International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 18, No. 5 (2004) 801–817

World Scientific
www.worldscientific.com

MULTICLASS CLASSIFIER FROM A COMBINATION OF LOCAL EXPERTS: TOWARD DISTRIBUTED COMPUTATION FOR REAL-PROBLEM CLASSIFIERS

CHRISTOPH KÖNIG*, GIUSEPPINA GINI† and MARIAN CRACIUN‡

Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo Da Vinci, 32, 20133, Milano, Italy
**chr_koenig@gmx.de*
†*gini@elet.polimi.it*
‡*mcraciun@ugal.ro*

EMILIO BENFENATI

Laboratory of Environmental Chemistry and Toxicology,
Istituto di Ricerche Farmacologiche "Mario Negri",
Via Eritrea, 62, 20157, Milano, Italy
benfenati@marionegri.it

In many real-world applications simple classifiers are too weak to have predictive power. Ensemble techniques, or mixture of experts, are a possible solution. We illustrate why mixture of experts are a natural choice in domains such as the prediction of environmental toxicity for chemicals, when a structural approach is pursued. The real data here used are derived from peer reviewed experiments, and are publicly available, but are difficult to model. We used them to predict aquatic toxicity for fish. Chemical information was coded into a set of about 160 descriptors; after reducing the dimensions of the feature vector through different techniques, we developed multivariate regression to build a model of the toxic effects of chemicals. Defining toxicity as a category, as in European Union (EU) regulations, we extended the study to predict toxicity class. Problems with poor predictive power of this simple approach have led us to reconsider the problem from a more theoretical angle. We have respected locality criterion to build different local classifiers, one for each chemical class, to achieve better results. Then we combined the classifiers to get a complete system to predict any chemical for the chemical classes studied.

Keywords: Mixture of experts; classification from regression; QSAR.

1. Introduction

Research in the past decade has shown that classification and regression problem ensembles are often much more accurate than the individual base learners that form them.[21] There are different ways to use several classifiers in a recognition problem, at least two main streams derived from the introduction of "ensembling" highly correct classifiers that disagree as much as possible, and "mixtures of experts",[17,19]

built on the idea of training individual networks on a subtask, and then combining these predictions with a gating function that depends on the input. In this paper we shall use the term combination of experts, in order to include all the aspects of integrating local experts.

While the combination of individual classifiers is still a matter of theoretical discussion,[16,30] sometimes the application domain itself can address specific ensemble solutions, as shown in the following. The application domain in which our research originated is the prediction of ecotoxicity for molecules,[12] a domain rapidly growing from chemometrics, and data mining. In traditional chemometrics and life sciences, regression models[10] are developed and evaluated and the predictive value is assessed through statistics. For regulatory purposes classifiers are a better solution, such that ecotoxicity authorities can choose a class label rather than a real number. The predictive assessment of such systems requires a more extensive view, to account for specificity and sensitivity of the classifiers, besides attaining a high confidence level.[10] This domain is difficult, the main reason being the need for selected, compatible data, avoiding arbitrary generalizations and the variability of biological data that can easily influence the output.

Of the many methods of combining classifiers, we chose the supervised learning paradigm; we clustered the data in a supervised way into different chemical classes, defining for each partition a set of training examples labeled with an output class tag, then we trained the individual classifiers and combined these predictions with a "gating" function, a classifier that learns how to allocate examples to the experts. Each expert is a model of a region of the input space, and the gating function has to decide from which model the data point originates.

In this paper we explain how to extend classical QSAR approaches to the prediction of environmental toxicity for chemical products through combinations of classifiers, and provide results on real data.

In Sec. 2, we illustrate some specific problems of toxicity prediction. We choose the EPA data set for toxicity on fish, and discuss its properties. We derive a large feature space from the chemical structures; chemical information is coded into a set of about 160 descriptors.

In Sec. 3, we illustrate the different stages of model building; having reduced the dimensions of the feature vector via different techniques, we developed multivariate regression to build a model of the toxic effects of chemicals. Defining toxicity as a category, according to EU regulations, we extended the study to predict toxicity class. Problems with the poor predictive power of this simple approach obliged us to reconsider the problem, with a stricter definition of the QSAR. We then built different local classifiers, one for each chemical class, that give better results, and combined them all together to produce a complete system to predict any chemical from the chemical classes studied.

In Sec. 4, we develop our combination scheme to accommodate different classifier experts for different parts of the domain, so as in principle to take advantage of distributed computation moving to very large data sets. We discuss the overall results, and conclude with our final considerations, in Sec. 5.

2. The Problem of Toxicity Prediction

2.1. *Chemometrics*

Chemometrics, the production and use of chemical information, is an area where pattern recognition techniques are extensively used. Chemometrics encompasses the basic steps of:

- Data analysis — Extracting information from chemical data.
- Experimental design — Yielding information from chemical data.
- Modeling — Investigating complicated relationships.

The basic Chemometrics strategies evolved from statistical experimental design, which gives the range of ways to generate a set of examples, reduce the range of attribute dimensions and transform data to simplify the response function, by linearizing, stabilizing the variance and making the distribution more normal.

