

Feature partitioning and boosting

Miklós Kurucz, Dávid Siklósi

Data Mining and Web Search Group
Computer and Automation Research Institute
Hungarian Academy of Sciences
joint work with several colleagues from Budapest

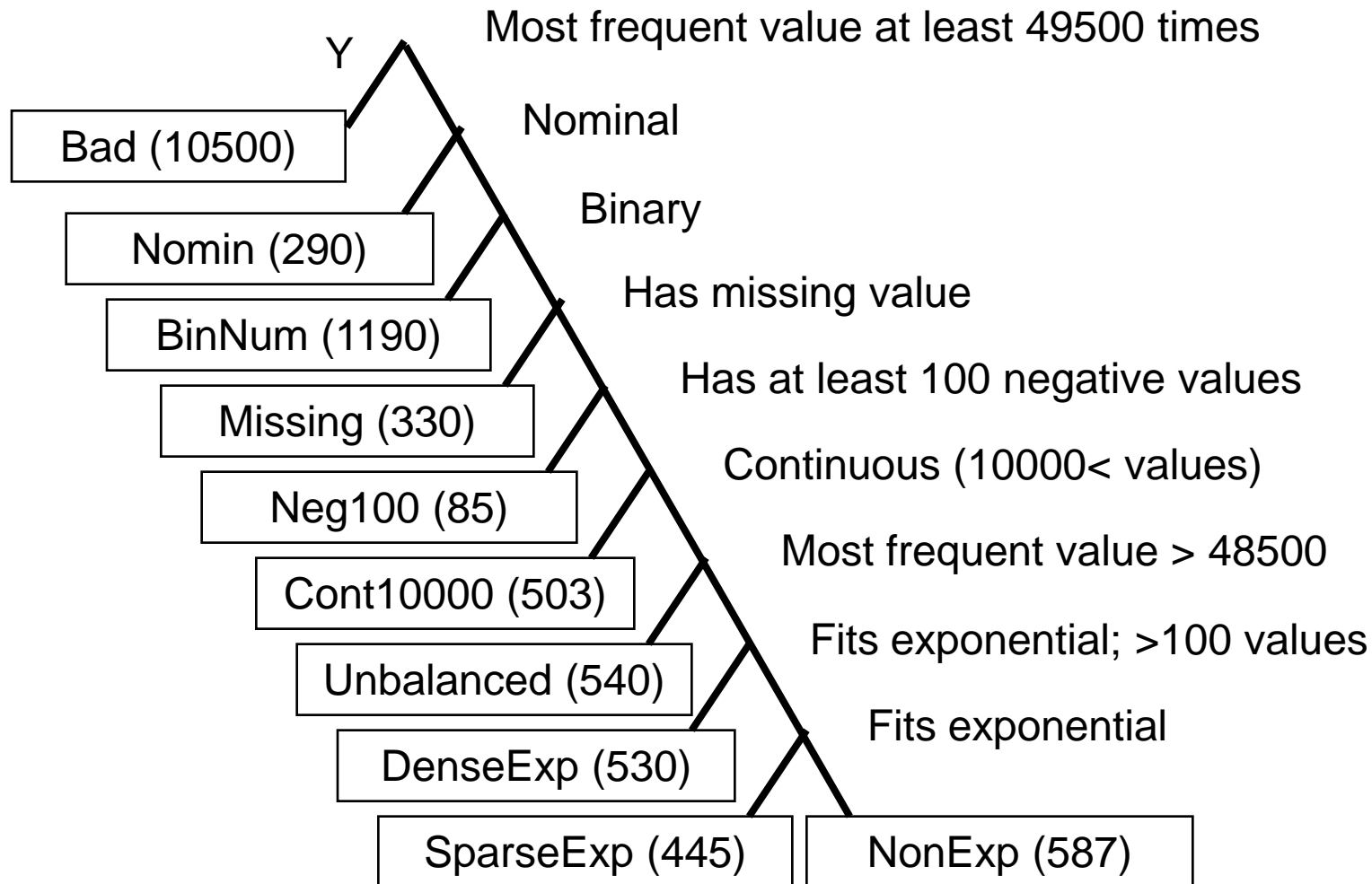


Methodology



- Large data set only
- 15,000 features -> partitioning, selection
- Feature evaluation as a weak pre-selection only
- Expected classifier combination to perform well over *partitioned feature set*
 - Might hold with knowledge of feature meaning
 - Did help in scaling, parallelization, exploration
- 10% heldout and 10% validation data set aside
- Access to large computational power, little additional time used after fast track
- Using Weka + scripts, tested many, many classifiers - *LogitBoost w/ decision stump* wins almost everywhere

