AlvsPK Challenge: FACT SHEET

Title: Report on Preliminary Experiments with Data Grid Models in the Agnostic Learning vs. Prior Knowledge Challenge

Name: Marc Boullé

Address: France Telecom R&D
2, avenue Pierre Marzin
22307 Lannion Cedex, France

Email: marc.boulle@orange-ftgroup.com

Acronym of your best entry: Data Grid (Coclustering)

Reference:
Paper published in IJCNN 2007

Method:
Data grids extend the MODL discretization and value grouping methods to the multivariate case.
They are based on a partitioning of each input variable, in intervals in the numerical case and in groups of values in the categorical case. The cross-product of the univariate partitions forms a multivariate partition of the input representation space into a set of cells. This multivariate partition, called data grid, allows to evaluate the correlation between the input variables and the output variable. The best data grid is searched owing to a Bayesian model selection approach and to combinatorial algorithms.
Three classification techniques exploiting data grids differently are presented and evaluated in the Agnostic Learning vs. Prior Knowledge Challenge:

- Data Grid (MAP): use the MAP data grid as a classifier
- Data Grid (CMA): use an ensemble of data grids
- Data Grid (Coclustering): apply a bivariate unsupervised data grid to learn a coclustering on the instance*variable space, using all the unlabelled train+valid+test data. The clusters of instances are used for prediction using the available labels (train+valid).

Summary of the method:

- Preprocessing : multivariate partition (discretization/value grouping)
- Feature selection: variables whose univariate partition contains at least two parts are selected
- Classification
  - Data Grid (MAP): the best multivariate partition forms a classifier
  - Data Grid (CMA): use an ensemble method
  - Data Grid (Coclustering): use learning from the unlabeled test set
- Model selection/hyperparameter selection: model are selected using a Bayesian approach (no hyper-parameter)

Results:

Table 1: Our methods best results

Dataset	Entry name	Entry ID	Test BER	Test AUC	Score	Track
ADA	Data Grid (CMA)	920	0.1756	0.8464	0.0245	Prior
GINA	Data Grid (Coclustering)	921	0.0516	0.9768	0.3718	Prior
HIVA	Data Grid (Coclustering)	921	0.3127	0.7077	0.5904	Agnos
NOVA	Data Grid (Coclustering)	921	0.0488	0.9813	0.141	Agnos
SYLVA	Data Grid (CMA)	918	0.0158	0.9873	0.6482	Agnos
Overall	Data Grid (Coclustering)	921	0.1223	0.8984	0.3813	Prior

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 compact feature subset, works with any variable type, ease of interpretation, no parameter tuning, use all the available data, computational efficiency
- qualitative advantages compute posterior probabilities, model selection based on a Bayesian approach, data grids are a new machine learning technique.