One of the most active areas in chemometrics is QSAR (Quantitative Structure-Activity Relationships),[15] developed in the last 40 years to assess the value of drugs, and now proposed as a method to assess general toxicity, as well as a way to obtain new knowledge from data. QSARs can be based either on regression[10] or classification[28]: for drug activity and toxicity to a given target, most QSAR models are regressions, mainly referring to the dose with toxic effect in 50% of the animals. Classification systems for QSAR or SAR (Structure-Activity Relationships) refer to regulatory bodies (NTP, EPA, IARC, etc.), that aim to use predictive methods for priority setting and for risk assessment. The correct modeling of QSAR derives from "postulates" as defined from evidence and theory, and is expressed as follows:

- The molecular structure is responsible for all the activities shown.
- Similar compounds have similar biological and physico-chemical properties.
- Congenericity: QSAR is applicable only to similar compounds.

From this definition of QSAR it is evident that the localness of the model must be preserved, and generalization requires attention. For toxicity prediction, considering the small amount of experimental data and the huge number of compounds, a way to maintain localness is to employ an ontological approach and divide the compounds into homogeneous sets.

Besides the ontological approach that can produce different classifications of chemicals, there is also a representation problem. Many molecular representations have been proposed, claiming to explain the properties of the molecule better (quantum similarity, spectral properties, descriptors, etc.), but no general conclusions can be reached.

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. Scientifically, predictive (eco)toxicology is a recent area. Knowledge about the causes of toxicity is unavailable, though there are some interesting cases, such as logP to describe narcosis.[6] Thus, a large number of features (in the hundreds)

should be tested. Classically QSAR models are (multi)linear equations; however some nonlinear approaches have also been used.[20]

In the present study we build models for predicting aquatic toxicity, both as QSAR models, which predict a continuous value,[15] and as classification methods, for toxic effect intervals — more directly applicable for regulation of chemicals. We show why they implicitly require *ensembling*, because QSAR models are local models. Moreover, we need to ensemble classifiers to improve the results of simple PLS methods, that have been applied in our laboratories[31] to the same data set, obtaining R^2 (cross-validated with leave-one-out) lower than 70%.

2.2. *Data set and molecular descriptors*

The US Environmental Protection Agency (EPA) studied toxicity in the fathead minnow (*Pimephales promelas*) using a series of industrially used organic compounds.[11,32] The measure for acute toxicity is LC50 (96h), i.e. the lethal concentration for 50% of a population within 96 hours.

The data set we built, called IMAGETOX fm, contains 568 different compounds for which the toxicity value was taken from EPA and a large number of features (chemical descriptors) were calculated at the Mario Negri Institute. The data is quite representative for most industrial chemicals, but they are still a very small percentage of the commercialized chemicals, and an even minor part of all possible chemicals humans can be exposed to. Nevertheless, they represent a unique collection of data, because experiments have been conducted according to a well-defined protocol, many observations have been collected to produce the information expressed as LC50, and many years of work and resources have been dedicated to this task. As typical in data from studies in the life sciences, the cost of experimental data is high. Thus, the number of experiments is quite low, as in other cases of life sciences in which classification methods have been used.[1] What is more convenient in our case is to produce many calculated values for the used chemicals, which are chemical descriptors.

To compute the descriptors, preliminary molecular modeling was done using HyperChem 5.0 (Hypercube Inc., Gainsville, Florida, USA) to generate 3D representations. These were then refined with the PM3 Hamiltonian, a semi-empirical method for energy minimization of the geometry. Accurate 3D representations of structures were necessary to generate descriptors dependent on geometry.

Most of the descriptors were calculated by CODESSA 2.2.1 (SemiChem Inc., Shawnee, Kansas, USA). Quantum-chemical descriptors, i.e. total energy of the molecule, HOMO and LUMO energies, ionization potentials, heat of formation, etc., were calculated using MOPAC (with the PM3 Hamiltonian). A class of descriptors largely used for QSAR studies along with the apparent partition coefficient (logD), was calculated by Pallas 2.1 (CompuDrug, Budapest, Hungary). These physico-chemical descriptors are the expression of lipophilicity of the molecule at various pH. We selected pH values of 3, 5, 6.5, 7, 7.4 and 9.

Table 1. Statistical information about the data set (toxicity value LC50 in mg/liter).

Parameter	Value
Maximum	75200.00
Minimum	0.00019
Geometrical mean	24.1313
Arithmetic average	1.0600e+003

Table 2. EU classification for fish (directive 92/32/EEC annex VI point 5.1).

Class	LC50 96h	Damage to the Environment
I	< 1 mg/L	Very toxic to aquatic organisms
II	1–10 mg/L	Toxic to aquatic organisms
III	10–100 mg/L	Harmful to aquatic organisms
IV	> 100 mg/L	May cause long-term adverse effects in the aquatic environment

After removing descriptors with missing and constant values, the entire data set consist of 156 descriptors. They can be split into six categories: constitutional descriptors (38), geometrical descriptors (12), topological descriptors (36), electrostatic descriptors (57), quantum-chemical descriptors (6), physico-chemical descriptors (7).

