The Korean Journal of Applied Statistics (2012)
25*(3), 471–483*

DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.471

A Variable Selection Procedure for K-Means Clustering

Sung-Soo Kim[1]

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

(Received February 23, 2012; Revised March 29, 2012; Accepted April 18, 2012)

Abstract

One of the most important problems in cluster analysis is the selection of variables that truly define cluster structure, while eliminating noisy variables that mask such structure. Brusco and Cradit (2001) present VS-KM(variable-selection heuristic for K-means clustering) procedure for selecting true variables for K-means clustering based on adjusted Rand index. This procedure starts with the fixed number of clusters in K-means and adds variables sequentially based on an adjusted Rand index. This paper presents an updated procedure combining the VS-KM with the automated K-means procedure provided by Kim (2009). This automated variable selection procedure for K-means clustering calculates the cluster number and initial cluster center whenever new variable is added and adds a variable based on adjusted Rand index. Simulation result indicates that the proposed procedure is very effective at selecting true variables and at eliminating noisy variables. Implemented program using R can be obtained on the website "http://faculty.knou.ac.kr/sskim/nvarkm.r and vnvarkm.r".

Keywords: K-means clustering, variable selection, Mojena's stopping rule, VS-KM, HINoV, adjusted Rand index.

1. 서론

군집분석은 데이터마이닝에서 말하는 자율학습(un-supervised learning)의 대표적인 방법 중의 하나로 군집을 형성하는데 있어서 경험적인 판단을 상당히 요구하고 있는 분석방법이다. 따라서 군집분석을 시작할 때 변수 선택에 대한 중요성에 대해서는 그리 신경을 쓰지 않고 진행되고 있는 것이 사실이다. 즉 군집분석에 이용되는 변수들의 선택에 대해서는 해당 분야 분석자의 경험에 전적으로 의존하고 있다고 해도 과언이 아니다. 더구나 군집분석을 하기 위해 실용적으로 이용되고 있는 SAS나 SPSS 등의 범용 통계패키지에서도 군집분석에 이용되는 변수선택에 대한 옵션이 없기 때문에 변수선택에 대한 중요성이 대부분 간과되고 있는 실정이고, 군집분석을 실행할 때는 가능한 한 얻을 수 있는 많은 변수들을 이용해서 실행되고 있다. 그러나 군집분석에 있어서는 한두 개의 관계없는 변수들이 포함되는 경우에도 내재된 군집을 밝히는데 실패할 수 있다 (Milligan, 1980a, 1980b, 1989). 이와 같이 군집 구조를 왜곡시키는 변수를 가면변수(masking variable)라고 한다 (Fowlkes와 Mallows, 1983). 가면변수를 밝히고 가면변수의 영향을 줄이는 방법으로는 변수 가중치 방법(variable weighting) 및 변수 선택(variable selection) 방법이 있다. 변수 가중치 방법은 군집 구조를 밝히는 상대적인 중요성을 고려하여 변수들

This research was supported by Korea National Open University Research Fund in 2009.
[1]Prefessor, Department of Information Statistics, Korea National Open University, Seoul 110-791, Korea.
E-mail: sskim@knou.ac.kr

의 가중치를 다르게 주는 방법이며, 변수선택 방법은 군집 분석에 유용한 변수만을 선택하는 방법이다. 변수 가중치 방법으로는 계층적 군집분석 방법에 이용된 가중치 방법 (De Soete, 1986), 비계층적 방법인 K-평균 군집방법에 적용된 SYNCLUS 방법 (DeSarbo 등, 1984) 등이 있고, 변수 선택 방법으로는 Fowlkes 등 (1987, 1988)가 다변량 분산분석의 분리 기준을 이용하여 제안한 앞으로부터의 변수선택 방법, Carmone 등 (1999)의 HINoV(heuristic identification of noisy variables) 방법, Brusco와 Cradit (2001)가 HINoV 방법을 수정하여 제안한 VS-KM(variable-selection heuristic for K-means clustering) 방법 등이 있다. 이러한 방법들은 다양한 연구들을 통해서 비교 검토되었는데 (Miligan, 1989; Gnanadesikan 등, 1995; Steinley와 Brusco, 2008), 특히 Gnanadesikan 등 (1995)은 변수 가중치 방법에 비해서 변수 선택 방법의 효과가 더 좋다는 사실을 언급하였다. 실제로 변수 선택방법은 군집분석을 위해 유의하지 않은 변수들에 대하여 데이터 수집에 대한 부담을 덜 수 있게 해주고, 또한 가중변수에 대한 의미를 해석하는 어려움을 덜어주는 장점을 가지고 있다. 변수선택방법 중에서 Carmone 등 (1999)이 제안한 HINoV 방법은 Hubert과 Arabie (1985)의 수정 Rand 지수 (Rand, 1971)를 이용하여 K-평균 군집에서 이용되는 변수 선택방법을 다루었고, Brusco와 Cradit (2001)이 제안한 VS-KM 방법은 HINoV 방법을 개선한 방법으로 K-평균 군집분석에서의 변수선택 방법을 다루고 있다. 최근에 들어서 군집분석의 변수선택방법은 모형 기반 군집분석에서 활발한 연구가 진행되고 있는데, 이를 위해서는 Raftery와 Dean (2006), Kim (2011) 등을 참조하기 바란다.

본 소고에서는 K-평균 군집분석에서 변수선택 방법을 다루고자 한다. K-평균 군집분석은 잘 알려진 바와 같이 대량자료의 군집분석에 널리 이용되는 방법으로 고객분류, 행동과학, 심리분류 등에 널리 이용된다. 특히 대량자료의 구조를 파악하기 위한 데이터마이닝의 방법으로도 널리 이용되고 있다. 그러나 K-평균 군집분석 절차에서 가장 큰 문제점은 군집의 수 K를 어떻게 정하느냐와 각 군집의 초기중심을 어떻게 구하느냐 하는 문제이다. 여기에 덧붙여 변수선택의 문제까지 포함된다면 K-평균 군집분석의 절차는 매우 어려워질 수 있을 것이다. 본 논문은 이러한 K-평균 군집분석의 절차와 변수 선택의 절차가 일련적인 과정으로 자동적으로 이루어질 수 있도록 하는 문제를 다루고 있다.

최근에 제안된 Kim (2009)의 자동화 K-평균 군집분석 절차는 군집의 수를 정하고, 초기 중심을 구하는 일련의 과정이 자동적으로, 연속적으로 이루어지도록 되어 있다. 또한 데이터의 수가 많은 대량의 자료의 경우에도 효율적으로 작동될 수 있도록 구성되어 있다. 따라서 K-평균 군집분석의 자동화절차에 변수선택 과정이 같이 연결되어 이루어진다면 탐색적인 방법으로서의 K-평균 군집분석의 활용이 매우 효과적으로 이루어질 수 있을 것이다.

