

Stacking prediction for a binary outcome

Charles Gomes, *L'Oréal*, cgomes@rd.loreal.com

Hicham Nocairi, *L'Oréal*, hnocairi@rd.loreal.com

Marie Thomas, *L'Oréal*, mthomas@rd.loreal.com

Fabien Ibanez, *KEYRUS biopharma*,

Jean-Francois Collin, *KEYRUS biopharma*,

Gilbert Saporta, *Chaire de Statistique Appliquée & CEDRIC, CNAM*, gilbert.saporta@cnam.fr

Abstract. 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 are combined through a specific "stacking" meta-model.

To deal with the case of a binary outcome and of categorical predictors, we introduce several improvements to stacking: combining models is done through PLS-DA instead of 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 variables selection.

Five very different models (Boosting, Naïve Bayes, SVM, Sparse PLS-DA and Expert Scoring) are combined through this improved stacking, and applied in the context of the development of alternative strategies for safety evaluation where multiple *in vitro*, *in silico* and physico-chemical parameters are used to classify substances in two classes : "Sensitizer" and "No Sensitizer".

Results show that stacking meta-models have better performances than each of the five models taken separately, and furthermore, stacking provides a better balance between sensitivity and specificity.

Keywords. Stacking meta-model, binary outcome, Prediction, Sparse-PLSDA, Boosting, Bayesian, Scoring, SVM.

1 Introduction

The purpose of the present study was to develop a model allowing to predict a binary outcome (Sensitizer/ No Sensitizer) on the basis of multiple *in vitro* and *in silico* variables.

For this supervised classification objective, many methods are in competition: linear or logistic discrimination, SVM, classification trees, boosting etc. 5 different methods were used in this study: Boosting [8, 27], Naïve Bayes [18], SVM, Sparse PLS-DA [10, 11, 12] and Expert Scoring.

These methods have strong differences, but they all produce posterior probability of belonging to the group of interest ("Sensitizer"). Instead of trying to choose the best method, we combine them by the stacking methodology of Wolpert [32] and Breiman [7].

We bring several new developments to stacking: the first one is to use PLS Discriminant analysis instead of OLS regression since the outcome is categorical and the predictions are highly correlated.

We also derive a specific methodology of variable selection for categorical predictors and a small number of cases, based upon a classical splitting into three samples with a condition of no missing categories.

Comparisons between the five methods and the combined model are done according to two features: global performance with ROC analysis ([20, 21]), and concordance assessed by Principal component analysis PCA [30].

A decision system with three intervals is finally proposed to the expert, with a no-decision region. Experimental results prove the feasibility and interest of our approach.

2 Stacking regression models

Also called stacked generalization, stacking is a kind of model averaging that combines several prediction models. It was proposed by Wolpert [32] and further studied by Breiman [7]. Let $f_m(x_i)$ be the prediction of y_i at point x_i using model m ($m = 1, \dots, M$), each model being linear or non-linear, complex or not. We look for a linear combination with weights w_m of the m predictors which performs best, according to the least squares criterion.

Instead of minimizing directly $\sum_{i=1}^n (y_i - \sum_{m=1}^M w_m f_m(x_i))^2$, stacking finds the weights which minimize $\sum_{i=1}^n (y_i - \sum_{m=1}^M w_m f_m^{-i}(x_i))^2$ where $f_m^{-i}(x_i)$ is the cross-validated prediction at x_i , not using x_i , which avoids giving high weights to complex models with a risk of overfitting.

The final model being $\hat{y} = \sum_{m=1}^M w_m f_m(x)$.

It is often recommended to constrain weights to be positive and to sum to 1, if we wish to interpret weights as posterior probability like in bayesian model averaging (BMA). Clarke [9] proved that stacking outperforms BMA in a large number of cases. Under the name of blending, stacking was successfully used in the Netflix competition. Unlike most applications of BMA, stacking does not need that all models be of the same kind, or that the true model belongs to the family.

Multicollinearity and partial least square regression (PLS)

Despite predictions coming from the M models are evidently highly positively correlated, there is no literature about weights stability. Like in regression when multicollinearity is present (Naes and Indahl, [22]), some kind of regularization is needed and we advocate here the use of PLS regression ([25, 31]) instead of OLS regression in order to find the weights. Using only one PLS component is generally enough and ensures the positivity of the weights.