We used log (1/LC50) to predict continuous values of acute toxicity. Because these values were widely spread (see Table 1) and to take account of regulations, the results were also transformed into the classification for toxicity to fish provided by Directive 92/32/EEC of the EU for dangerous substances (Table 2). For a discussion on toxicity classes for this data set, see Ref. 25.

The apparently simple classification of the output may hide the real scientific problem. Toxicity is a very complex phenomenon, involving many different and competing biochemical processes, which take place in different parts of the organism. Death, which is the endpoint considered in the data set, can be due to different causes. Also the chemical compound producing the toxic effect actually interacts with many biomolecules, and often undergoes metabolism, which generates new chemicals.

2.3. *Combination strategies*

Classifier combination strategies may reflect the local competence of individual experts as used in the mixture of experts paradigm. Beside simple averaging,[3] the output classifier can be trained separately using the outputs of the input classifiers as new features, as proposed by Ref. 24 developed in Ref. 27 and already applied to a similar problem in Ref. 5.

A strong point that makes combinations of classifiers attractive in QSAR, besides improved results, is the fact that they can be distributed both in time and space. Partitions of the data are distributed to different processors, each applies a learning algorithm to each subset, and then the learned results yield a single classifier.[7] The reduction in execution time results from the distribution of the expensive learning step to multiple processors, a research area we are currently actively pursuing.

In this paper, we focused on classifier combinations in this scenario. After studying the data set, and trying a simple classifier, we divided the problem domain into subsets (the chemical classes) and worked on them separately. We developed a set of disjointed simple classifiers, and combined them by selection of the most appropriate classifier.

3. Experiments and Results

3.1. *One monolithic system*

We used the WEKA data-mining workbench[34] created by the Department of Computer Science at the University of Waikato, New Zealand, for modeling. It comprises a wide range of data mining algorithms for regression, classification and clustering, as well as tools for preprocessing, evaluation and visualization of the data and results.

Initially we constructed a single linear model to predict the toxicity value, using all descriptors and the entire data set of 568 chemical compounds. We applied linear regression with a ten-fold cross validation (cv) in all our tests. This validation approach was discussed in related cases by Ambroise and McLachlan.[1] Results are listed in Table 3 and illustrated in Fig. 1, which summarizes the values obtained with the ten-fold cv models. Both the low R^2-value and the comparatively high error measures indicate that the model does not predict toxicity with satisfactory accuracy. This is mirrored again in the dispersion diagram (Fig. 1) with its wide spread. After regression the transformed results in classes of toxicity are not good (Table 4). The same poor results have been obtained using discriminant analysis. With so many variables there are often several that do not contribute to the final result, thereby reducing the prediction accuracy of a model. Furthermore with an increasing number of influencing factors the result becomes difficult to analyze and interpret and the risk of an overfitted model arises.[37]

Table 3. Accuracy of prediction of log (1/LC50) using a single model with all descriptors.

Model	Descriptors	Evaluation Parameter	Value
Linear regression	156	CORR	0.740
Ten-fold cross validation		R^2	0.548
		MAE (Mean Absolute Error)	0.643
		MSE (Mean Squared Error)	0.956

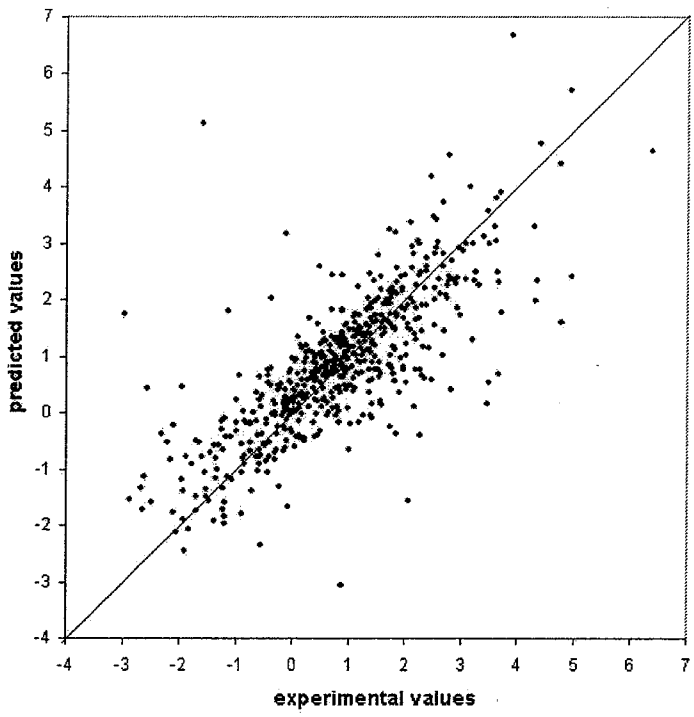


Fig. 1. Prediction dispersion of a single model with all descriptors; toxicity value is log(1/LC50).

Table 4. Accuracy of classification-from-regression into four toxicity classes using a single model and all descriptors.

	Number of Compounds	Percentage
Instances classified correctly	345	60.74%
Instances classified incorrectly	223	39.26%
Total	568	

Therefore, we decided to reduce the number of descriptors to obtain a more robust and accurate model. We used different selection algorithms of WEKA for the reduction,[35] i.e. the CFS-algorithm,[14] based on the correlation between attributes and the goal to be predicted, the ReliefF-algorithm[23] that addresses the discrimination ability of an attribute and the wrapper-routine,[22] which uses results of a learning method to select attributes. The latter gave the best result, selecting 20 of the descriptors distributed in all categories. Again, linear regression was performed.

