The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 497-508

A Study on K-Means Clustering[1)]

Whasoo Bae[2)] and Se Won Roh[3)]

Abstract

This paper aims at studying on K-means Clustering focusing on initialization which affect the clustering results in K-means cluster analysis. The four different methods(the MA method, the KA method, the Max-Min method and the Space Partition method) were compared and the clustering result shows that there were some differences among these methods, especially that the MA method sometimes leads to incorrect clustering due to the inappropriate initialization depending on the types of data and the Max-Min method is shown to be more effective than other methods especially when the data size is large.

Keywords : K-means clustering, initialization, KA, MA, Max-Min, Space Partition

1. 서론

군집분석(Anderberg(1973))은 비지도학습(Unsupervised Learning)의 한 방법으로서 상호연관성에 근거하여 서로 동질적인 집단으로 분류하는 기법으로 적용되는 분야가 많으며, 또한 군집분석을 통해 생성된 군집에 관한 정보는 하나의 설명변수로 또 다른 분석에 쓰일 수도 있다. 군집분석은 계층적 군집분석(Hierarchical Clustering)과 비계층적 군집분석(Non-Hierarchical Clustering)으로 나뉘는데, 비계층적 군집분석 중 하나인 K-평균 군집분석(K-Means Clustering)은 가장 많이 사용되는 기법 중 하나이다.

K-평균 군집분석(Hartigan(1974))은 자료에 있는 각 개체를 유사한 특성을 지니는 K개의 그룹으로 분할하는 방법으로, 각 군집에 속하는 개체들의 평균값을 중심점으로 하여 근접한 거리에 있는 개체를 묶어서 분할하게 된다. K-평균 군집분석은 거의 모든 형태의 자료에 적용이 가능하고 특별한 변환이 필요 없어서 적용이 쉽다는 장점이 있어 많은 분야에서 사용되고 있다. 허지만 이 방법은 사전에 정해져야 할 군집의 개수, K에 대한 정보가 주어져야 하고, K개 군집의 중심인 초기값(seed)들이 결정이 자료의 종류에 따라 군집의 형성에 상당한 영향을

1) This work was supported by the 2004 Inje University Research Grant.
2) Associate Professor, Department of Data Science, Inje University, Gimhae 621-749, KOREA
E-mail : wbae@stat.inje.ac.kr
3) Master in Data Science, Department of Data Science, Inje Universit, Gimhae 621-749, KOREA

주는 경향이 있다. 부적절한 초기값의 결정은 잘못된 군집의 생성과 군집 생성과정에서 많은 반복을 발생하여 군집생성에 많은 시간이 소요되고, 또한 군집분석의 성능에 상당한 영향을 미치게 되므로 이에 대한 많은 연구가 진행되어 왔다.

초기값 결정에 대하여 많은 방법이 연구되어 있는데, 가장 흔히 사용되는 방법으로는 MA(Macqueen Approach)방법(Macqueen, 1967)을 들 수 있다. MA방법은 자료에서 임의로 K개의 초기값을 선택하고 나머지 개체들은 초기값에 가장 가까운 군집으로 포함시킨 후, 군집의 중심을 다시 계산하여 군집의 중심의 변화량이 임계값(threshold) 이하가 될 때까지 반복하여 군집을 형성하게 한다. 이 방법은 랜덤하게 초기값을 선택하기 때문에 쉽고, 편하게 사용할 수 있는 반면 부적절한 값이 선택되었을 때는 잘못된 군집을 형성시킬 수 있다는 단점이 있다.

MA방법의 단점을 보완하기 위해 Kaufman and Rousseeuw(1990)은 KA(Kaufman Approach)방법을 소개했다. KA방법은 자료의 가장 중앙에 위치한 관측치를 첫 번째 초기값으로 설정하고, 나머지 초기값은 첫 번째 초기값과 일정한 거리 이상 떨어져 있으면서, 군집이 형성되기 쉽도록 초기값을 선택하게 했다. 이 방법은 MA방법보다 정교한 반면, 초기값을 구한 후 다음 단계의 초기값을 하나씩 구하는 과정에서 주변의 모든 관측값들을 고려하기 때문에 자료의 크기가 커지는 경우 계산 량이 많아지는 단점이 있다.

Peña, Lozano와 Larranaga (1999)는 KA방법, MA방법, Random 방법 그리고 FA(Forgy approach)방법(Forgy, 1965) 의 네가지 초기값 설정 방법에 대해 비교를 한 결과, KA가 유용하다고 결론을 내렸지만 자료의 크기가 큰 경우의 계산량의 증가를 고려할 때 이에 대해 보완된 초기치 설정방법에 대해 알아볼 필요가 있다.

본 논문에서는 MA 방법의 문제점과 KA 방법에서 필요한 계산량에 대한 부분을 보완할 수 있는 방법으로 Max-Min 방법과 Space Partition 방법의 두 방법을 제안하여 MA방법, KA방법과 비교해 보고자 한다.

2장에서는 MA방법, KA방법, Max-Min 방법 과 Space Partition 방법들에 의한 초기값 결정에 대한 알고리즘을 기술하고, 3장에서는 Fisher의 붓꽃 자료와 가상으로 만든 모의실험 자료를 이용하여 각 방법에 대한 비교를 하였으며, 결론 및 향후과제에 대해 4장에 기술하였다.

