

데이터마이닝

(Data Mining)

한국방송통신대학교
정보통계학과 장영재 교수

12 강 /

군집분석

목차

12. 군집분석

- 1) 군집분석이란?
- 2) 비유사성 척도
- 3) 계층적 군집분석
- 4) 비계층적 군집분석
- 5) 군집분석의 특징

1. 군집분석이란?

군집분석이란?

- 군집분석(cluster analysis)이란 관측값 또는 개체를 의미 있는 몇 개의 부분집단으로 나누는 과정을 의미
 - 군집분석은 대표적인 자율학습 방법으로 특성이 유사한 개별 개체들끼리 스스로 묶이도록 군집을 형성하게 되며, 이렇게 형성된 군집의 특징을 사후적으로 분석
 - 군집이란 군집분석(cluster analysis) 과정에서 나뉜 부분집단
 - 각 대상의 유사성을 측정하여 유사성이 높은 대상 집단을 분류하고, 군집에 속한 개체들의 유사성과 서로 다른 군집에 속한 객체들 간의 상이성을 규명
 - 유용한 군집이란 같은 집단 내의 관측값들은 서로 유사하고 다른 군집에 속한 관측값들 간에는 유사성이 적은 것을 의미
 - ex) 타겟 마케팅, 고객 세분화 등
- 군집분석은 자료의 사전정보 없이 자료를 파악하는 방법으로, 분석자의 주관에 결과가 달라질 수 있음

2. 비유사성 측도

2. 비유사성 측도

- ▶ 군집분석에서의 개체의 개수가 n 이고 변수의 개수가 p 개일 때의 자료구조는 <그림1>과 같음. 즉, 행렬 X 의 행은 개체, 열은 변수를 의미하며 x_{ij} 는 i 번째 개체에서 j 번째 연속형 변수의 관측값을 나타냄

<그림1> 군집분석의 자료구조

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} \quad x_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{ip})$$

2. 비유사성 측도

- 두 개체가 연속형 변수로 표현될 수 있다면 이들의 비유사성 (dissimilarity)은 개체 간 거리로 간주할 수 있음

- 비유사성(거리)의 측정 방법

- i번째 개체와 k번째 개체를 나타내는 두 연속형 변수를 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$

와 $x_k = (x_{k1}, x_{k2}, \dots, x_{kp})$ 라고 할 때, 비유사성은 각각 다음과 같이 정의

유클라디안(Euclidian)거리 : $d(x_i, x_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2}$

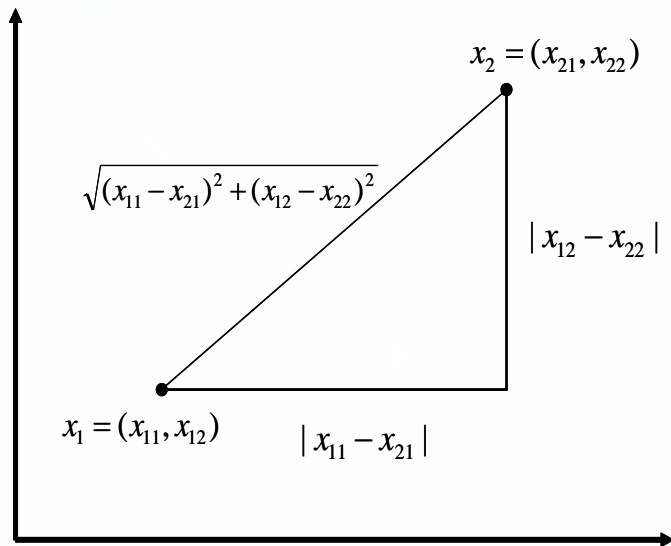
맨해튼(Manhattan) 거리 : $d(x_i, x_k) = \sum_{j=1}^p |x_{ij} - x_{kj}|$

민코브스키(Minkowski) 거리 : $d(x_i, x_k) = (\sum_{j=1}^p |x_{ij} - x_{kj}|^m)^{1/m}$

2. 비유사성 측도

- <그림 2>는 좌표평면에서 두 점간의 유클리디안 거리와 맨해튼 거리를 비교한 그림

<그림2> 두 점간의 유클리디안 거리와 맨해튼 거리 비교



2. 비유사성 측도

- 이밖에도 다차원 공간에서의 거리 측정방법인 마할라노비스 (Mahalanobis) 거리와 코사인 거리 등이 있음

- 마할라노비스 거리는 두 지점의 단순한 거리뿐만 아니라, 변수의 특성을 나타내는 분산과 공분산이 함께 고려되며 코사인 거리는 다차원의 양수 공간에서의 거리 측정에 많이 사용됨

