응용통계연구 (2009)
22(4), 723–733

자동화 K-평균 군집방법 및 R 구현

김성수[1]

[1]한국방송통신대학교 정보통계학과

(2009년 6월 접수, 2009년 6월 채택)

요 약

K-평균 군집분석이 가지는 두 가지 근본적인 어려움은 사전에 미리 군집 수를 정해야 하는 문제와 초기 군집중심에 따라 결과가 달라질 수 있는 문제이다. 본 연구에서는 이러한 문제를 해결하기 위한 자동화 K-평균 군집분석 절차를 제안하고, R을 이용하여 구현한 결과를 제공한다. 자동화 K-평균 군집분석에서 제안된 절차는 처음 단계로서 계층적 군집분석을 행한 후 이를 이용하여 군집 수와 초기 군집수를 자동으로 정하고, 다음 단계로 이 결과를 이용하여 K-평균 군집분석을 수행하는 방법을 택하였다. 처음 단계에서 이용된 계층적 군집분석 방법으로는 Ward의 군집분석을 한 후에 Mojena의 규칙을 이용하여 군집 수를 정하는 방법을 택하거나, 모형근거 군집분석방법을 수행한 후에 BIC 값을 이용하여 군집 수를 정하는 방법을 이용하였다. 제안된 자동화 K-평균 군집절차에는 대량자료의 분석에도 용이하게 이용될 수 있도록 반복된 표본추출 방법을 이용하여 군집 수 및 군집 중심을 구하는 절차를 포함하였다. 구현된 R 프로그램은 www.knou.ac.kr/ sskim/autokmeans.r에서 제공하고 있다.

주요용어: K-평균 군집분석, Ward 방법, Mojena 규칙, 모형근거 군집분석, BIC(Bayesian Information Criteria), 자동화 K-평균 군집분석.

1. 서론

K-평균 군집분석은 대량자료의 군집분석에 유용하게 이용되는 군집분석방법으로 고객분류, 행동과학, 심리 분류 등에 널리 이용되는 군집분석 방법이다. 특히 대량자료의 구조를 파악하기 위한 데이터마이닝의 방법으로도 널리 이용되고 있다. K-평균 군집방법은 단계적으로 군집을 구성하는 계층적 군집방법과는 달리 초기의 군집수를 미리 정하고 각 군집 중심과 각 관찰치와의 거리를 구하여 각 관찰치가 속하는 군집을 재배치하는 방법으로 군집을 결정하게 된다. 반면에 계층적 군집분석 방법은 각 관찰치들 사이의 유사성/거리 행렬을 구한 후에 관찰치들을 가까운 순서대로 연결해 나가는 방법이다. 이 방법은 관찰치의 수가 적은 경우에는 별 문제가 없다. 그러나 계층적 군집분석 방법은 관찰치들 사이의 유사성/거리 행렬을 구하여 군집을 결정하는 과정을 취하기 때문에 관찰치 수가 많은 경우에는 계층적 군집분석 방법은 계산량이 많고 오랜 시간이 소요되어 적절하지 않다. 이와 같이 관찰치 수가 큰 경우에 효율적으로 이용되는 방법이 비계층적 군집분석 방법인 K-평균 군집분석이다.

구체적으로 기본적인 K-평균 군집분석 절차는 다음과 같다.

1) 군집의 수 K를 정한다.

2) 임의의 K개 관찰치를 K개 각 군집에 임의로 지정한다. 이를 K개 각 군집의 중심으로 이용한다.

본 논문은 2007년 한국방송통신대학교 학술연구비 지원을 받아 작성된 것임.

[1](110-791) 서울특별시 종로구 동숭동 169, 한국방송통신대학교 정보통계학과, 교수. E-mail: sskim@knou.ac.kr

3) 모든 관찰치를 군집중심으로 부터 유클리디안 거리가 최소인 군집에 귀속시킨다.

4) 각 군집에 속한 관찰치들을 이용하여 군집중심을 새로 계산한다.

5) 군집간 관찰치이동의 변화가 없을 때까지 단계3과 단계4를 반복한다.

이와 같은 K-평균 군집분석 절차에서 가장 큰 문제점은 군집의 수 K를 어떻게 정하느냐와 각 군집의 초기중심을 어떻게 구하느냐 하는 문제이다. 실제로 군집분석은 군집의 수가 얼마인지도 모르는 상태에서 탐색적으로 데이터의 구조를 파악하여 각 관찰치를 분류하는 방법이기 때문에 군집의 수를 미리 정한다는 것은 어려운 문제라고 할 수 있다. 또한 K-평균 군집분석은 초기 중심에 따라 국소적인 최적해(local optimal solution)로 귀결될 수 있기 때문에 초기 중심에 따라 다른 군집분석 결과를 낳게 된다 (Everitt 등, 2001; Brusco와 Cradit, 2001; Chen 등, 2004). 이러한 의미에서 K-평균 군집분석을 자동화하기 위해서는 군집의 수를 정하고, 초기 중심을 구하는 일련의 과정이 자동적으로, 연속적으로 이루어져야 할 것이다. 또한 이러한 과정은 데이터의 수가 매우 많은 대량의 자료에 있어서도 효율적으로 작동될 수 있도록 구성되어야 한다.