2. K-평균 군집분석의 초기값 결정 방법

K-평균 군집분석에서는 주어진 자료를 K개의 동질적인 집단으로 분할하기 위해서 초기값을 선택해야 하는데, 초기값의 결정은 군집의 형성 및 군집형성에 걸리는 시간에 대해 중요한 요인이 된다.

초기값의 결정에 사용되는 방법들과 그 알고리즘에 대해서 알아보기로 한다.

2.1 MA 방법

MA 방법은 우리가 알고 있는 가장 일반적인 K-평균 군집분석 방법으로할 수 있는데 이 방법의 첫 단계에서는 자료에서 k개의 관측값을 랜덤하게 선택하여 초기값으로 결정하여 초기 군집의 중심을 이루게 한다. 다음 단계에서는 초기값으로 설정되지 않은 관측값들을 가장 가까운 초기값

이 중심인 군집에 배정하여 초기 군집을 형성한다. 형성된 군집의 중심을 다시 계산하여 관측값들과의 거리를 계산하여 가까운 중심의 군집으로 관측값을 이동하여 새로운 중심을 구하고, 군집의 중심의 이동이 임계값 이하가 될 때까지 이 과정을 반복한다.

알고리즘은 <표 1>에 요약되어 있다.

<표 1> MA 방법

1. 주어진 자료에서 랜덤하게 k개의 초기값$(s_1, s_2, \cdots, s_k)$을 선택한다.
2. 각 관측값(x_i)에 대해 초기값(s_i)까지의 거리를 계산한다.
   $d_j = \| x_i - s_j \|, j = 1, \cdots, k, i = 1, \cdots, n$
3. 각 x_i를 단계 2에서 계산된 k개의 d_j중 가장 작은 d_j를 주는 초기값 쪽의 군집으로 할당한다.
4. 군집의 중심을 다시 계산한다.
5. 군집 중심의 변화가 주어진 임계값 이하가 될 때까지 단계 3, 4를 반복 후 중지한다.

2.2 KA 방법

KA 방법은 MA 방법에 의해서 랜덤하게 초기값을 선택했을 때 발생할 수 있는 문제점을 해결하고 가능하면 형성될 군집의 내부에서 초기 군집의 중심이 선택되도록 초기값을 단계적으로 설정해 나간다..

이 방법의 첫 단계에서는 주어진 자료의 가장 중앙에 위치한 관측값을 첫 번째 초기값으로 선택하고 첫 번째 초기값을 제외한 나머지 모든 관측값에 대해서 초기값과 관측값과의 거리를 계산한다. 또 관측값과 자신의 주변에 있는 관측값과의 거리를 고려하도록 기준을 정해 선택된 초기값과 일정한 거리에서 떨어져 있으면서, 주변에 관측값들이 모여 있는 영역에서 다음 초기값을 설정하게 한다. 이 과정은 k개의 초기값이 단계적으로 모두 선택될 때까지 반복하게 된다.

KA 방법에 대한 알고리즘은 <표 2>에 설명되어 있다.

<표 2> KA 방법

1. 주어진 자료에서 가장 중앙에 위치한 값을 첫 번째 초기값 s_1으로 선택한다.
2. 나머지 모든 자료 $x_i (x_i \neq s_1), i = 1, \cdots, n$에 대하여
   a. 초기값으로 선택되지 않은 관측값 $x_j (x_j \neq s_1,\ i \neq j), j = 1, \cdots, n$에 대하여
      $C_{ji} = max(D_j - d_{ji}, 0)$를 계산한다.
      여기서 $d_{ji} = \| x_i - x_j \|$, $D_j = \min(d_{sj})$ (x_s는 선택된 초기값)
   b. x_i에 대해서 $\sum_j C_{ji}$를 계산한다.

(계속)

3. $\sum_j C_{ji}$를 최대화하는 x_i 를 두 번째 초기값으로 선택한다.

4. k개의 초기값이 모두 선택될 때까지 단계 2와 단계3을 반복한다.

5. 나머지 관측값들은 구해진 초기값들에 가장 가까운 쪽으로 군집을 형성 후, 다시 군집의 중심을 구하여, 군집의 중심의 이동이 임계값 이하가 될 때까지 반복한다.

2.3 Max-Min 방법

Max-Min 방법은 MA방법의 초기값을 랜덤하게 선택하였을 때 생기는 문제점을 해결하기 위하여 제안된 KA 방법이 정교하긴 하지만 자료가 많아짐에 따라 초기값 설정에 따른 시간이 많이 걸리는 단점을 보완하기 위한 방법으로 본 논문에서 제안하는 방법이다. 이 방법은 단계적으로 초기값을 선택하되 선택된 초기값들이 다음 초기값을 정하는 데 정보를 줄 수 있도록 하되, KA방법처럼 많은 계산을 하지 않도록 고안했다.

Max-Min 방법에서는 우선 자료에서 랜덤하게 하나의 관측값을 선택하여 첫 번째 초기값으로 선택하고, 첫 번째 초기값에서 나머지 관측값과의 거리를 구하여 그 거리를 최대로 하는 관측값을 두 번째 초기값으로 선택한다. 즉, 초기값을 선택함에 있어 초기값들이 한 곳에 모이는 현상을 방지하기 위해 처음 두 초기값은 멀리 있는 것을 택하게 한다.