The evaluation parameters were better than those obtained without reduction of the descriptors (Table 5) and the dispersion diagram shows that the points are closer to the ideal prediction (Fig. 2), even though an anisotropy with respect to the diagonal appears. However after transformation into classes of toxicity, the accuracy was still poor (see Table 6).

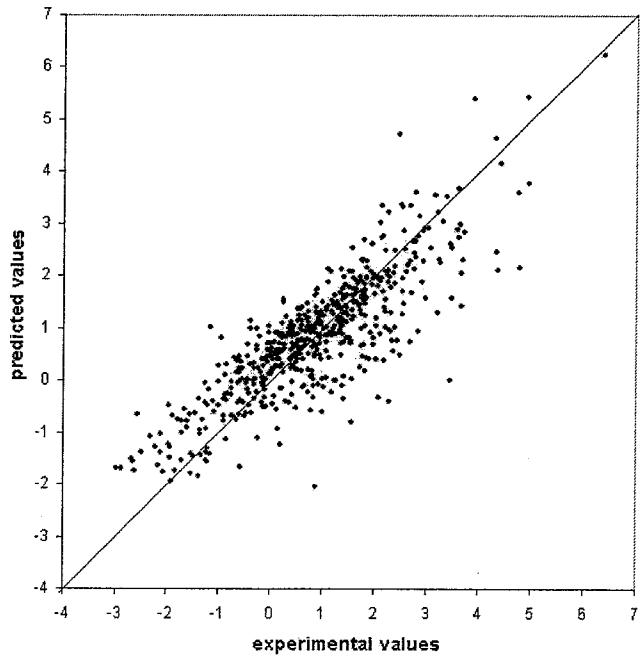


Fig. 2. Prediction dispersion of a single model with wrapper selection; toxicity value is log(1/LC50).

Table 5. Accuracy of the prediction of log (1/LC50) of a single linear regression model with wrapper selection.

Model	Descriptors	Evaluation Parameter	Value
Linear regression	20	CORR	0.838
Ten-fold cross validation		R^2	0.701
		MAE	0.565
		MSE	0.571

Table 6. Accuracy of classification into toxicity classes of a single linear regression model with wrapper selection.

	Number of Compounds	Percentage
Instances classified correctly	337	59.3 %
Instances classified incorrectly	231	40.7 %
Total	568	

3.2. *Models of chemical classes*

The data set contains a lot of completely different compounds, which are toxic in some way but structurally diverse. It is likely that no single model is able to yield good results for them all. Dividing the set into smaller groups is one way of dealing

Table 7. Subsets of chemical classes.

Chemical Classes in the Subset	EPA Classification Code	Number of Compounds
Hydrocarbons	2.0, 2.1	26
Ethers	3.0, 3.1, 3.3	24
Alcohols	4.0, 4.1, 4.2, 4.3	60
Aldehydes	5.0	44
Ketones	6.0, 6.1, 6.2	39
Acids	7.0, 8.0, 8.1, 8.2, 8.3	68
Nitriles, Sulfur Compounds	9.0, 12.1, 12.2, 12.3	33
Amines	10.0, 10.1, 10.2, 10.3, 10.4, 10.5	74
Benzenes	13.0, 13.1	33
Phenols	14.0, 14.1	49
Heterocyclics	15.0, 15.2, 15.3, 15.4, 15.5, 15.6	48
Carbamates, other pesticides	21.0, 22.0	28
Various classes (pasted)	1.0, 1.1, 11.1, 16.0, 17.0, 18.0, 19.0, 20.0, 23.0, 23.1, 24.0	42

Table 8. Linear regression results for 13 subsets of chemical classes.

Subset	Number of Descriptors	R^2	Mean Absolute Error (MAE)	Mean Squared Error (MSE)
Hydrocarbons	6	0.805	0.309	0.201
Ethers	12	0.986	0.130	0.031
Alcohols	7	0.721	0.551	0.555
Aldehydes	9	0.609	0.385	0.241
Ketones	2	0.734	0.476	0.497
Acids	10	0.840	0.342	0.183
Nitriles, Sulfur Compounds	4	0.812	0.364	0.281
Amines	16	0.932	0.255	0.105
Benzenes	5	0.820	0.310	0.161
Phenols	6	0.798	0.329	0.189
Heterocyclics	7	0.757	0.469	0.395
Carbamates, other pesticides	4	0.785	0.643	0.627
Various classes	3	0.732	0.562	0.509
Weighted mean (weighted by numbers of compounds)			0.395	0.302

with this diversity. Therefore we split the data set into 13 groups according to the EPA's chemical classification. Such groups contained between 24 and 74 compounds (Table 7). All groups apart from one contain only one or two chemical classes and a number of subclasses to keep the similarity of the compounds within a set. There remained a few classes with a very small number of compounds (fewer than ten). A model built up on such a small base is not reasonable, so these classes were merged. For each subset the number of descriptors was reduced using the wrapper method as already applied to the complete data set. Then a linear regression was performed on each set to achieve 13 models, each able to predict the toxicity of the corresponding data set. The ten-fold cross-validation results are summarized in Table 8. The results vary widely with the subset. Some models predicted the

toxicity values very well of the area of chemical compounds: these are the models for ethers and amines, the latter being the largest group. Other models gave less satisfactory results. The subset of aldehydes gave the worst results. But most models had lower errors with respect to the single model obtained previously. Even though the average selected descriptors were lower than for the single model, the models were able to predict the toxicity with greater accuracy. Thus a combination of these experts improved the result for the entire set.