3 Stacking for a binary response

Straightforward generalization

Surprisingly, few attention has been given to the case where y is a categorical binary variable. Stacking may be easily generalized to this case by combining probabilities of belonging to $y = 1$ category obtained by M different models, and using values 1 and 0 (or 1 and -1) for y , since it is well known that regression of a binary variable is equivalent to discriminant analysis.

PLS-DA proposed by Barker & Rayens [3] and Nocairi & al. [23] for a categorical response could have been used. However most users prefer logistic regression but logistic regression is also affected by multicollinearity: PLS- logistic regression from Bastien & al. [2] is thus a good alternative.

Family of models

The prediction method for a binary outcome corresponds to a variety of problems arising in statistics. The most known statistical methods of prediction are Fisher's linear discriminant function and logistic regression, which have proved their interest in many cases. However, for complex phenomena (example biology), these methods doesn't take into account some statistical problems such as non linear phenomena, multicollinearity... In order to counteract these problems a lot of statistical models have been developed, belonging to different categories such as expertise based scoring, decision trees, Bayesian networks, ensemble learning, machine learning ...

The solution proposed in this article is first to choose a member of each family: expert scoring, sparse PLS discriminant analysis by Chun & Keles [11], naïve Bayesian method by Hand & Yu [18], boosting by Bühlmann [8] and Support Vector Machine by Cortes & Vapnik [14].

Interpretation and proposals

Each model, including stacking, provides a probability to be "Sensitizer". The comparison of five models versus the meta-model will be performed by ROC curve analysis. The agreement between methods will be analyzed by principal component analysis (PCA). Instead of giving a unique threshold for this probability (eg 0.5), we use a partition into 3 intervals (red, green, orange) corresponding respectively to "Sensitizer" conclusion "No Sensitizer" conclusion and a no decision (unconclusive).

4 Proposal for the case of small-sized sample and qualitative outcome

Relevant selection of variables with 6 learning sets

Each method has its own variables selection based on a split into three sub samples: learning, test and validation.

In order to avoid any bias (not pertinent decision rules) that may be induced by a particular choice of subsets, the splitting procedure is repeated several times (times six here). The final selection consists in the variables which were selected in all subsets.

The following graph (figure 1) represents the process used to build a meta-model and compute its performances on a learning/test and validation split data set.

After splitting the data (step 1), the original part appeared in the second and third steps where the learning set is split again into n learning/test subsets. For each, a stacking model is built after a parametrization phase (as shown in figure 1) and a specific variable selection is done by cross validation.

The global stacking model (build under the complete learning set) is done using the variables that have been selected across all the previous meta-models. Then, the biologist/chimist experts analyze the results (the selected variables, different materials available from the models (rules from decision tree, PLS coefficients, score results etc)) to check if there is no inconsistency. Finally, the model performances are computed on the validation set.

