응용통계연구 제 13권 2호, 2000년, pp. 343-352

이중 K-평균 군집화

허명회[1)]

요 약

K-평균 군집화(K-means clustering)는 비계층적 군집화 방법의 하나로서 큰 자료에서 개체 군집화에 효율적인 것으로 알려져 있다. 그러나 종종 비교적 균일한 대군집의 일부를 소군집에 떼어주는 오류를 범하기도 한다. 이 연구에서는 그러한 현상을 정확히 인지하고 이에 대한 대책으로서 '이중 K-평균 군집화(double K-means clustering)' 방법을 제시한다. 또한 실증적 사례에 새 방법론을 적용해보고 토의한다.

주요용어: K-평균 군집화, 최적배속규칙, 혼합다변량정규분포.

1. 서론

MacQueen (1967)에서 시작된 K-평균 군집화는 비계층적 군집화 방법으로서 알고리즘은 대략 다음과 같다 (Hartigan 1975, p.102; Johnson and Wichern 1992, p.597; Sharma 1996, pp.202-207).

- 단계 0: k개의 각 군집에 1개씩의 개체를 심는다 (initial seeding). 또는 군집중심(cluster centroid)을 임시로 정한다.
- 단계 1: 모든 개체를 각각 가장 가까운 군집 중심을 찾아 배속시킨다.
- 단계 2: 각 군집의 중심을 산출한다.
- 단계 3: 단계 1과 단계 2를 변화가 거의 없을 때까지 반복한다.

K-평균 군집화는 알고리즘이 간단하여 특히 큰 자료의 개체 군집화에 효율적인 것으로 알려져 있다. 그러나 군집들이 대체로 비슷한 크기로 형성되는 현상이 자주 발견된다. 그림 1.1과 그림 1.2가 한 예이다 (Jin 1999, pp.16-17; SAS Institute 1990, pp.70-80 참조). 그림 1.1에서 그룹 1은 이변량 정규분포 $N_2(0,0,1,1,0)$으로부터 생성된 40개의 개체로 구성되었고 그룹 2는 이변량 정규분포 $N_2(4,4,1,1,0)$으로부터 생성된 4개의 개체로 구성되었다. 시각적으로 두 그룹은 명확히 구별된다. 그러나 K-평균 군집화 (첫 군집중심: (0,0)과 (3,3))를 시행한 결과는 실망스럽게도 두 그룹을 정확히 구분해내지 못하였다. 그림 1.2를 보라 (군집 수 k는 2로 지정됨). 그룹 2에 가까운 그룹 1의 4개 개체가 잘못 분류되어 있음을 볼 수 있다 (그러나 이 연구에서 제안하는 이중 K-평균 군집화로는 이런 오류들이 모두 옳게 잡혀진다).

1) (136-701) 서울시 성북구 안암동 5가 1, 고려대학교 정경대학 통계학과, 교수
E-mail: stat420@mail.korea.ac.kr

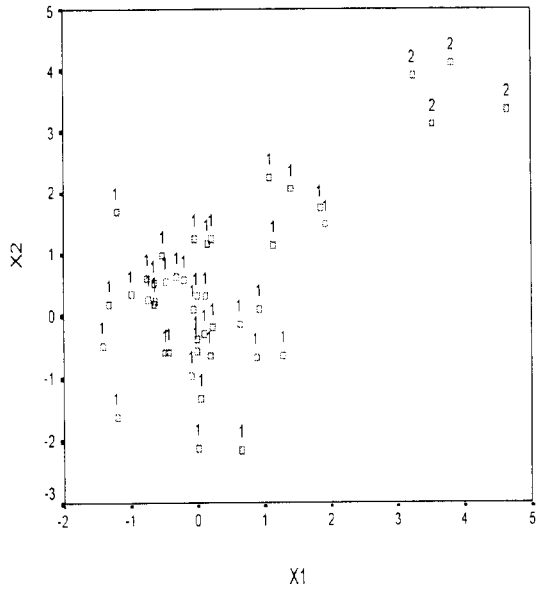

그림 1.1: 이변량 임의자료 : 그룹 1은 $N_2(0,0,1,1,0)$으로부터, 그룹 2는 $N_2(4,4,1,1,0)$으로부터 생성되었다.

