Introduction

PPforest is an ensemble learning method. It adapts the classic random forest to utilize combinations of variables, as produced by projection pursuit, in the tree construction. Utilizing linear combinations of variables to separate classes, we enhance the correlation between variables into account, and can outperform the classic forest when separations between groups exist in combinations of variables. This method is implemented in our package PPforest, i.e., https://github.com/natydasilva/PPforest.

Prototype PPforest Visualization for CRAB data

Motivation and tools

PPforest has a better prediction accuracy over other classification methods. However, the model interpretation is difficult. Model interpretation and diagnostic requires the analysis of different components of the same tree, the interactive and linked plots are key to an effective visualization. We develop an interactive tool to visualize PPforest model to find insights about data structure and PPforest attributes relation. We will see how well the model fits the data, we will identify cases where the fit is appropriate and where it is poor.

PPforest feature can be computed at independent observation level (proximity matrix and vote matrix).

Visualizaton is critical to help obtain an understanding of the class structure in the data and how the model fits it. Interactively related plots are key to get an effective model diagnostic visualization. The prototype visualization is done using CRAB data set. With these "out-of-bag" data in each tree we get the classification of them. At the end we consider to be the class with more votes every time a case is out-of-bag. The proportion of times that case is in wrong classified averaged over all cases is the OOB error estimate.

Each dot or line represents one individual observation (crab id), this level is useful to identify which observations are more difficult to classify by the PPforest model.

References


Contact information

Webpage: natsilva.public.iastate.edu
Email: natsilva@iastate.edu
Twitter: @natsilva
PPforest: github.com/natydasilva/PPforest

Discussion

- Visualization is critical to help obtain an understanding of the class structure in the data and how the model fits it.
- Interactive and linked plots are key to get an effective model diagnostic visualization.
- Different levels of analysis give a complete evaluation of PPforest object.
- Individual level allows to identify how the model fits the data and which are the wrong classify observations.
- Tree level allows to analyze each of the main components of the forest and understand the particular partitions and variables used for them.
- Forest level allows performance comparison with random forest.
- Effective, interactive and elegant visualization can be done in a simple way combining ggplot2, plotly and shiny.

CRAB data

The prototype visualization is done using CRAB data set. Measurements on rock crabs, 200 observations. 4 classes species-sex:

- FL: the size of the frontal lobe length, in mm
- RW: raw width, in mm
- CL: length mid-line of the carapace, in mm
- MW: maximum width of carapace, in mm
- BD: depth of the body. For females, measured after displacement of the abdomen, in mm