

6장. 군집분석





군집분석

1. 군집분석 개요
2. 유사성의 측정
3. 군집화의 방법
4. 군집분석시 유의사항

군집분석 개요

□ 군집(Cluster)

- 관련 다변량적 특성이 그룹 내적으로는 균일하고 외적으로는(타 그룹과는) 이질적인 관측개체의 모임

➔ 군집화의 기준 : 동일한 군집에 속한 개체(또는 개인)는 여러 속성이 유사하고, 서로 다른 군집에 속한 개체는 다른 속성을 갖도록 군집 구성

□ 군집화를 위한 변수

- 전체 개체(개인)의 속성을 판단하기 위한 기준

(예) 고객세분화

인구통계적 변인(성별, 나이, 거주지, 직업, 소득, 교육, 종교, ...)

구매패턴 변인(상품, 주기, 거래액, ...)


생활패턴 변인(라이프스타일, 성격, 취미, 가치관, ...)

군집분석 개요

□ 군집분석(Cluster Analysis) 정의

- 군집들의 개수나 구조에 관한 아무런 가정 없이 개체들 사이의 유사성 (similarity) 또는 거리에 근거하여 '자연스러운' 군집을 찾고 나아가 자료의 요약을 꾀하는 원시적이고 탐색적인 통계적 방법

□ 군집분석(Cluster Analysis) 목적

- 각 개체가 군집의 갯수, 내용, 구조 등이 사전에 정의되지 않은 상황에서
- 군집의 구성원이 됨을 개체 사이의 '유사성'(또는 비유사성)에 근거하여 식별함으로써
- 전체 다변량 자료의 구조를 파악하고
- 군집의 형성과정과 그 특성, 그리고 식별된 군집간의 관계 등을 체계적으로 연구, 분석하는 과정의 총체  적절한 군집으로 나누고, 각 군집의 특성, 군집간의 차이 등에 대한 탐색적 연구

군집분석 개요

□ 군집분석 의미

- 군집분석은 데이터 내에서 그룹을 찾아내는 다변량 분석의 한 기법
- 이 집단에서 대상은 케이스(개체)나 변수
 - 케이스(개체)의 군집분석 : 집단이나 분류에서 대상의 set을 분류한다는 측면에서 판별분석과 같음
 - 변수들의 군집분석 : 변수들의 관련된 집단으로 결부시키는 면에서 요인 분석과 같음

□ 군집분석의 특징

- **상관관계를 바탕으로 하지 않고** 단지 측정치의 차이를 이용
- 계층적(hierarchical) 군집과 비계층적(nonhierarchical) 군집

군집분석 개요

□ 군집분석의 활용 사례

- 생물분류학 : 생물을 특성에 따라 분류
- 의학 : 증세에 따라 분류된 환자에 대한 처방의 결정
- 심리학 : 성격유형에 따른 개인들의 분류
- 인류학 : 석기(stone tools)나 화석 등에 근거한 문화발달과정

□ vs. 판별분류분석

- 판별분류분석은 이미 알려진 집단정보가 각 개체에 대해 일단 주어져 있을 경우
- 이들 집단간의 차이를 분석하고
- 집단정보를 가지지 않은 새로운 개체를 이미 주어진 부분집단 중 하나에 분류하는 다변량 분석 방법

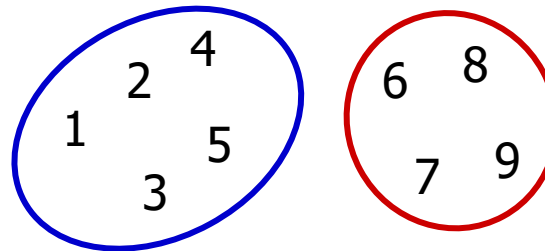
군집분석 개요

□ 군집의 유형

- 상호배반적(disjoint) 군집

각 개체가 상호배반적인 여러 군집 중, 오직 하나에만 속함

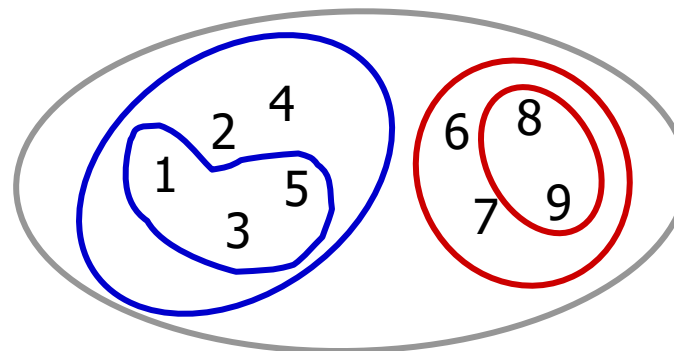
(예) 식품, 가전제품, 의류품



- 계층적 (hierarchical) 군집

한 군집이 다른 군집의 내부에 포함되는 형태로 군집간의 중복은 없으며, 군집들이 매단계 계층적인(나무) 구조를 이룸

(예) 전자제품 → 주방용 → 냉장고

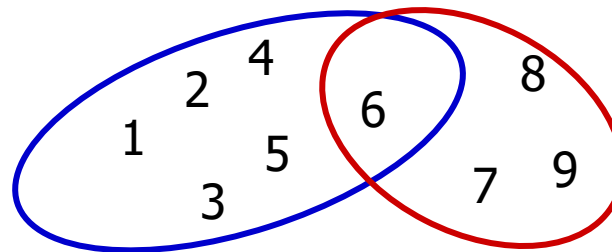


군집분석 개요

□ 군집의 유형

– 중복(overlapping) 군집

두 개 이상의 군집에 한 개체가 동시에 소속되는 것을 허용.



– 퍼지(fuzzy) 군집

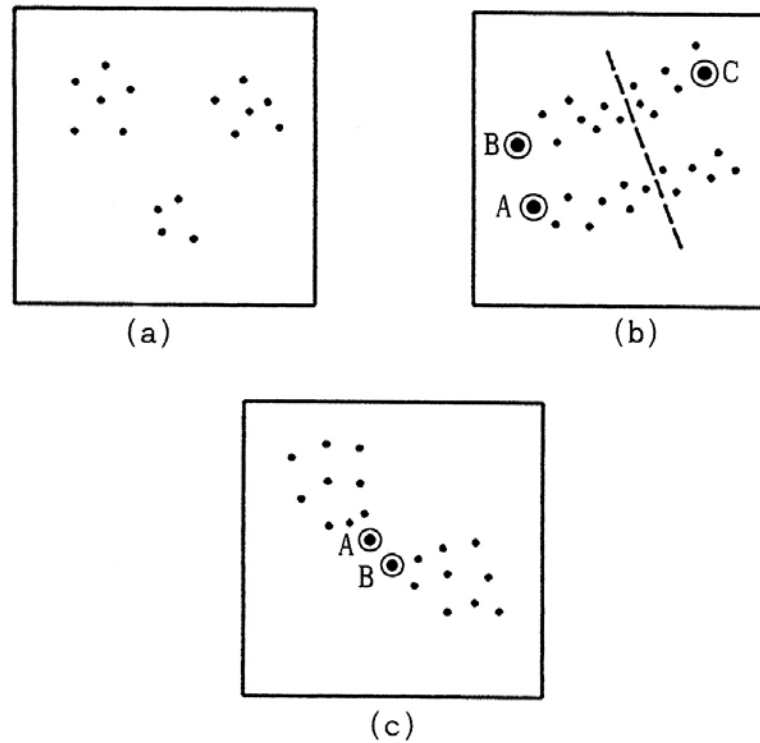
개체가 소속되는 특정한 군집을 표현하는 것이 아니라 각 군집에 속할 확률이나 지표로 규정(상호배반적, 계보적, 중복 등 어느 형태도 가능)

