Cluster Validity

Tania Cerquitelli
Politecnico di Torino

Measures of cluster validity

- The validation of clustering structures is the most difficult task
- To evaluate the "goodness" of the resulting clusters, some numerical measures can be exploited
- Numerical measures are classified into two main classes
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels
    - Entropy
    - Purity
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information
    - Sum of Squared Error (SSE)
    - Davies–Bouldin index
    - Sum of squares item distribution

External Indices

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Sum of Squared Error (SSE)

- For each point, the error is the distance to the representative point of its cluster
- To get SSE, these errors are squared and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i
  - m_i is the representative point for cluster C_i
    - m_i may corresponds to the center (mean) of the cluster
- Given two clustering results, the one with the smallest SSE is the best one

Davies-Bouldin index

- Let $M_{i,j}$ be a measure to evaluate the separation between the i^{th} and the j^{th} cluster, which as to be as large as possible
- Let S_i be a measure to evaluate the within cluster scatter for cluster i, which has to be as low as possible
- $R_{i,j}$ measures how good the cluster schema is as $R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$
  - The lower the value, the better the separation of clusters i and j and the tightness inside the clusters
- D_i chooses the worst case scenario $D_i = \max_{j \neq i} R_{i,j}$
- The Davies-Bouldin (DB) index is defined as $DB = \frac{1}{K} \sum_{i=1}^{K} D_i$
- Given two clustering results, the one with the smallest DB is the best one

Sum of squares item distribution

- The sum of squares item distribution (SSID) is defined as follows

$$SSID = \sum_{i=1}^{K} \left[\frac{|C_i|}{\sum_{j=1}^{K} |C_j|} \right]^2$$

  - K is the number of clusters
  - $|C_i|$ represents the number of points in cluster i
- When one cluster dominates and the others clusters are very small in comparison, the SSID will tend to 1.
- When the clusters have equal numbers of points, the SSID tends to 1/K where K is the number of clusters.