

Data Mining for Toxicity Prediction

Dr. Daniel Neagu

University of Bradford

Department of Computing, Richmond Road, Bradford BD7 1DP

D.Neagu@bradford.ac.uk

<http://www.comp.brad.ac.uk/contact/contact.php/staff/p/dneagu>

Data Mining is the process of search for patterns within large collections of (possible noisy) data. Key Data Mining problems addressed by intelligent systems are: Diagnosis, Pattern Recognition, Prediction, Classification, Clustering, Regression, Optimization and Control. The processes of data classification and regression are part of a direction known as Predictive Data Mining, having the goal to obtain predictive models for a specific target, based on relationships among large number of input variables. Predictive Toxicology is a multidisciplinary science that requires a close collaboration between researchers from Toxicology, Chemistry, Biology, Statistics and Artificial Intelligence – Machine Learning and Data Mining fields. The type of the problem addressed in our studies will be defined: to model and predict, for toxicological endpoints, the bio-chemical action of different classes of chemical compounds through relations dependent on their structures expressed by chemical (constitutional, geometrical, topological, electrostatic, quantum-chemicals, hydrophobic etc.) descriptors, known as the QSAR (Quantitative Structure Activity Relationship) problem.

Nowadays, in the domain of Predictive Toxicology, various Machine Learning techniques are applied: Decision and Regression Trees, k-Nearest Neighbour, Classification Rules, Artificial Neural Networks, Fuzzy Inference Systems. There are also approaches based on competitive or cooperative combinations of two or more global or local techniques (like Neuro-Fuzzy Systems) to build predictive models. Instead of concentrating on building the best expert for the whole domain, these techniques propose ensembles of some good local models, which conceptually are experts on different sub-sets of the domain. According to the results so far, there are reasons to affirm that individual models are sometimes too weak to have generalizing predictive power: mixture of experts is a way to improve the performance of single models, in the context of diversity.

A gentle introduction of key terms in Machine Learning and Data Mining will be given in order to identify the overlapping areas, applications and possible confusions generated by the various points of view and vocabulary terms. Some of the most used Machine Learning and Data Mining techniques for applications to predict toxicity of chemical products will be discussed. Some well known Machine Learning techniques will be compared in terms of their performances for predictive toxicology studies. Additional advantages of ensembles of experts will be reviewed. The need for benchmarks and a review of possible data resources will be given. Case studies of data pre-processing (normalization cases), feature selection and model development and integration will be discussed: their advantages and drawbacks will be highlighted.

Technical aspects of models development using existing software tools for predictive data mining applications will be presented, mainly regarding file formats, their exchangeability and performances. The challenge of data representation for homogeneous processing and data exchange will be addressed: some existing XML formats will be listed.

Some future research issues will be targeted mainly on data fusion, models re-usability and the place of the human expert in the process of automatic data mining: the need to develop software tools that carefully address the division of workload and expertise between the computer and the user.