Improved Time-Series Clustering with UMAP dimension reduction method

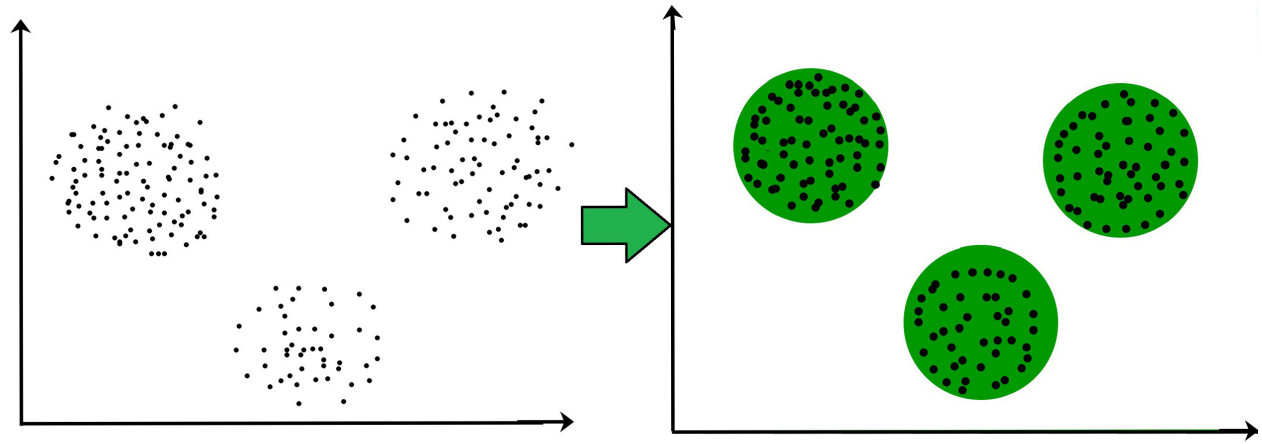
C. Pealat, G. Bouleux,V. Cheutet

University of Lyon - DISP, INSA Lyon - University Hospital of Saint Etienne

2020

Time series clustering

Clustering : Find groups in unlabelled data



1-D Time series : specific data type

Objective : Determine the efficiency of UMAP as a pre-processing step for clustering algorithm.

UMAP : Uniform Manifold Approximation and Projection

- Reduction of dimension algorithm, in general for visualization
- Methodology :

  ↪ Creation of a fuzzy graph G : Edges between [0,1] with respect to the distance

  ↪ Reduction of dimension of this graph with Laplacian Eigenmaps G'

