A.Z
Data Mining
Dec, 2023

lec 3 Clustering I

Abstract

Introduce Clustering and Clustering tendency

1 Clustering

Hard clustering: each point belongs to one cluster

Soft clustering: point can belong to multiple clusters with different probabilities

2 Objective functions(score)

Goal: minimize within-cluster-variation wc and maximize between-cluster variation bc

Let $\mathbf{C} = \{C_1, \ldots, C_K\}$ are clusters, $\mathbf{c}_1, \ldots, \mathbf{c}_K$ their centroids and d distance function.

Examples of wc :

$$wc(\mathbf{C}) = \sum_{i=1}^{K} wc\left(C_i\right) = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} d^2\left(\mathbf{x}, \mathbf{c}_i\right)$$

called hyperspherial clusters

$$wc\left(C_p\right) = \max_i \overbrace{\min_{\mathbf{y} \in C_p} \left\{d\left(\mathbf{x}_i, \mathbf{y}\right) \mid \mathbf{x}_i \in C_p, \mathbf{x}_i \neq \mathbf{y}\right\}}^{\mathbf{x_i}'\text{s distance to its nearest neighbour in } C_p}$$

called elongate clusters

Example of bc:

$$bc(\mathbf{C}) = \sum_{1 \leq i < j \leq K} d^2\left(\mathbf{c}_i, \mathbf{c}_j\right)$$

Example of overall measure is K-means criterion, Sum of Squared Errors(SSE):

$$\mathrm{SSE}(\mathbf{C})=\sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} L_2^2\left(\mathbf{x}, \mathbf{c}_i\right)$$

Minimizing SSE means minimizes within-cluster variance and maximizes between-cluster variance

3 Clustering tendency

3.1 Distance distributions

Plot a histogram of pairwise distances in data

No cluster data have only one peak, cluster data have more than one peaks.

3.2 Entropy-based measures

Idea: In random data (uniform distribution), the entropy is high and in clustered data low

Approach:

1. Calculate pairwise distances between points

2. Discretize distances onto m bins

$$E=-\sum_{i=1}^{m}\left[p_i \log \left(p_i\right)+\left(1-p_i\right) \log \left(1-p_i\right)\right]$$

where p_i = fraction of distances in the i th bin

paper: Feature Selection for Clustering – A FilterSolution. ICDM, 2002

3.3 Hopkins statistic

Idea: compare nearest neighbour distances from the original data and random data points

Approach:

1. Take a sample R of size r from original data $\mathcal{D}$

2. Generate random data (from uniform distribution) and take a sample S of size r from random data

3. Calculate for all $\mathbf{x} \in R$ distances to their nearest neighbours (in $\mathcal{D}$). Let these be $\alpha_1, \ldots, \alpha_r$

4. Calculate for all $\mathbf{x} \in S$ distances to their nearest neighbours (in $\mathcal{D}$). Let these be $\beta_1, \ldots, \beta_r$

$$H = \frac{\sum_{i=1}^{r} \beta_i}{\sum_{i=1}^{r} (\alpha_i + \beta_i)}$$

If $\mathcal{D}$ has uniform distribution, $H \approx 0.5$

If there are clusters, H approaches 1

Note: H follows $Beta(r, r)$ distribution

Problem: Distance distribution often very different in the center of data than on edges

Solution: choose sample points inside a hypersphere centered at the mean of data and containing 50

Problem: Results vary with different executions

Solution: repeat multiple times and calculate average

3.4 Wrapper models and validation indices

Idea: Iteratively cluster data with different feature sets and use validity indexes to find good features

Approach:

1. Cluster data and calculate some internal cluster validity index. Often use greedy methods, and the result depend on the validity criterion(and clustering method)

2. Create artificial class labels and identify discriminative features in a supervised manner, then evaluate each feature separately

But there is a circular definition: features are good if the clustering is good, but good clustering requires good features