Clustering and Classification Techniques to Assess Aquatic Toxicity

Giuseppina Gini, DEI, Politecnico di Milano, Italy - gini@elet.polimi.it
Emilio Benfenati, Istituto Mario Negri, Milano, Italy
Daniel Boley, CS Dept, University of Minnesota, MN, USA

ABSTRACT

The goal of toxicity prediction is to describe the relationship between chemical properties, on the one hand, and biological and toxicological processes, on the other. Knowledge about the causes of toxicity is incomplete. No single property can satisfy the requirement to model the toxic activity. In the present study we consider different methods to build up models useful for aquatic toxicity prediction. Our study is in the tradition of SAR and QSAR methods, but tries to predict a category. Due to the variability of the toxicity phenomenon, classification methods may present advantages because they refer to intervals of the observed toxic effect. Furthermore classification of compounds according to their toxicity has direct application for regulation of chemicals. In the paper we will report results obtained from the preparation and study of a data set of different classes of chemicals; starting from recursive partitioning algorithms we will test their results against clustering and classifiers.

1. Introduction

The study of the consequences of chemicals on the health of human beings and wildlife is now done through ad hoc experiments, which are very expensive, years long, and involve animal studies. The huge number of compounds to be studied makes this especially challenging. This research requires new and efficient computer-based approaches to analyse huge and complex amounts of information and to automatically discover and use new knowledge implicitly contained in the data.

The goal of toxicity prediction is to describe the relationship between chemical properties, on the one hand, and biological and toxicological processes, on the other. Knowledge about the causes of toxicity is incomplete. No single property can satisfy the requirement to model the toxic activity, with some interesting successful cases, as logP to describe narcosis. Thus, a large number of parameters are of potential interest. The problem is how to deal with this high-dimensional information [2, 3].

Since the sixties, the quantitative structure-activity relationship (QSAR) method has been applied to many drug and chemical design as well as to the prediction of specific toxicological endpoints. Finding structure-activity relationships is essentially a regression process and, historically, linear regression methods have been used to develop QSAR models.

Regression is an "ill-posed" problem in statistics, which sometimes results in QSAR models exhibiting instability when trained with noisy data. In addition, traditional regression techniques often require subjective decisions to be made as to the likely functional relationships between structure and activity.

In the nineties, regression methods based on neurall networks (NN) have been shown to overcome some of these problems as they can account for non-linear structure-activity relationships [10, 12].

A central objective of machine learning research is to develop algorithms that learn predictive relationships from data. This is a central component of data mining and knowledge discovery tasks. However, it is a difficult task, because inferring a predictive function from data is again an "ill-posed" problem; that is, many functions can often "fit" a given finite data set, and yet these functions might generalise very differently on new data drawn from the same distribution.

Several expert systems have been claimed to predict toxicity of chemicals. These ES use different approaches, usually incorporating a knowledge base of explicit rules derived from human experts, or relying on purely statistical approaches [7, 8, 9]. The advantage of rule-based systems is obvious, however it is very hard or impossible to obtain from experts a complete set of rules about toxicity problems not well understood, as in the case of many eco-toxicology problems. For this reason, a system able to reason on data and to extract usable rules would be valuable. This can be obtained using clustering and classification trees [4, 14, 15]..

2. The problem and its representation

2. 1. The chemical knowledge

The common practice in AI, starting with DENDRAL, has been to represent molecules as graphs. Nodes represent atoms, and arcs are the bonds. This view of the chemical structures is too weak, for many reasons. First of all, graphs represent only the planar topology of the molecule and are unable to consider the 3D structure. Second, other information, as the energy information, is lost.

0-7803-6400-7/00/$10.00 ©2000 IEEE

Given the compound structure, that can be graphically entered, there are different ways to compute descriptors that better account for the geometry, physics and properties of the molecule. We preferred to use calculated descriptors because in this case it is not necessary to synthesise the compound.

2.2. The toxicity endpoint

Aquatic toxicity has been addressed within this study. This toxicity has been usually of concern of the pesticide industry, and the available data expresses the effects as measured on plants and animals. Knowledge about how to transpose this knowledge to man or to the environment is out of our purposes.

