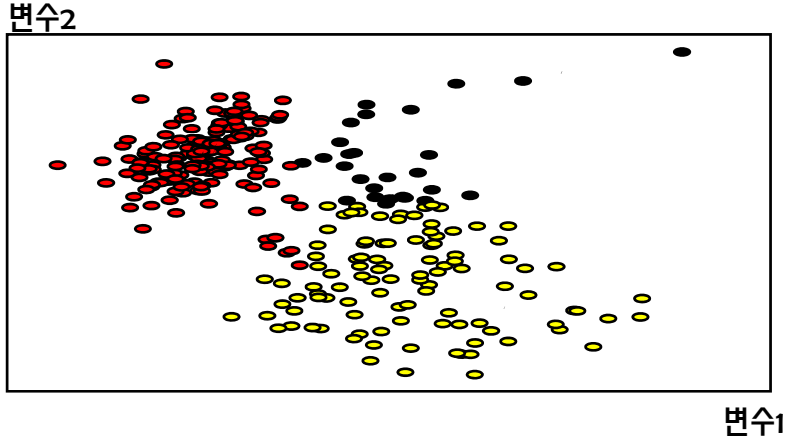
군집분석 개요 (페이지 248)

- Individual Directed Technique
  - 범주(그룹)에 대한 사전 정보가 없음
  - 다변량 측정치를 동시에 고려하여 데이터 개체 분류
    - 개체의 유사성(similarity, 거리의 반대 개념)을 측정변수들을 이용하여 계산
    - 유사성이 높은 개체를 군집으로 묶어간다.
  - 개체를 집단으로 그룹화 하여 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 얻는 분석 기법
    - 동일 군집 내의 관찰치는 서로 비슷한 속성을 갖도록 하고 서로 다른 군집에 속한 관찰치는 상이한 속성을 갖도록 군집을 구성

- 군집 원칙
  - 동일 군집에 속한 개체 유사한 속성 많음
  - 다른 군집에 속하면 유사성 매우 낮음

- 데이터 유형
  - 측정변수: 측정형(등간 척도 포함)
    - 개체의 속성을 판단하는 기준

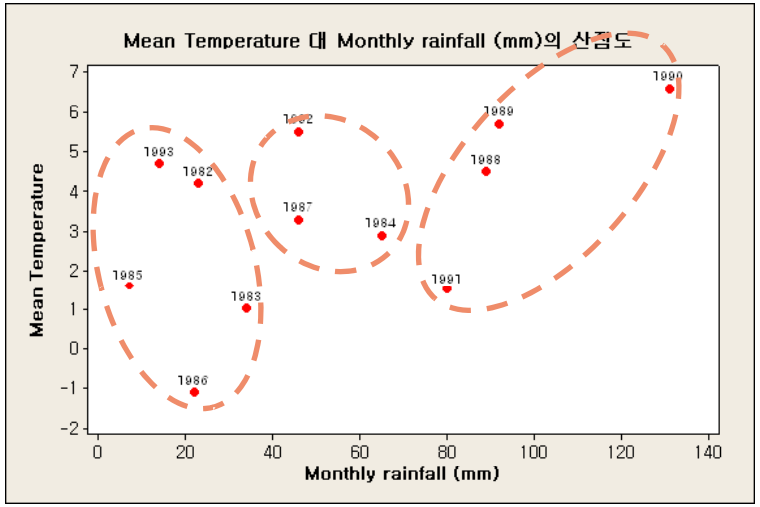
- 목적
  - 유사한 성향을 가진 개체를 모아 군집을 형성
  - 시각적 표현(주성분 분석 이용)을 통하여 군집간의 특성을 관찰하거나 목표변수와 관계를 파악
    - 개체를 동질적 속성에 의해 묶음으로써 데이터의 구조를 파악할 수 있음
    - 데이터의 차원을 축약하여 이용할 수 있음
    - 개체를 분류하기 위한 명확한 분류기준이 존재하지 않거나 기준이 밝혀지지 않은 상태에서 유용하게 이용