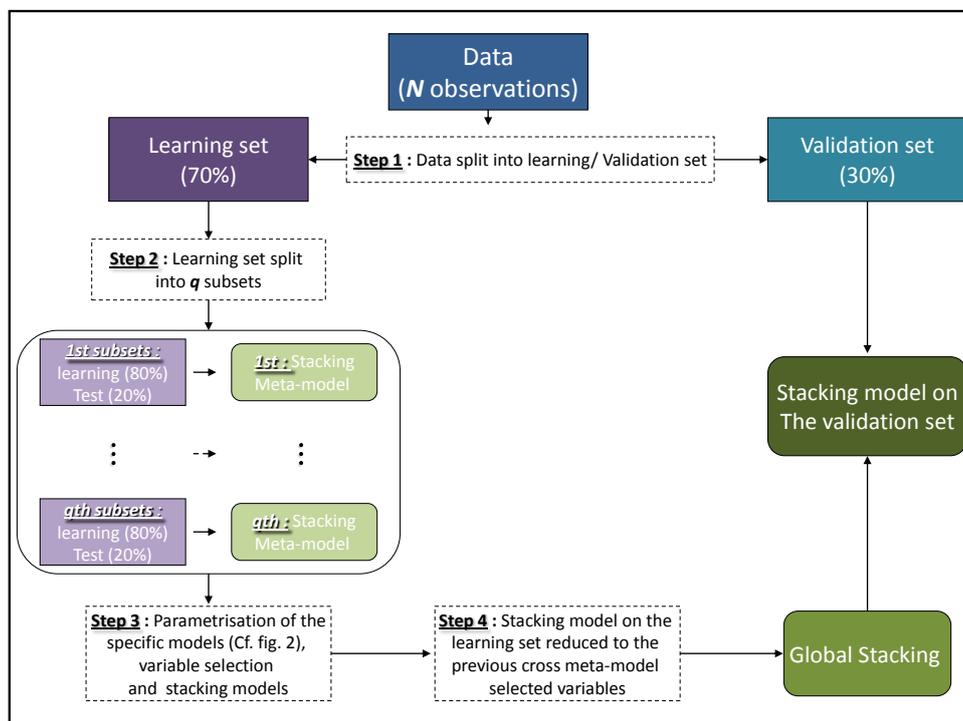


Figure 1: Process of validation rules

The figure 2 shows in details the parametrization process presented in the step 3 above.

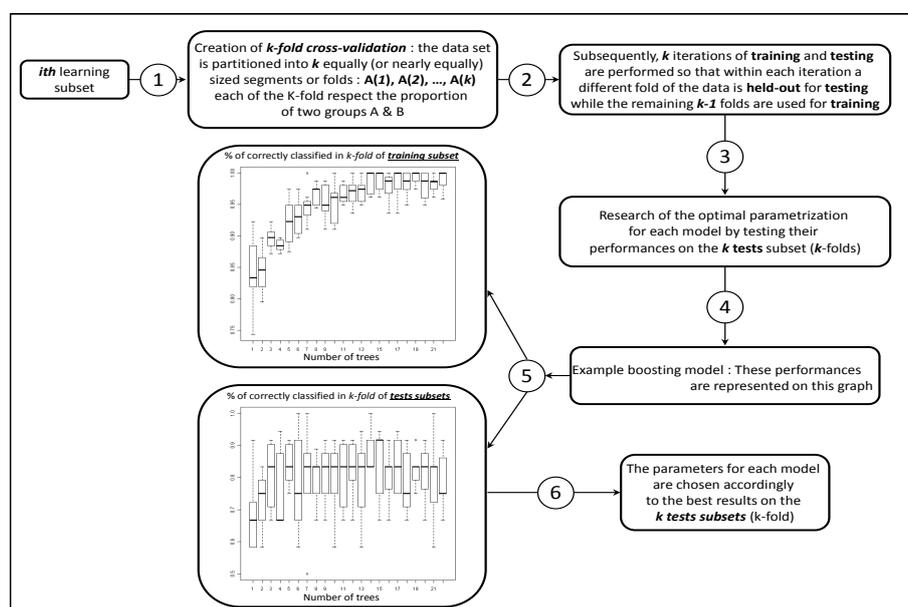


Figure 2: Parametrization process for each model

Construction of the learning and test samples

A problem occurs frequently for categorical predictors when the number of observations is small: some categories may have a null frequency, which makes estimation impossible. Some kind of stratified sampling is necessary to avoid this drawback. In our data set, the number of predictors and of their categories is too high to use balanced sampling [15]. The following heuristics is then used:

- Perform a sampling stratified upon the 2 categories of the outcome
- Reject all samples (learning or test) where a category is missing
- Draw at random *k* (here $k = 6$) samples among the remaining samples

5 Application to safety evaluation data

Statutory context

L'Oréal is developing approaches for safety evaluation (skin sensitization) of ingredients by combining multiple *in vitro* and *in silico* test methods.

Skin sensitization

Contact sensitizers are reactive molecules (*haptens*) that have the ability to modify skin proteins to form an antigen which will be recognized by specific *T* cells activated during the sensitization

process. In addition to the haptentation mechanism, contact sensitizers induce several phenotypic and functional changes of dendritic cells (*DC*) either directly or indirectly through intercellular signaling pathways implicating keratinocytes, fibroblasts and other skin cells. This rather complex and still not fully unraveled maturation process of *DC* induced by contact sensitizers, allow them, to migrate to the lymphnode, present antigen and prime efficiently hapten-specific *T* cells.

