No cut criteria for hierarchical clustering

EMILIE POISSON CAILLAULT, ERWAN VINCENT
Université du Littoral Côte d'Opale, Laboratoire d'Informatique Signal Image Côte d'Opale

Introduction

To do this, there is the development of new no cut criteria to determine whether a cut was good or not and whether it should be accepted.

The clustering is done with the Partition Around Medoïds (PAM) version of the Ng-Jordan-Weiss spectral clustering method.
If the no cut criteria are reached, the algorithm abord the cut and retry with some parameters improvements or experts improvements. Else the cut is accepted and the algorithm try to cut again the new clusters.

Method

- First the algorithm calculates the number of clusters to cut using the PEV method. If the number of cluster K = 1 then no cut.
- The dataset is cut using spectral PAM
- The Critsilp criteria is calculated on the result
- After :
  - If the cut is acceptable, then we continue the same method on the new clusters until the criteria is no longer satisfied or the cut becomes impossible
  - Else the clustering is restarted with some improvments
- The algorithm is finished if no more cuts are possible

Silhouette

For the cut criteria assesment part, we are using the silhouette value at an *i* point :

$$Sil(i) = \frac{b(i) - a(i)}{max(a(i), b(i)} \quad (1)$$

with *a(i)* : the mean distance between *i* and all other data points of the same cluster *C*

$$a(i) = \frac{1}{\#C - 1} * \sum_{i'=1}^{\#C} d(i, i') \quad (2)$$

and *b(i)* : the smallest mean distance of *i* to all points in any other cluster C_k

$$b(i) = \min(i, \frac{1}{\#C_k} * \sum_{i'=1}^{\#C_k} d(i, i')) \quad (3)$$

Silhouettes are calculated from the eigenspace or from the normalized eigenspace.

Silhouette criteria for a *C* cluster :

$$CritSil(C) = \#\{sil(i) < 0 \ / \ i \in C\} \quad (4)$$

Silhouette with their values under 0 indicates that the point higtly match with the neighboring cluster. More highly is the value of *CritSil(C)*, more there is confusion in the cut.