Structures and toxicity data of pesticides have been obtained from "the Pesticide Manual, eleventh edition" [11]. It contains data on 759 compounds. Missing data is common: for instance, some pesticides have toxicity data on trout, others on duck, others on both.

Lethal concentration for 50% of the animals (LC_{50}) on rainbow trout (*Onchorynkus mykiss*) and daphnia (*Daphnia magna*) were the two most common endpoints in aquatic toxicity. About 200 molecules presented toxicity data for these endpoints. We eliminated pesticides for which toxicity is referred to mixtures of diastereoiomers, because the toxic activity of the individual diastereoisomers is likely different; we kept data referred to a single diastereoisomer when available. We maintained pesticides with one chiral centre, even if mixtures of enantiomers. Polymers have not been considered. A set of 164 pesticides has been finally obtained (Set 1).

We chose to predict LC_{50}, computed as mmol/litre (millimoles per litre).

2. 3. The data sets

Chemical subsets of the compounds have been individuated; in particular in this study we also selected Organophosphate pesticides (OPs), which account for 27 compounds. OPs were the most numerous, however their number was too low. For this reason we collected data from other sources, building a set of 56 OPs (Set 2).

In this way we also studied a limited set (Set 2) of compounds more similar but smaller, and a larger set (Set 1) that is better for generalisation but poses the problem of the much more different toxic mechanisms and chemical classes it contains.

Preliminary molecular modelling has been done using HyperChem 5.0 (Hypercube, Inc, USA) to generate 3-D representations of the compounds. The 3-D structures have been refined with the PM3 Hamiltonian, a semiempirical method for energy minimization. Quantum-chemical descriptors have been calculated using HyperChem. Most of the descriptors have been calculated by CODESSA 2.2.1 (SemiChem, Inc., USA):

- constitutional descriptors, depending on the number and type of atoms, bonds and functional groups, 38 descriptors (18 as discrete values).
- Geometrical descriptors, which give molecular surface area and volume, moments of inertia, shadow area projections and gravitational indices, 12 descriptors.
- Topological descriptors, related to the degree of branching in the compounds, 38 descriptors. For some compounds some descriptors are not applicable (for instance: charge on a given atom, if the molecule does not have this atom).
- Electrostatic descriptors, such as partial atomic charges and others depending on the possibility to form hydrogen bonds, up to 77 descriptors (3 as discrete values).
- Quantum chemica: descriptors, related to the molecular orbitals and their properties.
- LogD, the apparent partition coefficient, (calculated with Pallas 2.1 by CompuDrug, Hungary). We selected the values at pH 3, 5, 7.4 and 9. These descriptors are the expression of the lipophilicity of the molecule at various pH.

A total set of about 170 descriptors has been built.

3. Clustering

Clustering has been done by Divisive Partitioning Principal. Direction. Divisive Partitioning (PDDP) [4] is an unsupervised clustering procedure which constructs a hierarchical classification tree top-down. Representing each pesticide by a vector of chemical descriptor values, this method splits the entire collection along the direction of maximal variance ("principal direction"), and then recursively splits each subpart along the local direction of maximal variance. This process builds a tree structure top-down with the root representing the entire collection. This method operates on the numerical values themselves and predicts numerical toxicity values, unlike CART [6] which predicts only a discrete class of toxicity out of a small collection of classes. To test PDDP as a predictor, we used a leave-one-out approach, in which one attempts to predict the toxicity of each pesticide based on the actual toxicity values for all the remaining pesticides. The PDDP splitting process was carried out until clusters of at most 5 entries were obtained, and the toxicity value predicted for each pesticide was defined to be the average of the toxicities for the other pesticides in the same cluster. The process takes only 2.5 sec on a Sun Sparc workstation using a

Matlab implementation available from the University of Minnesota [17].

We used the LC50 values for Trout, which on a logarithmic scale range from -5.23 to +1.11. Of the 164 pesticides tested in this example (Set 1), the predicted toxicity values for 100 (61%) were within 1 unit of the respective true values on this scale, while those for 45 (27%) differed by more than 1 but less than 2 units, and those for 19 (12%) differed by more than 2 units on this scale. The distribution of the actual numerical predicted values is illustrated in Fig 3.1 in which the dashed dotted lines enclose the regions where the predicted and actual values differ by at most 1 and 2 units, respectively, on this scale.