본 소고에서는 K-평균 군집분석에서 군집의 수를 결정하고 초기 중심을 구하는 일련의 절차를 거친 후 K-평균 군집분석을 행하고 군집의 결과를 그래프로 보여주는 자동화과정을 제안하고, R을 이용하여 구현한 결과를 보이고자 한다. 먼저 2장에서는 초기군집을 정하는 방법을 고찰하고, 3장에서는 자동화 K-평균 군집분석 절차를 소개한 후 4장에서 R 구현 결과를 보이고자 한다.

2. 초기 군집 설정

최적화 과정을 이용하는 비계층적 군집방법인 K-평균 군집분석에서 군집 수를 추정하는 것은 매우 중요한 일이다. 왜냐하면 정해진 군집 수에 따라 군집 분류결과가 확연히 달라질 것이기 때문이다. K-평균 군집분석에서 군집 수를 추정하는 방법으로 일반적으로 이용되는 방법은 계층적 군집분석을 행한 후 군집 단계의 거리 등을 이용한 끝내기 규칙(stopping rule)을 이용하여 군집 수를 정하는 방법이다. 군집 수를 정하는 방법은 이러한 끝내기 규칙 뿐만 아니라 계층적 군집분석 방법으로 어느 방법을 이용하였느냐에 따라 달라진다. 계층적 군집분석에서 끝내기 규칙의 비교, 검토를 위해서는 Milligan과 Cooper (1985), Everitt 등 (2001) 등을 참조하기 바란다.

2.1. Ward 방법을 이용한 군집 수 결정

K-평균 군집분석을 행하기 전에 사전에 군집 수를 정하는 방법으로 일반적으로 이용되는 방법은 Ward (1963)의 군집방법을 이용하여 계층적 군집분석을 행하고, 각 군집 간의 거리를 이용하여 군집 수를 정하는 방법이다. 왜냐하면 K-평균 군집분석은 적절한 조건하에서 trace(W) (W는 그룹내 제곱합 행렬)을 최소화하는 과정과 동일하고 (Everitt 등, 2001), Ward의 계층적 군집방법은 그룹내 잔차제곱합인 WESS(Within-Cluster error sum of squares)를 최소화하는 과정으로 이루어지므로 (Everitt 등, 2001; Krzanowski, 1988), K-평균 군집 수를 정할 때 Ward의 방법이 효율적으로 이용된다.

구체적으로 Ward 방법의 절차는 다음과 같다.

i) 초기 거리행렬 D를 구한다. 여기서 두 개체간의 거리는 유클리디안 제곱거리로 한다. 유클리디안 제곱거리를 반으로 나눈 값이 두 개체사이의 제곱합이 된다.

ii) 거리행렬에서 가장 값이 작은 두 개체를 선택하여 새로운 군으로 한다.

iii) 기존의 그룹과 새로운 군과의 거리를 다음 식을 이용하여 구한다.

$$D_{k.ij} = \frac{(n_i + n_k)D_{ki} + (n_j + n_k)D_{kj} - n_k D_{ij}}{(n_i + n_j + n_k)}$$

여기서 $D_{k.ij}$는 새로 형성된 군과 다른 군과의 거리이고, n_i는 개체수를 의미한다.

iv) 하나의 군이 될 때까지 단계 ii)와 iii)을 반복한다. 참고로 Ward의 잔차제곱합(Error Sum of Squares)은 각 단계에서 구한 거리를 누적시킨 값이 된다.

Ward 방법을 이용하여 계층적 군집분석을 행한 후, 우리가 접하는 또 하나의 문제는 적절한 군집 수를 정하는 방법이다. 계층적 군집방법에서 군집 수를 정하는 방법 중의 하나는 각 군을 합치는 과정에서 거리의 차가 현격히 나타나는 지점을 택하여 군집의 수로 정하는 방법이다. 이러한 방법 중의 하나가 Mojena (1977)의 규칙이다. Ward의 군집분석에 대한 Mojena의 규칙은 다음과 같다.

i) 각 군집을 병합하는 단계에서 최소 유클리디안 거리 $\underline{h} = (h_1, h_2, \ldots, h_{n-1})$를 구한다. 여기서 h_j는 j번째 병합단계의 최소 유클리디안 거리를 나타내며, n은 개체의 수를 의미한다. 실제적으로 Ward의 계층적 군집분석에서는 유클리디안 제곱거리를 이용하게 되나 Mojena의 규칙을 이용하여 군집의 수를 정하는 경우에는 유클리디안 거리가 더 효율적이다 (Mojena 등, 1980).

ii) $h_{j+1} > \bar{h} + ks_h$를 만족하는 단계에서 군집 수를 정한다. 여기서 $\bar{h}$와 s_h는 군집병합단계의 거리의 평균과 표준편차이다. 여기서 상수 k의 범위는 $1 \sim 3$의 값을 취한다. Mojena 등 (1980)은 2.5 이하의 값을 권유하나, Milligan과 Cooper (1985)는 1.25의 값을 권하고 있다.

