25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Generalising the raindrop plot for prediction and variable selection: An educational case study

D. Vasco [b], S. Low-Choy [a,b,c] and Parlo Singh [b]

[a] Arts, Education and Law Group, Griffith University, Brisbane, Australia
[b] Griffith Institute of Educational Research, Griffith University, Brisbane, Australia
[c] Centre for Planetary Health and Food Security, Brisbane, Australia
Email: d.vasco@griffith.edu.au

Abstract: Multi-model approaches are becoming more popular in statistics. Ensemble models are already popular in machine learning; however, they often lose the interpretability of a single model. Similarly, ensembles of classification trees optimise predictability at the expense of explainability, e.g. Random Forests, Gradient Boosting Machines, and Boosted Regression Trees. Yet individual trees are intuitive to interpret. We developed visual approaches to enhance the interpretability of ensemble approaches that involve many trees.

Visualising a classification tree (CT) as a decision tree shows important information about the model structure, such as variable selection and predictions. Branching shows a hierarchy of variable importance, with height reflecting the relative information provided by each branch. For each tree, this visualisation is easy to understand. However, for complex tree structures, it can be difficult to view and compare many trees simultaneously. To our knowledge, no existing visualisation techniques simultaneously show, for many trees, both the rich model structure (including variable selection) and predictive ability. Our work was motivated by a desire for easy communication on model performance, when relating socio-economic factors to non/participation in Australia's NAPLAN (National Assessment Program—Literacy and Numeracy).

We previously proposed a plot in a plant biosecurity setting (Vasco & Low-Choy, MODSIM 2017), to compactly show 10–20 trees, displaying model structure, predictions, and predictive performance indicators (PPIs). The improved raindrop plot (Figure 1) still shows predictions via colour: predicted response (hue), and strength (shade). Our diverse stakeholders prompted improvements. The terminology for PPIs is now easier to interpret and recall. PPIs are aligned so that optimal values are always high. Predictions are sharper (better PPIs) for moderate penalties on misclassifications (between 60 and 200), stabilising at penalty 70 (Figure 1).

Our work proposes a generalised class of Raindrop Plots, each compactly showing trees, for a specific model diagnostic. The enhanced original plot focuses on predictions (Figure 1), while another option focuses on variable inclusion. These plots were developed in R, to support compact visualisation of trees to support several quantitative approaches: machine learning (here), resampling (e.g., cross-validation or bootstrapping), Bayesian statistics (posterior distributions), or ensemble modelling. Overall, these plots support statistical thinking and an iterative approach to modelling and, hence may enhance machine learning and data science practices.

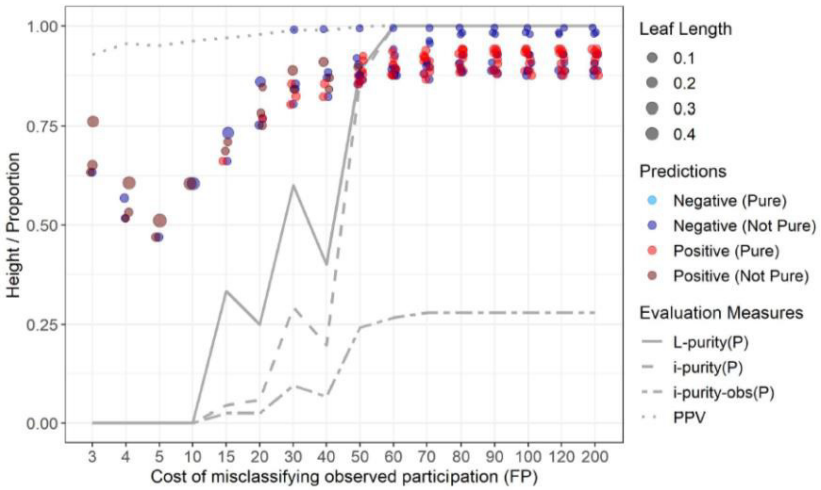

Figure 1. Raindrop plot to support choice of model options, penalty for misclassification, here (x-axis) for false negative (FN) classification, based on the stability of the leaves structure (points: y-coordinate, size and colour) and predictive indicators (grey lines)

Keywords: *Classification trees, model uncertainty, sensitivity analysis*