Objective

1. PPforest package implements a projection pursuit classification random forest.
2. Adapts random forest to utilize combinations of variables in the tree construction.
3. Projection pursuit classification trees are used to build the forest. (from PPtree/viz package)

PPtree algorithm

Combines tree structure methods with projection pursuit dimension reduction.

1. Each node a PP index is maximized to find the optimal 1-D projection.
2. Reduce the number of classes to two, by comparing means and assign new labels, G1 or G2.
3. Re-do PP with these new group labels finding the 1-D projection, y = αx + c.
4. Calculate the decision boundary c, keep α and c.
5. Separate data into two groups using new group labels G1 and G2.
6. For each group, stop if only one class else repeat the procedure, the splitting steps are iterated until the last two classes are separated.

Projection pursuit

Find interesting low-dimensional linear projections of high-dimensional data optimizing some specified function called the projection pursuit index.

Advantages:
- Able to bypass the curse of dimensionality, because work in low-dimensional linear projections.
- Items relevant projection pursuit indexes are able to ignore irrelevant variables.

We use LDA and PDA index in our PPforest.

PPforest algorithm

A random forest is an ensemble learning method, built on bagged trees. PPforest conducts a supervised classification using projection pursuit trees and random forest ideas.

Using this combination we are taking into account the association between variables to find separation that is not considered in a classic random forest.

PPforest R package

- PPforest: Run a Projection pursuit random forest object.
- predict.PPforest: Vector with predicted values from a PPforest object.
- pcf_importance: Plot a global measure of variable importance.
- pcf_plot: Proximity matrix visualization.
- pcf_oob_error: OOB error summary and visualization/verb.

PPforest package visualization

Heat map proximity matrix.

MDS for proximity matrix.

Importance variable

Cumulative oob error

Prediction accuracy comparison

Comparison of PPtree, CART, random forest and PPforest. We used various data sets included in PPforest package. In each case the mean of training and test error rates from 200 re-samples were computed. Based on prediction accuracy comparison, PPforest presents a better performance for NCICD and Crab data while Lymphoma, Lymphoma and Wine present the same performance than random forest.

Final comments

- PPforest uses the association between variables to find separation.
- The strength of each individual tree in the forest increases when classes are linearly separable, smaller error rate in the forest.
- The predictive performance is better than other classifiers for some of the analyzed data.
- PPforest is available on github: https://github.com/natydasilva/PPforest

References


Contact information

- Webpage: ndasilva.public.iastate.edu
- Email: ndasilva@iastate.edu
- Twitter: @pacoouak
- PPforest: github.com/natydasilva/PPforest