CHIMIOMÉTRIE 2013, 25-26 septembre 2013,
Brest, Université de Bretagne Occidentale

A Robust Strategy for Combining Several Classifiers for Small Samples and Heterogeneous Predictors

C. Gomes[1] H.Noçairi[1] M.Thomas[1] J.F.Collin[1] G.Saporta[2]
[1] L'Oréal Research and Innovation, Aulnay-Sous-Bois, France
[2] CEDRIC-CNAM, 292 rue Saint Martin, 75003 Paris, gilbert.saporta@cnam.fr

Keywords: supervised classification, stacking, PLS-DA, Boosting, Naïve Bayes, SVM, Safety evaluation

1 Introduction

Faced to safety constraints, one cannot rely on a single prediction method, especially when the sample size is low. Stacking introduced by Wolpert (1992) and Breiman (1996) is a successful way of combining several models. We modify the usual stacking methodology when the response is binary and predictions highly correlated, by combining predictions with PLS-Discriminant Analysis instead of ordinary least squares. A strategy based on repeated split samples is then developed to select relevant variables and ensure the robustness of the final model. This method is applied to the prediction of hazard of 165 chemicals, based upon 35 in vitro and in silico characteristics.

2 A modified Stacking method

Let $f_m(x_i)$ be the prediction of a numerical response y at point x_i using a regression model m (m=1,.., M). Each model may be of any kind: linear or non-linear, non parametric, tree, neural nets etc. Stacking linearly combines the M predictors, with optimal weights w_m according to the least squares criterion. It leads to a predictor which is better for the learning set than any of a single member of the family $f_m(x)$. But all models are not on the same foot (Hastie & al, 2009): the more complex a model is, the higher its weight and may lead to over-fitting. Instead of standard predicted values, stacking uses $f_m^{-i}(x_i)$ the cross-validated prediction at x_i, not using x_i. The weights minimize the criterium (1):

$$\sum_{i=1}^{n}\left(y_i - \sum_{m=1}^{M} w_m f_m^{-i}(x_i)\right)^2 \quad (1)$$

The final model is given by equation (2):

$$\hat{y} = \sum_{m=1}^{M} w_m f_m(x) \qquad (2)$$

Stacking looks like a frequentist version of Bayesian Model Averaging or BMA, when weights are constrained being positive and to sum 1, which is recommended, but, unlike BMA, stacking does not need that all models belong to the same kind, nor that the true model belongs to the family. Experiments proved that stacking outperforms BMA in a large number of cases (Clarke, 2003) involving much simpler computations.

In a binary classification context, and since predictions from the M models are highly correlated, our modified version of stacking combines posterior probabilities, through the PLS logistic regression of Bastien et al., 2005 instead of OLS regression

3 Applications

We first experimented our method on a data set from the UCI Machine Learning repository and then on cosmetic data, with a combination of five very different classifiers: sparse PLS-DA, Tree Boosting, SVM, Naïve Bayes and a Expert Scoring technique developed by L'Oréal Research. In both cases, stacking proved its efficiency: in a second application example, a prediction model was obtained for the development of alternative approaches in safety assessment of chemicals (skin irritation, sensitization) with better performances than each of the five initial models taken separately.

Moreover, in this specific example, we had to face two critical issues: the weak number of observations and the presence of categorical data. In the first issue, since the choice of the learning set may bring some bias in the choice of relevant variables, it was necessary to perform repeated sampling. Regarding the second issue, it was necessary to avoid that some categories become empty during each random sampling. This was done using a specific stratification technique, with an acceptance-rejection scheme in order to get test and validation samples with enough observations in each category. Finally, the following decision rule with 3 outputs, focusing on high probabilities, was adopted:

- chemicals with a probability $\geq 85\%$ are predicted "Hazardous",
- chemicals with a probability $\leq 15\%$ are predicted "Not hazardous"
- chemicals with a probability between those two thresholds are "Inconclusive"

Stacking then lead to a conclusion over more chemicals than all other models since the distribution of "hazard " probabilities provided by stacking is more bimodal:

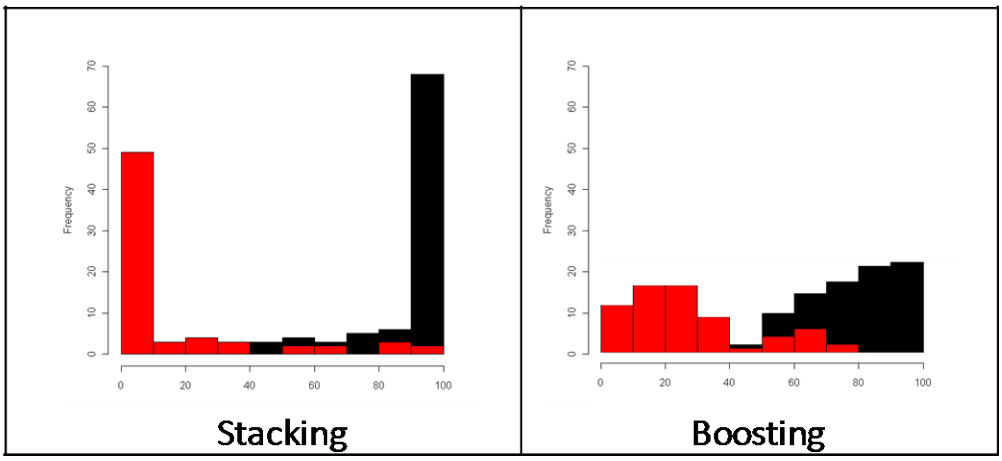

Figure 1: Distribution of scores

4 Bibliography

[1] D.Wolpert, Stacked Generalization. *Neural Networks*, 1992, 5:241-259

[2] L.Breiman, Stacked Regressions. *Machine Learning*, 1996, 24:49-64

[3] B.Clarke, Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored, *Journal of Machine Learning Research*, 2003, 4, 683-712

[4] P.Bastien,V.Esposito-Vinzi, M.Tenenhaus, PLS generalised linear regression. *Computational Statistics & Data Analysis*, 2005, 48:17–46