$$\text{Prob (개체 1 } \in \text{ 군집 A)} = 0.7$$

$$\text{Prob (개체 1 } \in \text{ 군집 B)} = 0.3$$

군집분석 개요

□ 군집의 형태



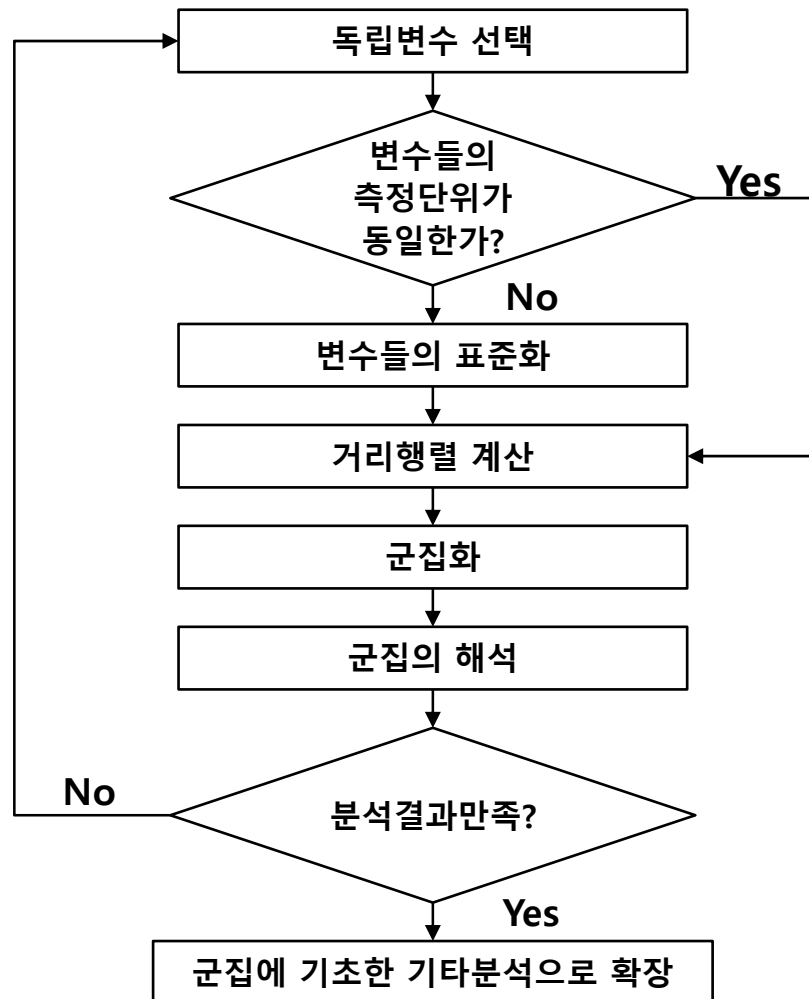
[a] 대부분의 군집방법들이 만족할 만한 결과 제공

[b] 개체 B와 C는 같은 군집에 속하는데도 불구하고 A와 B가 더 가깝다고 판단

[c] A와 B가 두 군집 사이의 고리 역할을 하여 군집방법에 따라서는 단순히 하나의 군집으로 결론 지을 수도 있음

군집분석 개요

□ 군집분석 절차 및 핵심 과제



핵심 과제

1. 어떠한 특성들을 비교할 것인가?
→ 독립변수 선정
2. 어떻게 유사성 거리를 측정할 것인가?
→ 유사성 거리 측정방법 선정
3. 어떻게 동질적인 집단으로 분류할 것인가?
→ 군집화 방법 선정

군집분석 개요

□ 독립변수 선정

- 군집분석은 독립변수 선정에 크게 의존하기에, 어떤 변수를 목적에 맞는 개체 특성으로 판단하는 것이 매우 중요함

1) 요인분석을 통해 변수들의 차원을 정리

: 다수의 변수로부터 추출된 소수의 요인들을 독립변수로 선정하는 방법

→ 새로 추출된 변수가 의미 있는 역할을 할 경우 효과적임

2) 구분 목적에 해당하는 변수를 독립변수로 선정

인구통계적 변인(성별, 나이, 거주지, 직업, 소득, 교육, 종교, ...)

구매패턴 변인(상품, 주기, 거래액, ...)

생활패턴 변인(라이프스타일, 성격, 취미, 가치관, ...)

유사성의 측정

□ 유사성 측정

- 측정한 변수들을 이용하여 모든 개체들 간의 거리(distance) 또는 비유사성(dissimilarity)을 계산하여, 모든 개체들 사이의 비유사성(dissimilarity)을 나타내는 거리행렬을 구함
- 유사성은 값이 클수록 두 개체 사이가 가깝다는 것을 의미하고, 비유사성(또는 거리)은 값이 작을수록 두 개체 사이가 가까움을 의미

□ 유사성의 측도

- 거리의 측도와는 달리 유사성(similarity)의 측도는 두 개체간의 측도값이 클수록 서로 가까운 것으로 인식
- 이러한 측도로는 상관계수, 두 벡터간의 코사인(cosine) 등이 이용
- 군집분석에서 유사성의 측도가 흔히 쓰이는 경우 중의 하나는 두 개체가 여러 가지 상황에서 같은 유(양성 또는 합격)를 나타내든지 또는 무(음성 또는 불합격)를 나타내는 분할표 등의 경우

유사성의 측정

□ 비유사성(거리)의 측도

- 거리의 측도는 두 개간의 측도값이 작을수록 가깝고 클수록 관련이 멀게 됨
- 비유사성(거리)측정방법에는
 - 유클리디안 (Euclidean) 거리
 - 유클리디안(Euclidean) 제곱거리
 - 도시블럭거리(City-block 또는 Manhattan distance)
 - 민코우스키 거리(Minkowski distance)
 - 마할라노비스 거리 (Mahalanobis distance)

※ 위 방법 중 하나를 선택하여 각 개체의 쌍에 대한 유사성/비유사성을 측정

이때 각 변수의 측정 단위가 서로 다르면 각 변수의 중요도가 달라지므로 각 변수들을 표준화한 후에 측정

유사성의 측정

※ 변수 표준화

- 변수들간의 단위, 크기에 민감한 거리척도는 변수들을 표준화시켜 진행

변수 표준화 방법

① i기준 변수변환 i 에 대해 x_{ij} 를 $\frac{(x_{ij} - \overline{x_{i.}})}{s_{i.}}$

② j기준 변수변환 j 에 대하여 $\frac{(x_{ij} - \overline{x_{.j}})}{s_{.j}}$

③ -1부터 1까지 범위로 변환

④ 0부터 1까지 범위로 변환

⑤ 평균 1로 변환

⑥ 표준편차 1로 변환

✓ 변수들 간의 단위 차이로 인해
발생하는 문제점을 제거

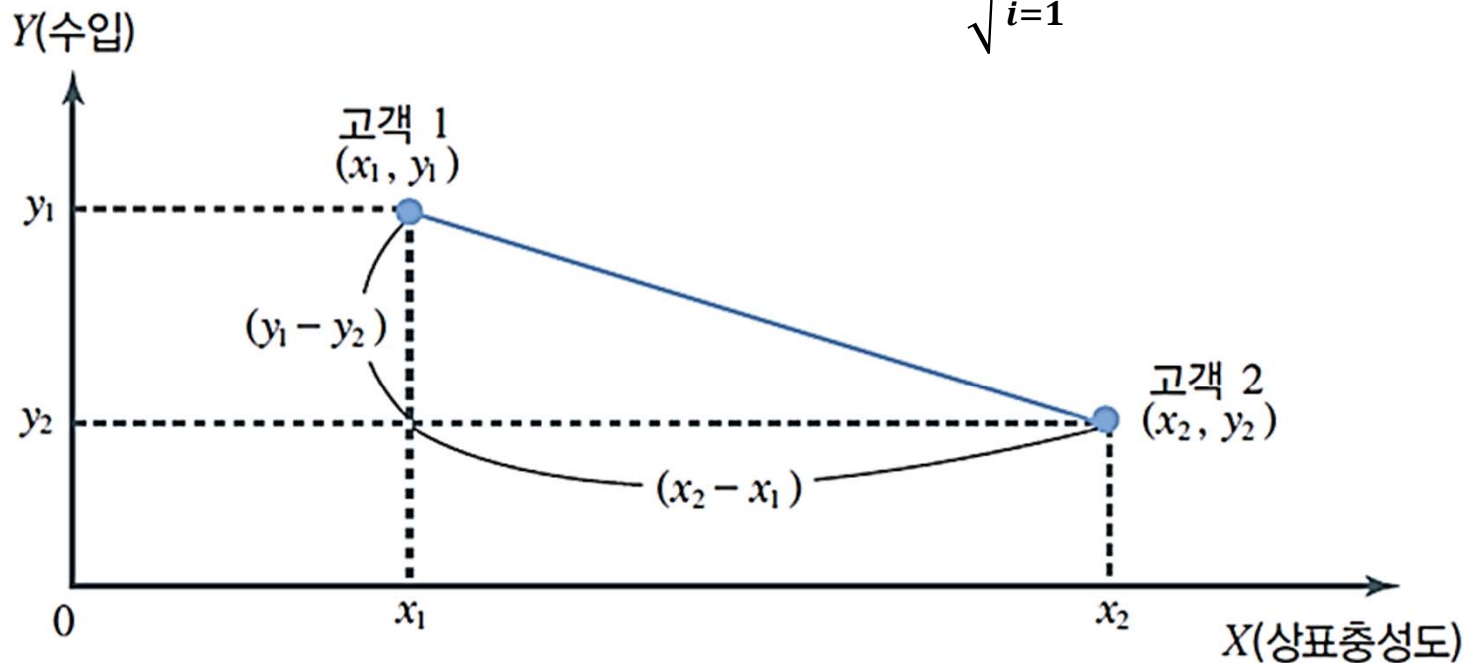
✓ 변수들에 내재되어 있는 구조를
정확하게 파악 가능

유사성의 측정

□ 거리 측도 방법

- 유클리디안 (Euclidean) 거리 : 변수들의 차이를 제곱하여 합산한 거리,
- 가장 일반적으로 사용

$$D(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$



$$d_{12} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

유사성의 측정

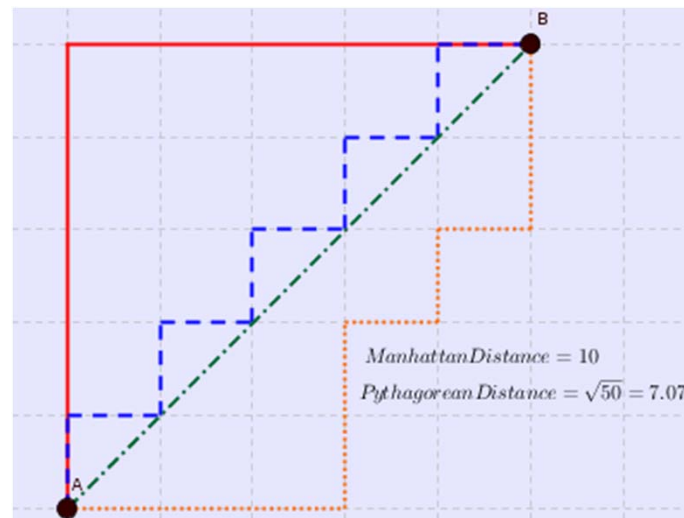
□ 거리 측도 방법

- 유클리디안 (Euclidean) 제곱거리 : 유클리디안 거리를 제공한 거리 측정방법

$$D(X, Y) = \sum_{i=1}^n (X_i - Y_i)^2$$

- 맨하튼 거리, 도시블럭 거리(City-block 또는 Manhattan distance)
: 변수값들의 차이를 절대값화하여 합한 거리 측정 방법

$$D(X, Y) = \sum_{i=1}^n |X_i - Y_i|$$



유사성의 측정

□ 거리 측도 방법

- 민코우스키 거리(Minkowski distance) : 거리를 산정하는 일반식으로 함수에 포함된 지수들을 조정해 줌으로써 다양한 방식의 거리 측정

$$D(X, Y) = \left(\sum_{i=1}^n (|X_i - Y_i|)^m \right)^{1/m} \quad m > 0$$

- 표준화 거리 또는 통계적 거리(statistical distance)

$$D(X, Y) = \left((X - Y)' D^{-1} (X - Y) \right)^{1/2} \quad D = \text{diag}(V_x, V_y) \text{ 는 표본분산행렬}$$

