Modelling and Prediction of Toxicity of Environmental Pollutants

Frank Lemke[1], Johann-Adolf Müller[1], and Emilio Benfenati[2]

[1] KnowledgeMiner Software, Dürerstr. 40
16341 Panketal, Germany
{frank,jamueller}@knowledgeminer.net
http://www.knowledgeminer.net
[2] Istituto di Ricerche Farmacologiche "Mario Negri" Milano, Via Eritrea 62
20157 Milano, Italy
benfenati@marionegri.it
http://www.marionegri.it

Abstract. This paper describes the problem of modelling toxicity of environmental pollutants using molecular descriptors from a systems theoretical viewpoint. It is shown that current toxicity modelling problems systematically incorporate very high levels of noise a priori. By means of a set of individual and combined models self-organised by KnowledgeMiner from a high-dimensional molecular descriptor data set calculated within the DEMETRA project we suggest a way how results interpretation and final decision making can effectively take into account the huge uncertainty of toxicity models.

1 Introduction

The global production of chemicals has increased from 1 million tonnes in 1930 to 400 million tonnes today. There are about 100.000 different substances registered in the EU market of which 10.000 are marketed in volumes of more than 10 tonnes, and a further 20.000 are marketed at 1-10 tonnes.

Besides the economical importance of the chemical industry as Europe's third largest manufacturing industry, it is also true that certain chemicals have caused serious damage to human health resulting in suffering and premature death and to the environment. The incidence of some diseases, e.g. testicular cancer in young men and allergies, has increased significantly over the last decades. While the underlying reasons for this have not yet been identified, there is justified concern that certain chemicals play a causative role for allergies.

The present system for general industrial chemicals distinguishes between "existing substances" i.e. all chemicals declared to be on the market in September 1981, and "new substances" i.e. those placed on the market since that date. There are some 2.700 new substances. Testing and assessing their risks to human health and the environment according to Directive 67/548 are required before marketing in volumes above 10 kg. In contrast, existing substances amount to more than 99% of the total volume of all substances on the market, and are not subject to the same testing requirements. In result, there is a general lack of knowledge about the properties and the uses of existing substances. The risk assessment process is slow and resource-intensive and does not allow the system to work efficiently and effectively [1].

J.A. López et al. (Eds.): KELSI 2004, LNAI 3303, pp. 221–234, 2004.

To address these problems and to achieve the overriding goal of sustainable development one political objective formulated by the European Commision in its White Paper [1] is the promotion of non-animal testing, which includes:

- Maximising use of non-animal test methods;
- Encouraging development of new non-animal test methods;
- Minimising test programmes.

A current way in that direction is building mathematical, Quantitative Structure-Activity Relationship (QSAR) models based on existing test data that aim on describing and predicting the short-term, acute impact of a chemical compound (pollutant) on the health of a population of a certain biological species. This impact can either be direct by injection or feeding or indirect by introducing a specific concentration of a chemical into the environment (air, water, soil). Representative for expressing the chemicals's impact on the population's health the lethal dose LD_{50} or the lethal concentration LC_{50} (toxicity) is measured correspondingly. LC_{50}, for example, specifies the experienced concentration of a chemical compound where 50% of the population died within a given time after intoduction of the chemical to the system.

In this work the Group Method of Data Handling (GMDH) [2] is used as a very effective and valuable modelling technology for building mathematical models and predictions of the lethal concentration.

2 The Problem of Modelling Toxicity

2.1 Systems Analysis

Generally, real-world systems are time-variant nonlinear dynamic systems [3]. Therefore, it should be useful to allow the modelling algorithm to generate systems of nonlinear difference equations. For toxicity modelling this system can be considered time-invariant due to the intentionally short-term effect of the pollutant.

A possible dynamic model of the ecotoxicological system is shown in figure 1,

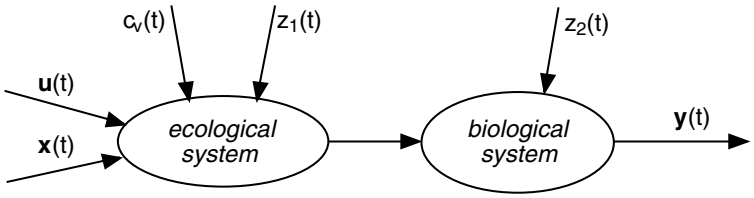

Fig. 1. Dynamic model of an aquatic ecotoxicological system

where
x(t) – state vector of the ecological system at time t,
u(t) – vector of external variables at time t,
$c_v(t)$ – concentration of the pollutant v at time t,
$z_1(t)$, $z_2(t)$ – external disturbances to the system at time t,

