Fuzzy and Evidential Contribution to Multilevel Clustering

LISIC, EA 4491, Université du Littoral Côte d’Opale, Calais, France
IFREMER, Lab. Environnement et Ressources, Boulogne-sur-mer, France

Martin CABOTTE
Pierre-Alexandre HÉBERT
Émilie POISSON CAILLAULT - Presented to Rainsmore

FCTA 2022
25th oct. 2022
La Valette, Malta

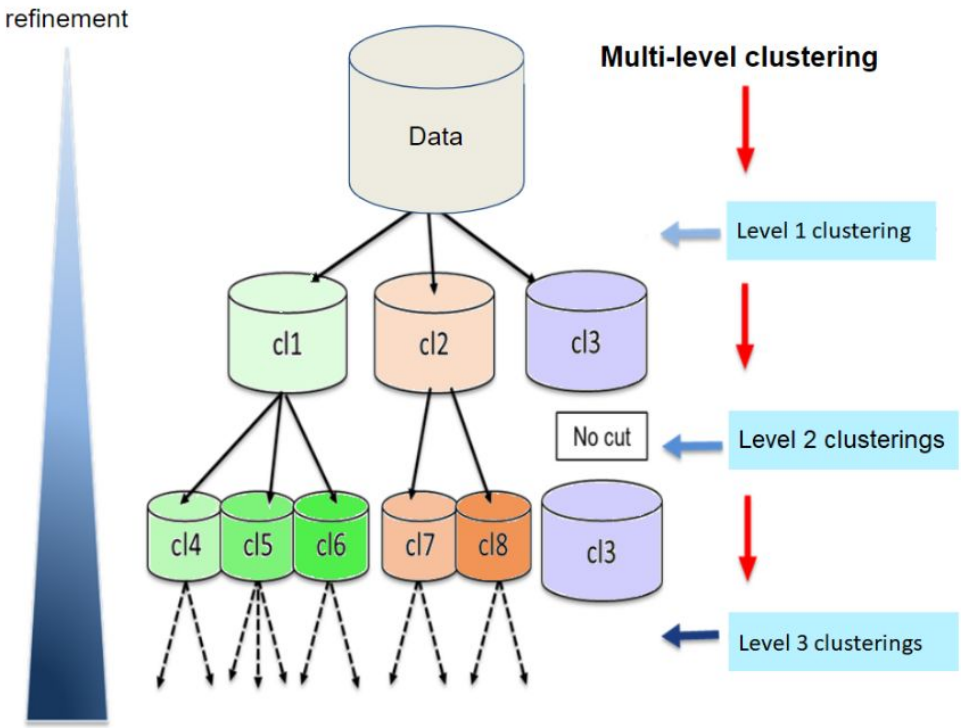
Content

Introduction

- Internship work with LISIC & IFREMER
  - by Martin Cabotte (Master 1, in ULCO's engineering school)
- Collaboration with IFREMER (France)
  - Application in marine resources monitoring
    - Phytoplancton analysis (INTERREG project Dymaphy)
    - Time series of water features (Jerico-Next H2020)

- LISIC Lab (Calais, France)
  - Semi/Unsupervised Classification, Spectral clustering, Fuzzy and Evidence theories
  - A **Multi-level Spectral Clustering** method: MSC (Poisson-Caillault, Grassi)
- Questions:
  - *May fuzzy or evidential framework improve multilevel clustering?*
  - *Which areas for improving multilevel clustering?*

Multilevel Clustering Approach



- Recursive process
- Multi-scale approach
- Key features, for each subdivision
  - Split (cut) criteria