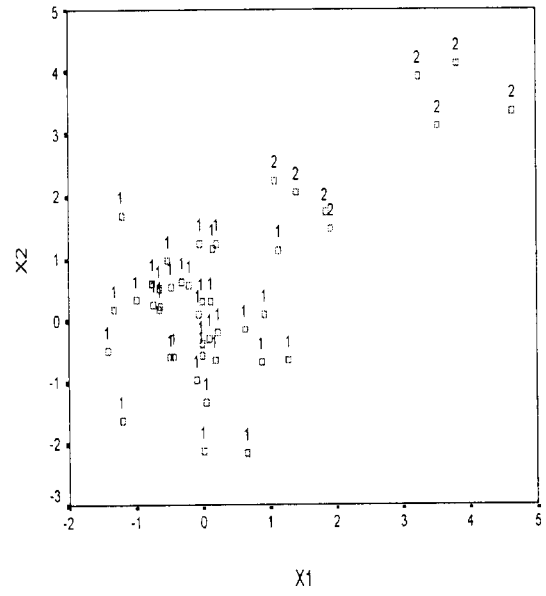

그림 1.2: 이변량 임의자료에 대한 K-평균 군집화 결과 : 그룹 2에 가까운 그룹 1의 4개 개체가 잘못 분류되어 있다 ($k = 2$).

이와 같은 현상이 드물고 우연한 것인가? 그렇지 않음을 2절에서 아주 단순한 실험을 통하여 보인다. 그렇다면 그 원인과 대책은 무엇인가? 3절에서 이에 대한 설명과 함께 이중 K-평균 군집화라는 K-평균 군집화의 변형 알고리즘을 제시한다. 4절에서는 Fisher의 붓꽃 자료에 새 알고리즘을 적용한 K-평균 군집화 결과를 옛 알고리즘에 의한 K-평균 군집화 결과와 비교하여 본다. 마지막으로 5절에서는 이중 K-평균 군집화의 위상을 전체 군집화 방법들 가운데서 살펴본다.

서론을 마치기 전에 이 연구의 범위를 제한하기 위하여 다음과 같은 몇 조건을 달기로 한다. 첫째, k개의 군집에 대한 첫 씨(initial seeds)는 임의로 정하는 것으로 한다. 따라서 자료 세트에서 관측들이 임의 순서로 배열된 경우에는 처음 k개의 관측이 첫 씨가 된다. 둘째, 특별한 언급이 없는 한, 자료 세트의 모두 변수에 대하여 평균 0, 표준편차 1로 표준화 변환을 한 뒤 군집화에 들어가는 것으로 한다. 물론, 모든 변수들의 측정 단위가 같으며 산포가 비슷한 경우에는 표준화 변환이 꼭 필요하지 않을 수 있다. 그러나, 이 연구에서는 그렇지 않은 경우를 일반적 상황으로 규정할 것이다.

2. 간단한 수치적 실험

다음과 같은 일변량 자료를 2개 군집으로 나누는 실험을 하여보았다: -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, r. 즉 자료크기는 $n = 20$이며, r의 값으로 10에서 45까지의 정수를 하나씩 고려하였다.

표 2.1이 K-평균 군집화를 적용하여 본 결과를 정리한 것이다 (첫 씨가 {-9, 9}인 경우). 예컨대 $r = 24$인 경우, 상식적으로는

군집 1*: -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
군집 2*: r(=24)

이 되어야 할 것 같지만, K-평균 군집화는

군집 1: -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1
군집 2: 2, 3, 4, 5, 6, 7, 8, 9, r(=24)

을 만들어낸다. r은 40 이상이 되어야 비로소 혼자만으로 독립 군집을 이룬다. 이것은 K-평균 군집화가 군집들을 대략 비슷한 크기로 만들려고 무리하는 경향이 있음을 보여준다.