$\mathbf{y}(t)$ – output vector of dimension p describing the health of the population at time t,
$\mathbf{y}(t)=[y_1(t), y_2(t),.., y_m(t), .., y_p(t)]^T$
$y_m(t)$ – the population's cumulated mortality rate at time t (see also fig. 3).
This dynamic model is described by the following system of equations:

$$\begin{aligned} \mathbf{x}(t+1) &= G(\mathbf{x}(t), \mathbf{u}(t), c_v(t), z_1(t), z_2(t)) \\ \mathbf{w}(t) &= H_1(\mathbf{x}(t), \mathbf{u}(t), c_v(t), z_1(t)) \\ \mathbf{y}(t) &= H_2(\mathbf{w}(t), z_2(t)) = H^*(\mathbf{x}(t), \mathbf{u}(t), c_v(t), z_1(t), z_2(t)) \end{aligned} \quad (1)$$

with $c_v(t)=\begin{cases} c_0 & t=t_0 \\ 0 & \text{else} \end{cases}$, and c_0 as the concentration of the test compound v in mg/l.

During the animal tests, however, the external variables $\mathbf{u}(t)$ and the state variables $\mathbf{x}(t)$ of the system are not observed, usually, or not observable and therefore they are considered constant so that for modelling the ecotoxicological system transforms into a nonlinear static system (fig. 2):

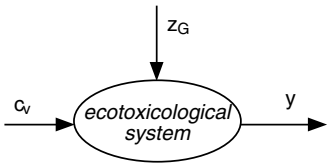

Fig. 2. Reduced model of the static system with noise $z_G = h_1(z_1, z_2)$

Additional noise z_3 is introduced to the static system by the missing information of external and state variables that now transforms to noise. Also the testing procedure itself adds some noise z_4 so that the static system's noise finally is $z_S = h_2(z_G, z_3, z_4)$, and the modelling task of the ecotoxicological system reduces to approximating the dependence of the experienced mortality rate y from the pollutant's concentration c_v:

$$y = f_1(c_v, z_S). \quad (2)$$

If an animal experiment is repeated several times using the same concentration $c_{i,v}$ of a chemical test compound v multiple experienced mortality rate values $\mathbf{y}_{c_{i,v}}$ are available (fig. 3). This means, for $c_{i,v}$ = const., the interval of the observed mortality rate values $\mathbf{y}_{c_{i,v}}$ can be seen as a direct expression of the static system's noise z_S. For the reverse case of measuring the concentration c_v for a constant mortality rate y_j = const. the problem transforms to

$$c_v = f_2(y_j, z_S) \text{ (fig. 3).} \quad (3)$$

For $y_j = 50\%$, c_v is the experienced lethal concentration LC_{50} for a pollutant v, which is actually used as the output variable in toxicity QSAR modelling. With a commonly observed rate $\frac{c_{v,max}}{c_{v,min}} \approx 4$ for a single compound v this output variable can be seen as highly noisy.

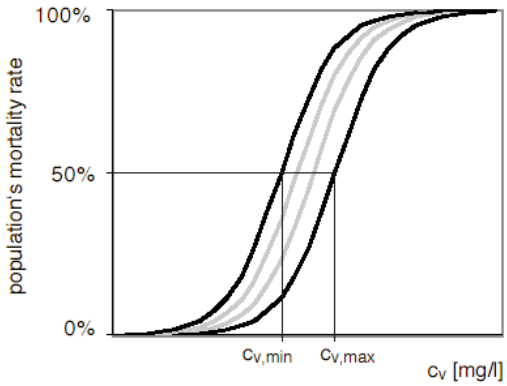

Fig. 3. Variation of LC_{50} resulting from a number of comparable tests

The initial task of modelling the observed mortatlity rate y from a pollutant's concentration c_v now shifts to finding a description of the dependence of a pollutant's lethal concentration LC_{50} for a specific species from the pollutant's molecular structure $\mathbf{s}_v$ (fig. 4):

$$LC_{50} = f_3(\mathbf{s}_v, z_M), \text{ with } z_M = h_3(z_S) \tag{4}$$

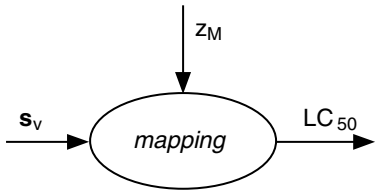

Fig. 4. The toxicity modelling problem as applied in practice. Note that the input variable c_v (LC_{50}) of the initial ecotoxicological system (fig. 1 and 2) has shifted to now being the objective of modelling

This finally means not to model the object itself – the ecotoxicological system – but one of its inputs – the external disturbance c_v. The initial system's input-output relation is mapped by just a single pair of observations (LC_{50}, y) so that it is described by a linear relationship a priori.