To compare the results with the CART method, we also divided the scale into 3 classes of equal length and counted the entries in each part: Class 1 from -5.23 to -3.12, in the logarithmic scale range, Class 2 from -3.12 to –1.00, Class 3 from -1.00 to +1.11. These classes are represented in Fig 3.1 by the vertical and horizontal lines. The class results are given in Table 3.1 below, showing an error rate of 37.2%. The correct values are on the diagonal.

Table 3.1: the predicted class obtained with PDDP using LC50 Trout values.

		Predicted class		
Real		**1**	**2**	**3**
	1	9	18	0
	2	13	72	22
	3	0	23	22

From Fig 3.1, it is seen that several pesticides were placed in the wrong class, but the predicted value differed very little from the true value. Thus using a few discrete classes (as in CART) instead of numerical values on a continuous scale (as in PDDP) gives a much cruder analysis of the performance of the method, as well as cruder predictions.

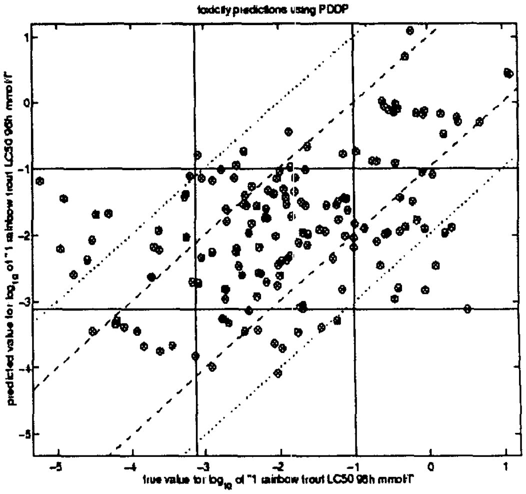

Fig 3.1. Diagram of numerical predicted (y) vs true values (x), showing the region where the two values differ by less than 1 (dashed line) and 2 (dotted line) units. The classes of Table 3.1 are delimited by the vertical and horizontal lines.

4. Classification

Two classifiers have been used to predict the toxicity class: CART [6] and Bayda [1].

4.1. CART

A classification tree is an empirical rule for predicting the class of an object from values of predictor variables [13, 16]. Common features of classification tree methods are

- Merging: relative to the target variable, non-significant predictor categories are grouped with the significant categories.
- Splitting: a variable to split population is chosen by comparison to all others. The method recursively splits nodes until a stopping rule is triggered
- Stopping: rules to determine how far to extend the splitting of nodes.
- Pruning: branches that add little to the predictive value of the tree are removed.
- Validation and error estimation: measurement of true error vs. apparent error, and validation using separate or resampled data are performed identically.

After a tree has been built, two verification methods have been used: partitioning and cross-validation (leave-one-out). This method uses all but one of the data to build the tree; the risk estimate is computed by partitioning the data into k separate groups or folds (where k =1). Next, k trees are built using the same growing criteria as the tree being evaluated. The first tree uses all folds except the first, the second tree uses all folds except the second, and so on, until each fold has been excluded once. Fore each of these trees, a risk estimate is computed (the proportion of all cases incorrectly classified.), and the cross-validated risk estimate is the average of these k risk estimates for the k trees, weighted by number of cases in each fold.

CART (Classification and Regression Trees) is a non-parametric classification method that constructs a binary decision tree. The high dimensional space of the objects in the training set is divided into subspaces such that each subspace can be associated with a single class. CART classification rule has a tree form that is easy to interpret, yet it takes into account the fact that different relationships may hold among variables in different parts of the data. CART does automatic stepwise variable selection. It performs well when the pattern space can be separated into pure class subspaces by few hyperplanes perpendicular to variable axes.