이에 반하여 다음 절에서 제안하는 이중 K-평균 군집화에는 이런 경향이 훨씬 덜하다. 표 2.2를 보라 (첫 씨가 {-9, 9}인 경우). 이중 K-평균 군집화는 $r \geq 22$이 되면 바로 이 한 점을 크기 1의 군집으로 독립시킨다.

만약 첫 씨를 {-9, r}로 하더라도 K-평균 군집화에서는 이런 현상은 지속되어 r이 혼자만으로 군집이 되기 위해서는 $r \geq 27$이여야 한다. 그러나 이중 K-평균 군집화의 경우는 $r \geq 20$이면 된다.

표 2.1: 일변량 자료의 K-평균 군집화 ($k = 2$) : 칸 내 1과 2는 군집 기호.

자료값	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	r
$10 \le r \le 19$	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
$20 \le r \le 27$	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
$28 \le r \le 33$	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
$34 \le r \le 37$	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
$38 \le r \le 39$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
$r \ge 40$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

표 2.2: 일변량 자료의 이중 K-평균 군집화 ($k = 2$) : 칸 내 1과 2는 군집 기호.

자료값	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	r
$10 \le r \le 19$	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
$20 \le r \le 21$	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
$r \ge 22$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2

3. 원인과 대책

왜 이와 같은 현상이 발생하는가? K-평균 군집화 알고리즘의 단계 1을 집중적으로 보자. 분류대상이 된 개체 i를

$$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^t, \quad i = 1, \cdots, N$$

이라고 하자. 그리고 전 단계에서 형성된 군집중심을

$$\boldsymbol{c}_j = (c_{j1}, \cdots, c_{jp})^t, \quad j = 1, \cdots, k$$

라고 하자. 이 때 각 개체마다 개체로부터 군집중심으로까지의 제곱 유클리드 거리가 가장 작은 군집을 찾게된다. 즉 개체 i가 군집 j^*에 배속될 조건은

$$(\boldsymbol{x}_i - \boldsymbol{c}_{j^*})^t (\boldsymbol{x}_i - \boldsymbol{c}_{j^*}) = \min_{j=1,\cdots,k} (\boldsymbol{x}_i - \boldsymbol{c}_j)^t (\boldsymbol{x}_i - \boldsymbol{c}_j) \tag{3.1}$$

에 의한다. 그러므로 두 군집의 중간쯤에 낀 개체들은 대략 반반씩 나뉘어지게 된다. 이러한 분류규칙은 직관적으로 타당해 보인다. 그러나 사실은 어떤 가정에 의해 강하게 지배되고 있다. 그 이유는 다음과 같다.

군집 j가 다변량 정규분포 $N_p(\boldsymbol{\mu}_j, \Sigma)$에 의하여 형성된다고 하자 ($j = 1, \cdots, k$). 즉, 다변량 정규성과 공통 공분산행렬을 가정하기로 한다. 그리고 군집 $1, \cdots, k$가 각각 확률 $\pi_1, \cdots, \pi_k$ ($\sum_{j=1}^{p} \pi_j = 1$)에 의하여 혼합된다고 하자. 그러면 개체 $\boldsymbol{x}_i$를 군집 j^*에 배속시키기 위한 최적규칙은

$$\begin{aligned} &\pi_{j^*} \cdot \exp\{-(\boldsymbol{x}_i - \boldsymbol{\mu}_{j^*})^t \, \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_{j^*})/2\} \\ &\quad = \max_{j=1,\cdots,k} \pi_j \cdot \exp\{-(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^t \, \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)/2\} \end{aligned} \tag{3.2}$$

이다 (Mardia, Kent and Bibby 1979, p.366). 따라서, $\boldsymbol{\mu}_j$가 $\boldsymbol{c}_j$에 의하여 추정되고 공통 공분산성의 가정이 성립하여 Σ가

$$S = \sum_{i=1}^{N}\sum_{j=1}^{k} d_{ij}\,(\boldsymbol{x}_i - \boldsymbol{c}_j)(\boldsymbol{x}_i - \boldsymbol{c}_j)^t/(N-k)$$