Due to the complexity of the skin sensitization process, it is now agreed that alternative hazard identification and risk assessment need to be addressed by combining a battery of methods. The aim of the statistic strategy was to combine *in silico* and *in vitro* tools, from chemical reactivity assay to *DC*-based assay, for the evaluation of skin sensitization.

Data

For this purpose we used a full data set on 165 chemicals composed of 35 different variables, representing the results from *in silico* predictions (*Derek*, *TIMES*, *Toxtree*), from *DPRA*, *MUSST*, *Nrf-2* and *PGE-2* *in vitro* assays as well as numerous physico chemical experimental or calculated parameters.

Results

The following figure shows how we combine 5 supervised classification methods by our stacking methodology. Four of them are well known (Naïve Bayes, Sparse PLS-DA, Tree boosting, SVM). The score method developed by field experts is described in the Appendix.

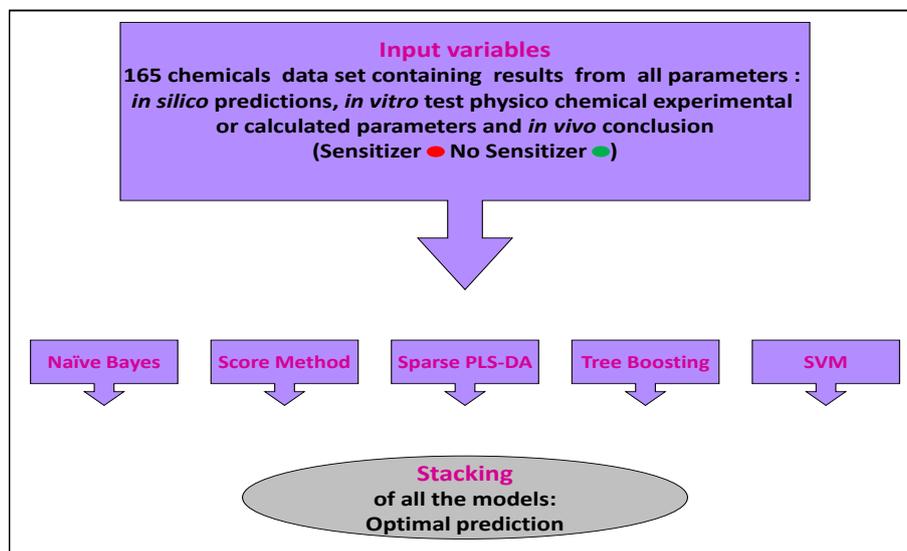


Figure 3: Visualization of the methodology

Each model, including stacking, provides a probability to be Sensitizer.

In this safety application, we have selected 10 variables, among the 35 included, thanks to the methodology described in figure 1.

As mentioned in the paragraph 1.2, the predictions provided by the five models are evidently highly positively correlated as shows the following *PCA* analysis:

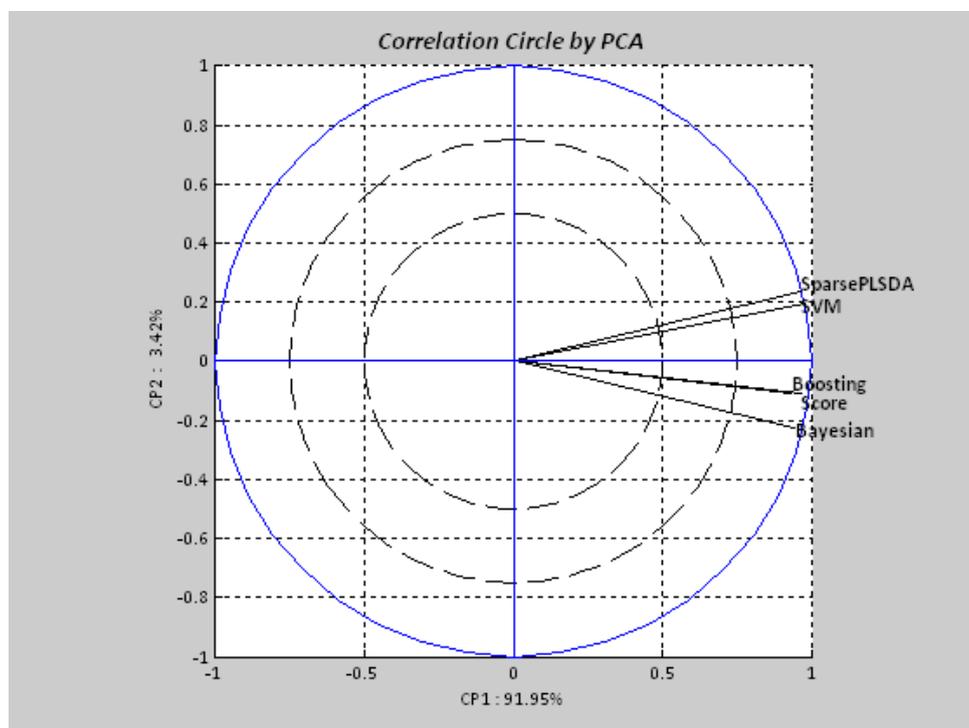


Figure 4: Correlation Circle by PCA

Thus, we used the logistics PLS DA [2, 5] instead of Logistic regression in the construction of stacking meta-model.

Results show that stacking model has better performance than all the other models taken separately on a larger set.

To illustrate this, the figure 5 present the *ROC* curves of the models used and the stacking meta-model. The stacking appears to be the most efficient (blue curve) with the highest area under the curve (0.949).

Furthermore, we observed that the distribution of the probabilities provided by stacking look more bimodal than all the other models. Thus stacking leads to the conclusion over a greater number of chemicals.

As the result, we decided to take into account only high probabilities:

- Chemicals with a probability $\geq 85\%$ are predicted "Sensitizer"
- Chemicals with a probability $\leq 15\%$ are predicted "No Sensitizer"
- Chemicals with a probability between those two thresholds are inconclusive

The thresholds of 85% and 15% were proposed by the Integrated Research European project called *OSIRIS* [24]. The main objective of *OSIRIS* being to optimize strategies for risk assessments of chemicals.

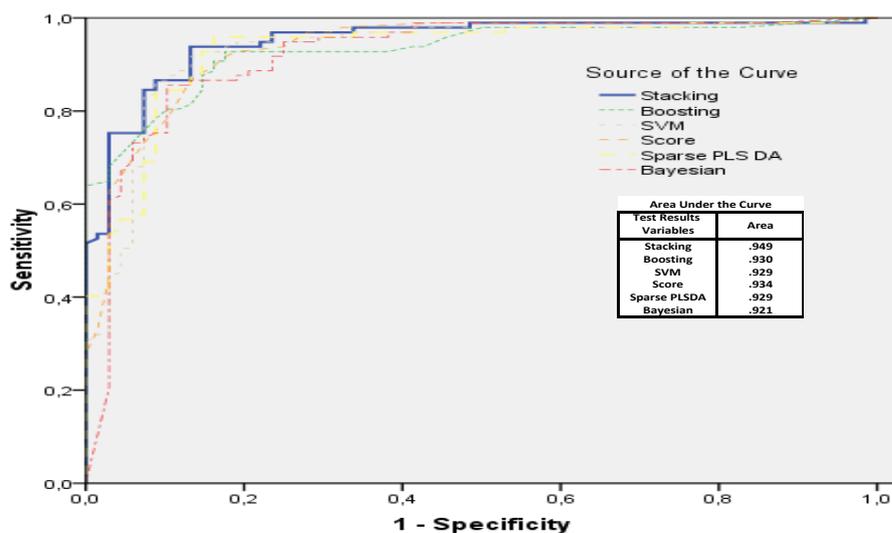


Figure 5: ROC curves of the models

This approach defines a confidence area. For example Boosting model concludes on 40% (67/165) of chemicals only, compared to Stacking model that concludes on 82% (135/165) of chemicals, as shown in figure 6.