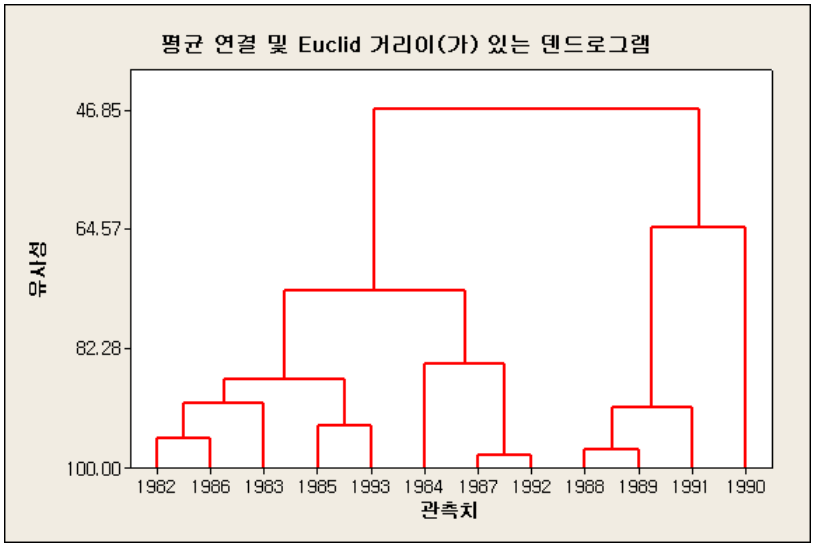


군집분석 개요(2)

- 장점
  - 탐색적인 기법: 주어진 자료의 내부구조에 대한 사전정보 없이 의미 있는 자료구조를 찾아낼 수 있음
  - 다양한 형태의 데이터에 적용가능: 유사성(거리)만 정의되면 모든 종류(텍스트 데이터)의 자료에 적용할 수 있음.
  - 분석방법 적용 용이성: 자료의 사전정보를 필요로 하지 않아서 누구나 쉽게 분석할 수 있음

- 단점
  - 가중치와 거리 정의: 가중치와 거리를 어떻게 정의하는가에 따라 군집분석의 결과가 아주 민감하게 반응함
  - 초기 군집 수의 결정이나, 군집 개수 결정이 쉽지 않음
  - 결과의 해석이 어려움: 찾아진 군집이 무엇을 의미하는지 데이터만을 이용해서는 알 수가 없는 경우가 많음
    - 주성분 분석을 이용하여 개체 집단 표현
    - 인구학적 특성에 의해 개체 특성 파악

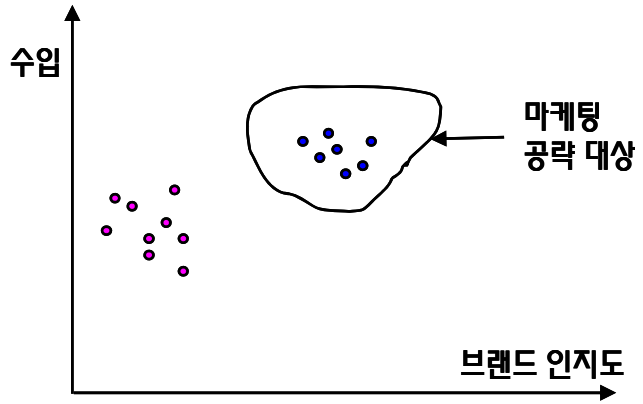




군집분석 활용 (페이지 248)

- 시장세분화 (마케팅)
  - 구매태도, 구매성향, 매체사용 습관 등과 같은 특성에 있어서 공통적인 특성을 공유하는 사람, 시장,조직 등의 대상들의 군집을 발견하여 시장 세분화에 활용
- 소비자를 분류함으로써 소비자 행동 이해 가능
  - 소비자들을 특성(측정 변량)에 따라 분류하여 내재된 특성 파악
- 잠재적인 신제품 기회 발견
  - 전체 시장에서의 경쟁상황과 경쟁관계에 있는 상표나 기업을 속성에 따른 군집 도출.
- 개체 데이터를 일반적이고 다루기 쉽게 축약
  - 표본추출을 위해 조사대상 지역이나 소비자들을 구분해야 하는 경우: 군집분석에 의해 서로 유사한 특성을 가진 대상끼리 묶어 계층이 구성되면 표본추출의 작업이 용이하고 정확성도 유지할 수 있음
- 고객 세분화
  - 고객이 기업의 수익에 기여하는 정도를 통한 고객세분화
  - 우수고객의 인구통계적 요인, 생활패턴 파악: 개별고객에 대한 맞춤관리
  - 고객의 구매패턴에 따른 고객세분화
  - 신상품 판촉, 교차 판매를 위한 표적집단 구성

변량	전체	추종자	EB	기본 기능
나이	28.6	27.2	19.3	53.0
교체의사	.429	.333	.833	.000
고등학생	.238	.0833	.667	.000
영업직	.238	.333	.000	.333
성별	.619	.750	.333	.667
대학생	.143	.167	.167	.000
전문직	.286	.333	.167	.333
거주지	1.33	1.50	1.17	1.00
생산직	.0476	.0833	.000	.000



군집 방법 (페이지 251)

Hierarchical 계층적 방법

- 유사성이 가까운 순서대로 개체들을 묶어(군집화) 가는 방법
- 한계점
  - 한 대상이 일단 어느 군집에 소속되면 다른 군집으로 이동될 수 없음
  - 이상 개체(outlier)는 제거되지 않고 반드시 어느 군집에 속한다.

