Downstream Performance Benefits of TreeCluster

Akash Atul Boghani, Inderjot Singh Saggu, Raghav K Subramanian

Introduction:

Phylogenetic Clustering of data can serve as a powerful tool for studying medical infections and diseases. Data on the percentage of Disease-Causing microbes in every patient and a machine learning model can help classify a patient as potentially sick or healthy. In this report, we evaluate the performance of Tree-Cluster, an optimal clustering algorithm, based on its efficacy in downstream applications.

Problem Definition:

TreeCluster[1] identifies the minimum number of leaf-clusters based on heterogeneity constraints and is useful for HIV transmission clustering. Here, we use stool samples of patients, with details of nucleotide percentages, to obtain Tree-Cluster Clustering result. This grouping is analyzed against machine learning models like Logistic Regression and Support Vector Machines with dependency on the threshold value for clustering also evaluated. Additionally, regressions are performed with clustering based on the maximum diameter of the cluster, sum of branch lengths and single linkage and with varying degrees of dimensionality reduction. The sparsity of the dataset implores us to attempt dimensionality reduction. We hypothesize that with either statistical/phylogenetic dimensionality reduction, we can filter the dataset to most relevant features with moderate accuracy predictions from our classification models.

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.
- FASTA file containing the mapping between features and leaves of the Phylogenetic tree

TreeCluster Threshold Value Variation:

As we can see from the plots, the dependency on the threshold distance for the number of clusters is a hyperbolic curve that flattens as the threshold value increases. In case of Single linkage and max diameter, since the threshold distance is dependent only on the distance between two arbitrary nucleotides in the tree, the value becomes with a hyperbolic curve, flattens, linearly decays and stabilizes. When we see the sum_branch method results, they are similar in that they reduce in the similar hyperbolic-flat-linear decay fashion but over a much longer period as shown below.