다음 단계의 초기값을 구하기 위해서 초기값에 선택되지 않은 나머지 관측값들에 대하여 첫 번째 초기값과의 거리와 두 번째 초기값과의 거리를 구하면 각 관측값에 대해 두 종류의 계산된 거리가 얻어진다. 이 두 거리 중, 최소값을 선택하여 각 관측값에 대해 두 초기값과의 거리로 대응하게 한다.(이 후 이 최소값을 선택된 초기값과의 관측값과의 거리로 정의하기로 한다.)

선택된 초기값과의 관측값과의 거리를 정함에 있어서 최소값을 거리로 정하는 것은 계보적 군집형성을 하는 경우 사용되는 최단연결법(single linkage)의 개념을 적용한다고 보면 된다. 각 관측값에 대응되어 있는 min(관측값과 두 초기값과의 거리)를 비교하여 이 값을 최대로 하는 관측값을 구하여 세 번째 초기값으로 선택함으로써 초기값들이 적절하게 떨어져서 선택되도록 하였다.

이 다음 단계의 초기값을 선택하기 위해서는 이전 단계까지 초기값으로 선택되지 않은 관측값에 대하여 이 과정을 반복적으로 시행하여 k개의 초기값이 모두 선택될 때까지 계속 실시한다.

<표 3>에서 Max-Min 방법에 대해 기술하고 있다.

<표 3> Max-Min 방법

1. 자료에서 랜덤하게 하나의 관측값을 선택하여 첫 번째 초기값 s_1 을 결정한다.
2. 나머지 관측값, $x_i\,(x_i \neq s_1), i = 1, \cdots, n$에 대하여 첫 번째 초기값($s_1$)과의 거리를 최대로 하는 관측값을 두 번째 초기값 s_2로 선택한다.

(계속)

3. 초기값으로 선택되지 않은 모든 관측값 $x_i (x_i \neq s_l \;\; l=1,2)$, $i=1,\cdots,n$에 대하여

a. 초기값 s_1 과 s_2와의 거리를 각각 구하여 이들의 최소값, sd_i을 계산하여 대응시킨다.

$$x_i \leftarrow sd_i = \min\{\| x_i - s_1 \|, \| x_i - s_2 \|\}, \; i=1,\cdots,n \;\; (x_i \neq s_l \;\; l=1,2)$$

b. 각 관측값 x_i에 대응하는 sd_i를 비교하여 이들의 값을 최대로 하는 관측값을 초기값 s_3로 선택한다.

$$s_3 = x_p \leftarrow \max_{1 \leq i \leq n}(sd_i) = sd_p$$

4. 다음 단계의 초기값 s_m, $m=4, \cdots k$을 구하기 위해서 이전 단계에서 구해진 초기값을 추가하면서 다음 단계를 반복하여 k개의 초기값들이 모두 선택되면 정지한다.

나머지 관측값, $x_i (x_i \neq s_l \;\; l=1,\cdots,m-1)$, $i=1,\cdots,n$에 대하여

a. 이미 구해진 초기값들과 x_j와의 거리를 각각 구하고 이들 거리의 최소값을 구한다.

$$x_i \leftarrow sd_i = \min\{\| x_i - s_1 \|, \| x_i - s_2 \|, \cdots, \| x_i - s_{m-1} \|\}, \; i=1,\cdots,n$$
$$(x_i \neq s_l \;\; l=1,\cdots,m-1)$$

b. 각 관측값 x_i에 대해서 얻어진 sd_i를 비교하여 이들의 값을 최대로 하는 관측값을 다음 초기값으로 선택한다.

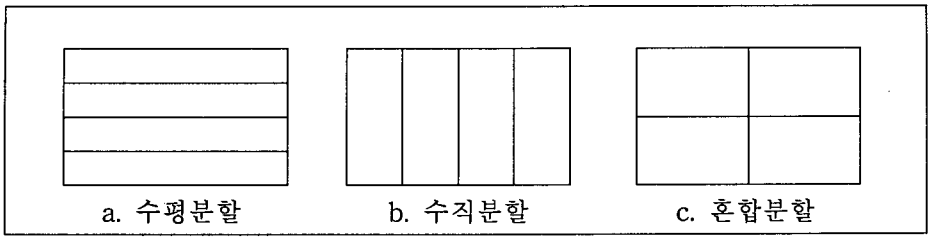
$$s_m = x_q \leftarrow \max_{1 \leq j \leq n}(sd_j) = sd_q$$

5. 나머지 관측값들은 구해진 초기값들에 가장 가까운 쪽으로 군집을 형성 한 후, 다시 군집의 중심을 구하여, 군집의 중심의 이동이 임계값 이하가 될 때까지 반복한다.