Non-hierarchical 비계층적 방법

- 군집의 중심이 되는 seed 점들 집합을 선택하여 그 seed 점과 유사성이 높은(거리가 가까운) 개체들을 그룹화 방법
- 문제점
  - 사전에 군집(그룹) 수에 대한 예상이 필요하다.
  - 개체 분류는 처음 선정한 seed 점들에 의해 영향을 많이 받아 분석에 따라 분류가 다를 가능성이 있다.
  - 군집의 수와 seed 값의 위치의 결합 조건이 너무 많아 계산이 분류를 위한 계산이 용이하지 않다.

- 계층적 방법에 의해 군집화 하여, 적절한 군집 수와 이상 개체를 결정한다.
- 이상 개체 제외하고 결정된 군집 수를 이용하여 비계층적 방법에 의해 군집화 한다.

판별분석과 차이

- 판별분석은 사전 집단 정보가 있는 경우 집단들간의 차별적 특성을 설명하는 변수들을 발견하여 판별식 유도
- 군집분석은 사전에 집단이 나누어져 있지 않으며 변수를 이용하여 개체 유사성 측정하고 개체를 집단화

요인분석과 차이

- 데이터의 구조를 평가한다는 점에서 요인분석에 비유될 수 있으나
- 요인분석은 변수 그룹화, 군집분석은 개체의 그룹화
  - 데이터를 전치하여 변수의 유사성으로 변수 분류 가능

유의사항

- 체계적인 통계적 추론에 의해 개발되지 않아 상대적으로 단순한 절차로 그 결과가 검정되지 못한 경우가 많음
- 동일한 표본에 대하여 상이한 군집분석 알고리즘을 사용하는 경우 상이한 결과가 만들어 질 수 있음
- 많은 경우에 있어서 거리측정방법이 달라지면 군집분석의 결과도 달라짐. 가능하면 몇 가지 측정방법을 사용하여서 이 결과를 이론적인 내용이나 기존의 연구결과와 비교해서 평가하는 것이 바람직함
- 변수간의 측정 척도가 상이한 경우에는 군집분석을 행하기 전에 표준화 하는 것이 바람직함. 이러한 표준화는 특정 변수의 변화 정도가 다른 변수에 비해 특히 큰 경우에 바람직함

계층적(hierarchical) 군집분석

- 개념
  - 데이터를 사용하여 유사성이 가장 큰 개체끼리 순차적으로 개체를 분류
  - 계층 군집분석의 결과인 덴드로그램 (Dendrogram)을 통해 개체 군집 현황과 전체 군집들간의 구조적 관계 파악
  - 군집 이름 부여, 군집 특성 파악: 주성분 분석 활용

- 주요 원리
  - 개체(집단)끼리 유사성(similarity) 측정하여 가장 유사한 개체(혹은 집단)끼리 순차적으로 묶음
    - 전체 대상을 하나의 군집으로 해서 출발하여 개체들을 분할해 나가는 방법: 분할 (Division) 방법
  - 개체간 유사성 정도를 측정하는 개념 필요: 유사성을 거리로 정의
  - 집단과 개체(개체) 유사성 정의 필요: 연결(linkage) 방법

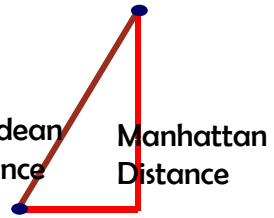
- 유사성 개념
  - 데이터 내 속성(변수)면에서 개체의 유사 정도를 나타냄
  - 군집분석에서는 비유사성 척도인 거리(distance)를 이용

