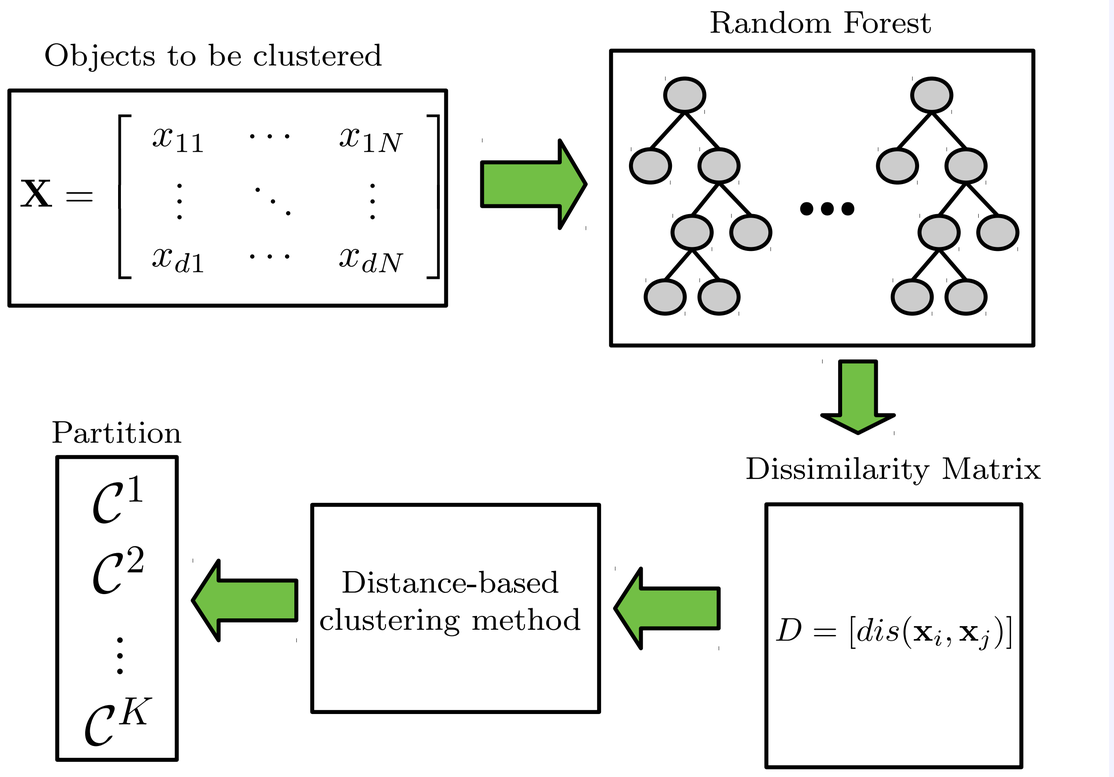
On learning Random Forests for Random Forest-clustering

Manuele Bicego[1], Francisco Escolano[2]

[1] Università degli Studi di Verona, Verona (Italy), [2] Universidad de Alicante, Alicante (Spain)

Random Forest Clustering

- Random Forests have been mainly used for classification and regression;
- Less attention has been paid to the **clustering** scenario
- The most employed approach to RF-clustering exploits the description capabilities of RF to define a dissimilarity measure between points, to be used within a classic distance-based clustering method [1-4].



Motivation and Contributions

Motivation

- In RF-clustering, the problem of learning the RF has received poor attention (main efforts on the derivation of the distance):
- Classic solutions:
  1. (Most used) generation of a synthetic negative class plus training of a standard classification RF;
  2. (Few): usage of a completely randomized RF;
- Our point: **this step is crucial!!**

Contributions

1. Two novel solutions for learning RFs in RF-clustering:
   - **Gaussian Density RF**: RFs designed for density estimation [5] but never used for RF-clustering; the learning is performed by optimizing Gaussian entropy
   - **Rényi RF**: novel RFs introduced in this paper; the learning is performed by optimizing the Renyi entropy, estimated using a non parametric bypass entropy estimator [6] (appropriate when the Gaussianity assumption is too strict – details in the paper!)
2. A thorough experimental evaluation to show that a proper learning of RF is fundamental in RF clustering.
3. A set of guidelines for the different aspects of RF-clustering.

Experimental Evaluation

- We employed 8 standard UCI-ML datasets
- We analyse different options for all the steps of RF-clustering (for a total of 96 configurations):
  - 4 learning strategies (Classification RF, Randomized RF, Gaussian Density RF and Rényi RF), with different RF parametrizations (number of trees, feature subsampling),
  - 4 different distances: Shi [1,2], Zhu2 [3], Zhu3 [3], Ting [4]
  - 4 different distance-based methods: Spectral clustering, Affinity Propagation, Hierarchical clustering (Ward-Link)
- All the results (with statistical tests) are in the paper!

Main findings

1. The classic learning scheme (Classification RFs) is hardly the best solution (only in 2 cases over 96)
2. RFs based on data entropy (Gaussian Density RFs and Rényi-RFs) seem to be an excellent option in this context (better than alternatives in 52 cases over 96)
3. Randomized RFs are a reasonable option (better than others in 16 cases over 96), especially in high dimensional spaces

Guidelines

- **Number of Trees**: few trees (50) seem to be enough;
- **Feature subsampling**: subsampling of features is beneficial;
- **Learning**:
  - if the problem is high dimensional (e.g. dimensionality larger than 10), then use Randomized RF;
  - in the other cases use the Gaussian Density RFs, and check the Gaussianity of the resulting clusters using the Royston's test [7]; if all clusters are non-Gaussian, then re-train the forest with Rényi RF;
- **Distance**: Zhu2 and Zhu3 are both adequate;
- **Clustering**: Spectral clustering.

Results with guidelines

Dataset	Guidelines	Average	Best
Iris	0.8893	0.6173	0.9019 (Gauss,100,0.5,Zhu3,HC)
Wine	0.8426	0.5605	0.8973 (Rand,100,0.5,Zhu3,SC)
glass	0.2430	0.1890	0.3253 (Rényi,100,1,Zhu2,HC)
BTissue	0.4365	0.3389	0.4536 (Gauss,100,0.5,Zhu3,AP)
heart	0.3796	0.1518	0.3797 (Rand,100,0.5,Ting,SC)
Lung	0.1831	0.1430	0.2213 (Gauss,50,1,Zhu2,HC)
Parkinsons	0.1547	0.1003	0.3868 (Rényi,100,0.5,Shi, SC)
Auto-mpg	0.4919	0.3037	0.5224 (Rényi,50,1,Zhu2,SC)

Well above average and not so distant from best!

References.

[1] L. Breiman: *Random forests.* ML 2001. [2] T. Shi et al.: *Unsupervised learning with random forest predictors.* JCGS 2006. [3] X. Zhu et al.: *Constructing robust affinity graphs for spectral clustering.* CVPR 2014. [4] K.M. Ting et al.: *Overcoming Key Weaknesses of Distance-based Neighbourhood Methods using a Data Dependent Dissimilarity Measure.* KDD 2016. [5] Criminisi et al.: *Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning*, Found.Tr.CG.Vis., 2012 [6] Pal et al.: *Estimation of Renyi entropy and mutual information based on generalized nearest-neighbor graphs.* Nips 2010. [7] J. Royston, *An extension of Shapiro and Wilk's test for normality to large samples*, App.Stat. 1982.