PERFORMANCE PREDICTION CHALLENGE FACT SHEET

Title: LogitBoost with trees

Name, address, email: Roman Werner Lutz, Seminar for Statistics, ETH Zurich, CH-8092 Zurich, Switzerland, lutz@stat.math.ethz.ch

Acronym of your best entry: LB tree mix cut adapted

Reference: LogitBoost with Trees Applied to the WCCI 2006 Performance Prediction Challenge, Roman Werner Lutz, In Proceedings IJCNN06, to appear.

Method:

As preprocessing we used PCA for Nova with centered and scaled variables and took the first 400 principal components. No preprocessing was used for the other datasets. Then we applied LogitBoost with trees of prefixed depth. The number of iterations, the tree depth (in each iteration a tree of the same depth is fitted) and the BER guess were chosen/computed by 10-fold cross-validation. Shrinkage was added to make LogitBoost more stable: in each iteration only a fraction ? (0.3, 0.1 or 0.03) of the fitted learner was added. ? was chosen by visual inspection of the cross-validated BER curve (as a function of the boosting iteration). As a result, LogitBoost yielded probabilities of class membership for each sample. The cut point for the final classification was the proportion of class +1 in the data.

For the second entry we used the Wilcoxon test (for continuous variables) and the Fisher exact test (for binary variables) for variable pre-selection (variables with a p-value above 0.1 were dropped). For the third entry we averaged the predicted probabilities of LogitBoot with and without variable pre-selection. For the fourth entry we made an intercept adaption (on the logit scale) so that the average of the predicted probabilities on the test set equals the proportion of class +1 in the data.

Results:

In the challenge, we rank 1st as a group and our best entry (our fourth) is the 1st, according to the average rank computed by the organizers. Our method is quite simple: no preprocessing is needed (except for Nova) and the tuning parameters are chosen by cross-validation. Additionally, LogitBoost with trees does variable selection, because in each iteration only a few variables are chosen.

Dataset	Our best entry					The challenge best entry				
	Test AUC	Test BER	BER guess	Guess error	Test score (rank)	Test AUC	Test BER	BER guess	Guess error	Test score (rank)
ADA	0.8304	0.1696	0.1550	0.0146	0.1843 (3)	0.9149	0.1723	0.1650	0.0073	0.1793 (1)
GINA	0.9639	0.0361	0.0388	0.0027	0.0386 (5)	0.9712	0.0288	0.0305	0.0017	0.0302 (1)
HIVA	0.7129	0.2871	0.2700	0.0171	0.3029 (8)	0.7671	0.2757	0.2692	0.0065	0.2797 (1)
NOVA	0.9542	0.0458	0.0503	0.0045	0.0499 (8)	0.9914	0.0445	0.0436	0.0009	0.0448 (1)
SYLVA	0.9937	0.0063	0.0058	0.0005	0.0067 (7)	0.9991	0.0061	0.0060	0.0001	0.0062 (1)
Overall	0.8910	0.1090	0.1040	0.0079	0.1165(6.2)	0.8910	0.1090	0.1040	0.0079	0.1165(6.2)

Code: Our implementation was done in R.

Keywords: PCA, Wilcoxon test, Fisher exact test, LogitBoost, trees of fixed depth, 10-fold cross-validation, shrinkage.