- 거리의 종류
  - 개체 i, 개체 k, j=1,2, ..., p: 군집 변수
  - 클리드(Euclidian) 거리: 최단 거리, 가장 많이 사용
  - 맨하탄(Manhattan) 거리: 직선 이동 거리, 이상치 비중 약해짐
  - 피어슨(Pearson) 거리: 거리를 변수 분산으로 나누어 표준화 개념

$$d(i,k) = \sqrt{\sum_j (x_{ij} - x_{kj})^2}$$

$$d(i,k) = \sum_j \left| x_{ij} - x_{kj} \right|$$

$$d(i,k) = \sqrt{\sum_j \left(x_{ij} - x_{kj} \right)^2 / v_j}$$

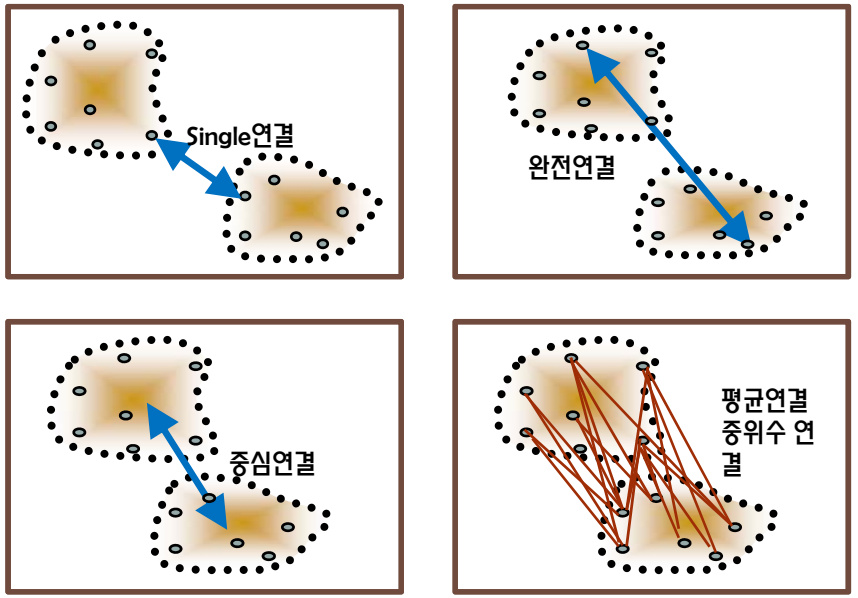


$$Z_{ij} = \frac{X_{ij} - \bar{X}_{\cdot j}}{S(X_{\cdot j})}$$

- 변수 표준화
  - 군집 변수의 단위가 다르면(분산의 크기 다름) 단위 큰 변량이 개체 거리(유사성)에 영향을 준다.
  - 그러므로 변량 단위 통일을 위한 변량 표준화 필요
  - Pearson 거리는 표준화 개념이 고려됨

계층적 군집분석 개체(군집)연결방법(페이지 252)

- **Linkage?** 가까운 개체 집단끼리 순차적으로 묶어갈 때 집단과 개체 (혹은 집단) 거리 측정을 위한 개념
- **거리 측정 방법**
  - Nearest neighbor (single 단일): 두 군집의 각 개체 중 가장 가까이 있는 개체의 거리
  - Furthest neighbor (완전 complete)r: 두 군집의 각 개체 중 가장 멀리 있는 개체의 거리
  - Centroid neighbor (중심연결): 군집의 평균 간의 거리
  - Average neighbor (평균연결): 한 군집의 개체와 다른 군집 개체들의 각 거리 평균
  - Median neighbor (중위수 연결): 평균 대신 거리 중위수 사용, 이상치의 영향 적음
  - Ward's minimum variance: 군집의 평균간 거리를 각 군집의 개체 개수의 역의 합으로 나눈 제곱근을 구한 거리
- **어떤 방법을 사용하는 것이 좋은가?**
  - Nearest 방법은 군집의 수가 줄어들고 이상 개체 판단에 유리
  - Furthest는 군집간 거리를 최소화 하는 경향이 있어 개체 수가 적은 군집을 얻음
  - 가장 많이 사용하는 방법은 Average neighbor 방법
  - 여러 방법 사용하여 군집간 평균 거리, 군집 내 개체간 평균 거리가 작은 군집 방법