    → decision to subdivise
  - Cluster number (K) estimation

Multilevel Clustering – Split-Criteria

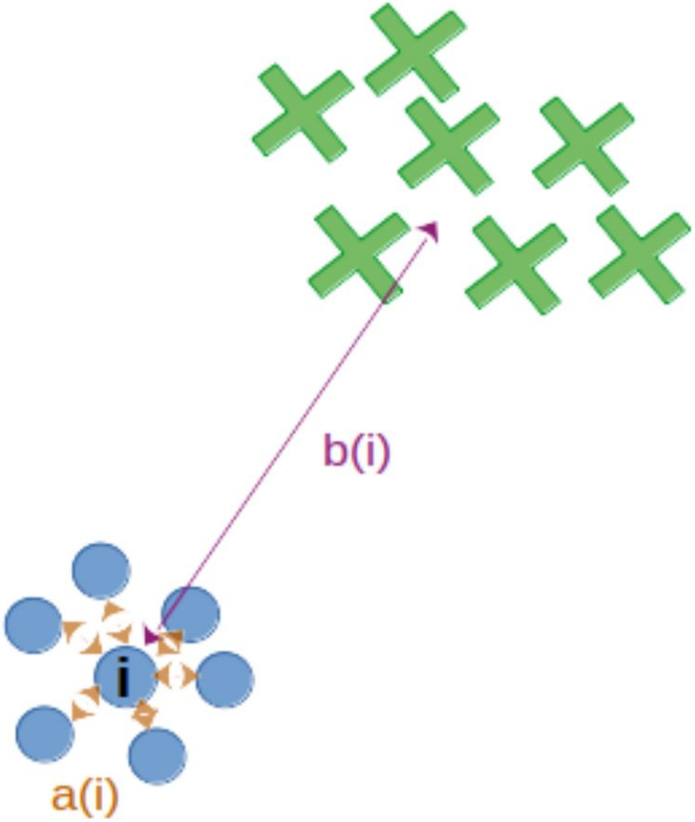
- Crisp split-criteria (a priori)
  - Silhouette (P. Rousseeuw, 1987)

$$Sil_i = \frac{b(i) - a(i)}{max\left(a(i), b(i)\right)} \in [-1, 1]$$

$$Sil_k = mean\left(silhouette_i\right) |i \in C_k$$

$$CardSil_k = \#\left(silhouette_i < 0\right) |i \in C_k$$

  - Degrees of **cohesion** + **separation**
    - If Degree > *Threshold* Then **Stop split**



Multilevel Clustering – Split-Criteria

- Soft split-criterion (Campello, Hruschka)

$$FS = \frac{\sum_{ik} \left(\mu_{pi} - \mu_{qi}\right)^{\alpha} .Sil_i}{\sum_{i} \left(\mu_{pi} - \mu_{qi}\right)^{\alpha}}$$

- **Proposed** soft split-criteria (a posteriori)
  - Degrees of **Non-ambiguity** → **Separation** only

$$Mass100_k = mean_{i \in C_k} m_i(C_k)$$

$$Mass25_k\text{: lowest 25\% masses only}$$

  - Averaged over all clusters: *Mass100* and *Mass25*
- *If Degree* > *Threshold Then* ***Split*** (that is: keep the clustering done)

Multilevel Clustering – Spectral Embedding

- To deal with **non-linearly separable** or **non-globular clusters**
  - Spectral Embedding = Spectral Clustering - K-means
  - Aims at:
    - Concentrating similar objects
    - Making more suitable methods of the K-means family
  - Computation: at **each** subclustering
    - Requires K as input
    - But it may be estimated by some specific methods

Multilevel Clustering – K estimations

- Initial features space
  - A posteriori estimation of K
    - Set as the number between 2 and 10 which maximizes the global *Silhouette* measure of the partition obtained
- Spectral space
  - K obtained from the spectral embedding computation
    - K = Number of “top” eigenvalues
    - K = Dimension of the embedded space

Comparison Protocol

- Algorithms
  - Direct (crisp + soft)
    - K-means (KM), c-means (CM), Evidential-cmeans (ECM)
  - Hierarchical
    - Ward-HClustering, HDBSCAN
  - Multilevel
    - Recursive « Direct » algorithms
- For each algorithm, 2 spaces considered
  - Initial features space
  - Spectral embedding space

Comparison Protocol: Quality Criteria

- Comparison to the **ground-truth** classes
  - For ML methods: "terminal subclusters" only
- Unsupervised criteria
  - **Adjusted Rand Index**: corrected for-chance Rand Index
  - "Non-overlap" score
    - part of the Rand Index which counts the number of pairs of separated points (distinct classes) which are – correctly - assigned to distinct clusters
- "Supervised" criteria
  - Precision: $$\frac{1}{K^*} \cdot \sum_{i \in \{1, \ldots, K^*\}} \frac{TP_i}{TP_i + FP_i}$$ Recall: $$\frac{1}{K^*} \cdot \sum_{i \in \{1, \ldots, K^*\}} \frac{TP_i}{TP_i + FN_i}$$

