Clustering Evaluation/Model Assessment

Javier Béjar

CS - MAI

AMLT - 2014/2015
1. Cluster Evaluation

2. Internal criteria

3. External criteria

4. Number of clusters

5. Application

6. Cluster Visualization
1 Cluster Evaluation

2 Internal criteria

3 External criteria

4 Number of clusters

5 Application

6 Cluster Visualization
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
Model evaluation

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
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).

**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 points are uniformly distributed the value of \( H \) will be around 0.5.
We can use different methodologies/criterion to evaluate the quality of a clustering:

- Comparison with a model partition/labeled data (External criteria)
- Quality measures based on the examples/quality of the partition (Internal criteria)
- Comparison with other clusterings (Relative criteria)
1. Cluster Evaluation

2. Internal criteria

3. External criteria

4. Number of clusters

5. Application

6. Cluster Visualization
Internal criteria

- We can measure some properties that are 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
Some of the indices correspond directly to the objective function optimized:

- Quadratic error/Distorsion (k-means)

\[
SSE = \sum_{k=1}^{k} \sum_{\forall x_i \in C_k} \| x_i - \mu_k \|^2
\]

- Log likelihood (Mixture of gaussians/EM)
For prototype based algorithms several measures can be used to compute quality indices.

Scatter matrices: interclass distance, intraclass distance, separation

\[
S_{W_k} = \sum_{\forall x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T
\]

\[
S_{B_k} = |C_k|(\mu_k - \mu)(\mu_k - \mu)^T
\]

\[
S_{M_{k,l}} = \sum_{\forall i \in C_k} \sum_{\forall j \in C_l} (x_i - x_j)(x_i - x_j)^T
\]
Internal criteria

Internal criteria - Indices

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

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

- Determinant criteria (higher cross-intercluster distance)

\[
Det(S_M) = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} S_{M_{i,j}}
\]
Internal criteria - Indices

- Condorcet criteria (total interclass/intraclass distances)

\[ \sum_{C_i \in C} \sum_{x_j, x_k \in C_i, x_j \neq x_k} s(x_j, x_k) + \sum_{C_i \in C} \sum_{x_j \in C_i; x_k \notin C_i} s(x_j, x_k) \]

- Kalinski 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

\[
R_{ij} = \frac{S_{W_i} + S_{W_j}}{S_{M_{ij}}}
\]

\[
R_i = \max_{j:j \neq i} R_{ij}
\]
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

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

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

being \( x_i \in C_j, H = \{h : 1 \leq h \leq K\} \)
Internal criteria - Indices

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

Data with 5 clusters

![Graph showing interclass distance](image-url)
Data with 5 clusters
1. Cluster Evaluation

2. Internal criteria

3. External criteria

4. Number of clusters

5. Application

6. Cluster Visualization
External criteria

- These indices measure if a clustering is similar to a model partition $P$
- It is equivalent to have a labeled dataset (ground truth)
- If we do not have a model these criteria can be used to compare the results of using different parameters of 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
- These means that 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 we have four values:
  - The two examples belong to the same class in both partitions ($a$)
  - The two examples belong to the same class in $C$, but not in $P$ ($b$)
  - The two examples belong to the same class in $P$, but not in $C(c)$
  - The two examples belong to different classes in both partitions ($d$)
External criteria

External criteria - Indices

- **Rand/Adjusted Rand statistic:**
  \[
  R = \frac{(a + d)}{(a + b + c + d)}; \quad ARand = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+b)(a+c)}{a+b+c+d}}
  \]

- **Jaccard Coefficient:**
  \[
  J = \frac{a}{(a + b + c)}
  \]

- **Folkes and Mallow index:**
  \[
  FM = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}}
  \]
Defining Mutual Information between two partitions as:

$$MI(Y_i, Y_k) = \sum_{X^i_c \in Y_i} \sum_{X^k_{c'} \in Y_k} \frac{|X^i_c \cap X^k_{c'}|}{N} \log_2 \left( \frac{N|X^i_c \cap X^k_{c'}|}{|X^i_c||X^k_{c'}|} \right)$$

and Entropy of a partition as

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

where $X^i_c \cap X^k_{c'}$ 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:**
  \[ AMI(U, V) = \frac{MI(U, V) - E(MI(U, V))}{\max(H(U), H(V)) - E(MI(U, V))} \]
1. Cluster Evaluation
2. Internal criteria
3. External criteria
4. Number of clusters
5. Application
6. Cluster Visualization
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:
  - Calinski 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)
- Different clusters of the data increasing the number of clusters and compared to datasets (B) generated with a uniform distribution
- 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:

\[
Gap(k) \geq Gap(k + 1) - s_{k+1}
\]
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 and cluster them and test their stability.

Using disjoint samples:
- The dataset is divided into 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:
- The dataset is divided into three disjoint samples \(S_1, S_2, S_3\).
- Two clusterings are obtained from \(S_1 \cup S_3\) and \(S_2 \cup S_3\).
- Indices can be defined about the coincidence of the common examples in both partitions.
1. Cluster Evaluation

2. Internal criteria

3. External criteria

4. Number of clusters

5. Application

6. Cluster Visualization
Application: Wheelchair Control Dataset

- Wheelchair with shared control (patient/computer)
- Recorded trajectories of several patients in different situations
  - Angle/distance to the goal, Angle/distance to the nearest obstacle from around the chair (210 degrees)
- Characterization about how the computer helps the patients with different handicaps
- How many clusters characterize the trajectories?
Wheel chair control characterization

PCA 88 $\rightarrow$ 3 dimensions
Wheel chair control characterization

Silhouette index for k-means
Wheel chair control characterization
Stability/Consensus ARAND and ANMI for k-means

ARAND

ANMI
Cluster Evaluation

Internal criteria

External criteria

Number of clusters

Application

Cluster Visualization
Another method for model assessment is to visualize the data and look for clusters.

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

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

Both methodologies are computationally expensive.
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

![Distance Matrix](image-url)
Cluster visualization - Distance matrix

Dataset with five noisy and overlapping clusters
Random Data
Three rings dataset (euclidean distance, cosine similarity)