A next problem is how to express the structure $\mathbf{s}_v$ of the chemical v. Commonly, it is a complex chemical object, but for building a mathematical model that describes the dependence of the toxicity from the chemical structure a formal transformation into a set of numerical properties - descriptors - is required. This transformation is based on chemical and/or biological domain knowledge implemented in some software (fig. 5):

$$\mathbf{d}_v = f_4(\mathbf{s}_v, z_T) \tag{5}$$

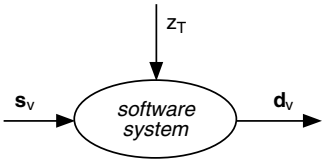

Fig. 5. Model of the chemical structure to molecular descriptor transformation

In the chemical domain, for example, input of the software system can be a 2-dimensional or a 3-dimensional drawing of the chemical structure, but also SMILES coded strings or other expressions may be possible. Output of the system is a certain set of molecular descriptors depending on the software used and the theoretical model implemented. Applying different software provides different sets of descriptors that may intersect to some extent but may not necessarily have identical values though. Also, the interpretational power of descriptors can be low or difficult when they loose chemical meaning.

The process of descriptor calculation also adds noise. Not only software bugs or manual failures may introduce noise, more important for introduction of uncertainty should be the interpretational clearance of domain knowledge for properly formalising an appropriate set of molecular descriptors, different starting condition assumptions (conformation) for descriptor calculation, or several different optimisation options. Not always is their chemical meaning very strong or theoretically accounted.

The final, simplified nonlinear static model used in QSAR modelling to describe acute toxicity is (fig. 6):

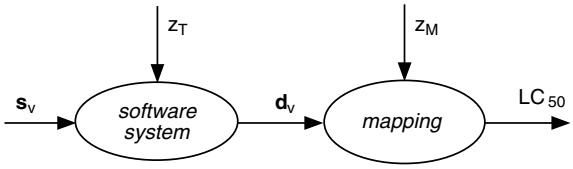

Fig. 6. Simplified model for describing acute toxicity

with

$$LC_{50} = f_5(f_4(\mathbf{s}_v, z_T), z_M) = f(\mathbf{s}_v, z_T, z_M), \tag{6}$$

LC_{50} – experienced lethal concentration for a certain species and chemical compound,

$\mathbf{s}_v$ – the structure of the tested chemical compound in the chemical domain,

z_T – noise of the chemical structure to molecular descriptor transformation process,

z_M – noise transformed from the ecotoxicological test system,

$\mathbf{d}_v$ - vector of numerical molecular descriptors of the test compound

The external disturbance z_T which adds noise to descriptor input space used for modelling can be reduced by fixing bugs and manual failures and by finding a most consistent chemical structure to descriptor transformation – although it is not clear a priori which transformation or optimisation will add and which will reduce noise. The

disturbance z_M, which finally results from the experimental tests, in contrast, adds noise to the output LC_{50} and is a given fact that cannot be changed afterwards.

2.2 Modelling Methods

Apparently, toxicity QSAR modelling implies dealing with very noisy data. Data sets are generally not perfect reflections of the world. The measuring process necessarily captures uncertainty, distortion and noise. Noise is not errors that can infect data but part of the world. Therefore, a modelling tool, but also results and decisions, must deal with the noise in the data. Information about the noise dispersion can be useful for choosing adequate modelling technologies by referencing the ideas of Stafford Beer's adequacy law [4]: The "black boxes" of the objects have to be compensated by corresponding "black boxes" in the information or control algorithm. Based on this idea, the following general classification of modelling algorithms is suggested in [2]: For a small level of noise dispersion, all regression-based methods using some internal criterion can be applied:

- GMDH with internal selection criteria,
- Statistical methods, or
- Neural Networks.

For considerably noisy data – which always includes small data samples – GMDH or other algorithms based on external criteria are preferable. For a high level of noise dispersion, i.e., processes that show a highly random or chaotic behavior, finally, nonparametric algorithms of clustering, Analog Complexing, or fuzzy modelling should be applied to satisfy the adequateness law. This implies also that with increasing noise in the data the model results and their descriptive language become fuzzier and more qualitative.