- 마할라노비스 거리(Mahalanobis distance)

$$D(X, Y) = \left((X - Y)' S^{-1} (X - Y) \right)^{1/2} \quad S = \{S_{ij}\} \text{ 는 표본공분산행렬}$$

유사성의 측정

□ 거리 측도 방법

- 캔버라 거리(Canberra distance)

$$D(X, Y) = \sum_{i=1}^n \frac{|X_i - Y_i|}{(X_i + Y_i)}$$

- 체비셰프 거리(Chebyshev distance)

$$D(X, Y) = \max_i |X_i - Y_i|$$

거리가 가까울수록 유사성이 크고, 거리가 멀수록 비유사성 (dissimilarity)이 큰 사실을 군집화 단계에서 적용할 수 있다.

군집화 방법

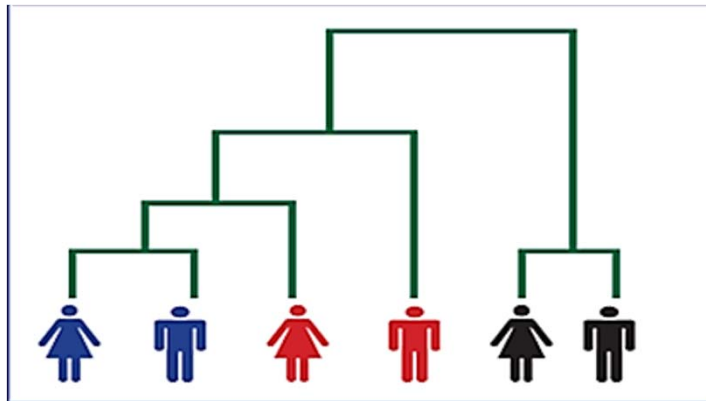
□ 군집화 방법

– 군집 대상의 중복 유, 무에 따라 상이한 방법론의 적용

• 중복이 없는 경우

✓ 계층적인 방법(hierarchical cluster)

- **자료의 크기가 작은 경우** 활용하는 방법
- 개별대상간의 거리에 의하여 가장 가까이 있는 대상들로부터 시작하여 결합하여 감으로써 나무모양의 계층구조를 형성해가는 방법
- **덴드로그램**을 그려 줌으로써 군집이 형성되는 과정을 정확하게 파악할 수 있으나 자료의 크기가 크면 분석하기 어려움
- 계층군집의 수를 정하지 않고, 대분류에서 소분류까지의 계층적 구조를 얻는 방법



각 개체가 군집으로 묶이면서 소속이 바뀌지 않으며, 최종적으로는 1개의 군집으로 묶여지는 방법임

계층적 군집분석

- ❖ 병합적 계층 군집분석 (agglomerative hierarchical cluster analysis): 모든 개체가 서로 다른 군집에 속하는 경우에서 시작하여 모든 개체가 한 개의 군집에 속할 때까지 매 단계마다 두 개의 개체(또는 군집)를 병합
- ❖ 분할적 계층 군집분석 (divisive hierarchical cluster analysis): 모든 개체가 한 개의 군집에 속하는 경우에서 시작하여 모든 개체가 서로 다른 군집에 속할 때까지 매 단계마다 두 개의 개체(또는 군집)를 분할
- ❖ 한 번 진행한 분할 또는 병합은 되돌릴 수 없음.
- ❖ 연구자가 적절한 그룹의 숫자를 결정해야 함.

□ Data : crime.csv

□ 문제 제기

– 1975년 미국 대도시의 강력범죄에 관한 자료에 대해 군집 분석을 하고자 함.

(인구 100,000명당 사건 발생률)

```
> crime.data<-
read.csv("crime.csv",header=T)
> crime.data<- crime.data[,-1]
> crime.data
```

	city	murder	rape
1	Atlanta	16.5	24.8
2	Boston	4.2	13.3
3	Chicago	11.6	24.7
4	Dallas	18.9	34.2
5	Denver	6.9	41.5
6	Detroit	13.0	35.7
7	Hartford	2.5	8.8
8	Honolulu	3.6	12.7
9	Houston	16.8	26.6
10	Kansas City	10.8	43.2
11	Los Angeles	9.7	51.8
12	New Orleans	10.3	39.7
13	New York	9.4	19.4
14	Portland	5.9	23.0
15	Tucson	5.1	22.9
16	Washington	12.5	27.6

```
> dim(crime.data)
```

```
[1] 16  3
```

16개의 자료로 분석

– 체비셰프 거리(Chebyshev distance) $D(X,Y) = \max_i |X_i - Y_i|$

```
> dist(crime.data[,c(2:3)],method="max",diag=T)
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.0															
2	12.3	0.0														
3	4.9	11.4	0.0													
4	9.4	20.9	9.5	0.0												
5	16.7	28.2	16.8	12.0	0.0											
6	10.9	22.4	11.0	5.9	6.1	0.0										
7	16.0	4.5	15.9	25.4	32.7	26.9	0.0									
8	12.9	0.6	12.0	21.5	28.8	23.0	3.9	0.0								
9	1.8	13.3	5.2	7.6	14.9	9.1	17.8	13.9	0.0							
10	18.4	29.9	18.5	9.0	3.9	7.5	34.4	30.5	16.6	0.0						
11	27.0	38.5	27.1	17.6	10.3	16.1	43.0	39.1	25.2	8.6	0.0					
12	14.9	26.4	15.0	8.6	3.4	4.0	30.9	27.0	13.1	3.5	12.1	0.0				
13	7.1	6.1	5.3	14.8	22.1	16.3	10.6	6.7	7.4	23.8	32.4	20.3	0.0			
14	10.6	9.7	5.7	13.0	18.5	12.7	14.2	10.3	10.9	20.2	28.8	16.7	3.6	0.0		
15	11.4	9.6	6.5	13.8	18.6	12.8	14.1	10.2	11.7	20.3	28.9	16.8	4.3	0.8	0.0	
16	4.0	14.3	2.9	6.6	13.9	8.1	18.8	14.9	4.3	15.6	24.2	12.1	8.2	6.6	7.4	0.0

method 를 이용하여 여러가지 거리 행렬을 구할 수 있다.

method: the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski"

