Classification and Clustering fusion

Emilie Poisson Caillault.

LISIC, Laboratoire Informatique Signal Image Côte d’Opale

Univ. Littoral Côte d’Opale

Short cv

2001 : Polytech'Nantes engineer in computer science

2001 : DEA Automation and Computer Science, Ecole Centrale Nantes

2005 : PhD, Univ. Nantes at LC2N (IRCCYN Lab).

Architecture and Training of a hybrid Neuro-Markovian System for On-Line Handwriting Recognition

Keywords : TDNN, SDNN, SD-TDNN, MS-TDNN, global discriminant training, MLE-MMI, Mask/Filter in convolution layer.

2006 : Assistant Professor - Univ Littoral in data science and machine learning

2014 : IFREMER delegation

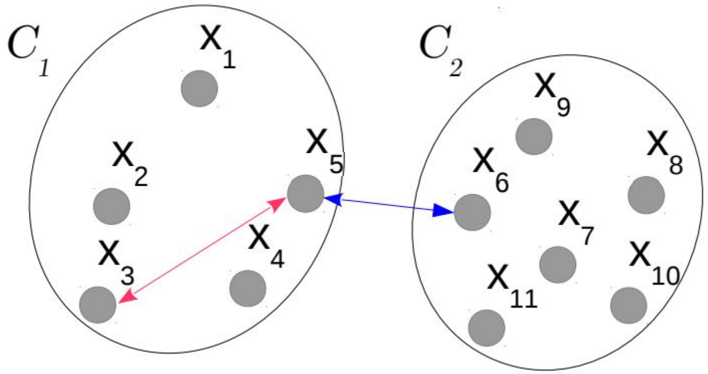
2020 : HDR degree

Contributions to the classification and segmentation of Time series by statistical unsupervised or guided learning

Keywords : similarity, DTW-criteria, DTW-imputation, spectral clustering and multi level approach

JERICO project
CPER IDEAL
ORIENTOI application
LISIC/IFREMER PhD supervision

a. Pattern clustering and classification
b. time series
c. convolutional neural networks
d. hidden markov models
e. fully unsupervised or constrained spectral clustering
f. elastic distance metrics for signal comparison
g. environmental science computing

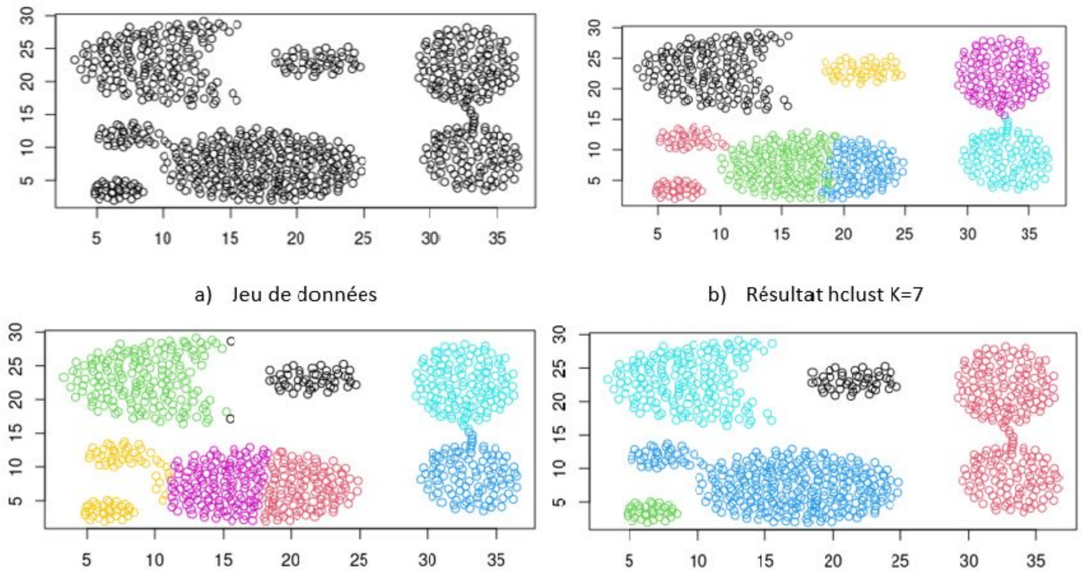


Clustering/Classification:

- two similar points would be in the same class.
- Not well-defined problem
- Similarity relation is transitive within a group
- Between-class Similarity is not transitive: Two objects similar to the same third may not be similar to each other.

