AlvsPK Challenge: FACT SHEET

Title:
Feature selection with redundancy elimination + gradient boosted trees.

Name, address, email:
ASML team, INTEL Corporation, alexander.borisov@intel.com

Acronym of your best entry:
out1-fs-nored-val (Intel final 1)
out3-fs-red-valid (Intel final 2)
out5-valid-no-fs (Intel final 3)

Reference:
Paper to be resubmitted to the JMLR

Method:
No preprocessing was done.
The method consists of the following steps
1. Feature selection using ensemble classifiers (ACE FS). Random probes that are permutation of original features are added. Importance of each variable in RF ensemble is compared versus importance of probes using t-test over several ensembles. Variables that are more important in statistical sense then most of probes are selected as important. Variables are ordered according to sum of gini index reduction in tree splits.
2. Variable masking it estimated on important variables with GBT ensemble using surrogate splits (if a more important variable has surrogate on less important one, the second variable is masked by the first). Again, statistically significant masking pairs are selected, then subset of mutually non-masked variables with high importance is chosen
3. Effect of found variables is removed using RF ensemble.
Steps 1-3 are repeated until no more important variables remain.
Next GBT with embedded feature selection (to prevent over fitting) is built on selected variable set. The following parameters of GBT were optimized : number of trees, tree depth, shrinkage, number of selected features per tree node and importance adjust rate (for embedded FS), stratified sampling 0/1 class proportions, priors. For FS, #of trees in series, importance and masking quantile were chosen.
Optimization strategy (manual) was to set reasonable parameter values, then try to adjust each parameter (sequentially), so that test error decreases (model was trained on 60%.of training data during parameter optimization). Several passes over all GBT parameters was done, one for FS parameters.
Priors were selected using cross validation (FS+GBT run was done on K partitions of the data, optimal priors were selected on remaining part).

Results:

Table 1: Our methods best results

Dataset	Entry name	Entry ID	Test BER	Test AUC	Score	Track
ADA	out1-fs-nored-val (Intel final 1)	1051	0.1737	0.8259	0.0143	Agnos
GINA	out1-fs-nored-val (Intel final 1)	1051	0.0373	0.9631	0.2436	Agnos
HIVA	out3-fs-red-valid (Intel final 2)	1052	0.2899	0.7123	0.1124	Agnos
NOVA	out1-fs-nored-val (Intel final 1)	1051	0.0547	0.9468	0.2756	Agnos
SYLVA	out1-fs-nored-val (Intel final 1)	1051	0.0135	0.9865	0.5126	Agnos
Overall	out1-fs-nored-val (Intel final 1)	1051	0.1142	0.8859	0.2373	Agnos

Table 2: Winning entries of the AlvsPK challenge

Best results agnostic learning track						
Dataset	**Entrant name**	**Entry name**	**Entry ID**	**Test BER**	**Test AUC**	**Score**
ADA	Roman Lutz	LogitBoost with trees	13, 18	0.166	0.9168	0.002
GINA	Roman Lutz	LogitBoost/Doubleboost	892, 893	0.0339	0.9668	0.2308
HIVA	Vojtech Franc	RBF SVM	734, 933, 934	0.2827	0.7707	0.0763
NOVA	Mehreen Saeed	Submit E final	1038	0.0456	0.9552	0.0385
SYLVA	Roman Lutz	LogitBoost with trees	892	0.0062	0.9938	0.0302
Overall	Roman Lutz	LogitBoost with trees	892	0.1117	0.8892	0.1431
Best results prior knowledge track						
Dataset	**Entrant name**	**Entry name**	**Entry ID**	**Test BER**	**Test AUC**	**Score**
ADA	Marc Boulle	Data Grid	920, 921, 1047	0.1756	0.8464	0.0245
GINA	Vladimir Nikulin	vn2	1023	0.0226	0.9777	0.0385
HIVA	Chloe Azencott	SVM	992	0.2693	0.7643	0.008
NOVA	Jorge Sueiras	Boost mix	915	0.0659	0.9712	0.3974
SYLVA	Roman Lutz	Doubleboost	893	0.0043	0.9957	0.005
Overall	Vladimir Nikulin	vn3	1024	0.1095	0.8949	0.095967

- quantitative advantages

Method is very fast (~a minute for one FS iteration on NOVA dataset with 16K+ vars) (20 ensembles with 70 trees))
(faster than all known to us minimal subset selection methods). Complexity is proportional to
to (Fsel+Fimpvar)*N*logN*Ntrees*Nensembles*Niter + Niter*Fimpvar^2,
Niter - #of iteration of ACE FS algorithm always < 10, usually 3-4
Nensembles = 20 (number of ensembles for t-test)
Ntrees = 20-100 (number of trees in RF or ensemble)
N - number of samples,
Fsel = number of selected important vars per tree split (sqrt(total number features) or less)
Fimvar – total number of selected important variable, for NOVA – 400-800 depending on parameters).
Works with any variable types, mixed values, requires no preprocessing.

- qualitative advantages

This method allows to find a small subset of features with the same predictive capacity as the original set.

Original # of features, CV-err using all features / best subset size, CV-err using best subset

Ada: 47 , 0.190902 / 16 ,0.185584

Gina : 970, 0.052740 / 75 ,0.050629

Hiva:1617,0.284723 / 221, 0.255898

Nova: 12993,0.059070 / 400, 0.051794

Sylva: 212 , 0.013268 /69, 0.012852

Keywords:

- Preprocessing or feature construction: ----
- Feature selection approach: embedded feature selection.
- Feature selection engine:miscellaneous classifiers (RF, GBT).
- Feature selection search: variable masking estimation, redundancy elimination, statistical test.
- Feature selection criterion: 5-fold cross-validation.
- Classifier: RF, Gradient boosting trees.
- Hyper-parameter selection: manual optimization.
- Other: ------.