```
> dist(crime.data[,c(2:3)],method="manhattan")
      1      2      3      4      5      6      7      8      9     10     11     12     13     14     15
2  23.8
3   5.0 18.8
4  11.8 35.6 16.8
5  26.3 30.9 21.5 19.3
6  14.4 31.2 12.4  7.4 11.9
7  30.0  6.2 25.0 41.8 37.1 37.4
8  25.0  1.2 20.0 36.8 32.1 32.4  5.0
9   2.1 25.9  7.1  9.7 24.8 12.9 32.1 27.1
10 24.1 36.5 19.3 17.1  5.6  9.7 42.7 37.7 22.6
11 33.8 44.0 29.0 26.8 13.1 19.4 50.2 45.2 32.3  9.7
12 21.1 32.5 16.3 14.1  5.2  6.7 38.7 33.7 19.6  4.0 12.7
13 12.5 11.3  7.5 24.3 24.6 19.9 17.5 12.5 14.6 25.2 32.7 21.2
14 12.4 11.4  7.4 24.2 19.5 19.8 17.6 12.6 14.5 25.1 32.6 21.1  7.1
15 13.3 10.5  8.3 25.1 20.4 20.7 16.7 11.7 15.4 26.0 33.5 22.0  7.8  0.9
16  6.8 22.6  3.8 13.0 19.5  8.6 28.8 23.8  5.3 17.3 27.0 14.3 11.3 11.2 12.1
>
```

군집화 방법

□ 군집화 방법

- 중복이 없는 경우

- ✓ 비계층적인 방법(non-hierarchical cluster) ⇒ **K-means Clustering**

- 자료의 크기가 제한이 없는 경우 활용하는 방법
- 구하고자 하는 군집의 수를 정한 상태에서 설정된 군집의 중심에 가장 가까운 개체를 하나씩 포함해 가는 방식으로 군집을 형성하는 방법
- 많은 자료를 빠르고 쉽게 분류할 수 있으나 군집의 수를 미리 정해 주어야 하고, 군집을 형성하기 위한 **임의의 초기값**에 따라 군집결과가 달라지는 단점이 있음



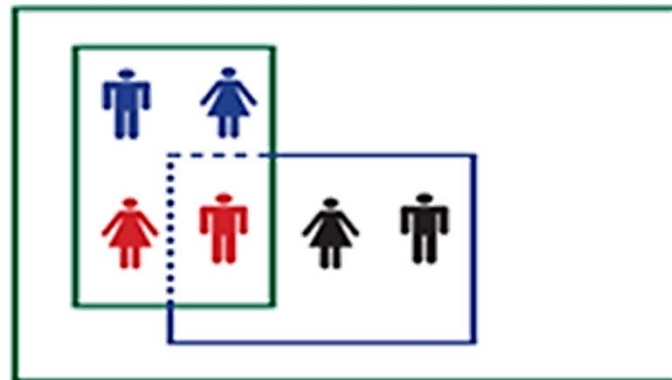
군집화 방법

□ 군집화 방법

- 중복이 있는 경우

- ✓ 중복군집분석 ⇒ 프림(PRIM : patient rule induction methods)

- 몇 개의 군집화 규칙을 상이하게 적용하며 군집화하는 방법으로 하나의 개체가 여러 군집에 동시에 포함될 수 있음



군집화 방법

□ 계층적 군집의 방법

계층적 군집화 알고리즘

1단계) 각 개체가 군집인 상태로 시작
(첫 단계의 군집수는 n)

2단계) 다양한 군집화 기준에 따라
가장 유사한 두 개체 군집을 합병
→ 군집 수가 줄어듦 (군집수 $n-1$)

3단계) 단계2에서 형성된 군집 중
가장 유사한 두 군집을 합병
→ 군집 수가 줄어듦

4단계) 최종적으로 모든 개체가
단일 군집으로 묶여짐 (군집수 1)

어떤 계층적 군집화 기준을
선정할 것인가?

몇 개의 군집으로 분류할
것인가?

군집화 방법

□ 계층적 군집의 방법

- 최단 연결법(Nearest-Neighbor Method) [단일연결법:Single Linkage Method] (R에서 hclust 의 method="single"사용)
 - 각 군집에 속하는 개체 사이의 거리 중 최단 거리 기준
 - 속도가 빠르고, 트리 구조가 변하지 않기 때문에 순서적 의미를 갖는 자료에 적합
 - 고립된 군집을 찾는데 유용
 - 그러나, 몇 개의 개체에 의해 군집이 고리로 연결될 수가 있고 이들은 실제로는 다른 군집이지만 다음 단계에서는 보다 큰 하나의 군집으로 묶일 수 있음
 - 즉, 두 군집 사이에 어중간하게 속해 있는 개체 때문에 마치 전체적으로 보면 두 군집이 하나인 것처럼 연결된 것으로 인식

군집화 방법

최단연결법(single linkage : nearest neighbor)에
의한 두 군집 C_1 과 C_2 의 거리는

$$d\{C_1, C_2\} = \min\{d(x, y) | x \in C_1, y \in C_2\}$$

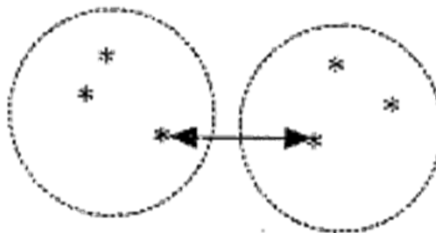
로 두 군집간의 최단 거리를 군집간 거리로 정의한다.

최단연결법으로 군집화하는 방법은 거리행렬 $D = \{d_{ik}\}$ 로부터 최단 거리의 쌍 U, V 를 찾아 한 군집으로 병합하는 것이다.

(UV) 로 묶인 군집과 군집 W 와의 거리는

$$d\{(UV)W\} = \min\{d_{UW}, d_{VW}\}$$

로 구한다.



군집화 방법

□ 계층적 군집의 방법

– 최단 연결법(Nearest-Neighbor Method) 예

5개 개체의 거리행렬 D로부터 군집 형성

$$D = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{matrix}$$

- 1) 개체 개수만큼의 군집수부터 시작하여 가장 거리가 가까운 쌍을 찾는다

$$\min\{d_{ik}\} = d_{53} = 2$$

5번 개체와 3번 개체를 묶어 군집(35)로 한다

군집화 방법

2) 군집(35)와 나머지 개체 1,2,4와의 거리 계산

$$d\{(35)1\} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$$

$$d\{(35)2\} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$$

$$d\{(35)4\} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$$

4개 군집간의 거리행렬

$$D_1 = d_{ik} = \begin{matrix} & \begin{matrix} (35) & 1 & 2 & 4 \end{matrix} \\ \begin{matrix} (35) & 1 & 2 & 4 \end{matrix} & \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{matrix}$$

(35)와 (1)이 가장 가까우므로 같은 군집 (1(35))로 묶임

군집화 방법

3) 군집(1(35))와 나머지 개체 2,4와의 거리 계산

$$\begin{aligned} d\{(135)2\} &= \min\{d_{12}, d_{(35)2}\} = \min\{9, 7\} = 7 \\ d\{(135)4\} &= \min\{d_{14}, d_{(35)4}\} = \min\{6, 8\} = 6 \end{aligned} \quad d\{(2)(4)\} = 5$$

3개 군집간의 거리행렬

$$D_2 = d_{ik} = \begin{matrix} & \begin{matrix} 1(35) & 2 & 4 \end{matrix} \\ \begin{matrix} 1(35) & 2 & 4 \end{matrix} & \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{matrix}$$

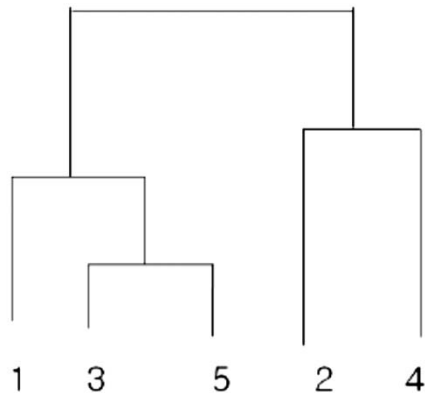
개체 2와 4가 가장 가까우므로 군집(24)로 묶임

군집화 방법

4) 군집(1(35))와 (24)와의 거리 계산

$$d\{(135)(24)\} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7, 6\} = 6$$

5) 나무구조그림



2개 군집 : (135)와 (24)

3개 군집 : (135), (2) 와 (4)

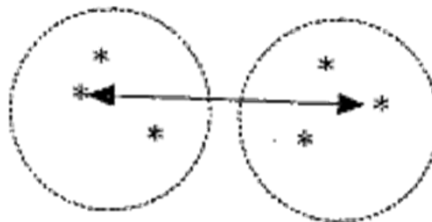
4개 군집 : (1), (35), (2)와 (4)

군집화 방법

□ 계층적 군집의 방법

- 최장 연결법(Furthest-Neighbor Method), [완전연결법(Complete Linkage Method)] (R에서 hclust 의 method="complete"사용)
 - 군집에 속한 개체들의 최장거리를 사용하여 분석
 - 군집의 내부적 응집성에 중점을 두는 방법
 - 가장 바깥에 퍼져있는 점들을 기준으로 하여 다른 군집(A,B)에 속하지 않는 것들은 바로 이 군집(C)에 속한다는 식으로 분류하는 방법

$$d\{C_1, C_2\} = \max\{d(x, y) | x \in C_1, y \in C_2\}$$



군집화 방법

– 최장 연결법(Furthest-Neighbor Method)의 예

5개 개체의 거리행렬 D로부터 군집 형성

$$D = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{matrix}$$

- 1) 개체 개수만큼의 군집수부터 시작하여 가장 거리가 가까운 쌍을 찾는다

$$\min \{d_{ik}\} = d_{53} = 2$$

5번 개체와 3번 개체를 묶어 군집(35)로 한다

군집화 방법

2) 군집(35)와 나머지 개체 1, 2, 4와의 거리 계산

$$d\{(35)1\} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11$$

$$d\{(35)2\} = \max\{d_{32}, d_{52}\} = \max\{7, 10\} = 10$$

$$d\{(35)4\} = \max\{d_{34}, d_{54}\} = \max\{9, 8\} = 9$$

$$d_{12} = 9 \quad d_{14} = 6 \quad d_{24} = 5$$

4개 군집간의 거리행렬

$$D_1 = \begin{matrix} & \begin{matrix} (35) & 1 & 2 & 4 \end{matrix} \\ \begin{matrix} (35) & 1 & 2 & 4 \end{matrix} & \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{bmatrix} \end{matrix}$$

(2)와 (4)가 가장 가까우므로 같은 군집 (24) 로 묶임

군집화 방법

3) 군집(35)와 군집(24)와 나머지 개체 1과의 거리 계산

$$d\{(35)1\} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11$$

$$d\{(35)(24)\} = \max\{d_{(35)2}, d_{(35)4}\} = \max\{10, 9\} = 10$$

$$d\{1(24)\} = \max\{d_{12}, d_{14}\} = \max\{9, 6\} = 9$$

3개 군집간의 거리행렬

$$D_2 = \begin{matrix} & \begin{matrix} (35) & 1 & (24) \end{matrix} \\ \begin{matrix} (35) & 1 & (24) \end{matrix} & \begin{bmatrix} 0 & & \\ 11 & 0 & \\ 10 & 9 & 0 \end{bmatrix} \end{matrix}$$