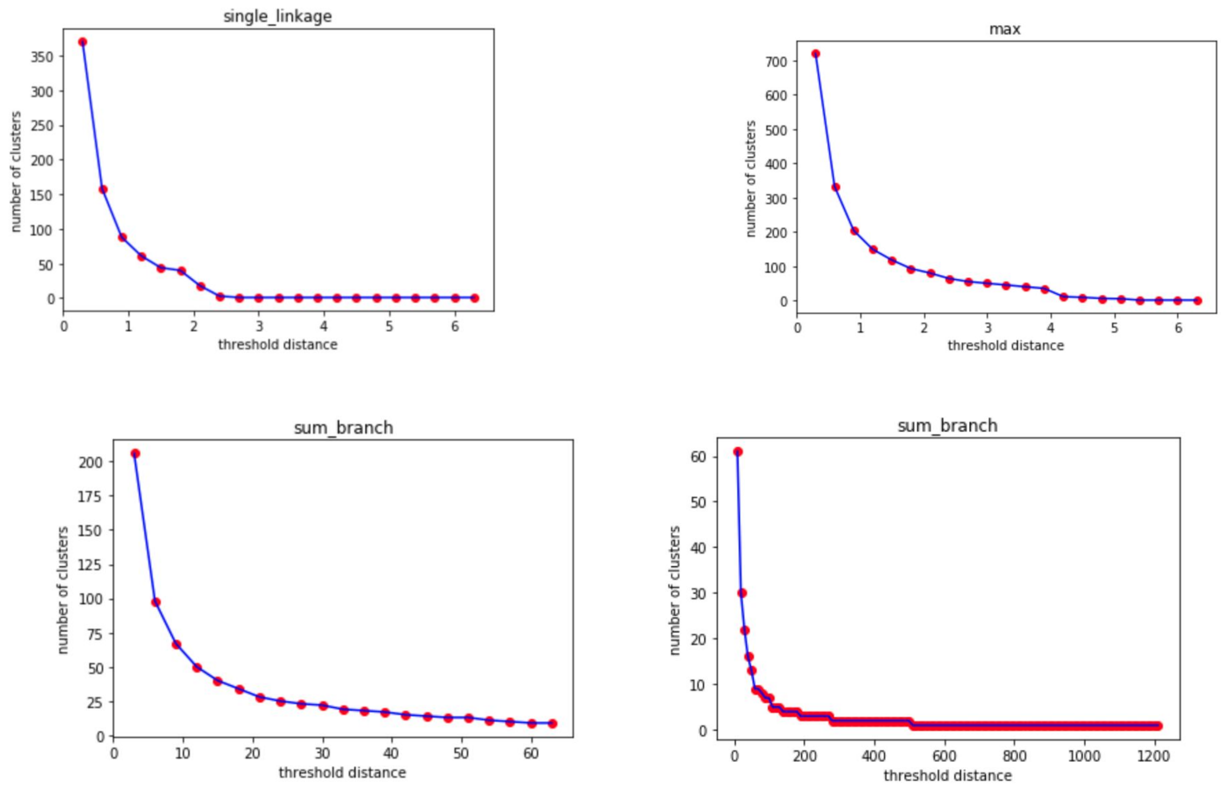


Fig (i) : Tree Cluster method vs Threshold Value plots

Algorithm for Closest Threshold Value Search:

For the search algorithm to work, we use a random start input in the range of expected values of threshold distance from TreeCluster, some number of coarse iterations and fine iterations to identify a good result. Random start values are sampled along the distribution of possible values for the method. We use an objective function

$$|1(expected_{clusters}) - 1(actua_{clusters})|$$

here, for estimating a loss value. Also, the gradient of the loss is taken to be

$$0.01 \times loss \times \frac{(10^{-order(loss)})}{10^{j}}$$

where j is the number of fine variants. This keeps incrementing, until j is equal to the number of fine adjustments, and for every such case, we run i coarse iterations till i is equal to the number of coarse adjustments. In every iteration, we add the gradient to the loss value and move on. We store the best possible values for loss and minimum epochs, and this is our result that we return.

This algorithm will give us a randomized result based on the starting point, the number of coarse and fine iterations, the range of the expected threshold distance values from TreeCluster. With the right number of epochs, we can always converge to a good solution, and non-determinism helps us explore the sample space for cases where support is not $-\infty$, or more complex trees.

Results:

Given DNA fragments collected from biopsy and stool samples we want to predict the presence/absence of Inflammatory Bowel Disease (IBD) disease in the participant. We model this problem as a binary classification task and use Logistic Regression and Support Vector Machines (SVMs) models to learn the necessary mapping. The given data is extremely sparse which necessitates the need for feature dimensionality reduction. We use mainly three techniques for this purpose:

- Principal Component Analysis (PCA)
- Truncated Singular Value Decomposition (T-SVD)
- Phylogenetic Clustering (Max, Sum Branches and Single Linkage methods) followed by PCA in each output cluster so as to extract the max-variance feature vectors.

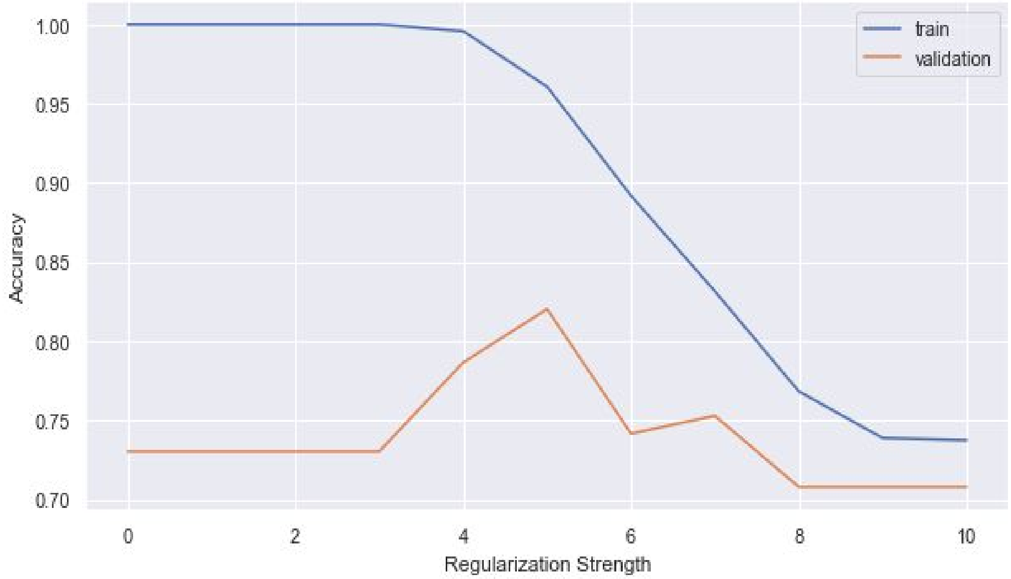
For comparison between the techniques, we used the algorithm described earlier that determines distance threshold that would result in a desired number of clusters.

