On Consensus Clustering Validation

João M. M. Duarte[1,2] Ana L. N. Fred[1] André Lourenço[1]
F. Jorge F. Duarte[2]

[1] Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal

[2] GECAD - Knowledge Engineering and Decision Support Group, Instituto Superior de Engenharia do Porto, Porto, Portugal

Outline

Cluster Ensembles

- Clustering ensemble approaches have been proposed aiming:
  - to improve data clustering robustness and quality;
  - reuse clustering solutions;
  - cluster data in a distributed way.
- There are many alternative ways of building the clustering ensemble, defining the combination strategy and extraction algorithm, and choosing the final number of clusters.

Cluster Ensembles (2)

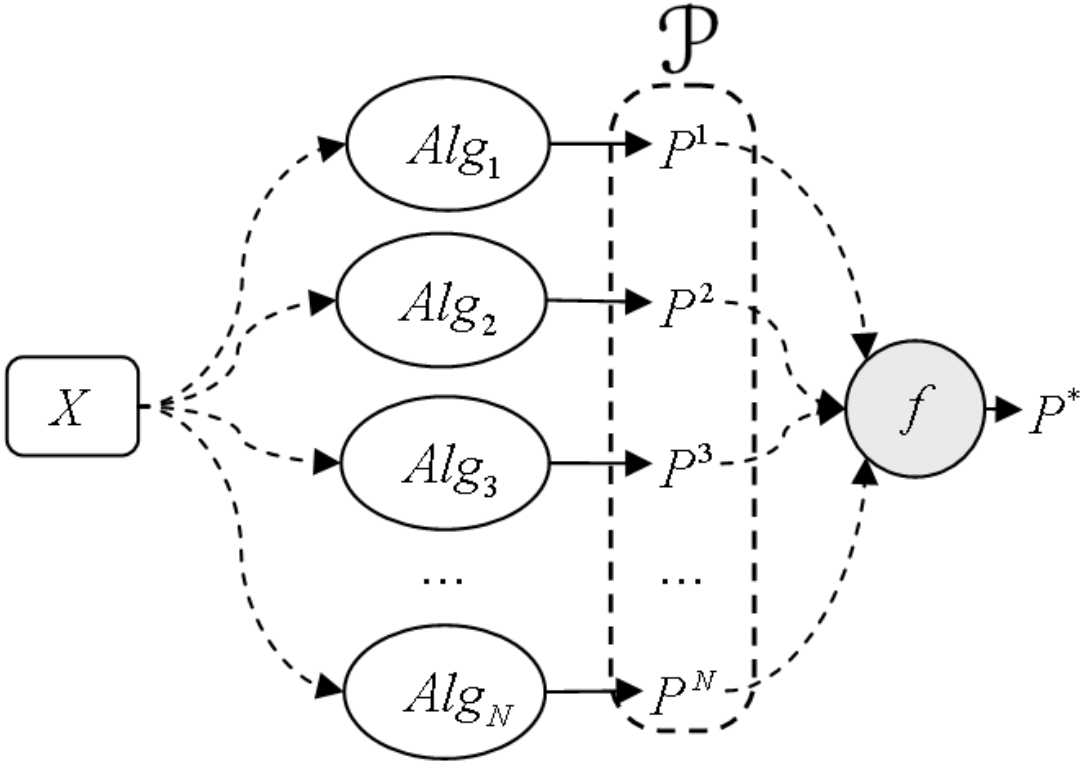


Figure: Combining multiple data partitions

How To Validate Consensus Partitions?