3.3. *Combination of the chemical class models*

As mentioned in Sec. 2.3, the most common methods for combining results are averaging (with or without weighting) or voting; another method that uses the output of a learning algorithm as input is "stacking".[36] But all these combinations are "ensemble" methods, which means that they combine outputs redundantly. In our case such linear combinations are not very useful because we would lose the local influence of our experts, built and adapted on strictly separate areas of the data space. This modularity implies that in further studies we can improve these experts and easily exchange or supplement them.

To maintain this advantage we decided to use a competitive strategy,[33] thus selecting the appropriate expert for each instance. This selection benefits from the fact that we used the chemical classification to separate the data space. The classification strongly depends on the structure of a chemical compound, i.e. the existence of specific atoms or functional groups within a molecule. The structure is expressed mainly by constitutional descriptors (see 2.2) that already exist in the data set. Since the other descriptors in the set depend on the constitutional ones as well, they can also improve this classification.

This is an original approach, because the classifiers are used not to predict toxicity classes, but to identify the most fruitful models, on the basis of the automatic selection of chemical classes. Chemical classes can be defined by human experts, and on the same data set we used separate local models obtained by splitting the data set into in chemical classes.[31] Another study[2] used discriminant analysis in order to classify toxic mechanisms of action of phenols, which are a subset of our data set. The novelty of the present approach is that we need neither human knowledge for the selection of chemical classes nor experimental data on the toxic mode of action. The classification done by human experts can be difficult because quite frequently more than one functional group is present in the same structure. In this case, the same chemical can belong to more than one chemical class. Our approach is directed to an automatic classification of chemicals into nominal classes, using chemical descriptors. Then, specific toxicity models can be used, based on these nominal classification.

In order to develop the classifier, we trained different algorithms included in WEKA. The best results were obtained by applying a meta classifier scheme that was able to handle multiclass data sets with two-class classifiers. It was applied to

Table 9. Accuracy (ten-fold cross-validated) of classifying compounds in 13 chemical subsets.

	Number of Compounds	Percentage
Instances classified correctly	485	85.4%
Instances classified incorrectly	83	14.6%
Total	568	

Table 10. Accuracy of prediction of log (1/LC50) of the combined model.

Model	Descriptors	Evaluation Parameter	Value
Linear regression	20	CORR	0.896
Ten-fold cross validation		R^2	0.802
		MAE	0.417
		MSE	0.390

Table 11. Accuracy of classification into toxicity classes of the combined model.

	Number of Compounds	Percentage
Instances classified correctly	409	72.0%
Instances classified incorrectly	159	28.0%
Total	568	

the J48-algorithm, which implements the C4.5 algorithm of Quinlan,[29] using the error correction code to improve the accuracy. The results are listed in Table 9. Differentiating between all subsets showed that there were just three of 13 subsets that presented low accuracy for the classification into chemical classes, one being the mixed subset with chemical classes of less than 10 compounds. The remaining subsets were predicted very well, most with accuracy above 90%. With these limitations, the classifier was able to discriminate between the subsets.

The output of this classifier was then used to select the appropriate toxicity model for each compound in the data set. After combining the different sub-models, the results improved considerably. R^2 typically increased an order of magnitude and the error values fell by nearly 30% compared with the best result of one single linear model as used before (Table 10 as opposed to Tables 3 and 5). This time the classification in toxicity classes was also better. About 12% more instances were classified correctly, which amounts to more than 70 chemical compounds out of 568 (see Table 11).

3.4. *Prediction of an external set*

We compared the results obtained by ten-fold cross-validation by predicting new, unseen data. Thus we split the data set in the ratio of 80:20 into a training set (456 cases) and a test set (112 cases), according to the distribution in chemical and toxicological classes, respected even for the smallest sets.

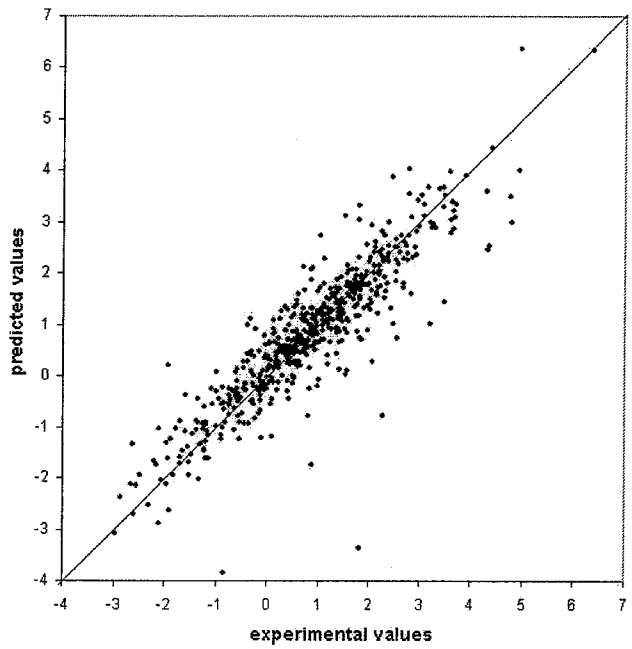


Fig. 3. Prediction dispersion of a combined model of 13 chemical classes; toxicity value is log(1/LC50).