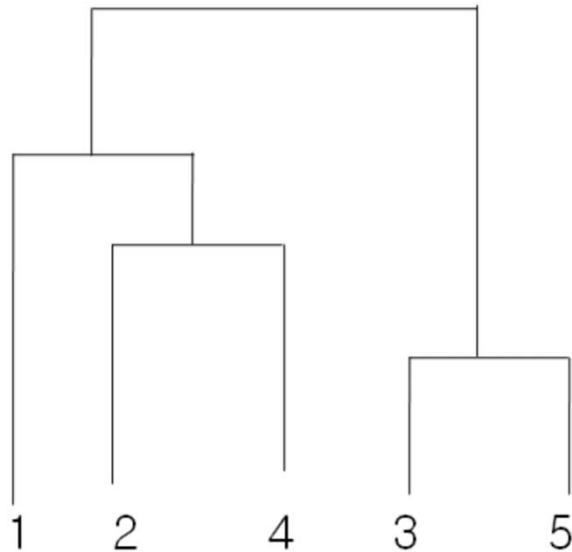
개체 1과 군집(24)가 가장 가까우므로 군집(1(24))로 묶임

군집화 방법

4) 군집(35)와 (1(24))와의 거리 계산

$$d\{(35)(1(24))\} = \max\{d_{(35)1}, d_{(35)(24)}\} = \max\{11, 10\} = 11$$

5) 나무구조그림



2개 군집 : (124)와 (35)

3개 군집 : (1), (24) 와 (35)

4개 군집 : (1), (2), (4)와 (35)

군집화 방법

□ 계층적 군집의 방법

– 중심 연결법(Centroid Linkage Method)

- 군집들 사이의 유클리디안 중심거리를 사용
- 두 군집 사이의 거리는 두 군집의 중심간 거리로 계산된다

$$d(G_1, G_2) = \| \overline{x_1} - \overline{x_2} \|^2$$

- 두 군집이 결합되면 새로운 군집의 중심은 가중평균을 이용

$$\overline{x} = \frac{n_1 \overline{x_1} + n_2 \overline{x_2}}{n_1 + n_2}$$

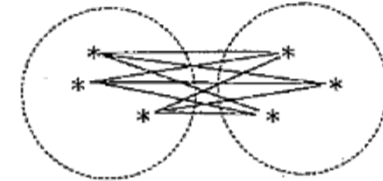
- 각 군집 사이의 거리를 구한 후 중심거리가 가장 가까운 데이터와 다시 새로운 군집을 형성

군집화 방법

□ 계층적 군집의 방법

– 평균 연결법(Average Linkage Method)

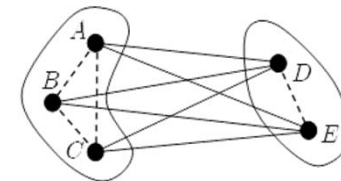
- 각 군집의 개체 대 개체간 거리를 평균함
- 모든 가능한 쌍이 다 계산되므로 시간이 오래 걸림



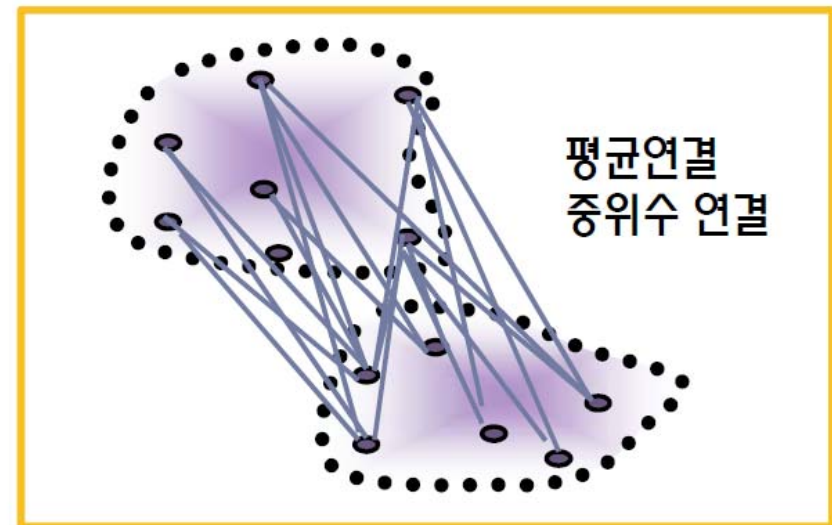
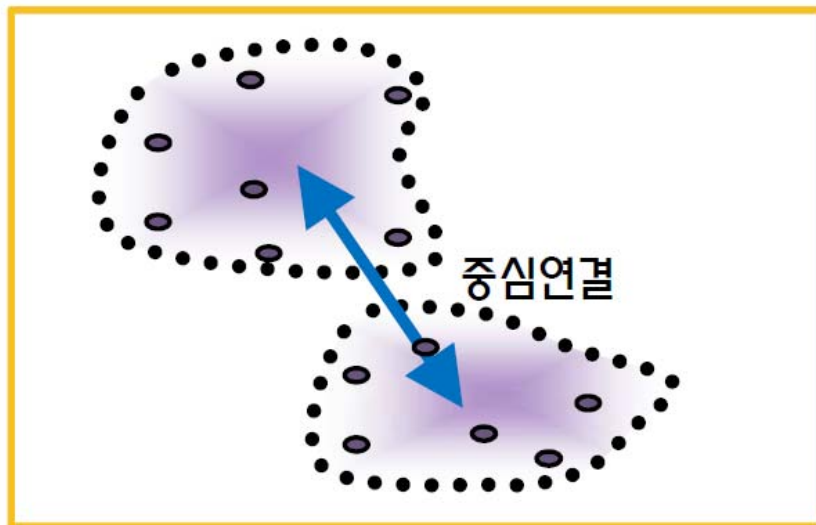
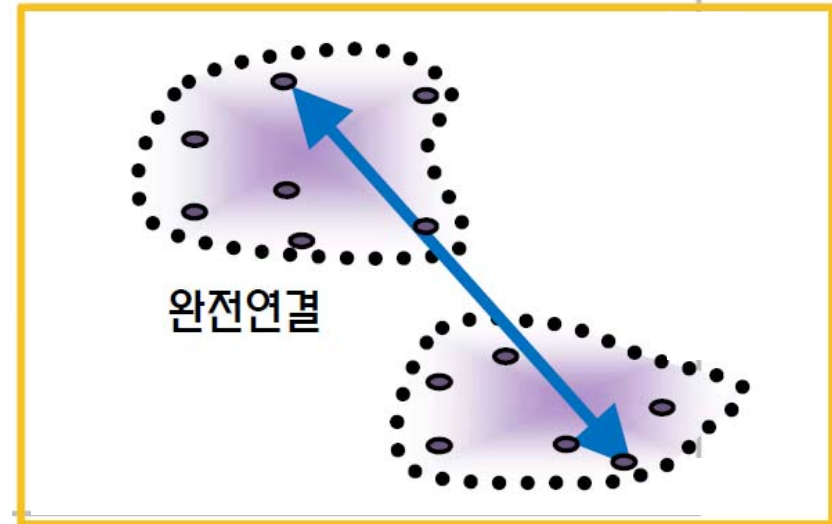
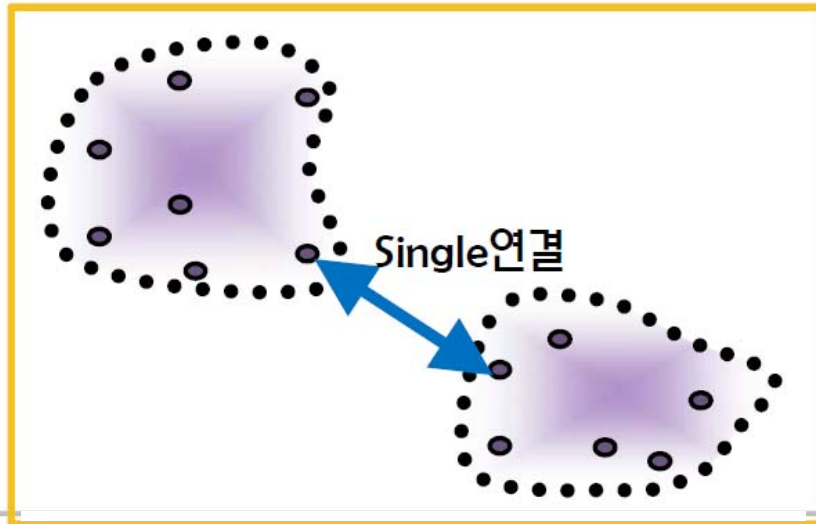
$$d\{C_1, C_2\} = \frac{1}{n_1 n_2} \sum_i \sum_j d_{ij}$$

n_1 은 군집 C_1 의 개체 수
 n_2 은 군집 C_2 의 개체 수

- ✓ 군집 간 평균연결법: 군집간 각 개체간의 거리를 평균하여 이 평균거리가 가장 가까운 집단을 연결하는 방식
- ✓ 군집 내 평균연결법 : 군집 간 평균방식의 변형으로 다른 군집에 있는 개체간의 거리 뿐만 아니라 같은 집단에 속한 개체간의 거리도 포함하여 평균을 구하는 방식



군집화 방법



군집화 방법

– 평균 연결법(Average Linkage Method)의 예

5개 개체의 거리행렬 D로부터 군집 형성

$$D = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{matrix}$$

- 1) 개체 개수만큼의 군집수부터 시작하여 가장 거리가 가까운 쌍을 찾는다

$$\min \{d_{ik}\} = d_{53} = 2$$

5번 개체와 3번 개체를 묶어 군집(35)로 한다

군집화 방법

2) 군집(35)와 나머지 개체 1, 2, 4와의 거리 계산

$$d\{(35), 1\} = \frac{1}{2 \cdot 1} (d_{31} + d_{51}) = \frac{1}{2} (3 + 11) = 7$$

$$d\{(35), 2\} = \frac{1}{2 \cdot 1} (d_{32} + d_{52}) = \frac{1}{2} (7 + 10) = 8.5$$

$$d\{(35), 4\} = \frac{1}{2 \cdot 1} (d_{34} + d_{54}) = \frac{1}{2} (9 + 8) = 8.5$$

$$d_{12} = 9$$

$$d_{14} = 6$$

$$d_{24} = 5$$

4개 군집간의 거리행렬

$$D_1 = \begin{matrix} & \begin{matrix} (35) & 1 & 2 & 4 \end{matrix} \\ \begin{matrix} (35) \\ 1 \\ 2 \\ 4 \end{matrix} & \begin{bmatrix} 0 & & & \\ 7 & 0 & & \\ 8.5 & 9 & 0 & \\ 8.5 & 6 & 5 & 0 \end{bmatrix} \end{matrix}$$

(2)와 (4)가 가장 가까우므로 같은 군집 (24) 로 묶임

군집화 방법

3) 군집(35)와 군집(24)와 나머지 개체 1과의 거리 계산

$$d\{(35), (24)\} = \frac{1}{2 \cdot 2} (d_{32} + d_{52} + d_{34} + d_{54}) = \frac{1}{4} (7 + 10 + 9 + 8) = 8.5$$

$$d\{(35), 1\} = 7$$

$$d\{(24), 1\} = \frac{1}{2 \cdot 1} (d_{21} + d_{41}) = \frac{1}{2} (9 + 6) = 7.5$$

3개 군집간의 거리행렬

$$D_2 = \begin{matrix} & \begin{matrix} (35) & 1 & (24) \end{matrix} \\ \begin{matrix} (35) \\ 1 \\ (24) \end{matrix} & \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 8.5 & 7.5 & 0 \end{bmatrix} \end{matrix}$$

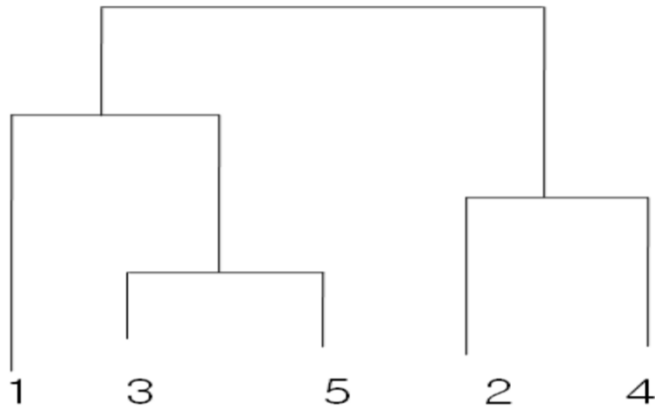
개체 1과 군집(35)가 가장 가까우므로 군집(1(35))로 묶임

군집화 방법

4) 군집(135)와 (24)와의 거리 계산

$$\begin{aligned} d\{(135)(24)\} &= \frac{1}{3 \cdot 2} (d_{12} + d_{14} + d_{32} + d_{34} + d_{52} + d_{54}) \\ &= \frac{1}{6} (9 + 6 + 7 + 9 + 10 + 8) = 8.17 \end{aligned}$$

5) 나무구조그림



2개 군집 : (135)와 (24)

3개 군집 : (1), (24) 와 (35)

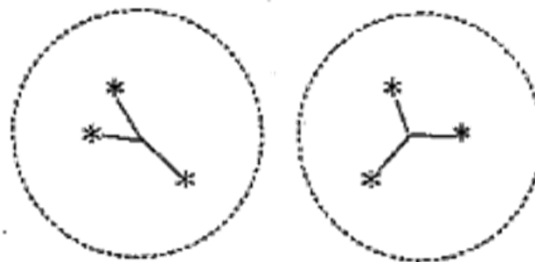
4개 군집 : (1), (2), (4)와 (35)

군집화 방법

□ 계층적 군집의 방법

– Ward의 방법(1963)

- 각 단계에서 군집을 묶음으로써 발생하는 정보의 손실을 측정
- 군집내 편차들의 제곱합(within group sum of squares)에 근거를 두고 군집들을 병합시키는데, 편차가 너무 커지면 군집으로 묶지 않음
- 크기가 작은 군집을 먼저 병합함으로써 비슷한 크기의 군집들을 유도하는 경향



군집화 방법

– Ward의 방법(1963)

- 군집내 제곱합 증분과 군집간 제곱합을 고려한 계층적 군집 방법
- 군집간 정보의 손실을 최소화하도록 군집화를 하는데 여기서 군집간의 정보란 편차제곱합 ESS (error sum of squares)로 나타낸다.
- 군집 A와 군집 B의 군집내 거리(within-cluster distance)는 군집내 제곱합.

$$ESS_A = \sum_{j=1}^{n_A} (\mathbf{X}_{Aj} - \overline{\mathbf{X}}_A)' (\mathbf{X}_{Aj} - \overline{\mathbf{X}}_A) = \sum_{j=1}^{n_A} \sum_{t=1}^p (X_{Ajt} - \overline{X}_{At})^2$$

$$ESS_B = \sum_{j=1}^{n_B} (\mathbf{X}_{Bj} - \overline{\mathbf{X}}_B)' (\mathbf{X}_{Bj} - \overline{\mathbf{X}}_B) = \sum_{j=1}^{n_B} \sum_{t=1}^p (X_{Bjt} - \overline{X}_{Bt})^2$$

여기서 $\overline{\mathbf{X}}_A$ 와 $\overline{\mathbf{X}}_B$ 는 각 군집에서의 평균관측값벡터

군집화 방법

- Ward의 방법(1963)
- 군집 A와 군집 B를 합친 경우 군집내 제곱합은 다음과 같다.

$$ESS_{AB} = \sum_{j=1}^{n_{AB}} (\mathbf{X}_{ABj} - \overline{\mathbf{X}}_{AB})' (\mathbf{X}_{ABj} - \overline{\mathbf{X}}_{AB}) = \sum_{j=1}^{n_{AB}} \sum_{t=1}^p (X_{ABjt} - \overline{X}_{ABt})^2$$

여기서

$$\overline{\mathbf{X}}_{AB} = \frac{n_A \overline{\mathbf{X}}_A + n_B \overline{\mathbf{X}}_B}{n_A + n_B}$$

은 합친 군집의 중심으로 군집 A와 군집 B간의 중심으로 표현된다.

군집화 방법

- Ward의 방법(1963) (R에서 `hclust` 의 `method="ward.D2"` 사용)
- 군집 A와 군집B는 군집형성으로부터 생기는 편차제곱합의 증가분을 최소화하도록 형성된다.

$$I_{AB} = ESS_{AB} - (ESS_A + ESS_B)$$

- 위의 식을 최소화하는 것은 군집간 거리(Between cluster distance)를 최소화하는 것과 같다.
- 즉, 군집 A와 군집B가 멀리 떨어져 있을수록 병합하면서 생기는 I_{AB} 가 크다. 또는 군집 A와 군집 B가 가까울수록 I_{AB} 가 작게 되어 정보의 손실이 작다.
- 개체와 군집 중심과의 편차제곱합 ESS가 작을수록 군집내 개체가 모여 있음을 알 수 있다.

군집화 방법

$$I_{AB} = ESS_{AB} - (ESS_A + ESS_B)$$

I_{AB} 를 다시 정리해보면

$$\begin{aligned} I_{AB} &= n_A (\overline{X}_A - \overline{X}_{AB})' (\overline{X}_A - \overline{X}_{AB}) + n_B (\overline{X}_B - \overline{X}_{AB})' (\overline{X}_B - \overline{X}_{AB}) \\ &= \frac{n_A n_B}{n_A + n_B} (\overline{X}_A - \overline{X}_B)' (\overline{X}_A - \overline{X}_B) \\ &= \frac{(\overline{X}_A - \overline{X}_B)' (\overline{X}_A - \overline{X}_B)}{\frac{1}{n_A} + \frac{1}{n_B}} \end{aligned}$$

I_{AB} 를 최소화하는 것은 군집간 거리(between cluster distance)를 최소화하는 것과 같다.

군집A와 군집B가 가까울수록 I_{AB} 가 작게 되어 정보의 손실이 작게 된다

민코우스키 거리(Minkowski distance)행렬을 이용한 평균연결법.

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

Step 1. 거리 행렬 정의

