The Boruta Wrapper for Assessing Variable Importance

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

"Boruta" "Shadow" variables

The "Boruta" methodology of Kursa and Rudnicki for identification of all coordinates x_j of an input $\mathbf{x}$ that have "statistically detectable" variable importance builds on the importance measure for a bagging predictor I^j, usually derived from random forests. (Boruta is named for the mythological Slavic god of the forest.)

The method is aimed at judging which I^js are "clearly more than noise." To enable this, when r predictors are currently under consideration some (say $s \geq \max(5, r)$) additional "shadow" (plausible noise) predictors are considered along with the actual predictors. These shadow predictors are made by randomly permuting elements in columns of the original input matrix for the r predictors under consideration. These "should" prove to be of no importance in the prediction of y.

Standardized importance scores

Boruta operates in stages in a "backwards elimination" fashion, beginning with consideration of all p original predictors and at a given stage dropping from the set of remaining potentially important variables those that are "clearly no better" than the best shadow variable at the stage.

To make decisions about elimination, the set of importance values I_b^j for a given j (newly indexing both those actual predictors still under consideration as $1, \ldots, r$ and those shadow predictors newly generated at the beginning of the stage as $r+1, \ldots, r+s$), and compute both their mean, I^j, and their sample standard deviation, call it S^j. Then for all $j = 1, 2 \ldots, r+s$ let

$$Z^j \equiv \frac{I^j}{S^j}$$

“Tests” of significance

Then some kind of rough test of "statistical significance" is based on comparison of the scores (possibly accumulated across stages) Z^j for real inputs (i.e. for $j = 1, \ldots, r$) against the best score for a shadow variable

$$\max_{j=r+1,\ldots,r+s} Z^j$$

The elimination process is intended to ultimately drop from consideration all those predictors whose scores are not clearly bigger than those of (by construction useless) shadow predictors.

Comments

This is, of course, a heuristic and exact details vary with implementation. But the central idea is above and makes sense. It can be applied to any bagging context, and variants of it could be applied where one is not bagging, but other forms of holding out a test set are employed. Typically, the prediction method used is the random forest, because of its reputation for broad effectiveness and its independence of scaling of coordinates of the input. But there is nothing preventing its use with, say, a linear prediction or smoothing methodology.