organized_02

Hyeonho Lee

2018년 10월 7일

Contents

세그멘테이션(다차원 척도법, 군집분석)

다차원 척도법 Multidimensional Scaling(MDS)

개요

1. 다차원 관측값 또는 개체들 간의 거리 또는 비유사성을 이용하여 개체들을 원래의 차원보다 낮은 차원(2차원 or 3차원)의 공간상의 점으로 표현(spatial configuration)하는 통계적 분석방법
2. 목적 : 차원의 축소를 통해 개체들 사이의 관계를 쉽게 파악
   예제) 정치 후보자, 소비자 제품들의 성향에 대한 구조를 파악하고자 할 때, 이들 개체들의 특성을 측정한 후에 개체들의 거리 또는 비유사성을 구한 뒤, 이들 개체들을 2차원 또는 3차원 공간상에 표현하여 개체들 사이의 관계를 파악하는데 이용
3. MDS구분
   1) 메트릭 MDS(metric MDS) : 등간척도나 비율척도 자료에 근거하여 비유사성 이루어지는 경우
   2) 넌메트릭 MDS(nonmetric MDS) : 순서척도 자료에 근거하여 비유사성 측정 되는 경우
      * metric(측정기준에 의해 발생하는 총 카운트)
   3) 적합성
      (1) kruskal의 STRESS or S-STRESS : 공간상의 표현이 주어진 비유사성에 어느 정도 적합한가를 측정하는 기준
   4) 최적모형의 적합
      (1) 부적합도 : STRESS or S-STRESS 이용 각 개체들을 공간상에 표현

$$STRESS = \sqrt{\frac{\sum_{i<j}(S_{ij} - \hat{S_{ij}})^2}{\sum_{i<j} S_{ij}^2}}$$

$$S - STRESS = \sqrt{\frac{\sum_{i<j}(S_{ij} - \hat{S_{ij}})^2}{\sum_{i<j}(S_{ij}^2)^2}}$$

      (2) $\hat{S_{ij}}$: 측정모형에서 구한 S_{ij}의 적합값
      (3) 부적합도를 최소로 하는 방법으로 반복알고리즘을 이용하게 적합
      (4) 부적합도 값 일정한 수준 이하로 될 때 최종적으로 적합된 모형으로 제시
      (5) 부적합도 값은 0과 1사이의 값을 취한다.(0에 가까울수록 적합된 모형이 적절하다고 판단)
      (6) STRESS $\geq$ 0.10 : STRESS의 크기가 적정 수준이 될 때까지 차원을 높인다. 그러나 표현 공간이 커질수록 STRESS는 작아지지만 결과의 해석이 복잡하다.
      (7) 일반적으로 2차원 또는 3차원 정가 적당하다.

Stress	적합도 수준
0	완벽 (Perfect)
0.05 이내	매우좋음 (Excellent)
0.05 - 0.10	만족 (Satisfactory)
0.10 - 0.15	보통 (Acceptable, but doubt)
0.15 이상	나쁨 (Poor)

Table 1: Strees에 따른 적합도 수준

Cluster Analysis(군집 분석)

1 정의

1. 군집분석은 모집단 또는 범주에 대한 사전 정보가 없는 경우에 주어진 관측값들 사이의 거리 또는 유사성을 이용하여 전체를 몇 개의 집단으로 그룹화하여 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 분석법이다.

2 군집화

1. 기준 : 동일한 군집에 속하는 개체(또는 개인)는 여러 속성이 비슷하고 서로 다른 군집에 속한 관찰치는 그렇지 않도록 군집을 구성
2. 군집화를 위한 변수 : 전체 개체(개인)의 속성을 판단하기 위한 기준
   예시) 고객세분화 : 인구통계적 변인(성별, 나이, 거주지, 직업, 소득, 교육, 종교, …)
   구매패턴 변인(상품, 주기, 거래액, …)
   생활패턴 변인(라이프스타일, 성격, 취미, 가치관, …)

3 활용

1. 고객세분화
   1) 고객이 기업의 수익에 기여하는 정도를 통한 고객 세분화
   2) 우수고객의 인구통계적 요인, 생활패턴 파악, 개별고객에 대한 맞춤관리
   3) 고객의 구매패턴에 따른 고객세분화
   4) 신상품 판촉, 교차판매를 위한 표적집단 구성

4 비유사성의 척도 : 거리

1. 군집분석에서는 관측값들이 서로 얼마나 유사한지 또는 유사하지 않은지를 측정 할 수 있는 측도가 필요하다.
2. 군집분석에서는 보통 유사성보다는 비유사성을 기준으로하며 거리를 사용한다.

