Clustering Evaluation

Javier Béjar

URL - Spring 2019

CS - MAI

Cluster Evaluation

Model evaluation

- The evaluation of unsupervised learning is difficult
- There is no goal model to compare with
- The true result is unknown, it may depend on the context, the task to perform, ...
- Why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare different models/parameters

What can be evaluated?

- Cluster tendency, there are clusters in the data?
- Compare the clusters to the true partition of the data
- Quality of the clusters without reference to external information
- Compare the results of different clustering algorithms
- Evaluate algorithm parameters
    - For instance, to determine the *correct* number of clusters

Model evaluation - Cluster Tendency

- Before clustering a dataset we can test if there are actually clusters
- We have to test the hypothesis of the existence of patterns in the data versus a dataset uniformly distributed (homogeneous distribution)

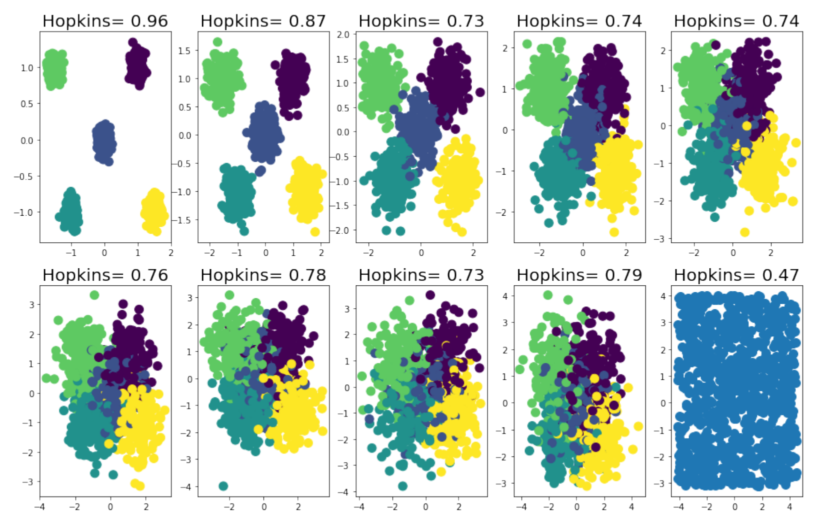
Model evaluation - Cluster Tendency

- **Hopkins Statistic**
  1. Sample n points (p_i) from the dataset (D) uniformly and compute the distance to their nearest neighbor ($d(p_i)$)
  2. Generate n points (q_i) uniformly distributed in the space of the dataset and compute their distance to nearest neighbors in D ($d(q_i)$)
  3. Compute the quotient:

$$H = \frac{\sum_{i=1}^{n} d(p_i)}{\sum_{i=1}^{n} d(p_i) + \sum_{i=1}^{n} d(q_i)}$$

  4. If data are uniformly distributed the value of H will be around 0.5

Hopkins Statistic - Example



Cluster Quality criteria

- We can use different methodologies/criterion to evaluate the quality of a clustering:
  - External criteria: Comparison with a model partition/labeled data
  - Internal criteria: Quality measures based on the examples/quality of the partition
  - Relative criteria: Comparison with other clusterings

Internal criteria

Internal criteria

- Measure properties expected in a good clustering
    - Compact groups
    - Well separated groups
- The indices are based on the model of the groups
- We can use indices based on the attributes values measuring the properties of a good clustering
- These indices are based on statistical properties of the attributes of the model
    - Values distribution
    - Distances distribution

Internal criteria - Indices

- Some of the indices correspond directly to the objective function optimized:
  - Quadratic error/Distorsion (k-means)

$$SSE = \sum_{k=1}^{k} \sum_{\forall x_i \in C_k} \parallel x_i - \mu_k \parallel^2$$

  - Log likelihood (Mixture of gaussians/EM)