마할라노비스(Mahalanobis)거리 : $d(x_i, x_k) = (x_i - x_k)^T \Sigma^{-1} (x_i - x_k)$

코사인 (Cosine) 거리 : $1 - \frac{\sum_{j=1}^p x_{ij} \times x_{kj}}{\sqrt{\sum_{j=1}^p x_{ij}^2} \times \sqrt{\sum_{j=1}^p x_{kj}^2}}$

3. 계층적 군집분석

3. 계층적 군집분석

- 계층적 군집분석에는 가까운 관측 값들끼리 묶는 응집분석 (agglomerative analysis)과 먼 관측 값들을 나누어가는 분할분석 (divisive analysis)이 있음
 - 계층적 군집화는 비계층적 군집화에 비하여 군집의 수에 대한 사전 지식이 필요하지 않다는 장점
 - 한 계층에서 어떤 군집에 할당되면 그 계층 아래에서는 다른 위 계층에서 나뉘는 다른 군집으로 할당될 수 없다는 특징이 있음

3. 계층적 군집분석

1) 응집분석

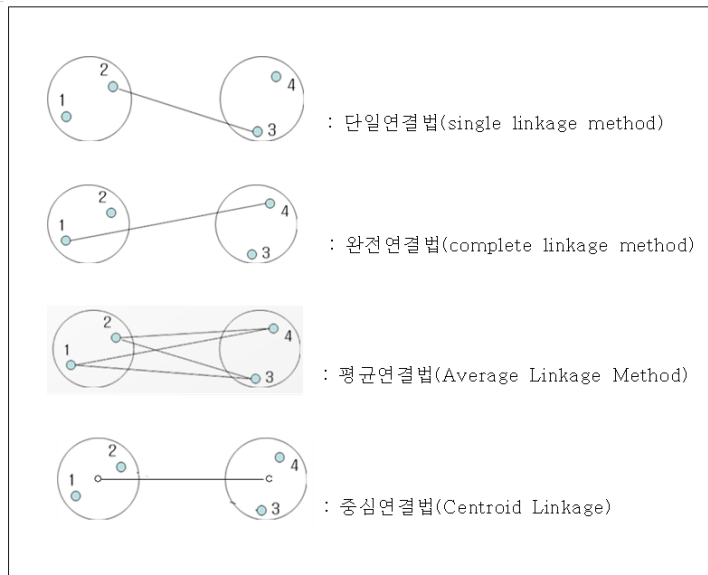
➤ 응집분석의 알고리즘

- ① 각 개체를 하나의 군집으로 하여 전체 n 개의 군집을 형성
- ② 각 군집 간의 거리를 기준으로 가장 가까운 두 개의 군집을 병합하여 $n-1$ 개의 군집을 형성
- ③ $n-1$ 개의 군집 중 가장 가까운 두 군집을 병합하여 군집을 $n-2$ 개로 축소
- ④ 군집의 수를 줄여나가며 전체가 하나의 군집을 이룰 때까지 이 과정을 반복

3. 계층적 군집분석

1) 응집분석

- 자료에서 i 번째 개체와 k 번째 개체의 거리를 $d(i, k)$ 라 하고, 군집 P 와 군집 Q 의 거리를 $d(P, Q)$ 라 할 때 개체 간의 거리 $d(i, k)$ 는 앞서 살펴본 여러 가지 방법 중 하나를 선택



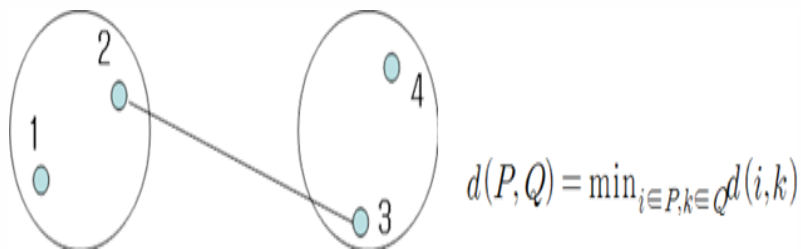
〈그림3〉 거리 계산 방법에 따른 주요 응집분석의 종류

3. 계층적 군집분석

- 두 군집 간의 거리 $d(P, Q)$ 를 계산하는 방법에 따라 응집분석의 형태가 달라지는데 응집분석의 방법으로는 다음과 같은 방법이 있음

① 단일연결법 (Single Linkage Method)

