Unsupervised Randomer Forests

***Outline**: The unsupervised randomer forest, or UrerF, is an algorithm that is based on the random forest. The main idea is similar, where a root node is recursively divided into two nodes until some boundary condition is reached, but some operating methods, like splitting criteria, feature generation and algorithm output. For a normal random forest, the outcome is a classification or regression result, but for a URerf, the result is a proximity matrix of all the samples where the distance is depicted by the probability of two samples falling into the same leaf node.*

***Feature Generation**: The feature generation method of URerf is similar to that of SPORF, which I have written a summary about, so here I am not going to talk about it. The link of summary is here.*

***Splitting Criteria and Boundary conditions**: The URerf uses the Fast-BIC, which is a combination of Mclust-BIC and k-means algorithms, as the splitting criteria. Instead of fitting a two-component Gaussian mixture model directly to the whole dataset like we do in the Mclust-BIC approach, we search for every possible split point and fit two Gaussian model to the two sides of the split separately. The split with the smallest BIC value is considered the best one.*

***Proximity Matrix**: The output of URerf is a n ×n matrix of tree distances of every two samples, and the 'distance' is defined using the following equation:*

$$\mathrm{D} = \frac{L_{i,\mathrm{j}}}{S_{i,j}}$$

In which, $L_{\mathrm{i,j}}$ is the count of sample i and sample j fall into the same leaf node, and $S_{\mathrm{i},j}$ is the count of all the trees that takes sample i and sample j in the bootstrap sampling.

***Sidenote**: Because of the lack of labels in unsupervised learning, the splitting criteria is changed from the gain of 'node purity' to most likelihood estimation.*