본 소고에서는 2장에서 K-평균 군집분석의 변수선택 방법으로서 HINoV 방법 및 이를 개선한 VS-KM 방법을 소개하고, 3장에서 자동화 K-평균 절차에 VS-KM 방법을 연결하여 변수를 선택하는 방법을 제안하고, 4장에서 R을 이용하여 구현한 결과 및 시뮬레이션 결과를 보이고자 한다.

2. K-평균 군집 변수 선택 방법

2.1. HINoV 방법

Carmone 등 (1999)은 군집분석의 재현성에 근거하여 군집분석에 포함될 변수선택 방법으로서 HINoV(heuristic identification of noisy variables) 방법을 제안하였다. 이 방법은 각각의 변수들을 이용하여 K-평균 군집을 구성한 다음에 재현성을 측정하기 위한 측도로서 Hubert와 Arabie (1985)의 수정 Rand 지수 (Rand, 1971)를 이용하여 변수를 선택하고, 선택된 변수들을 이용하여 K-평균 군집분석을 시행하는 절차로 이루어져 있다. 이 방법은 각각의 변수에 대한 수정 Rand 지수의 합을 TOPR_j라 할 때, 이 값이 상대적으로 큰 변수들을 선택하는 절차로 구성되어 있다.

HINoV 절차는 한 번의 수행으로 군집분석에 이용될 변수들을 모두 선택하는 절차로서 구현하기가 매우 간단하고, 효율적이나 다음과 같은 제약점을 가지고 있다. 첫째로 HINoV 방법은 TOPR_j 값의 스크리 그림을 이용하기 때문에 분석자에 따라 주관적으로 흐르기 쉽고, 둘째로 참변수(true variables)간의 수정 Rand 지수는 크고 반면에 가면변수간(또는 가면변수와 참변수간)의 수정 Rand 지수는 작다는 가정을 전제로 한다. 여기서 가면변수간의 상관관계가 매우 큰 경우에는 각각의 변수들을 이용한 군집 결과는 거의 유사한 구조를 따를 것이므로 이들 변수간의 수정 Rand 지수는 클 것이고, 따라서 잘못된 변수들이 포함될 경우가 많아질 수 있는 단점을 가지고 있다.

2.2. VS-KM 방법

Brusco와 Cradit (2001)은 HINoV 방법의 단점을 보완하고, Fowlkes 등 (1988)이 제안한 앞으로부터의 선택방법을 가미하여 K-평균 군집분석을 위한 변수선택 방법인 VS-KM(variable-selection heuristic for K-means clustering) 방법을 제안하였다. 이 방법도 기본적으로 수정 Rand 지수를 이용하고 있다. 예를 들어 3개의 변수가 선택되어 있다고 하자. 이 경우에 선택된 3개의 변수를 이용한 군집과 다른 4번째 변수만을 이용하여 분류된 군집간의 수정 Rand 지수를 계산하여 이 값이 큰 경우에는 4번째 변수를 포함시키고, 반대로 이 값이 작으면 포함시키지 않는 방법이다. 또한 이 방법에서는 Fowlkes 등 (1988)이 변수선택방법에 이용했던 군집간 제곱합에 대한 총제곱합의 비율을 이용한다. 구체적인 절차는 다음과 같다.

$M =$ 개체의수, $(i = 1, 2, \ldots, M)$

$D =$ 변수 수, $(j = 1, 2, \ldots, D)$, $D = D_1 + D_2$

$D_1 =$ 참변수(true variable)의 개수

$D_2 =$ 가면변수(masking variable)의 개수

$\boldsymbol{X} = M \times D$관측값 행렬, $x_{ij} = i$개체의 j변수의 관측값

$C =$ 군집의 수, $(c = 1, 2, \ldots, C)$

$\boldsymbol{p_j} =$ 변수 j를 이용한 소속을 나타내는 $1 \times M$ 행벡터,

$p + ji =$ 개체 i가 속하는 군집

$\boldsymbol{R} = D \times D$ 수정 Rand 지수,

$(r_{jk} = r_{kj})$: 소속 군집 $\boldsymbol{p_j}$와 $\boldsymbol{p_k}$의 수정 Rand 지수,

$j = 1, \ldots, D-1$, $k = j+1, \ldots, D$

$\boldsymbol{S} =$ 선택된 변수군

$\boldsymbol{U} =$ 선택되지 않은 변수군, $\boldsymbol{S} \cup \boldsymbol{U} = \{1, 2, \ldots, D\}, \boldsymbol{S} \cap \boldsymbol{U} = \{\phi\}$

$\boldsymbol{w_{jk}} =$ 변수 (j, k)를 이용한 소속 군집을 나타내는 $(1 \times M)$ 행벡터,

$j = 1, \ldots, D-1$, $k = j+1, \ldots, D$

$w_{jki} =$ 개체 i의 소속 군집

$\boldsymbol{Q} =$ 분류된 군집 $\boldsymbol{w_{jk}}$에 대하여 총제곱합에 대한 군집간 제곱합 비율,

$q_{jk} = q_{kj}$, $j = 1, \ldots, D-1$, $k = j+1, \ldots, D$

$T =$ 처음 두 변수 쌍을 선택하기 위한 기준값

$\boldsymbol{y}$ = 선택 변수군 $j \in S$을 이용하여 분류된 $1 \times M$ 행벡터,
y_i: i번째 개체의 소속 군집
G_i = 두 분류 $\boldsymbol{p_j}$와 $\boldsymbol{y}$의 수정 Rand 지수,
$j = 1, \ldots, D$
G_{min} = 변수 선택을 위한 G_j의 최소 허용값
G_{jac} = 다음 번 변수 선택을 위해 G_j에 곱하는 요인값

단계 0: 초기화한다.
$\boldsymbol{y} = \boldsymbol{0}$; $\boldsymbol{p_j} = \boldsymbol{0}$, $j = 1, \ldots, D$
$\boldsymbol{w_{jk}} = \boldsymbol{0}$, $j = 1, \ldots, D-1$, $k = j+1, \ldots, D$,
$\boldsymbol{S} = \{\phi\}$, $\boldsymbol{U} = \{1, 2, 3, \ldots, D\}$

단계 1: 각각의 변수를 이용하여 고정된 군집 수 C에 대한 분류를 수행하고 $\boldsymbol{p_j}$ $(j = 1, \ldots, D)$를 구한다.

단계 2: 두 분류군 $\boldsymbol{p_j}$,$\boldsymbol{p_k}$ $(j = 1, \ldots, D-1, k = j+1, \ldots, D)$의 $D(D-1)/2$ 쌍에 대한 수정 Rand 지수를 계산한다.

