Stacking Prediction for Multiclass Outcomes

Hicham Nocairi*
L'Oréal Recherche, Aulnay-sous-Bois, France – hnocairi@rd.loreal.com

Gilbert Saporta
CEDRIC-CNAM, Paris, France – gilbert.saporta@cnam.fr

A large number of supervised classification models have been proposed in the literature. In order to avoid any bias induced by the use of one single statistical approach, they may be combined through a specific "stacking" meta-model.

To deal with the case of multiclass response, we introduce several improvements to stacking: combining models is done through partial least squares discriminant analysis (PLS-DA) instead of ordinary (OLS) due to the strong correlation between predictions, and a specific methodology is developed for the case of a small number of observations, using repeated sub-sampling for variable selection.

Five very different models (Bagging, Naive Bayes, Support vector machine (SVM), Sparse PLS-DA, Weighted k-Nearest-Neighbor Techniques (kknn)) are combined through this improved stacking, and applied in the following context: according to the 7th Amendment of the European Cosmetic Directive, L'Oréal is committed to stop animal testing. Consequently, l'Oréal must develop alternative approaches for safety evaluation of chemicals (Sensitization Potency).

Stacking methodology is applied to the prediction of potential sensitization of 103 chemicals, based upon in vitro and in silico characteristics in order to predict a four categories response based upon in vivo test results (Extreme/Strong/Moderate/Weak).

Results show that the stacking meta-model have better performances than each of the five models taken separately.

Keywords: Supervised classification; OLS; Sparse-PLSDA; Bagging; Bayesian; SVM; kknn; Safety evaluation.