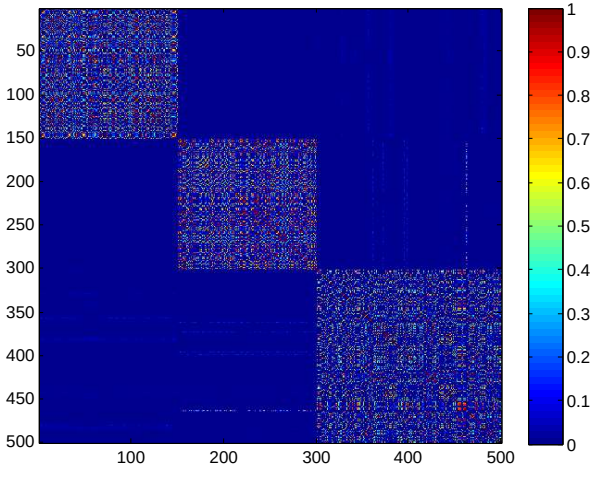
- *For a given data set, which consensus clustering solution should be selected?*
- We propose the validation of clustering combination results at three levels:
  - original data representation
  - clustering ensemble level
  - learned pairwise similarity

Consensus Clustering Validation

- Usually, the assessment of the quality of a consensus partition, P^*, is performed:
  - on the **original data space** - evaluates a data partition using an utility/cost function usually measuring intra-cluster compactness and inter-cluster data patterns separation (e.g. Silhouette, Dunn's, Davies-Bouldin indices), or
  - on the **clustering ensemble space** - rely on the agreement between the consensus partition and the partitions in the clustering ensemble (e.g. ANMI, ACC indices).

Validation on the Similarity Space

- We propose to validate P^* on a learned similarity space:
  - similarities between data patterns are induced from the clustering ensemble using the EAC method - the co-association matrix is used as a similarity matrix;



$$\mathbf{c}_{ij} = \frac{\sum_{l=1}^{N} vote_{ij}^{l}}{N},$$

where $vote_{ij}^{l}$ means that x_i and x_j belong to the same cluster cluster.

  - modifications to the Silhouette, Dunn's, and Davies-Bouldin indices were performed.

Statistical Validity Index based on Pairwise Similarity

- We propose a new validity measure inspired in the Parzen window density estimation technique.
- The likelihood of the data $\mathcal{X}$ given a partition P, is defined as:

$$L(\mathcal{X}|P) = \prod_{i=1}^{N} p(x_i|P), \quad p(x_i|P) = \sum_{k=1}^{K} p(x_i|C_k \in P) \cdot \Pr(C_k).$$

Statistical Validity Index based on Pairwise Similarity

- Following the idea behind the Parzen-window density estimation method, we define:

$$p(x_i|C_k) = \frac{K_N}{|C_k| \cdot V_k(x_i)} \tag{1}$$

- Since we rely only on pairwise similarities, we approximate $V_k(x_i)$ by a quantity proportional to it:

$$V_k(x_i) \triangleq \operatorname{diam}_k(x_i), \qquad \operatorname{diam}_k(x_i) = 2\left(1 - \min_{x_j \in KNN_k(x_i)} \mathbf{C}_{ij}\right) \tag{2}$$

- Using previous equations and estimating $\Pr(C_k)$ as $\frac{1}{n}|C_k|$:

$$L(\mathcal{X}|P) = \prod_{i=1}^{N} \sum_{k=1}^{K} \frac{K_N}{n \cdot V_k(x_i)}. \tag{3}$$

Data Sets

- Five real data sets available at the UCI repository:
  - Iris, Std Yeast, Optdigits, House Votes and Wine;
- Nine synthetic data sets