2.4 Space Partition 방법

Space Partition 방법에 의한 초기값의 선택은 주어진 자료의 분포를 이용하여 자료의 공간을 군집의 개수인 k개로 분할하여 초기값을 선택하는 방법이다. 물론 주어진 설명변수의 수가 2개라면 산점도를 통해 2차원상에서 자료의 분포를 쉽게 확인할 수 있으므로 자료의 응집현황을 판단하여 산점도 평면을 k개로 분할하면 된다. 설명변수가 3개이상이 될 경우엔 주성분분석(Principal Component Analysis)을 실시하여 얻어진 제 1주성분 점수과 제 2주성분 점수를 이용하여 산점도를 그려서 자료의 분포를 확인한 후, 자료의 응집에 따라 주성분 평면을 k개로 분할한다. 분할된 k 개 공간의 중심에 가장 가까운 값을 가각 초기값으로 선택하고 주성분 점수에 의해 표기된 초기값은 원래 관측값으로 초기값을 환원하여 군집형성에 사용한다. SAS 등의 통계패키지에서는 개체별 주성분점수가 설명변수의 선형변환에 의해 계산되어 설명변수에 대응되어 함께 주어지므로 초기값으로 선정된 개체의 주성분점수에 대응하는 설명변수의 값으로 환원시키는 것이 가능하

는 작성된 산점도를 통해 주어진 k의 수만큼 주관적으로 판단하여 선택하게 한다. 자료의 분포에 따라 균등분할, 비례분할 등의 방법이 선택될 수 있다. <그림 1>은 공간을 분할하는 방법 중 균등분할의 몇 가지 예를 나타내고 있다.



<그림 1> 공간 분할의 방법

Space Partition 방법에 대한 알고리즘은 <표 4>와 같다.

<표 4> Space Partition 방법

1. 주어진 자료에서 설명변수의 개수(p)를 파악한다.
   a. p=2 이면 주어진 자료를 이용해서 2차원 산점도를 작성한다.
   b. p>2 이면 주성분분석을 통한 제 1, 2 주성분을 이용해서 2차원 산점도를 작성한다.
2. 군집의 개수 k 만큼 공간을 분할한다.
3. 각 공간의 중심점(c_i)을 구한다. ($i = 1,2,...,k$)
4. 각 중심점(c_i)과 관측값사이의 거리를 구한다.
5. 각 공간에서 중심점(c_i)과 거리를 최소화하는 관측값 s_i, $i = 1,2,...,k$를 초기값으로 선택한다.
6. 선택된 관측값 s_i에서
   a. p=2 이면 $s_i(i = 1,2,...,k)$를 초기값으로 선택한다.
   b. p>2 이면 s_i에 대한 원래의 관측값으로 환원 후 초기값으로 선택한다.
7. 나머지 관측값들은 구해진 초기값들에 가장 가까운 쪽으로 군집을 형성한 후, 다시 군집의 중심을 구하여, 군집의 중심의 이동이 임계값이하가 될 때까지 반복한다.

3. 사례를 이용한 초기값 결정방법의 비교

3.1 Iris 자료를 통한 비교

Iris 자료는 다변량 자료 분석에서 많이 적용되는 자료로 꽃받침의 길이(sepal length), 꽃받침의 두께(sepal width), 꽃잎의 길이(petal length)와 꽃잎의 두께(petal width)를 나타내는 4개의 변수로 이루어진 150개의 관측값이 Setosa, Versicolour, Virginica의 3가지 품종으로 나뉘어져 품종별

로 50개의 관측값이 할당되어있다.

Iris 자료에 대해 3개의 군집으로 나누기 위하여 MA 방법, KA 방법, Max-Min 방법 그리고 Space Partition 방법에 의한 초기값 설정을 MATLAB으로 프로그래밍하여 구하고 <표 5>에 각 방법에 의해 선택된 초기 군집의 중심을 나타내 두었다. 주어진 변수는 4개이나 이 중심을 Space Partition 방법에서 필요한 산점도와 연결하여 비교하기 위하여 주성분분석 결과 주요한 변수로 나타나는 꽃잎의 길이와 꽃잎의 폭, 두 변수에 대해 나타내었다. 각 방법에 따라 초기 군집 중심은 차이가 나는 것으로 나타나 있다.

<표 5> 각 방법에 의한 초기 군집의 중심

방법 / 초기값	MA 방법		KA 방법		Max-Min 방법		Space Partition 방법	
	꽃잎길이	꽃잎폭	꽃잎길이	꽃잎폭	꽃잎길이	꽃잎폭	꽃잎길이	꽃잎폭
1	5.6	2.4	3.8	1.1	4.0	1.3	1.3	0.2
2	5.8	1.8	1.5	0.2	1.0	0.2	4.0	1.3
3	3.9	1.4	5.3	1.9	6.9	2.3	5.1	1.9

<표 5>에 구해진 초기 중심값을 이용하여 통계패키지 SAS(1999)의 FASTCLUS PROCEDURE를 사용하여 군집이 형성된 결과를 비교해 보았는데 <표 6>에 각 방법에 의해서 얻어진 수렴한 군집의 중심을, <표 7>은 수렴할 때까지 걸린 반복의 횟수를 나타내었고, 형성된 군집 내부에서 각 관측치와 해당군집의 최종 중심과의 차이의 제곱합을 rP산해서 나타내는 오차제곱합은 <표 8>과 같다. MA 방법은 랜덤하게 초기값을 선택하기 때문에 그 때마다 달라지므로 100번의 반복을 통해서 얻어진 평균값으로 나타내었다. <그림 2>, <그림 3>, <그림 4> 그리고 <그림 5>는 네 방법에 의해 형성된 초기 군집의 중심, 수렴한 군집의 중심 그리고 군집을 보여준다.