Linkage/split process way leads to different structures. (single, average, complete linkage).

Each method is adapted for a kind of shapes.



a) Jeu de données

b) Résultat hclust K=7

c) Résultat K-means K=7

d) Résultat HDBSCAN minPts=4

Linkage/split process way leads to different structures. (single, average, complete linkage).

Each method is adapted for a kind of shapes.

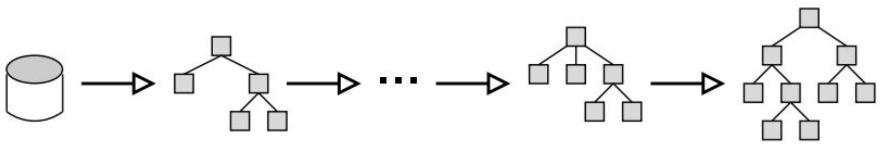
No assumption about dataset → idea: combine many methods.

S(N,K) : Number of possible partitions - N observations to partition in K groups

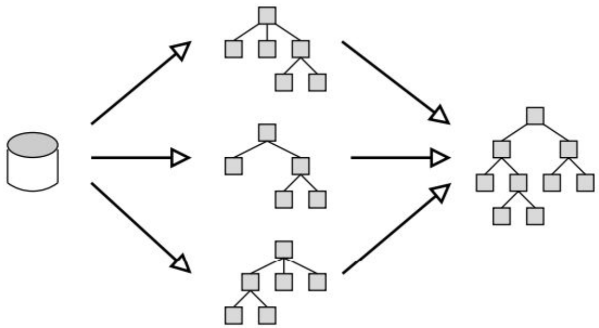
$$S_{N,K} = \frac{1}{K!} \sum_{k=0}^{K} (-1)^k \, (K-k)^N \binom{K}{k} \simeq \frac{K^N}{K!} \text{ quand } N \to \infty$$

Kinds of fusion and collaboration

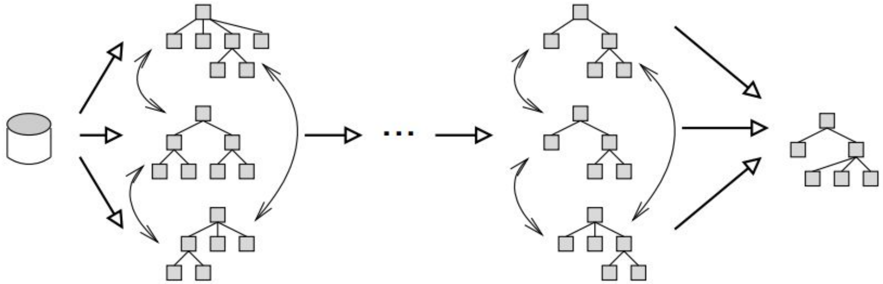
From BDA2017
Gançarski et al.
Clustering collaboratif



