

8강. Clustering Analysis

◆ 담당교수 : 김 동 하

■ 학습개요

이번 강의에서는 비지도 학습 중에 하나인 군집 분석에 대해서 배우도록 한다. 계층적 군집화를 위해 사용되는 계층적 군집 분석법, 그리고 비계층적 군집화를 위해 사용되는 K-평균 군집법에 대해 다룬다.

■ 학습목표

1	거리 혹은 비유사도의 개념에 대해 학습한다.
2	계층적 군집분석에 대해 학습한다.
3	K-군집 분석법에 대해 학습한다.

■ 주요용어

용어	해설
최장,최단 연결법	병합을 이용한 계층적 군집 분석법의 일종으로, 병합시 군집과 군집 사이의 거리를 샘플들 사이의 최장 거리 혹은 최단 거리로 사용한다.
덴드로그램	각 단계에서 관측치의 군집화를 통해 형성된 그룹과 이들의 유사성 수준을 표시하는 나무 형태의 다이어그램.
K-평균 군집법	대표적인 비계층적 군집 분석 방법론으로 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다.

■ 학습하기

01. 군집분석

군집분석의 정의

- 군집분석은 모집단 또는 범주에 대한 사전 정보가 없는 경우에 사용 -> 비지도 학습법

- 주어진 관측값들 사이의 거리 또는 유사성을 이용하여 전체를 몇 개의 집단으로 그룹화
- 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 것이 목표

군집화

- 군집화의 기준
 - > 동일한 군집에 속하는 개체는 비슷하고, 다른 군집에 속한 개체는 그렇지 않도록 군집을 구성
- 군집화를 위한 변수
 - > 예: 고객 세분화
 - > 인구통계적 변수 (성별, 나이, 거주지, 소득 등)
 - > 구매패턴 변수 (구매상품, 주기, 거래액 등)

군집분석의 활용

- 고객 세분화
 - > 고객이 기업의 수익에 기여하는 정도를 통한 고객 세분화
 - > 개별 고객에 대한 맞춤 관리
- 고객의 구매패턴에 따른 고객 세분화
 - > 신상품 판촉, 교차판매를 위한 표적 집단 구성

군집분석의 유형

- 상호 배반적 (disjoint) 군집
 - > 각 관측치가 오직 하나의 군집에만 속함
 - > 예: 한국인, 중국인, 일본인, ...
- 계층적 (hierarchical) 군집
 - > 한 군집이 다른 군집의 내부에 포함
 - > 군집간의 중복은 없으며, 군집들이 매 단계 계층적인 구조를 이룸.
 - > 예: 전자제품 -> 주방용 -> 냉장고
- 중복 (overlapping) 군집
 - > 두 개 이상의 군집에 한 관측치가 동시에 포함.

군집분석의 특징

- 군집 분석을 위한 기준의 설정에 있어서 다양한 방법이 존재.
 - > 유사성 혹은 비유사성의 다양한 정의
- 분석자의 주관에 따라 결과가 달라질 수 있음.

02. 거리

비유사성의 척도: 거리

- 관측값들이 서로 얼마나 유사한지 측정할 수 있는 척도가 필요
- 군집분석에서는 비유사성을 기준으로 하여 유사정도를 측정.

연속형 자료의 거리 측도

- 유클리디안 거리 (Euclidean distance) : p 차원 공간에서 주어진 두 점 $x = (x_1, \dots, x_p)$ 와 $y = (y_1, \dots, y_p)$ 사이의 유클리디안 거리는 다음과 같이 정의된다:

$$d(x, y) = \left[\sum_{i=1}^p (x_i - y_i)^2 \right]^{1/2}$$

- 민코프스키 거리 (Minkowski distance) : 양수 $q > 0$ 에 대해서 p 차원의 두 점 $x = (x_1, \dots, x_p)$ 와 $y = (y_1, \dots, y_p)$ 사이의 민코프스키 거리는 다음과 같이 정의된다:

$$d(x, y) = \left[\sum_{i=1}^p (x_i - y_i)^q \right]^{1/q}$$

- 표준화 거리 : p 차원의 두 점 $x = (x_1, \dots, x_p)$ 와 $y = (y_1, \dots, y_p)$ 사이의 표준화 거리는 다음과 같이 정의된다:

$$d(x, y) = \left[\sum_{i=1}^p \left(\frac{x_i - y_i}{S_i} \right)^2 \right]^{1/2}$$

범주형 자료의 거리 측도

- 불일치 항목 수

-> 예: 세 개체 A, B, C 사이의 거리는 다음과 같이 계산할 수 있다:

개체	성별	학력	출신지역
A	남자	고졸	경기
B	여자	고졸	전남
C	남자	대졸	경기

-> $d(A, B) = 2$, $d(A, C) = 1$, $d(B, C) = 3$

03. 계층적 군집 분석

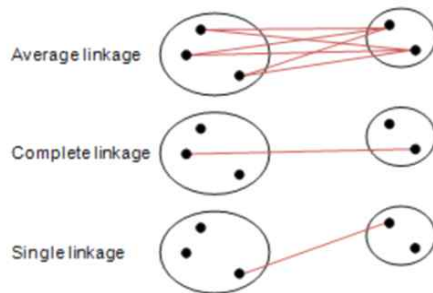
계층적 군집 분석

- 가까운 관측값들 끼리 묶는 병합 방법과 먼 관측값들을 나누어가는 분할 방법으로 나눌 수 있다.
- 계층적 군집분석에서는 주로 병합 방법이 주로 사용됨.
- 나무 구조인 덴드로그램을 통해 나타낼 수 있음.

병합 방법

- 처음에 n 개의 자료를 각각 하나의 군집으로 생각한다 (n 개의 군집).
- 가장 거리가 가까운 두 개의 군집을 병합하여 $n-1$ 개의 군집으로 군집을 줄인다.
- 그 다음, $n-1$ 개의 군집 중 가장 가까운 두 군집을 병합하여 군집을 $n-2$ 개로 줄인다.

- 이를 반복하여 계속하여 군집의 수를 줄여 나간다.
- 이 과정은 시작부분에는 군집의 크기는 작고 동질적이며, 끝부분에서는 군집의 크기는 커지고 이질적이 된다.
- 군집들간의 거리를 측정하는 방법에 따라 다양한 종류의 방법이 있다.
-> 최단거리, 최장거리, 평균거리 방법 등.



최단 연결법 (Single linkage method)

- 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최소값으로 측정
- 군집이 고리 형태로 연결되어 있는 경우에는 부적절한 결과를 제공.
- 고립된 군집을 찾는데 중점을 둔 방법.
- 주어진 5개의 관측값에 대한 거리(비유사성) 행렬에 대하여 최단연결법으로 군집을 얻고 덴드로그램으로 나타내보자.

1	0				
2	7	0			
3	1	6	0		
4	9	3	8	0	
5	8	5	7	4	0

- 1단계
-> 거리 행렬에서 $d(1,3)=1$ 이 최소이므로 관측값 1과 3을 묶어 군집 (1,3)을 만든다.
- 군집 (1,3)과 관측값 2,4,5와의 거리

$$d((2), (1,3)) = \min\{d(2,1), d(2,3)\} = d(2,3) = 6$$

$$d((4), (1,3)) = \min\{d(4,1), d(4,3)\} = d(4,3) = 8$$

$$d((5), (1,3)) = \min\{d(5,1), d(5,3)\} = d(5,3) = 7$$
 -> 이를 이용하여 다음과 같은 거리 행렬을 만들 수 있다.
- 2단계
-> $d(2,4) = 3$ 이 최소이므로 관측값 2와 4를 묶어 군집 (2,4)를 만들자.

(1,3)	0			
2	6	0		
4	8	3	0	
5	7	5	4	0

-> 군집 (2,4), (1,3), (5) 사이의 거리행렬 구하기.

$$d((2,4), (1,3)) = \min\{d((2), (1,3)), d((4), (1,3))\} = d((2), (1,3)) = 6$$

$$d((5), (2,4)) = \min\{d(5,2), d(5,4)\} = d(5,4) = 4$$

(1,3)	0		
(2,4)	6	0	
5	7	4	0

- 3단계

-> $d((5), (2,4)) = 4$ 이 최소값을 가지므로 군집 (2,4)와 (5)를 묶어 (2,4,5)를 만든다.

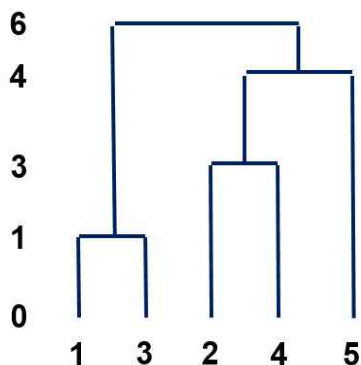
-> $d((1,3), (2,4,5)) = d(2,3) = 6$ 으로 이용하여 다음의 거리 행렬을 얻는다.

(1,3)	0	
(2,4,5)	6	0

- 4단계

-> 마지막으로 전체가 하나의 군집을 이룬다.