There is a broad spectrum of possible algorithms to use, because it is not possible to define the characteristics of the controlled object in advance, exactly. Therefore, it is helpful to try several modelling algorithms, first, and then decide which algorithms suit the given type of object best or most appropriately combine the results of different modelling runs in a hybrid model. In QSAR modelling, for several reasons, predominantely algorithms have been used for modelling linear static systems (linear regression, PLS, especially), which is an additional significant simplification of the highly disturbed ecotoxicological system model. One reason surely is connected with problems in creating and validating reliable descriptive and predictive nonlinear models. Even in cases where it was possible to create to some meaning good predictive nonlinear models (Neural Networks) – not looking at the special validation requirements of nonlinear models in general – they commonly have no or only low descriptive power which, however, turns out being an important feature for applicability and acceptability in real-world scenarios. Users usually don't want to rely decisions on kind of "black boxes". Due to the large noise level in toxicity modelling descriptive power might also be part of the model evaluation procedure, because models that can be interpreted from a theoretical viewpoint can be judged using domain knowledge. Another reason for preferring linear models in toxicity QSAR modelling is the high-dimensional descriptor space and/or the comparingly low number of tested com-

pounds, which always implies state space dimension reduction. Linear approaches are widely used here in preprocessing to obtain a small set of “best” descriptors, where “best” then relates to building linear models.

2.3 Modelling Technologies Used

2.3.1 High-Dimensional Modelling

A new approach to high-dimensional state space modelling we have been developing and using is based on multileveled self-organisation. The basic idea here is dividing high-dimensional modelling problems into smaller, more manageable problems by creating a new self-organising network level composed of active neurons, where an active neuron is represented by an inductive learning algorithm (lower levels of self-organisiation) applied to disjunct data sets. The objective of this approach is based on the principle of regularisation of ill-posed tasks, especially the requirement of defining the actual task of modelling a priori to be able to select a set of best models. In the context of a knowledge discovery from databases, however, this also implies using this principle in every stage of the knowledge extraction process – data preselection, preprocessing including dimension reduction, modelling (data mining), and model evaluation – consistently. The proposed approach of multileveled self-organisation integrates preprocessing, modelling, and model evaluation into a single, automatically running process and it therefore allows for directly building reliable models from high-dimensional data sets (up to 30.000 variables) objectively. The external information necessary to run the new level of self-organisation is provided by the corresponding algorithm’s noise sensitivity characteristic as explained in [5, 6].

2.3.2 Inductive Learning Algorithm

The inductive learning algorithm we used in this work in the network’s active neurons is the Group Method of Data Handling (GMDH) as described in more detail in [2]. The theory of GMDH Neural Networks was first developed by A.G. Ivakhnenko [7, 8] in 1968 based on Statistical Learning Network theory and on the principle of induction, where induction consists of

- The cybernetic principle of self-organization as an adaptive creation of a network without subjective points given;
- The principle of external complement enabling an objective selection of a model of optimal complexity and
- The principle of regularization of ill-posed tasks.

This different foundation compared to traditional Backpropagation Neural Networks allows for autonomous and systematical creation of optimal complex models by employing both parameter and structure identification. An optimal complex model is a model that optimally balances model quality on a given learning data set ("closeness of fit") and its generalisation power on new, not previously seen data with respect to the data's noise level and the task of modelling (prediction, classification, modelling, etc.). It thus solves the basic problem of experimental systems analysis of systematically avoiding "overfitted" models based on the data's information only. This makes GMDH a most automated, fast and very efficient supplement and alternative to other data mining methods. Also, in result of modelling an analytical model in form of

algebraic formulas, difference equations, or systems of equations is available on the fly for interpretation and for gaining insight into the system. In our work the GMDH implementation of the KnowledgeMiner software was used, exclusively [9].

2.3.3 Model Combining

Another focus is on model combining. There are several reasons to combine models or their results [2]:

1. All kinds of parametric, nonparametric, algebraic, binary/fuzzy logic models are only simplified reflections of reality. There are always several models with a sufficient same degree of adequacy for a given data sample. However, every model is a specific abstraction, a one-sided reflection of some important features of reality only. A synthesis of alternative model results gives a more thorough reflection.
2. Although models are self-organised, there is still some freedom of choice in several areas due to the regularisation requirement of ill-posed tasks. This freedom of choice concerns, for example, the type of model (linear/nonlinear) and the choice of some modelling settings (threshold values, normalisation etc.). To reduce this unavoidable subjectivity, it can be helpful to generate several alternative models and then, in a third level of self-organisation, improving the model outputs by synthesising (combining) all alternative models in a new network.
3. In many fields, such as toxicology, there are only a small number of observations, which is the reason for uncertain results. To improve model results the artificial generation of more training cases by means of jittering, randomisation, for example, is a powerful way here.
4. All methods of automatic model selection lead to a single "best" model while the accuracy of model result depends on the variance of the data. A common way for variance reduction is aggregation of similar model results by means of resampling and other methods (bagging, boosting) following the idea: Generate many versions of the same predictor/classifier and combine them.
5. If modelling aims at prediction, it is helpful to use alternative models to estimate alternative forecasts. These forecasts can be combined using several methods to yield a composite forecast of a smaller error variance than any of the components have individually. The desire to get a composite forecast is motivated by the pragmatic reason of improving decision-making rather than by the scientific one of seeking better explanatory models. Composite forecasts can provide more informative inputs for a decision analysis, and therefore, they make sense within decision theory, although they are often unacceptable as scientific models in their own right, because they frequently represent an agglomeration of often conflict theories.