이와 같이 Ward 방법을 이용하여 군집분석을 행하고, Mojena의 규칙을 이용하여 군집 수를 정한 다음, 각 군집 중심을 구하여 K-평균 군집의 초기값으로 활용할 수 있다.

2.2. 모형근거 군집방법을 이용한 군집 수 결정

K-평균 군집 수를 정하기 위한 방법으로 유용하게 이용될 수 있는 방법은 비교적 최근에 활발히 활용되고 있는 방법으로서 모형근거 군집방법(model-based clustering methods)을 고려할 수 있다 (Banfield와 Raftery, 1993; Fraley와 Raftery, 1998). 모형근거 군집방법은 각 군집이 다변량 정규분포를 따르는 가우시안 혼합 분포(Gaussian mixed distribution)를 따른다는 가정하에서 우도함수를 최대화하는 군집을 찾는 방법이다. 군집분석을 위한 확률모형은 다음과 같다.

관찰치 $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$이 주어질 때, 군집의 수를 G개라 하고, k번째 군집에서 관찰치 $\boldsymbol{x}_i$의 분포를 $f_k(\boldsymbol{x}_i|\theta_k)$라 하자. 여기서 관찰치 $\boldsymbol{x}_i$가 k번째 군집에 속할 확률을 τ_k라 하면, 혼합우도 함수는

$$L_M(\theta_1, \theta_2, \ldots, \theta_G; \tau_1, \ldots, \tau_G|\boldsymbol{x}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\boldsymbol{x}_i|\theta_k)$$

과 같다. 여기서 일반적으로 접근하는 분포 $f_k(\boldsymbol{x}_i|\theta_k)$는 다음과 같은 다변량 정규분포를 가정한다.

$$f_k(\boldsymbol{x}_i|\mu_k, \Sigma_k) = \frac{\exp\left\{-1/2(x_i - \mu_k)'{\Sigma_k}^{-1}(x_i - \mu_k)\right\}}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}}$$

Banfield와 Raftery (1993)는 $\Sigma_k = \lambda_k D_k A_k D_k'$ 분해에 근거하여 군집 분석을 위한 방법을 제안하였다. 여기서 D_k는 분포의 방향(Orientation)을, A_k는 형태(Shape)를, λ_k는 크기(Volume)를 나타내는 특징을 지닌다. 참고로 $\Sigma_k = \lambda I$, 즉 각 군집 분포의 형태가 구형(spherical)이고, 크기와 형태가 같

은(equal) 경우에는 Ward 방법의 그룹내 잔차제곱합의 기준과 동일한 접근방법이 되며, K-평균 군집방법의 알고리즘은 $\Sigma_k = \lambda I$인 경우의 CEM(Classification Expectation-Maximization) 알고리즘과 같은 접근방법을 따르게 된다 (Fraley와 Raftery, 1998).

모형근거 군집방법은 우도함수를 이용하여 계층적 군집을 형성하게 되는데, 각 단계에서 우도함수를 최대화하는 군집쌍이 결합되게 된다. 가우시안 혼합모형에서 $\Sigma_k = \lambda_k D_k A_k D_k'$의 여러 조건에서 계층적 군집방법에 대한 알고리즘은 Fraley (1998)를 참조하면 된다. 가우시안 혼합모형에서 계층적 군집방법의 문제 중 하나는 최적 군집모형을 선택하는 것이다. 이를 위해서는 여러 조건하에서 생성되는 군집모형을 비교하기 위한 측도가 필요하게 되는데 효과적으로 이용되는 방법이 BIC(Bayesian Information Criteria)이다 (Fraley와 Raftery, 1998; Dasgupta와 Raftery, 1998; Stanford와 Raftery, 2000).

구체적으로 모형근거 군집방법의 단계는 다음과 같다.

i) 초기단계로 모형근거 군집방법을 수행한다.

ii) 군집의 공분산의 여러 조건들과 군집 수에 따라 각 개체들의 군집 소속을 결정하기 위해 EM 알고리즘을 실행한다.

iii) 여러 군집들에 대하여 BIC 값이 최대가 되는 군집을 채택한다.

이와 같이 가우시안 혼합 모형을 이용하여 초기 군집 수와 초기 군집 중심을 형성한 다음 이를 토대로 다음 K-평균 군집 절차로 들어가면 된다.