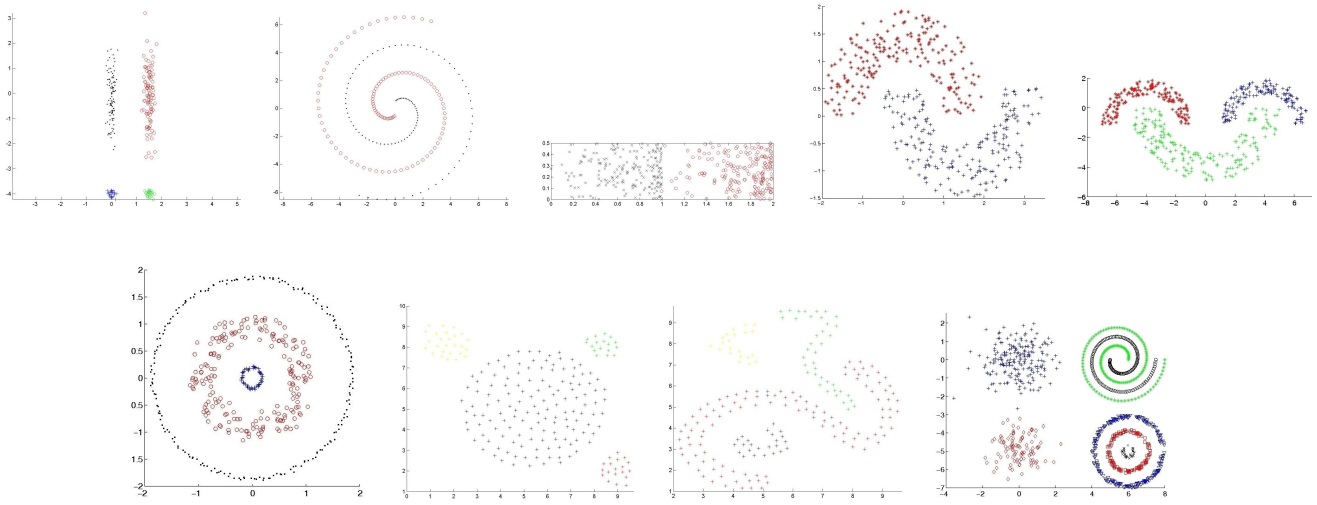
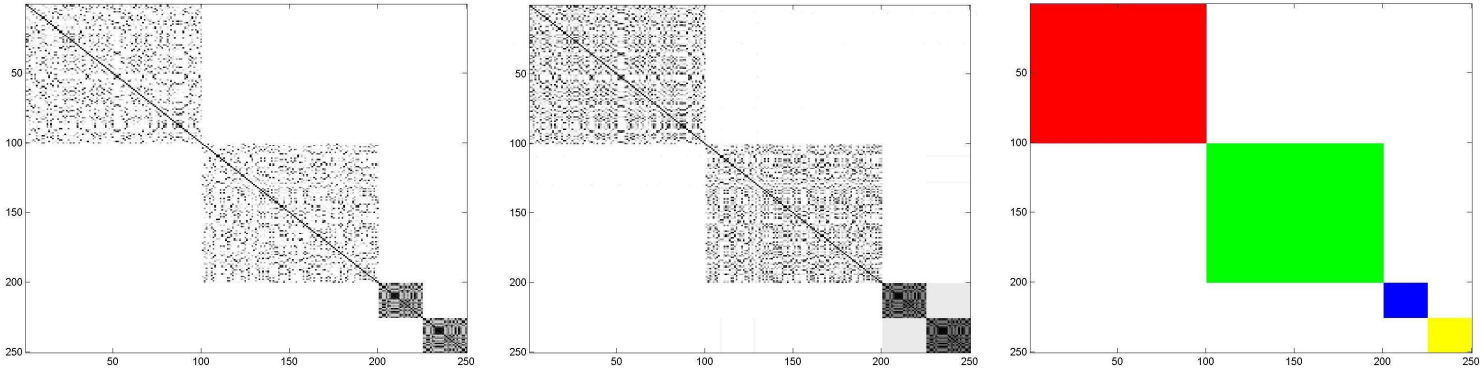


Figure: Synthetic data sets.

Experimental Setup

- Two different methods were used to build the CEs using K-means algorithm and producing 150 data partitions:
  - **A** - with exactly $K = 20$ clusters for Iris data set, $K = 50$ for Concentric data set, $K = 120$ for Complex data set and $K = 30$ for all the other data sets;
  - **B** - was randomly chosen to be an integer in the interval $[10 : 30]$.



(a) Method A (b) Method B (c) Natural Partition

Figure: Co-association matrices

Results

Figure: $NMI(P^*, P^0)$ for the consensus selected by each validity measure.