Internal criteria - Indices

- For prototype based algorithms several measures can be use to compute quality indices
- Scatter matrices: interclass distance, intraclass distance, separation

$$
\begin{aligned}
S_{W_k} &= \sum_{\forall x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \\
S_{B_k} &= |C_k|(\mu_k - \mu)(\mu_k - \mu)^T \\
S_{M_{k,l}} &= \sum_{\forall i \in C_k} \sum_{\forall j \in C_l} (x_i - x_j)(x_i - x_j)^T
\end{aligned}
$$

Internal criteria - Indices

- Trace criteria (lower overall intracluster distance/higher overall intercluster distance)

$$Tr(S_W) = \frac{1}{K}\sum_{i=1}^{K} S_{W_k} \quad Tr(S_B) = \frac{1}{K}\sum_{i=1}^{K} S_{B_k}$$

- Calinski-Harabasz index (interclass-intraclass distance ratio)

$$CH = \frac{\sum_{i=0}^{K} |C_i| \times \|\mu_i - \mu\|^2/(K-1)}{\sum_{k=1}^{K}\sum_{i=0}^{|C_i|} \|x_i - \mu_i\|^2/(N-K)}$$

Internal criteria - Indices

- Davies-Bouldin criteria (maximum interclass-intraclass distance ratio)

$$\bar{R} = \frac{1}{K} \sum_{i=1}^{K} R_i$$

where

$$\begin{aligned} R_{ij} &= \frac{S_{W_i} + S_{W_j}}{S_{M_{ij}}} \\ R_i &= \max_{j:j \neq i} R_{ij} \end{aligned}$$

Internal criteria - Indices

- Silhouette index (maximum class spread/variance)

$$S = \frac{1}{N} \sum_{i=0}^{N} \frac{b_i - a_i}{max(a_i, b_i)}$$

Where

$$a_i = \frac{1}{|C_j| - 1} \sum_{y \in C_j, y \neq x_i} \|y - x_i\|$$

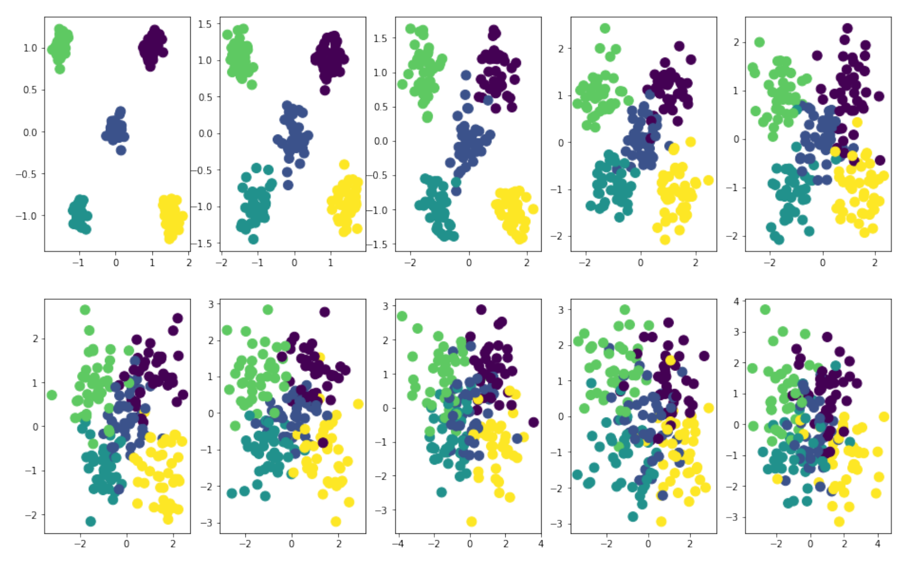
$$b_i = \min_{l \in H, l \neq j} \frac{1}{|C_l|} \sum_{y \in C_l} \|y - x_i\|$$

