12장 군집분석(cluster analysis)

- 군집분석은 개체들의 특징에 따라 서로 비슷한 것끼리 몇 개의 군집으로 묶는 다변량 기법
- 자료 특징: 훈련자료에 목적그룹 (target variable)이 없다.
  cf : 판별분석: 훈련자료에 목적그룹이 있다.
- 자율학습(unsupvervised learning) vs. 지도학습(supervised)
- 군집분석 : 그룹 개수 정해져 있지 않다.

- 군집분석 기원 : Driver and Kroeber (1932), Tryon (1939) 기원 : Wikipedia
- 생물학자 Sokal & Sneath의 'Principle of Numerical Taxonomy'는 군집화 방법 개발 자극. 생물학적 분류를 위해 생명체에 대한 유사성

을 측정하여 유사성이 큰 것들은 동일한 군집을 형성하며, 군집의 패턴이 인식된 후에는 새로운 개체를 패턴인식을 통해 분류할 수 있다고 가정하였다. (김재희, 다변량통계)

- 사회과학분야 : 데이터에 근거한 인류학적 분류, 심리학분야에서 심리시험결과에 의거한 집단분류, 사회학에서 사회경제활동지표를 근거로 한 계급 분류

- 예 : 자동차 브랜드를 몇 개의 변수(차 비용, 연비, 승차감, 안정성, 가속성)를 수집하여 자동차를 몇 개의 군집으로 분류하여 마케팅 대상의 범주를 정한다.
- 예 : 고객집단별 인구통계학적, 사회적, 행태적 특성파악을 통하여 집단의 프로필 작성 및 특성파악 (고객의 세분화를 통한 DM, 맞춤형 고객관리)

- 군집분석은 탐색적 데이터 마이닝의 중요 처리 기법 중의 하나
- 통계자료처리, 기계학습, 패턴인식, 영상분석, 바이오정보학에서 널리 사용하고 있음

- 군집분석 목적

1. 개체를 특징이 비슷한 개체들의 그룹으로 분리
2. 각 군집의 특성, 군집간의 차이 등에 관하여 통계적 처리

- 군집분석은 목표 그룹이 정해져 있지 않기 때문에 군집분석의 방법에 따라 결과가 차이가 날 수 있다. 따라서 평가의 기준을 정하는 것이 중요한 과제 가운데 하나이다.

12.1 유사성의 측도

- 유사성(similarity) 측도 : 관측된 벡터간의 거리(distance)
  거리가 가까울수록 유사성이 크고, 거리가 멀수록 유사성이 작다.

- 거리의 종류

$x=(x_1,\cdots,x_p), y=(y_1,\cdots,y_p)$

1. $\sqrt{(x-y)^T(x-y)}=\left\{\sum_{i=1}^{p}(x_i-y_i)^2\right\}^{1/2}$: 유클리디안 거리
2. $\sqrt{(x-y)^T S^{-1}(x-y)}$: 마하라노비스 거리 (변수들 공분산고려)

3. $\left\{\sum_{i=1}^{p}|x_i-y_i|^m\right\}^{1/m}$: 민코우스키 거리 (유클리디안 거리의 일반화)

4. $\sum_{i=1}^{p}|x_i-y_i|$: 맨하탄 (Manhattan) 거리 (Minkowski에서 $m=1$)

5. $\max_i|x_i-y_i|$: 체비세프 (Chebychev) 거리

6 binary 거리 : 성분의 이진으로 표현. 0보다 크면 1, 그렇지 않으면 0
x->(1,1,1) y->(1,0,1) x-y->(0,1,0) 평균: 1/3

7. $\sum_i \frac{|x_i-y_i|}{|x_i+y_i|}$: Canberra 거리

8. $\sqrt{(x-y)^T S_D (x-y)}=\left(\sum_{i=1}^{p}\frac{(x_i-y_i)^2}{s_i^2}\right)^{1/2}$: Karl Pearson 거리

(예) Euro.d : 1994 유럽연합에 가입한 나라의 GNP에 대한 농업의존도 (x1), GNP (x2)
첫 번째 열은 나라 이름으로 처리하기 위한 작업이 필요

```
> a=read.table("Euro.d",header=T)
> Euro.d=a[,2:3]
> rownames(Euro.d)=a[,1]
> head(Euro.d)
    farm   GNP
B    2.7  16.8
DK   5.7  21.3
D    3.5  18.7
```

```
> round(dist(Euro.d,method="euclidean",diag=F,upper=F,p=2),2)
          B     DK      D     GR      E      F    IRL      I      L     NL      P
DK     5.41
D      2.06   3.41
GR    22.34  22.57  22.66
E     18.03  17.31  17.95   5.96
F      3.45   3.51   2.66  20.10  15.30
IRL   12.75  13.31  13.08   9.60   5.92  10.56
I      5.80   5.47   5.42  17.38  12.53   2.77   7.92
L      4.28   2.22   2.30  24.04  19.00   4.06  14.57   6.66
NL     1.65   5.10   2.44  20.75  16.38   2.20  11.15   4.20   4.67
P     17.24  17.86  17.66   5.16   4.38  15.16   4.60  12.52  19.17  15.67
UK     2.83   8.05   4.85  21.49  17.79   5.30  12.10   6.72   7.10   3.12  16.32
```

빨간색은 최대값 : L , GR 최소값 : NL, B

- method : This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski".

```
> dist(Euro.d,method="maximum",diag=F,upper=F,p=2)
       B   DK    D   GR    E    F  IRL    I    L   NL    P
DK   4.5
D    1.9  2.6
GR  19.5 16.5 18.7
E   17.2 14.2 16.4  5.5
F    3.3  3.5  2.5 16.2 13.9
IRL 11.3 10.4 10.5  8.2  5.9  8.0
I    5.8  4.7  5.0 13.7 11.4  2.5  5.7
L    4.2  2.2  2.3 18.7 16.4  3.2 10.5  5.0
NL   1.6  4.9  2.3 17.9 15.6  1.7  9.7  4.2  4.6
P   14.7 13.5 13.9  4.8  3.6 11.4  3.4  8.9 13.9 13.1
UK   2.8  7.3  4.7 19.9 17.6  3.8 11.7  6.2  7.0  2.4 15.1
```