```
> dist.minkowski<-  
dist(crime.data[,c(2:3)],method="minkowski")
```

Step 2. 군집화 방법 선택 후 계층적 군집분석 실시

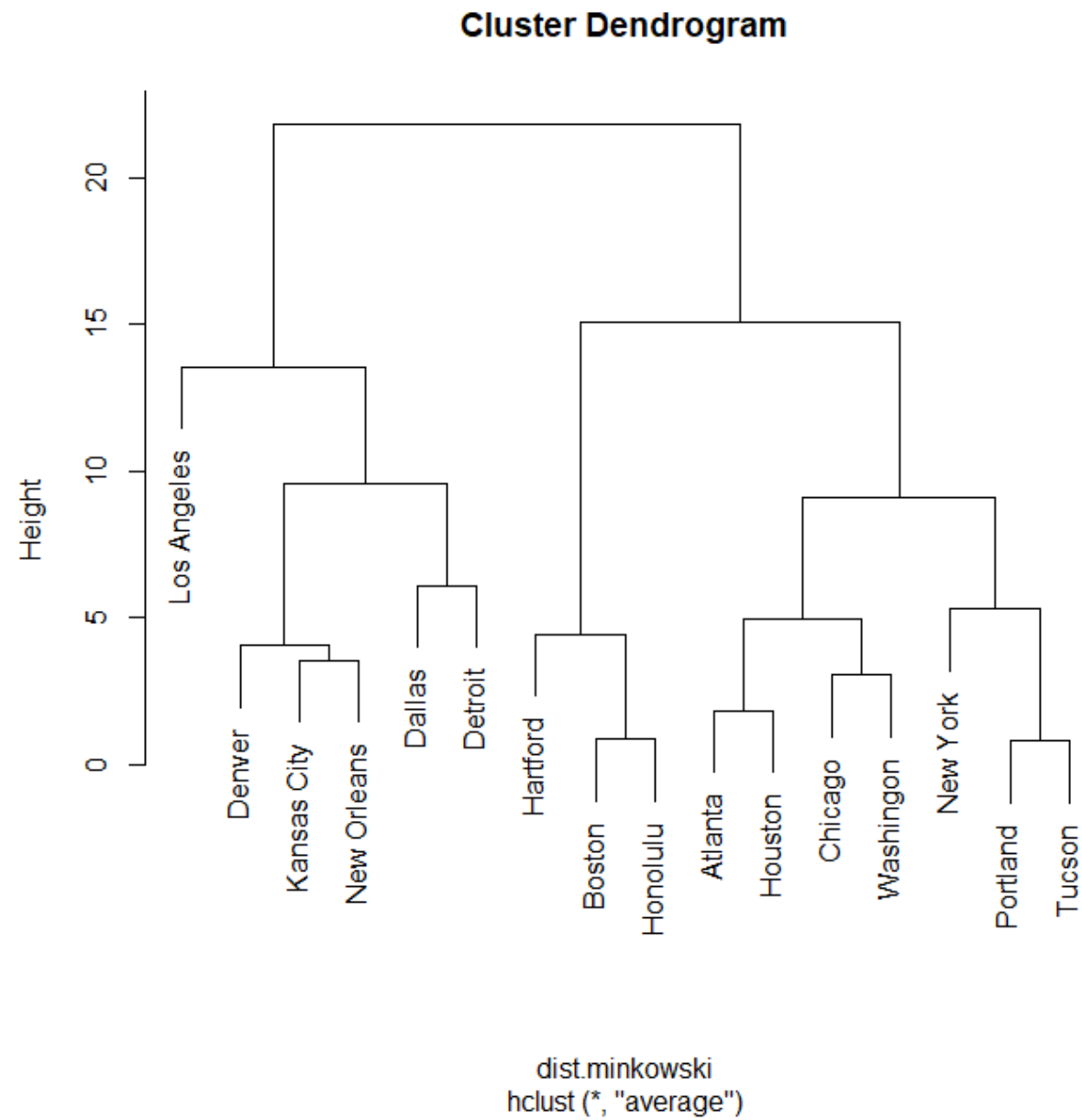
```
> minkowski.average<- hclust(dist.minkowski,method="average")
```

Step 3. 덴드로그램 그리기