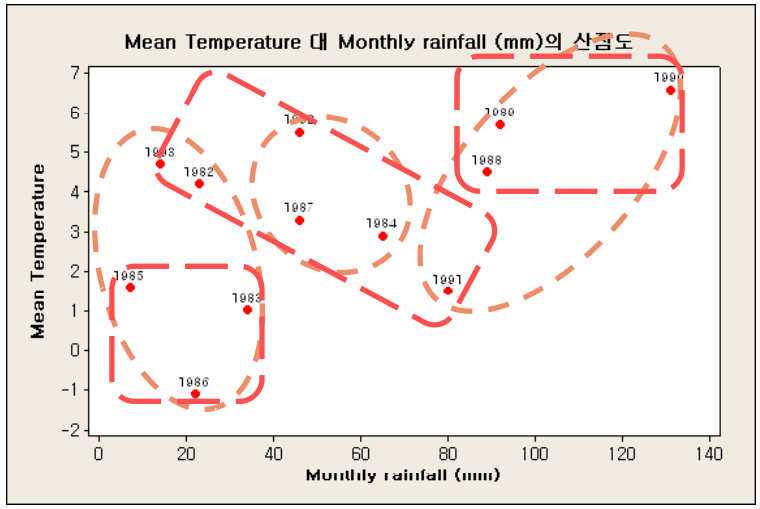
계층적 군집분석 Dendrogram 방법 (페이지 255)

■ 정의

- 군집의 병합 과정 및 집단간 거리를 이차원 도면을 사용하여 간략히 표현
- 유사성이 높은(거리가 가까운) 순서대로 개체를 순차적 연결
- 덴드로그램에서 선의 높이는 유사성 크기를 표시

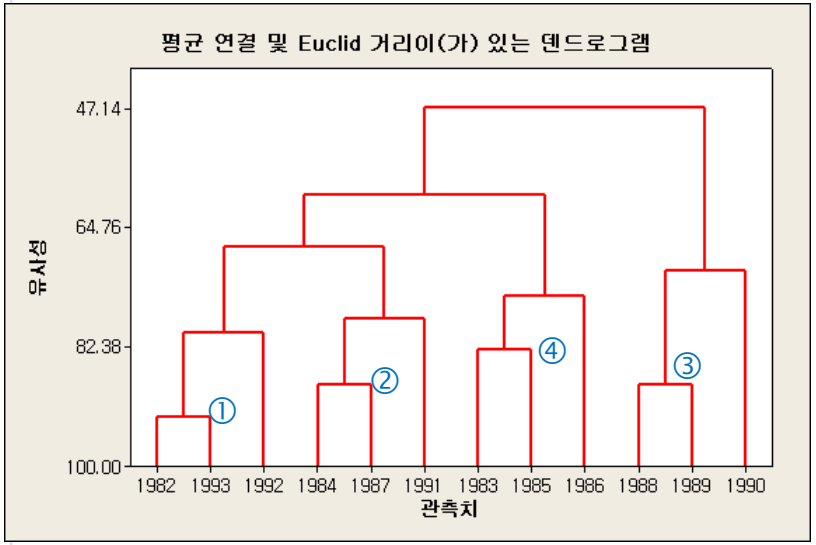
■ 예제

- 1982~1993년 2월(n=12)을 평균 강수량과 평균 온도 속성으로 군집화 하자. (변수 표준화)



■ 군집의 개수 결정

- 덴드로그램 이용
  - 시각적 판단 eye ball judgement
- Hotelling T 값
  - 집단 간 평균 차이 검정
  - 값이 큰 부분에서 집단 결정
- CCC(Cubic Clustering Criterion, Searle; 1983)
  - 3 이상이며, 갑자기 줄어드는 곳에서 군집 개수 결정



비계층적 (non-hierarchical) 군집분석

▪비계층적 군집 정의

- 군집의 개수를 분석 전에 정해야 한다.
  - 계층적 군집, 사전 정보, 분석자의 결정에 의해 군집의 개수 분석 전 결정
- 군집의 중심을 결정
  - 우선 seed(군집의 중심)를 정하고 이 seed에 가까운 개체들을 군집으로 묶는다.
  - 군집이 결합되면, 각 군집별 군집화 과정 오류를 계산한다.
- 군집화 단계에서 오류가 발생하면 seed를 조정하고 오류를 재계산한다.
- 군집화의 각 단계가 끝나면서 발생하는 오류가 발생하지 않으면 군집화를 종료한다.

▪방법

- 군집의 중심 결정
  - 집단 내 개체 평균: K-means 비계층적 방법
  - Euclidian 거리 중심
- 군집의 크기 결정 (fast cluster)
  - 반지름(radius) 길이

▪K-평균 군집 방법

- 사전에 결정된 군집 수 K에 기초하여 각 관측 값을 군집의 중심들 중에서 가장 가까운 군집에 할당하는 방법

단계 1 : 군집의 수 K를 결정

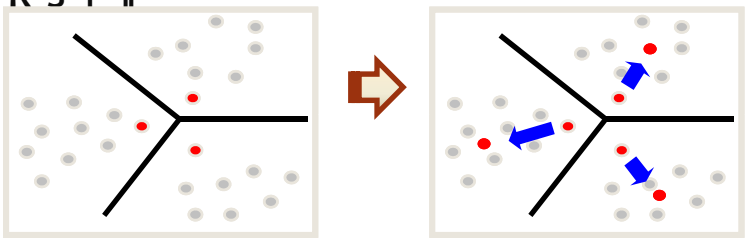
단계 2 : 초기 K개 군집의 중심을 랜덤하게 선택함

단계 3 : 각 관측 값들을 가장 가까운 중심의 군집에 할당함

단계 4 : 새로운 군집에 할당된 관측 값들로 새로운 중심을 계산

단계5: 개체 군집 변동이 없을 때까지 단계 3, 4를 반복한다.

