PERFORMANCE PREDICTION CHALLENGE: FACT SHEET FORMAT

Title: Random Forests
Name, address, email: Corinne Dahinden, Seminar for Statistics, ETH Zurich, CH-8092 Zurich, Switzerland, dahinden@stat.math.ethz.ch
Acronym of your best entry: RF

Reference: Classification with Tree-Based Ensembles Applied to the WCCI 2006 Performance Challenge Datasets, Corinne Dahinden, In Proceedings IJCNN06, to appear.

Method:
For the NOVA dataset, PCA with centered and scaled variables were computed and the first 400 principal components were taken. For the other datasets, no preprocessing was applied. No variable-selection was performed for any dataset. Then Random Forests was used with 4000 trees fitted with CART. Instead of the theoretical cutoff, which is the proportion of labels with +1 in the dataset, this cutoff is estimated by Cross-Validation, which considerably improves the performance of Random Forests for unbalanced datasets.

The performance prediction was guessed by 10-fold Cross-Validation.

Results:
In the challenge, we rank 4th as a group and our best entry is the 11th, according to the average rank computed by the organizers. The Random Forests algorithm can be applied to a wide range of datasets and is not subject to the "small n – large p" problem. Plain standard Random Forests is very simple, yet highly efficient. The cutoff-adaptation has even shown to improve this performance, still the computational cost is kept low. The procedure requires minimal human interaction, can be used for variable selection and internally computes unbiased estimate of the generalization error. Random Forests achieves results which can keep up with the most sophisticated algorithms.

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	0.8200	0.1800	0.1650	0.0150	0.1950 (16)	0.9149	0.1723	0.1650	0.0073	0.1793 (1)
GINA	0.9587	0.0413	0.0490	0.0077	0.0490 (17)	0.9712	0.0288	0.0305	0.0017	0.0302(1)
HIVA	0.7009	0.2994	0.2700	0.0294	0.3284 (32)	0.7671	0.2757	0.2692	0.0065	0.2797 (1)
NOVA	0.9470	0.0530	0.0530	0.0000	0.0530 (15)	0.9914	0.0445	0.0436	0.0009	0.0448 (1)
SYLVA	0.9946	0.0054	0.0065	0.0011	0.0065 (3)	0.9991	0.0061	0.0060	0.0001	0.0062 (1)
Overall	0.8842	0.1158	0.1087	0.0106	16.6	0.8910	0.1090	0.1040	0.0079	0.1165 (6.2)

Code: The computations were done in R. Random Forests is implemented in the randomForest package available from CRAN (http://cran.r-project.org). In addition, we have also implemented the cutoff-adaptation in R.

Keywords: PCA, embedded feature selection, RF, 10-fold cross-validation, ensemble method, CART