Comparison Protocol: Parameters Tuning

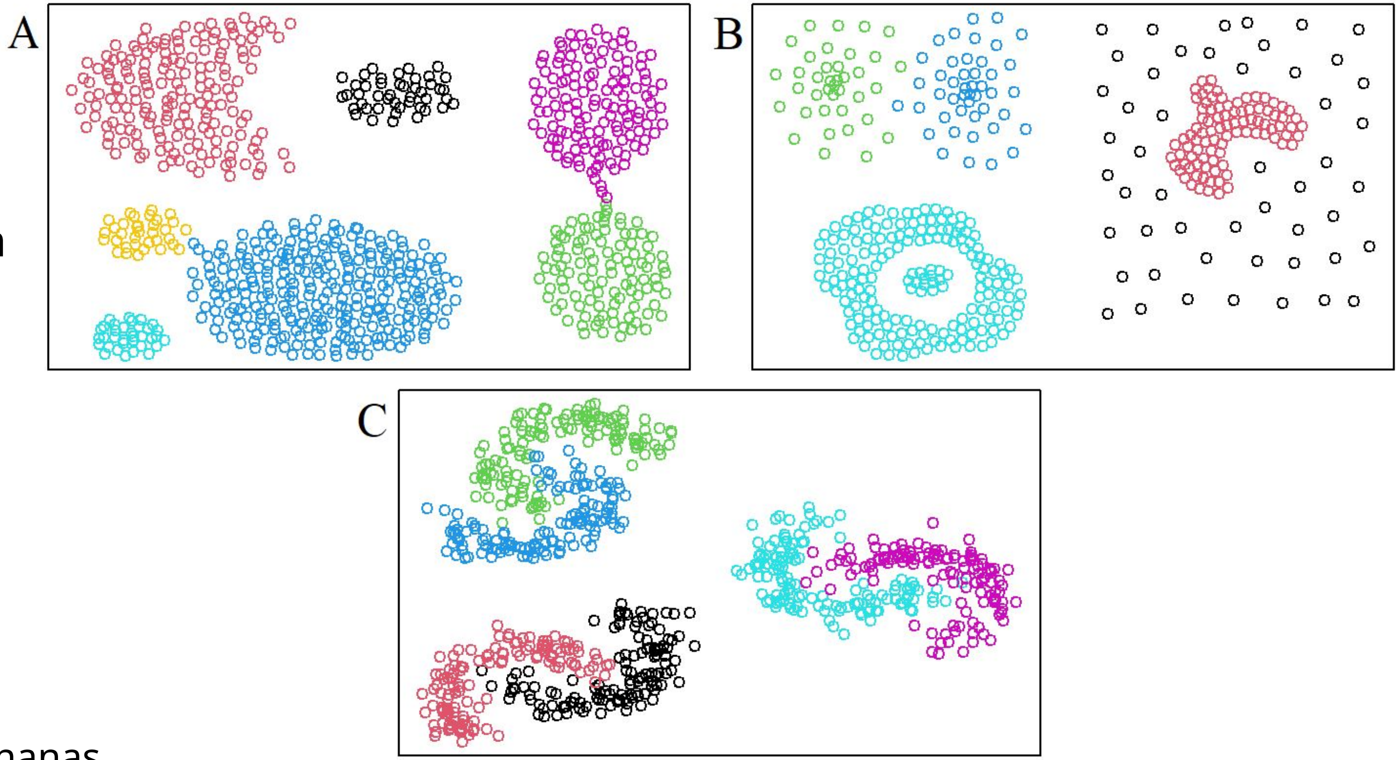
- Direct + Hierarchical clusterings
  - K = ground-truth K^*
- ML-clusterings
  - For each clustering, K is not tuned but estimated
  - Terminal K is set as close as possible to ground-truth K^*, by a **split-criterion tuning**
    - Threshold domain is sampled in 20 values, and best value is kept:

$$v = argmin_v |K(v) - K^*|$$

- HDBSCAN
  - Similar method to tune its *minPoints* parameter

Comparison Protocol: 3 Datasets

- (A) Aggregation
  - ~ Globular clusters
  - Small vs large clusters
  - Some contacts between clusters
- (B) Coumpound
  - Hierarchical structure
    - 3 x 2 clusters
- (C) 6-Bananas
  - High ambiguity
    - 3 x 2 neighbouring bananas