단계 3: 변수 (j, k)를 이용하여 고정된 군집 수 C에 대한 분류를 수행하고 소속 군집 $\boldsymbol{w_{jk}}$를 구한 후, 분류된 군집 $\boldsymbol{w_{jk}}$에 대하여 총제곱합에 대한 군집간 제곱합 비율인 q_{jk}를 계산한다.
$(j = 1, \ldots, D-1, k = j+1, \ldots, D)$

단계 4: $\mathrm{Max}_{j,k}(r_{jk}) \geq T$이면 $\delta = \mathrm{Max}_{j,k}(q_{jk} \mid r_{jk} \geq T)$, 아니면 $\delta = \mathrm{Max}_{j,k}(q_{jk})$. $q_{j'k'} = \delta$인 두 변수 (j', k')를 선택하고, $\eta = r_{j',k'}$, $\boldsymbol{S} = \boldsymbol{S} \cup \{j', k'\}$, $\boldsymbol{U} = \boldsymbol{U} - \{j', k'\}$으로 한다.

단계 5: 선택된 변수들 $j \in \boldsymbol{S}$를 이용하여 고정된 군집 수 C에 대한 분류를 수행하고, 군집결과인 $\boldsymbol{y}$를 얻는다.

단계 6: 선택되지 않은 변수들에 대한 분류 결과인 $\boldsymbol{p_j}$와 기존의 선택된 변수들의 분류 결과인 $\boldsymbol{y}$와의 수정 Rand 지수 G_j를 계산한다.

단계 7: $\lambda = \mathrm{Max}_{j \in U}(G_j)$, $\lambda < G_{\min}$이거나 $\lambda < \eta \cdot G_{jac}$이면 다음 단계 8로 간다. 아니면 최대 값을 갖는 변수 j'을 택하여, $G_{j'} = \lambda$, $\eta = \lambda$로 하고, $\boldsymbol{S} = \boldsymbol{S} \in \{j'\}$, $\boldsymbol{U} = \boldsymbol{U} - \{j'\}$로 한다. $\boldsymbol{U} = \phi$이면 단계 8로 가고, 아니면 단계 5로 간다.

단계 8: 선택된 변수 $\boldsymbol{S}$를 이용하여 K-평균 군집분석을 수행한다.

VS-KM 절차는 HINoV 절차의 가장 큰 단점인 상관관계가 큰 두 가면변수의 사전 선택을 방지하기 위해 군집 효과를 나타내는 측도인 군집간 제곱합 비율을 이용하고 있다. 또한 HINoV 절차는 한 번의 수행으로 군집분석에 이용될 변수들을 모두 선택하는 반면에 VS-KM 절차는 앞으로부터의 변수선택 방법을 사용하여 추가해 나가므로써 군집분석에 유용한 모든 변수들을 선택할 가능성을 높여주고 있다. VS-KM 절차에서는 처음에 선택되는 두 변수가 매우 중요한 역할을 한다. 처음 선택되는 두 쌍의 변수군과 동일한 군집형태를 지니는 변수부터 차례로 더해 나가기 때문이다. VS-KM 절차에서는 군집 효과가 떨어지지만 상관관계가 큰 두 가면변수의 초기 선택을 방지하기 위하여 단계 3에서 가능한 모든 두 쌍에 대하여 총제곱합에 대한 군집 간 제곱합의 비율을 계산하고 있다. 그러나 기본적으로 이용되는 측도가 수정 Rand 지수이기 때문에 모든 가능한 두 쌍에 대하여 이를 계산할 필요는 없고, 수정 Rand 지수가 큰 상위 몇 쌍을 고르고, 이에 대한 제곱합의 비율을 고려하게 된다면 계산량이 많이 줄어들게 된다.

VS-KM 방법은 HINoV 절차에 비해 여러 가지 장점을 가지고 있지만, 이 방법에서도 K-평균 군집방법이 가지는 근본적인 문제점, 즉 몇 개의 군집으로 구성되어 있고, 또한 초기 군집 중심에 대한 문제가 여전히 남아 있는 것이 사실이다. 군집분석에서 적정 군집 수는 선택된 변수에 따라 달라질 수 있다. VS-KM 방법에서는 초기부터 고정된 군집수를 이용하여 K-평균 군집분석을 수행하고 변수를 더해나가는 과정을 택하기 때문에 선택된 변수에 따라 적정 군집수가 달라질 수 있는 문제를 간과하고 있다. 따라서 K-평균 군집분석에 있어서 군집 수를 고정하고 변수선택 과정을 진행하는 것보다는 변수선택 과정을 군집 수의 결정과 동시에 연계하여 진행하는 것이 바람직하다고 할 수 있다.

3. 자동화 K-평균 군집과 VS-KM 방법의 연계

K-평균 군집분석은 데이터의 수가 많은 경우에도 효율적으로 활용되고 있지만, 초기 군집 수를 사전에 정해야 하고, 또한 초기 군집중심에 따라서 군집 결정이 달라질 수 있는 문제를 안고 있다. 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 초기 중심을 임의로 선택하여 이용하기도 한다 (Everitt 등, 2001).

K-평균 군집분석이 가진 이러한 어려움을 줄이기 위해 최근에 제안된 절차로는 Kim (2009)의 자동화 K-평균 군집절차를 들 수 있다. 이 절차는 K-평균 군집분석에 초기값으로 제공되어야 하는 군집 수 및 군집중심이 자동으로 계산되고, 이를 이용하여 K-평균 군집분석이 수행되도록 구성되어 있다. Kim (2009)의 자동화 K-평균 군집절차는 대량 데이터의 경우에도 효율적으로 수행될 수 있도록 일부 표본을 랜덤추출하는 과정을 활용하고 있다. 대량 데이터를 고려한 Kim (2009)의 자동화 K-평균 군집분석 절차는 다음과 같다.

<자동화 K-평균 군집분석 절차>

i) 단순랜덤추출 또는 계통추출을 통해서 표본을 추출한다.

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

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

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

자동화 K-평균 군집분석 절차에서 데이터의 수가 적당한 경우에는 모든 데이터를 이용하여 계층적 군집분석을 행하여 군집 수 및 초기 군집중심을 구한 뒤에 이를 이용하여 K-평균 군집분석을 수행하도록 되어있다. 이러한 자동화 K-평균 군집 절차는 기본적으로 모든 변수를 활용하여 처리하도록 되어있다. 앞에서 언급한 바와 같이 가면변수가 포함되어 있는 경우에 모든 변수를 이용하여 군집분석을 행하는 경우에는 처음부터 잘못된 군집 수를 정하여 시작할 수 있다. 따라서 자동화 K-평균 군집분석 절차에 변수선택 절차가 추가되는 것이 바람직하다.

