How to evaluate (unsupervised) clustering

Yujin Kim 12-15-2021

Answers:

From *stackexchange*

“how well a particular unsupervised method performs will largely depend on ***why one is doing unsupervised learning in the first place****”*

Answers:

From *stackexchange*

“Using a supervised approach as a proxy to how well an unsupervised approach works doesn't require the discovery of new features. For example, clustering doesn't learn new features, yet clustering is often used ***to improve the prediction accuracy of a supervised learner****, with the added benefit of explaining why this may be so. For example, k-means clustering can produce k predictions that are each improved by way of exploiting the discovered structure and compression from clustering.”*

Determining the quality of a clustering algorithm

Several steps for validation of algorithm results

Internal or unsupervised validation

- Determining the clustering tendency in the data
- Determining the correct number of clusters
- Assessing the quality of the clustering results without external information.

External or supervised validation

- Comparing the results obtained with external information.

Both supervised and unsupervised validation

- Comparing two sets of clusters to determine which one is better.

Internal or unsupervised validation

(1) Determining the clustering tendency in the data

- whether data exhibits some tendency to form actual clusters
- e.g., null hypothesis testing with bootstrapping
  - H0 : the randomness of data
  - H1 : the non-randomness (clustering) exists

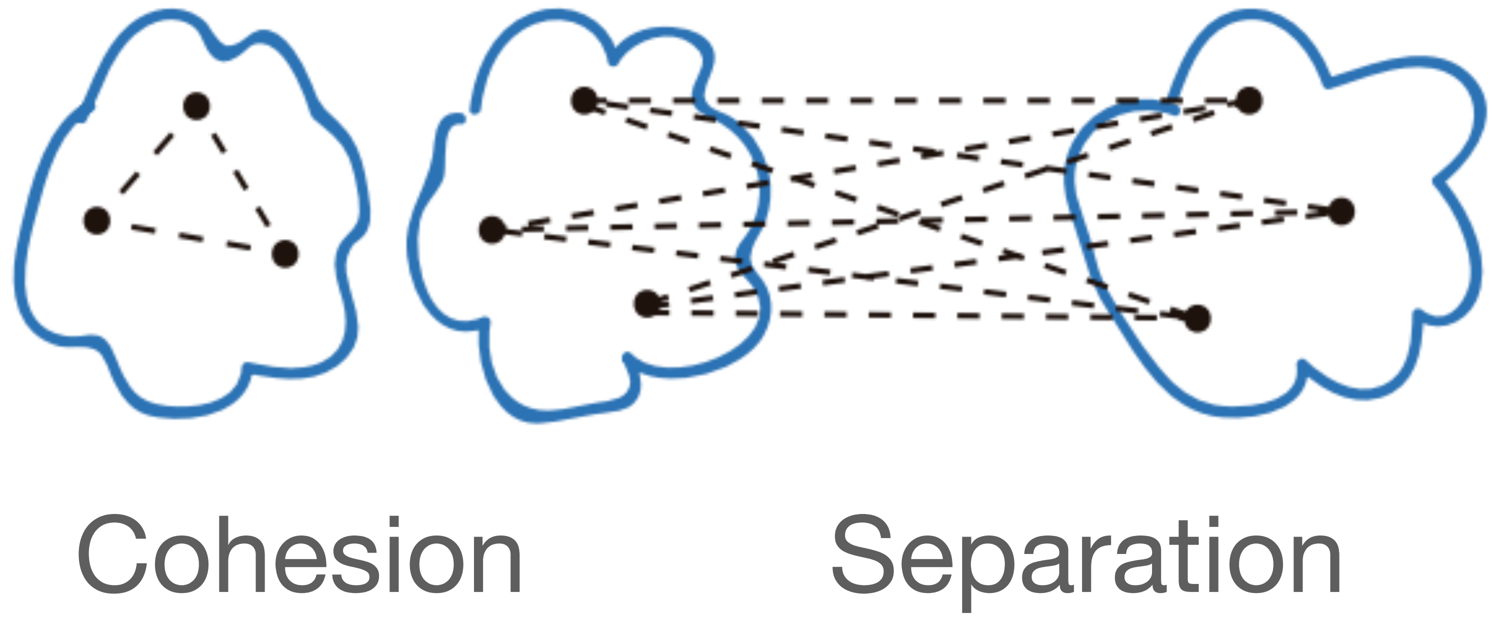
- Random plot hypothesis H_0: All proximity matrices of order $n \times n$ are equally likely.
- Random label hypothesis H_0: All permutations of labels on n objects are equally likely.
- Random position hypothesis H_0: All sets of n locations in some region of a d-dimensional space are equally likely.

Context of data

Internal or unsupervised validation

(2) Determining the correct number of clusters

- Cohesion : how closely the elements of the same cluster are to each other
- Separation: measures qualify the level of separation between clusters.
- e.g., Partitional algorithm — proximity metrics as well as metrics of cohesion and separation (e.g., Silhouette coefficient)
- e.g., Cophenetic coefficient for hierarchical algorithms (CPCC)
- When it has a high separation between clusters and a high cohesion within clusters, a clustering is considered to be good



$$cohesion(C_i) = \sum_{x \in C_i, y \in C_i} proximity(x, y)$$

$$separation(C_i, C_j) = \sum_{x \in C_i, y \in C_j} proximity(x, y)$$

Internal or unsupervised validation

(2) Determining the correct number of clusters (cont.)