with $x_i \in C_j, H = \{h : 1 \leq h \leq K\}$

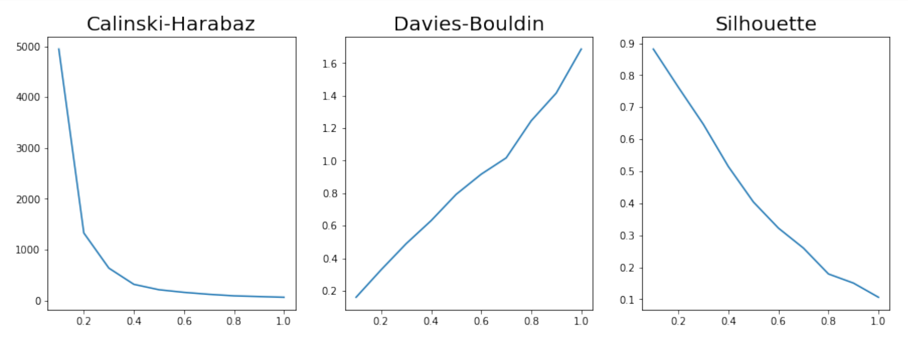
Internal criteria - Indices

- More than 30 indices can be found in the literature
- Several studies and comparisons have been performed
- Recent studies (Arbelatiz et al, 2013) have exhaustively tested these indices, some have a performance significativelly better that others
- Some of the indices show a similar performance (not statistically different)
- The study concludes that Silhouette, Davies-Bouldin and Calinski Harabasz perform well in a wide range of situations

Internal criteria - 5 clusters different variance



Internal criteria - 5 clusters different variance - Scores



External criteria

External criteria

- These indices measure the similarity of a clustering to a model partition P
- Without a model they can be used to compare the results of using different parameters or different algorithms
  - For instance, can be used to assess the sensitivity to initialization
- The main advantage is that these indices are independent of the examples/cluster description
- That means that they can be used to assess any clustering algorithm

External criteria - Indices

- All the indices are based on the coincidence of each pair of examples in the groups of two clusterings
- The computations are based on four values:
  - The two examples in the same cluster in both partitions (a)
  - The two examples in the same cluster in C, but not in P (b)
  - The two examples in the same cluster in P, but not in C (c)
  - The two examples in different cluster in both partitions (d)

External criteria - Indices

- Rand/Adjusted Rand statistic:

$$Rand = \frac{(a+d)}{(a+b+c+d)}; \quad ARand = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+b)(a+c)}{a+b+c+d}}$$

- Jaccard Coefficient:

$$J = \frac{a}{(a+b+c)}$$

- Folkes and Mallow index:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

External criteria - Indices - Information Theory

- Defining Mutual Information between two partitions as:

$$MI(Y_i, Y_k) = \sum_{X_c^i \in Y_i} \sum_{X_{c'}^k \in Y_k} \frac{|X_c^i \cap X_{c'}^k|}{N} \log_2\left(\frac{N|X_c^i \cap X_{c'}^k|}{|X_c^i||X_{c'}^k|}\right)$$

- and Entropy of a partition as

$$H(Y_i) = -\sum_{X_c^i \in Y_i} \frac{|X_c^i|}{N} \log_2\left(\frac{|X_c^i|}{N}\right)$$

where $X_c^i \cap X_{c'}^k$ is the number of objects that are in the intersection of the two groups

External criteria - Indices - Information Theory

- Normalized Mutual Information:

$$NMI(Y_i, Y_k) = \frac{MI(Y_i, Y_k)}{\sqrt{H(Y_i)H(Y_k)}}$$

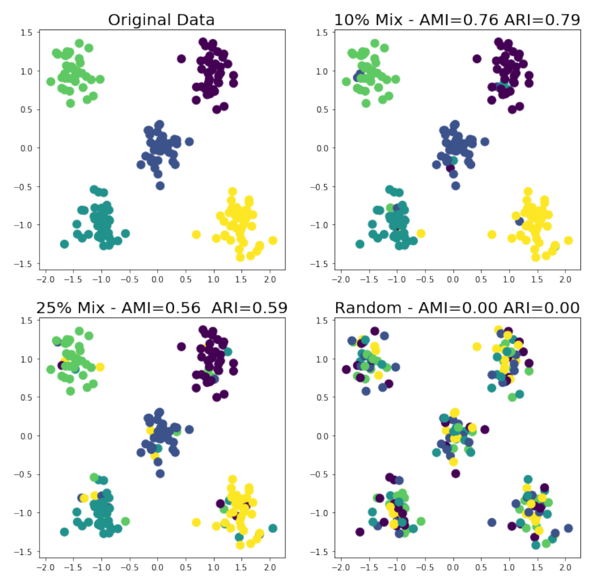
- Variation of Information:

$$VI(C, C') = H(C) + H(C') - 2I(C, C')$$

- Adjusted Mutual Information:

$$AMI(U, V) = \frac{MI(U, V) - E(MI(U, V))}{\max(H(U), H(V)) - E(MI(U, V))}$$

External criteria - ARI/AMI Scores



Number of clusters

Number of clusters

- A topic related to cluster validation is to decide if the number of clusters obtained is the correct one
- This point is important specially for the algorithms that need this value as a parameter
- The usual procedure is to compare the characteristics of clusterings of different sizes
- Usually internal criteria indices are used in this comparison
- A graphic of this indices for different number of clusters can show what number of clusters is more probable

Number of clusters - Indices

- Some of the internal validity indices can be used for this purpose: Calinsky Harabasz index, Silhouette index
- Using the within class scatter matrix (S_W) other criteria can be defined:
  - Hartigan index:

$$H(k) = \left[\frac{S_W(k)}{S_W(k+1)} - 1\right](n-k-1)$$

  - Krzanowski Lai index:

$$KL(k) = \left|\frac{DIFF(k)}{DIFF(k+1)}\right|$$

  being $DIFF(k) = (k-1)^{2/p} S_W(k-1) - k^{2/p} S_W(k)$

Number of clusters - The Gap Statistic

- Assess the number of clusters comparing a clustering with the expected distribution of data given the null hypothesis (no clusters)
- Compute different clusterings of the data increasing the number of clusters and compare to clusters of data (B) generated with a uniform distribution

Number of clusters - The Gap Statistic

- The Gap statistic:

$$Gap(k) = (1/B)\sum_b log(S_W(k)_b) - log(S_W(k))$$

- From the st. dev. (sd_k) of $\sum_b log(S_W(k)_b)$ is defined s_k as:

$$s_k = sd_k\sqrt{1 + 1/B}$$

- The probable number of clusters is the smallest number that holds:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

Number of clusters - Cluster Stability

- The idea is that if the model chosen for clustering a dataset is correct, it should be stable for different samplings of the data
- The procedure is to obtain different subsamples of the data, cluster them and test their stability

Number of clusters - Cluster Stability

- Using disjoint samples:
  - Dataset divided in two disjoint samples that are clustered separately
  - Indices can be defined to assess stability, for example using the distribution of the number of neighbors that belong to the complementary sample
- Using non disjoint samples:
  - Dataset divided in three disjoint samples (S_1,S_2, S_3)
  - Two clusterings are obtained from $S_1 \cup S_3$, $S_2 \cup S_3$
  - Indices can be defined about the coincidence of the common examples in both partitions

Python Notebooks

This Python Notebook has examples for Measures of Clustering Validation

- Clustering Validation Notebook (click here to go to the url)

If you have downloaded the code from the repository you will able to play with the notebooks (run jupyter notebook to open the notebooks)

Python Code

- In the code from the repository inside subdirectory `Validation` you have the python program `ValidationAuthors`,
- The `authors` dataset is clustered with different algorithms (K-means, GMM, Spectral) and different validity indices are plotted for the number of clusters