- 군집

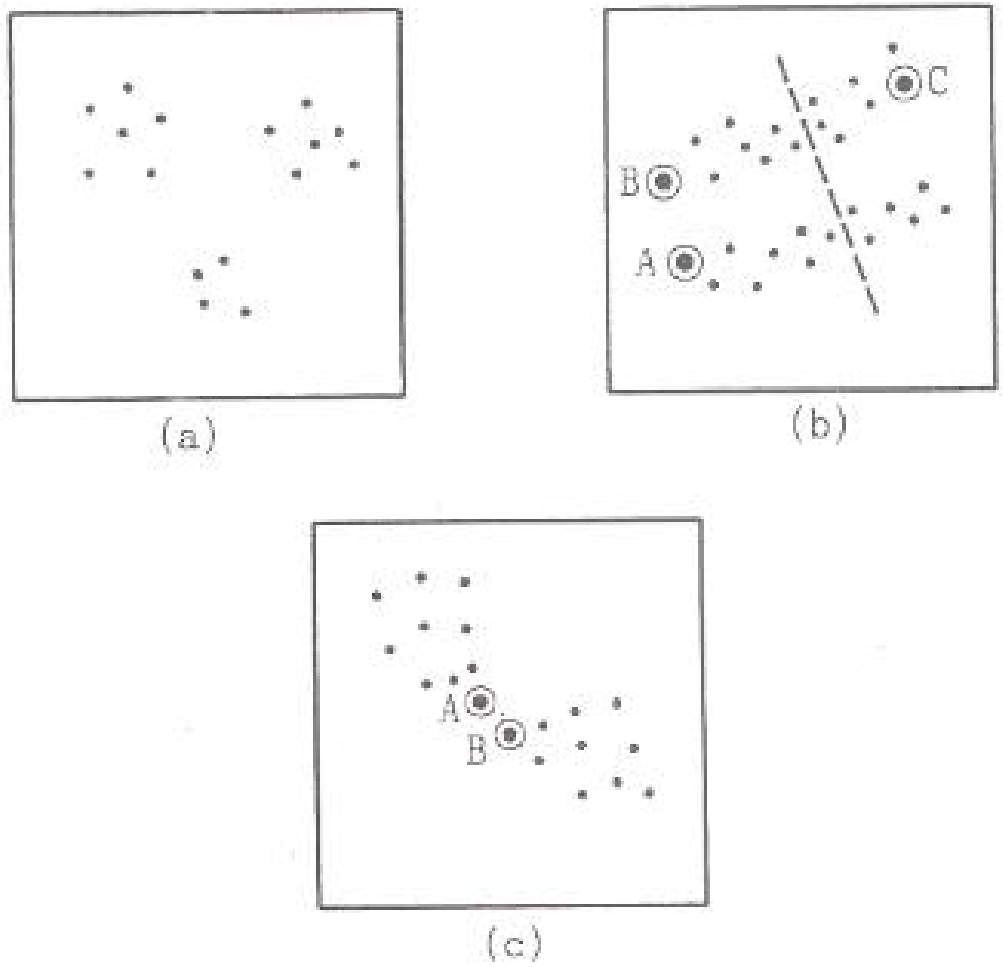
1. 계층적 군집(hierarchical clustering) : 계보적 군집
   군집의 수를 정하지 않고, 군집을 수를 2개로 분리하고, 기존의 구조를 유지한 채 하나의 그룹에서 다시 2개로 분리하여 총 3개의 군집을 구성하는 군집분석방법. 나뭇가지와 같은 군집분석 방법.
2. 분리군집 (partitioning clustering) : 상호배반적인 군집.
   각 개체는 서로 중복되지 않는 군집에 속한다.
3. 중복군집 (overlapping clustering) :
   각 개체는 두 개 이상의 군집에 소속된다.
4. 퍼지 군집 (fuzzy clustering) :
   각 개체가 어떤 군집에 속할 확률을 멤버쉽 함수로 표현

● 군집분석의 어려움

– 구형군집 (a) : 군집분석 방법들의 결과가 우수

– 긴 타원형군집 (b) : 개체들사이의 단순 유클리드 거리로 측정하면 잘 못된 결과를 보인다. 개체 A, B의 거리는 개체 A, C의 거리보다 작기 때문에 개체 A, B를 하나의 군집으로 처리한다.
Q : 해결방법은 무엇인가?

– 군집이 붙어있는 경우 : 개체 A, B가 두 군집 사이의 고리역할을 하여 군집방법에 따라서는 하나의 군집으로 결론지을 수 있다.



출처 : 다변량통계분석(김기영,전명식)

- 군집 : 군집 내에 속한 개체들의 특성은 동질적이고, 서로 다른 군집에 속한 개체들간의 특성은 서로 이질적이 되도록 군집을 형성해야.
  : Fisher의 판별분석 원리와 비슷

- 군집분석방법

1. 계층적 군집방법 : 병합법(agglomerative), 분리법(divisive)
2. 분리군집방법 : K-평균 군집

12.2 계층분류법 (hierarchical clustering method)

- 분할적 계층분류법(divisive) : 처음에는 두 개로 쪼개고, 두 개 중 한 개를 택하여 두 개로 쪼개면 세 개의 그룹이 되고, 세 개의 그룹 가운데 한 개를 택하여 두 개로 쪼개면 네 개의 그룹이 되고, ...
- 병합적 계층분류법(agglomoerative method) : 가장 가까운 두 개를 하나의 그룹으로 묶으면 처음 n 개의 그룹에서 (n−1)개의 그룹이 되고, (n−1)개의 그룹에서 가장 가까운 그룹을 두 개로 묶으면 (n−2)개의 그룹이 되고 ...
- 분할과 병합은 방향만 다를 뿐. 병합적 방법만 설명.

Q : 개체와 개체는 거리가 짧은 것이 유사성이 높다. 개체와 병합된 그룹과의 거리를 어떻게 측정할 것인가?

– 그룹 연결법 : 최단결합법(single), 최장결합법(complete), 군평균결합법(group average), 중심법(centroid), 중앙치법(median), Ward법 등.

12.2.1 최단결합법 (single linkage, nearest neighbor)

- 두 개의 그룹 A, B 사이의 거리 : $\min_{i \in A, j \in B} d_{ij}$
- 두 그룹사이의 거리 : 두 그룹의 모든 거리들 중에서 가장 짧은 것
- 예 : 5개의 개체가 있다. 두 개체들간의 거리는 다음과 같다.

표 12.1: 인공거리자료

	1	2	3	4	5
1	-				
2	1	-			
3	5	4	-		
4	9	8	3	-	
5	8	7	4	2	-

Step 0 : 두 개체들간의 거리 가운데 가장 짧은 것은 개체 1, 2이다.
따라서 (1,2)가 하나의 그룹이 된다.

Step 1 : {(1,2), 3, 4, 5} 4개의 개체들 사이의 거리를 구한다.

표 12.2: 인공거리자료의 최단결합법 1 단계

	(1,2)	3	4	5
(1,2)	-			
3	4	-		
4	8	3	-	
5	7	4	2	-

: 여기서 (1,2)와 3 사이의 거리 계산을 살펴보면 두 개체사이의 거리는 { d(1,3), d(2,3) } 으로 표 12.1에서 {5, 4}이다. 따라서 (1,2)와 3사이의 거리는 min{5,4}=4 이다. 표 12.2의 거리가 가장 짧은 것은 4, 5 이다. 4,5를 병합한다.

Step 2 : {(1,2), 3, (4,5)} 3개의 개체들 사이의 거리를 구한다.