3 Results on Modelling Toxicity of Pesticide Residues

3.1 The Data Set

We used a data set calculated within the DEMETRA project [10]. It contains 281 chemical compounds – pesticides - and given corresponding experienced lethal concentrations LC_{50} for trout. 1061 2D molecular descriptors were calculated by different

commercial or publicly available software. This descriptors set is highly redundant so that by means of clustering a non-redundant nucleus of 647 potential 2D descriptors showing a diversity of at least 2% was obtained. 46 chemical compounds were hold out for out-of-sample testing (N_C) of the generated models so that 235 pesticides were used for modelling ($N_{A,B}$).

3.2 Individual Models

A set of 13 different linear and non-linear QSAR models M_1 to M_{13} was self-organised directly from this data set by the KnowledgeMiner data mining software [9]. The necessary workflow of accessing data from the database, preprocessing (missing values detection, data transformation), and modelling (data mining) was automated by applying AppleScript integrating various software tools running under Mac OS X in that way (MS Excel, MS Word, TextEdit, Valentina DB, AppleWorks, KnowledgeMiner).

For each model we calculated three different model performance measures: Descriptive Power (DP) as described in [5], the Coefficient of Determination (R^2), and the Mean Absolute Percentage Error (MAPE) as follows:

$$R^2 = 1 - \delta^2, \delta^2 = \frac{\sum_{i \in N} (y_i - \hat{y}_i)^2}{\sum_{i \in N} (y_i - \bar{y})^2} \leq 1, \tag{7}$$

$$MAPE = \frac{\sum_{i \in N} |y_i - \hat{y}_i|}{\sum_{i \in N} |y_i|} \times 100\%, \tag{8}$$

where y_i, $\hat{y}_i$, and $\bar{y}$ are the true, estimated, and mean values of the output variable, respectively, and δ^2 is the Approximation Error Variance criterion [2].

The corresponding results are listed in table 1.

Table 1. Performance parameters for 13 individual models self-organised by KnowledgeMiner

	Calculated on $N_{A,B}$			Calculated on N_C		Calculated on $N_{A,B,C}$	
MODEL	R^2	DP [%]	MAPE [%]	R^2	MAPE [%]	R^2	MAPE [%]
M1 (linear)	0,69	43	28	0,54	34	0,67	28
M2 (linear)	0,71	44	28	0,42	37	0,66	29
M3 (nonlinear)	0,71	40	26	0,49	34	0,68	28
M4 (nonlinear)	0,74	43	25	0,41	37	0,63	28
M5 (nonlinear)	0,68	40	n.a.	0,31	47	0,62	31
M6 (linear)	0,71	45	26	0,36	40	0,64	30
M7 (linear)	0,71	45	26	0,33	42	0,63	31
M8 (nonlinear)	0,76	47	23	0,30	39	0,66	28
M9 (nonlinear)	0,75	46	24	0,21	43	0,64	29
M10 (linear)	0,70	45	27	0,58	31	0,68	28
M11 (linear)	0,69	44	28	0,54	33	0,66	29
M12 (nonlinear)	0,72	44	26	0,49	33	0,68	28
M13 (nonlinear)	0,76	48	25	0,42	37	0,69	28

3.3 The Combined Model

Finally, a combined model M_{comb} out of the 13 individual models was generated likewise. The combined model is built on the predicted toxicity values of the individual models M_1 to M_{13} as input information. To introduce new independent information for this second model optimization level, all chemical compounds of the initial data set including those hold-out for testing were used for modelling so that all 281 compounds built the learning data set here ($N_{A,B}$). This is possible and reasonable, because the modelling task is set to work under conditions for which the generalization power of the external cross-validation selection criterion of the GMDH algorithm [2] works properly according to the algorithms's noise sensitivity characteristic [5, 6]. Table 2 shows the performance improvements of the combined model.