5 거리 측도의 종류들

1. 유클리드(Euclid) 거리
2. Minkowski 거리
3. 표준화거리
4. Mahalanobis 거리
5. 범주형 자료의 거리(불일치 항목 수)
6. Symbolic String 사이의 거리

6 군집분석의유형 및 특징

1. 상호배반적(disjoint) 군집 : 각 관찰치가 상호배반적인 여러 군집 중 오직 하나에만 속함
   예시) 한국인, 중국인, 일본인
2. 계보적(hierarchical) 군집 : 한 군집이 다른 군집의 내부에 포함되는 형태로 군집간의 중복은 없으며, 군집들이 매 단계 계층적인 (나무)구조를 이룬다.
   예시) 전자제품 -> 주방용 -> 냉장고
3. 중복(overlapping) 군집 : 두 개 이상의 군집에 한 관찰자가 동시에 포함되는 것을 허용
4. 퍼지(fuzzy) 군집
   1) 관찰치가 소속되는 특정한 군집을 표현하는 것이 아니라, 각 군집에 속할 가능성을 표현
   2) Pr(개체가 군집A에 속함) = 0.7, Pr(개체가 군집B에 속함) = 0.3
5. 군집분석은 그 기준의 설정, 즉 유사성이나 혹은 비유사성의 정의나 군집의 형태 등 매우 다양한 방법이 있다. 군집분석은 자료의 사전정보 없이 자료를 파악하는 방법으로, 분석자의 주관에 결과가 달라질 수 있다. 따라서, 군집분석은 한번에 분석이 끝나는 것이 아니고, 매회 결과를 잘 관찰하여 의미 있는 정보요약을 얻어내야 한다. 특이값을 갖는 개체의 발견, 결측값의 보정 등에 군집분석이 사용될 수 있다. 군집분석에서 군집을 분석하는 중요한 변수의 선택이 중요하다.

7 계층적 군집분석

1. 개요 : 가까운 관측값들 끼리 묶는 병합방법과 먼 관측값들을 나누어가지는 분할방법으로 나눌 수 있다. 계층적 군집분석에서는 주로 병합 방법이 주로 사용된다. 계층적 군집분석의 결과는 나무구조인 덴드로그램을 통해 간단하게 나타낼 수 있고, 이를 이용하여 전체 군집들간의 구조적 관계를 쉽게 살펴볼 수 있다.
2. 병합방법
    1) 1단계 : 처음에 n개의 자료를 각각 하나의 군집으로 생각한다. 즉 군집의 수는 n이다.
    2) 2단계 : 이 n개의 군집 중 가장 거리가 가까운 두개의 군집을 병합하여 n-1개의 군집으로 군집을 줄인다.
    3) 3단계 : n-1개의 군집 중 가장 가까운 두 군집을 병합하여 군집을 n-2개로 줄인다.
    4) 이를 반복한다. 이 과정은 시작부분에는 군집의 크기는 작고 동질적이며, 끝부분에서는 군집의 크기는 커지고 이질적이 된다.
3. 거리측정방법
    1) 최단거리(최단연결법, Single Linkage Method) : 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최소값으로 측정한다. 유리 위에 떨어진 물방울들이 서로 뭉치는 현상과 비슷하다. 같은 군집에 속하는 관측치는 다른 군집에 속하는 관측치에 비하여 거리가 가까운 변수를 적어도 하나는 갖고있다. 군집이 고리형태로 연결되어 있는 경우에는 부적절한 결과를 제공한다. 고립된 군집을 찾는데 중점을 둔 방법이다.
    2) 최장연결법(완전연결법, Complete Linkage Method) : 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최대값으로 측정한다. 같은 군집에 속하는 관측치는 알려진 최대 거리보다 짧다. 군집들의 내부 응집성에 중점을 둔다.