표 12.3: 인공거리자료의 최단결합법 2 단계

	(1,2)	3	(4,5)
(1,2)	-		
3	4	-	
(4,5)	7	3	-

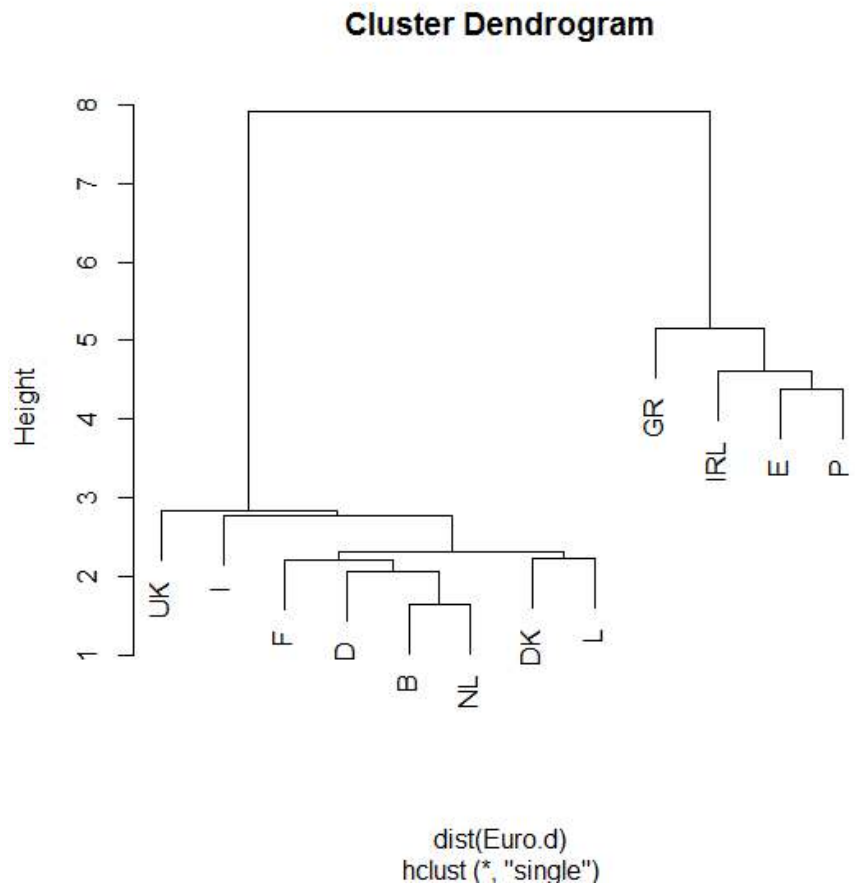
: 개체 (1,2), (4,5) 사이의 거리를 구하면 {d(1,4), d(1,5), d(2,4), d(2,5)}이고 값은 {9, 8, 8, 7} 이므로 두 개체사이의 거리는 7이다.

: 표 12.3에서 거리가 가장 짧은 3, (4,5)가 하나로 병합된다.

: {(1,2), (3,4,5)}로 병합된다.

Step 3 : 최종적으로 (1,2,3,4,5)로 병합된다.

```
> a1=hclust(dist(Euro.d),method="single")
> plot(a1)
```



12.2.2 최장결합법 (complete linkage, farthest neighbor)

- 두 개의 그룹 A, B 사이의 거리 : $\max_{i \in A, j \in B} d_{ij}$
- 두 그룹사이의 거리 : 두 그룹의 모든 거리들 중에서 가장 긴 것
- 예 : 5개의 개체가 있다. 두 개체들간의 거리는 다음과 같다.

표 12.1: 인공거리자료

	1	2	3	4	5
1	-				
2	1	-			
3	5	4	-		
4	9	8	3	-	
5	8	7	4	2	-

Step 0 : 두 개체들간의 거리 가운데 가장 짧은 것은 개체 1, 2이다.
따라서 (1,2)가 하나의 그룹이 된다.

Step 1 : {(1,2), 3, 4, 5} 4개의 개체들 사이의 거리를 구한다.

표 12.5: 인공거리자료의 최장결합법 1 단계

	(1,2)	3	4	5
(1,2)	-			
3	5	-		
4	9	3	-	
5	8	4	2	-

: 여기서 (1,2)와 3 사이의 거리 계산을 살펴보면 두 개체사이의 거리는 { d(1,3), d(2,3) } 으로 표 12.1에서 {5, 4}이다. 따라서 (1,2)와 3사이의 거리는 max{5,4}=5 이다. 표 12.5의 거리가 가장 짧은 것은 4, 5 이다. 4,5를 병합한다.

Step 2 : {(1,2), 3, (4,5)} 3개의 개체들 사이의 거리를 구한다.

표 12.6: 인공거리자료의 최장결합법 2 단계

	(1,2)	3	(4,5)
(1,2)	-		
3	5	-	
(4,5)	9	4	-

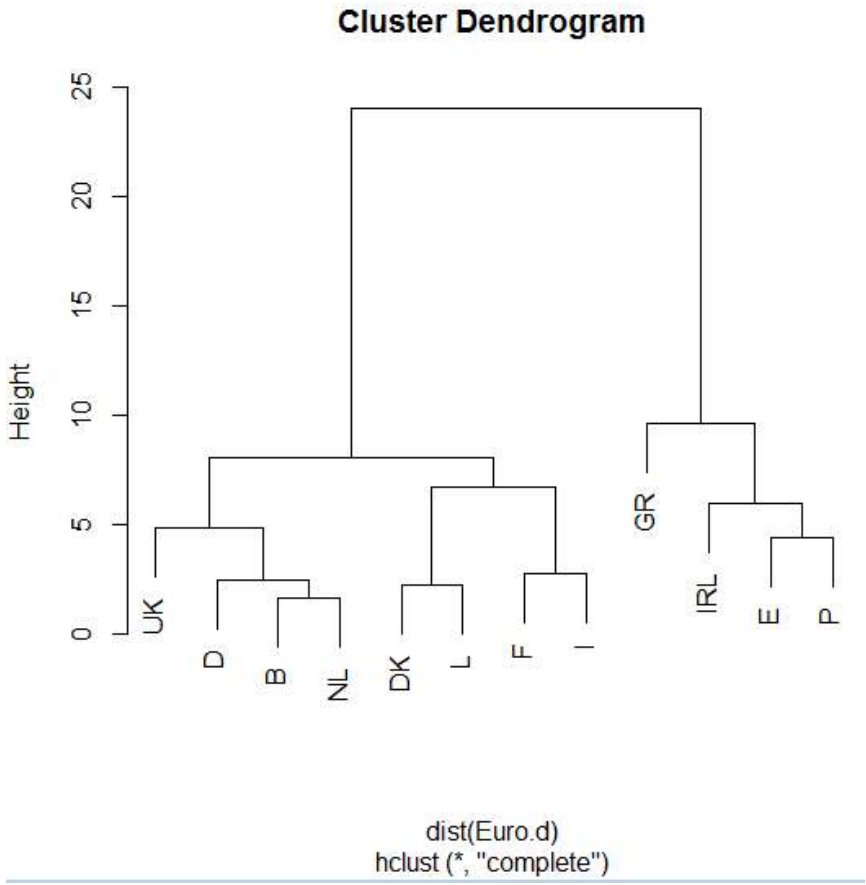
: 개체 (1,2), (4,5) 사이의 거리를 구하면 {d(1,4), d(1,5), d(2,4), d(2,5)}이고 값은 {9, 8, 8, 7} 이므로 두 개체사이의 거리는 9이다.

: 표 12.6에서 거리가 가장 짧은 3, (4,5)가 하나로 병합된다.

: {(1,2), (3,4,5)}로 병합된다.

Step 3 : 최종적으로 (1,2,3,4,5)로 병합된다.

```
> a2=hclust(dist(Euro.d),method="complete")
> plot(a2)
```



12.2.3 군평균결합법 (average)

- 두 개의 그룹 A, B 사이의 거리 : $\frac{1}{n_A n_B} \sum_{i \in A, j \in B} d_{ij}$
- 두 그룹사이의 거리 : 두 그룹의 모든 거리들의 평균
- 예 : 5개의 개체가 있다. 두 개체들간의 거리는 다음과 같다.

표 12.1: 인공거리자료

	1	2	3	4	5
1	-				
2	1	-			
3	5	4	-		
4	9	8	3	-	
5	8	7	4	2	-

Step 0 : 두 개체들간의 거리 가운데 가장 짧은 것은 개체 1, 2이다.
따라서 (1,2)가 하나의 그룹이 된다.

Step 1 : {(1,2), 3, 4, 5} 4개의 개체들 사이의 거리를 구한다.

