Boosting Flexible Learning Ensembles with Dynamic Feature Selection

Alexander Borisov, Victor Eruhimov, Eugene Tuv [*]

Abstract

Models in industrial applications are often challenged by complexity of underlying data. This includes mixed type variables with blocks of non-randomly missing data, categorical predictors with very large number of levels (hundreds or thousands). Very often datasets are extremely saturated - small number of observations and huge number of variables (tens of thousands). Both predictors and responses normally are less than clean. Both regression and multilevel classification models are of interest. Thus, an universal learner is needed ...

Recent advances in tree based methods such as MART (Freidman's Gradient Tree Boosting) and RF (Breiman's Random Forests) are proven to be effective in addressing most of the issues listed above. Both are resistent to outliers in X-space, both have efficient mechanism to handle missing data, both are competitive in accuracy with the best known learning algorithms in regression and classification settings, mixed type data is handled naturally. However, MART uses exhaustive search on all input variables for every split and every tree in ensemble, and it becomes computationally extremely expensive to handle very large number of predictors. At the same time RF shows significant degradation in accuracy in the presence of many noise variables.

We propose a simple improvement to both ensembles: only a small subset of features is considered at every construction step of an individual learner in ensemble. Sampling distribution of features is dynamically modified to reflect currently learned feature importance. This distribution is initialized as uniform, and progresses with adjustable rate to prevent initial overweighting of a few variables. Feature importance are dynamically recalculated over the current ensemble (we used reduction in impurity due to splits on the feature as measure of it's importance).

This method makes tree gradient boosting feasible (actually very fast) for the data with large number of predictors without loss of accuracy. It also adds bias correction element to RF in the presence of many noise variables. Actually our experiments showed slight improvement of predictive accuracy for MART on average and very significant for RF in the presence of noise.

Note that feature selection challenge results were obtained using stochastic gradient boosting with dynamic feature selection implemented in IDEAL (internal tool) practically out of box with a few runs.

[*] Intel Corporation, alexander.borisov@intel.com,victor.eruhimov@intel.com,eugene.tuv@intel.c om