Multi-Level clustering

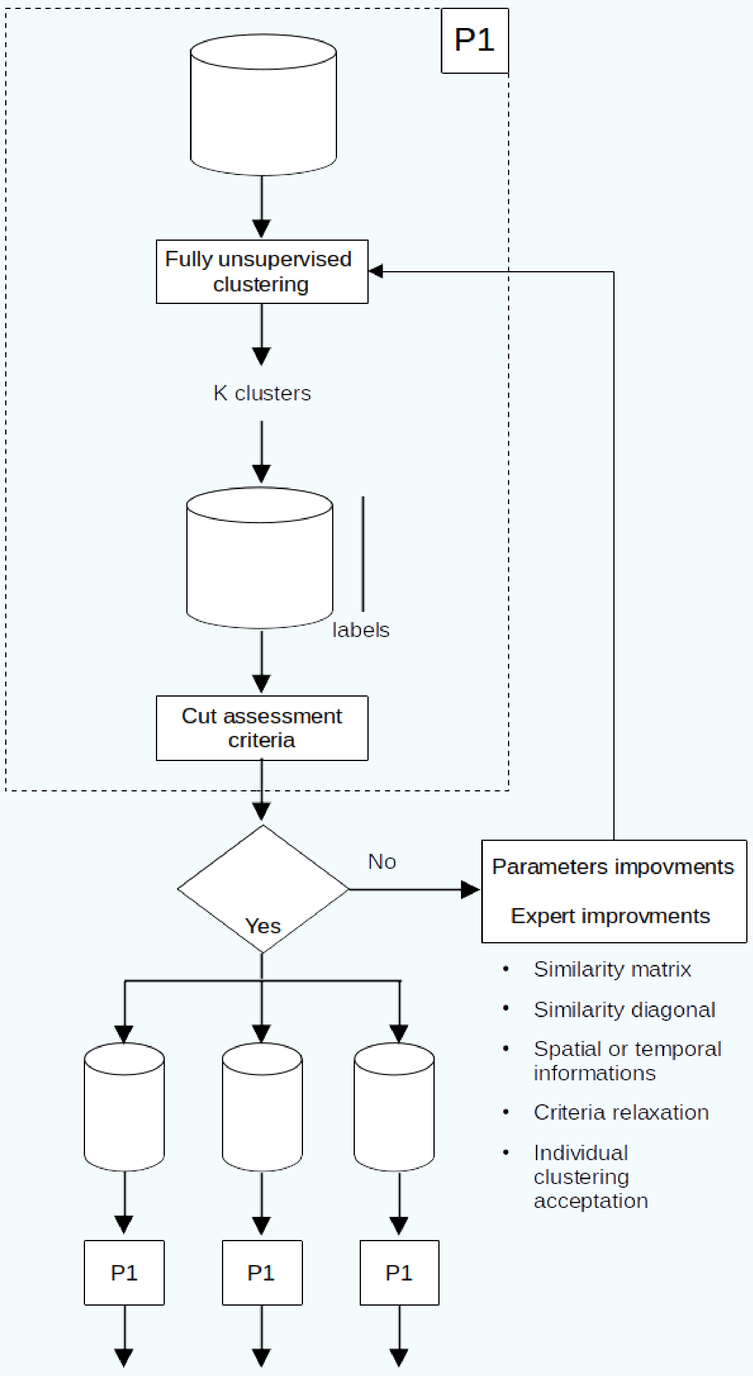

Fig. 1: Multi-Level clustering method

Principal EigenValue (PEV)

The fully unsipervised clustering is using the PEV method to select the number K of clusters to cut.

The PEV method is a method to determine the number K of cluster to cut. This method count the principal eigenvalues (eigenvalues = 1) and this number is the number K of cluster to cut.

The PEV method works as follows :

- If the number n of eigenvalues >= 0.999 is superior to 1 then K = n
- Else K = the number of eigenvalues >= 0.99

Results

This results are made for a similarity neighborhood of 7 and a maximum clustering level of 5 :

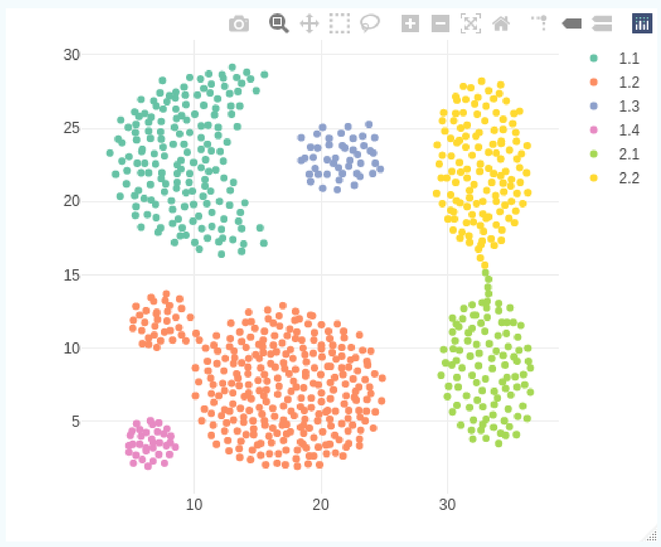

Fig. 2: Aggregation dataset

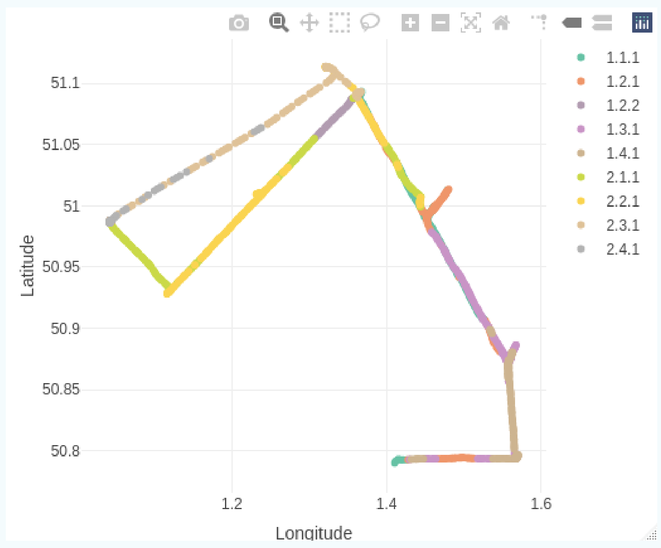

Fig. 3: Phytoplankton cluster

References

[1] Peter J. Rousseeuw: *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, Volume 20, November 1987, Pages 53-65.

[2] Andrew Y. Ng; Michael I. Jordan; Yair Weiss : *On Spectral Clustering: Analysis and algorithm*, NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and SyntheticJanuary, 2001, Pages 849–856.

Partners