표 12.7: 인공거리자료의 군평균결합법 1 단계

	(1,2)	3	4	5
(1,2)	-			
3	4.5	-		
4	8.5	3	-	
5	7.5	4	2	-

: 여기서 (1,2)와 3 사이의 거리 계산을 살펴보면 두 개체사이의 거리는 { d(1,3), d(2,3) } 으로 표 12.1에서 {5, 4}이다. 따라서 (1,2)와 3사이의 거리는 (5+4)/2=4.5 이다. 표 12.5의 거리가 가장 짧은 것은 4, 5 이다. 4,5를 병합한다.

Step 2 : {(1,2), 3, (4,5)} 3개의 개체들 사이의 거리를 구한다.

표 12.8: 인공거리자료의 군평균결합법 2 단계

	(1,2)	3	(4,5)
(1,2)	-		
3	4.5	-	
(4,5)	8.0	3.5	-

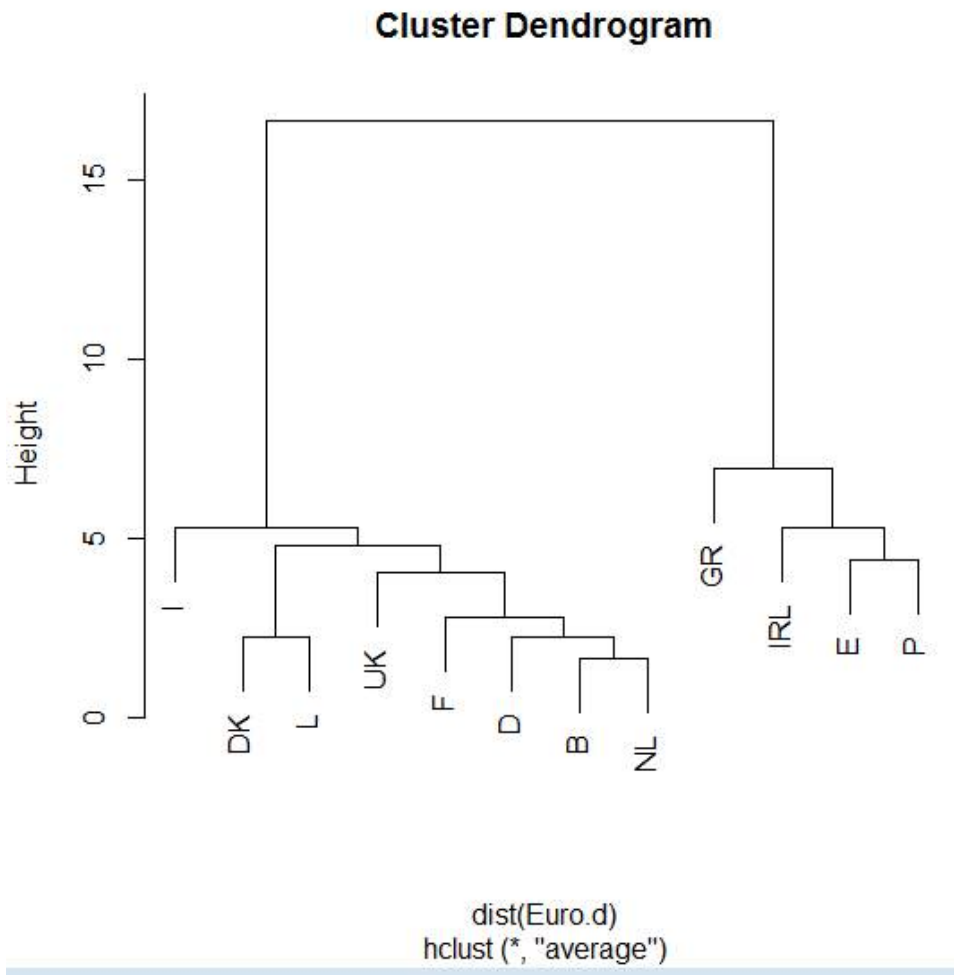
: 개체 (1,2), (4,5) 사이의 거리를 구하면 {d(1,4), d(1,5), d(2,4), d(2,5)}이고 값은 {9, 8, 8, 7} 이므로 두 개체사이의 거리는 (9+8+8+7)/4=8.0 이다.

: 표 12.6에서 거리가 가장 짧은 3, (4,5)가 하나로 병합된다.

: {(1,2), (3,4,5)}로 병합된다.

Step 3 : 최종적으로 (1,2,3,4,5)로 병합된다.

```
>plot(hclust(dist(Euro.d),method="average")) # dendrogram
```



12.2.4 중심법, 중앙치법, Ward법

- 중심법 : 각 그룹의 중심(평균)을 구하고, 그 중심사이의 거리

$$d(A,B) = \sqrt{(\overline{x_A} - \overline{x_B})^T(\overline{x_A} - \overline{x_B})}$$

  : 합쳐진 그룹의 중심은 $(n_A\overline{x_A} + n_B\overline{x_B})/(n_A + n_B)$

- 중앙치법 : 중심법과 다른 점은 합쳐진 그룹의 중심을

  : $(\overline{x_A} + \overline{x_B})/2$

- Ward법 : 군집내 제곱합 증분과 군집간 제곱합을 고려한다.
  - 군집 A와 군집 B의 군집내 거리(within-cluster distance)는

$$ESS_A = \sum_{j=1}^{n_A}(x_{Aj} - \overline{x_A})^T(x_{Aj} - \overline{x_A})$$

  - 합쳐진 군집 AB의 군집내 제곱합은

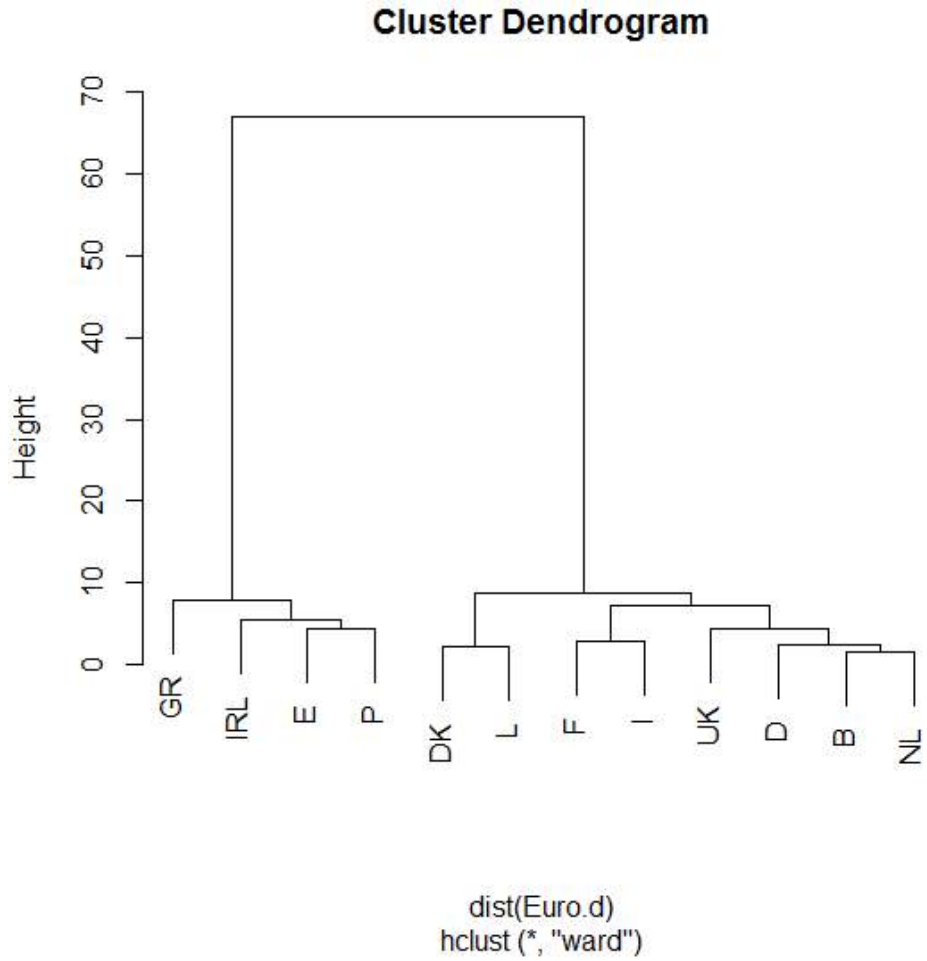
$$ESS_{AB} = \sum_{j=1}^{n_{AB}} (x_{ABj} - \overline{x_{AB}})^T (x_{ABj} - \overline{x_{AB}}),\ \overline{x_{AB}} = \frac{n_A \overline{x_A} + n_B \overline{x_B}}{n_A + n_B}$$