Table 2. Performance parameters for the combined model

		Calculated on $N_{A,B}$	
MODEL	R^2	MAPE [%]	MAPE [%]
M_{comb} (linear)	0,76	50	25

The self-organised model equation for M_{comb}, $y = f_1(M_5, M_{10}, M_{11}, M_{12}, M_{13})$, is:

$$\text{Lg}(LC_{50}\ [\text{mmol/l}]) = 0.131 - 0.243\ M_{11} + 0.242\ M_5 + 0.384\ M_{10} + 0.301\ M_{12} + 0.364\ M_{13} \quad (9)$$

and it is finally composed of 53 different descriptors.

3.4 Model Uncertainty and Prediction Interval

As pointed out in section 2, toxicity data are highly noisy and therefore require adequate modelling and results interpretation methods. Additionally, all methods of automatic model selection lead to a single "best" model. On this base are made conclusions and decisions as if the model was the true model. However, this ignores the major component of uncertainty, namely uncertainty about the model itself. In toxicity modelling it is not possible that a single crisp prediction value can cover and reflect the uncertainty given by the initial object's data. If models can be obtained in a comparingly short time it is useful to create several alternative reliable models on different data subsets or using different modelling methods and then to span a prediction interval from the models' various predictions for describing the object's uncertainty more appropriately. In this way a most likely, a most pessimistic (or most save), and a most optimistic (or least save) prediction is obtained, naturally, based on the already given models only, i.e., no additional (statistical) model has to be introduced for confidence interval estimation, for example, which would had to make some new assumptions about the predicted data, and therefore, would include the confidence about that assumptions, which, however, is not known a priori.

A prediction interval has two implications:

1. The decision maker is provided a set of predicted values that are possible and likely representations of a virtual experimental animal test including the uncertainty once observed in corresponding past real-world experiments. The decision maker can base its decision on any value of this interval according to importance, reliability, safety, impact or effect or other properties of the actual decision. This keeps the principle of freedom of choice for the decision process.
2. Depending on which value used, a prediction interval also results in different model quality values starting from the highest accuracy for most likely predictions.

Figure 7 displays the prediction intervals for test set compounds (N_C) from the models contained in the combined model M_{comb} reported in 3.3.

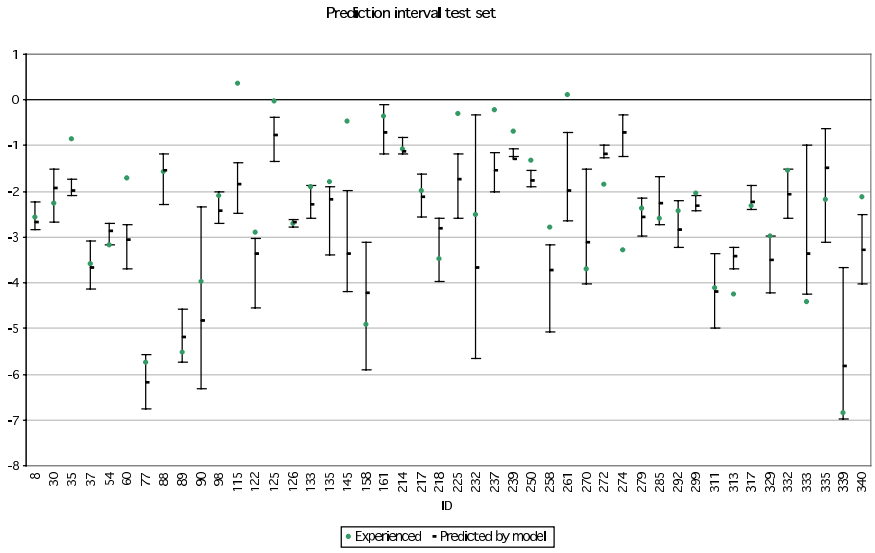

Fig. 7. Prediction interval for the test set from the models contained in the combined model

In a real-world application scenario evaluation and decision-making can only base on predictions; no experienced toxicity value is given, usually, except those available from past tests. A supplement to providing prediction intervals that covers model uncertainty for decision making from another perspective can be the following approach:

1. For N compounds create a list of pairs $(y_i, \hat{y}_i)$ with y_i as the observed toxicity for a compound i and $\hat{y}_i$ as the predicted toxicity for a compound i. N preferably equals the total number of compounds available for a data set, i.e., learning and testing data. The estimated/predicted values $\hat{y}_i$ can be any values of the prediction interval, minimum, maximum, mean, for example.
2. Sort the matrix $\left[\mathbf{y}\ \hat{\mathbf{y}}\right]$ with respect to column $\hat{\mathbf{y}}$.
3. Create q equidistant intervals (classes) based on $\hat{\mathbf{y}}$.