```
> plot(minkowski.average,labels=crime.data[,1])
```

*** 여러가지 거리행렬과 여러가지 군집화 방법에 따른 군집화를 비교해본다.**

```
> plot(minkowski.average, labels=crime.data[,1])
```



군집화 방법

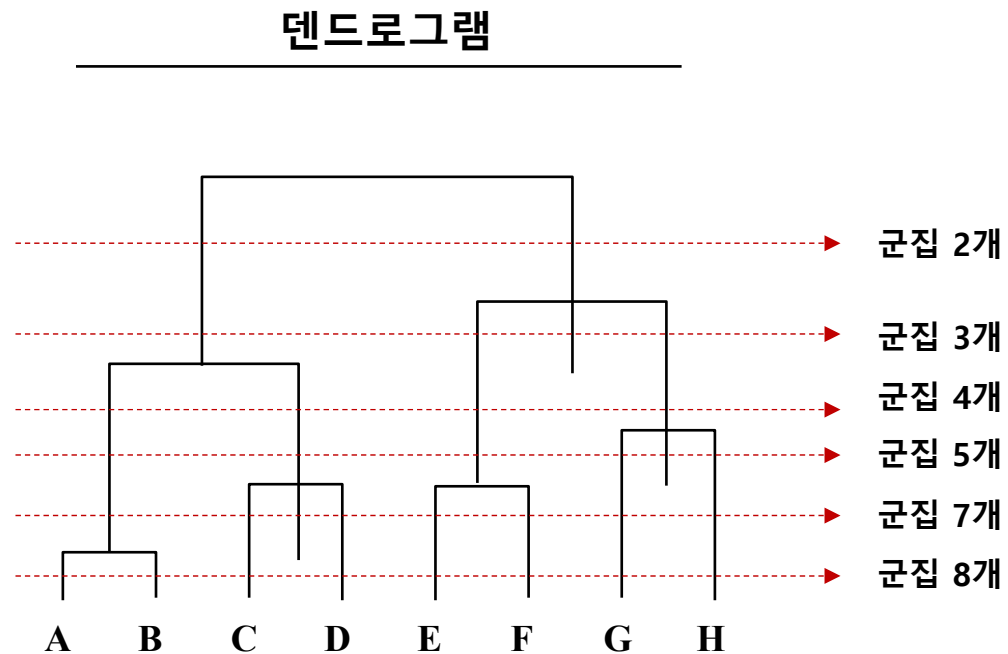
□ 계층적 군집의 방법 선택

- 같은 분산, 표본의 크기를 가진 구형 집단들에 대해서는 군집 간 평균연결법이나 Ward 방법을 사용
- 분산과 표본의 크기가 같지 않은 늘어난 군집에 대해서는 가장 가까운 항목 (최단연결법) 사용
- 분산과 표본의 크기가 같지 않은 구형 집단들에 대해서는 가장 먼 항목 (최장연결법)을 사용

군집화 방법_결과 표현

□ 덴드로그램 (Dendrogram, 나무형 그림)

군집화 결과를 표현하기에 효과적인 방법으로 덴드로그램을 사용함



→ 최종적으로 (A, B, C, D) 그룹과 (E, F, G, H) 그룹으로 나뉘짐

군집화 방법

□ 비계층적 군집의 방법

- 비계층적 군집화 중 대표적인 것은 K-평균 군집화이며, k는 군집 수로써

미리 지정

K-평균 군집화 알고리즘

1단계) k개의 각 군집에 1개씩의 개체 넣기
(initial seeding)

2단계) 모든 개체를 각각 가장 가까운
군집중심을 찾아 배속

3단계) 군집중심을 새로 계산

4단계) 변화가 없을 때까지 단계2/단계3 반복

군집수 k를 얼마로 지정할 것인가?

K개의 각 군집에 어떤 개체를
1개씩 넣을 것인가?

Initial seeding에 따라 결과가
크게 좌우됨

➔ 계층적 군집화의 결과 이용

군집화 방법 □ 비계층적 군집의 방법

– K-Means 방법

• 특징

- ✓ 각 개체를 상호배반적인 K개의 군집을 형성
- ✓ 초기에 부적절한 병합(분리)이 일어났을 때 회복 가능
- ✓ 군집의 수 K를 사전에 정의
- ✓ 대용량 자료의 경우 유용

• 초기 군집수

- ✓ 사전에 아무 정보 없이 전체 데이터에 알맞은 군집수 찾기는 어려움
- ✓ 군집수 k에 따른 여러 번의 군집분석 수행하여 종합적으로 결과를 판단하여 최종 모형 선택

• 군집 수 결정 방법

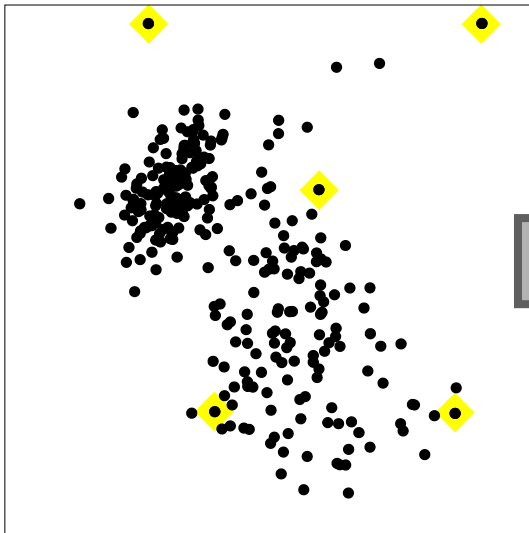
- ✓ 적절한 다변량 통계분석법을 이용
- ✓ 관찰치의 위치를 시각화해 이를 관찰
 - 주성분분석으로 차원 축소하고 2차원 혹은 3차원 그래프로 판단
- ✓ 계보적 군집분석의 결과로 얻어지는 여러 통계량의 변화 관찰하여 결정

군집화 방법

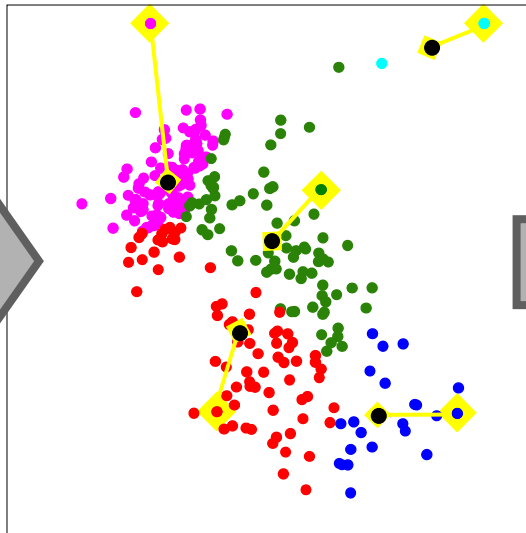
□ 비계층적 군집의 방법

– K-Means 방법

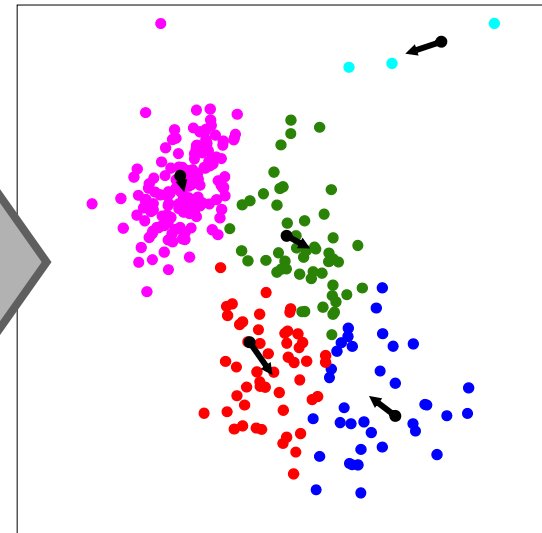
군집의 수 K 결정 : $K=5$
최초 군집기준값 결정



개체의 할당
군집중심 재 산출



개체의 할당
군집중심 재 산출



군집화 방법

□ 비계층적 군집의 방법

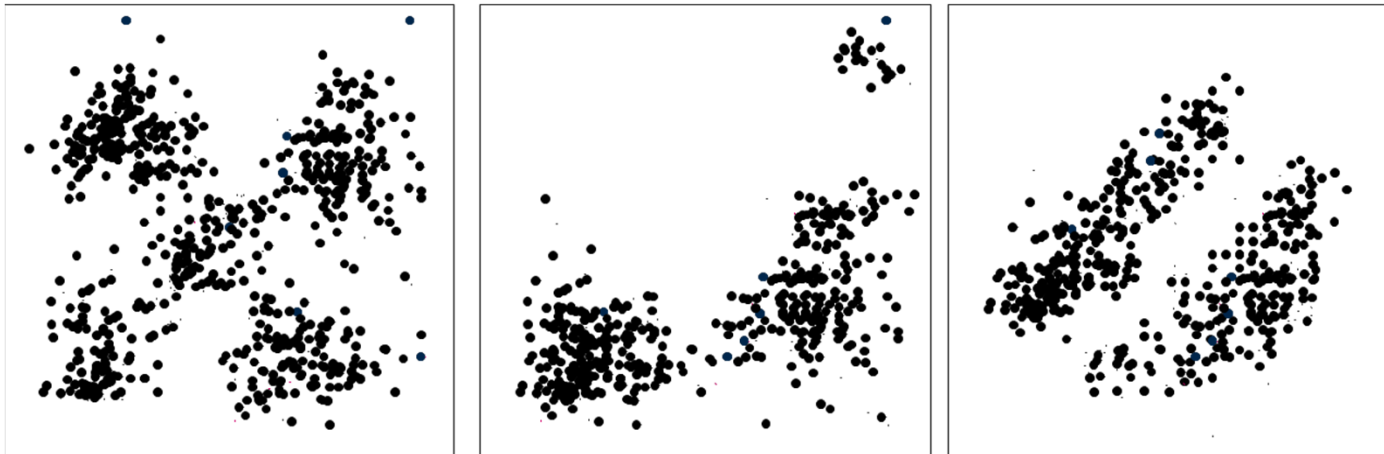
– K-Means 방법

• 주의점

✓ 군집 수 K 의 사전 결정

✓ 초기 군집중심의 설정

✓자료가 내포한 특이한 군집구조



K-mean clustering in R

