PERFORMANCE PREDICTION CHALLENGE: FACT SHEET

Title: A Study of Supervised Learning with Multivariate Analysis on Unbalanced Datasets

Name, address, email: Yu-Yen Ou, Department of Computer Science and Engineering, Yuan-Ze University, Chung-Li, Taiwan, yien@csie.org

Acronym of your best entry svm+ica

Reference:

Yu-Yen Ou, Hao-Geng Hung and Yen-Jen Oyang, A Study of Supervised Learning with Multivariate Analysis on Unbalanced Datasets, IJCNN06.

Method:

Our study aimed at providing effective solutions to these two challenges. For handling unbalanced datasets, we proposed that a different value of the cost parameter in Support Vector Machine (SVM) is employed for each class of samples. For handling high-dimensional datasets, we resorted to Independent Components Analysis (ICA), which is a multivariate analysis algorithm, along with the conventional univariate analysis.

<u>Preprocessing</u>

Independent Components Analysis (ICA)

Noise Reduction

<u>Feature selection</u>

Univariate Analysis

<u>Classification</u>

Support Vector Machine (SVM)

- RBF kernel and linear kernel
- give different cost parameter to the each class of data

<u>Model selection/hyperparameter selection</u>

Cross Validation

<u>Performance prediction guess.</u>

Cross Validation

Results:

In the challenge, we rank 16th as a group and our best entry is the 46th, according to the average rank computed by the organizers. Also, our method yields the second best results for GINA dataset.

Dataset	Our best entry					The challenge best entry				
	Test AUC	Test BER	BER guess	Guess error	Test score	Test AUC	Test BER	BER guess	Guess error	Test score
ADA	0.8041	0.1959	0.151	0.0449	0.2408	0.8965	0.1845	0.1742	0.0103	0.1947
GINA	0.9672	0.0328	0.04	0.0072	0.04	0.99	0.0461	0.047	0.0009	0.0466
HIVA	0.676	0.324	0.24	0.084	0.4081	0.7464	0.2804	0.2776	0.0028	0.2814
NOVA	0.9347	0.0653	0.05	0.0153	0,0805	0.9914	0.0445	0.047	0.0025	0.0464
SYLVA	0.9812	0.0188	0.002	0.0168	0.0356	0.999	0.0067	0.0065	0.0002	0.0067
Overall	0.8727	0.1273	0.0966	0.0336	0.161 (46)	0.9246	0.1124	0.1105	0.0034	0.1152 (1)

Keywords:
centering, scaling, ICA, univeriate feature selection, Chi-square,
F-score, training error, leave-one-out, K-fold cross-validation, SVM,
kernel-method, grid-search, cross-validation