3. 자동화 K-평균 군집

K-평균 군집분석은 계층적 군집분석에 비해 컴퓨팅 속도가 빠르고 메모리의 할당이 작아 실제적으로 데이터의 수가 많은 경우에 효율적으로 활용되고 있지만, 초기 군집 수를 사전에 알고 있어야만 하는 문제가 내재하고, 또한 초기 군집중심에 따라서 군집 결정이 달라질 수 있는 문제를 안고 있다 (Everitt 등, 2001). 따라서 K-평균 군집분석을 행할 때는 군집 수를 여러 개로 바꾸어 가면서 행한 후에 군집결과를 비교하여 선택하기도 하고, 군집중심에 있어서도 사전 정보를 이용하거나, 계층적 군집분석의 결과를 이용하거나 또는 임의로 k개를 임의로 선택하여 초기중심으로 이용하기도 한다 (Everitt 등, 2001). 이러한 절차로 인해서 K-평균 군집분석이 가진 효율성에 비해서 실제로 활용면에서는 많은 어려움을 가지고 있다.

K-평균 군집분석을 좀 더 효과적으로 이용하기 위해서는 K-평균 군집분석에 초기값으로 제공되어야하는 군집 수 및 군집중심이 자동적으로 계산되고, 이를 이용하여 K-평균 군집분석이 수행되면 될 것이다. 이러한 면에서 자동화 K-평균 군집 절차가 요구된다고 할 수 있다. K-평균 군집분석을 자동화하기 위해서는 1단계에서 초기 군집 수와 초기 군집 중심이 결정된 후, 다음 단계로 K-평균 군집분석이 자동적으로 수행되도록 해야 된다. 이러한 자동화 절차에는 데이터의 수가 아무리 많아도 효율적으로 K-평균 군집분석이 수행될 수 있는 절차가 요구된다. 왜냐하면 데이터의 수가 많은 경우에 계층적 군집분석을 통해서 군집 수를 정하기 위해서는 많은 계산시간이 요구되고 비효율적이기 때문이다.

데이터의 수가 많은 경우에 직관적이면서도 쉽게 이용될 수 있는 접근 방법은 일부 표본을 랜덤 추출해서 군집분석을 행한 후, 이를 토대로 나머지 데이터에 대해서 판별분석 등을 통해서 군집을 형성하는 것이다 (Banfield와 Raftery, 1993; Brusco와 Cradit, 2001; Wehrens 등, 2004). K-평균 군집분석의 경우에는 일부 표본을 랜덤추출해서 계층적 군집분석을 통해서 군집 수 및 군집 중심을 구한 후, 이를 초기값으로 해서 전체 데이터에 대한 K-평균 군집분석 절차를 수행할 수 있다. 여기서 랜덤추출된 표본이 군집을 더 잘 나타낼 수 있도록 하기 위해서는 랜덤추출된 표본을 이용하여 군집 수 및 군집 중심을 구

하는 과정을 여러 번 반복 수행하여 가장 많이 결정되는 군집 수를 정하고, 이들의 군집중심을 이용하여 초기 군집 수 및 군집중심이 선택되도록 제안될 수 있다.

따라서 대량 데이터를 고려하여 제안된 자동화 K-평균 군집분석 절차는 다음과 같다.

i) 단순랜덤추출 또는 계통추출을 통해서 표본을 추출한다.

ii) 추출된 표본에 계층적 군집분석을 행한 후, 군집 수 및 군집중심을 구한다.

iii) 위 i), ii) 단계를 반복 수행한 후, 가장 많이 나타나는 군집 수를 초기 군집 수로 하고, 각 군집 중심의 평균을 통하여 초기 군집중심을 구한다.

iv) iii) 단계에서 결정된 군집 수 및 군집중심을 초기값으로 하여 K-평균 군집분석을 행한다.

이러한 K-평균 군집분석 절차에서 데이터의 수가 적당한 경우에는 모든 데이터를 이용하여 계층적 군집분석을 행한 후, 군집 수 및 군집중심을 구한 뒤에 K-평균 군집분석을 행하면 된다. 초기 군집수 및 군집중심에 대해서는 2장에서 소개한 방법을 이용하면 되고, 이외에도 다른 방법을 활용할 수 있다.

4. R 구현 및 실행 예

R 시스템은 자료처리, 통계분석, 그래픽 분야 등에 탁월한 기능을 가지고 있는 통계시스템이다. R 시스템은 미국 AT&T사의 Bell 연구소에서 개발된 S 언어를 기반으로 통계분석 모듈 및 다양한 그래픽 도구로 이루어져 있다. 초기 R 버전은 Auckland 대학의 Robert Gentleman과 Ross Ihaka에 의해 발전되었으며, 이후 전세계 R 개발팀에 의해 꾸준히 확장되고 발전이 이루어지고 있다. R 시스템은 S 언어를 기반으로 한 통계분석시스템인 S-PLUS와는 달리 누구나 자유롭게 쓸 수 있는 시스템으로 GNU(www.gnu.org) 규약하에 소스가 개방되어 꾸준히 발전되고 있는 시스템으로, 일련의 데이터처리 및 분석 작업을 대화형으로 처리할 수 있다.

