The Area under the ROC Curve as a Criterion for Clustering Evaluation

Helena Aidos[1], Robert P. W. Duin[2] and Ana Fred[1]

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal
[2]Pattern Recognition Laboratory, Delft University of Technology, The Netherlands
{haidos, afred}@lx.it.pt , r.duin@ieee.org

Validation Criteria for Clustering

based on the data only, such as the average intra-cluster distance

Internal criteria

e.g., Silhouette Index, Davies-Bouldin Index, Dunn Index

Advantages/disadvantages:

- Do not need to have the true class label.
- Biased towards one clustering algorithm.

External criteria

e.g., Rand Statistics, Jaccard Coefficient, Fowlkes and Mallows Index

Advantages/disadvantages:

- Need to have the true class label for each object.

Designed for a fixed number of clusters

ROC curve and AUC

(to study the robustness of clustering algorithms for several number of clusters, *k*)

Consider two given points $\mathbf{x}_a$, $\mathbf{x}_b$.

➢ Type I error:

$$\varepsilon_1 \equiv P(\mathbf{x}_a \in C_i, \mathbf{x}_b \in C_j | \mathbf{x}_a, \mathbf{x}_b \in P_l), i \neq j$$

➢ Type II error:

$$\varepsilon_2 \equiv P(\mathbf{x}_a, \mathbf{x}_b \in C_i | \mathbf{x}_a \in P_j, \mathbf{x}_b \in P_l), j \neq l$$

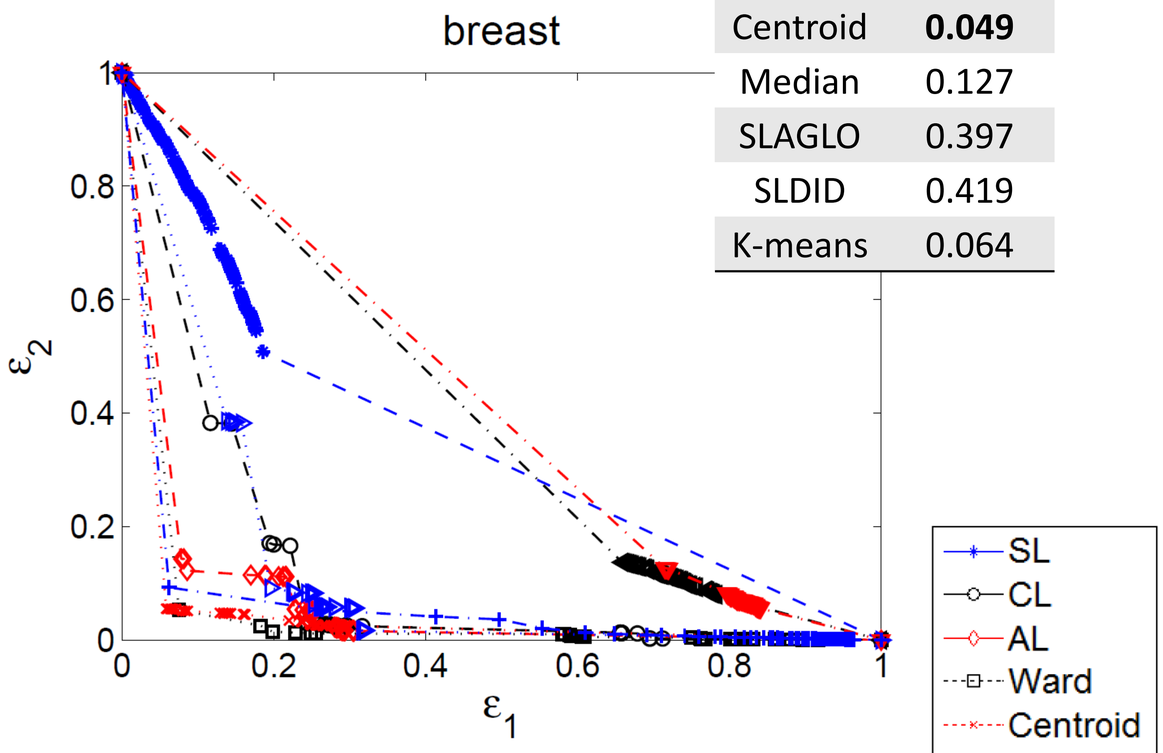
➢ A clustering partition *C* is **concordant** with the true labeling, *P*, of the data if

$$\begin{cases} \varepsilon_1 = 0 & \text{if } k \leq m \\ \varepsilon_2 = 0 & \text{if } k \geq m \\ \varepsilon_1 = \varepsilon_2 = 0 & \text{if } k = m. \end{cases}$$

➢ A ROC curve is **proper** if, when varying *k*, ε_1 increases whenever ε_2 decreases and vice-versa.