<표 6> 수렴한 군집 중심의 비교 (※100번 반복의 평균)

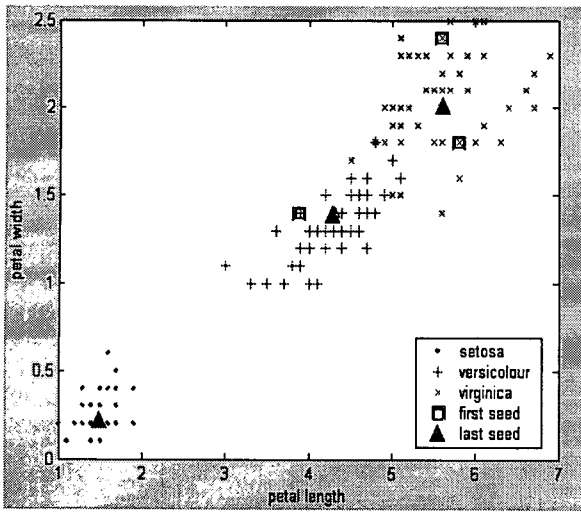
방법 / 수렴된값	MA 방법※		KA 방법		Max-Min 방법		Space Partition 방법	
	꽃잎길이	꽃잎폭	꽃잎길이	꽃잎폭	꽃잎길이	꽃잎폭	꽃잎길이	꽃잎폭
1	1.5	0.2	1.5	0.2	1.5	0.2	1.5	0.2
2	4.3	1.4	4.3	1.4	4.3	1.4	4.3	1.4
3	5.6	2.0	5.6	2.0	5.6	2.0	5.6	2.0

<표 7> 수렴할 때까지의 반복의 수 (※100번 반복의 평균)

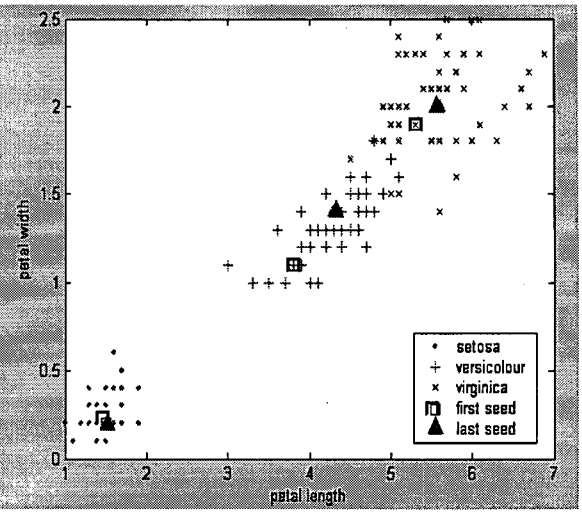
MA 방법※	KA 방법	Max-Min 방법	Space Partition 방법
7.5	6	7	6

<표 8> 생성된 군집의 오차제곱합 (※100번 반복의 평균)

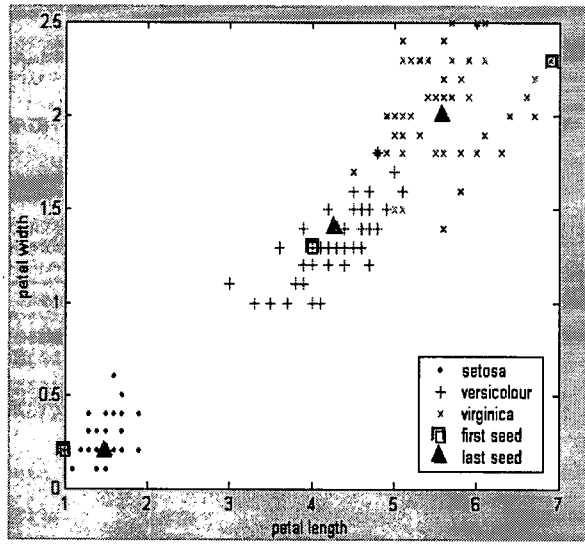
MA 방법※	KA 방법	Max_Min 방법	Space Partition 방법
31.4321	31.4293	31.4293	31.4293