자동화 K-평균 군집분석을 위해 개발된 R 프로그램 구현을 위해 이용된 방법과 프로그램에서 활용된 내장 R 함수는 다음과 같다.

4.1. 군집 설정을 위한 프로그램

초기 군집 설정을 위해 활용된 방법은 Ward의 계층군집분석을 행한 뒤, Mojena가 제안한 규칙 (Mojena 등, 1980)을 이용해 군집수를 정하는 방법과 모형근거 군집방법 (Banfield와 Raftery, 1993)을 행하고, BIC(Bayesian Information Criteria)를 이용하여 군집 수를 정하는 방법을 이용하였다. Ward 방법 및 Mojena의 규칙을 구현한 함수 및 모형근거에서 이용한 R 내장함수 등은 다음과 같다.

a) Ward 계층적 군집방법과 Mojena 규칙 구현

R에서 제공되는 계층적군집분석 함수로는 hclust 함수가 있다. 여기서는 Ward 군집분석을 수행하고, 각 군이 병합되는 군집간 거리를 이용한 Mojena 규칙을 이용하여 군집 수를 정하기 위하여 WardClustering 및 DecideClusterCall.ward 함수를 개발하였다.

b) 모형근거 군집분석 R 내장함수

모형근거 군집방법은 R 패키지 MCLUST (Fraley와 Raftery, 2006)를 이용한다. MCLUST는 정규혼합모형에 근거한 모형근거 군집분석을 위한 R 패키지로서 Mclust, mclustBIC 등 다양한 함수를 제공하고 있다. 자세한 내용은 www.stat.washington.edu/mclust를 참조바란다.

c) 표본추출 및 반복 수 설정

데이터의 수가 아주 많은 경우에 모든 데이터를 이용하여 계층적 군집분석을 행한 후, 초기 군집수와 군집 중심을 구하기 위해서는 컴퓨팅의 부담이 많아 비효율적이다. 따라서 데이터의 수가 많은 경우에는 일부를 랜덤추출하여 추출된 표본을 이용하여 초기 군집수와 군집 중심을 구하는 방법이 이용된다. 여기서는 추출된 표본의 편의를 줄이기 위하여 표본을 추출한 후 군집 수를 정하는 과정을 반복한 후, 빈도가 가장 많이 나타나는 군집 수를 정하고, 이들 군집 중심의 평균으로 초기 군집을 정하는 방법을 이용한다. R에서 표본추출로 이용된 패키지는 sampling이고, 단순임의추출함수로는 srswor 함수를 계통추출함수로는 UPsystematic 함수를 이용하였다.

d) K-평균 군집분석

K-평균 군집분석은 각 개체들과 군집중심간의 제곱합이 최소가 되도록 K-군집을 형성하는 방법이다. K-평균 군집분석방법의 알고리즘으로는 Hartigan과 Wong (1979)의 방법, MacQueen 방법, Lloyd 방법, Forgy 방법 등이 있다. R 시스템에서 제공되는 K-평균 군집분석 함수로는 kmeans가 있으며 디폴트로 Hartigan과 Wong (1979)의 방법을 이용한다. 프로그램에서는 초기 계층적 군집분석 단계에서 구한 군집 수와 군집 중심을 이용하여 kmeans 함수를 실행한다.

4.2. R 구현 예

K-평균 군집을 실행하기 위한 데모 예로서 클레멘타인 Telco CAT 자료 셋 churn.txt를 활용하고자 한다 (SPSS, 2000). 이 자료는 1,477명의 고객의 통화관련자료로서 군집화에 이용될 변수는 장거리통화량(LONGDIST), 국제통화량(international), 시내통화량(local) 등 3개 변수로서, 허명회와 이용구 (2004)와 마찬가지로 local에 로그변환, $\text{LOG10local} = \log_{10}(\text{local} + 1)$을 적용하였다. 허명회와 이용구 (2004)는 이 자료를 K-평균 군집분석의 재현성 평가에 활용한 바, 재현성 기준에 근거하여 군집 수를 정할 때 군집 수가 4개 또는 5개가 적절함을 보인 바 있다. 이 자료를 이용하여 R 구현 예를 살펴보도록 하자.

그림 4.1은 개발된 R 프로그램(http://www.knou.ac.kr/~ sskim/autokmeans.r)을 실행한 화면이다. 실행순서는 다음과 같다.