자동화 K-평균 군집분석 절차에 변수선택 절차가 가미되기 위해서는 선택된 변수에 따라 적정 군집 수가 달라지는 절차를 고려하여 군집 수를 정하고, 나머지 변수에 대해서도 여기서 정해진 군집 수를 이용하는 것이 필요하다. VS-KM 절차에서는 처음부터 고정된 군집 수로 K-평균 군집분석을 행하면서 변수를 하나씩 더해 나가는 과정을 택하고 있으나 선택된 변수에 따라 적정 군집 수가 달라지는 문제를 고려할 필요가 있다. VS-KM 절차를 살펴보면 처음 두 변수 쌍을 어떻게 선택하느냐가 매우 중요하다는 것을 알 수 있다. 처음 두 변수 쌍이 선택된 후, 여기에 다른 변수들이 추가적으로 더해지는 과정으로 이루어져 있기 때문이다. 따라서 각각의 개별 변수를 이용하여 자동화 K-평균 군집분석을 수행한 뒤, 각

두 변수별로 수정 Rand 지수 및 총제곱에 대한 군집간 제곱합 비율을 계산하여 초기 변수쌍을 선택하고 자동화 K-평균 군집 절차를 수행하면서 변수를 더해나가는 과정을 반복하면 될 것이다. 변수선택 과정을 가미한 자동화 K-평균 군집 분석 절차는 다음과 같다.

단계 0: VS-KM 절차와 같이 초기화한다.
$\boldsymbol{y} = 0; \boldsymbol{p}_j = 0,\ j = 1, \ldots, D$
$\boldsymbol{w}_{jk} = 0,\ j = 1, \ldots, D-1,\ k = j+1, \ldots, D,$
$\boldsymbol{S} = \{\phi\}, \boldsymbol{U} = \{1, 2, 3, \ldots, D\}$

단계 1: 전 개체 또는 추출된 표본을 이용하여 각각의 변수에 대하여 자동화 K-평균 군집 절차를 시행한다.

단계 2: 가능한 모든 두 변수쌍에 대하여 수정 Rand 지수를 구한다.

단계 3: 수정 Rand 지수가 큰 상위 변수군에 대하여 총제곱합에 대한 군집간 제곱합 비율을 계산한다.

단계 4: 수정 Rand 지수가 지정값 T 이상이면서 총제곱합에 대한 군집간 제곱합 비율이 큰 두 변수 쌍을 선택한다. 두 변수 (j', k')가 선택된 경우, $\boldsymbol{S} = \boldsymbol{S} \cup \{j', k'\}, \boldsymbol{U} = \boldsymbol{U} - \{j', k'\}$으로 한다.

단계 5: 선택된 변수들 $j \in \boldsymbol{S}$을 이용하여 자동화 K-평균 군집 절차를 수행하고, 군집결과인 $\boldsymbol{y}$를 얻는다.

단계 6: 선택되지 않은 변수들에 대하여 단계 5에서 얻어진 군집수로 군집분석을 실시하고 분류결과인 $\boldsymbol{p}_j$와 기존의 선택된 변수들의 분류결과인 $\boldsymbol{y}$와의 수정 Rand 지수 G_j를 계산한다.

단계 7: 단계 6에서 얻어진 수정 Rand 지수 중 최대값 $\lambda = \mathrm{Max}_{j \in \boldsymbol{U}}(G_j)$를 구한다. $\lambda < G_{\min}$이거나, $\lambda < \eta \cdot G_{jac}$이면 변수선택 과정을 끝내고 단계 8로 간다. 아니면 최대값을 갖는 변수 j'을 택하여 $\eta = \lambda$으로 하고, $\boldsymbol{S} = \boldsymbol{S} \cup \{j'\}, \boldsymbol{U} = \boldsymbol{U} - \{j'\}$로 한다. $\boldsymbol{U} = \phi$이면 단계 8로 가고, 아니면 단계 5로 간다.

단계 8: 선택된 변수 $\boldsymbol{S}$를 이용하여 자동화 K-평균 군집분석을 수행한다.

4. R 구현 및 시뮬레이션 결과

K-평균 군집분석에서 적정 군집 수를 구하기 위해 활용된 방법은 Ward (1963)의 계층군집분석을 행한 뒤, Mojena 등 (1980)이 제안한 규칙을 이용해 군집수를 정하고 각 군집 중심을 구하여 K-평균 군집의 초기값으로 활용하는 방법을 이용하였다. 여기서 데이터의 수가 대량인 경우에는 랜덤추출된 표본을 이용하여 초기 군집수와 군집 중심을 구하는 과정을 이용하였다. 군집 수를 정하는 방법은 이외에도 다양한 방법이 있으므로 다른 방법을 사용하는 것도 좋을 것이다. 이러한 예로서는 모형근거 군집방법(Banfield와 Raftery, 1993)을 행하고, BIC(Bayesian Information Criteria)를 이용하여 군집 수를 정하는 방법도 한 예라 할 수 있다.

군집방법들의 효과를 살펴보기 위해서는 가상 데이터가 많이 이용된다. 군집분석에 이용되는 가상데이터를 만드는 알고리즘으로는 Miligan (1985), Waller 등 (1999), Qui와 Joe (2006) 등의 알고리즘을 들 수 있다. 특히 Miligan (1985)은 실험계획법 관점에서 군집 데이터를 생산했는데, 예를 들어 군집 수, 변수 수, 군집 크기, 이상치, 잡음(noisy) 변수, 측정 오류 등을 고려하여 군집 데이터를 만드는 알고리즘을 제안하였다. Qui와 Joe (2006)는 군집간의 이격도(degree of separation)개념을 도입하여 Miligan의 알고리즘을 개선하였고, 특히 R 패키지 "clusterGeneration"을 제공하여 다양한 군집 데이터를 생산할 수 있게 하였다.

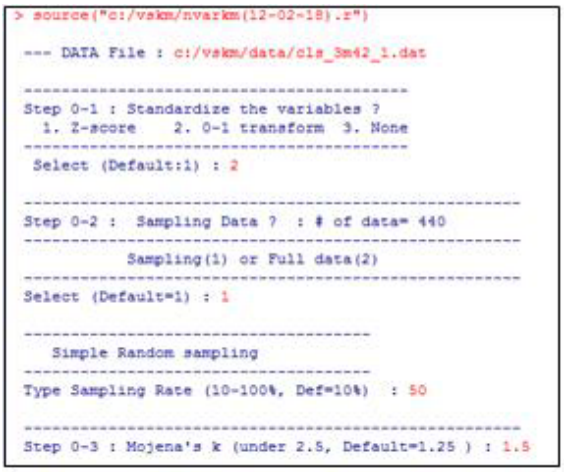

Figure 4.1. Running R program

본 소고에서는 제안된 자동화 변수선택 방법의 효과를 살펴보기 위하여 Qui와 Joe (2006)의 다음과 같은 3가지 요인을 고려하여 가상 군집 데이터 $27(3 \times 3 \times 3)$개의 데이터셋을 만들어 자동화 K-평균 군집방법에서 제안된 변수선택방법의 효과를 살펴보았다. 데이터 수는 400~1000개로 하였다.

① 군집수 $C = 3, 4, 5$.