The optimal tree is the one that has the minimum cross-validated risk. For each non-terminal node of the optimal tree, are displayed:

- the risk associated with the node;
- the variable and its threshold value where the objects are split to form the left and the right child nodes;

For each terminal node, are shown:

- the number of objects in the node;
- the probability of the node;
- the class associated with the node (the class that contains most of the objects in the node);
- the risk associated with the node;
- the pureness of the node

The pureness is reported as the number of objects in each class and the class probabilities. In an ideal case, only objects from a single class are in a terminal node, i.e. one of the classes has probability 1.0 and the rest have probability 0.0.

The misclassification matrix, calculated without and with cross-validation, has rows corresponding to the true classes, and columns corresponding to the assigned classes (calculated with and without cross-validation). In a perfect classification, all the off-diagonal elements of the misclassification matrix are zero.

The cross-validated error rate and cross-validated risk are measures of the goodness of class prediction. If the default priors and loss function are used, the error rate equals the risk.

The classification analysis has been performed with SCAN (Minitab Inc., USA) on both the full set and the OPs.

4.1.1. Analysis of the full set

For the set of the 164 pesticides, we defined three toxicity classes, with toxicity from 1 to 0.66, from 0.66 to 0.33 and from 0.33 to 0, in the normalised logarithmic scale. In this way the population of each class is similar. The inputs were all descriptors. We obtained an error in validation of 26.3%. Table 4.1. shows the real class and the class predicted in validation (leave-one-out).

Table 4.1. The predicted classes for the 164 pesticides

		Predicted class		
		1	2	3
Real	1	16	13	0
	2	11	73	9
	3	0	10	32

4.1.2. Analysis of the OPs

In a similar way we used CART with the OPs subset. We used only 21 descriptors, selected in a previous study [9], and defined four classes:

- with toxicity values (antilog of LC50 for trout, scaled between -1 and 1) between -1 and --0.5;
- with toxicity values between -0.5 and 0;
- with toxicity values between 0 and 0.5
- with toxicity values between 0.5 and 1.

The four classes are quite balanced (6 elements for the first class, 13 for the second one, 15 for the third one, 9 for the fourth). Table 4.2. shows the results for the 43 molecules assigned in validation using leave-one-out. The Error Rate was low: 0.12.

Table 4.2. The predicted class for the 56 OPs, using leave-one-out method.

		Predicted			
		1	2	3	4
Real	1	4	2	0	0
	2	0	9	2	2
	3	1	1	9	4
	4	0	0	5	4

However, this classification model used to predict the validation set of 13 molecules gave an error rate of 0.38, as in Table 4.3.

Table 4.3. The predicted class for the 56 OPs, using the external validation set.

		Predicted			
		1	2	3	4
Real	1	0	0	0	0
	2	0	1	2	0
	3	0	0	5	0
	4	0	2	1	2

The class 3 (the most represented in the training set) has been correctly assigned. There are two mistakes for the second class and three for the fourth class. In the test set there were not first class molecules and the model correctly assigned no one object in this class.

The descriptors selected by CART in this case are different from those selected in the previous model built with 27 compounds.

The classification tree for trout, using the OPs, is illustrated in Figure 4.1. We can see that for a class there may be more than one leaf.

The chemical descriptors used in the tree may be useful to have information on the molecular features involved in the toxic mechanism. For instance, in the tree illustrated in Figure 4.1 some descriptors are topological, as "Average2", "Kier sh1", "Randic 2". They give information on atomic connectivity in the molecule. Other descriptors are constitutional, such as "Number O", the number of oxygen atoms. Others are electrostatic, such as "HA depen" and "PNSA-1P", and reflect characteristics of the charge distribution of the molecule. Finally, some descriptors are geometric, referring to the moment of inertia "Moment o", or to the molecular surface area "Molecula".