8 비계층적 군집분석

1. 개요 : 비계층적 군집분석에서는 흔히 관측값들을 몇 개의 군집으로 나누기 위하여 주어진 판정을 최적화 한다. 따라서, 최적분리 군집분석이라고 한다. 대표적인 비계층적 군집분석 방법이 k-means method가 있다.
2. k-평균 군집방법 : 사전에 결정된 군집수 k에 기초하여 전체 데이터를 상대적으로 유사한 k개의 군집으로 구분한다. k-means clustering은 계보적 군집법에 비하여 계산량이 적다. 따라서, 대용량 데이터를 빠르게 처리할 수 있다.
3. k-평균 군집방법 알고리즘
    1) 군집수 k를 결정한다.
    2) 초기 k개 군집의 중심을 선책한다.
    3) 각 관찰치를 그 중심과 가장 가까운 거리에 있는 군집에 할당한다.
    4) 위의 과정을 기존의 중심과 새로운 중심의 차이가 없을 때까지 반복한다.
4. k-평균 군집방법의 초기군집수정 결정 : k-평균 군집분석법의 결과는 초기 군집수 k의 결정에 민감하게 반응한다. 실제 자료의 분석에서는 여러 가지의 k값을 선택하여 군집분석을 수행한 후 가장 좋다고 생각되는 k값을 이용한다. 여러 개의 군집분석 결과 중 어떤 결과가 좋은가 하는 문제는 관측값 간의 평균 거리와 군집간의 평균거리를 비교함으로써 수행한다. 가장 좋은 방법은 자료의 시각화를 통한 최적 군집수의 결정인데, 자료의 시각화를 위하여는 차원의 축소가 필수적이고, 이를 위하여 주성분 분석방법이 널리 사용된다. 시각화가 어려운 경우에는 , 여러 가지 통계량을 사용하는데, 예를 들면, 각 그룹의 산포행렬의 행렬식을 최소로 하는 군집수를 찾는다.
5. k-평균방법의 단점
    1) 군집이 겹치는 경우에 좋지 않다.
    2) 이상치에 민감하다.
    3) 각 관찰치가 할당된 군집에 속하지 않을 불확실성에 대한 측정치가 없다.
    - 단점 극복 방법 : 가우스 혼합모형(Gaussian mixture model)
6. 가우스 혼합모형(Gaussian mixture model)
    1) 주어진 군집수 k에 대하여, 각 군집의 관측치의 분포가 미지의 평균과 분산을 따르는 정규분포라고 가정한다.
    2) 자료를 가장 잘 분리할 수 있는 최적의 평균과 분산을 추정과 최대화의 두 단계를 반복하여 구한다.
    3) 결과물로는 각 관찰치에 대하여 그 관찰치가 각 군집에 속할 확률이 계산된다.
    4) 마지막으로, 각 관찰치는 가장 높은 확률을 갖는 집단으로 할당된다.
    5) 이러한 방법을 소프트 군집화(soft clustering)이라 한다.

9 단위변환, 가중치 부여

1. 단위변환
    1) 군집분석은 자료 사이의 거리를 이용하여 수행되기 때문에, 각 자료의 단위가 결과에 큰 영향을 미친다.
    예시) (x,y,z) 세 개의 변수가 어떤 거리를 측정하였다고 했을 때, 그 단위가 x는 야드, y는 센티미터, z는 마일로 측정되었다면, 그 거리의 계산에 유의 하여야 한다. z의 단위 1의 차이는 y의 단위 185,200의 차이와 같고, x의 2,025

와 같다. 만약 서로 다른 종류의 측정치로 자료가 구성되어 잇으면, 위와 같은 상대적인 평가도 불가능 하다. 즉 data normailzation이 필요하다.

2) 위의 문제를 해결하기 위해 표준화 방법을 많이 사용한다. 표준화 방법이란 각 변수의 관찰값으로부터 그 변수의 평균을 빼고, 그 변수의 표준편차로 나누는 것이다. 표준화된 자료는 모든 변수가 평균이 0이고 표준편차가 1이 된다. 표준화된 자료의 유클리드거리는 표준화거리와 같다.