② 군집 이격도(이격도 $= 0.01$은 두 군집이 $N(0, 1)$과 $N(0, A)$에서 $A = 4$의 분포에서 생성된 가까운 군집을 의미하며, 이격도 $= 0.21$은 $A = 6$에서 생성된 분리된 군집, 이격도 $= 0.34$는 $A = 8$에서 생성된 잘 분리된 군집을 나타낸다. 본 실험에서는 잘 분리된 군집에서 이격도 $= 0.40$을 시용하였다.) Separation $= 0.01$, 0.21, 0.40.

③ 변수 수 $D = 6(2), 8(3), 10(4)$; ( )안은 잡음(noisy) 변수 수임.

Figure 4.1은 R프로그램(http://faculty.knou.ac.kr/sskim/nvarkm.r)(대량자료의 경우에는 R 프로그램 http://faculty.knou.ac.kr/sskim/vnvarkm.r을 이용하기 바란다)을 실행한 화면이다. 실행순서는 다음과 같다.

① 데이터 파일 입력

② 변수 표준화 여부

③ 표본 추출 여부

④ 군집수를 정하기 위한 Mojena k값 입력

Figure 4.1의 실행절차에서는 표준화 절차로서 0–1 변환 절차를 택하고 있고, 케이스 선택은 단순임의 표본 추출방법으로 50%를 택하고, Mojena k값은 1.5를 입력하는 과정을 보여주고 있다. 대량자료의 경우에는 표본추출률을 10% 이하로 하기 바란다.

Figure 4.2는 변수선택절차가 가미된 자동화 K-평균 군집분석 결과를 단계별로 보여주고 있다. 먼저 단계 1에서는 각 개별 변수들을 이용한 K-평균 군집 수를 보여주고 있다. 단계 2에서는 수정 Rand 지수가 가장 큰 5개의 변수쌍을 차례대로 보여주고 있고, 단계 3에서는 단계2에서 선택된 다섯 개의 변

```
=== Variable Selection Process in Auto K-means ====

 Step 1 : Size of Cluster for each variable
[1] 5 5 6 3 5 4

 Step 2 : Highest adjusted Rand Index(Top 5)
     [,1] [,2]
[1,]    4    1
[2,]    6    4
[3,]    6    1
[4,]    4    3
[5,]    3    1

 Step3 : Ratio of SSB/SST (Top 5)
[1] 0.8023 0.8286 0.8126 0.8557 0.8107

 Step 4: First Selected Vars = ( 4 3 )
 Number of Cluster =  6
 Stpe 5-6 : Adjusted Rand Index between Y and Unselected vars ===
[1] 0.1698 0.0030 0.0000 0.0000 0.0026 0.1727

 Step 7 : Select Variables
 adj.max =  0.1727  which =  6

 Selected Vars = ( 4 3 6 )
 UnSelected Vars = ( 1 2 5 )
 Number of Cluster =  4
 Step5-6 : u.adj.max =  0.3196  which =  1

 adj.max =  0.3196 which=  1
 Selected Vars = ( 4 3 6 1 )
 UnSelected Vars = ( 2 5 )
 Number of Cluster =  3
 Step5-6 : u.adj.max =  0.007  which =  5

 adj.max =  0.007 which=  5

 Step 8 : Result of Last K-Means Using Selected Variables

 Selected Vars = ( 4 3 6 1 )
 UnSelected Vars = ( 2 5 )

 Cluster Size =  177 138 125
```