Evaluation: We split our data in 80:10:10 split with a fixed seed value for consistency, we use ROC score and F (F_1) score as evaluation metrics since it considers both the precision and recall for computing the score.

Logistic Regression												
Clusters/Features	Full Dataset		PCA		Truncated SVD		TreeCluster Max selection		TreeCluster Sum Branches selection		TreeCluster Single Linkage selection	
Entry Type	ROC	F score	ROC	F score	ROC	F score	ROC	F score	ROC	F score	ROC	F score
500	**0.65**	**0.72**	0.66	0.69	0.68	0.73	0.51	0.58	0.44	0.51	0.50	.57
250	0.65	0.72	0.63	0.67	0.58	0.64	0.52	0.59	**0.57**	**0.64**	0.45	0.52
100	0.65	0.72	0.70	0.75	0.64	0.70	0.49	0.54	0.44	0.50	0.52	.58
50	0.65	0.72	**0.77**	**0.79**	0.59	0.66	0.49	0.55	0.46	0.51	0.50	0.55
25	0.65	0.72	0.57	0.63	0.56	0.62	0.48	0.53	0.51	0.55	0.51	.55
10	0.65	0.72	0.53	0.63	0.51	0.61	0.48	0.52	0.48	0.52	0.5	.53

Table (i) : Logistic Regression Results

‘**Full Dataset**’ has no clustering or dimensionality reduction and is our baseline. To understand the extent of overfitting, we plotted training and validation accuracy as a function of regularization strength (L2).



We can observe the severity of overfitting which furthers the need for feature space dimensionality reduction. Without dimensionality reduction, no amount of regularization can help us bridge the gap between training and validation performance without compromising overall accuracy. Though we do observe that regularization strength of 5 (C = 10^5) gives best validation accuracy.

Support Vector Machine (SVM)												
Clusters	Full Dataset		PCA		Truncated SVD		TreeCluster Max selection		TreeCluster Sum Branches selection		TreeCluster Single Linkage selection	
Entry Type	ROC	F score	ROC	F score	ROC	F score	ROC	F score	ROC	F score	ROC	F score
500	**0.79**	**0.83**	0.65	0.69	**0.80**	**0.83**	0.49	0.52	0.47	0.51	0.49	0.56
250	0.79	0.83	0.66	0.70	0.79	0.83	0.52	0.57	0.49	0.56	0.50	0.56
100	0.79	0.83	0.64	0.68	0.69	0.74	0.48	0.52	0.53	0.59	0.51	0.55
50	0.79	0.83	0.63	0.68	0.67	0.72	0.55	0.60	0.48	0.52	0.50	0.53
25	0.79	0.83	0.66	0.69	0.69	0.73	0.55	0.68	0.49	0.52	0.50	0.53
10	0.79	0.83	0.67	0.69	0.68	0.73	0.49	0.52	0.5	0.53	0.50	0.53

Table (ii) : Support Vector Machine Results

Discussion:

- When we look for grouping with a small number of clusters, we see that it is very hard to obtain a sensible statistical grouping. Dimensionality reduction methods like Principal Component Analysis do not decorrelate the data well, and results for a low number of clusters provide features with negative/complex values which is meaningless here. Therefore, especially for dense clustering of features, phylogenetic clustering provides us results where conventional statistical methods fail.

- It is interesting to note that conventional PCA based dimensionality reduction showed good results in Logistic Regression, but the Truncated SVD reduction showed better performance when used with an SVM.

- The phylogenetic feature selection, in general, shows poorer results than we would have expected from it, pointing to perhaps a missing, non-trivial mapping from clusters to selected features. While we performed PCA for each of the clusters hoping to extract the maximum variance from each cluster, it might be prudent to explore the possibility of differential weighting in the future; where singleton clusters are weighed lower to larger clusters.

- Statistical methods like SVD and PCA require intensive steps like identifying the decomposition of a matrix. In comparison, the runtime of phylogenetic clustering is minimal, and typically around O(n) time complexity for different clustering methods on TreeCluster.

- The amount of data available for clustering may also be an indicator of the quality of results obtained. Since we have a highly sparse dataset here, with a higher percentage of data, it is possible that the benefits of training the model are seen in the form of better classification F-score and ROC metrics.

- Both Statistical and Phylogenetic feature reduction have their own advantages in terms of run times and accuracy. For accurate predictions towards medical applications, an appropriate combination of these may prove to be useful compared to the individual approaches for carrying out a swift, and reliable classification task.

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, #1 (2013): 1.