- 군집 A와 군집B가 형성하면서 생기는 편차제곱합의 증분 $I_{AB} = ESS_{AB} - (ESS_A + ESS_B)$ 를 최소화하도록 하는 군집분석

$$I_{AB} = (\overline{x_A} - \overline{x_B})^T (\overline{x_A} - \overline{x_B}) / (1/n_A + 1/n_B)$$

- I_{AB}를 최소화하는 것은 군집간의 거리를 최소화하는 것
- 군집 A, 군집 B가 멀리 떨어져 있을수록 병합하면서 생기는 I_{AB}가 크고, 가까울수록 I_{AB}가 작다.
- 개체와 군집 중심과의 편차제곱합 ESS가 작을수록 군집내 개체가 모여있다.

```
>plot(hclust(dist(Euro.d),method="ward")) # dendrogram
```



12.2.5 적합도

1. 코페네틱 상관계수: d_{rs}와 δ_{rs}의 상관계수.

(cophenetic correlation coefficient)

- d_{rs} : r, s번째 개체의 거리
- δ_{rs}: 개체 r,s가 처음으로 같은 그룹으로 분류될 때의 분류수준
- $\binom{n}{2}$개의 d_{rs},δ_{rs}사이의 상관계수

● 코페네틱 상관계수가 1에 가까울수록 적합한 군집분석방법이다.

(예) 인공자료에 대한 코페네틱 상관계수 (최단결합법)

$d_{12}=1, \delta_{12}=1$: 개체 1,2는 1단계에서 결합

$d_{14} = 9, \delta_{14} = 4$: 개체 1,4는 4단계에서 결합

표 12.9: 인공거리자료의 최단결합법에서 코페네틱 상관계수

(r,s)	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
d_{rs}	1	5	9	8	4	8	7	3	4	2
δ_{rs}	1	4	4	4	4	4	4	3	3	2

- 표 12.9의 두 변수 d_{rs}, δ_{rs}의 상관계수를 구하면 0.82 이다. 따라서 원래의 거리를 최단결합법에 의한 덴드로그램이 잘 표현했다.
- Euro.d의 코페네틱 상수

```
>Euro.dist=dist(Euro.d)
>cop2=cophenetic(hclust(Euro.dist,method="single"))
>cor(Euro.dist, cop2)
```

```
[1] 0.8927221
> cop2=cophenetic(hclust(Euro.dist,method="complete"))
> cor(Euro.dist,cop2)
[1] 0.9025177
```

따라서 최장결합법이 최단결합법보다 더 우수하다고 할 수 있다. 그런데 그림 12.2와 그림 12.4를 비교하면, 군집의 수를 2라고 할 때 두 군집방법은 결과의 차이가 없다.

2. Rousseeuw의 결합계수(agglomerative coefficient)

- $d(i)$를 i번째 관측치가 처음으로 그룹을 형성할 때의 거리를 최종적으로 하나의 그룹이 될 때의 거리로 나눈값.

(예) 인공자료에서 1번째 관측값의 처음 그룹을 형성할 때의 거리는 1이고, 최종적으로 하나의 그룹이 될 때의 거리는 (1,2), (3,4,5)의 거리는 4이므로 $d(1)=1/4$이다. 마찬가지로 $d(2)=1/4$이다.

관측치 4,5는 각각 그룹을 형성할 때의 최소값은 2이므로 $d(4)=d(5)=2/4$이다. 표 12.3에서 $d(3)=3/4$ 이다.

- 결합계수는 $\frac{1}{n}\sum_{i=1}^{n}(1-d(i))$ 이다.

(예) 결합계수는 $\frac{1}{5}(\frac{3}{4}+\frac{3}{4}+\frac{1}{4}+\frac{2}{4}+\frac{2}{4})=\frac{11}{20}=0.55$

```
>library(cluster)
> agnes(Euro.d,method="single")$ac
[1] 0.6198451
```

```
> agnes(Euro.d,method="complete")$ac
[1] 0.8444569
```

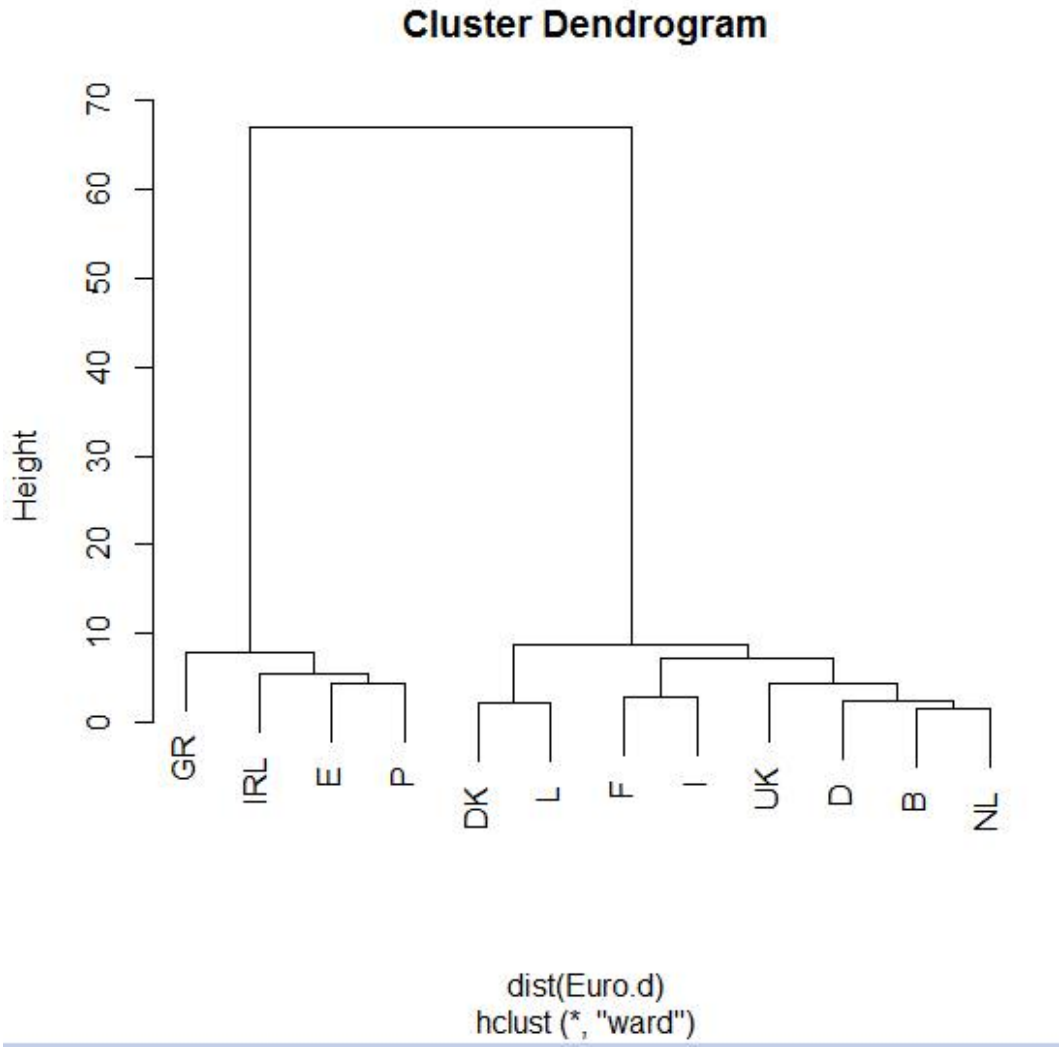
- 위의 두 결과 (코페네틱 상관계수, 결합계수)를 바탕으로 최단결합법과 최장결합법의 성능을 평가하면 다음과 같다.

