PERFORMANCE PREDICTION CHALLENGE: FACT SHEET

Title: Advanced Analytical Methods, INTEL

Name, address, email: Borisov Alexander (alexander.borisov@intel.com) & Eugene Tuv (eugene.tuv@intel.com)

Acronym of your best entry: IDEAL

Reference: Borisov A., Eruhimov V. and Tuv, E. *Tree-Based Ensembles with Dynamic Soft Feature Selection*, In Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti Zadeh, editors, Feature Extraction, Foundations and Applications. Springer, 2006. (in press)

Method: Gradient Tree Boosting with Dynamic Feature Selection.

No preprocessing was done. We used gradient tree boosting with dynamic feature selection. The method builds a serial binomial logistic regression tree ensemble. Each new expert – a shallow tree is built on the residuals from the previous iteration using a random subset of cases. For very unbalanced datasets a stratified sampling was used to up weight the rare class. At each node, a random small subset of variables [sqrt(total number of variables)] is selected. The vars sampling probabilities are proportional to sum of priors (initially equal, then decreasing influence as trees are added to the ensemble) and current variable importances computed using split scores evaluated over the ensemble.

All hyper-parameters (tree depth, sampling scheme, regularization, importance adjustment rate for the dynamic FS, class priors) were selected using test sample estimation (over multiple train/test partitions). Performance prediction guess error was done using the same test sample estimation.

Learning from the unlabeled test set was not used.

Results: In the challenge, we ranked 7th as a group and our best entry is the 17.6 th , according to the average rank computed by the organizers. Our method is accurate and fast on wide variety of datasets with complex dependencies (for example, we ranked in top ten on NIPS2003 also with the same method). Our method is generally more accurate and incomparably faster (on massive in both dimensions datasets) than original Freidman's MART and Breiman's RF. It does not require any data preprocessing.

Dataset	Our best entry					The challenge best entry				
	Test AUC	Test BER	BER guess	Guess error	Test score	Test AUC	Test BER	BER guess	Guess error	Test score
ADA	0.9110	0.1779	0.162	0.0159	0.1939(12)	0.8304	0.1696	0.155	0.0146	0.1843(3)
GINA	0.9893	0.035	0.044	0.009	0.044(7)	0.9639	0.0361	0.0388	0.0027	0.0386(5)
HIVA	0.7193	0.3085	0.285	0.0235	0.3312(3)	0.7129	0.2871	0.27	0.0171	0.3029(8)
NOVA	0.9837	0.0622	0.062	0.0002	0.0622(21)	0.9542	0.0458	0.0503	0.0045	0.0499(8)
SYLVA	0.9986	0.0132	0.0085	0.0047	0.0179(45)	0.9937	0.0063	0.0058	0.0005	0.0067(7)
Overall	0.92038	0.11936	0.1123	0.01066	0.12984(17.6)	0.891	0.109	0.104	0.0079	0.1165(6.2)

Code: Method is implemented in C++ as a part of Intel's IDEAL ML software product . It is not publicly available.

Keywords: Ensembles, boosting, feature selection, tree