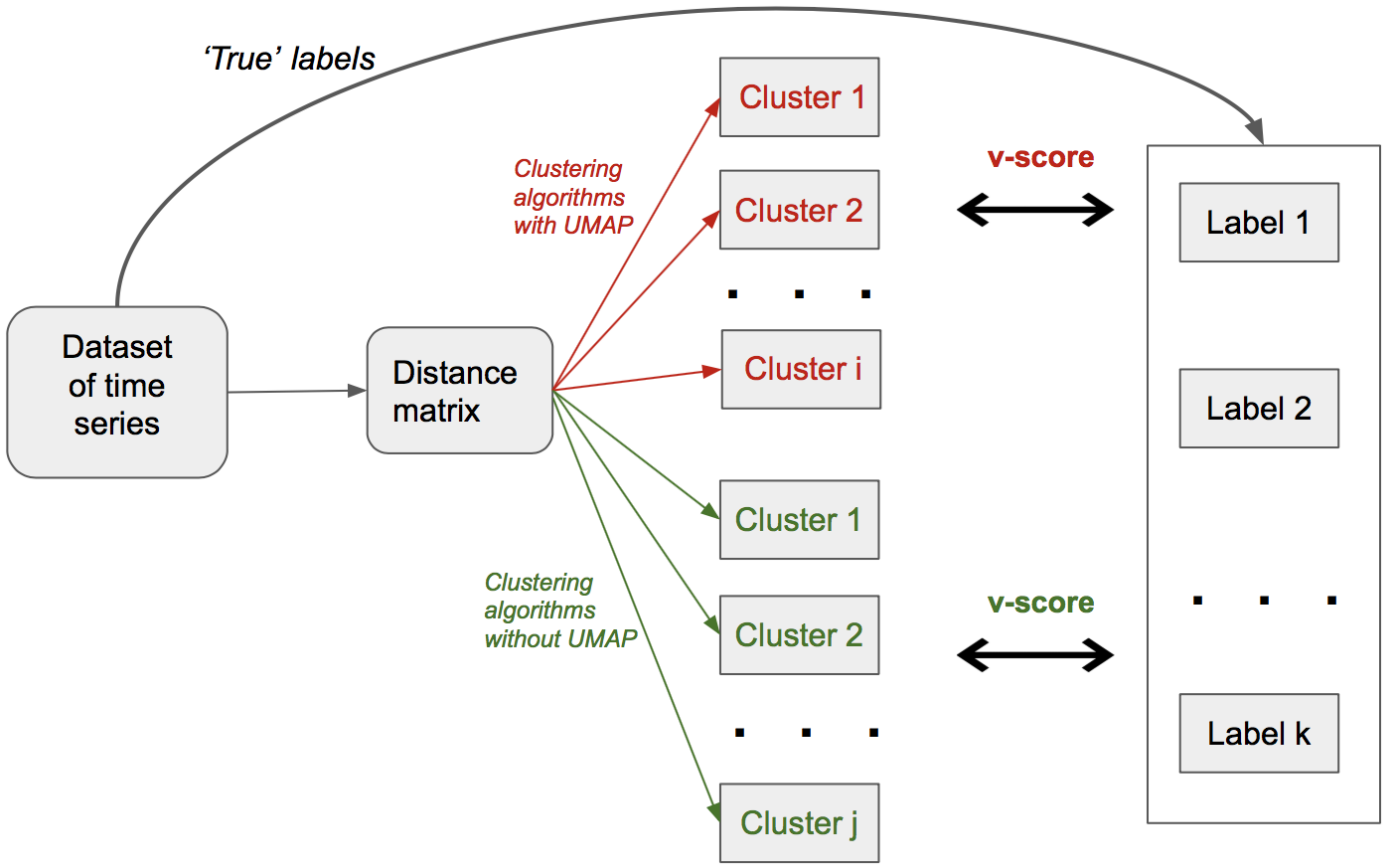
  ↪ Forced directed graph layout to minimized the entropy between G and G'
- What about clustering ?

  ↪ Controversy : It can create pseudo group

  ↪ Test on real data

Creation of a benchmark

- UCR Time Series Classification Archive
- 85 databases of time series with same length, with labels
- Benchmark of clustering results, using v-measure score to compare clustering results and true labels



Three Clustering algorithms

- **K-means**
  - ↪ Centroid-based clustering
  - ↪ Needs : *Distance, definition of a mean, number of clusters* [1]
- **Hierarchic**
  - ↪ Connectivity-based clustering (Dendogram)
  - ↪ Needs : *Distance, number of clusters* [1]
- **HDBSCAN**
  - ↪ Density-based clustering
  - ↪ Needs : *Distance*

1. Determined through silhouette score

Results

Results summarized :

	Kmeans	**Hier.**	**HDB.**	**U.+Kmeans**	**U.+Hier.**	**U.+HDB.**
Mean	0.253	0.212	0.218	0.284	0.273	0.292
Best with U.				61%	66%	75%

↪ With UMAP, improvement of the mean for all 3 algorithms

↪ Improvement for at least 61%of the databases

↪ Particularly efficient with HDBSCAN

And with an other distance ?

Results on the Stiefel manifold

Methodology new distance :

1. Embedding on $\mathbf{R^{n \times p}}$ with delay coordinate embedding
2. Orthogonalization to obtain an element of $V_{n,p} = \{A \in \mathbb{R}^{n \times p} : A^T A = Id\}$
3. New similarity measure : geodesic on the Stiefel manifold (Principal Angle)
4. Karcher mean with respect to the geodesic, allows to apply K-means

	Kmeans	**Hier.**	**HDB.**	**U.+Kmeans**	**U.+Hier.**	**U.+HDB.**
Mean	0.203	0.173	0.182	0.292	0.213	0.301
Best with U.				50%	67%	83%

What to retain so

- Three clustering algorithms and two distances have been tested
- UMAP increased the v-measure score for all possible combinations
- In particular, UMAP coupled with HDBSCAN gave the best results
- UMAP and HDBSCAN can be still better fitted