① 데이터 파일 입력

② 첫 번째 행이 변수이름인지를 물음

③ 변수 표준화 여부

④ 초기 군집수와 군집중심을 구하기 위한 방법 선택

⑤ 표본 추출 여부를 물음

⑥ 표본 추출시 비율 및 반복 수 설정

이와 같은 실행절차를 거친 후의 결과는 그림 4.2와 같다. 그림 4.2는 초기 군집 수를 구하는 방법으로 Ward방법을 택하고, 단순랜덤추출을 이용하여 10% 표본을 추출하여 군집 수를 구하는 과정을 10번 반복한 결과에서 Mojena의 규칙을 적용한 군집의 수가 각각 $(3, 5, 4, 4, 4, 3, 4, 4, 4, 3)$으로 나타난 것을 알 수 있다. 따라서 가장 빈번하게 나타나는 군집의 수인 4로 정해지고, 군집의 수가 4인 군집 중심들의 평균을 이용한 초기군집중심을 보여주고 있다. 이는 허명회와 이용구 (2004)가 K-평균 군집분석의 재현

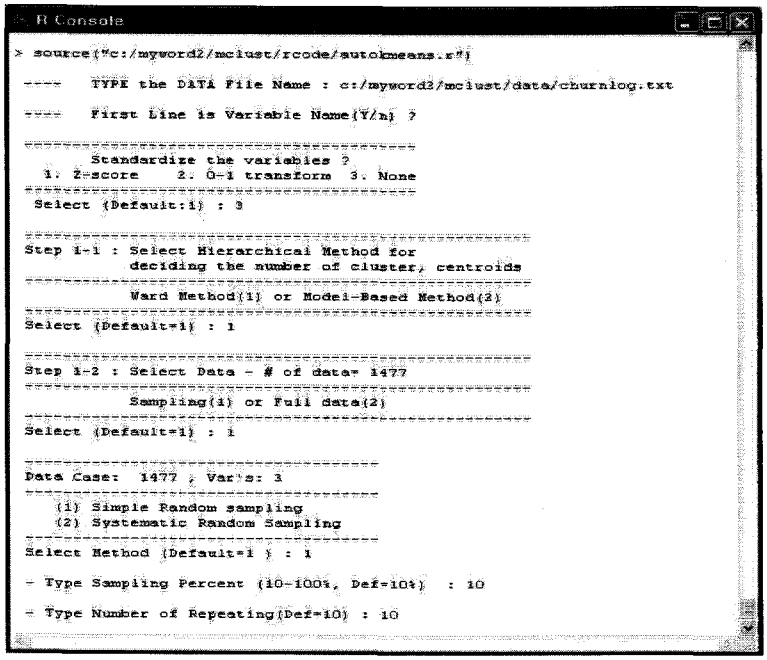

그림 **4.1.** R 실행 화면 예

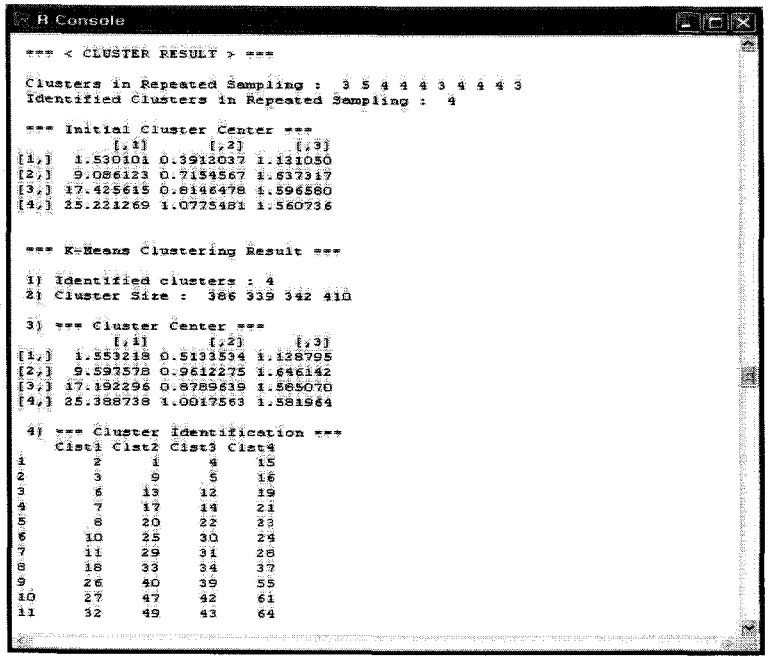

그림 **4.2.** K-평균 군집 결과

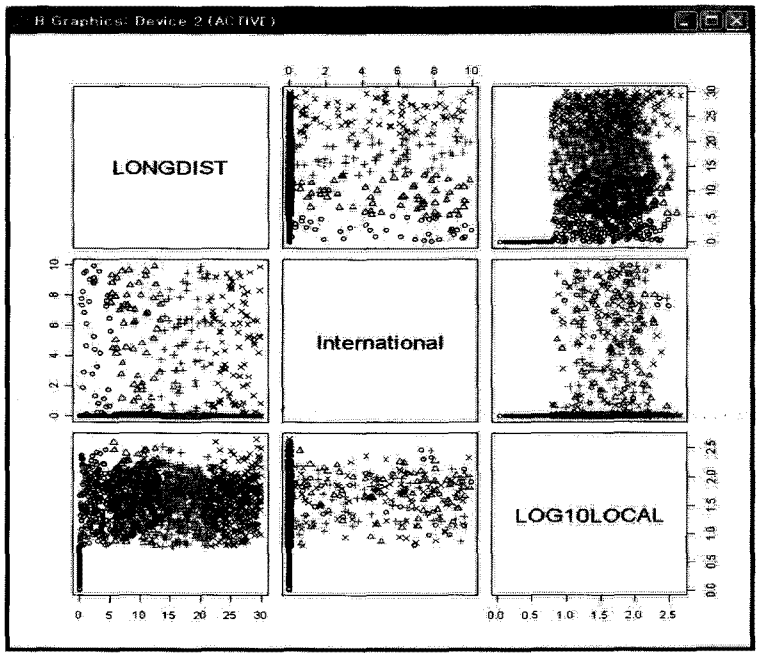

그림 **4.3.** 군집상태를 보여주는 산점도 행렬

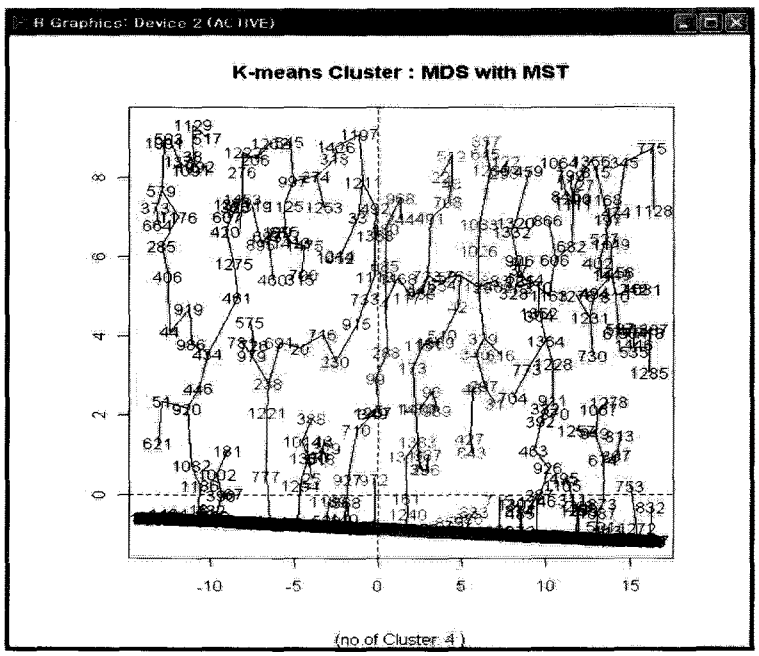

그림 **4.4.** MDS와 MST를 결합한 K-평균 군집분석 표시

성 평가지표에서 가장 높은 Rand Index (Rand, 1971)를 보이는 군집 수 4와 일치함을 알 수 있다. 그림 4.2는 또한 K-평균 군집결과와 군집중심, 각 케이스의 군집레이블을 나타내고 있다.