The result is q disjoint classes of corresponding observed and estimated toxicity values. For each class j, j=1, 2, .., q, the estimated toxicity mean and the minimum, maximum, and mean of the observed toxicities can be calculated. This means that here an interval of observed toxicity values for a given interval of predicted toxicities is obtained that describes the prediction's uncertainty for a related class or interval. Using a new compound's most likely prediction from the prediction interval, for example, this value would decide in which prediction class the compound would fit into along with the class' uncertainty given by the interval of past experienced toxicity values. Figure 8 plots the results of a derived decision model for q=12 classes from the predictions of the combined model reported in 3.3 and table 3 lists the underlying data of fig. 8 for reference. For comparison, the results based on the minimum (most toxic) predictions of the 13 individual models of section 3.2 are shown in fig. 9. Table 4 shows the accuracy values for these two decision models compared to a mean-based model.

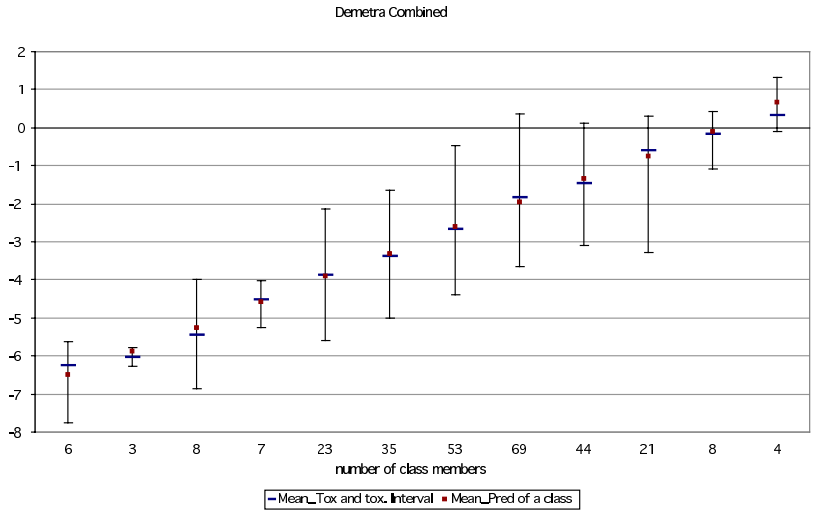

Fig. 8. Decision model based on the predictions of the combined model

Table 3. Underlying data of the decision model of fig. 8

Class	Number of class members	From predicted toxicity	To predicted toxicity	Min. observed toxicity	Mean observed toxicity	Max. observed toxicity	Mean predicted toxicity
1	6	-6.90	-6.24	-7.74	-6.23	-5.62	-6.49
2	3	-6.24	-5.59	-6.27	-6.03	-5.79	-5.88
3	8	-5.59	-4.93	-6.84	-5.45	-3.98	-5.26
4	7	-4.93	-4.27	-5.24	-4.50	-4.02	-4.57
5	23	-4.27	-3.61	-5.58	-3.87	-2.13	-3.89
6	35	-3.61	-2.95	-5.02	-3.37	-1.64	-3.31
7	53	-2.95	-2.29	-4.40	-2.66	-0.47	-2.61
8	69	-2.29	-1.63	-3.66	-1.83	0.36	-1.95
9	44	-1.63	-0.97	-3.10	-1.46	0.12	-1.33
10	21	-0.97	-0.31	-3.27	-0.60	0.30	-0.74
11	8	-0.31	0.35	-1.09	-0.15	0.43	-0.10
12	4	0.35	1.01	-0.10	0.32	1.33	0.66