방법	코페네틱 계수	결합계수
최단결합법	0.892	0.619
최장결합법	0.902	0.844
군평균법	0.907	0.784

따라서 코페네틱 계수는 3 방법이 거의 차이가 없고, 결합계수를 고찰하면 가장 좋은 군집분석 방법은 최장결합법이라고 할 수 있다.

12.2.6 그룹의 수

- 그룹의 개수 결정 문제 : 인자분석의 인자의 개수를 결정과 비슷
- 나무구조그림을 이용하여 병합되는 과정에서의 거리가 상대적으로 큰 변화를 보일 경우에 대한 그룹의 개수를 정한다.
- Ward의 방법을 이용하는 경우 그룹의 개수에 대한 ESS 증분을 검토하여 급격한 변화가 일어나는 위치에서 대응되는 그룹의 개수 (Scree plot과 유사)
- 데이터의 상황을 최적화시키는 그룹의 개수는 아직 방법이 없다고 해도 무방하다.


Cluster Dendrogram
Height
0
10
20
30
40
50
60
70
GR
IRL
E
P
DK
L
F
I
UK
D
B
NL
dist(Euro.d)
hclust (*, "ward")

12.3 비계층분류법 (non-hierarchical) : 분리군집방법

- 계층적 군집분석의 단점
  - 어떤 개체가 특정한 군집에 할당되면 다른 군집에 다시 할당될 수 없다.
  - 자료의 개수 n이 많아지면 그룹의 개수 g에 따라 계산량이 많다.
  - 덴드로그램으로 분석하기 어렵다.
- 분리군집법 : 미리 설정된 최적화에 근거하여 개체를 분리한다. 단, 최종 그룹의 개수는 미리 정해졌다고 가정한다.
  1. 초기 군집을 어떻게 설정할 것인가
  2. 개체를 어떤 기준에 의하여 군집에 할당할 것인가?
  3. 특정군집에 속하는 개체의 일부 또는 전체를 다른 군집에 어떤 기준에 의하여 재할당할 것인가?

12.3.1 K-평균법 (K means algorithm)

- c번째 군집평균 = $\overline{x_c}$,
- $i \in c$ 일 때, i번째 개체에서 c번째 군집평균까지 거리 제곱 :

  $d_{ic}^2 == (x_i - \overline{x_c})^T (x_i - \overline{x_c})$
- 각 개체를 c번째 군집에 재할당할 때 오차제곱합 :

  $E = \sum_i d_{ic}^2$
- E를 최소화하게 i번째 개체를 군집 c로 할당.
- 분리군집방법 : 각 개체를 어느 한 군집으로부터 다른 군집으로 재할당할 때 E를 계산하여 비교하면서 더 이상 움직일 개체가 없을 때까지 반복 수행한다.

● K-평균법의 절차

1. 자료를 K개 초기 군집으로 나눈다.
2. K개의 군집의 군집평균을 구한다.
3. 각 개체에서 각 군집의 군집평균까지의 거리를 계산하여 최소가 되게 하는 거리에 해당하는 군집으로 그 개체를 할당한다.
4. 3의 결과에 의해 구성된 새로운 K개의 군집을 구성한다.
5. 이전의 군집과 새로운 군집이 변화가 없을 때까지 2~4를 반복한다.

```
>kmeans(Euro.d, 2)
K-means clustering with 2 clusters of sizes 4, 8

Cluster means:
        farm      GNP
```

```
1 18.3750   9.000
2  4.5625 17.825

Clustering vector:
  B  DK   D  GR   E   F IRL   I   L  NL   P  UK
  2   2   2   1   1   2   1   2   2   2   1   2

Within cluster sum of squares by cluster:
[1] 57.46750 71.91375
 (between_SS / total_SS =  84.7 %)

Available components:
```

```
[1] "cluster"      "centers"      "totss"          "withinss"
[5] "tot.withinss" "betweenss"    "size"
```

[해석] cluster means (결과중 centers) : 군집평균
clustering vectors (결과중 cluster) : 할당된 군집 번호
totss (total sum of squares) : 총제곱합
withinss (within ss, W) : 군집내 잔차제곱합
tot.withinss : sum of withinss
betweenss (between ss, B): 군집간 잔차제곱합
size : 할당된 각 군집의 자료 개수

```
> a=kmeans(Euro.d,2)
> a$bet/(a$bet + a$tot.within) # 이 값이 클수록 분리가 잘 되었다.
[1] 0.8470351
```

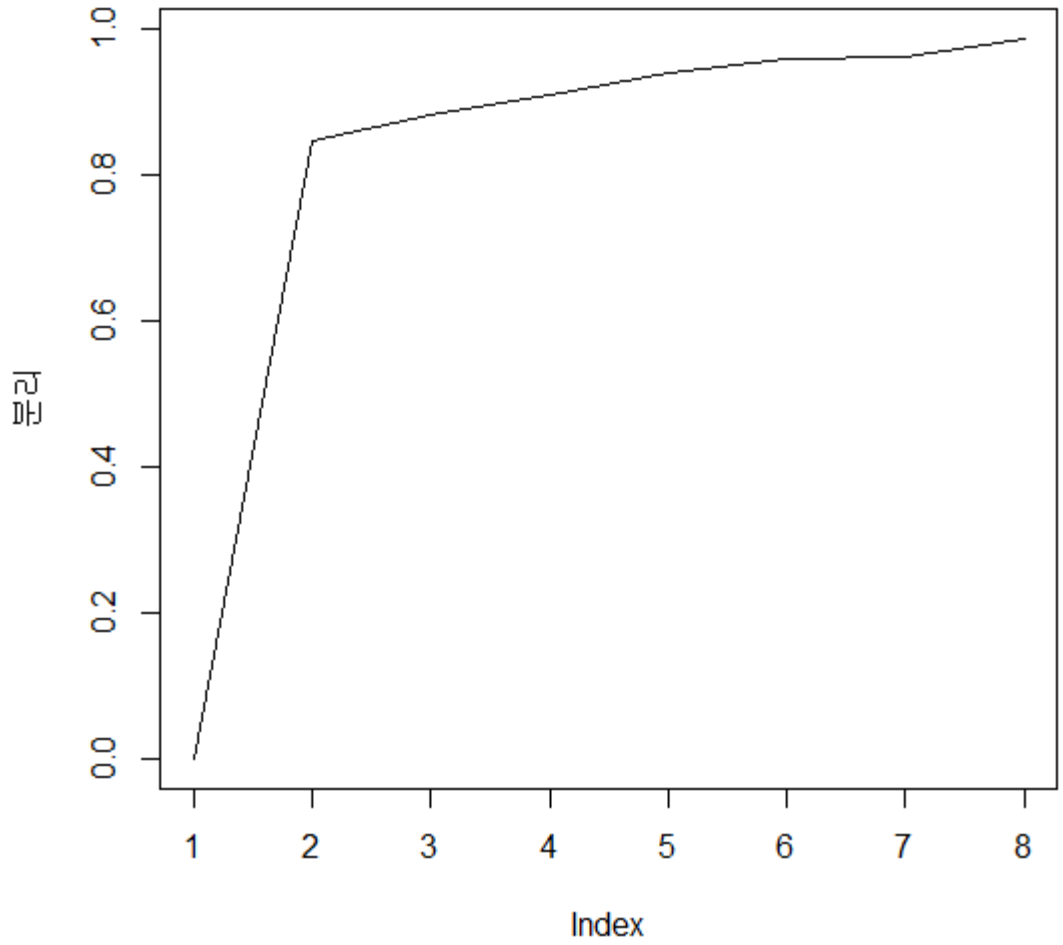
```
> b=kmeans(Euro.d,3)
> b$bet/(b$bet + b$tot.within)
[1] 0.8852474
> b=kmeans(Euro.d,4)
> b$bet/(b$bet + b$tot.within)
[1] 0.9108413
> b=kmeans(Euro.d,5)
> b$bet/(b$bet + b$tot.within)
[1] 0.9490536
> b=kmeans(Euro.d,6)
> b$bet/(b$bet + b$tot.within)
[1] 0.9553768
```

