Chapter 4
Optimal Clustering

1

목차

1 Motivation

- 대상을 해당 특징을 지닌 그룹으로 나누고, 내부적 동질성을 최대화되며, 외부적 동질성은 최소화하는 방법을 말함
- 금융 시장에서 클러스터링은 중요함
  - 상대 지표를 보기 위해 peer를 선정할 때에도 사용

2 Proximity Matrix

- **모델**
  - $\boldsymbol{X}(\boldsymbol{N} \times \boldsymbol{F})$
    - N : # of objects
    - F : # of features
  - **$\boldsymbol{N}$개의 objects의 $\boldsymbol{F}$를 통해 proximity를 계산할 수 있음**
    - $N \times N$ 행렬 생성
    - Proximity measures
      - Correlation
      - Mutual information
      - Distance Metric
    - Standardize the input data를 해야 한다
      - Scale이 큰 하나의 변수가 다른 값을 잡아먹을 수 있음

3

Types of Clustering

- Types
  - Partitional
    - One-level partitioning
    - 서로 배타적
  - Hierarchical
    - Partition을 반복적으로 진행
    - 분할적일수도, 응집적일수도 있음

3 Types of Clustering

- Connectivity
  - Distance connectivity(hierarchical)
- Centroids
  - Vector quantization(k-means)
- Distribution
  - Statistical Distributions
- Density
  - Search for connected dense region in the data space
  - DBSCAN , OPTICS
- Subspace
  - Cluster는 2차원(features, observation)으로 나뉨
  - Biclustering(coclustering) 진행 가능
    - cluster observations and features simultaneously.

3

Types of Clustering

- 문제상황
  - # of Freatures exceeds # of observation
    - Curse of Dimensionality
    - Most of the space spanning the observation will be empty
- 해결
  - Project the data matrix onto a low-dimensional space
    - PCA
  - Project the proximity matrix onto a low-dimensional space
    - 해당 값을 new X matrix로 사용

4 Number of Clusters

- 문제상황
  - Partitioning algorithms
    - 연구자들이 correct # of clusters를 제시해줘야 함
- 해결
  - Elbow method
    - Marginal percentage of variance가 predefined threshold를 넘어서지 못할 때 멈춤
    - Percentage of Variance Explained : $\frac{\boldsymbol{between\ group\ variance}}{\boldsymbol{Total\ Variance}}$: $\boldsymbol{F-test}$
  - Optimal Number of Clusters
    - Silhouette method
      - Correlation matrix뿐만 아니라 다양하게 적용될 수 있음

4

Number of Clusters

- Observation Matrix
  - Modeling
    - N variables : 다변량 정규분포
      - $\rho_{i,j}$: $correlation\ between\ i, j$
        - strong common component가 있다면 detoning이 필요함
  - 방법론
    - 1) Define distance matrix
      - $d_{i,j} = \sqrt{\frac{1}{2}(1 - \rho_{i,j})}$
    - 2) use Correlation matrix
    - 3) derive the X matrix as $X_{i,j} = \sqrt{\frac{1}{2}(1 - \rho_{i,j})}$
      - 앞으로는 3번 방법 사용
    - 장점
      - $\rho_{i,j} = 0.9, \rho_{i,j} = 1.0$이 $\rho_{i,j} = 0.1$에서 $\rho_{i,j} = 0.2$ 보다 큼

4

Number of Clusters

- Base Clustering
  - K-means algorithm
    - 장점
      - 쉽고, 효과적
    - 단점
      - User-set # of clusters K
      - Initialization is random

4 Number of Clusters

- Base Clustering
  - K-means algorithm 개선 방향
    - 1. optimal K 찾기
      - Silhouette score 사용
        - $\boldsymbol{S_i} = \frac{\boldsymbol{b_i - a_i}}{\boldsymbol{max\{a_i, b_i\}}}$
          - $\boldsymbol{a_i}$: average distance between $\boldsymbol{i}$ and other elements (intracluster distance)
          - $\boldsymbol{b_i}$: average distance between $\boldsymbol{i}$ and all the elements in the nearest cluster of which $\boldsymbol{i}$ is not a member (intercluster distance)
        - $\boldsymbol{S_i = 1}$: cluster well
        - $\boldsymbol{S_i = -1}$: cluster poor
      - q : Cluster Quality
        - $\boldsymbol{q} = \frac{\boldsymbol{E[\{S_i\}]}}{\sqrt{\boldsymbol{V[\{S_i\}]}}}$
          - $\boldsymbol{E[\{S_i\}]}$: mean of the silhouette coefficients
          - $\boldsymbol{V[\{S_i\}]}$: variance of the silhouette coefficients

4

Number of Clusters

- Base Clustering
  - K-means algorithm 개선 방향
    - 2. K-means의 initialization 문제 개선
      - 1) Evaluate the observation matrix
      - 2) double for
        - 2-1) different k
          - Evaluate the quality q for each clustering
        - 2-2) repeat first loop multiple time
      - 3) select the clustering with the highest q

4 Number of Clusters

- Base Clustering
  - K-means algorithm 개선 방향
    - 3. Higher-Level Clustering
      - Cluster of inconsistent quality
        - $\overline{q}$: quality의 평균
        - $\{ q_k | \, q_k < \overline{q_k} \}$

Thank you