그림 4.3은 산점도 행렬에서 각 케이스의 군집상태를 나타낸 그림이다(실제 실행결과에서는 군집별로 색깔로도 구분함). 변수군(LOGNDIST, international), (LONGDIST, LOG10local)에서 선명히 구분되는 군집상태를 보여주고 있다. 군집이 표시된 산점도행렬은 변수선택에 대해서도 간접적으로 이용될 수 있음을 보여준다. 그림 4.4는 MDS(multidimensional scaling)를 수행한 후에 MST(minimal spanning tree)을 결합하여 군집을 표시한 결과를 보여주고 있다(실제 실행결과에서는 군집별로 색깔로 구분함). MDS와 MST를 이용한 군집결과 표시방법에 대해서는 김성수 (1999)와 Kim 등 (2000)을 참조하기 바란다.

5. 맺음 말

본 연구에서는 K-평균 군집분석을 행하는데 있어서 가장 큰 어려움인 군집 수의 결정 및 초기 군집 중심을 자동으로 구한 뒤에 K-평균 군집분석을 행하기 위한 자동화 K-평균 군집분석 절차를 제안하고, R을 이용하여 구현한 결과를 보이고 있다. K-평균 군집 수를 정하기 위한 방법으로는 Ward 군집분석을 행하고 Mojena가 제안한 끝내기 규칙을 이용하거나, 모형근거 군집분석을 행한 뒤에 BIC를 이용하여 군집 수를 정하는 방법을 구현하였다. 군집 수를 정하기 위한 방법으로는 이외에도 다른 다양한 방법을 이용할 수도 있을 것이다. 이러한 절차들은 제공된 R 소스프로그램에 첨가 구축하여 사용하길 바란다. 군집 수를 정하는 방법은 어느 방법이 절대적으로 가장 좋은 결과를 나타낸다고 할 수는 없다. 왜냐하면 데이터의 구조에 따라 우선적으로 선호되는 방법들이 있을 것이기 때문이다. 이러한 의미에서 군집 수를 정하기 위한 다양한 방법들을 R 소스프로그램에 추가하고 데이터의 구조에 따른 시뮬레이션 결과들을 보이는 것도 좋은 연구가 될 것이다.

