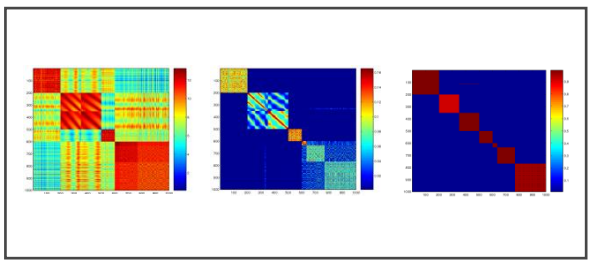
instituto de telecomunicações

Our goal is to learn the pairwise similarity between patterns in order to facilitate a proper partitioning of the data without the *a priori* knowledge of the number of clusters, and of the shape of these clusters.

We propose a clustering ensemble approach combined with cluster stability criteria to selectively learn the similarity from a collection of different clustering algorithms with various parameter configurations.



Research team
> Ana Fred
> André Lourenço

Learning Pairwise Similarity

Overview

Different vectorial and / or (dis)similarity representations can be produced for a given data. These distinct representations or data generating models have typically been used individually, in single classifiers or single clustering algorithms, or simultaneously, as in classifier combination techniques or cluster ensemble methods, depending, respectively, on whether working under a supervised or unsupervised learning approach.

A different perspective to consider consists of learning the intrinsic (dis)similarity relations between patterns, either using a single or multiple possible data representations. In this project, in collaboration with Prof. Anil K. Jain (Michigan State University, USA) we have proposed the learning of pairwise similarity in an unsupervised manner, combining a cluster ensemble approach with cluster stability criteria [1] . The underlying fundamental ideas are:

- each clustering algorithm induces a similarity between data points;
- each clustering algorithm may have different levels of performance in different regions of the (embedded) feature space;
- meaningful clusters can be identified based on cluster stability criteria.
- each stable cluster is viewed as an independent evidence of data organization

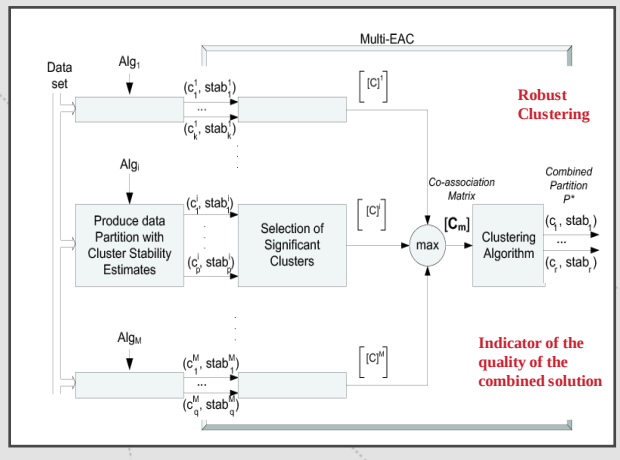


Figure 1. Cluster ensemble approach to learn pairwise similarity and the Multi-EAC method.

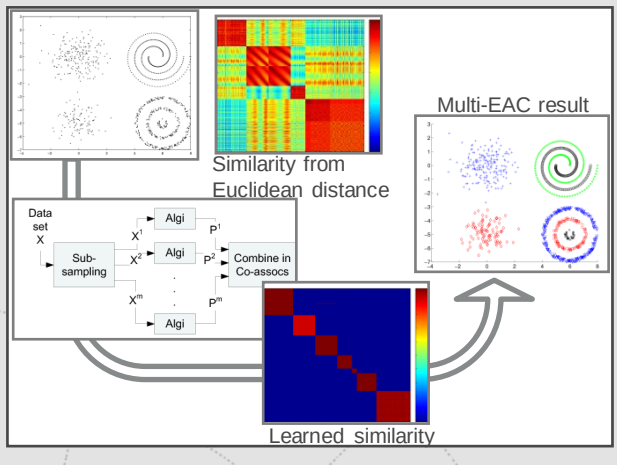


Figure 2. Illustration of the method on a synthetic data set with complex cluster structures.

Contributions

Producing cluster ensembles (by exploring different clustering algorithms or parameter values) and measuring cluster stability based on a subsampling approach, the proposed method selectively learns pairwise similarity by combining only clusters passing a stability test. A schematic representation of the method is given in fig. 1. According to this approach based on cluster stability criteria, it is possible to identify more significant local similarities, that are further combined into a global similarity that better represents the intrinsic organization of the data. This approach was explored in the context of unsupervised learning, and the method was named Multi-EAC, extending the previously proposed Evidence Accumulation Clustering (EAC) method [2] by locally and selectively combining multiple clustering criteria under the clustering ensemble approach. Experimental results of the application of a clustering technique over the learned similarity, have shown that a greater consistency of resulting data partitions is obtained (meaning that different clustering algorithms lead to the same, or approximate, clustering result), with better performance indices, as assessed by consistency measurement of the final data partition with known labeled data, used for validation purposes, when compared with clustering results of individual algorithms and with the previous EAC method. The method is illustrated in figure 2, showing the learned similarity (see pictorial representation of the matrix, in a gradient of colors from blue to red, the red corresponding to highest similarity) and the final data partition obtained over a synthetic complex data set; notice the distinctiveness of the block wise structure of the learned similarity matrix (each block corresponding to a ”natural” cluster), as compared to the Euclidean-based similarity computed over the original feature space. Thus, the proposed method unveils intrinsic pattern similarity, as perceived by stable cluster solutions produced by multiple clustering criteria. The approach has been applied in the analysis of electrocardiographic data, revealing temporal patterns associated with increasing stress levels in individuals performing a concentration task on the computer [3].

Key references

[1] Ana L. Fred and A. K. Jain, “Learning Pairwise Similarity for Data Clustering”, *18th Intl. Conference on Pattern Recognition, ICPR 2006*, Hong Kong, 2006 (recipient of the Best paper award in Pattern Recognition and Basic Technologies from the IAPR).

[2] Ana L. Fred and Anil K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):835-850, 2005.

[3] André Lourenço and Ana L. Fred, “Unveiling Intrinsic Similarity: application to temporal analysis of the ECG”, *Intl. Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2008*, Funchal, 2008.

Contact information: afred@lx.it.pt