- K=3의 예



- 군집 수 K 값 결정
  - 계층적 분석 방법에 의해 k 결정
  - 여러 k 사용하여 군집간 평균 거리나 군집 내 개체 평균 거리를 활용하여 최적 k 결정

계층적 군집분석 예제 (페이지 256)

- 데이터 PIZZA.txt
  - 56개 피자, 7개 성분 함유량(id mois prot fat ash sodium carb cal)

- 군집변수 별 평균
  - 군집분석은 개체의 유사성(일반적으로 거리로 측정)에 의해 개체를 묶는 것이므로 변수의 단위 차이가 있으면? 그래서 표준화 먼저...

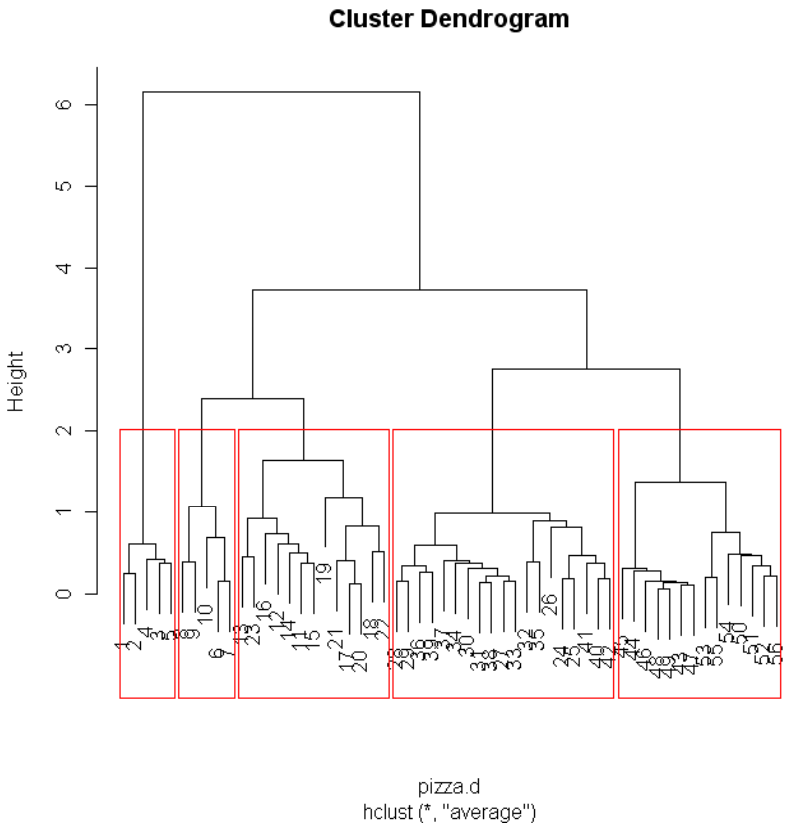
- 계층적 군집 (Euclidean 거리, average linkage)
  - 반드시 군집 변수만 가지고 개체 군집

```
> pizza=read.table("pizza.txt",header=T)
> pizza0=pizza[,2:8]
> pizza.s=scale(pizza0)
> pizza.d=dist(pizza.s,method="euclidean")
> pizza.ca=hclust(pizza.d,method="average")
> plot(pizza.ca) # display dendogram
```

  - 덴드로그램의 의한 군집 개수 결정
    - 군집의 개수 5개?

```
> groups=cutree(pizza.ca,k=5)
> rect.hclust(pizza.ca,k=5,border="red")
```

