A Comparison of Probabilistic, Neural, and Fuzzy Modeling in Ecotoxicity

Giuseppina Gini, Marco Giumelli
DEI, Politecnico di Milano, piazza L. da Vinci 32, Milan, Italy
Emilio Benfenati, Nadège Piclin
Istituto Mario Negri, via Eritrea 62, Milan, Italy
Jacques Chrétien, Marco Pintore
Lab Chemometrics & Bioinformatics, University of Orléans, France

Abstract. The common practice in inducing toxicity models from data is regression analysis. The predictive power of such models is usually poor with very different molecules and toxicity end-points. In the present work, we study toxicity classification of pesticides: we discuss about knowledge representation, and we test probabilistic and softcomputing techniques. We conclude with the interpretability of the induced models.

1. Introduction

The study of the consequences of chemicals on the health of human beings and wildlife requires new computer-based approaches to analyse complex information and to automatically discover knowledge implicitly contained in data [4]. The goal of toxicity prediction is to describe the relationship between chemical properties and biological and toxicological processes. So far most of the studies are Quantitative Structure-Activity Relationship (QSAR) models, which predict a continuous value [5]. However, due to the variability of the toxicity phenomenon [1], classification may present advantages because it determines intervals of the toxic effect, and has direct application for regulation of chemicals.

In the present study we report on different methods to build models for aquatic toxicity prediction, as linear and non linear regression, Bayesian networks [7], Self-Organizing Map (SOM) [8], and Support Vector Machine (SVM) [2]. Fuzzy logic [12] is finally used. Some major points are discussed:

- Choice of the *chemical class before studying toxic activity*: it is difficult to obtain good predictive models for heterogeneous sets of compounds.;
- *Reliability of prediction*; neither the means square error, nor the correlation, are good measures of the predictive power of a model. A commonly used validation method is the leave-one-out (LOO) cross-validated coefficient R^2_{cv}, calculated as 1 – PRESS/SS_Y, where PRESS is the prediction error sum of squares of the model and SS_Y is the sum of squares for the Y variable. With this method results are optimistic compared to the validation with an external set; however a major problem in toxicity prediction is the availability of too few experimental data to split them in two sets (training and test);
- *Molecule description*: our hypothesis is to consider the *whole molecule* and not some of its fragments.

The road here proposed is to see how to use the available knowledge, not to produce a symbolic system to integrate the induced models.

2. Knowledge Representation: Eco-Toxicity and Molecules

In QSAR models, logP is commonly used to model the bioaccumulation potential of chemicals, but alone is not enough to model toxicity . Two of the open problems in QSAR are defining the relevant molecular descriptor,s and the availability of general models. We faced these problems studying pesticides. Structures and toxicity data of pesticides have been obtained from "The Pesticide Manual (1997)". Lethal concentration for 50% of the animals (LC_{50}) is the chosen endpoint. A set of 235 pesticides (Table 1). has been obtained, and checked, for rat and aquatic toxicity. Modelling has been done on the whole data set or on chemically homogeneous subsets, which share the same biological mechanism. Correlation analysis of the toxicity is in Table 2. About chemical knowledge, the practice in AI has been to represent molecules as graphs. Nodes represent atoms, arcs are the bonds. This view of the chemical structures is too weak: graphs represent only the planar topology of the molecule, and are unable to consider the 3D structure or energy formation. Our model building uses more chemical information, as in Figure 1.

Table 1: Number of compounds for each chemical class.

Chemical Class	**Total**	**Training Set**	**Test Set**
Anilines	39	21	18
Aromatic halogenated	83	57	26
Carbamates	26	23	3
Heterocycles	119	93	26
Organophosphorous	59	27	32
Ureas	31	24	7
Different Class	5	4	1
Total	**362**	**249**	**113**

Table 2: Correlation matrix of LC_{50} (correlation coefficient r).