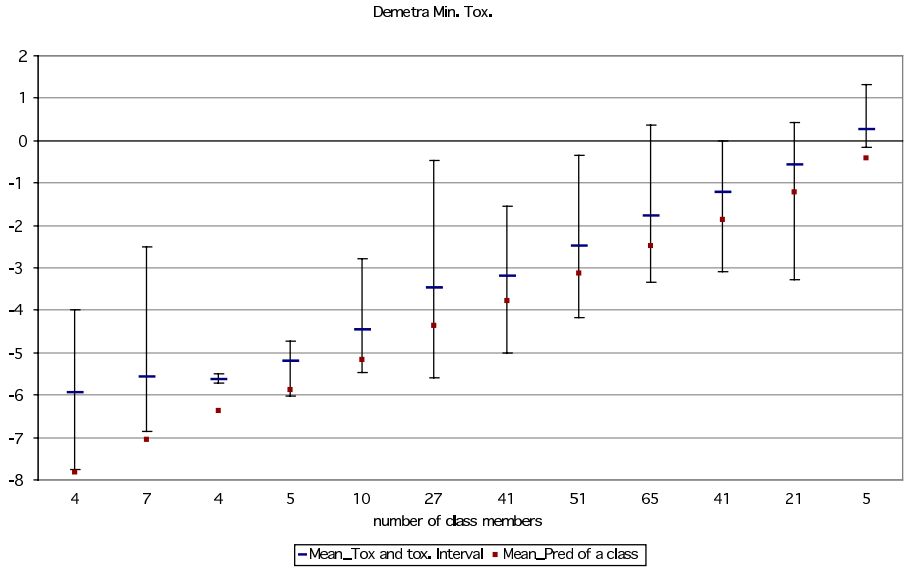

Fig. 9. Decision model of 12 classes based on the minimum predictions of 13 individual models

Table 4. Accuracy of three decision models for trout

	Min. Tox. vs. Mean Pred.	Mean Tox. vs. Mean Pred.	Max. Tox. vs. Mean Pred.
R^2 decision model fig. 8	0.51	0.99	0.68
R^2 decision model fig. 9	0.75	0.79	0.0
R^2 decision model using the mean prediction of 13 models (not displayed)	0.4	0.97	0.5

The result in table 4 confirms the expectation that the combined model shows a higher performance than just using the mean of a number of individual models.

4 Conclusions

The current results and conclusions are primarily based on the Demetra data set, but several other toxicity data sets have been investigated, also.

1. Animal tests run to obtain the data source for toxicity QSAR modelling are described by a complex, nonlinear dynamic ecotoxicological system. The mortality rate of a certain species as an observed output variable of this system, however, is not object of toxicity modelling. Instead, an input variable of the test system – the external disturbance LC_{50} (lethal concentration or dose) – is modelled by a pollutant's molecular structure. The system's observed output variable, the mortality rate y, is mapped by a single pair of observations (LC_{50}, y) and, therefore, is described by a linear static model a priori. This, in fact, is a strong simplification of the ecotoxicological system.

2. Since different values are measured for LC_{50} that can vary up to a factor of 4 when running multiple tests it is also not exactly clear, which of these values can be seen as the "true" value for modelling. This value as the models' target variable, however, has an important impact on model results both predictive and descriptive, which finally means uncertain model results.
3. The used input information for modelling does not reflect very appropriated the desired input-output relation of the complex ecotoxicological system and this results in highly noisy data. Observing additional characteristical state or external variables of the test system during the animal tests may significantly reduce the data's noise and thus the models' uncertainty. The modelling approach should be improved to better cover the system's non-linear and dynamic behaviour.
4. Applying GMDH for multileveled self-organisation and model combining turns out a very effective and valuable knowledge extraction technology for building reliable and interpretable models, objectively, in short time frames from noisy and high-dimensional data sets, directly. Also, the obtained models are easy to implement in other runtime environments for application and reuse.
5. Decision-making has to take into account the models' uncertainty. Prediction and toxicity intervals obtained by applying many alternative models are one efficient way to fit this goal inherently.

Acknowledgement

The work has been done within the project DEMETRA, funded by the European Commission under the contract QLK5-CT-2002-00691.

References

1. European Commission: White Paper. Strategy for a future Chemicals Policy, 27.02.2001
2. Müller, J.-A., Lemke, F.: Self-Organising Data Mining. Extracting Knowledge From Data, BoD, Hamburg, 2000
3. Müller, J.-A.: Systems Engineering. FORTIS Wien, 2000
4. Beer, S.: Cybernetics and Management, English University Press, London 1959
5. Lemke, F., Müller, J.-A.: Validation in self-organising data mining, ICIM 2002, Lvov (http://www.knowledgeminer.net/pdf/validation.pdf)
6. Lemke, F.: Does my model reflect a causal relationship? http://www.knowledgeminer.net/isvalid.htm, 2002
7. Ivakhnenko, A.G., Müller, J.-A.: Selbstorganisation von Vorhersagemodellen, Verlag Technik 1984
8. Farlow, S.J. (Ed.): Self-organizing Methods in Modeling: GMDH-Type Algorithms. Marcel Dekker, New York 1984
9. KnowledgeMiner: Self-organising data mining and prediction tool, http://www.knowledgeminer.net, version X 5.0.8, 2004
10. DEMETRA, EC project, http://www.demetra-tox.net, 2004