- 덴드로그램 결과



최장(완전) 연결법 (Complete linkage method)

- 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의

최대값으로 측정.

- 같은 군집에 속하는 관측치는 알려진 최대 거리보다 짧다.
- 군집들의 내부 응집성에 중점을 둔다.

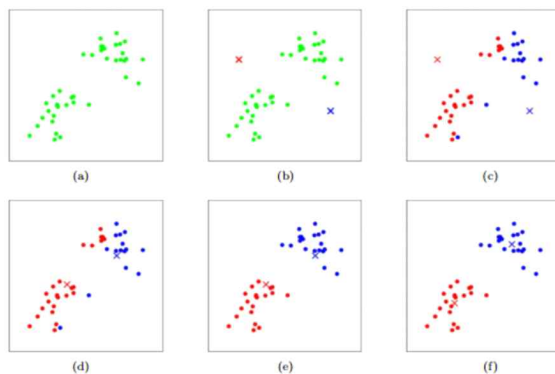
04. 비계층적 군집 분석

비계층적 군집 분석

- 비계층적 군집분석에는 흔히 관측값들을 몇 개의 군집으로 나누기 위하여 주어진 판정기준을 최적화 한다.
- 따라서, 최적분리 군집분석이라고도 한다.
- 대표적인 비계층적 군집분석 방법
 - > K-평균 방법.

K-평균 군집법

- 사전에 결정된 군집수 K에 기초하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 구분.
- 계층적 군집법에 비하여 적은 계산량.
 - > 대용량 데이터를 빠르게 처리할 수 있음.
- 알고리즘:
 1. 군집수 K를 결정한다.
 2. 초기 K개 군집의 중심을 선택한다.
 3. 각 관측치를 그 중심과 가장 가까운 거리에 있는 군집에 할당한다.
 4. 위의 과정을 기존의 중심과 새로운 중심의 차이가 없을 때까지 2번과 3번을 반복한다.



군집 수의 결정

- K-평균 군집분석법의 결과는 군집수 K의 결정에 민감하게 반응.
- 실제 자료의 분석에서는 여러 가지의 K값에 대해 군집분석을 수행한 후 가장 좋다고 생각되는 K값을 이용.
- 일반적으로 K값이 증가함에 따라 거리제곱합은 감소.
 - > 자료의 단위가 큰 영향을 미치기 때문에 표준화 작업이 먼저 수행되어야 함.
- K값을 증가시키면서 거리제곱합의 감소가 완만해지는 군집의 수를 최적의 군집 수로 결정

K-평균 군집 분석의 단점

- 군집이 겹치는 경우에 좋지 않음.
- 이상치에 민감

■ 연습문제

(단답형)1. 군집 유형 중 하나로, 각 관측치가 오직 하나의 군집에만 속하는 것을 무엇이라 하는가?

정답 : 상호 배반적 군집

해설 : 군집 유형에는 상호 배반적 군집, 계층적 군집, 중복 군집 등이 있으며, 관측치가 오직 하나의 군집에만 속하는 것은 상호 배반적 군집이라고 한다.

(객관식)2. 다음 중 비계층적 군집 분석법에 대해 잘못 설명한 것을 고르시오.

- ① 범주형 자료 사이의 비유사성 척도로 불일치 항목 수를 많이 사용한다.
- ② 가까운 관측값들 끼리 묶는 병합 방법과 먼 관측값들을 나누어가는 분할 방법이 존재한다.
- ③ 계층적 군집 분석에서는 주로 분할 방법을 많이 사용한다.
- ④ 군집 분석 결과를 덴드로그램을 통해 나타낼 수 없다.

정답) ③

해설) 계층적 군집 분석에서는 주로 병합 방법을 많이 사용한다.

(O/X)3. K-평균 군집법에서 일반적으로 군집의 수 K가 증가함에 따라 거리제곱합은 감소한다.

정답 : O

해설 : 군집의 수가 증가할수록 거리제곱합은 감소한다.

■ 정리하기

1. 군집 분석은 주어진 관측값들 사이의 거리 또는 유사성을 이용하여 전체를 몇 개의 집단으로 그룹화하는 분석법을 의미한다.

2. 계층적 군집 분석은 보통 병합 방법을 주로 사용하며, 최단거리, 최단거리, 평균거리 방법 등을 사용할 수 있다.
3. K-평균 군집법은 사전에 정의된 군집 수 K 를 이용해 각 군집별 분산을 최소화하는 방향으로 군집과 군집별 평균을 계산하는 방법이다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

박창이, 김용대, 김진석, 송종우, 최호식. 『R을 이용한 데이터마이닝』. 서울:교우사, 2018.