Feature Partitioning



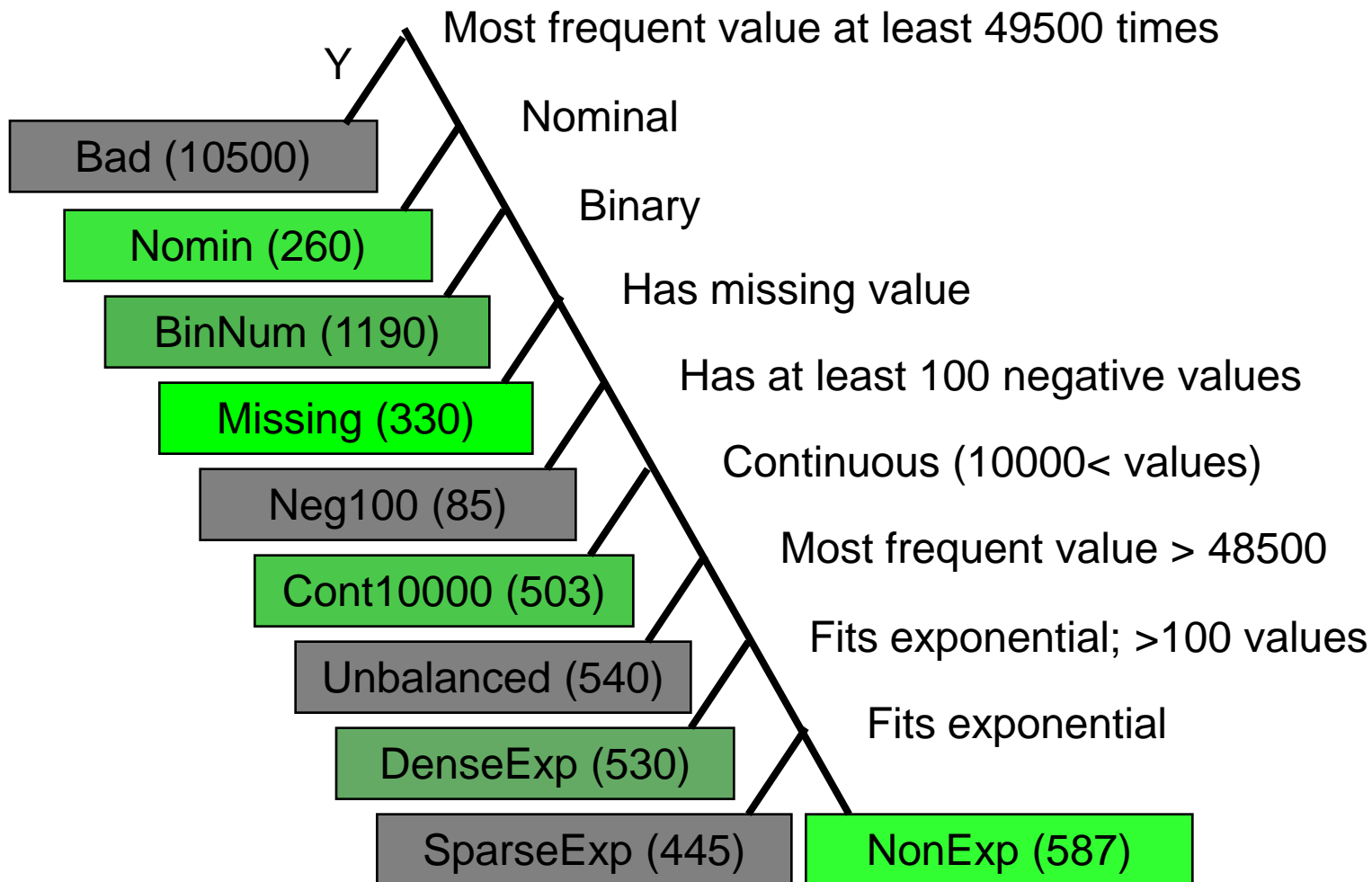
Performance of feature subsets



	churn		appetency		upselling	
	heldout	valid	heldout	valid	heldout	valid
Missing w/ LogitBoost	0.7232	0.7318	0.8394	0.8217	0.8855	0.8931
NonExp w/ AdaBoost	0.7188	0.7359	0.8551	0.8332	0.8835	0.8815
Nominal w/ LogitBoost	0.6657	0.6696	0.8385	0.7868	0.7623	0.7649
Cont10000 w/ Logitboost	0.6465	0.6631	0.6564	0.6712	0.7419	0.7474
BinNum w/ Logitboost	0.6369	0.6187	0.7204	0.7233	0.8016	0.8126
DenseExp w/ LogitBoost	0.6294	0.6473	0.6398	0.6591	0.7251	0.7391
NonExp w/ Bayes	0.6230	0.6531	0.5870	0.6393	0.7330	0.7224
Combination w/ LogitBoost		0.7667		0.8537		0.9100
Combination of log-odds w/ LogitBoost		0.7583		0.8361		0.9026

Table 1: The AUC value of feature subsets and the classifier combination over our 10+10% heldout and validation sets.

Feature Partitioning





1. Feature evaluation: weak pre-selection
 - Many non-predictive, highly correlated features
 - Threshold hard to set
 - IG, Chi² overscore many unique values
 - Gain Ratio overscore few unique values
2. LogitBoost itself uses a few selected features
 - Superlinear time, even 1000 features too much
 - Used over our feature partitioning
 - Used over random partition

Partitioned vs global over our heldout ...



	churn		appetency		upselling	
	heldout	valid	heldout	valid	heldout	valid
Combination LogitBoost		0.7667		0.8537		0.9100
