Using R to evaluate the variable importance

April 7, 2015

This summary reviews the keypoints in the paper ”A new, conditional variable-importane measure for random forests available in the package party” by Strobl, C., Hothorn, T., and Zeileis, A., 2009.

1. Function `cforest` in the R package `party`. The resulting forests are unbiased and thus preferable to the `randomForest` implementation available in the R pacakge `randomForest` if the predictors are of different types. A conditional permutation importance measure has been added to the `party` package, which helps to evalute the importance of correlated predictors.

2. Recursive partitioning methods are amongst the most popular and widely used statistical learning tools for non-parametric regression and classification.

3. The scope of recursive partitioning methods in R:

   - Standard classification and regression trees in the R package `rpart`.
   - Random forests in the R package `randomForest`.
   - Unbiased tree algorithm in the R package `party`. Useful functions include `ctree` and `cforest`.

4. The major weak spot of classical approaches in `rpart` and `randomForest` is the variable-selection bias, which refers to the fact that in standard tree algorithms, variable selection is biased in favor of variables offering many potential cut-poiints, so that variables with many categories and continuous variables are artificially preferred.

5. The variable importance is usually evaluated using the measures such as Gini importance, the mean decrease in accuracy or permutation importance. The Gini importance is accessible in the `randomForest` with `importance(obj, type=2)`. Noted that for variables of different types, such measure is biased in favor of continuous variables and variables with many categories.

6. Permutation importance or the mean decrease in accuracy is accessible in both `randomForest` and `party`, e.g., `randomForest::importance(obj, type=1)` or `party::varimp(obj)`. Note that for variables of different types, unbiased measure is only achieved when subsampling is used as in `cforest(controls=cforest_unbiased())`.

7. Permutation importance is a reliable measure for *uncorrelated* predictors when sub-sampling without replacement (instead of bootrap sampling). This is the default setting for `party::cforest_ control` as `party::cforest_ unbiased`.

8. The rationale of the original permutation importance is as follows: by randomly permuting the predicotr varibale X_j, its original association with the response Y is broken. When the permuted X_j and the remaining non-permuted predictors are used to predict the response for the out-of-bag observations, the prediction accuracy will decrease substantially if the original variable X_j was associated with the response Y. Thus, Breiman (2001) suggests the difference in prediction accuracy before and after permuting X_j, averaged over all trees, as a measure for variable importance.

9. Suggestions are as follows:

   - If predicotrs are of different types, then use `party::cforest` with default option `controls=cforest_unbiased()`. Then use `party::varimp()`.

- Otherwise, feel free to use `party::cforest` with `party::varimp`; or `randomForest::randomForest` with `importance(obj, type=1)` or `importance(obj, type=2)`. Note always set `scale=FALSE`.
- If predictors are highly correlated, use `party::cforest` with `party:varimp(conditional=TRUE)`.
- Always check whether the result is the same with a different random seed. If the ranking is dependent on the random seed, try to increase the number of trees `ntree` in `randomForest` and `cforest_ control`.