- 단일연결법은 최단연결법이라고도 하며 두 군집 P와 Q간의 거리는 P에 속한 개체와 Q에 속한 개체 하나씩을 뽑았을 때 나타날 수 있는 거리의 최솟값으로 측정



〈그림4〉 단일연결법에서의 두 군집 간 거리

3. 계층적 군집분석

〈예제 7-2〉

개체	X1(변수1)	X2(변수2)
1	1	1
2	1	2
3	3	4
4	5	5
5	7	5.5

$$\begin{matrix} & 1 & 2 & 3 & 4 & 5 \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 0 & 1 & 5 & 8 & 10.5 \\ 1 & 0 & 4 & 7 & 9.5 \\ 5 & 4 & 0 & 3 & 5.5 \\ 8 & 7 & 3 & 0 & 2.5 \\ 10.5 & 9.5 & 5.5 & 2.5 & 0 \end{pmatrix}
 \end{matrix}$$

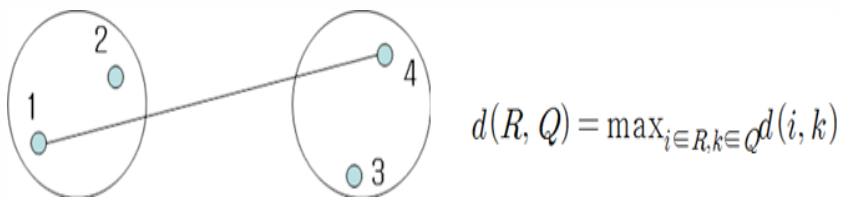
$$\begin{matrix} \{1,2\} & 3 & 4 & 5 \\ \{1,2\} & \begin{pmatrix} 0 & 4 & 7 & 9.5 \\ 4 & 0 & 3 & 5.5 \\ 7 & 3 & 0 & 2.5 \\ 9.5 & 5.5 & 2.5 & 0 \end{pmatrix} \\ 3 & & & \\ 4 & & & \\ 5 & & & \end{matrix} \quad \rightarrow \quad \begin{matrix} \{1,2\} & 3 & \{4,5\} \\ \{1,2\} & \begin{pmatrix} 0 & 4 & 7 \\ 4 & 0 & 3 \\ 7 & 3 & 0 \end{pmatrix} \\ 3 & & \\ \{4,5\} & & \end{matrix}$$

$$\begin{matrix} \{1,2\} & \{3,4,5\} \\ \{1,2\} & \begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix} \\ \{3,4,5\} & \end{matrix}$$

3. 계층적 군집분석

② 완전 연결법 (Complete Linkage Method)

- 완전연결법은 최장 연결법이라고도 하며 두 군집 P와 Q간의 거리는 P에 속한 개체와 Q에 속한 개체 하나를 뽑았을 때 나타날 수 있는 거리의 최대값으로 정의



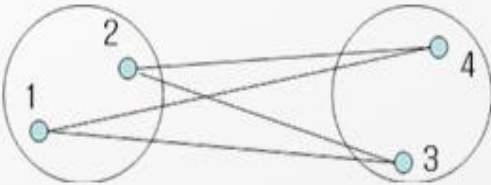
〈그림5〉 완전연결법에서의 두 군집 간 거리

- 완전연결법은 단일연결법에 비해 이상치나 잡음의 존재에 영향을 덜 받는다고 알려져 있고 전형적으로 관측값이 서로 매우 유사하여 밀집되어 있는 군집을 구별하는데 이용

3. 계층적 군집분석

③ 평균연결법 (Average Linkage Method)