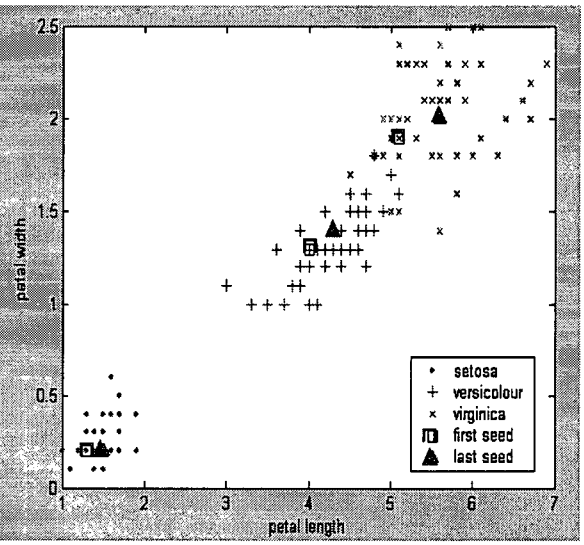
<그림 2> MA 방법 적용 결과



<그림 3> KA 방법 적용 결과



<그림 4> Max-Min 방법 적용 결과



<그림 5> Space Partition 방법 적용 결과

군집 분석 결과를 살펴보면, KA 방법, Max-Min 방법 그리고 Space Partition 방법은 일정한 학습 규칙에 따라서 초기값을 선택하기 때문에 항상 수렴한 군집의 중심이 같은 반면에 MA 방법은 초기값을 랜덤하게 선택하기 때문에 선택된 초기값에 따라서 수렴한 값이 다르게 나타나 100번의 반복 중 첫 번째 결과만 표시하였으며, 오차제곱합은 100번 반복의 평균값을 사용하였다. <그림 3>, <그림 4> 그리고 <그림 5>를 통해서 KA 방법, Max-Min 방법, Space Partition 방법은 초기 군집 중심과 수렴한 군집 중심사이에 변화가 거의 없는 것을 알 수 있다. 수렴하기까지 반복한 횟수는 KA 방법과 Space Partition 방법이 6번으로 가장 작았으며, Max-Min 방법, MA 방법의 순으로 나타나고 있음을 <표 3>에서 볼 수 있다. 군집분석을 통해서 수렴한 각 군집의 중

심과 각 군집에 속한 관측값들 사이의 오차제곱합을 나타내는 <표 8>의 결과를 살펴보면 KA 방법과 Space Partition 방법이 반복의 횟수와 오차제곱합이 가장 낮게 나타났지만, 초기값을 구하는 과정을 전반적으로 고려했을 때 Space Partition 방법이 가장 좋은 결과를 보이고 있음을 알 수가 있었다.

3.2 모의실험 자료를 통한 비교

앞에서 Iris 자료를 이용하여 초기값 설정 방법에 따라 군집분석에 미치는 영향에 대하여 비교 분석 하였다. 좀 더 정확한 비교를 위해서 가상의 자료를 만들었다. 가상의 자료는 x와 y의 2개의 변수를 가지며, 4개의 그룹으로 나뉘어져 있다. 각각의 그룹은 250개씩 총 1000개의 관측값을 가지게 하였으며, 이들은 모두 다 정규분포를 이용해 분산은 동일하게 주고, 평균은 각 그룹마다 다르게 주었다. <표 9>는 가상의 자료를 생성하는 MATLAB 코드이다.

<표 9> 모의실험 자료의 생성

```
x1=randn(250, 2)*0.5;
x2=3.3+randn(250, 2)*0.5;
x3=5+randn(250, 2)*0.5;
x4=[x1(:, 1), 8+x1(:, 2)];
a=[x1; x2; x3; x4];
```

x1은 첫 번째 그룹의 자료를 형성하는 것으로 x변수와 y변수 각각 $N(0,0.5^2)$의 난수를 생성시키고, x2는 두 번째 그룹으로 $N(3.3,0.5^2)$, x3은 세 번째 그룹으로 $N(5,0.5^2)$으로부터 생성시켰으며, x4의 경우는 x1과 동일한 분포를 따르지만, y변수에만 8을 더하여 네 번째 그룹을 형성하게 만들어 주었다.

이 자료를 이용하여 MA 방법, KA 방법, Max-Min 방법 그리고 Space Partition 방법에 의한 초기값을 구한 결과가 <표 10>에 나타나 있다.

<표 10> 각 방법에 의해 선택된 초기 군집의 중심

방법 / 초기값	MA 방법		KA 방법		Max-Min 방법		Space Partition 방법	
	x	y	x	y	x	y	x	y
1	0.3572	-0.6283	2.1762	3.8824	0.1620	8.3571	0.3282	0.9687
2	-0.2770	7.8268	0.0416	7.9803	0.0038	-1.2600	3.6766	1.9779
3	5.4942	5.8513	-0.0599	-0.0358	6.6130	4.7817	0.2532	6.8438
4	-0.0192	-0.2642	4.8742	4.9916	1.9825	3.6102	4.4716	6.3934

각 방법에 의해 선택된 초기값을 이용하여 SAS의 FASTCLUS PROCEDURE에 의한 군집분석을 실시하였으며 <표 11>은 각 방법의 적용 결과로 얻어진 수렴한 군집의 중심을 나타내고 있으며, <표 12>는 수렴하기까지 반복의 횟수를 나타낸다. 형성된 군집의 적절성을 평가하기 위한 오

차제곱합은 <표 13>과 같다. MA 방법은 랜덤하게 초기값을 선택하기 때문에 100번 반복을 통해서 평균값을 구하였다. <그림 6>, <그림 7>, <그림 8> 그리고 <그림 9>는 각 방법에 의해서 선택된 초기값을 이용해 실시한 군집분석결과이다.

<표 11> 수렴한 군집 중심의 비교

방법 / 수렴된값	MA 방법※		KA 방법		Max-Min 방법		Space Partition 방법	
	x	y	x	y	x	y	x	y
1	0.2176	-0.3755	-0.0188	-0.0441	-0.0188	-0.0441	-0.0188	-0.0441
2	4.1328	4.1689	3.2937	3.2903	3.2937	3.2903	3.2937	3.2903
3	-0.2515	0.2819	4.9853	5.0615	4.9853	5.0615	4.9853	5.0615
4	0.0570	8.0077	0.0570	8.0077	0.0570	8.0077	0.0570	8.0077