(a) Clustering séquentiel

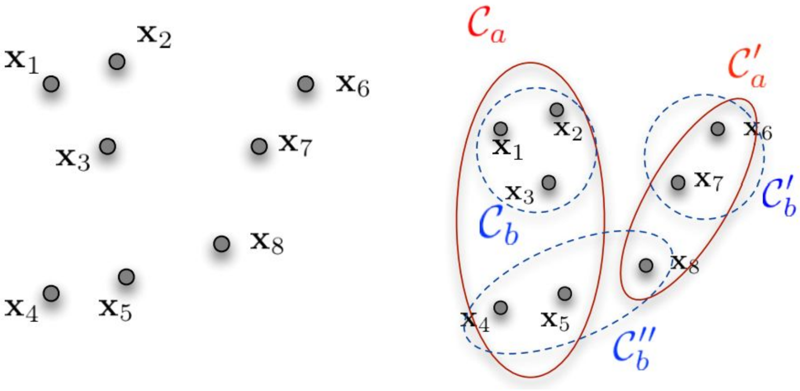


(b) Clustering coopératif



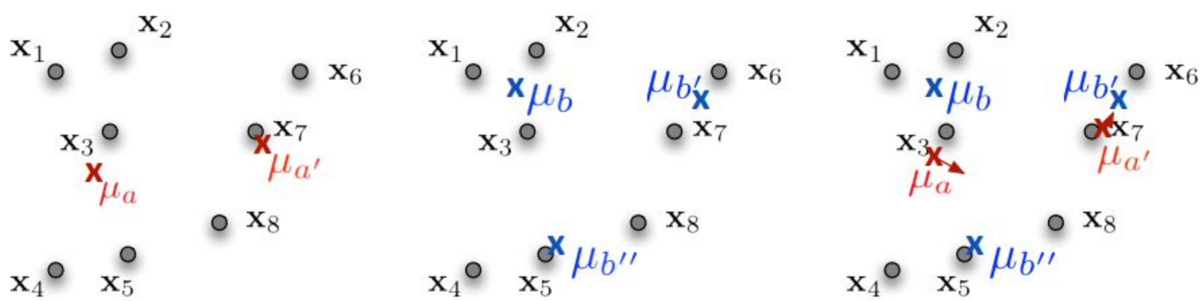
(c) Clustering collaboratif

Easy solution of sequential clustering.



Same dataset, same features.

→ Information transfer by centroid information.

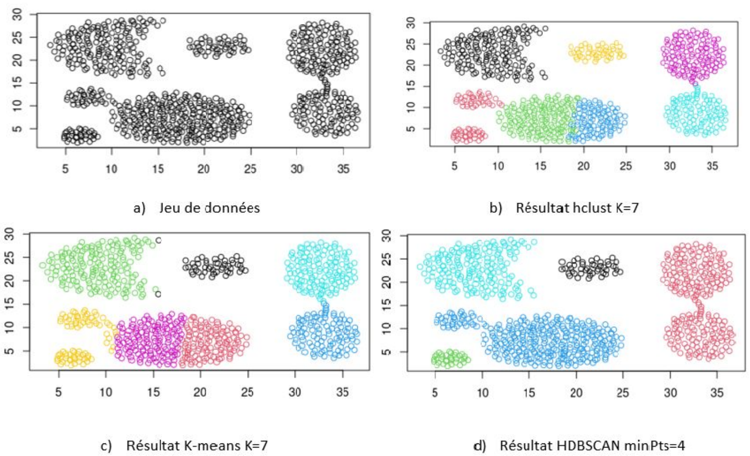


(a) Centres des clusters A (b) Centres des clusters B (c) Nouveaux centres A

But which partition to adapt? based on which criteria?

Cooperative fusion

Find a consensus between all methods



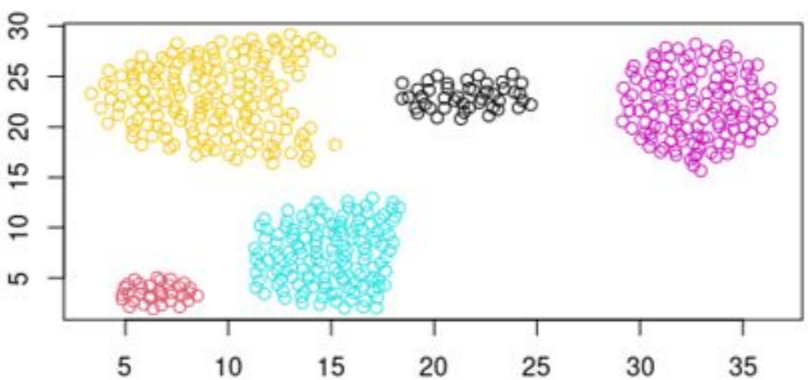
a) Jeu de données

b) Résultat hclust K=7

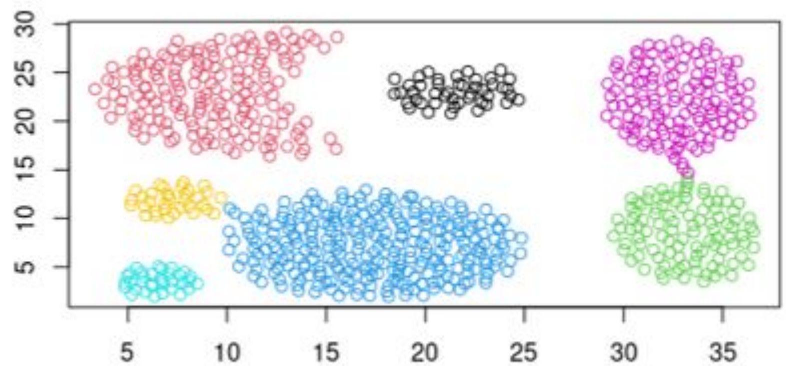
c) Résultat K-means K=7

d) Résultat HDBSCAN minPts=4

Fusion - (KM-HC-HDBSCAN)