- Other metrics — CH (the Calisnki-Harabasz coefficient; the variance ratio criterion), the Dunn index, etc etc.
- The silhouette coefficient; computing a particular point and computing the global silhouette coefficient [-1, 1] by the average of the particular silhouette coefficients for each example.
- Drawback; high computational complexity: O(dn^2); impractical for huge data sets.

1) For each example, the average distance $a(i)$ to all the examples in the same cluster is computed:

$$a(i) = \frac{1}{|C_a|} \sum_{j \in C_a, i \neq j} d(i, j)$$

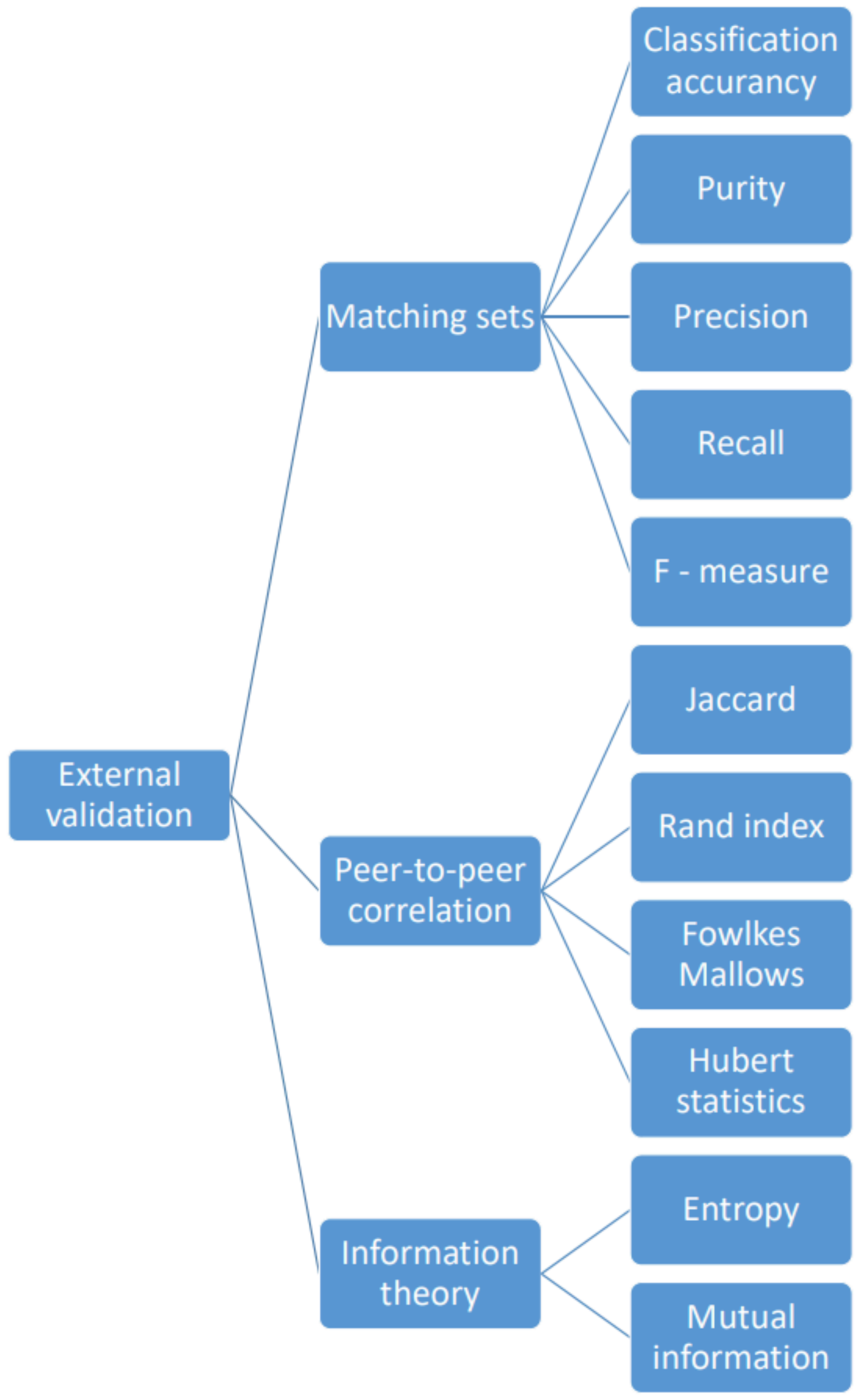
2) For each example, the minimum average distance $b(i)$ between the example and the examples contained in each cluster not containing the analyzed example:

$$b(i) = \min_{C_b \neq C_a} \frac{1}{|C_b|} \sum_{j \in C_b} d(i, j)$$

3) For each example, the silhouette coefficient is determined by the following expression:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

External or supervised validation



- Matching sets; comparing two partitions of data (with true data)
  - precision, recall, F-1 etc
- Peer-to-peer correlation; seeking to measure the similarity between two partitions under equal conditions
  - Jaccard coefficient (only TP), Rand coefficient (TP + TN, similar to accuracy)
- Information theory; capturing existing uncertainty in the prediction of the natural classes
  - entropy, mutual information

Reference

- Palacio-Niño, J. O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*. https://arxiv.org/pdf/1905.05667.pdf
- https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation
- https://stats.stackexchange.com/questions/79028/performance-metrics-to-evaluate-unsupervised-learning