Results and Analysis

- Aggregation & Coumpound Results
  - Spectral space
    - **when** the final K remains close enough to ground-truth K*
      - ML performs well
      - soft ML-Cmeans slightly outperforms ML-Kmeans, particularly with the Mass criteria
  - Initial space
    - Aggregation: direct methods and Ward-HC are better here
    - Coumpound: ML-CM and ML-ECM perform best (with criteria Mass100)
  - Limits of *CardSil* (K<K*), and also *Silhouette* & *Fuzzy Silhouette* (K>>K*)

Results and Analysis

- 6-bananas dataset Results
  - **No true success** (complex dataset: non-separability + noise)
  - Spectral space
    - direct methods and Ward-HC are better here
      - The ambiguity between pairs of bananas is too high, this disturbs the estimation of the spectral space dimension = K
  - Initial space
    - Ward-HC is best
    - ML-CM and ML-KM are not far away
  - Mass criteria: less overclustering than Fuzzy Silhouette

Conclusion

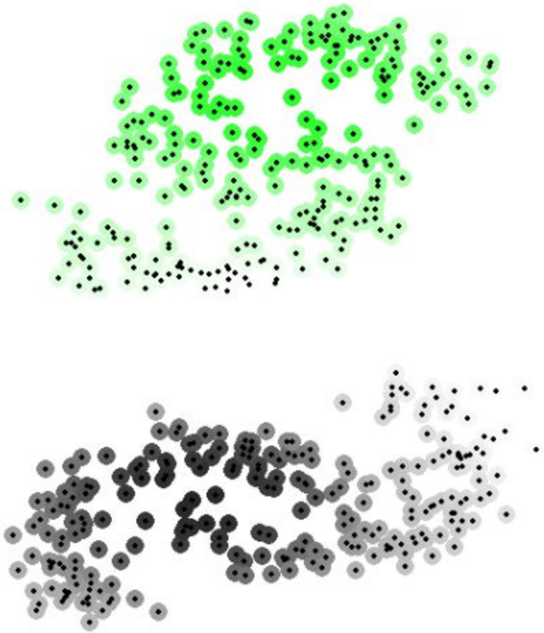
- Toy datasets, not so easy
  - Some clusters are nested, very close to each other, noisy
  - This makes the estimation of K and the decision to split hard (for each sub-clustering of ML methods)
    - A lot of “overclusterings” in ML methods, which leads to low quality scores
    - Non-ML methods do not suffer from this drawback (input)
- Compared to ML-KM, soft ML-CM and ML-ECM can improve results (Coumpound, Aggregate)
- Split-criteria
  - Silhouette variants seems to not perform very well
  - Mass criteria help avoiding overclustering

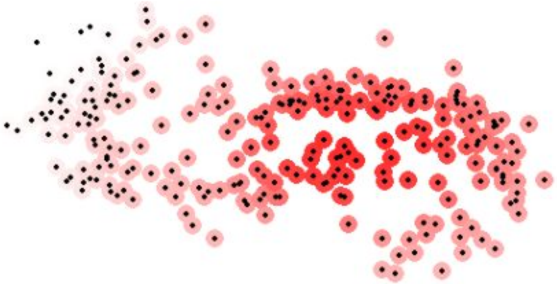
Future works

- Towards soft clustering, with **split & merge** process
  - Here overclustering is a drawback; but a merge process should be able to rebuild fragmented classes
    - This is indicated by the good “non-overlap” scores: points in a same cluster tend to belong to the same class
  - **Use more soft information**, by subclustering points with a weight equal to their non-ambiguity; then re-assign ambiguous points
- Test other split-criteria, and improve the research of the optimal thresholds
  - Obtained K should be contrained to be closer to ground-truth K*
- Improve agreement measurement between ML-clustering and simple clustering
- Look for more convenient ECM methods
  - Compared to C-means, ECM tends to push cluster centers to the border of the space