- 평균연결법에서는 두 군집 P와 Q간의 거리가 P에 속한 개체와 Q에 속한 개체 간의 모든 거리의 평균

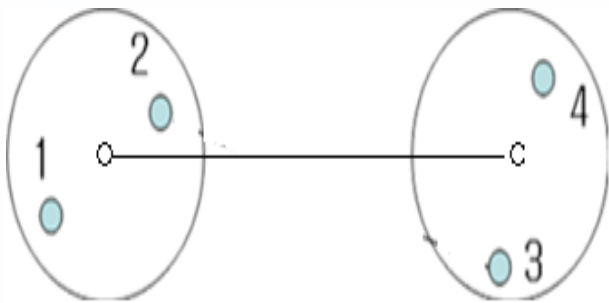
$$d(P, Q) = \frac{\sum_{i \in P, k \in Q} d(i, k)}{(\text{군집 } P \text{에서의 개체수}) \times (\text{군집 } Q \text{에서의 개체수})}$$


〈그림6〉 평균연결법에서의 두 군집 간 거리

3. 계층적 군집분석

④ 중심연결법 (Centroid Linkage Method)

- 중심연결법에서는 두 군집 P와 Q간의 거리가 P에 속한 개체와 Q에 속한 개체의 중심점 간의 거리



〈그림7〉 중심연결법에서의 두 군집 간 거리

3. 계층적 군집분석

2) 분할분석(Divisive Analysis)

- 개체의 수가 많을 경우 계산시간이 오래 걸린다는 응집분석의 단점을 보완하기 위하여 Macnaughton-smith(1964)가 분할분석을 제안
 - 대표적인 분할분석으로는 다음에 요약된 Kaufman과 Rousseeuw(1990)의 DIANA(Divisive Analysis)알고리즘을 꼽을 수 있음

3. 계층적 군집분석

1) 분할분석(Divisive Analysis)

- ① 군집 Q에 속한 각 개체에서 다른 개체까지의 평균거리를 계산하여 평균 거리가 가장 긴 개체를 선택. 이 개체를 군집에서 제거하여 군집 A(new)를 만들고 제거된 개체를 갖는 군집 B(new)라 정의
 - ② 군집 A(new)와 B(new)를 군집 A(old)와 B(old)로 정함
 - ③ 군집 A(old)에서 다른 개체와의 평균거리가 가장 긴 개체를 찾아 이 개체를 i라 정의
 - ④ ③에서 찾은 개체 i와 군집 B(old)에 속한 개체들 간에 평균거리를 계산
 - ⑤ 개체 i와 군집 A(old)에서, 다른 개체의 평균거리에서 개체 i와 군집 B(old)의 개체의 평균거리를 뺀 값이 0보다 크면 개체 A(old)에서 i를 제거한 군집 A(new)를 생성. 그리고 B(old)에서 i를 추가한 군집 B(new)를 생성. 이후 ②-④번 과정을 계속 반복.
- 만약 거리의 차가 0 이하면 군집형성을 중지하고 A(old)와 B(old)를 최종 군집으로 선택.

3. 계층적 군집분석

〈예제 7-5〉

개체	X1(변수1)	X2(변수2)
1	1	1
2	1	2
3	3	4
4	5	5
5	7	5.5

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 0 & 1 & 5 & 8 & 10.5 \\ 1 & 0 & 4 & 7 & 9.5 \\ 5 & 4 & 0 & 3 & 5.5 \\ 8 & 7 & 3 & 0 & 2.5 \\ 10.5 & 9.5 & 5.5 & 2.5 & 0 \end{pmatrix} \end{matrix}$$

개체	다른 개체와의 평균 거리
1	$(1 + 5 + 8 + 10.5)/4 = 6.125$
2	$(1 + 4 + 7 + 9.5)/4 = 5.375$
3	$(5 + 4 + 3 + 5.5)/4 = 4.375$
4	$(8 + 7 + 3 + 2.5)/4 = 5.125$
5	$(10.5 + 9.5 + 5.5 + 2.5)/4 = 7$

가장 먼 개체 : 5번

$A(\text{new}) = \{1, 2, 3, 4\}$

$B(\text{new}) = \{5\}$

▷ $A(\text{old}) = A(\text{new})$,

$B(\text{old}) = B(\text{new})$ 로 함

3. 계층적 군집분석

개체	{1,2,3,4}에서 다른 개체와의 거리	{5}와의 거리	거리의 차이
1	$(1+5+8)/3=4.67$	10.5	-5.83
2	$(1+4+7)/3=4$	9.5	-5.5
3	$(5+4+3)/3=4$	5.5	-1.5
4	$(8+7+3)/3=6$	2.5	3.5

