Clustering for HDLSS using distance vectors

2021311169 서현기

Introduction

- Key of clustering in HDLSS : how to measure the distance between clusters.
- Classical clustering method does not always work well for high dimensional data.
- existing method : MDP clustering - label consistency depends on the sample size and variance of two clusters while MDP clustering only focuses on the difference between the mean vectors

  $\rightarrow$ Poor performance when two clusters have same mean value but different variances
- In order to overcome this, Tereda et al. uses Euclidean distance that contain information regarding the cluster structure in high-dimensional space called distance vector clustering

Preliminaries

- Condition
  1. $p^{-1}\sum_{s=1}^{p} E[X_{ks}]^2 \rightarrow \mu_k^2 \quad \text{as } p \rightarrow \infty$ (mean vector of kth cluster is converge)
  2. $p^{-1}\sum_{s=1}^{p} var[X_{ks}]^2 \rightarrow \sigma_k^2 \quad \text{as } p \rightarrow \infty$ (variance of kth cluster is converge)
  3. $p^{-1}\sum_{s=1}^{p} \{E[X_{ks}]^2 - E[X_{ls}]^2\} \rightarrow \delta_{kl}^2 \quad \text{as } p \rightarrow \infty$ (difference of mean values is converge)
  4. There exists a permutation of variables which is ρ-mixing for functions that are dominated by quadratics (mild condition for law of large number)
- $\eta_{kl} := \underset{p\rightarrow\infty}{lim} p^{-1}\sum_{s=1}^{p} E[X_{ks}]E[X_{ls}]$ (kind of covariance)

Distance vector clustering

Algorithm

1. Compute the usual Euclidean distance matrix $D := (d_{ij}^{(p)})_{N\times N}$ (or inner product matrix $S := XX'$) from the centered data matrix $X := (x_{is})_{N\times p}$
2. Compute the following distance matrix $\Xi := (\xi_{ij}^{(p)})_{N\times N}$

$$\xi_{ij}^{(p)} = \sqrt{\sum_{t\neq i,j}(d_{it}^{(p)} - d_{jt}^{(p)})^2}$$

3. For the matrix Ξ, apply a usual clustering method (e.g, Ward's method)

- Ξ reflects the difference in distance from all other observations.

Theoretical properties

- K-means type
  - objective function of the k-means type distance vector clustering method

$$Q(\mathcal{C}_K|K) := \sum_{i=1}^{N} \min \sum_{j \neq i} (d_{ij}^{(p)} - \bar{d}_{ij}^{(p)})^2$$

$$\text{where } \mathcal{C}_K \text{ is a partition of objects}$$

- Lemma 1

  Let K be the true number of clusters, for an arbitrary $K^* \geq K$,

  $$\underset{\mathcal{C}_{K^*}}{min} Q(\mathcal{C}_{K^*}|K^*) \overset{\mathbb{P}}{\to} 0 \quad as\ p \to \infty$$

  Which means K-means clustering by Ξ can make object functions 0 asymptotically

- Lemma 2

  If $n_k \geq 2$ and true number of clusters K is given, then

  $$if\ \forall k, l(k \neq l);\ \delta_{kl}^2 > 0,$$
  $$\text{then the estimated cluster label vector with the k-means type distance vector}$$
  $$\text{based on } \Xi \text{ converges to the true label vector in probability}$$
  $$as\ p \to \infty$$

- Hierarchical clustering type
  - proposition

    Assume that general conditions a) ~ d) hold, $n_k \geq 2$, let $\mathcal{C}_k := \{C_1, \cdots, C_K\}$ be the rue cluster partition, then

    $$if \forall k, l(k \neq l); \sigma_k \neq \sigma_l\ or\ \delta_{kl}^2 > 0,\ then$$
    $$P\left(\forall k, l(k \neq l); \max_{i,j \in C_k} \xi_{ik}^{(p)} < \min_{i \in C_k, j \in C_l} \xi_{ik}^{(p)}\right) \to 1\ \ as\ p \to \infty$$

- Sufficient condition of Distance vector clustering

  1. When using Distance matrix D

     $$\sigma_k \neq \sigma_k\ \ or\ \ \delta_{kl} > 0$$

  2. when using inner product matrix S

     $$\delta_{kl} > 0$$

  $\to$ those conditions **do not depend on the sample size!!**

  - When using D, We can capture the clusters whose only difference is variance not mean
  - When using S, We can capture the mean-different clusters regardless of its variance(less susceptible than MDP method)