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].

Objects to be clustered

\[ X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{d1} & \cdots & x_{dN} \end{bmatrix} \]

Random Forest

Partition:

\[ C^1 \quad C^2 \quad \ldots \quad C^K \]

Distance-based clustering method

Dissimilarity Matrix

\[ D = [\text{dis}(x_i, x_j)] \]

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

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

Well above average and not so distant from best!

References.