Logitboost by partition	0.7557	0.7649	0.8668	0.8509	0.9122	0.9099
Logitboost random	0.7540	0.7612			0.9064	0.9069
Combination log-odds LogitBoost		0.7583		0.8361		0.9026
feature evaluation LogitBoost	0.7335	0.7414	0.8033	0.7924	0.8935	0.8868

Table 2: AUC values over our 10+10% heldout and validation sets.

... and the Cup test set



	churn	appetency	upselling	score
Winner (University of Melbourne)	0.7570	0.8836	0.9048	0.8484
LogitBoost + ADTree by partition (final)	0.7567	0.8736	0.9065	0.8456
LogitBoost by partition	0.7496	0.8683	0.9042	0.8407
Combination LogitBoost	0.7409	0.8561	0.8894	0.8288

Table 3: The AUC value of selected final methods over the test set.



Final best solution



- LogitBoost and ADTree
- Plain average turns out better than combination by classifiers
- Final results use all training set (combination by cross-validation)
- Final results (less than 20) evaluated over the 10% feedback - no overtraining, no difference in relative order
- Understand the variance (difference between 10% and full test set +0.02% for us but lot more for other teams)?



Further directions



- Partitioning by meaning (traffic, socio-demographic etc) might work better
- Would the same methods scale for larger data (M's of users instead of 50K)?
- Staying power (prediction for future)?
- Evaluate graph stacking? Needs call graph

Questions ?

Miklós Kurucz

mkurucz@ilab.sztaki.hu

<http://datamining.sztaki.hu>