# Cluster ensemble based on Random Forests for genetic data

September 2019

Clustering is very useful in population structure analysis, where individuals are grouped based on shared genetic variation. This population structure analysis is crucial for any further analysis of genetic data. Random forests are suitable for this as they can handle high dimensional data. Random Forest derived proximity measure combined with clustering technique can determine the underlying structure of genetic data.

Single nucleotide polymorphisms are the most common type of variation used to infer population structure. A cluster ensemble approach will be very efficient at this. SNPs can be homozygous or heterozygous.

Two types of clustering methods can be employed here - distance based or dimension reduction based approach. As RF provides a natural method for measuring proximities between individuals, it can handle the linkage nature among genetic markers. For using RF for unsupervised learning, RF algorithm must randomly generate synthetic data based on the original dataset.

The m ethodology has three parts

1. unsupervised Random Forest
2. ensemble construction
3. consensus function

In ensemble construction, forest is constructed in an unsupervised fashion and the constrcuted forest is parsed to compute the proximities between the individuals. Clustering technique is applied onm the resulting proximity matrix. A final clustering is done in consensus function by making use of co-association matrix with agglomerative clustering.