organized_02

Hyeonho Lee

2018년 10월 7일

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
- 목적
  1. 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
- 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)
- 적합성
  1. kruskal의 STRESS or S-STRESS : 공간상의 표현이 주어진 비유사성에 어느 정도 적합한가를 측정하는 기준
- 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_ij의 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.

Table 1: Strees에 따른 적합도 수준

Stress	적합도수준
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)

Cluster Analysis(군집 분석)

- 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.
- 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, …)
- 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
- 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.
- 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
- 군집분석의유형 및 특징
  1. 상호배반적(disjoint) 군집 : 각 관찰치가 상호배반적인 여러 군집 중 오직 하나에만 속함 - 예) 한국인, 중국인, 일본인
  2. 계보적(hierarchical) 군집 : 한 군집이 다른 군집의 내부에 포함되는 형태로 군집간의 중복은 없으며, 군집들이 매 단계 계층적인 (나무)구조를 이룬다. - 예) 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 잇는 정보요약을 얻어내야 한다. 특이값을 갖는 개체의 발견, 결측값의 보정 등에 군집분석이 사용될 수 있따. 군집분석에서 군집을 분석하는 중요한 변수의 선택이 중요하다.
- 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)
        (1) 1단계 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.