에 의하여 잘 추정된다면 (여기서 N은 전체 자료의 크기이고 d_{ij}는 개체 $\boldsymbol{x}_i$가 군집 j에 속하는 경우엔 1이고 그 외의 경우엔 0으로 정의) 그리고 π_j가 군집 j에 배속된 개체 수 n_j에 비례한다면 (즉 $\pi_j \simeq n_j/N,\ j=1,\cdots,k$), (3.2)는

$$\begin{aligned}&\tfrac{n_{j^*}}{N}\cdot\exp\{-(\boldsymbol{x}_i-\boldsymbol{c}_{j^*})^t\,S^{-1}\,(\boldsymbol{x}_i-\boldsymbol{c}_{j^*})/2\}\\ &\quad=\max_{j=1,\cdots,k}\tfrac{n_j}{N}\cdot\exp\{-(\boldsymbol{x}_i-\boldsymbol{c}_j)^t\,S^{-1}\,(\boldsymbol{x}_i-\boldsymbol{c}_j)/2\}\end{aligned}\qquad(3.3)$$

로 대치 가능하다. 그러므로 K-평균 군집화에서의 배속규칙(allocation rule) (3.1)은 대략

- $n_1 = \cdots = n_k$이고,
- $S \propto I_p$인 경우

에서만 최적규칙에 근사함을 알 수 있다. 따라서 규칙 (3.3)을 채택한 다음의 알고리즘을 이중 K-평균 군집화 방법으로 제안한다.

- 단계 I: 통상적인 K-평균 군집화를 시행하여 $\pi_1,\cdots,\pi_k$와 $\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k$ 및 Σ에 대한 초기 추정값 $n_1/N,\cdots,n_k/N$과 $\boldsymbol{c}_1,\cdots,\boldsymbol{c}_k$ 및 S를 얻는다.
- 단계 II-1: 규칙 (3.3)에 의하여 각 개체를 k개 군집에 재배속시킨다.
- 단계 II-2: 군집중심 $\boldsymbol{c}_1,\cdots,\boldsymbol{c}_k$, 공분산행렬 S, 군집의 크기비율 $n_1/N,\cdots,n_k/N$을 새로 계산한다.
- 단계 II-3: 단계 II-1과 II-2를 변화가 거의 없을 때까지 반복한다.

따라서 이중 K-평균 군집화는 보통 K-평균 군집화 (단계 I)와 변형 K-평균 군집화 (단계 II)를 수행한다. 이 때문에 이중 K-평균 군집화(double K-means clustering)라고 명명한 것이다.

이중 K-평균 군집화는 K-평균 군집화를 두 번 정도 시행하는 정도의 계산을 필요로 하므로 데이터 마이닝(data mining)에서와 같이 대규모 자료의 개체 군집화에 적용되는 경우에도 그다지 큰 계산적인 부담을 주지 않을 것이다.

4. 사례 : Fisher의 붓꽃 자료

Fisher의 붓꽃 자료(iris data)에는 3개 품종(1:setosa, 2:versicolor, 3:verginica), 각 품종당 50개 개체의 4개 변수(x1:sepal length, x2:sepal width, x3:petal length, x4:petal width)가

있다. 붓꽃자료는 흔히 판별분석에 예제자료로 활용되지만 군집분석에서도 예제자료로서 활용가치가 있다. 전체 $N = 150$개의 개체를 $k = 3$개의 군집으로 나눌 때 원품종 분류와 어느 정도 일치하게 되는가를 봄으로써 군집화 방법이 제대로 작동하는지를 가늠하여 볼 수 있기 때문이다. 그림 4.1을 보라. 일반적으로 K-평균 군집화에서는 단순히 제곱 유클리드 거리에 근거하므로 분석에 앞서 모든 군집분석 변수들을 표준화한다. 여기서도 그렇게 하여 K-평균 군집화를 적용해본 결과가 표 4.1이다. 품종 1은 모두 단일 군집에 의하여 한 군데 묶였으나 품종 2는 11개가 주 군집에서 이탈하였고 품종 3은 14개가 이탈함으로써, 총 150개 개체 중에서 25개가 잘못 묶여졌다. 표 4.2는 이에 덧붙여 이중 K-평균 군집화를 덧붙여 적용한 결과이다. 오분류(misclassification) 개체 수가 25개에서 6개로 줄어들었음을 보라.