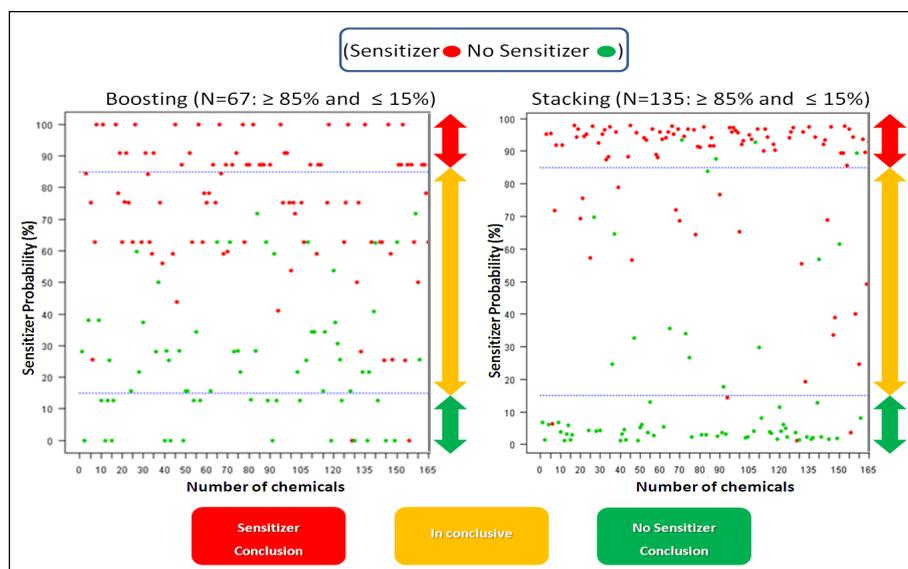


Figure 6: Confidence area of the boosting model and of the stacking model

As described in the methodology (figure 1), we have evaluated the performances on the validation set ($N = 50$). The table below shows that the stacking model provides the highest response rate (80% of validation set) and the best balance between sensitivity (91%) and specificity (94.44%).

True class	Predicted class	Boosting	SVM	Score	Sparse	Bayesian	Stacking
Sensitizer	Sensitizer	11	19	7	16	19	20
No Sensitizer	Sensitizer	0	1	0	1	2	1
Sensitizer	No Sensitizer	2	3	1	3	1	2
No Sensitizer	No Sensitizer	7	15	13	15	16	17
N		20/50	38/50	21/50	35/50	38/50	40/50
Sensitivity		84.61	86.36	87.50	84.21	95.00	91.00
Specificity		100	93.75	100	93.75	89.00	94.44
Concordance		90.00	89.47	95.24	88.57	92.00	92.50
Kappa		0.79	0.78	0.89	0.77	0.84	0.85

Table 1: Performance comparisons (take into account only high probabilities ($\geq 85\%$ and $\leq 15\%$)) on a validation set ($N = 50$)

the score technique has the best kappa, but eliminates most of the data, which is not acceptable from an industrial point of view.

To achieve this methodology, we have chosen the R environment combined with a web interface where most of these methods were already available. We used the following packages for: SparsePLS [13], pls [26], rpart [29], adabag [17], SVM [16], penalizedSVM [4] and Stacking [5].

6 Conclusion and perspectives

We have presented in this article improvements and applications of stacking meta-model for a binary toxicological outcome. This methodology gives a prediction model with better performances for the development of alternative approaches in safety evaluation of chemicals than each of the initial five models separately.

The binary outcome revealed the difficulty to split the data into learning/test subsets. To overcome this problem we have proposed a specific methodology based on samples stratification.

Various improvements are possible, for example we could add several models such as: Decision Tree based model on C4.5 [28], Neural Network [1], Multiblock Redundancy Analysis [6], partitioning around medoids (PAM: [19])...

Moreover, this methodology can be extended to the multi-class case, in the framework of safety data: such methodology could potentially predict several ordered categories of potency.

Acknowledgement

The authors are very grateful to Cécile PIROIRD, Silvia TEISSIER, Thierry PAULOIN, Nathalie ALEPEE, Valérie MICHAUT, Jean-Roch MEUNIER and Frédéric LEROY for fruitful discussions and careful and critical reading of the manuscript.