	Quail	Trout	Daphnia
Trout	-0.02		
Daphnia	0.21	0.06	
Duck	0.55	0.44	0.14

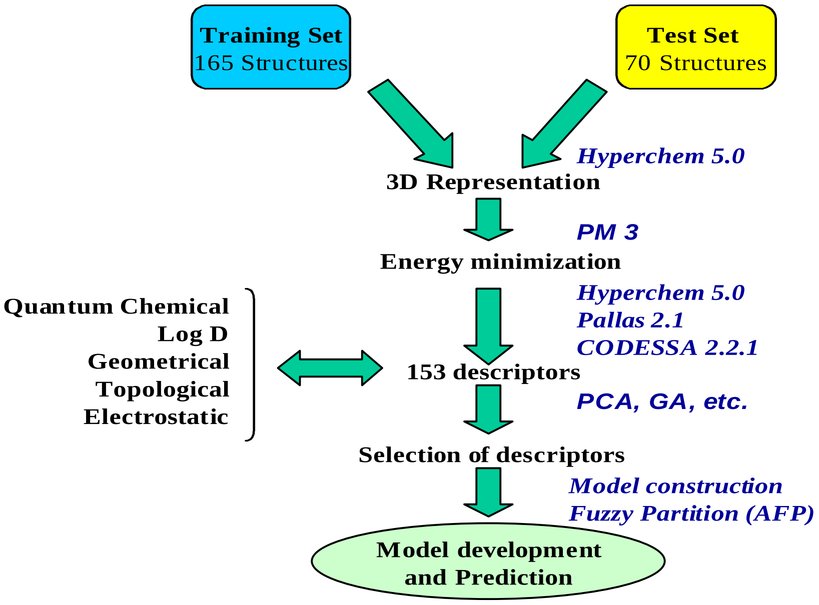

Figure 1: Construction of the models.

3. Regression studies

3.1. Multivariate Analysis

Data have been analyzed using principal component regression (PCR) and partial least squares (PLS). Table 3 shows, for internal validation, that no results were obtained modelling all pesticides. PLS models on single chemical classes showed variable results: no toxicological end-point was predicted for all chemical classes, and no chemical class gave good results for all end-points (Table 3).

Table 3: Linear regression for LC_{50}using PLS. Figures indicate R^2_{cv} , when > 0.5.

Chemical Class	**rainbow trout**	**daphnia**	**rat**	**duck**	**quail**
Aniline	0.78	0.72	No results	No results	No results
Carbamate	No results	No results	No results	No results	No results
Organophosphorus	No results	0.69	No results	No results	No results
Urea	0.78	0.85	0.59	No results	No results
Heterocyclic	No results	0.56	No results	0.55	No results
Halogenated aromatic	No results	No results	No results	No results	0.55

3.2. NN

We used NN as non-linear regression. Descriptors selection was necessary. In order to statistically study the input variables, Principal Component Analysis (PCA) has been used, since it permits to evaluate correlation and relevance of variables, to see the multivariate information characterizing the objects in a two dimensional orthogonal space, to synthesise data and eliminate noise. The loading plot graph permits to identify the role of each variable (molecular descriptor) towards the two principal components considered and their direct or inverse correlations. Considering loading plots on I and II component, 15 variables with the greatest weight have been selected. The same approach has been used on the III and IV principal component yielding a selection of six other descriptors. The total variance expressed by the four components is 90%. With the selected variables we built a predictor with back-propagation NN. Results were good in LOO, not for an external validation set of 13 molecules. A problem may be the coupling of PCA and NN: PCA compresses data on the basis of linear behavior, but does not keep into account non-linearity assumed for aquatic toxicity.

So we switched from regression to classification, a matter addressed also in other studies [11].

4. Classifiers

We used the log values of LC_{50}, scaled between 0 and 1, and we divided this interval into four classes (ClassA 0÷0.25; ClassB 0.25÷0.5; ClassC 0.5÷0.75; ClassD 0.75÷1).

4.3 Bayesian

Bayes theorem expresses a way to update beliefs in a hypothesis given additional evidence and the background context. Bayesian networks are a graphical formalism for reasoning about probability distributions: they use a directed acyclic graph (DAG) to encode conditional independence assumptions about the domain. Each variable is represented as a node, an arc between two nodes denotes the existence of a direct probabilistic dependency between the two variables. A Bayesian Network classifier contains a node C for the class variable and a node X for each of the domain feature. Given an instance vector x, the network computes the probability $P(C = c_k | X = x)$ for each class c_k. If the true distribution $P(C | X)$ is known we achieve the optimal classification by selecting the class c_k for which the probability is maximum. Unfortunately the true distribution can only be approximated from the training set, a process NP-hard for general networks; so we chose the Naï ve Bayes Classifier, where all the features nodes are directly connected to the class node. Table 4 shows results using LOO with Bayda (www.cs.Helsinki.FI/research/cosco.Projects/).