	Clustering Ensemble Construction Method A											Clustering Ensemble Construction Method B										
Data Set	L_o	S_o	D_o	DB_o	L_s	S_s	D_s	DB_s	ANMI	ACC	Best	L_o	S_o	D_o	DB_o	L_s	S_s	D_s	DB_s	ANMI	ACC	Best
Iris	**0.81**	**0.81**	0.71	0.71	**0.81**	**0.81**	0.71	**0.81**	**0.81**	**0.81**	0.81	**0.81**	**0.81**	0.72	0.72	**0.81**	**0.81**	0.72	0.72	**0.81**	**0.81**	0.81
Std Yeast	0.49	0.49	0.08	**0.53**	0.49	**0.53**	0.24	0.08	0.49	0.49	0.53	0.48	0.48	0.37	0.32	0.48	**0.53**	0.23	0.48	0.48	0.48	0.53
Optdigits	**0.81**	**0.81**	0.71	0.63	**0.81**	**0.81**	0.63	**0.81**	**0.81**	**0.81**	0.81	0.81	**0.83**	**0.83**	0.72	0.81	0.81	0.72	**0.83**	0.81	**0.83**	0.83
House Votes	**0.50**	**0.50**	0.03	0.03	**0.50**	0.14	0.14	0.14	**0.50**	**0.50**	0.50	**0.49**	**0.49**	0.02	**0.49**	**0.49**	**0.49**	0.14	0.14	**0.49**	**0.49**	0.49
Wine	**0.77**	**0.77**	0.66	0.08	**0.77**	0.66	0.08	0.66	**0.77**	**0.77**	0.77	0.77	**0.80**	0.06	0.17	**0.80**	0.77	0.17	0.06	0.77	0.77	0.80
Cigar	**1.00**	**1.00**	**1.00**	0.23	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	0.84	**1.00**	**1.00**	**1.00**	0.84	**1.00**	0.38	**1.00**	0.84	0.84	1.00
Spiral	**1.00**	0.00	0.05	0.05	**1.00**	0.00	**1.00**	**1.00**	0.00	0.00	1.00	0.01	0.01	0.08	0.08	0.01	0.01	**1.00**	**1.00**	0.01	0.01	1.00
Bars	**0.94**	**0.94**	0.06	0.06	**0.94**	**0.94**	0.06	**0.94**	**0.94**	**0.94**	0.94	**0.94**	**0.94**	0.21	0.21	**0.94**	**0.94**	0.21	0.21	**0.94**	**0.94**	0.94
2 Half Rings	**0.99**	**0.99**	0.17	0.17	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.99	0.87	**0.99**	0.21	0.21	0.87	0.87	**0.99**	**0.99**	0.87	0.87	0.99
3 Half Rings	**1.00**	**1.00**	0.08	0.08	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00
Concentric	**1.00**	**1.00**	0.09	0.09	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	0.70	**1.00**	0.14	0.14	0.70	0.70	**1.00**	**1.00**	0.70	0.70	1.00
D1	**1.00**	**1.00**	**1.00**	0.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	0.40	**1.00**	**1.00**	**1.00**	0.40	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00
D2	**1.00**	**1.00**	0.14	0.14	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	0.57	0.71	0.34	0.34	0.57	0.57	**1.00**	**1.00**	0.71	0.71	1.00
Complex	**0.83**	0.44	**0.83**	0.44	**0.83**	**0.83**	0.82	0.82	**0.83**	**0.83**	0.83	**0.70**	**0.70**	**0.70**	0.63	**0.70**	0.63	0.56	**0.70**	**0.70**	**0.70**	0.87
#Best criterion	13	11	3	1	13	11	7	10	12	12		5	11	5	4	6	7	6	9	6	7	

- K_N was defined as $\lceil \sqrt{n} \rceil$.
- Consensus partitions were produced by Average-Link, Single-Link, Complete-Link and Ward's-Link algorithms.

Conclusions

- The learned similarity-based criteria can be used instead of the traditional CE measures;
- The similarity-based criteria are a good option when the original data representation is not available;
- The proposed validity measure is a good choice for consensus clustering validation when the clusters belonging to the CE are not likely to contain patterns of different "natural" clusters.

Questions/Challenges

- How to choose between consensus partitions with different the number of clusters?
- How to assess the quality of a clustering ensemble?
- How to address the validation of overlapping clusterings?