We constructed the appropriate models for chemical classification and for toxicity prediction, using 456 cases, and we used these models to predict the toxicity of the other 112 cases. The best result for chemical classification (compare Table 9) had a prediction accuracy of about 88.4% correctly classified instances on the test set and nearly 99.1% on the training set.

Following combination with the predictions of the sub-models, the training/test results for regression and classification into toxicity classes were a little inferior than the cross-validated results. We obtained a value for R^2 for the combined model of 0.745 compared to 0.509 with a linear regression using all descriptors and 0.630 with a selected set of attributes (compare 3.1). The classification results in toxicity classes, at first sight, do not follow the same trend. For single linear regression with all attributes we achieved 66.1% correctly classified instances (74 out of 112). Reduction of the number of descriptors improved this to 70.6% (79 instances), but the combined model "classified" 69.6% of the instances (78) correctly. At point 2.1, we noted that it is important not to look merely at the overall classification accuracy into chemical classes but also to scrutinize where the misclassified instances are. Indeed it is more dangerous to classify a highly toxic compound as less toxic, than the other way round. In that sense, despite the slight deterioration overall there were less instances of molecules classified as less toxic than originally detected. The same can be said for the ten-fold cross-validation results. These results confirm that this approach gives good results in the complex field of toxicity prediction.

The classifier of chemical classes, with correct prediction of about 85%, lessens the final combination result. To verify that the approach takes advantage of the good results within the sub-models, we calculated a weighted mean of the errors of each sub-model weighted by the number of instances occurring in each subset, as shown in Table 8. Comparison with Table 10 shows that the combination attains levels very close to the weighted mean of the errors.

The classification of compounds into chemical classes is ambiguous. Often there is more than one functional group in a molecule so it can potentially be classified into two or more chemical classes. This is a general problem and does not apply specifically to our study. Thus for some incorrectly classified compounds the selected model is almost able to predict the correct way, but for others the prediction is false, so the results deteriorate. In our case it is difficult to estimate the influence of this, but spot checks showed that in most cases a bad choice of chemical classifier worsened the final result. One way to improve the outcome of the selection could be to use more constitutional descriptors in the set, considering that the cost and time needed to calculate them is relatively low.

4. Discussion

The prediction of toxicity using advanced models is a topic that merits attention for the possible advantages offered by these models, compared to laboratory experimental methods. However, the issue of prediction is complex. The target of these models, the definition of the inputs, the software to be used and how to assess the results are all matters for discussion by the regulatory bodies. For this reason, it is useful for them to have studies at hand comparing different approaches. Specific problems for ecotoxicity are:

(1) incomplete knowledge of the toxicity,
(2) large variable quality of the experimental data,
(3) limited number of experimental data.

Chemical representation is also questionable, since many thousands of chemical descriptors can be used, but a lot of them are highly redundant and have no clear *a priori* relationship with the toxic phenomenon under study.

The two main criteria to obtain the sub-models are:

- to split the compounds according to their chemical classification (as we propose here);
- to split the compounds according to their toxic mode of action (MOA).

In a study on toxicity prediction in the fathead minnow, Russom *et al.*[32] encoded human expert knowledge in the definition of MOA (Mode Of Action). They classified eight mechanisms involved in aquatic toxicity. They consequently identified specific chemical fragments and developed a heuristic to find the MOA on the basis

of these fragments. Finally, they produced simple QSAR models to predict toxicity for specific MOA.

Indeed, several definitions of MOA have been proposed. Also about narcosis, a simple MOA, different opinions exist. Moreover, MOA should be considered as a continuum. For these problems, and for the *a posteriori* empirical definition of MOA, many studies have instead addressed models on specific chemical classes, instead than MOA.

There are problems basically related to the possibility that a chemical can belong to more than one chemical class, and when organizing the sub-models with a unique architecture. Our approach relies on an automatic system to define the chemical classes and then combines specific models to predict toxicity. Our integrated system compares well with previous methods requiring greater human knowledge.

Of course our approach is far from complete, but it has the advantage of being modular, which makes it more flexible than other holistic approaches. Some papers have reported similar results with holistic models for toxicity classes[28] and another approach used probabilistic neural networks.[20] Holistic approaches have the advantage of simplicity, but in order to make progress they require a complete rebuilding of the model. The combination of local experts presented here has the advantage that

(1) the reliability of the models for different chemical classes is more clearly recognized and defined, and
(2) it offers a simple opportunity to study the weaker sub-models.