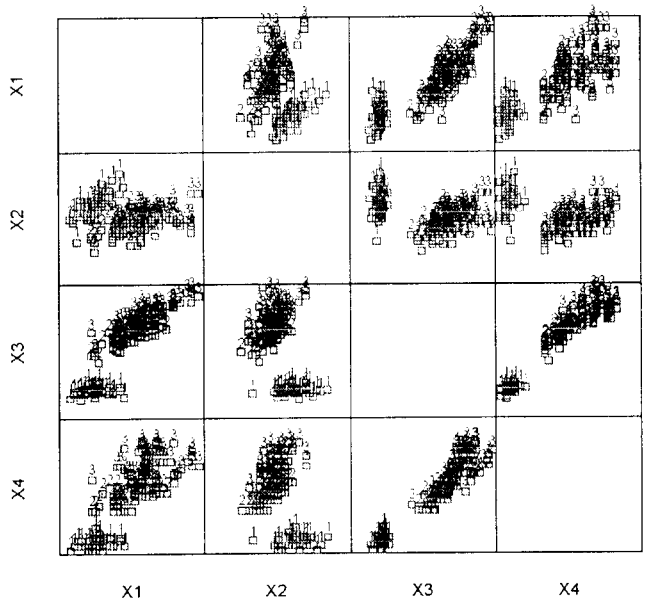

그림 4.1: 붓꽃 자료의 산점도 행렬 : 숫자는 품종 표시

표 4.1: 붓꽃자료의 K-평균 군집화

	군집 1	군집 2	군집 3
품종 1	50	0	0
품종 2	0	39	11
품종 3	0	14	36

표 4.2: 붓꽃자료의 이중 K-평균 군집화

	군집 1	군집 2	군집 3
품종 1	50	0	0
품종 2	0	48	2
품종 3	0	4	46

표 4.3: 부분자료의 K-평균 군집화

	군집 2	군집 3
품종 2	47	3
품종 3	14	36

표 4.4: 부분자료의 이중 K-평균 군집화

	군집 2	군집 3
품종 2	48	2
품종 3	2	48

이 사례에서는 이중 K-평균 군집화를 적용함으로써 오히려 군집의 크기가 균일하게 되었다. 유추하건대, 그 원인은 1) Fisher의 붓꽃자료에는 특이점(특이군집)이 없으므로 앞의 사례들과는 다르며 2) 이중 K-평균 군집화가 근사적 최적 규칙으로써 참 군집크기에 접근하게 되기 때문으로 생각된다.

Fisher의 붓꽃자료에 또 하나의 실험을 해보기로 한다. 그림 4.1의 산점도 행렬에서 보듯이 품종 1(setosa)과 품종 2·3(versicolor, verginica)은 시각적으로도 뚜렷이 구분된다. 특히 변수 x3(petal length)과 x4(petal width)의 산점도에서 그것이 명확히 드러난다. 또한 앞의 두 군집화 결과도 그것을 뒷받침한다. 따라서 품종 1을 제외한 품종 2·3의 부분자료($N = 100$)에 K-평균 군집화와 이중 K-평균 군집화를 적용해보자. 물론 여기서도 군집화에 앞서 표준화 처리를 하였다. 표 4.3와 표 4.4이 그 결과이다.

부분자료에 대한 표 4.3와 표 4.4의 군집화가 전체자료의 군집화인 표 4.1과 표 4.2의 결과와 각각 비교하여 향상된 결과를 보여준다. 그러나 여기서 K-평균 군집화가 전체자료의 3-군집화에서 25개의 오류를 보이고 부분자료의 2-군집화에서 17개의 오류를 보이는 등 불안정성(unstability)을 보였다는 데 주목할 필요가 있다. 반면, 이중 K-평균 군집화는 훨씬 안정된 군집화 결과를 보였다. 그 원인은 무엇인가?