2. 가중치 부여
   1) 자료의 분석 전에 각 변수의 중요도가 같지 않음을 안다면, 적절한 가중치를 이용하여 각 변수의 중요도를 조절할 수 있다.
      예시) 같은 수입을 가지는 두 가족이 같은 대지면적을 가지는 두 가족보다 공통점이 많다고 생각이 드는 경우가 있다. 이 경우에는, 수입 변수에 높은 가중치를 주고 대지면적 변수에 낮은 가중치를 줌으로써 해결한다. 가중치는 대부분의 경우 단위변환(표준화)를 수행한 우 부여한다. 가중치의 대한 군집의 영향을 평가 하기 위하여는 여러 가지의 가중치에 대하여 군집분석의 결과를 구하고 이 결과들을 비교한다.

10 군집평가 및 변수선택

1. 군집평가
   1) 군집분석에는 분석 전에 정해야 하는 사항이 많다.
      예시) 초기군집수, 가중치 등
   2) 분석자의 주관에 의하여 결정되는 이러한 사항들이 군집분석의 결과에 어떻게 영향을 미치는 가를 알아보기 위하여는, 군집분석 평가가 필수적이다.
   3) 좋은 결과는 각 군집 안에서의 분산이 최소로 되는 것이다. 또는 사용되어진 거리의 측도를 이용하여 군집내의 거리의 평균과 군집간의 거리의 평균을 비교할 수 있다.
   4) 즉, 군집내의 거리의 평균이 군집간의 거리의 평균 보다 작으면 좋은 결과이다.
   5) 군내 변동은 작고, 군간 변동은 크면 좋다는 이야기 이다.
2. 변수선택
   1) 찾아진 각 군집은 어떠한 변수에 의하여 군집이 형성됐는가를 파악하는 것을 목적으로 한다.
   2) 각 변수에 대한 그룹내의 거리의 평균과 그룹간의 거리의 평균을 측정한다.
   3) 그룹내의 거리가 그룹간의 거리에 비하여 아주 작은 변수가, 그 군집을 형성하는 데 크게 기여하는 변수이다.
      예시) 군집분석을 수행한 겨로가 특정한 군집에는 소득이 비슷한 사람들이 많이 모여 있음을 알 수 있따. 이를 통하여, 소득이 자료의 패턴에 큰 영향을 주는 것을 확인 할 수 있다.
   4) k-means는 스토리라인을 만들기가 어렵다. 그러므로 사용을 지향하는 것이 좋다. 또한 군집분석으로 변수선택을 하기엔 너무 어렵다. serious한 변수선택을 하고 분석을 하는 것이 좋고, 전문가와 상의하는 것을 추천한다.

11 자기영상 군집분석

1. 군집분석에서 오직 하나만의 군집이 존재하는 경우에 아주 유용하게 사용될 수 있다.
   예시) 모터제조 공장에서의 모터의 불량원인을 알려고 할 때, 정상적인 모터의 자료를 이용하여 군집분석을 수행하면, 하나의 군집이 찾아진다. 이 때, 새로운 모터와 이 군집과의 거리가 크면, 이 새로운 모터를 불량모터라고 의심할 수 있다. 또 다른 예로는, 위조지폐 탐지가 있다.

12 군집분석의 장단점

1. 장점
   1) 탐색적인 기법 : 자료의 내부구조에 대한 사전정보 없이 의미 있는 자료구조를 찾아낼 수 있다.
   2) 다양한 형태의 데이터에 적용가능 : 거리만 잘 정의되면, 모든 종류의 자료에 적용할 수 있다.
      예시) 신문기사와 같은 텍스트 자료도 그 거리만 잘 정의하면 얼마든지 군집분석을 사용 할 수 있다.
   3) 분석방법의 적용 용이성 : 자료의 사전정보를 필요로 하지 않아서 누구나 쉽게 분석가능
2. 단점
   1) 가중치과 거리 정의 : 가중치와 거리를 어떻게 정의하는가에 따라 분석의 결과가 민감하게 반응.
   2) 초기 군집수 k의 결정이 쉽지 않다.
   3) 결과의 해석이 어렵다. 특히, 찾아진 군집이 무엇을 의미 하는지 데이터만으로는 알 수 없다.