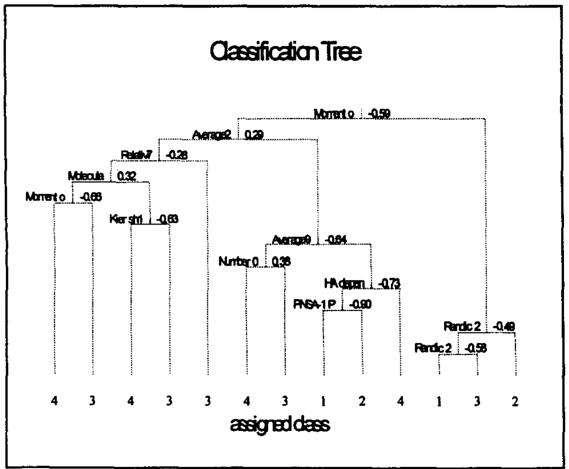

Figure 4. 1. A tree obtained with CART

However, it is not easy to obtain simple and stable rules. Similar models can be obtained with CART starting from different selected descriptors, and even if the trees perform equally well, the descriptors and the critical values for the nodes may change. For this reason it may be useful to perform several attempts, in order to find confirmation of important descriptors.

4. 2. Bayesian Classifier

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 in the network. An arc between two nodes denotes the existence of a direct probabilistic dependency between the two variables. The lack of an arc between two nodes implies that no direct probabilistic influence exists between those variables.

A Bayesian Network classifier is a Bayesian Network applied to a classification problem. It contains a node C for the class variable and a node X for each of the domain feature. Each node has a finite number of states: the class node has a state for each class, the feature nodes have a discrete set of values.

Given an instance vector x, a Bayesian network allows to compute the probability $P(C = c_k \mid X = x)$ for each possible class c_k. If the true distribution $P(C \mid X)$ is known, we achieve the optimal classification by selecting the class c_k for which the probability is maximised. Unfortunately the true distribution is not available and can only be approximated from the training set.

The simplest such classifier is the Naïve Bayes Classifier, which makes the strong assumption that all the features are conditionally independent from one another. In this case the features nodes are all directly connected to the class node. The initial probability is computed from the Mutual Available Information from the training set. Several approaches have been made to improve the classification abilities of Bayesian networks, but the introduction of interaction between feature variables makes the problem of inducing an optimal classifier NP-hard.

Bayda 1.31, from the University of Helsinki, is a Java implementation of a Bayesian predictive discriminant analysis based on a Naïve Bayes model build from the data set. It supports continuous-valued nodes, missing data handling, forward/backward variable selection, and external leave-one-out cross-validation.

The experiments with BAYDA are based on class subdivision studied to make similar the cardinality of the classes in the training set.

4.2.1. Rats

Taking the normalised values for rats, we defined four classes:

Class 1, toxicity 0 through 0.02 (39 elements),
Class 2, toxicity 0.02 through 0.13 (39),
Class 3, toxicity 0.13 through 0.45 (44)
Class 4, toxicity 0.45 through 1 (42).

Accuracy: **43.9%** of the classifications are correct, as in Table 4.4.

Table 4.4. The predicted class for rats

	55%	33%	21%	58%	Success
class	1	2	3	4	
	53	45	28	38	# predicted

4.2.2. Daphnia

Taking the normalised values for daphnia, we defined four classes:

Class 1, toxicity 0 through 0.255 (40 elements),
Class 2, toxicity 0.255 through 0.39 (41),
Class 3, toxicity 0.39 through 0.58 (41)
Class 4, toxicity 0.58 through 1 (42).

Accuracy: **45.7%** of the classifications are correct, as in Table 4.5.

Table 4.5. The predicted class forDaphnia.

	55%	45%	29%	54%	Success
class	1	2	3	4	
	38	44	41	41	# predicted

Results were poor compared to those obtained with CART. We may only comment on those results concluding that a naïve Bayes classifier is unable to capture the real structure of the data.

5. Conclusions

The prediction of toxicity using advanced models is a topic that deserves study for the possible advantages offered by these models, compared to laboratory experimental methods. However, the matter of prediction is a complex one. The target of these models, the definition of the inputs, the software to be used, the way to assess the results, are matter of discussion. In this sense it is highly useful to have studies comparing different approaches on more data.

In the present study we compare methods to predict categories of toxic effects. On a general point of view it is preferable to have methods offering a better resolution of the predicted value. However, the predicted value should always be evaluated considering that the experimental values used for training (resulting from laboratory experiments on animals) are affected by a variability, which may be quite high, due to animal variability and to variability for the experimental procedure. This is the first reason for considering methods that provide categorical values. The second one is that the obtained classification can help for a first screening of toxic properties of chemicals, according to regulatory schemes that classify chemicals.