※100번 반복의 첫 번째 결과

<표 12> 수렴할 때까지의 반복의 수

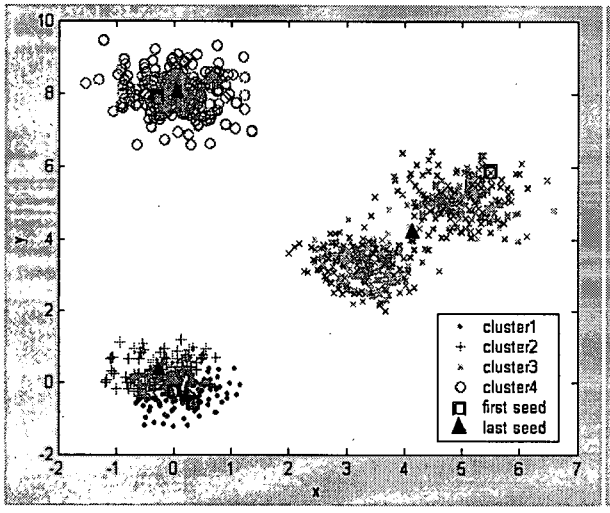
MA 방법※	KA 방법	Max-Min 방법	Space Partition 방법
8	3	3	3

※100번 반복의 평균

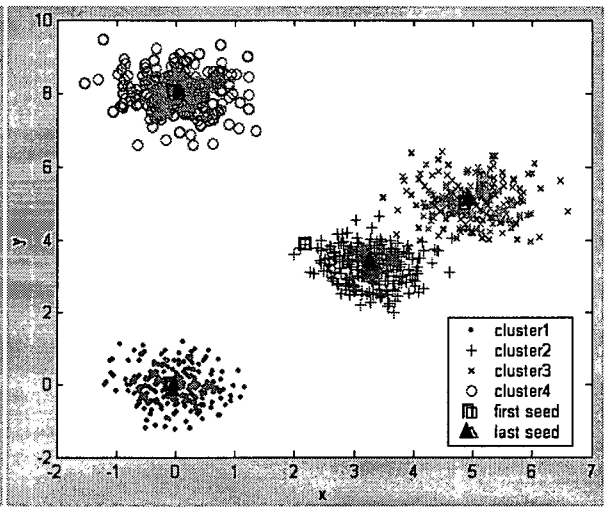
<표 13> 형성된 군집의 오차제곱합

MA 방법※	KA 방법	Max-Min 방법	Space Partition 방법
764.4968	481.0342	481.0342	481.0342

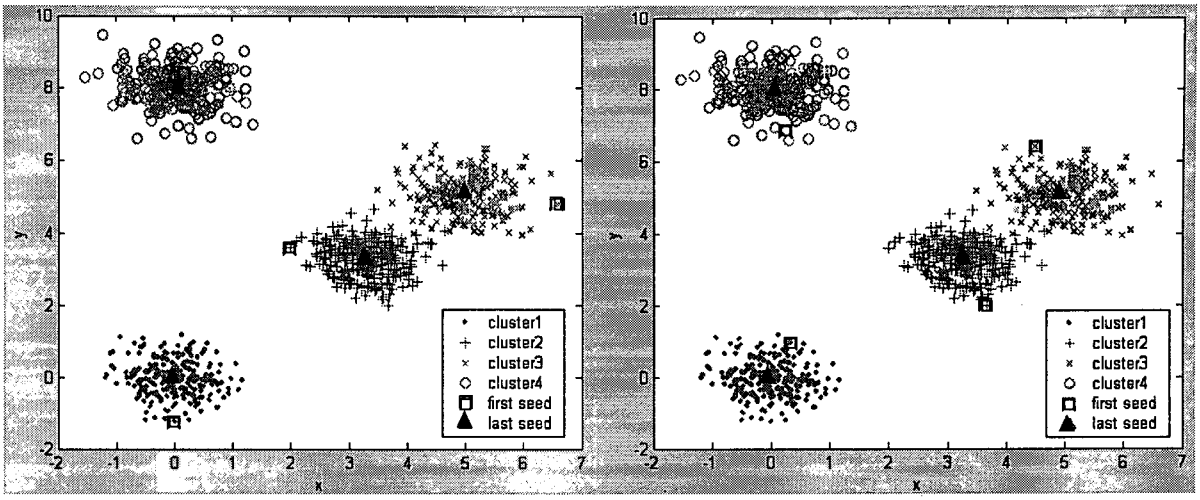
※100번 반복의 평균



<그림 6> MA 방법 적용 결과



<그림 7> KA 방법 적용 결과



<그림 8> Max-Min 방법 적용 결과

<그림 9> Space Partition 방법 적용 결과

모의실험 자료에 대한 군집분석결과를 살펴보면, MA 방법은 초기값을 랜덤하게 선택하기 때문에 선택된 초기값에 따라서 수렴한 값이 다르게 나타나 100번 반복의 평균값을 사용하였다. 수렴할 때까지 반복한 횟수는 KA 방법, Max-Min 방법 그리고 Space Partition 방법이 3번으로 MA 방법의 8번에 비해 작게 나타났다. 부적절한 초기값 때문에 <표 13>에서 보이는 것과 같이 오차제곱합이 다른 방법에 의해 생성된 값보다 2배 정도 크게 나타났으며, 그룹 2와 그룹 3을 동일한 군집으로 분류하는 경우도 나타났다. 초기값 선택 시 수렴까지 반복의 수, 오차제곱합을 종합적으로 비교해 볼 때 KA방법, Max-Min 방법과 Space Partition 방법이 똑같은 결과를 주었다.