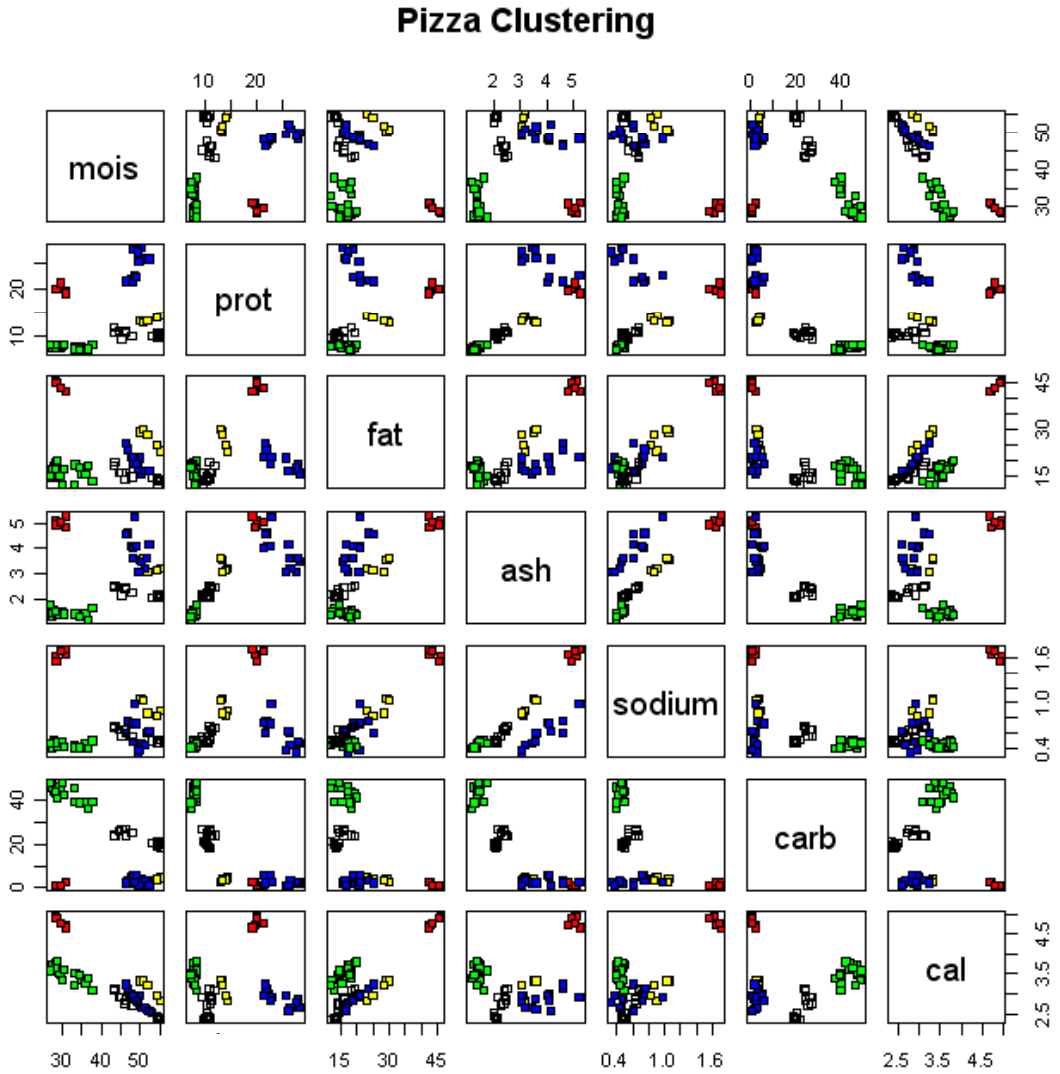
    - 군집 박스는 rest.hclust() 함수에 의해 나타남



군집분석 결과 표현 (페이지 263)

▪개체 군집별 군집변수 간 산점도

```
> groups=cutree(pizza.ca,k=5)
> rect.hclust(pizza.ca,k=5,border="red")
> pizza.fin=cbind(pizza0,groups)
> pizza.fin=cbind(pizza0,groups)
> pairs(pizza.fin[c("mois","prot","fat",
+ "ash","sodium","carb","cal")],
+ main="Pizza Clustering", pch=22,
+ bg=c("red", "yellow", "blue","green")
+ [unclass(pizza.fin$groups)])
```



군집분석 결과 이름 부여 (페이지 261)

▪군집별 평균 구하기

```
> library(doBy)
> summaryBy(mois+prot+fat+ash+sodium+carb+cal
+ ~groups,data=pizza.fin,
+ FUN = function(x){c(m=mean(x))})
  groups    mois.m    prot.m    fat.m    ash.m
1      1 29.72000 19.958000 43.80200 5.028000
2      2 52.30800 13.556000 27.08000 3.298000
3      3 49.07308 24.576923 19.47385 3.899231
4      4 31.12684  7.901579 16.29316 1.434211
5      5 49.71857 10.555714 14.94643 2.230000
   sodium.m    carb.m    cal.m
1 1.6480000  1.618000 4.804000
2 0.9320000  3.758000 3.130000
3 0.5976923  2.984615 2.856154
4 0.4421053 43.258421 3.513684
5 0.5500000 22.549286 2.667143
```