Table 4: Correct classification (%) by Bayda using LOO for all or selected descriptors

	Trout		Daphnia		Rat	
Chemical class	All descr.	Sel. descr.	All descr.	Sel. descr	All descr.	Sel. descr
Anilines	46	38	**63**	**60**	51	51
Carbamates	50	**100**	42	54	35	35
Heterocycles	47	42	51	**60**	45	49
Halogenated Aromatics	49	35	45	54	42	39
Organophosphorous	39	37	51	49	27	37
Ureas	**68**	**61**	55	34	**64**	**71**

4.2 Self Organizing Maps (SOM)

Vector Quantization (VQ) networks are unsupervised density estimators. Each competitive unit corresponds to a cluster, the center of which is called a "codebook vector". Kohonen's learning law finds the codebook vector closest to each training case and moves the "winning" codebook vector closer to the training case, according to a learning rate. In a SOM, the neurons (clusters) are organized into a grid. The grid exists in a space that is separate from the input space; any number of inputs may be used as long as the number of inputs is greater than the dimensionality of the grid space. A SOM tries to find clusters such that any two clusters that are close to each other in the grid space have codebook vectors close to each other in the input space. After SOM have been trained, the quality of prediction has been evaluated on the mapping precision (Table 5). However with an external validation set, no results could be obtained. For this reason, subsets of chemical classes were considered for daphnia (Table 6). The results are correct, (about 70%) but difficult to generalize with an external validation set.

4. 3 Support Vector Machines (SVM)

A common problem of neural networks is overfitting. SVM try to avoid this by finding the hyperplane with maximal distance from the hyper plane to data in the feature space. The hyper plane is represented by an expansion of a subset of the training data, called support vector. Support Vector Machines non-linearly map their n-dimensional input space into a high dimensional feature space, where a linear classifier is constructed. Two features make this approach successful: the generalisation ability, and the

Table 5: Correct classification (%) obtained with SOM using LOO for all the compounds on five species.

Animal	Daphnia	Duck	Quail	Rat	Trout
Correct %	65	71	76	63	55

Table 6: Comparison between %values obtained with LOO and validation set for Daphnia (SOM).

	LOO	**Validation Set**	
Chemical Class		**Training Set**	**Test Set**
Anilines	69	71	21
Aromatic halogenated	71	75	38
Carbamates	73	83	0
Heterocycles	65	70	48
Organophosphorous	67	78	45
Ureas	65	75	60

simplicity (construction of the classifier only needs to evaluate an inner product between two vectors of the training data). For classification, SVM operate by finding a hyper surface in the space of possible inputs, to split the positive examples from the negative ones. The split will be chosen to have the largest distance from the hyper surface to the nearest of the positive and negative examples. Although the separating hyper plane is linear, it is on a feature space induced by a kerne, hence can separate data, which are linearly no separable in input space. For our experiments (table 7) we applied the MatLab OSU_SVM toolbox, (http://www.eng.ohio-state.edu/ maj/osu_svm).

Table 7: Comparison between % values obtained with LOO and validation set for Daphnia (SVM).

	LOO	**Validation Set**	
Chemical Class		**Training Set**	**Test Set**
Anilines	86	87.5	54.5
Aromatic halogenated	53	64	42
Carbamates	85	88	67
Heterocycles	64	61	62
Organophosphorous	84	53	41
Ureas	97	90	60

5. The new Adaptative Fuzzy Partition (AFP) classifier

AFP is a supervised classification method [9]. It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces. In a first phase, the global descriptor hyperspace is cut into two subspaces where the fuzzy rules are derived. These two subspaces are divided step by step into smaller subspaces until certain conditions are satisfied, namely:

- the number of molecular vectors within a subspace attains a minimum threshold number;
- the difference between two generated subspaces is negligible in terms of chemical activities;
- the number of subspaces exceeds a maximum threshold number.

The aim of the algorithm is to select the descriptor and the cut position which allow getting the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace A and in its neighbouring subspaces. If the number of trial cuts per descriptor is defined by N_cut, the number of trial partitions equals (N_cut + 1) N. Only the best cut is selected to subdivide the original subspace.

All rules created during the fuzzy procedure are considered to establish the model between descriptors hyperspace and chemical activities. Indicating with P($x_{1,}$ x_2, ... x_n) a molecular vector, a rule for a subspace S_k is defined by:

if x_1 is associated with $m_{1k}(x_1)$ ***and*** *x_2 is associated with $m_{2k}(x_2)$ …* ***and*** *x_N is associated with $m_{Nk}(x_N)$ Þ the score of the activity O for P is O_{kP},*

where x_i represents the value of the i^{th} descriptor for the molecule P, μ_{1k} is the membership function for the subspace k related to descriptor i and O_{kP} is the activity value related to the subspace S_k. The “and” is represented by the *Min operator* and the membership functions are defined by trapezoidal shapes. If the width of a subspace S_k on the i^{th} dimension, after each cut, is represented by w_i, the p and q parameters defining the shape of the trapezoid are calculated as $p = \lambda_i w_i$ and $q = \nu_i w_i$, where the to parameters λ_i and ν_i vary so that $p \geq 1$ and $q \leq 1$. All the rules created during the fuzzy procedure are considered to build the model. After establishing AFP model, a centroid defuzzification procedure [5] determines the chemical activity of a new test molecule. All the subspaces k are considered and the the membership degree for the activity O and a molecule Pj is:

$O(P_j) = \dfrac{\sum_{k=1}^{N\_subsp} (\mathrm{Min}_i^N \, \mu_{ik}(x_i)_{P_j}).(O_k)}{\sum_{k=1}^{N\_subsp} (\mathrm{Min}_i^N \, \mu_{ik}(x_i)_{P_j})}$	$\mu_{ik}(x_i)_{P_j}$ membership function for descriptor i in molecule Pj N total number of molecular descriptors N_subsp total number of subspaces O_k, global score of the activity in the subspace S_k

The full set of the pesticides was used for the prediction of toxicity against the trout, considering two different classifications. The first one includes four toxicity classes as established by the EU Directive 92/32/EEC, the second defines three toxicity classes so to include in the training set a similar number of compounds. The relevant descriptors were selected by an innovative procedure based on genetic algorithms [10]. After establishing the AFP model on the training set, the toxicity classes for the test set were predicted. The statistical results for the validation tests are reported in Table 8. For the homogeneous intervals, the AFP model predicts the correct activity for 71% of the test set compounds. Similar results are derived for the training set s, confirming the goodness of the model. Moreover, class 3, including high toxicity compounds, was the best correctly predicted (86%). In the case of the EU intervals, AFP established a model able to predict correctly 60% of the test set compounds and 78% of the training set compounds (Table 8). The most toxic class was better predicted (69%). AFP builds up a scheme of the rules used for each toxicity class, as for example:

> if 0 < x(log D-pH5) < 0.26 and 0 < x(Balaban Index) < 0.51 and x(Randic Index) > 0.81…. ⇒ the membership degree of class 1, for the compound 34, is 0.5.

Table 8: Statistical values (%) for classification established on the homogeneous and EU intervals.

Classes	**range EU**	**EU %**		**range hom.**	**Homogeneous %**	
	LC50 (mg/l)	Training Set	Test Set	LC50 (mg/l)	Training se	Test Set
Class1	> 100	77	58	> 12	75	57
Class2	10 – 100	79	46	1.2 –12	76	63
Class3	1 – 10	74	58	< 1.2	64	86
Class4	< 1	83	69			
Total		**78**	**60**		**72**	**71**

6. Conclusions

Specific problems for ecotoxicity are the incomplete knowledge on toxicity, the quality of the experimental data affected by a high variability, their limited number. Here we show results with LOO and validation set; and observe that LOO is quite optimistic, probably for the high discontinuity of the function to be learned. AFP shows the best capability to generalize, and allows deriving rules which could give insights in the biochemical processes.

Acknowledgments

We acknowledge the financial contribution of the European Commission, project HPRN-CT-1999-00015.

References

[1] Benfenati, E., Grasso, P., Pelagatti, S. and Gini, G. On variables and variability in predictive toxicology. *IV Girona Seninar on Molecular Similarity*, July 5-7, 1999, Girona, Spain.

[2] Cristianini, N. and Shawe-Taylor, J. An Introduction to Support Vector Machines, Cambridge Univ. Press, 2000.

[3] Gini, G., Lorenzini, M., Benfenati, E., Grasso, P. and Bruschi, M.. Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network, *J. Chem. Inf. Comp. Sci.*, **39** (1999) 1076-1080.

[4] Gini, G. and. Katrizky, A. (eds). Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools. SS-99-01, AAAI Press, Menlo Park, California, 1999.

[5] Gupta, M. M. and Qi, J. Theory of T-norms and fuzzy inference methods, *Fuzzy Sets and Systems*, **40** (1991).

[6] Hansch, C., Hoekman, D., Leo, A., Zhang, L. and Li, P. The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicology Letters* **79** (1995) 45-53.

[7] Heckerman, D., Geiger, D. and Chickering, D. Learning Bayesian Networks: the combination of knowledge and statistical data. *Machine Learning* **20**, (1995) 3, p 197-243.

[8] Kohonen, T. The self-organizing map, *Neurocomputing*, **21** (1998) 1-6.

[9] Pintore, M., Ros, F., Audouze, K. and Chrétien, J.R. Adaptive Fuzzy Partition in Data Base Mining, 2001.

[10] Ros, F., Pintore, M. and Chrétien, J. R. Molecular Descriptor Selection Combining Genetic Algorithms and Fuzzy Logic, submitted, 2001.

[11] Torgo, L. and Gama, I. Regression Using Classification Algorithms. *Intelligent Data Analysis*, 1997.

[12] Zadeh, L.A. Fuzzy sets and their applications to classification and clustering. In Van Ryzin J. Ed. *Classification and Clustering*. Academic Press, New York,1977, pp. 251-299.