Using a flexible architecture, sub-models can be easily modified, introducing better models, or even integrated with new, independent sub-models.

5. Conclusion

We are increasingly aware of the need to understand and predict the consequences of chemicals on human health and the environment.[12] This is now studied through *ad hoc* experiments, which are very expensive, may take years and involve animals. The huge number of compounds makes this especially challenging: there are more than 23 million listed in the Chemical Abstract Service. However, data on toxicity is available only for 10% of industrially produced chemicals. Moreover, the increasingly widespread use of the newly introduced Combinatorial Chemistry by chemical companies will increase this number by an order of magnitude.

At the moment we can address the scientific and economic importance of the predictive models developed to date by considering two main areas: combinatorial chemistry, the process of building new compounds and ecotoxicology, the process of regulating their use. Combinatorial chemistry works through economy of scale: libraries of thousands of compounds are built up and screened in quantities of a few atoms. Libraries can be designed to be targeted, i.e. compiled from compounds that are likely to be relevant to the target compound, or diverse, to maximize the

discovery potential. Researchers use QSAR analysis in order to choose compounds that are likely to have the required characteristics. For ecotoxicology the primary concern, even before that of cost reduction, is the reduction of animal experiments.

In the present study we compared methods to predict toxic effects both as a number and as a class label. From a general point of view QSAR prefers methods that compute a continuous value; however, this value must be evaluated considering that the experimental data used for training (from animal experiments) may present high variability, because of animal variability and experimental procedure. This is the first reason for considering methods that provide categorical values. The second is that the classification can help for a first screening of toxic properties of chemicals, on the basis of regulatory schemes that classify chemicals.

We developed a method to automatically split a wide and heterogeneous data set of pollutants into chemical classes according to chemical descriptors. Then we developed local models for the so-defined chemical classes and we combined the local models into a classifier. We also discussed the advantages of flexible modular classifiers; one reason being to keep as simple as possible any improvement in one of the classifiers in case new experimental data become available, another being the reuse of good models in new classifiers.

Acknowledgments

This work was partially funded by EU contract QLK5-CT-2002-00691 Demetra; Christoph König and Marian Craciun were supported by EU contract HPRN-CT-1999-00015 Imagetox.

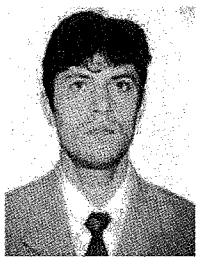
References

1. C. Ambroise and G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, *Proc. Natl. Acad. Sci.* **99** (1999) 6562–6566.
2. A. O. Aptula, T. I. Netzeva, I. V. Vlakova, M. T. D. Cronin, T. W. Schultz, R. Kühne and G. Schürmann, Multivariate discrimination between models of toxic action of phenols, *QSAR* **21** (2002) 12–22.
3. E. Bauer and R. Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, *Mach. Learn.* **36**(1–2) (1999) 105–139.
4. E. Benfenati and G. Gini, Computational predictive programs (expert systems) in toxicology, *Toxicology* **119** (1997) 213–225.
5. E. Benfenati, P. Mazzatorta, D. Neagu and G. Gini, Combining classifiers of pesticides toxicity through a neuro-fuzzy approach, in *Multiple Classifier Systems*, eds. F. Roli and J. Kittler, Lecture Notes in Computer Science, Vol. 2364 (Springer, 2002) pp. 293–303.
6. S. P. Bradbury and R. L. Lipnick, Introduction: structural properties for determining mechanisms of toxic action, *Environ. Health Persp.* **87** (1990) 181–182.
7. P. Chan and S. Stolfo, Learning arbiter and combiner trees from partitioned data for scaling machine learning, in *Proc. 1st Int. Conf. Knowledge Discovery and Data Mining*, August 1995, pp. 39–44.
8. R. D. Combes and P. Judson, The use of artificial intelligence systems for predicting toxicity, *Pestic. Sci.* **45** (1995) 179–194.

