldenderfer, M.S. and Blashfield, R. K. (1978), "Computer Programs for Performing Iterative Partitioning Cluster Analysis," Applied Psychological Measurement, 2, 533-541.

Anderberg, M. R. (1973), Cluster analysis for researchers. New York: Academic Press.

Arabie, P., Carroll, J., DeSarbo, W.S., and Wind, Y. (1981), "Overlapping Clustering: A New Methodology for Product Positioning," Journal of Marketing Research, 18, 310-7.

Berry, M. and Linoff, G. (1997), Data Mining Techniques, Wiley Computer Publishing.

Berson, A., Smith, S. and Thearling, K. (2000), Building Data Mining Applications for CRM, McGraw-Hill.

Blashfield, R. K. (1976), "Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods," Psychological Bulletin, 83, 377-88.

Carroll, J. and Arabie, P. (1983), "INDCLUS:An Individual Differences Generation of the ADCLUS Model and the MAPCLUS Algorithm," Psychometrika, 48, pp. 157-169.

Chaturvedi, A. and Carroll, J. (1994), "An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models," Journal of Classification, 11, pp. 155-170.

Chaturvedi, A., Carroll, J., Green, P. and Rotondo J. (1997), "A Feature-Based Approach to Market Segmentation via Overlapping K-Centroids Clustering," Journal of Marketing Research, 34 (August), 370-7.

DeSarbo, W. (1982), "GENNCLUS: New Models for General Nonhierarchical Clustering Analysis," Psychometrika, Vol. 47, pp. 449-475.

Frank, R. E. and Green, P. E. (1968), "Numerical Taxonomy in Marketing Analysis: A Review Article," Journal of Marketing Research, 5, 83-98.

Green, P. E., Frank, R. E., and Robinson, P. J. (1967), "Cluster Analysis in Test Market Selection," Management Science, 13, B-387-400.

Green, P. E. and Krieger, A. (1991), "Segmenting Markets with Conjoint Analysis," Journal of Marketing, 55, 20-31.

Hair, J., Anderson, R., Tatham, R. and Black, W. (1992), Multivariate Data Analysis with Readings, 3rd Ed., Macmillan Publishing Company, NY, NY.

Helsen, K. and Green, P.E. (1991), "A Computational Study of Replicated Clustering with

an Application to Market Segmentation, "Decision Science, 22, 1124-41.

Hubert, L. and Arabie, P. (1985), "Comparing Patitions," Journal of Classification, 2, pp. 193-218.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999), "Data Clustering: A Review," ACM Computing Surveys, 31(3), pp. 264 -323.

Joycce, T. and Channon, C. (1966), "Classifying Market Survey Respondents," Applied Statistics, 15, 191-215.

McLachlan, G. J. and Basford, K. E. (1988), Mixture Models: Inferences and Applications to Clustering, New York: Marcel Dekker.

Milligan, G.W. and Sokol L.M. (1980), "A Two-Stage Clustering Algorithm with Robust Recovery Characteristics," Educational and Psychological Measurement, 40, 755-9.

Milligan, G.W. (1985), "An Algorithm for Generating Artificial Test Clusters," Psychometrica, 50, pp. 123-127.

Milligan, G.W. and Cooper, M.C. (1986), "A Study of Comparability of External Criteria for Hierarchical Cluster Analysis," Multivariate Behavioral Research, 21, 441-58.

Neidell, L.A. (1970), Procedures and Pitfalls in Cluster Analysis, Proceedings, Fall Conference, Chicago: American Marketing Association.

Sawtooth Software (1990), CCA System for Convergent Cluster Analysis, Ketchum, ID: Sawtooth Software.

Sawtooth Software (1986), ACA System for Adaptive Conjoint Analysis, Ketchum ID: Sawtooth Software.

Schaffer, C. and Green, P. (1998), "Cluster-Based Market Segmentation: Some Further Comparisons of Alternative Approaches," Journal of the Market Research Society, 40(2), pp. 155-163.

Smith, S.M. (1990), PC-MDS: Multidimensional Statistics Package, Institute of Business Management, Brigham Young University.

Srivastava, R. K., Alpert, M.I. and Shocker, A.P. (1984), "A Customer-Oriented Approach for Determining Market Structures,"Journal of Marketing Research, 48, 32-48.

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, 58, 236-44.

Wolfe, J. H. (1970), "Pattern Clustering by Multivariate Mixture Analysis," Multivariate Behavioral Research, 5, 329-50.

Abstract

A Methodology for Comparing Clustering Programs

Sung-ho Kim*
Seung-ik Baek*

Over the years, cluster analysis has become a popular tool for marketing and segmentation researchers. There are various methods for cluster analysis. Among them, K-means partitioning cluster analysis is the most popular segmentation method. However, because the cluster analysis is very sensitive to the initial configurations of the data set at hand, it becomes an important issue to select an appropriate starting configuration that is comparable with the clustering of the whole data so as to improve the reliability of the clustering results. Many programs for K-mean cluster analysis employ various methods to choose the initial seeds and compute the centroids of clusters. In this paper, we suggest a methodology to evaluate various clustering programs. Furthermore, to explore the usability of the methodology, we evaluate four clustering programs by using the methodology.

Key words: 군집분석, K-평균 군집분석, 데이터 마이닝

* College of Business Administration, Hanyang University