Cluster Visualization

Cluster visualization

- Dimensionality reduction
    - Project the dataset to 2 or 3 dimensions
    - The clusters in the new space could represent clusters in the original space
    - The confidence depends on the reconstruction error of the transformed data and that the transformation maintains the relations in the original space

Cluster visualization

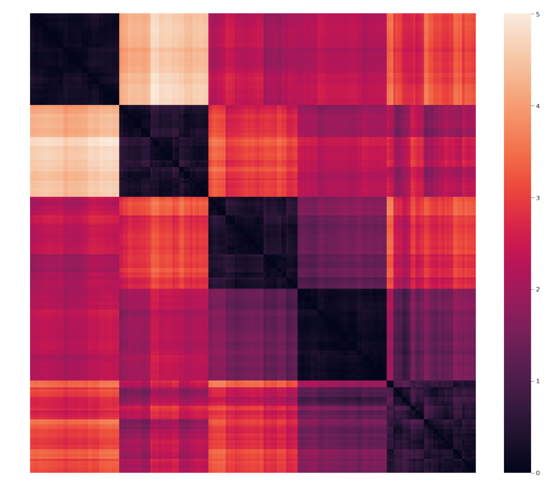
- Distance matrix visualization
    - The distance matrix represents the examples relationships
    - Can be rearranged so the closer examples appear in adjacent columns
    - Patterns in the rearranged matrix can show cluster tendency

Cluster visualization - Distance matrix

- There are several methods
- The simplest one is to use a hierarchical clustering algorithm and rearrange the matrix using a inorder traversal of the tree
- Results will depend on the algorithm used and the distance/similarity function
- Can be applied to quantitative and qualitative data
- See patterns in the distance matrix is not always guarantee of clusters in the data

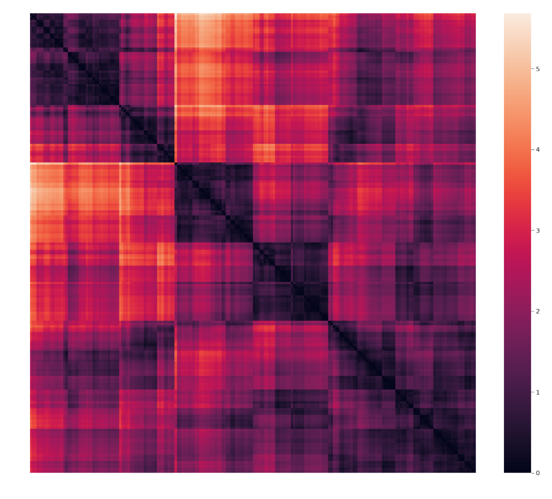
Cluster visualization - Distance matrix

Dataset with five well separated clusters



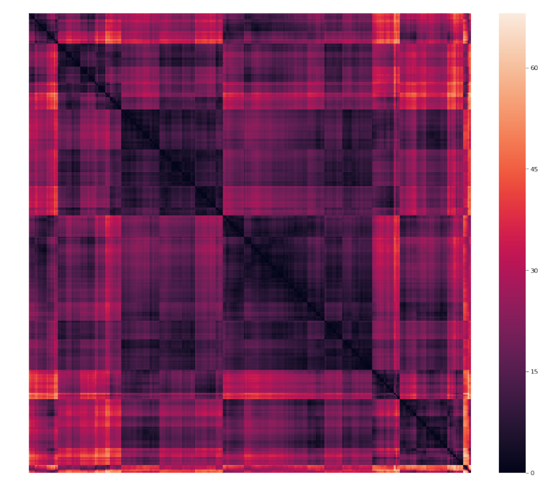
Cluster visualization - Distance matrix

Dataset with five noisy and overlapping clusters



Cluster visualization - Distance matrix

Random Data



Cluster visualization - Distance matrix

Two circles dataset (euclidean distance, cosine similarity)