Figure 4.2. Process results of variable selection for K-means clustering

수쌍에 대한 총제곱합에 대한 군집간 제곱합 비율을 보여주고 있다. 이 결과에서 보면 변수 $(4,1)$ 쌍이 가장 큰 수정 Rand 지수값을 갖는 반면에 변수 $(4,3)$쌍이 총제곱합에 대한 군집간 제곱합 비율이 가장 크다는 것을 보여주고 있다. 수정 Rand 지수가 일정값 이상이면서 총제곱합에 대한 군집간 제곱합 비율이 가장 큰 두 변수를 기준으로 선택할 때, 처음으로 선택되는 두 변수군은 $(4,3)$이고, K-평균 군집을 한 결과 군집 수는 6개가 된다. 다음으로 변수 6의 수정 Rand 지수값이 지정된 기준값(여기서는 단계 7에서의 최소 기준값을 0.05로 함)보다 크므로 선택된 변수군에 합해진다. 이제 변수 $(4,3,6)$을 이용한 K-평균 군집을 한 결과 군집수는 4개가 되고, 나머지 변수군중에서 변수 1이 더해지는 것을 알 수 있다. 마지막으로 변수 5의 경우에는 수정 Rand 지수값이 0.007로서 기준값 0.05보다 작아지므로 이 단계에서 변수를 선택하는 절차가 끝나게 된다. 따라서 K-평균 군집변수로 선택되는 최종 변수는 $(4,3,6,1)$이고, K-평균 군집수는 3개로 변하고, 잡음변수는 $(2,5)$가 됨을 보여준다. 이 결과에서 흥미있는 사실은 선택되는 변수군에 따라 K-평균 적정 군집 수가 달라지는 것을 알 수 있다. 즉, 변수 개별로 할 때는 군집수가 각 변수별로 $(5,5,6,3,5,4)$개에서 변수 $(4,3)$이 이용되는 경우에는 6개로 결정이 되고, 변수 $(4,3,6)$에서는 군집수가 4개, 최종 선택되는 변수 $(4,3,6,1)$에서는 군집수가 3개로 결정되는 것을 알 수 있다. 따라서 K-평균 군집분석의 가장 큰 어려움인 군집수를 정해야하는 문제도 변수를

```
 === Simulation Result ===
 Original True Vars = ( 1 3 4 6 )
 Original Noisy Vars = ( 2 5 )
 Cluster Size =  126 137 177
===  Original Cluster Group ===
  [1] 3 3 2 2 3 2 2 1 1 3 1 1 3 1 1 3 2 1 1 2 2 2 2 2 1 1 3 1 2 2 3 1 2 3 1 2
 [37] 1 3 2 1 2 1 2 2 3 3 3 1 1 1 1 3 3 1 2 3 1 1 1 1 2 2 3 1 2 3 1 2 3 2 3 3
 [73] 2 1 3 3 3 1 1 3 3 3 3 1 2 1 3 2 2 2 3 2 3 3 1 3 2 3 1 1 3 3 1 3 2 2 3 1
[109] 3 1 2 2 2 3 1 1 2 3 2 3 3 2 3 2 3 3 3 3 1 3 3 2 2 2 2 2 3 2 3 3 3 3 1 1
[145] 3 1 3 3 2 1 1 1 3 3 2 3 2 2 3 3 1 3 3 3 2 3 2 3 1 1 2 1 1 2 2 2 1 1 1 3
[181] 3 1 3 1 1 1 1 3 1 3 3 2 2 1 1 3 1 3 3 2 3 3 1 3 1 2 2 2 1 3 2 2 3 3 3 3
[217] 2 2 1 2 1 3 3 2 3 3 3 3 1 1 1 2 3 3 3 3 2 3 2 1 3 3 2 2 1 3 2 1 2 1 1 3
[253] 2 3 2 2 1 3 3 1 2 2 3 1 3 3 3 3 3 2 1 3 3 2 3 2 2 3 1 3 3 3 1 3 2 2 2 1
[289] 2 2 1 2 3 3 3 1 2 1 1 1 3 2 3 3 1 2 3 3 1 3 3 3 1 1 3 3 2 2 1 1 2 3 1 3
[325] 3 2 3 3 1 3 2 2 1 1 1 3 1 2 3 3 2 2 2 1 2 1 1 3 2 2 1 2 3 1 3 2 2 2 2 1
[361] 3 3 3 2 1 1 1 3 2 2 3 1 2 2 3 3 2 1 1 3 2 1 2 2 3 3 3 3 1 3 3 3 2 3 2 3
[397] 1 2 1 2 3 3 3 1 3 3 3 1 3 3 1 2 2 3 2 2 3 3 1 3 2 2 1 2 2 1 1 2 2 2 3 2
[433] 3 1 3 3 1 3 3 2

 === Result of K-Means Using Selected Variables ===
 Selected Vars = ( 4 3 6 1 )
 UnSelected Vars = ( 2 5 )
 Cluster Size =  177 138 125
 === Clustered Group ===
  [1] 1 1 2 2 1 2 2 3 3 1 3 3 1 3 3 1 2 3 3 2 2 2 2 2 3 3 1 3 2 2 1 3 2 1 3 2
 [37] 3 1 2 3 2 3 2 2 1 1 1 3 3 2 3 1 1 3 2 1 3 3 3 3 2 2 1 3 2 1 3 2 1 2 1 1
 [73] 2 3 1 1 1 3 3 1 1 1 1 3 2 3 1 2 2 2 1 2 1 1 3 1 2 1 3 3 1 1 3 1 2 2 1 3
[109] 1 3 2 2 2 1 3 3 2 1 2 1 1 2 1 2 1 1 1 3 1 1 2 2 2 2 2 1 2 1 1 1 1 3 3
[145] 1 3 1 1 2 3 3 3 1 1 2 1 2 2 1 1 3 1 1 1 2 1 2 1 3 3 2 3 3 3 2 2 2 3 3 3 1
[181] 1 3 1 3 3 3 3 1 3 1 1 2 2 3 3 1 3 1 1 2 1 1 3 1 3 2 2 2 3 1 2 2 1 1 1 1
[217] 2 2 3 2 3 1 1 2 1 1 1 1 3 3 3 2 1 1 1 1 2 1 2 3 1 1 2 2 3 1 2 3 2 3 3 1
[253] 2 1 2 2 3 1 1 3 2 2 1 3 1 1 1 1 1 2 3 1 1 2 1 2 2 1 3 1 1 1 3 1 2 2 2 3
[289] 2 2 3 2 1 1 1 3 2 3 3 3 1 2 1 1 3 2 1 1 3 1 1 1 3 3 1 1 2 2 3 3 2 1 3 1
[325] 1 2 1 1 3 1 2 2 3 3 3 1 3 2 1 1 2 2 2 3 2 3 3 1 2 2 3 2 1 3 1 2 2 2 2 3
[361] 1 1 1 2 3 3 3 1 2 2 1 3 2 2 1 1 2 3 3 1 2 3 2 2 1 1 1 1 3 1 1 1 2 1 2 1
[397] 3 2 3 2 1 1 1 3 1 1 1 3 1 1 3 2 2 1 2 2 1 1 3 1 2 2 3 2 2 3 3 2 2 2 1 2
[433] 1 3 1 1 3 1 1 2

 Adjusted Rand Index between original group and K-mean group ===
 Adjusted Rand Index =  0.9939
```

Figure 4.3. Simulation results of clustering data

선택하는 절차와 더불어 자동으로 해결된다는 것을 알 수 있다. 물론 이러한 결과는 데이터를 랜덤으로 추출하는 과정을 거치기 때문에 수행할 때마다 결과가 달라지므로, 선택되는 변수의 순서가 중요도를 나타내지는 않는다는 사실을 유념하기 바란다.

Table 4.1은 총 27개 가상 군집데이터를 이용하여 변수 선택 절차를 거친 K-평균 군집분석 결과(Figure 4.3 참조)를 정리한 표이다. Table 4.1의 데이터셋 "cls_3s42"는 군집수 = 3, 이격도 = 0.01을 나타내고, 참변수 수는 4개, 잡음변수는 2개를 나타낸다. Figure 4.1의 실행절차와 같이 표준화 절차로서 0–1 변환 절차(이격도 = 0.01의 경우에는 이웃한 군집이 가까이 있으므로 원자료의 구조를 유지하기 위해 변환하지 않은 원자료를 선택)를 택하고, 단순임의표본 추출방법으로 50%의 케이스를 임의로 선택한 데이터를 이용하여 Ward 군집분석을 행하여 초기 군집 수와 군집 중심을 구한 후 K-평균 군집분석을 수행하였다. Ward 군집분석 방법에서 군집 수를 결정하기 위한 Mojena 규칙에서는 $k = 1.5$(이격도 = 0.01의 경우에는 $k = 2.5$를 사용)를 사용하였다. Mojena 규칙은 군집 수를 정하기 위한 값이므로 1∼2.5 범위내의 값을 사용하면 된다. 변수선택 절차에서는 처음 두 변수 쌍을 얻기 위한 기준으로 $T = 0$을 사용하여 총제곱합에 대한 군집간 제곱합 비율이 가장 큰 두 변수를 택하고, 단계 7에서 변수

Table 4.1. Results of variable selection for K-means clustering($G_{min} = 0.05$, $G_{jac} = 0.5$)