K-평균 군집화는 군집변수들의 척도에 크게 의존한다. 왜냐하면 제곱 유클리드 거리를 사용하기 때문이다. 품종 1이 포함된 전체자료에서와 품종 1이 제외된 부분자료에서의 표준화(척도화)는 다를 수밖에 없는데 K-평균 군집화가 그것들로부터 민감한 영향을 받게 된 것이다. 그러나 이중 K-평균 군집화는 덜 그러하다. 왜냐하면 최적규칙 (3.2)와 그것의 조작화인 (3.3)이 관측벡터의 아핀 변환(affine transform)에 대하여 불변적(invariant)이기 때문이다. 구체적으로

$$\boldsymbol{x} \;\rightarrow\; A\,\boldsymbol{x} + \boldsymbol{b} \quad (\text{여기서 } A\text{는 비정칙 행렬}) \tag{4.1}$$

로 변환되는 경우

$$\boldsymbol{\mu} \;\rightarrow\; A\,\boldsymbol{\mu} + \boldsymbol{b}, \quad \Sigma \;\rightarrow\; A\,\Sigma\,A^t, \quad S \;\rightarrow\; A\,S\,A^t$$

이기 때문에 (3.2)와 (3.3)은 (4.1)의 변환에 대하여 불변적이다. 따라서 이중 K-평균 군집화의 단계 II는 위치-척도 변환으로부터 영향을 받지 않는다. 다만, 단계 I에서 통상적인 K-평균 군집화를 채택하기 때문에 이중 K-평균 군집화도 척도화 문제에서 완전히 자유롭지는 않다.

5. 맺음 말

군집분석의 목적이 k개의 군집으로 다변량 자료를 분할하는 것이라면, 자연스럽게 K-평균 군집화를 혼합 다변량정규분포의 문제로 환원하여 볼 수 있다. 즉 p-변량 관측벡터 $\boldsymbol{x}$에 대한 확률밀도로서

$$f(\boldsymbol{x}\ ;\ \pi_1,\cdots,\pi_k\ ,\ \boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k\ ,\ \Sigma_1,\cdots,\Sigma_k)\ =\ \sum_{j=1}^{k}\pi_j\cdot N_p(\boldsymbol{\mu}_j,\Sigma_j),\qquad \sum_{j=1}^{k}\pi_j\ =\ 1$$

을 가정하고 $f(\boldsymbol{x}\ ;\cdot)$을 추정해냄으로써 k개의 군집을 추출해내는 것이다 (Everitt 1974, pp. 34-35; Everitt and Dunn 1991, pp. 113-115). 혼합분포에서 미지의 파라미터들은 최대가능도추정(maximum likelihood estimation) 또는 EM 알고리즘에 의해 얻어질 수 있지만 많은 계산이 요구된다. 이에 반하여 K-평균 군집화는 자료크기에 크게 관계없이 비교적 계산이 간단하므로 기본적으로 탐색적 자료분석(exploratory data analysis; EDA) 기법이다. 이 점에서는 이중 K-평균 군집화도 마찬가지이다.

본 연구에서 제안된 이중 K-평균 군집화는 EDA 기법이면서도 혼합 정규분포 밀도함수 추정에 의한 군집화가 갖는 최적성을 내재화하고 있으므로 매우 상이한 두 군집화 방법의 합성(hybrid)으로 볼 수 있다. 단, 공분산 행렬 $\Sigma_1,\cdots,\Sigma_k$에 대하여 공통성을 가정한 이유는 안정적 군집화를 위한 것이었다. 자료의 크기가 충분히 크다면 군집별로 각기 다른 공분산 행렬을 추정하여 활용할 수도 있을 것이다.

