AlvsPK Challenge: FACT SHEET

Title: "Classification with Random Sets, Boosting and Distance-based Clustering"
Name: Vladimir Nikulin
Address: Airservices Australia, 25 Constitution Ave., Canberra, ACT 2601
Emails: vnikulin@digisurf.com.au, vladimir.nikulin@airservicesaustralia.com
Acronym of the best entry: vn3
Reference: The paper of 24 pages will be resubmitted to JMLR

Table 1: Our best results (used only last 5 complete entries)

Dataset	Entry ID	Entry Name	Track	Test BER	Test AUC
ADA	1026	vn5	Agnos	0.1751	0.8331
ADA	1024	vn3	Prior	0.1788	0.8225
GINA	1023	vn2	Prior	0.0226	0.9777
GINA	1025	vn4	Agnos	0.0503	0.9507
HIVA	1024	vn3	Agnos	0.2904	0.7343
NOVA	1026	vn5	Agnos	0.0471	0.9456
SYLVA	1024	vn3	Prior	0.0071	0.9959
SYLVA	1025	vn4	Agnos	0.0096	0.9933
Overall	1024	vn3	Prior	0.1095	0.8949
Overall	1026	vn5	Agnos	0.1177	0.8891

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

Method:
Overfitting represents usual problem associated with classification of high-dimensional data. According to the proposed approach we can use large number of classifiers where any single classifier is based on the subset of relatively small number of randomly selected features or random sets (RS) of features.
Consequently, any single RS-classifier 1) will not overfit, and 2) may be evaluated very quickly. The property of limited overfitting is a very important. As a result, feature selection in the final model will be made according to several best performing subsets of features.

The proposed method is an essentially different comparing with Breiman's Random Forests (voting system)[1] where the final classifier represents a sample average of the single classifiers. Note, also, that any single RS-classifier may be evaluated using different methods and it is not necessarily a decision tree.

Secondly, we propose a new boosting approach, which is based on experience-innovation principles. Assuming that overfitting is limited, it is logical to increase weights of randomly selected mis-classified patterns (innovation) in order to improve training results. As a starting point for any iteration we can use weights, which correspond to the best past result (experience). Again, the proposed system is not a voting one in difference to AdaBoost or LogitBoost[2].

Thirdly, using some criterion we can split data under expectation that the corresponding clusters will be more uniform in the sense of relations between features and target variable. The final model may be constructed as an ensemble of several models, which were evaluated independently using particular data from the corresponding clusters.

Keywords: random forests, gradient-based optimization, boosting, cross-validation, distance-based clustering.

We used an opportunity of the Challenge to test CLOP Version 1.1 -- October, 2006.
The most basic (and sufficient) instructions may be found on the last page of Ref.[3] The package is a quite efficient and can produce competitive results in application to any dataset of the Challenge. It is very easy to arrange suitable cross validations with required number of folds in order to evaluate any particular model, and there is a wide range of choices. For example, we can recommend my_model='boosting'. All necessary details in relation to this model may be found in the file "model_examples.m" in the directory ../CLOP/sample_code.
Definitely, Isabelle Guyon and her team have done an excellent work.

Also, it is worth to mention that ADA-prior task is a very similar to the recent task of PAKDD-2007 Data Mining Competition[4]. Accordingly, we applied the same preprocessing technique. Firstly, using standard methods we reduced categorical features to the numerical (dummy) values. Also, we normalized continuous values to the range [0..1]. As a result of the above transformation we created totally numerical dataset with 127 features. Then, using soft Mean-Variance Filtering[5] the number of features was reduced to 108.

Some concluding remarks:
Certainly, practical experience is the best way to learn, and I am pleased with results of the Table 1, which demonstrate significant improvement over all previous results dated July 2006.
The proper feature selection is a very essential in order reduce overfitting.
The following models appears to be the most suitable:

---

[1] L. Breiman (2001) "Random Forests", Machine Learning, 45, 1, pp.5-32.
[2] J. Friedman and T. Hastie and R. Tibshirani (2000) "Additive logistic regression: a statistical view of boosting", Annals of Statistics, 28, pp.337-374.
[3] I. Guyon and A. Alamdari and G. Dror and J. Buhmann (2006) "Performance Prediction Challenge", IJCNN, Vancouver, BC, Canada, July 16-21, pp.2958-2965.
[4] http://lamda.nju.edu.cn/conf/pakdd07/dmc07/
[5] V. Nikulin (2006) "Learning with mean-variance filtering, SVM and gradient-based optimization", IJCNN, Vancouver, BC, Canada, July 16-21, pp.4195-4202.

LogitBoost for ADA and SYLVA; RBF-SVM for GINA and LinearSVM for NOVA; regularized linear model for HIVA.

Currently, the areas of my primary interests are decision trees, random forest and a variety of boosting modifications. According to my experience, such existing packages as "randomForest" or "ADA" (R-environment) are efficient, but there may be problems with memory allocation. The performance of "TreeNet", Salford Systems, is a very good in the case of regression in difference to classification. Also, it is not easy to arrange a satisfactory cross-validation using TreeNet. Respectively, a new package (written in C with dynamic memory allocation) is under construction at the moment.

I was in Orlando, FL, twice in 2005 and 2006 and wish all participants of IJCNN-2007 very pleasant and productive work during the Conference.

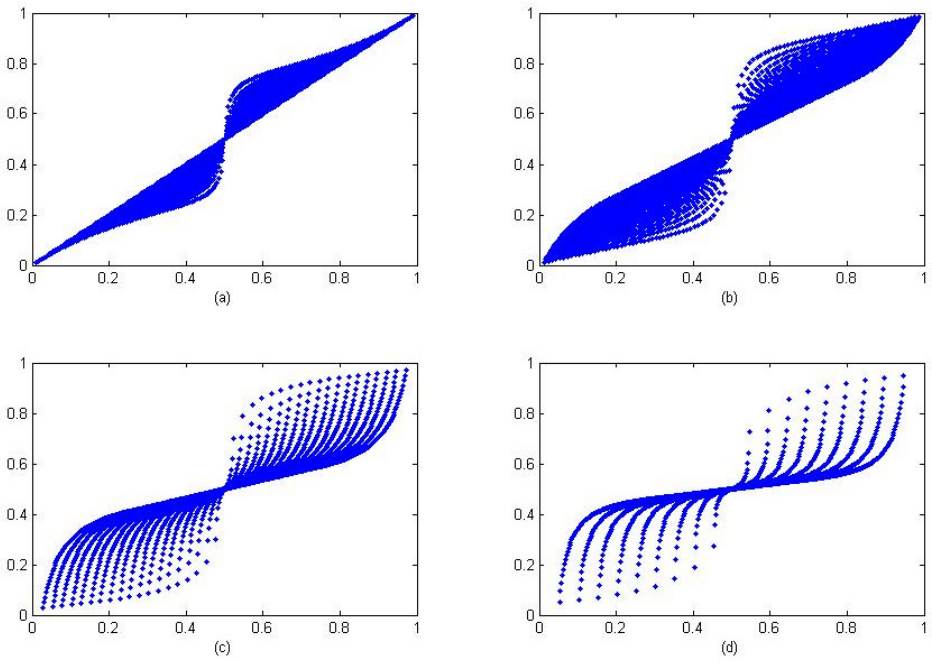

Figure1: BER vs BER where true-labels and expected-labels were replaced with each other; (a) balanced case, .. , (d) imbalanced case. These nice figures illustrate non-symmetrical properties of the BER loss function.