Q : 가장 최선의 K는? A: scree plot으로 결정할 수 있다.

```
> b=rep(0,8);
> for(i in 2:8){
  a=kmeans(Euro.d,i);
  b[i]=a$bet/(a$bet+a$tot.within);
  }
> round(b,3)
[1] 0.000 0.847 0.883 0.911 0.940 0.958 0.962 0.987
```

스크리플롯과 유사하게 급격한 증가있은 후에, 점점 분리에 따른 효과가 줄어드는 것을 볼 수 있다. Euro.d는 군집의 개수를 2개로 하는 것이 타당하다.

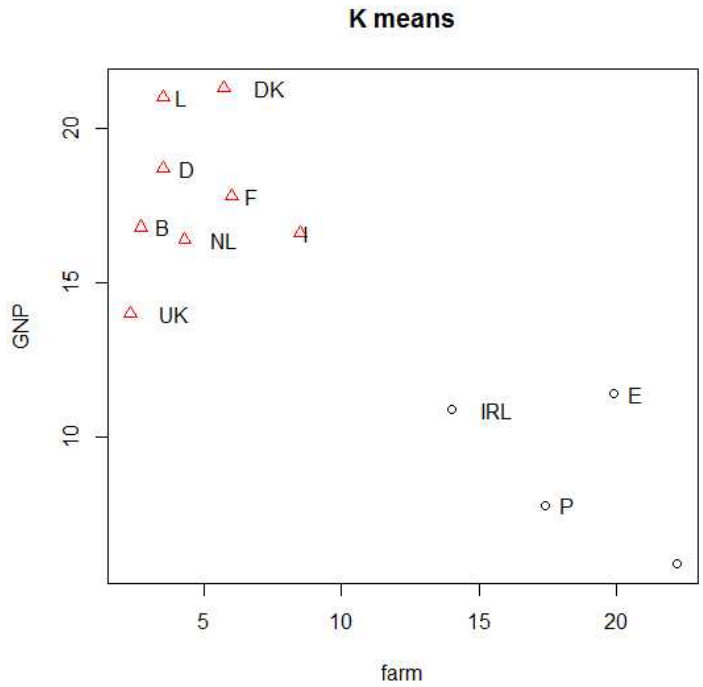
```
> plot(b,type="l",ylab="분리")
```


분리
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
Index

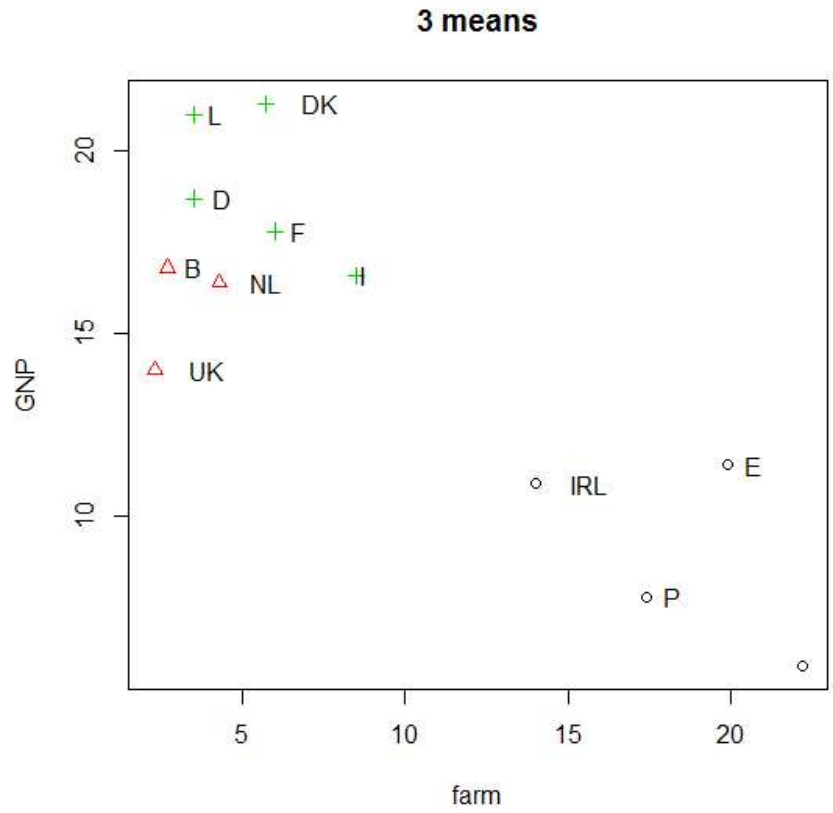
[결과 분석]

2-평균법

```
> plot(Euro.d,pch=a$clus,col=a$clus,main="K means")
> text(Euro.d,rownames(Euro.d),adj=-1)
```