군집화 방법에 대한 평가는 매우 다양한 상황에서 수행될 수 있다 (Milligan 1981; SAS Institute 1990, pp.56-97). 일반적인 결론은 K-평균 군집화가 대체로 양호하지만 모든 상황에서 두루 좋은 결과를 보이지는 않는다는 것이다. 이중 K-평균 군집화도 이 점에서 예외가 될 수는 없다. 이중 K-평균 군집화가 K-평균 군집화에 대한 부분적인 개선일 수는 있어도 모태인 K-평균 군집화의 모든 취약점을 극복하지는 못할 것이다. 연구자가 실험적으로 확인한 한 가지 사실은 이중 K-평균 군집화가 K-평균 군집화와 마찬가지로 폭이 좁고 긴 군집 패턴에 대하여는 취약하다는 점이다.

[추기] 1) 본 연구에 사용된 SAS IML 코드를 얻고자 하는 한국통계학회 회원께서는 저자에 e-mail로 연락하기 바란다. 2) 본 연구에 사용된 연구자의 K-평균 군집화 프로그램에서는 1절 마지막 단락에서 언급한대로 첫 씨(initial seeds)를 임의 배열 자료 세트의 첫 k개 관측으로 하였다. 따라서 첫 씨 심기를 위하여 특별한 알고리즘을 적용한 SAS의 K-평균 군집화인 PROC FASTCLUS와 일부 다른 결과를 낼 수 있다.

참고문헌

[1] Everitt, B.S. (1974). *Cluster Analysis.* Wiley, New York.

[2] Everitt, B.S. and Dunn, G. (1991). *Applied Multivariate Data Analysis.* Edward Arnold, London.

[3] Hartigan, J.A. (1975). *Clustering Algorithms.* Wiley, New York.

[4] Jin, Seohoon (1999). *A Study on the Partitioning Method for Cluster Analysis.* 고려대학교 통계학과 박사학위 논문.

[5] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis,* Third Edition. Prentice Hall, NJ: Englewood Cliffs.

[6] MacQueen, J.B. (1967). "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* 1, 281-297.

[7] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

[8] Milligan, G.W. (1981). "A review of Monte Carlo tests of cluster analysis," *Multivariate Behavioral Research,* 16, 379-407.

[9] SAS Institute (1990). *SAS/STAT User's Guide* (Vol. 1), Version 6 Fourth Edition. SAS Institute, NC: Cary.

[10] Sharma, S. (1996). *Applied Multivariate Techniques.* Wiley, New York.

[1999년 10월 접수, 2000년 3월 채택]

Double K-Means Clustering

Myung-Hoe Huh[1)]

ABSTRACT

In this study, the author proposes a nonhierarchical clustering method, called the "Double K-Means Clustering", which performs clustering of multivariate observations with the following algorithm:

- Step I: Carry out the ordinary K-means clustering and obtain k temporary clusters with sizes $n_1, \cdots, n_k$, centroids $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_k$ and pooled covariance matrix S.
- Step II-1: Allocate the observation $\boldsymbol{x}_i$ to the cluster j^* if it satisfies

$$\frac{n_{j^*}}{N} \cdot \exp\{-(\boldsymbol{x}_i - \boldsymbol{c}_{j^*})' S^{-1} (\boldsymbol{x}_i - \boldsymbol{c}_{j^*})/2\} = \max_{j=1,\cdots,k} \frac{n_j}{N} \cdot \exp\{-(\boldsymbol{x}_i - \boldsymbol{c}_j)' S^{-1} (\boldsymbol{x}_i - \boldsymbol{c}_j)/2\},$$

  where N is the total number of observations, for $i = 1, \cdots, N$.
- Step II-2: Update cluster sizes $n_1, \cdots, n_k$, centroids $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_k$ and pooled covariance matrix S.
- Step II-3: Repeat Steps II-1 and II-2 until the change becomes negligible.

The double K-means clustering is nearly "optimal" under the mixture of k multivariate normal distributions with the common covariance matrix. Also, it is nearly affine invariant, with the data-analytic implication that variable standardizations are not that required. The method is numerically demonstrated on Fisher's iris data.

Keywords: K-means clustering; Optimal allocation rule; Mixture of multivariate normal distributions.

1) Professor, Dept. of Statistics, Korea University. E-mail: stat420@mail.korea.ac.kr