Appendix

Scoring: Each variable is transformed into a discrete "sensitizer" score with values -3 -2- 1 +1 +2 +3. The (empirical) rule is the following for a categorical variable: let m be a category of this variable; if the number of "sensitizer" chemicals in m is larger than twice the number of "no sensitizer" chemicals in the same category, but less than 3 times, the score is equal to 1. If

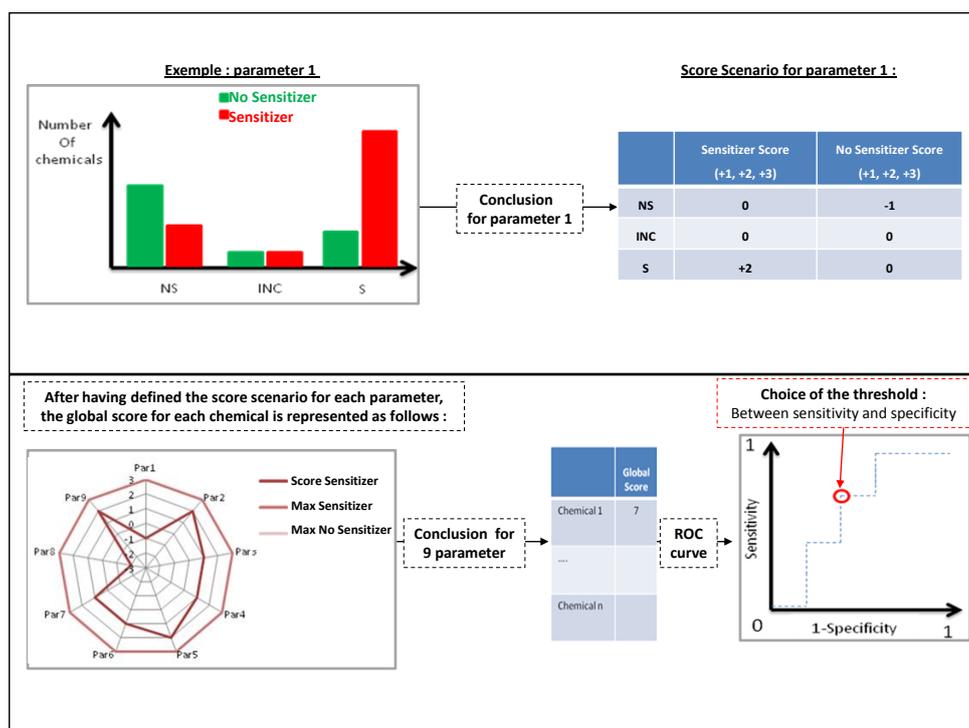


Figure 7: Figure 7. Principle for Score Models

the number of "sensitizer" chemicals in m is larger than 3 times the number of "no Sensitizer" chemicals in the same category, but less than 4 times, the score is equal to 2 etc. A symmetric rule is applied for negative values. For continuous variables, we use thresholds instead of categories. Partial scores for each chemical may be displayed in a radar chart. The global score is the sum of partial scores, which leads to a ROC curve.

Bibliography

- [1] Anderson, J. A. (1995). Early network models. In J. A. Anderson (Ed.). *An introduction to neural networks*, 209–238. Cambridge: The Mit Press.
- [2] Bastien, P., Vinzi V.E. & Tenenhaus, M. (2005) *PLS generalised linear regression*. Computational Statistics & Data Analysis. **48**, 17–46.
- [3] Barker, M. & Rayens, W. (2003) *PLS for discrimination*. Journal of Chemometrics, **17**, 166–173.
- [4] Becker, N., Werft, W., & Benner, A. (2010) *penalizedSVM: Feature Selection SVM using penalty functions*, R package version 1.1. <http://CRAN.R-project.org/package=penalizedSVM>.