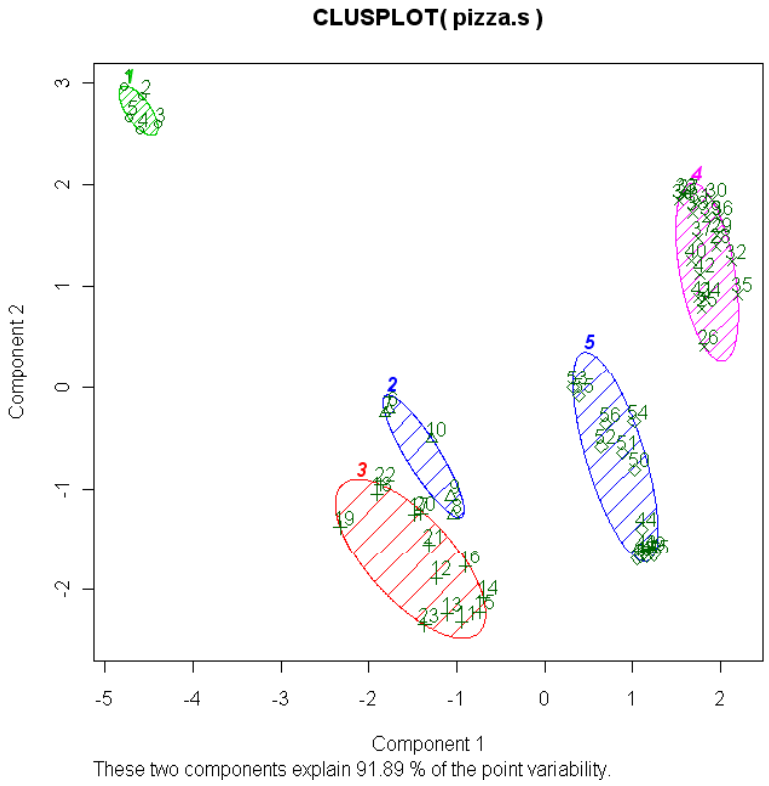
▪주성분 이름 부여

```
> prcomp(pizza0,scale=T)
Standard deviations:
[1] 1.977158827 1.588502006 0.
[5] 0.163030529 0.007954041 0.

Rotation:
                PC1         PC2
mois     0.07246752 -0.5914399
prot     0.37660199 -0.2877471
fat      0.43916372  0.2812532
ash      0.48641175 -0.1085911
sodium   0.44020235  0.2271298
carb    -0.43137342  0.3199851
cal      0.20879920  0.5679143
```

▪주성분 변수 표현

```
> library(cluster)
> clusplot(pizza.s,groups,
+ color=TRUE, shade=TRUE,labels=2, lines=0)
```



비계층적 군집분석 방법 (페이지 276)

▪K-means 방법

```
> # K-Means Clustering with 5 clusters
> pizza.kc=kmeans(pizza.s,5)
> # Cluster Plot against 1st 2 principal components
> library(cluster)
> clusplot(pizza.s, pizza.kc$cluster, color=TRUE,
+ shade=TRUE,labels=2, lines=0)
```

• 평균에 의한 군집 해석

```
> pizza.fin=cbind(pizza,pizza.kc$cluster)
> library(doBy)
> summaryBy(mois+prot+fat+ash+sodium+carb+cal
+ ~pizza.kc$cluster,data=pizza.fin,
+ FUN = function(x){c(m=mean(x))})
  pizza.kc$cluster   mois.m    prot.m    fat.m
1                1 28.71500  8.010833 16.55500
2                2 49.71857 10.555714 14.94643
3                3 49.97167 21.515556 21.58667
4                4 35.26143  7.714286 15.84429
5                5 29.72000 19.958000 43.80200
     ash.m  sodium.m    carb.m    cal.m
1 1.463333 0.4516667 45.278333 3.622500
2 2.230000 0.5500000 22.549286 2.667143
3 3.732222 0.6905556  3.199444 2.932222
4 1.384286 0.4257143 39.795714 3.327143
5 5.028000 1.6480000  1.618000 4.804000
```

• 주성분 변수에 의한 군집 해석