Future works

- ECM Drawback: center space is empty





Results: initial space

Feature space

	Direct (1)			Agglomerative (2)		ML KM (3)		ML CM (4)			ML ECM (5)		
	KM	CM	ECM	HC	HDBSCAN	CardSil	Sil	FS	Mass25	Mass100	FS	Mass25	Mass100
Coumpound with class fusion K*=5													
ARI	0.57	0.51	0.48	0.59	0.76 - 0.84	0.5	0.28	0.28	0.45	0.8	0.35	0.47	0.83
NonOverlap	0.94	0.95	0.93	0.94	0.94-0.98	0.74	0.94	0.94	0.79	0.97	0.94	0.79	0.94
Precision*	0.84	0.64	0.63	0.91	0.89-0.94	0.47	0.93	0.93	0.49	0.67	0.92	0.44	0.92
Recall*	0.74	0.6	0.59	0.79	0.76-0.9	0.4	0.8	0.8	0.5	0.7	0.8	0.48	0.8
NbClusters	5*	5*	5*	5*	6-9	2	23	24	6	7	12	6	11
Aggregation K*=7													
ARI	0.76	0.74	0.55	0.81	0.81-0.67	0.66	0.56	0.52	0.63	0.59	0.55	0.52	0.52
NonOverlap	0.99	0.99	0.92	1	0.93-0.93	0.98	0.99	0.97	0.93	0.94	0.97	0.95	0.95
Precision*	0.76	0.76	0.47	0.79	0.64-0.64	0.95	0.97	0.79	0.65	0.66	0.76	0.67	0.67
Recall*	0.83	0.83	0.54	0.86	0.71-0.71	0.89	0.93	0.83	0.61	0.7	0.82	0.66	0.66
NbClusters	7*	7*	7*	7*	5-55	14	18	15	13	17	25	14	14
6-Bananas K*=6													
ARI	0.57	0.59	0.57	0.67	0.57-0.03	0.57	0.37	0.37	0.54	0.54	0.38	0.49	0.51
NonOverlap	0.94	0.94	0.93	0.94	0.83-0.98	0.83	0.94	0.94	0.96	0.96	0.94	0.96	0.92
Precision*	0.76	0.78	0.79	0.86	0.25-0.92	0.25	0.73	0.73	0.84	0.84	0.72	0.8	0.63
Recall*	0.76	0.78	0.75	0.82	0.5-0.87	0.5	0.81	0.81	0.84	0.84	0.8	0.79	0.72
NbClusters	6*	6*	6*	6*	3-218	3	75	64	10	10	50	14	14

Results: spectral space

Embedded spectral space

	Direct (1)			Agglomerative (2)		ML KM (3)		ML CM (4)			ML ECM (5)		
	KM	CM	ECM	HC	HDBSCAN	CardSil	Sil	FS	Mass25	Mass100	FS	Mass25	Mass100
Coumpound K*=6													
ARI	**0.49**	**0.43**	**0.43**	0.51	0.86-0.45	0.81	0.36	0.26	0.85	0.85	0.26	0.58	0.58
NonOverlap	0.92	0.91	0.91	0.92	0.94	0.92	1	1	0.94	0.94	1	0.99	0.94
Precision*	0.7	0.52	0.52	0.7	0.92	0.7	0.99	1	0.94	0.94	0.99	0.97	0.94
Recall*	0.67	0.5	0.5	0.67	0.79-0.78	0.67	0.99	1	0.83	0.83	0.99	0.93	0.83
NbClusters	6*	6*	6*	6*	5-7	4	17	28	7	7	21	14	13
Aggregation K*=7													
ARI	0.96	0.95	0.77	0.99	0.99-0.44	0.81	0.33	0.29	0.85	0.29	0.29	0.96	0.45
NonOverlap	1	1	0.99	1	1-0.97	0.93	1	1	0.97	1	1	1	1
Precision*	0.96	0.94	0.77	0.99	0.99-0.96	0.64	0.95	1	0.84	1	1	1	1
Recall*	0.99	0.98	0.85	0.99	1-0.89	0.71	0.99	0.99	0.85	0.99	0.99	0.99	0.99
NbClusters	7*	7*	7*	7*	7-20	5	21	38	14	37	38	8	26
6-Bananas K*=6													
ARI	0.65	0.63	0.64	0.66	0.59-0.57	0.57	0.35	0.32	0.55	0.49	0.32	0.41	0.49
NonOverlap	0.95	0.95	0.95	0.95	0.93-0.93	0.83	0.99	0.99	0.88	0.93	0.99	0.98	0.93
Precision*	0.82	0.81	0.82	0.84	0.63- 0.63	0.25	0.92	0.93	0.46	0.68	0.92	0.87	0.67
Recall*	0.82	0.81	0.82	0.83	0.72-0.71	0.5	0.91	0.92	0.61	0.74	0.91	0.85	0.74
NbClusters	6*	6*	6*	6*	6-8	3	23	24	7	13	24	18	13