데이터셋	가상 군집 데이터				K-평균 변수선택 결과		
	군집수	이격도	참변수	잡음변수	군집수	선택된 변수	수정 Rand 지수
cls_3s42	3	0.01	(2, 3, 4, 6)	(1, 5)	3	(3, 2, 6, 4)	0.8808
cls_3m42	3	0.21	(1, 3, 4, 6)	(2, 5)	3	(4, 3, 6, 1)	0.9939
cls_3v42	3	0.40	(1, 2, 3, 6)	(4, 5)	3	(3, 2, 6, 1)	1.0000
cls_3s53	3	0.01	(1, 2, 6, 7, 8)	(3, 4, 5)	3	(6, 1, 8, 2)∗	0.9062
cls_3m53	3	0.21	(1, 2, 3, 4, 5)	(6, 7, 8)	3	(2, 1, 5, 4)∗	1.0000
cls_3v53	3	0.40	(1, 5, 6, 7, 8)	(2, 3, 4)	3	(5, 1, 6, 8, 7)	1.0000
cls_3s64	3	0.01	(1, 3, 6, 8, 9, 10)	(2, 4, 5, 7)	3	(6, 1, 8, 10, 9, 3)	0.8157
cls_3m64	3	0.21	(2, 3, 4, 5, 7, 10)	(1, 6, 8, 9)	3	(4, 2, 10, 7, 3)∗	0.9850
cls_3v64	3	0.40	(2, 3, 4, 5, 6, 10)	(1, 7, 8, 9)	3	(4, 3, 2, 5, 10)∗	1.0000
cls_4s42	4	0.01	(2, 3, 4, 5)	(1, 6)	4	(5, 2, 4, 3)	1.8568
cls_4m42	4	0.21	(2, 3, 5, 6)	(1, 4)	4	(5, 2, 6, 3)	1.0000
cls_4v42	4	0.40	(1, 2, 3, 4)	(5, 6)	4	(4, 1, 2, 3)	1.0000
cls_4s53	4	0.01	(3, 4, 5, 6, 7)	(1, 2, 8)	4	(7, 6, 3, 4, 5)	0.7859
cls_4m53	4	0.21	(1, 5, 6, 7, 8)	(2, 3, 4)	4	(5, 1, 7)∗	0.9943
cls_4v53	4	0.40	(2, 3, 5, 7, 8)	(1, 4, 6)	4	(7, 5, 3, 2, 8)	1.0000
cls_4s64	4	0.01	(3, 5, 7, 8, 9, 10)	(1, 2, 4, 6)	4	(9, 7, 10, 5, 3)	0.8597
cls_4m64	4	0.21	(1, 2, 3, 4, 5, 6)	(7, 8, 9, 10)	4	(5, 4, 2, 6, 1, 3)	0.9951
cls_4v64	4	0.40	(1, 5, 6, 7, 8, 10)	(2, 3, 4, 9)	4	(10, 6, 1, 5, 8, 7)	1.0000
cls_5s42	5	0.01	(1, 4, 5, 6)	(2, 3)	5	(4, 1, 6, 5)	0.8478
cls_5m42	5	0.21	(1, 3, 4, 5)	(2, 6)	5	(5, 4, 1, 3)	0.9585
cls_5v42	5	0.40	(2, 3, 4, 6)	(1, 5)	5	(6, 3, 2, 4)	1.0000
cls_5s53	5	0.01	(1, 3, 4, 5, 6)	(2, 7, 8)	5	(5, 4, 6, 1)∗	0.8628
cls_5m53	5	0.21	(1, 3, 5, 7, 8)	(2, 4, 6)	5	(7, 3, 8, 5, 1)	0.9895
cls_5v53	5	0.40	(2, 5, 6, 7, 8)	(1, 3, 4)	5	(8, 2, 5, 7, 6)	1.0000
cls_5s64	5	0.01	(1, 3, 5, 6, 8, 9)	(2, 4, 7, 10)	5	(6, 1, 3, 8)∗	0.8522
cls_5m64	5	0.21	(1, 2, 4, 5, 6, 10)	(3, 7, 8, 9)	5	(10, 6, 4, 2, 5, 1)	0.9927
cls_5v64	5	0.40	(1, 2, 4, 6, 7, 9)	(3, 5, 8, 10)	5	(4, 1, 9, 6, 2, 7)	1.0000

주) ∗: 참변수 일부가 포함되어 있지 않은 경우

를 추가적으로 포함시키기 위한 기준으로는 $G_{\min} = 0.05$, $G_{jac} = 0.5$를 사용하였다.

두 번째 데이터셋 cls_3m42를 예로 들면, 군집 수가 3이고, 군집간 이격도 = 0.21, 참변수는 4개로서 변수$(1, 3, 4, 6)$이고, 잡음변수는 2개로서 변수$(2, 5)$로 생성된 가상 군집 데이터인데, K-평균 변수선택 결과에서 보면 선택된 변수는 4개로 $(4, 3, 6, 1)$은 처음에 선택된 두 변수는 $(4, 3)$이고 변수 6과 변수 1이 순서대로 선택된 것을 나타낸다. 여기서 K-평균 군집 수 결과는 3개이며, 원 데이터와의 수정 Rand 지수는 0.9939로 가상데이터와 변수선택절차를 거친 K-평균 군집분석과의 결과가 거의 일치하는 것을 보여준다.

Table 4.1의 결과에서 군집간 이격도가 0.01인 경우는 가상으로 생성된 군집데이터의 구조를 유지하기 위해서 자료를 변환하지 않고 원시 자료를 이용하여 분석하였다. 이격도가 0.21, 0.40인 경우에는 0–1 변환한 자료를 이용하여 분석하였다. 이 결과에서 보면 전체적으로 참변수가 잘 선택된 것을 알 수 있다. 일부 데이터셋에서 참변수의 일부가 포함되지 않고 있는 경우를 알 수 있는데(∗ 표시), 군집분석 결과의 일치성을 나타내는 수정 Rand 지수를 보면 모두 값이 크다는 것을 알 수 있어 선택된 변수를 이용한 K-평균 군집분석을 한 결과가 좋음을 확인할 수 있다. 특히 군집간 이격도가 0.21, 0.40인 경우에는

수정 Rand 지수값이 모두 0.95 이상이어서 선택된 변수만을 이용하여 군집분석을 한 결과도 매우 뛰어남을 알 수 있다. 또한 가상 데이터를 이용한 변수 선택 결과에서 잡음변수는 어느 데이터셋에서도 포함되어 있지 않다는 것을 알 수 있다. 이는 선택된 변수군에 잡음변수가 포함되는 Type II 오류가 없다는 것을 보여주고 있어 변수선택 절차가 가미된 자동화 K-평균 군집분석이 매우 효용적임을 알 수 있다. 변수를 선택하기 위한 기준으로 이용되는 단계 7에서의 기준값 G_{min}, G_{jac}는 경험적으로 선택하면 된다 (Brusco와 Cradit, 2001). 실제적으로 기준을 강화하여 $G_{min} = 0.03$, $G_{jac} = 0.3$을 사용하는 경우에도 변수선택 결과는 Table 4.1과 거의 유사한 것을 알 수 있다.

Table 4.1의 변수선택결과에서 보면 참변수의 일부가 선택되지 못하는 경우에도 군집분석 재현성 결과가 좋음을 알 수 있고, 반면에 잡음변수는 어느 데이터셋에도 포함되는 경우는 하나도 없음을 알 수 있다. 이러한 사실은 Miligan (1985) 및 Brusco와 Cradit (2001)이 언급한 것처럼 참변수의 일부가 포함되지 못하는 Type I 오류인 경우에는 군집분석 결과의 재현성에 문제가 별로 없는 반면에, 잡음변수가 포함되는 Type II 오류인 경우 심각한 군집분석 결과의 오류가 발생할 수 있다는 사실을 감안할 때도 제안된 변수 선택 절차가 매우 효율적임을 알 수 있다.