Most of the published studies on modelling of properties of chemical compounds (SAR and QSAR) use the leave-one-out validation method, or, in several cases, only show the values in fitting, without any validation. In the present study we show results both with leave-one-out and validation set. We can observe that the leave-one-out method is quite optimistic, and the only way to adapt to a large number of new individuals is to re-train the network for updating the model. The real problem is how representative can be a model built on the basis of a limited population. Even if the results appear very good according to the leave-one-out method, a dramatic reduction of the prediction appears using a larger set of compounds for validation. This problem is probably not due to over-fitting, but to the high discontinuity of the function to be learned.

However, the hope that a symbolic description could help in understanding the rationale of the toxic activity results is now fable. The classification trees obtained from CART are so variable in their structure to leave a little space for automatic scientific discovery.

No single approach is likely enough to solve the problem. The knowledge and the simulation of the interaction between the chemical compound and the cell is still the best way to asses the mechanism of the toxic effect, but a QSAR study, as in our experiment, has valuable results for a first screening.

Besides that, in our classification experiments CART gave good results; it may be worthwhile to mix different classifiers obtained through CART and other approaches.

Acknowledgements

We acknowledge the financial contribution of the European Commission, project COMET, ENV4-CT97-0508 and, partially, of NATO (CRG 971505), and NSF grant IIS-9811229.

References

1. BAYDA software, http://www.cs.Helsinki.FI/research/cosco.Projects/
2. E. Benfenati, P. Grasso, S. Pelagatti, and G. Gini, “On Variables and Variability in Predictive Toxicology”, IV Girona Seminar on Molecular Similarity, Girona, Spain, July 5-7, 1999.
3. E. Benfenati, and G. Gini, “Computational predictive programs (expert systems) in toxicology”, *Toxicology,* 119:213-225, 1997.
4. M. Berthold, and D. J. Hand, “Intelligent Data Analysis – An introduction”, Springer, Berlin, 1999.
5. D. L. Boley: “Principal Direction Divisive Partitioning”, Data Mining and Knowledge Discovery, vol 2 #4, p 325-344, Dec. 1998.

6. L. Breiman et al., "Classification and Regression Trees (CART)", Wadsworth & Brooks, 1984.
7. R.D. Combes, and P. Judson, "The use of artificial intelligence systems for predicting toxicity", *Pestic. Sci.*, 45:179-194, 1995.
8. J.C. Dearden, M.D. Barratt, R. Benigni, et al., "The development and validation of expert systems for predicting toxicity", *ATLA*, 25:223-252, 1997.
9. G. Gini, and A. Katritzky, (Eds.) "Predictive Toxicology of Chemicals: Experiences and Impact of Artificial Intelligence Tools", Proc. AAAI Spring Symposium on Predictive Toxicology, Report SS-99-01, AAAI Press, Menlo Park, California, 1999.
10. G. Gini, E. Benfenati, P. Grasso, M. Lorenzini, and A. Vittore. "Some results for the prediction of carcinogenicity using hybrid systems", Proc. AAAI Spring Symposium on Predictive Toxicology, Report SS-99-01, AAAI Press, Menlo Park, California, 1999, pp 138-143.
11. The Pesticide Manual, Eleventh Edition, British Crop Protection Council, Berks (UK), 1997.
12. G. Gini, M. Lorenzini, E. Benfenati, and P. Grasso, "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, pp. 1076-1080, 1999.
13. J.S.U. Hjorth, (1994), "Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap", London: Chapman & Hall.
14. D. J. Hand, "Construction and Assessment of Classification Rules", Wiley, 1997.
15. D. Michie, D.J. Spiegelhalter, and C.C. Taylor (eds.), "Machine Learning, Neural and Statistical Classification", Ellis Horwood, 1994.
16. J. Mingers, '"An Empirical Comparison of Pruning Methods for Decision Tree Induction", Machine Learning, 4, 227-243 1989.
17. Principal Direction Divisive Partitioning Software http://www.cs.umn.edu/~boley/PDDP.html