- [5] Bertrand, F., Meyer, N., & Maumy-Bertrand, M. (2011) *plsRglm : Partial least squares Regression for generalized linear models, R package version 0.6.5*. <http://cran.r-project.org/web/packages/plsRglm/>.
- [6] Bougeard, S., Hanafi, M., Nocairi, H., & Qannari, E.M. (2006) *Multiblock canonical correlation analysis for categorical variables: application to epidemiological data* Multiple correspondence analysis and related methods (Greenacre, M., Blasius, J. Editors). Chapman & Hall/CRC.
- [7] Breiman, L. (1996) *Stacked Regressions*. Machine Learning, **24**, 49–64.
- [8] Bühlmann, P., & Hothorn, T. (2007) *Boosting Algorithms: Regularization, Prediction and Model Fitting*. Institute of Mathematical Statistics, Statistical Science. Vol 22, **7**, 477–505.
- [9] Clarke, B. (2003) *Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored*. Journal of Machine Learning Research, **4**, 683–712.
- [10] Chung, D. & Keles, S. (2010) *Sparse Partial Least Squares Classification for High Dimensional Data*. Statistical Applications in Genetics and Molecular Biology, Vol. 9, Article **17**.
- [11] Chun, H. & Keles, S. (2010) *Sparse partial least squares for simultaneous dimension reduction and variable selection*. Journal of the Royal Statistical Society, Series B. **72**, 3–25.
- [12] Chun, H. & Keles, S. (2009) *Expression quantitative trait loci mapping with multivariate sparse partial least squares regression*. Genetics. **182**, 79–90.
- [13] Chung, D., Chun, H. & Keles, S. (2009) *SPLS : Sparse Partial Least Squares (SPLS) Regression and Classification, R Package version 2.1-0*. <http://CRAN.R-project.org/package=spls>.
- [14] Cortes, C. & Vapnik, V. (1995) *Support-vector network*. Machine Learning, **20**, 1–25.
- [15] Deville, J.C. & Tillé, Y. (2004) *Efficient balanced sampling: The cube method*. Biometrika, **91**, **4**, 893–912.
- [16] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. (2011) *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, R package version 1.5-26*. <http://CRAN.R-project.org/package=e1071>.
- [17] Cortes E.A., Martinez M.G. & Rubio N.G. (2011) *adabag: Applies Adaboost.M1 and Bagging, R package version 1.1*. <http://cran.r-project.org/web/packages/adabag>.
- [18] Hand, D.J. & Yu, K. (2001) *Idiots Bayes—not so stupid after all?* International Statistical review, **69**, 385–398.
- [19] Kaufman, L. & Rousseeuw, P.J. (1987) *Clustering by means of medoids (PAM)*. In Dodge, Y. (ed.), *Statistical Data Analysis Based on the L1-norm and Related Methods*. North Holland, Amsterdam, 405–416.
- [20] Lusted L.B. (1960) *Logical analysis in roentgen diagnosis*. Radiology, **74**, 178–93.

- [21] Lusted L.B. (1971) *Signal detectability and medical decision making*. *Sciences*, **171** 1217–9.
- [22] Naes, T. & Indahl, U. (1998) *A unified description of classical classification methods for multicollinear data*. *Journal of Chemometrics*. **12**, 205–220.
- [23] Nocairi, H., Qannari, E.M., Vigneau E., & Bertrand D. (2005) *Discrimination on latent components with respect to patterns. Application to multicollinear data*. *Computational Statistics & Data Analysis*. **48**, 139–147.
- [24] OSIRIS: Stakeholder Workshop, Integrated Testing Strategies. EU Integrated Research Project, 8–9 March 2011, Leipzig. Germany.
- [25] Stahle, L. & Wold, S. (1987) *Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study*. *Journal of Chemometrics*. **1**, 185–196.
- [26] Ron, W. & Bjrn-Helge, M. (2007) *PLS:Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*, *R package version 2.1-0*. <http://mevik.net/work/software/pls.html>.
- [27] Ron M., Gunnar R. (2003) *An Introduction to Boosting and Leveraging*. *Advanced Lectures on Machine Learning*, Springer, Review Paper.
- [28] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- [29] Therneau T.M. & Atkinson, B (2010) *Rpart : Recursive Partitioning*, *R package version 3.1-48*. <http://CRAN.R-project.org/package=rpart>.
- [30] Wold H. (1966) *Estimation of principal components and related models by iterative least squares*, in *Multivariate Analysis*, Krishnaiah P. R. (Ed.), Academic Press, New York, 391–420.
- [31] Wold S., Martens, H. & Wold, H. (1983) *The multivariate calibration problem in chemistry solved by the PLS method*, in *Proc. Conf. Matrix Pencils*, Ruhe A. & Kagström B. (Eds), March 1982, *lecture Notes in Mathematics*, Springer Verlag, Heidelberg, 286–293.
- [32] Wolpert, D.(1992) *Stacked Generalization*. *Neural Networks*, **5**, 241–259.