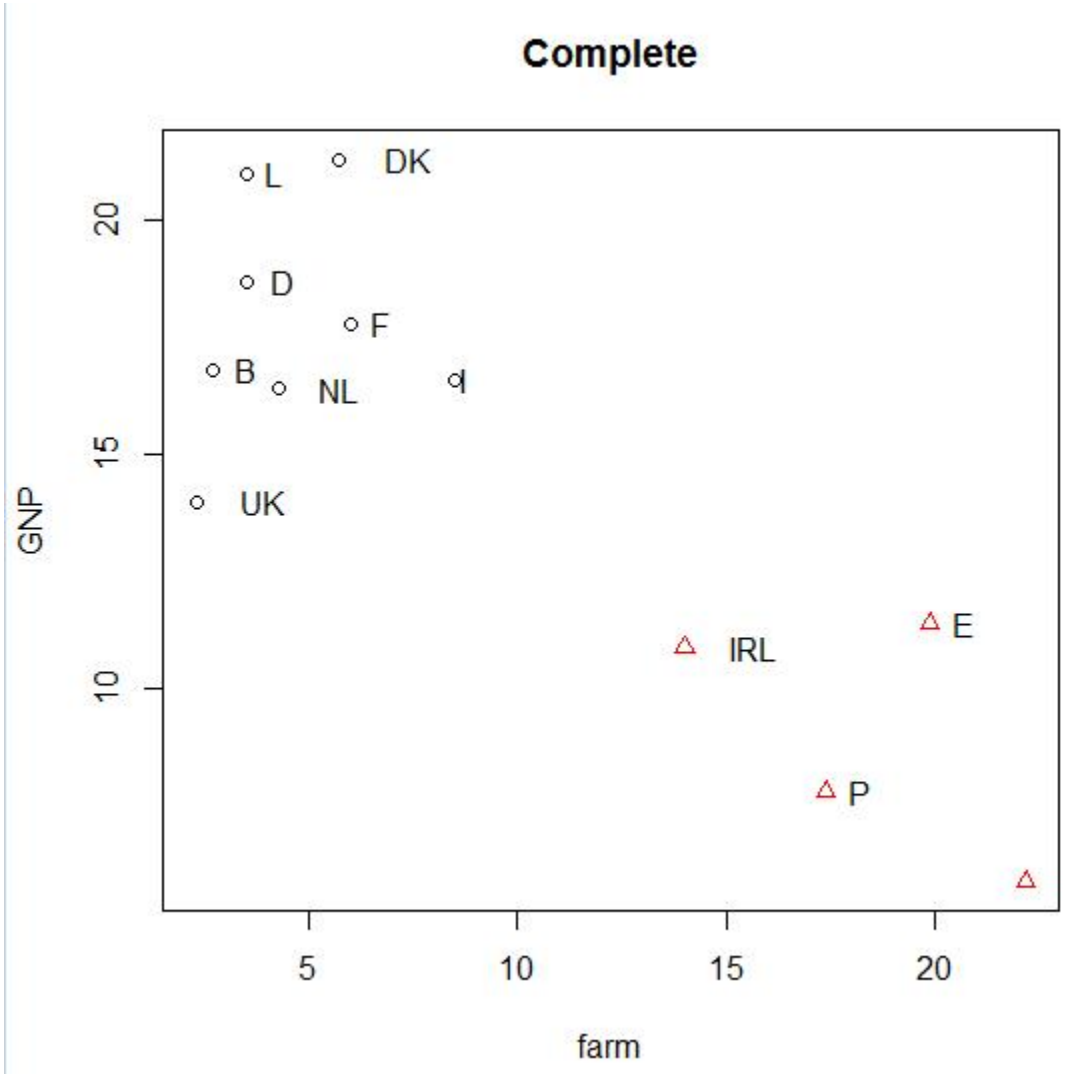
```
> a=kmeans(Euro.d,3)
> plot(Euro.d,pch=a$clus,col=a$clus,main="3 means ")
> text(Euro.d,rownames(Euro.d),adj=-1)
```



최장결합법

```
>a=hclust(dist(Euro.d),method="complete")
>  two.rts=cutree(a,2)
>  two.rts
  B  DK   D  GR   E   F IRL   I   L  NL   P  UK
  1   1   1   2   2   1   2   1   1   1   2   1
>  plot(Euro.d,pch=two.rts,col=two.rts,main="Complete")
>  text(Euro.d,rownames(Euro.d),adj=-1)
```

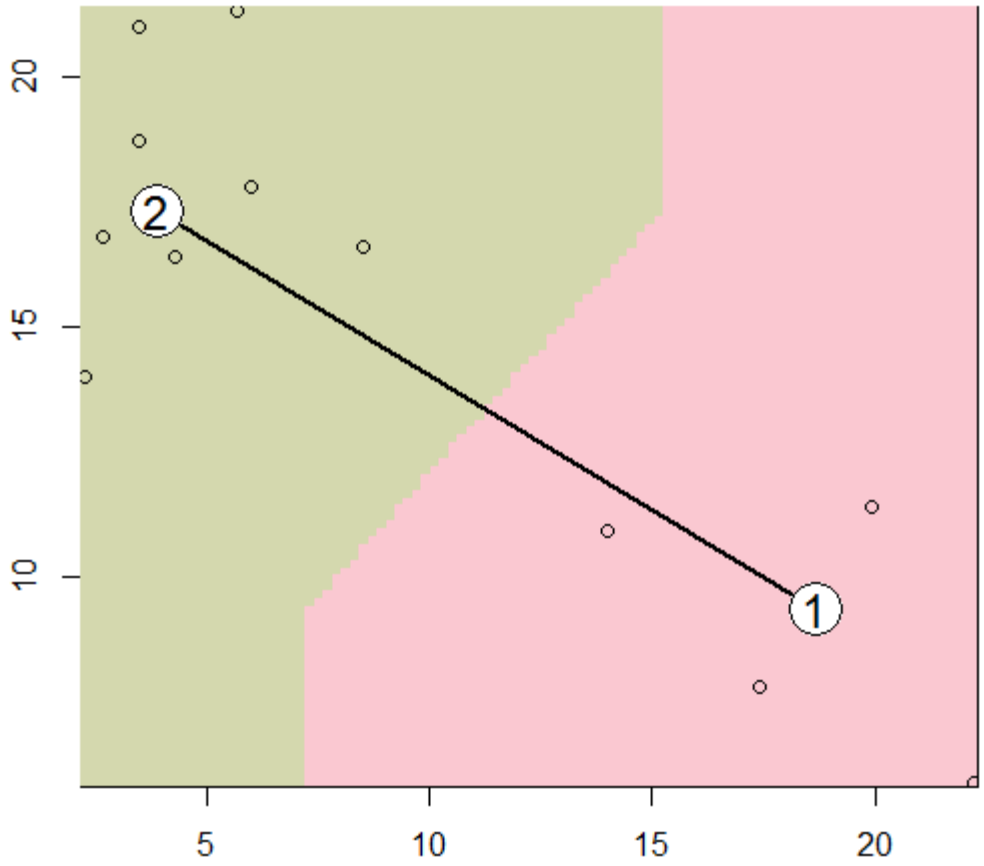
K-평균법과 결과가 일치한다.


Complete
DK
L
D
F
B
NL
UK
IRL
E
P
GNP
farm
20
15
10
5
10
15
20

12.3.2 K-medians

- 평균은 이상치에 영향을 크게 받는다.
- 평균대신에 중앙값을 사용하여 K-means를 대신한다.

```
>library(flexclust)
> a=kcca(Euro.d,k=2,family=kccaFamily("kmedians"))
> a@cluster
 B  DK   D  GR   E   F IRL   I   L  NL   P  UK
 2   2   2   1   1   2   1   2   2   2   1   2
> image(a)
> points(Euro.d)
```


20
15
10
5
10
15
20
2
1

● 군집분석시 유의사항

군집분석은 소속 집단에 대한 정보가 없을 때, 개체들사이의 유사성을 근거로 자율적으로 군집을 형성시키는 다변량분석 기법이다.

1. 만약 자료 자체의 뚜렷한 군집이 없을 경우에는 군집분석의 성능이 우수하지 못하다.
2. 군집분석은 이상치에 큰 영향을 받는다. 이상치의 존재유무를 반드시 확인하여야 한다.
3. 군집분석은 변수들의 측정척도가 상이한 경우 많은 영향을 받으므로, 이 경우에는 표준화를 사전에 수행하여야 한다.
4. 군집의 타당성을 검토하기 위하여, 여러 방법을 시행한 후 결과들을 비교하여 유사한 결과를 가지는 경우를 최종결과로 받아드린다.