Comparative Study of two Classification Algorithms for the Prediction of Drug-induced Phospholipidosis

Marcel Youmbi Foka, *Timothy Clark*
Computer-Chemie-Centrum, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraße. 25, 91052 Erlangen, Germany

Phospholipidosis (PPL) are structural components of mammalian cytoskeleton and cell membranes [1,2]. They define an excessive accumulation of intracellular phospholipids.
In this work, two Machine Learning techniques are successfully applied to classify phospholipidosis as active or inactive with respect to a specific target biological system. A comparison of NaiveBayes and RandomForest algorithms, which are fully implemented in Weka [3], is presented, in an effort to identify drugs inducing phospholipidosis.
Based on applying Bayes' theorem which is credited to the British Mathematician Thomas Bayes, the Naïve Bayes classifier is constructed from the probability model, which is abstractly a conditional model

$$p(C|F_1,\ldots,F_n)$$

over a dependent class variable C with a smaller number of outcomes or classes, conditional on several feature variable F_1 through F_n.
Random Forests, proposed by Breiman [4] is a collection of tree predictors

$$h(x;\Theta_k),\ k = 1,\ldots,K$$

x is the observed input (covariate) vector of length p with associated random vector X.
The parasurf descriptors and the additional water-octanol partition coefficient, calculated with AM1, AM1*, MNDO, MNDO/d, PM3, PM6, using the default isodensity and the solvent excluded surfaces were used to capture chemical informations. The total data set was randomized, then 50% were used as training set and the remainder were used to assess the generalization performance.
In general using the 10-fold cross validation, the RandomForest algorithm was able to produce models with best predictive performance (83%). Models obtained from the NaiveBayes algorithm were totally surface dependent. For the same Hamiltonian, the use of the solvent excluded surface always yielded models with better accuracies than the default isodensity surface.

[1] R. Lowe, R. C. Glen, and J. B. O. Mitchell, *Molecular Pharmaceutics*, **2010**, 7, 1708-1714.
[2] Chatman LA, Morton D, Johnson TO and Anway SD, *Toxicol Pathol,* **2009**, 37, 997-1005.
[3] Ovidiu Ivanciuc, *Curr. Topics Med. Chem*, **2008**, 8, 1691-1709.
[4] L. Breiman, Random Forests, *Machine Learning,* **2001**, 45, 5-32.