9. J. C. Dearden, M. D. Barratt, R. Benigni *et al.*, The development and validation of expert systems for predicting toxicity, *ATLA* **25** (1997) 223–252.
10. N. R. Draper and H. Smith, *Applied Regression Analysis*, 2nd edn. (John Wiley, NY, 1981).
11. ECOTOX, *ECOTOXicology Database System,* Prepared for the US Environmental Protection Agency, by OAO Corporation, Duluth, Minnesota, February 2000.
12. G. Gini and A. Katritzky (eds.), *Predictive Toxicology of Chemicals: Experiences and Impact of Artificial Intelligence Tools* (AAAI Press, Menlo Park, CA, USA, 1999).
13. G. Gini *et al.*, Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network, *J. Chem. Inform. Comput. Sci.* **39** (1999) 1076–1080.
14. M. A. Hall, Correlation-based feature selection for machine learning, Ph.D. thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.
15. C. Hansch, D. Hoekman, A. Leo, L. Zhang and P. Li, The expanding role of quantitative structure-activity relationships (QSAR) in toxicology, *Toxicol. Lett.* **79** (1995) 45–53.
16. T. K. Ho, Multiple classifier combination: lessons and next steps, in *Hybrid Methods in Pattern Recognition*, eds. A. Kandel and H. Bunke (World Scientific, 2002).
17. R. A. Jacob, M. I. Jordan, S. J. Nowlan and G. E. Hinton, Adaptive mixtures of local experts, *Neural Comput.* **3** (1991) 79–87.
18. A. Jain, P. Duin and J. Mao, Statistical pattern recognition: a review, *IEEE Trans. PAMI* **22**(1) (2000) 4–37.
19. M. I. Jordan and R. A. Jacob, Mixtures of experts and the EM algorithm, *Neural Comput.* **6** (1994) 181–214.
20. K. L. E. Kaiser and S. P. Niculescu, Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*pimephales promelas*): a study based on 865 compounds, *Chemosphere* **38** (1999) 3237–3245.
21. J. Kittler, M. Hatef, R. Duin and J. Matas, On combining classifiers, *IEEE Trans. Patt. Anal. Mach. Intell.* **20**(3) (1998) 226–239.
22. R. Kohavi and G. H. John, Wrappers for feature subset selection, *Artif. Intell.*, special issue on relevance **97**(1–2) (1996) 273–324.
23. I. Kononenko, Estimating attributes: analysis and extension of RELIEF, *European Conf. Machine Learning*, Catania, Italy, April 1994, pp. 171–182.
24. A. Krogh and J. Vedelsby, Neural network ensembles, cross validation and active learning, in *Advances in Neural Information Processing Systems*, Vol. 7, eds. G. Tesauro, D. S. Touretzky and T. K. Leen (MIT Press, Cambridge, MA, 1995).
25. P. Mazzatorta, E. Benfenati, D. Neagu and G. Gini, The importance of scaling in data mining for toxicity prediction, *J. Chem. Inform. Comput. Sci.* **42** (2002) 1250–1255.
26. J. M. McKim, S. P. Bradbury and G. J. Niemi, Fish acute toxicity syndromes and their use in the QSAR approach to hazard assessment, *Environ. Health Persp.* **87** (1987) 171–186.
27. D. Neagu and G. Gini, Neuro-fuzzy knowledge integration applied to toxicity prediction, in *Innovations in Knowledge Engineering*, eds. R. Jain *et al.* (Advanced Knowledge International Pte Ltd., Australia, 2003).
28. M. Pintore, N. Piclin, E. Benfenati, G. Gini and J. R. Chrétien, Predicting toxicity against the fathead minnow by adaptive fuzzy partition, *QSAR Comb. Sci.* **22** (2003) 210–219.
29. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Mateo, CA, 1993).

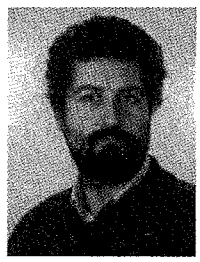
30. F. Roli and G. Fumera, Analysis of linear and order statistics combiners for fusion of imbalanced classifiers, in *Multiple Classifier Systems*, eds. F. Roli and J. Kittler, Lecture Notes in Computer Science, Vol. 2364 (Springer, 2002) pp. 252–261.
31. A. Roncaglioni, A. Colombo, U. Maran, M. Karelson and E. Benfenati, Prediction of acute aquatic toxicity to fish comparing different QSAR approaches, *Toxicol. Lett.* **144** (2003) 52.
32. C. L. Russom, S. P. Bradbury, D. E. Hammermeister and S. J. Drummond, Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*pimephales promelas*), *Environ. Toxicol. Chem.* **16** (1997) 948–967.
33. A. J. C. Sharkey, *Combining Artificial Neural Nets — Ensemble and Modular Multi-Net Systems* (Springer, London, 1999).
34. WEKA 3.2.3, www.cs.waikato.ac.nz/ml/weka.
35. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (Morgan Kaufmann Publishers, San Francisco, CA, 1999).
36. D. H. Wolpert, Stacked generalization, *Neural Networks* **5** (1992) 241–259.
37. L. Xu and W.-J. Zhang, Comparison of different methods for variable selection, *Anal. Chim. Acta* **446** (2001) 475–481.

Cristoph König graduated in informatics for business from the University of Applied Sciences, Dresden, Germany, in 2002. After training in data mining and a fellowship in the research program "IMAGETOX" at the "Mario-Negri" (2002) and at Politecnico di Milano (2002/2003), he is now a researcher at BMW Group AG, Munich, Germany.



Marian Viorel Craciun received the B.S. degree in mathematics and informatics in 1998 and the M.Eng. degree in computer engineering in 2002 from University Dunarea de Jos of Galati, Romania. Currently, he is a Ph.D. student in the same university.

Giuseppina Gini received the doctorate in physics from Milan University, Italy, in 1972. After various appointments as Assistant Professor at Politecnico di Milano and as Research Fellow at Stanford University (AI Laboratory and NMR Laboratory), she is an Associate Professor at Politecnico di Milano, Italy.



Emilio Benfenati received the doctorate in chemistry from Milan University, Italy in 1979. After research in pharmaceutical industry at Stanford University, California, currently, he is Head of the Laboratory of Environmental Chemistry and Toxicology, at "Mario Negri" Institute, Milan, Italy.