```
> kmeans.two<-
kmeans(crime.data[,2:3],centers=2,iter.max=10,nstart=2)
> plot(crime.data[,2:3],col=kmeans.two$cluster)
> points(kmeans.two$centers,col=1:2,pch=9)
```

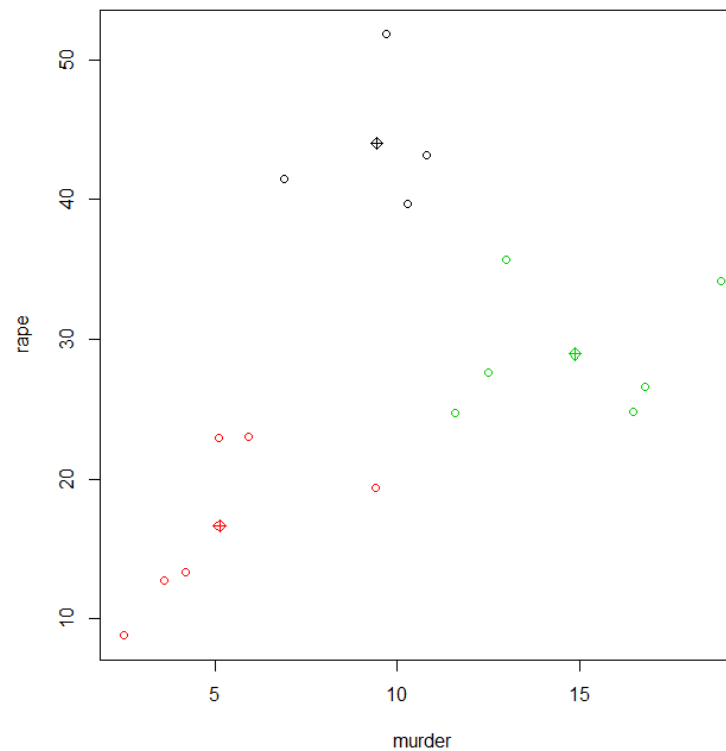
nstart=2 # random initial value의 시도 횟수

```
> kmeans.three<-
kmeans(crime.data[,2:3],centers=3,iter.max=10,nstart=2)
> plot(crime.data[,2:3],col=kmeans.three$cluster)
> points(kmeans.three$centers,col=1:3,pch=9)
```

```
> kmeans.three$iter # 몇번만에 알고리즘이 멈췄는지 보여줌.
[1] 2
> kmeans.three$cluster # 16개의 샘플이 어떻게 묶여있는지 보여줌.
[1] 3 2 3 3 1 3 2 2 3 1 1 1 2 2 2 3
> kmeans.three$centers # 각 군집의 평균값을 변수별로 출력.
      murder      rape
1  9.425000 44.05000
2  5.116667 16.68333
3 14.883333 28.93333
```

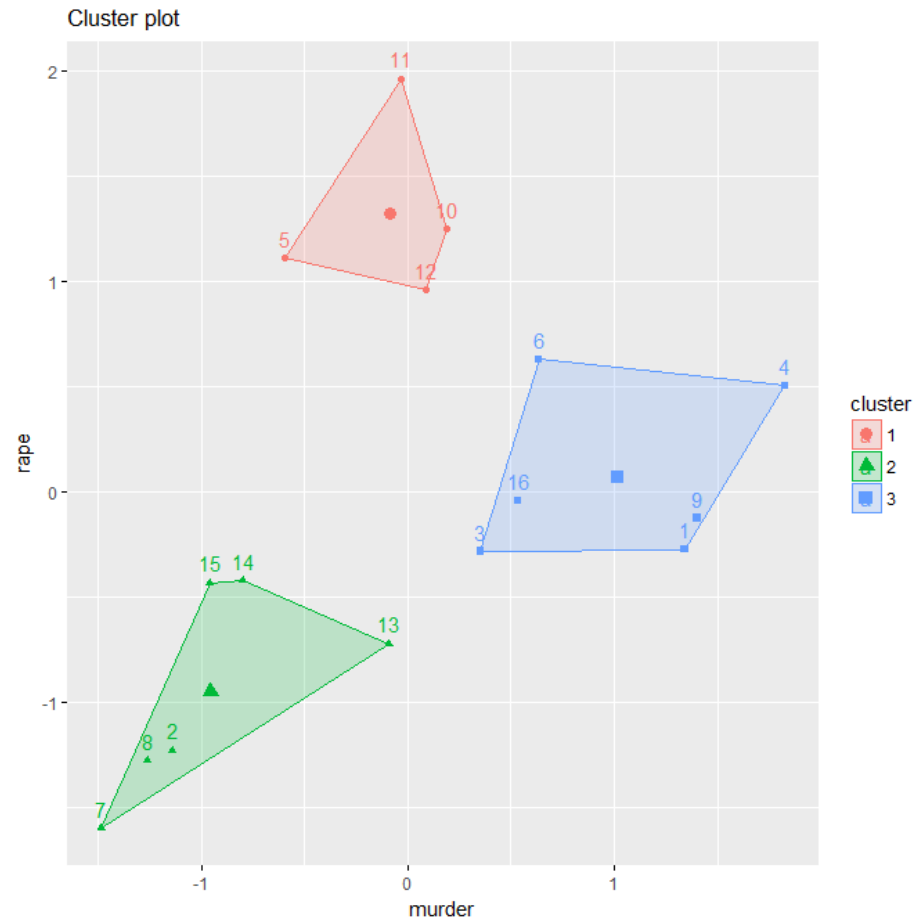
K-mean clustering in R

```
> kmeans.three<-  
kmeans(crime.data[,2:3],centers=3,iter.max=10,nstart=2)  
> plot(crime.data[,2:3],col=kmeans.three$cluster)  
> points(kmeans.three$centers,col=1:3,pch=9)
```



K-mean clustering in R (Same results, different graphical illustration)

```
> require("factoextra")  
> kmeans.three<-  
kmeans(crime.data[,2:3],centers=3,iter.max=10,nstart=2)  
> fviz_cluster(kmeans.three,data=crime.data[,2:3])
```



군집분석시 유의 사항

□ 유의 사항

- 군집분석은 비슷한 속성을 가진 대상이나 변인을 집단으로 묶는다.
- 군집분석에서 나온 결과는 이론적 근거에 따라서 해석해야 한다.
 - 특정의 데이터에서 어떤 결과가 나왔다고 그것을 무조건 일반화하고자 하는 태도는 피해야 함
- 군집분석은 회귀분석과 달리 필요 없는 변인을 제거하는 기능이 없음
 - 그러므로 연구자가 판단하여 중요한 관련변인은 다 넣어야 하고 관계 없는 변인들은 제거해야 함
- 분석 방법을 여러 개 써서 일관된 결과가 나오면 어느 정도 신뢰할 수 있음
 - 주어진 자료를 임의로 반으로 나누어 분석해도 비슷한 결과가 나오면 믿을 수 있음
- 척도에 따라 다른 결과가 나올 수 있음
 - 자료를 표준화시켜서 사용해야 함