➢ Evaluate Robustness: A clustering algorithm is more **robust** to the choice of *k* than another algorithm if the former's AUC is smaller than the latter's.

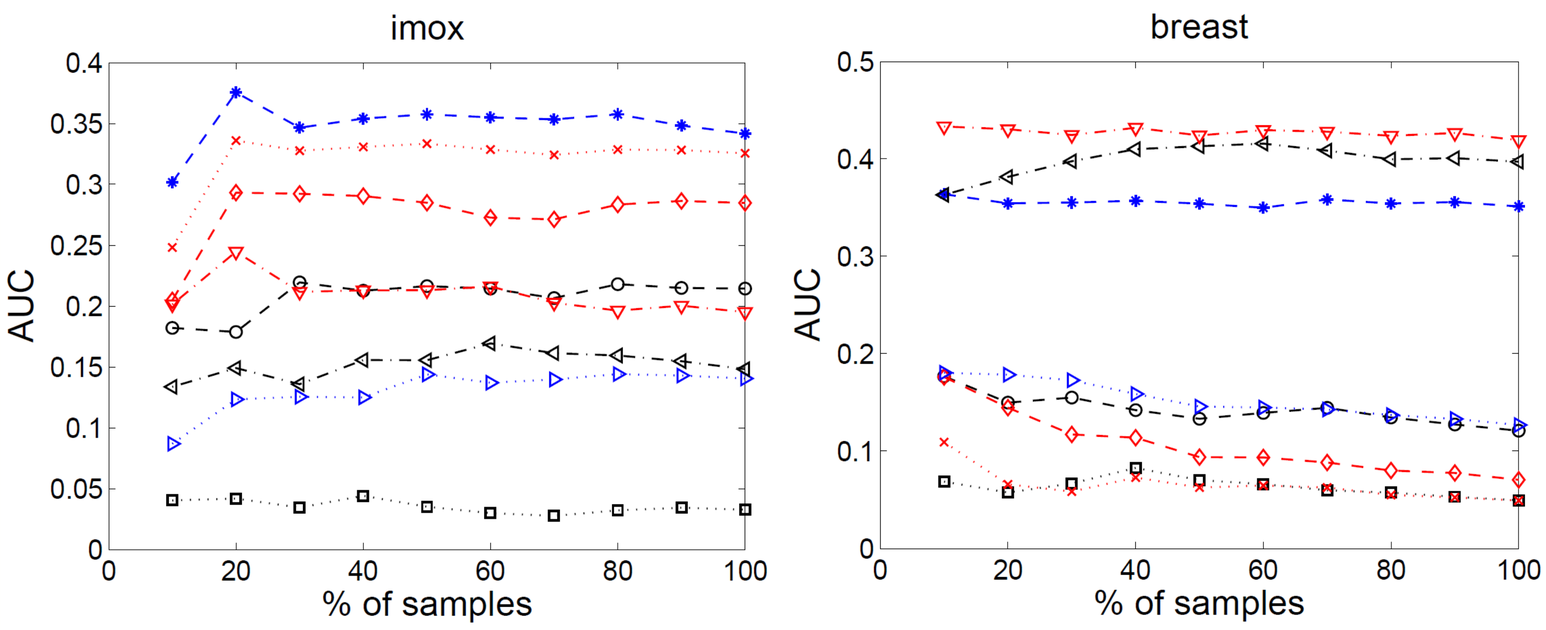
ROC curve and AUC when we have access to all labeling information



	AUC
SL	0.351
CL	0.121
AL	0.070
Ward	**0.049**
Centroid	**0.049**
Median	0.127
SLAGLO	0.397
SLDID	0.419
K-means	0.064

Semi-supervised Context

We might have only labels for a small part of the entire dataset.



Conclusions

➢ In the literature, external and internal criteria are designed to **evaluate clustering** algorithms for a **fixed number of clusters**.

➢ The proposed measure quantifies the performance of an algorithm for **several *k* simultaneously**.

- This allows measuring how robust a clustering algorithm is to the choice of *k*.

➢ In the semi-supervised context, the whole dataset is used to perform clustering, whereas the AUC is computed with only a part of the data.

➢ The measure proposed can be used to automatically detect whether the currently labeled data is already enough.

➢ Allow us to extrapolate classes from the labeled data to the unlabeled data, if one can find a clustering algorithm which yields low and consistent AUC value for the labeled portion.

This work was supported by the Portuguese Foundation for Science and Technology, grant PTDC/EIA-CCO/103230/2008.