4. 결론 및 향후과제

비계층적 군집분석에서 일반적으로 많이 사용되고 있는 방법이 K-평균 군집분석이다. K-평균 군집분석은 거의 모든 형태의 자료에 적용이 가능하며, 적용이 쉬운 반면 초기값에 민감한 단점이 있다. 본 논문에서는 K-평균 군집분석에서 초기값을 결정하는 방법들 중에서 MA 방법, KA 방법, Max-Min 방법 그리고 Space Partition 방법을 비교 분석함으로서 초기값 선택의 중요성 및 군집 분석 결과에 미치는 영향에 대해서 알아보았다.

Iris 자료, 모의실험 자료를 이용한 비교 연구를 통해서 전반적인 군집분석의 결과를 살펴보면, KA 방법의 경우는 예제 자료를 통해서 볼 수 있었듯이 초기값이 형성될 군집에 위치하여 수렴할 때까지의 반복의 수가 작았으며, 생성된 군집의 오차 제곱합도 작게 나타난 반면, 초기값을 구하는 과정에서 주변의 모든 관측값들의 거리를 고려하기 때문에 많은 시간이 소요되었다. Iris 자료의 경우처럼 관측값의 수가 작은 경우에는 MA 방법, Max-Min 방법, Space Partition 방법 모두 별 차이가 없었으나, 모의실험 자료의 경우처럼 비교적 관측값의 수가 많은 경우는 상당한 차이를 보였다.

MA 방법의 경우는 랜덤하게 초기값을 선택하기 때문에 초기값을 선택하기까지의 시간이 적게 걸리는 장점이 있지만 부적절한 초기값이 선택될 경우 수렴할 때까지의 반복의 수가 증가하거나 모의실험 자료의 분류 결과와 같이 잘못된 결과를 야기할 수도 있다.

Space Partition 방법은 산점도를 통한 자료의 분포를 바탕으로 동일하게 자료의 공간을 분할하

여 초기값을 선택하는데, MA 방법, Max-Min 방법보다는 조금 더 많은 시간이 소요되었으나, 관측값의 수가 증가함에 따라 많은 시간이 소요되는 KA 방법에 비해서 아주 적은 시간이 소요되었다. 또한 형성될 군집의 주변으로 초기값이 선택되었기 때문에 수렴까지의 반복의 수나 오차제곱합도 작게 나타났지만, 자료의 설명변수가 많을 경우 주성분분석을 통해 초기값의 정보를 찾은 후, 다시 원래의 자료로 환원해야 하는 번거로움이 있었다.

Max-Min 방법의 경우는 초기값을 선택할 때 기존의 선택된 초기값과의 거리만을 고려하기 때문에 MA 방법에 비해서는 많은 시간이 소요되지만 관측값의 수가 증가하더라도 KA 방법에서 소요되는 시간만큼 많이 증가하지 않았으며, 비교적 형성될 군집의 주변으로 초기값이 선택되기 때문에 수렴할 때까지의 반복의 수나 형성된 군집의 오차 제곱합도 작게 나타났다.

비교 분석을 통한 전반적인 군집분석의 결과는 KA 방법, Max-Min 방법, Space Partition 방법은 선택된 초기값의 분포와 수렴까지의 반복의 수와 오차 제곱합이 비슷하게 나타난 반면, MA 방법의 경우는 부적절한 초기값이 선택될 경우 잘못된 결과를 야기할 수 있음을 보았다.

사용되는 자료의 크기가 점점 증가하는 대용량의 자료에서 군집분석을 통해 유용한 정보를 찾는 경우에는 형성될 군집의 주변에 초기값이 선택되면서 초기값 선택 시에 소요되는 시간을 줄일 수 있는 Max-Min 방법이 유용하게 사용될 수 있을 것이다.

본 논문에서 사용되었던 자료들은 이상치를 포함하고 있지 않는 자료지만 실제로 사용되는 자료들은 입력 오류 등으로 인한 많은 이상치를 포함하고 있다. 이상치에 대한 사전 확인 및 처리도 필요하지만 이상치를 포함하는 자료의 경우의 군집 분석을 위한 초기값 선택에 대한 연구가 필요하다고 생각된다. 또한 K-평균 군집분석에서는 평균을 군집의 중심을 생각하기 때문에 이상치에 영향을 많이 받으므로 평균 대신 중앙값을 이용하는 K-Median Clustering을 이용하는 경우의 초기값에 설정 및 군집 형성에 대한 연구도 병행할 필요가 있다고 생각한다.

참고문헌

[1] Anderberg M.R (1973). *Cluster Analysis for Applications*. Academic Press, New York.

[2] Forgy, E. (1965). Cluster Analysis of Multivaruate Data; Efficiebcy vs. Interpretability of Classification. *Biometrics*, 21, 768.

[3] Hartigan J.A (1974). *Clustering Algorithms*. John Wiley & Sons, New York.

[4] Kaufman L and Rousseeuw P.J(1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. John Wiley & Sons, Canada.

[5] Macqueen J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proc. Symp. Math. and Probability*, 5th, Berkeley, 1, 281-297, AD 669871. University of California Press, Berkeley, CA.

[6] Peña J.M., Lozano J.A. and Larranaga P. (1999). An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Lett*, 20 : 1027-1040.

[7] SAS/STAT User's Guide Version 8(1999), *SAS Publishing*, 1193-1244.

[2005년 5월 접수, 2005년 6월 채택]