Abstract

A novel algorithm, Tensor Iterative Random Forest (TiRF), is able to effectively build forests that can be mined for interactions within a multidimensional X matrix, a multidimensional Y matrix, and interactions between multiple dimensions in X and Y, simultaneously. TiRF uses dimension reduction techniques (such as Lasso or a nested call to iRF) on Y for the given subsets of X, thus ensuring that the new subset of Y dimensions is highly connected to the X being split upon. The dimension-reduced Y matrix can then be used in measuring node purity, ensuring that noise is reduced and that the given features in X are measured for the ability to split the appropriate dimensions in Y. The resulting trees and forest now contain paths in X, and each node also has an associated set of Y. This means that random intersection trees can be used to find sets of interacting Xs from the forest and sets of interacting Ys conditional on those sets of X. As the data sets grow exponentially, peta-scale and exascale deployments of TiRFs are being targeted to perform systems-wide analyses of biological or other complex systems that can be represented as matrices.

Iterative Random Forest (iRF)

Random Forest (RF) begins with the construction of decision trees with a random subset of the training data and a random set of variables to split on at each node. These decision trees are bagged, or averaged together, to produce a forest with both predictions and feature importance, with a much lower chance of overfitting than found in the trees that are decision trees[1]. After an initial RF, iRF uses the resulting importance measurements as feature weights in the new iteration. These subsequent forests are run on a new, smaller, more efficient feature set by eliminating features with zero importance. This process is repeated until the model no longer improves prediction of a designated test data set or out-of-bag error within the model[2].

Example Results

The results from iRF and TiRF can be used to produce a network by connecting the targets of each model (metabolites) to the features (SNPs) and groups of features (SNP groups) if they meet a predetermined importance cutoff. The SNP groups can also be connected to associated genes. For this network, the SNPUs and associated genes come from the phytozone[6] dataset for Pupulus riciprurus[5]. The resulting fully constructed network can be seen in Figure 4 with the green nodes representing metabolites, purple nodes representing SNPs, orange nodes representing groups of SNPs, and blue nodes representing genes. The zoomed portion of Figure 4 shows a more detailed view of a local network topology surrounding a metabolite. Using this style of view it is possible to see what SNPs and combinations of SNPs (and the genes corresponding to those SNPs) may have an effect on a given metabolite and what possible relationships metabolites may share through SNPs or genes.

References


Acknowledgements

Funding provided by The BioEnergy Science Center (BESC) and The Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. This research used resources of the Oak Ridge Leadership Computing Facility and the Oak Ridge National Laboratory Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.