Ground truth



Initial Method:

- Apply n clustering methods from observations (same features or not).
- From the n Partitions, compute a Consensus Matrix to adapt

step 1 - Apply clusterings
step 2- Compute confusion matrix M(l,k) et M(k,l) between each clustering partition . C_i,k the class number of the k methods

$$\alpha_{k,l}^{i,j} = \frac{||C_k^i \cap C_l^j||}{||C_k^i||}$$

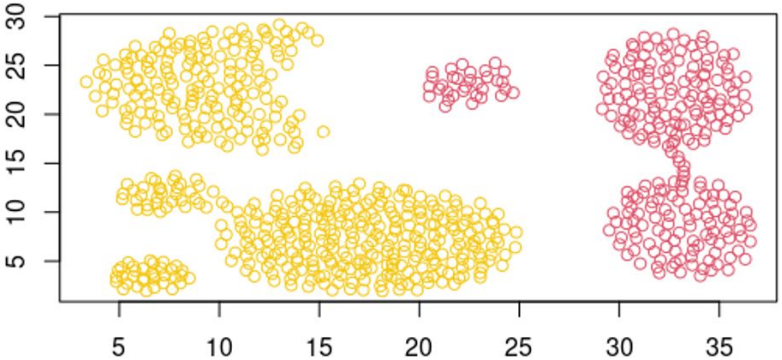
step 3- Compute similarity S(k,l)=M(Ck,Cl)xM(Cl,Ck)
step 4- Compute vote

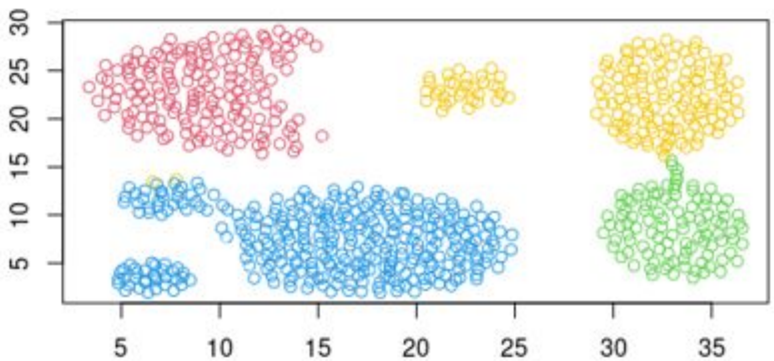
$$vote(pointp, classe_i) = \sum_{m_1=1}^{n} \sum_{m_2=1}^{n} \mathbf{1}|si\ argmax_i(s(m_1, m_2)s(classe(p, m_1), classe_i(m_2)))$$

Improvements:
- weighted vote by confidence score in clustering method
- weighted vote by silhouette score of each class/cluster.

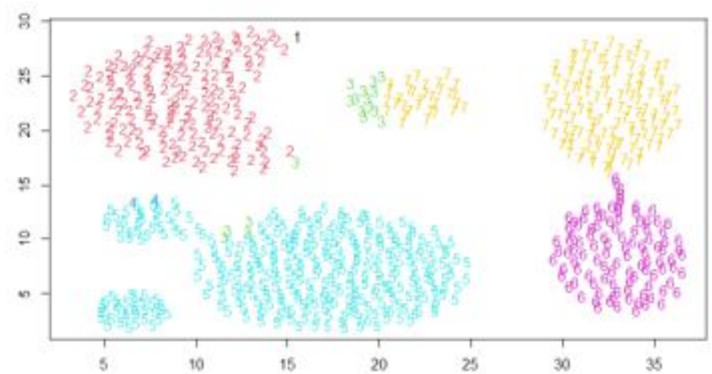
Multi K-means Fusion.

Fusion - (multi-KM, K=1 to 12)





a) Silhouette globale



b) Silhouette cluster par cluster

Other idea ?
Do you want to explore this also ?

Contact us :
emilie.poisson@univ-littoral.fr