K-평균 군집분석에서 계층적 군집분석의 결과를 이용하여 초기 군집 수를 정하기 위한 절차는 대량의 자료에 적용하기에는 많은 자원의 어려움을 안고 있다. 따라서 제안된 자동화 K-평균 군집분석 절차에는 대량 자료의 경우에도 효율적으로 이용될 수 있도록 표본추출을 통하여 초기 군집 수를 정하는 절차를 포함하고 있으며, 표본추출이 갖는 편의를 보완하기 위하여 이러한 과정을 반복 수행하여 결과를 구하는 절차를 제안하고 있다.

자동화 K-평균 군집분석을 위하여 개발된 R 프로그램은 www.knou.ac.kr/~ sskim/autokmeans.r에서 다운받아 사용하면 된다. 개발된 프로그램에서 발견되는 오류는 전적으로 저자의 책임이며, 누구나 이를 수정하고 보완하여 활용하기를 바란다. 자동화 K-평균 군집분석이 더 나아가기 위해서는 변수선택 기능이 추가되고, 특이점 검출 기능이 보완되면서 동시에 대량자료의 그래픽 표현 기능이 추가되어야 할 것이다. 이러한 주제는 각각이 모두 커다란 연구주제이며, 이러한 연구들의 결과가 개발된 R 프로그램에 추가되기를 기대한다.

참고문헌

김성수 (1999). 통계그래픽스를 이용한 K-평균 및 계층적 군집분석, <한국분류학회지>, **3**, 13–27.

허명회, 이용구 (2004). K-평균 군집화의 재현성 평가 및 응용, <응용통계연구>, **17**, 135–144.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–821.

Brusco, M. J. and Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering, *Psychometrika*, **66**, 249–270.

Chen, J. S., Ching, R. K. H. and Lin, Y. S. (2004). An extended study of the K-means algorithm for data clustering and its applications, *The Journal of the Operational Research Society*, **55**, 976–987.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering, *Journal of the American Statistical Association*, **93**, 294–302.

Everitt, B. S., Landau, S. and Leese, M. (2001). *Cluster Analysis*, Arnold, London.

Fraley, C. (1998). Algorithms for model-based gaussian hierarchical clustering, *SIAM Journal on Scientific Computing*, **20**, 270–281.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis, *The Computer Journal*, **41**, 578–588.

Fraley, C. and Raftery, A. E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, Technical Report No. 504, Department of Statistics University of Washington.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm, *Applied Statistics*, **28**, 100–108.

Kim, S. S., Kwon, S. and Cook, D. (2000). Interactive visualization of hierarchical clusters using MDS and MST, *Metrika*, **51**, 39–51.

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis*, Oxford Science, Oxford.

Milligan, G. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, **50**, 159–179.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation, *The Computer Journal*, **20**, 359–363.

Mojena, R., Wishart, D. and Andrews, G. B. (1980). Stopping rules for Wards'clustering method, COMPSTAT, 426–432.

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

SPSS (2000). Clementine Application Templates for Telecommunication Industries(Telco CAT), Chicago, SPSS Inc.

Stanford, D. C. and Raftery, A. E. (2000). Principal curve clustering with noise, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 601–609.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function, *Journal of American Statistical Association*, **58**, 236–244.

Wehrens, R., Buydens, L. M. C., Fraley, C. and Raftery, A. E. (2004). Model-based clustering for image segmentation and large data sets via sampling, *Journal of Classification*, **21**, 231–253.

Automated K-Means Clustering and R Implementation

Sung-Soo Kim[1]

[1]Department of Information Statistics, Korea National Open University

(Received June 2009; accepted June 2009)

Abstract

The crucial problems of K-means clustering are deciding the number of clusters and initial centroids of clusters. Hence, the steps of K-means clustering are generally consisted of two-stage clustering procedure. The first stage is to run hierarchical clusters to obtain the number of clusters and cluster centroids and second stage is to run nonhierarchical K-means clustering using the results of first stage. Here we provide automated K-means clustering procedure to be useful to obtain initial centroids of clusters which can also be useful for large data sets, and provide software program implemented using R.

Keywords: K-means clustering, Ward's method, Mojena's stopping rule, model-based clustering, BIC(Bayesian Information Criteria), automated K-means clustering.

This research was supported by Korea National Open University Research Fund in 2007.

[1]Prefessor, Department of Information Statistics, Korea National Open University, Seoul 110-791, Korea. E-mail:sskim@knou.ac.kr