개체	{1,2,3}에서 다른 개체와의 거리	{4,5}와의 거리	거리의 차이
1	$(1+5)/2=3$	$(8+10.5)/2=9.25$	-6.25
2	$(1+4)/2=2.5$	$(7+9.5)/2=8.25$	-5.75
3	$(5+4)/2=4.5$	$(3+5.5)/2=4.25$	0.25

3. 계층적 군집분석

개체	{1,2}에서 다른 개체와의 거리	{3,4,5}와의 거리	거리의 차이
1	1	$(5+8+10.5)/3=7.83$	-6.83
2	1	$(4+7+9.5)/3=5.83$	-5.83

개체	다른 개체와의 평균 거리
3	$(3+5.5)/2=4.25$
4	$(3+2.5)/2=2.75$
5	$(5.5+2.5)/2=4$

개체	{4,5}에서 다른 개체와의 거리	{3}과의 거리	거리의 차이
4	2.5	3	-0.5
5	2.5	5.5	-3

4. 비계층적 군집분석

4. 비계층적 군집분석

- 계층적 군집분석은 관찰치의 수가 적은 경우에 적당하지만 대용량의 데이터를 가지고 군집분석을 실시할 때에는 비계층적 군집분석 방법인 K-평균 군집분석을 사용

- K-평균 군집분석은 매퀸(1967)이 제안한 것으로 계층적 군집분석과 달리 군집수를 미리 정하고 분석을 실행

- ① 군집의 수 K를 설정
- ② 임의의 K개 관찰치를 K개 각 군집에 임의로 지정. 이를 K개 각 군집의 중심으로 이용
- ③ 모든 관찰치를 군집중심으로부터 유클리디안 거리가 최소인 군집에 귀속
- ④ 각 군집에 속한 관찰치들을 이용하여 군집중심을 새로 계산
- ⑤ 변화(군집간 관찰치이동)가 없을 때까지 단계3 과 단계4 를 반복

4. 비계층적 군집분석

- K-평균 군집분석 군집의 수 결정과 초기값 설정이 중요
 - 군집의 수를 결정하기 위해서는 각 군집 수와 산출된 평균 거리를 이용할 수 있음
 - 초기값 설정을 위한 방법은
 - 1) 다양한 초기값을 가지고 주어진 군집 수에 대하여 K-평균 군집화를 수행하고 최종 군집에서 중심점까지의 평균거리를 구하여 이들 중 가장 작은 평균거리를 갖는 초기값에 대한 군집을 선택하거나
 - 2) 계층적 군집화를 시행하고 계층적 군집화로부터 군집의 수와 형성된 군집으로부터 중심점을 구한 후, 이 결과를 가지고 K-평균 군집화를 수행, 또는
 - 3) 주성분을이용

5. 군집분석의 특징

5. 군집분석의 특징

- 군집분석은 자료 사이의 거리를 이용하여 수행되기 때문에 자료의 단위가 결과에 큰 영향을 미치므로 자료를 표준화하는 경우가 있음
 - 각 변수의 중요도가 다를 경우 가중치를 이용하여 각 변수의 중요도를 조절. 가중치는 대부분의 경우 단위변환(표준화)를 수행한 후 부여
- 각 변수의 중요도가 다를 경우 가중치를 이용하여 각 변수의 중요도를 조절. 가중치는 대부분의 경우 단위변환(표준화)를 수행한 후 부여
 - 장점은 군집분석이 탐색적인 기법으로 주어진 자료에 대한 사전정보 없이 의미 있는 자료구조를 찾아낼 수 있다는 것이고 다양한 형태의 데이터에 적용가능하며 분석방법의 적용이 쉽다는 점
 - 단점은 복잡한 자료에 대해서는 유의미한 군집을 찾기가 힘들다는 것인데, 군집이 이상치에 영향을 받기 때문. 변수의 개수가 많은 경우에도 좋은 군집을 찾기 힘든 경우가 많으며 가중치와 거리 정의가 어렵고 초기 군집수 k 의 결정이 어려움 결과의 해석이 어렵다는 단점도 존재

The background is a vibrant abstract composition featuring various shades of blue and purple. It includes large, soft-edged organic shapes, several circles with diagonal hatching patterns, and smaller circles with halftone dot patterns. A central white rounded rectangle contains the text.

강의를 마쳤습니다.
다음시간에는...