Downstream Performance Benefits of TreeCluster

Akash Atul Boghani, Inderjot Singh Saggu, Raghav K. Subramanian

Problem Definition:

Medical problems like HIV transmission, and gut infections are studied using phylogenetic clustering of data from stool samples. The clustering helps identify disease-causing microbes as well as the percentage of each microbe in the patient. With a huge amount of data, and a machine learning model, we can then learn to classify a patient as potentially sick or healthy. In this project, we will evaluate the performance of TreeCluster, an optimal-clustering algorithm, based on its efficacy in downstream applications.

Objective:

TreeCluster is a phylogenetic tree clustering tool that finds the minimum number of leaf-clusters based on heterogeneity constraints.[1] Such an algorithm has several benefits for downstream applications like divide-and-conquer multiple sequence alignment and HIV transmission clustering. Our objective is to analyse performance improvements using common machine learning models like Random Forests, SVMs and Linear Regression. Additionally, we plan to evaluate the threshold phylogenetic parameter values that fits best for the clustering based on downstream applications for the given dataset.

Data:

The dataset, provided to us by Prof. Mirarab includes:

- Phylogenetic tree of test data in Newick format.
- BIOM test samples of 800 individuals with 20,000 features.
- Pre-known healthy/sick labels for each of the 800 individuals.

Methodology:

Our work will focus on three algorithmic techniques: Random Forests, SVMs and Linear Regression - all of which are proven techniques for binary classification. Our base case will consist of these evaluations without dimensionality reduction, and performance will be evaluated by cross validation and AUC metrics. This will done for two clustering methods - one based on minimum diameter of the cluster, and another based on sum of branch lengths as the limitation for clustering. The third approach which clustering is based on in the paper is single linkage is ignored for further analysis.

We will then attempt to analyse performance with varying degrees of dimensionality reduction, since it is known that our original dataset is rather sparse and we can find better features for classification. Dimensionality reduction in this case can be achieved by either statistical or phylogenetic feature selection, and we will evaluate the performance of either approach. Our aim will be to identify, in each case, the optimal amount of dimensionality reduction (filtering).

Given the availability of time, we will explore the possibility of augmenting the capabilities of the TreeCluster tool so as to generate an arbitrary number of optimized solutions, rather than the one solution that it presently generates.

References:

[1] Balaban, Metin, Niema Moshiri, Uyen Mai, and Siavash Mirarab. “TreeCluster : Clustering Biological Sequences Using Phylogenetic Trees.” *BioRxiv*, 2019, 591388.

[2] Statnikov, Alexander, Mikael Henaff, Varun Narendra, Kranti Konganti, Zhiguo Li, Liying Yang, Zhiheng Pei, Martin J. Blaser, Constantin F. Aliferis, and Alexander V. Alekseyenko. “A Comprehensive Evaluation of Multicategory Classification Methods for Microbiomic Data.” *Microbiome* 1, no. 1 (2013): 11.