본 실험 결과는 데이터 수를 1000 이하로 한 경우이나 데이터의 수가 큰 대량자료인 경우에는 추출된 표본만을 이용하여 군집분석을 행하고, 수정 Rand 지수를 구하여 변수를 선택하는 과정으로 압축할 수 있다. 대량자료 변수선택프로그램 http://faculty.knou.ac.kr/sskim/vnvarkm.r(R에서 메모리 크기를 알려면 memory.size, memory.limit 명령을 참조하기 바란다)에서 추출률을 5% 이하로 하여도 변수선택 효과는 매우 뛰어남을 알 수 있다.

5. 맺음말

K-평균 군집분석을 위한 변수 선택 방법으로서 Brusco와 Cradit (2001)이 제안한 VS-KM 방법은 미리 고정된 군집 수를 정하고 변수를 선택하는 과정을 거치게 된다. 잘 알려진 바와 같이 K-평균 군집분석을 행하는데 있어서 가장 큰 어려움은 군집 수를 결정 하는 문제와 초기 군집 중심을 결정하는 문제라고 할 수 있다. Figure 4.2의 분석 절차에서 나타내는 바와 같이 선택된 변수에 따라 적정 군집 수가 달라질 수 있는 문제를 안고 있기 때문에 사전에 군집수를 고정하고 변수선택을 실행하는 VS-KM 방법은 효율성이 떨어진다고 할 수 있다.

Brusco와 Cradit (2001)는 K-평균 군집분석을 위한 초기 군집수를 구하기 위하여 먼저 일부 추출된 표본을 이용하여 Ward의 계층적 군집 분석을 행하고 이를 이용하여 고정된 군집수를 정하고 변수 선택 절차를 수행하였다. 본 연구에서는 이러한 방법을 개선하여 변수선택 절차를 Kim (2009)이 제안한 자동화 K-평균 군집분석 절차와 연계한 방법을 제안하고 있다.

자동화 K-평균 군집분석 절차를 위하여 본 소고에서 사용한 방법은 대량 자료의 경우에도 효율적으로 이용될 수 있도록 표본추출을 통하여 얻은 데이터를 이용하여 Ward 군집분석을 행하고 Mojena의 규칙을 이용하는 방법을 사용하였으나, 군집 수를 정하는 방법으로는 이외에도 모형근거 군집분석을 행한 뒤에 BIC(Bayesian Information Criteria)를 이용하여 군집 수를 정하는 방법 (Banfield와 Raftery, 1993; Fraley와 Raftery, 1998)이나 다른 다양한 방법을 이용할 수도 있을 것이다. 또는 그래프를 이용하는 방법으로서 Kim 등 (2000)이 활용한 방법도 응용할 수 있을 것이다. 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 비교하고, 데이터의 구조에 따른 시뮬레이션 결과들을 보이는 것도 좋은 연구가 될 것이다. 또한 대량자료의 경우에 다양한 군집구조를 가지는 데이터를 생성하여 추출률 등을 변수로 한 시뮬레이션 결과를 보이

는 것도 좋은 응용 연구가 될 것이다.

본 연구에서는 K-평균 군집분석에서 변수를 선택하는 알고리즘을 구현하는데 있어서 특이값(outlier) 검출에 대한 문제는 고려하지 않았다. 실제로 K-평균 군집분석을 행하는데 있어서 특이값 검출에 대한 연구 자체도 완성되지 않은 과제이기 때문에 이러한 방법들을 연구하고, 변수선택과 특이값 검출을 동시에 연계한 K-평균 군집분석을 실행하는 방법도 도전해 볼만한 좋은 과제라고 할 수 있을 것이다. 이러한 절차들은 본 연구자가 제공하는 R 프로그램(http://faculty.knou.ac.kr/sskim/nvarkm.r; 대량자료의 경우 vnvarkm.r)에 첨가 구축하여 사용하길 바란다. 변수선택과 연계한 자동화 K-평균 군집분석 프로그램에서 발견되는 오류는 전적으로 저자의 책임이며, 누구나 이를 수정하고 보완하여 활용하기를 바란다. 또한 이 연구를 계기로 더 나은 변수선택 방법이 연구, 보완되고 특이점 검출 기능이 추가된 자동화 K-평균 군집 방법 및 R 프로그램이 개발되기를 기대해본다.

References

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.

Carmone, F. J., Kara, A. and Maxwell, S. (1999). HINoV; A new model to improve market segmentation by identifying noisy variables, *Journal of Marketing Research*, **36**, 501–509.

De Sarbo, W. S., Carroll, J. D., Clark, L. A. and Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with different weighting of variables, *Psychometrika*, **49**, 57–78.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering, *Quality and Quantity*, **20**, 169–180.

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

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1987). Variable selection in clustering other contexts, In C.L. Mallows(Ed.), *Design, Data and Analysis*, 13–34.

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1988). Variable selection in clustering, *Journal of Classification*, **5**, 205–228.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings (with comments and rejoinder), *Journal of the American Statistical Association*, **78**, 553–584.

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

Gnanadesikan, R., Kettenring, J. R. and Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis, *Journal of Classification*, **7**, 271–285.

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

Kim, S. (1999). Interactive visualization of K-means and Hierarchical clusters, *The Journal of Data Science and Classification*, **3**, 13–27.

Kim, S. (2009). Automated K-means clustering and R implementation, *The Korean Journal of Applied Statistics*, **22**, 723–733.

Kim, S.-G. (2011). Variable selection in normal mixture model based clustering under heteroscedasticity, *The Korean Journal of Applied Statistics*, **24**, 1213–1224.

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

Milligan, G. W. (1980a). An examination of six types of the effects of error perturbation on fifteen clustering algorithms, *Psychometrika*, **45**, 325–342.

Milligan, G. W. (1980b). An algorithm for generating artificial test clusters, *Psychometrika*, **50**, 123–127.

Milligan, G. W. (1989). A validation study of a variable-weighting algorithm for cluster analysis, *Journal of Classification*, **6**, 53–71.

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**, 259–363.

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

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering, *Journal of the American Statistical Assocation*, **101**, 168–178.

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

Qui, W.-L. and Joe, H. (2006). Generation of random clusters with specified degree of separation, *Journal of Classification*, **23**, 315–334.

Steinley, D. and Brusco, M. J. (2008). A new variable weighting and selection procedure for K-means cluster analysis, *Multivariate Behavioral Research*, **43**, 77–108.

Waller, N. G., Underhill, J. M. and Kaiser, H. (1999). A method for generating simulated plasmodes and artificial test clusters with user-defined shape, size, and orientation, *Multivariate Behavioral Research*, **34**, 123–142.

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