Carcinogenicity Prediction of Aromatic Compounds Using Self-organising Data Mining

Frank LEMKE*, Emilio BENFENATI**
**KnowledgeMiner Software, Germany*
*** Istituto di Ricerche Farmacologiche "Mario Negri", Italy*

Abstract. Self-organising data mining is a new approach that supports the workflow process of a Knowledge Discovery more comprehensively and that targets on increasing both reliability and predictive and descriptive power of generated models of ill-defined systems such as ecotoxicological systems. This paper reports results from applying a self-organising data mining for describing and predicting carcinogenicity of aromatic compounds from molecular descriptors.

1. Self-organising Data Mining

Knowledge Discovery from Data [1] is of increasing importance to model, describe, predict, and understand the complex behaviour of real-world systems in many domains. For an objectively working and sophisticated knowledge discovery, it is necessary to limit the user involvement in the knowledge extraction process to a contributing inclusion of well known a priori or domain knowledge. Soft computing, i.e., Fuzzy Modelling, Neural Networks, Genetic Algorithms, Inductive Modelling, and other methods of a more or less automatic model generation, is a widely used tool within knowledge discovery. However, its application has been almost exclusively focused on core data mining. A new approach that is going to reflect the workflow character of a knowledge discovery – data pre-processing, dimension reduction, data mining, model evaluation, and combining models - more comprehensively, is called self-organising data mining [2] [3] [9]. Of special importance for further application of an obtained model is its final evaluation. From data analysis, only, it is impossible to decide whether the estimated model reflects the causal relationship between input and output, adequately, or if it is just a stochastic model of noncausal correlations. An automated procedure introduced in self-organising data mining that may help evaluating the usefulness of a created model is application of an algorithms' noise sensitivity characteristic [2] [4].

2. Carcinogenicity Prediction of Aromatic Compounds

Man is exposed to many chemicals of natural and synthetic origin. An urgent question concerns their potential negative effects on human health. To identify chemicals inducing toxicity and to limit the incidence of human cancers and other diseases, rodent bioassays are the principal methods used today. However, this approach is not altogether problem-free, on several accounts: (1) the cost of the assay (>1 mill. U.S. dollars/chemical); (2) the time needed for the tests (3-5 years); (3) ethical considerations and public pressure to reduce or eliminate the use of animals in research and testing; (4) difficulties in the extrapolation to man [5].

Additionally, ecotoxicological systems such as the effect of aromatic compounds on the induction of cancer or other tumours are complex, ill-defined systems. These systems are characterised by (1) inadequate a priori information about the system; (2) large number of potential, often immeasurable variables; (3) noisy and few data samples; and (4) fuzzy objects [2]. The economical, ethical, and methodological problems resulting from applying theory-driven methods or even dedicated experts systems [5] suggest using a data-driven approach as outlined in section 1.

In a first analysis, we used an initial data set of 104 aromatic compounds and 34 molecular descriptors [6] and the KnowledgeMiner data mining tool [7] for modelling and knowledge extraction. Analog Complexing, GMDH Neural Networks, and Fuzzy Rule Induction [2] was applied to generate a set of single models as well as a combined solution. A summarised report of this work can be found in [8].

Concluding from the results of this analysis, in a second modelling run, a revised data set of only 92 aromatic compounds was used. Here, experts removed some compounds whose experimental carcinogenicity is not available, finally. Since almost all models in our initial work also constantly reported the largest errors on 7 of the 12 removed compounds (marked as: O2, O50, O55, O68, O83, O85, and O94 in [8]), a goal of the repeated run was to test if accuracy and reliability of models will rise, and to predict carcinogenicity values for the removed 12 compounds. For that purpose, we applied a prototype of KnowledgeMiner that implements a new algorithm for creating a nucleus of input variables by means of a multileveled self-organised state space dimension reduction and an automated, second level

validation procedure for linear GMDH models (noise sensitivity characteristic) [2] [4]. The basic idea here is that evaluation of models has been based on the assumption that data mining algorithms may have the power of ideal noise filters. However, as shown in [4], even an automated leave-one-out cross-validation driven inductive learning approach as implemented in KnowledgeMiner, for example (first level of validation to avoid overfitting), does not show quasi ideal noise filtering. Therefore, evaluation of a final model needs additional justification with the corresponding noise sensitivity characteristic of the algorithm it was created with. We identified this characteristic by means of a Monte Carlo simulation for KnowledgeMiner's linear GMDH models, until now. Therefore, modelling of carcinogenicity was limited to the generation of linear regression models in our recent analysis.

To allow models to express a nonlinear relation, anyhow, we added several synthesised variables to the initial 34 descriptors x_i:

$$u_i = \frac{1}{x_i}; v_i = \frac{1}{x_i^2}; w_i = x_i^2,$$

which, differently combined, built 5 distinct data sets for modelling.

First, to show the impact of the reduced data set on model accuracy, we used a corresponding set of input variables $\{\underline{x}, \underline{u}\}$ and the same algorithm as reported for the linear GMDH model in our initial analysis (called GMDHM1 in [8], reported AEV of 0.54). It turns out that, in fact, a decreased approximation error variance (AEV) is obtained for this model M_0 (table 1), which is described by this self-organised equation:

$$\begin{aligned} Y = {} & 1.43 - 7.9 \times \text{wiener}^{-1} + 0.001 \times \text{heat} + 0.127 \times \text{flex} - 0.034 \times \text{polariz} - 27.11 \times \text{polariz}^{-1} \\ & - 0.002 \times (\text{pH}=2)^{-1} - 0.035 \times \text{homo} + 0.463 \times \text{kappa2}^{-1} + 0.3 \times \text{kappa3}^{-1} + 1.02 \times \text{kA2}^{-1} \end{aligned}$$

An advantage of GMDH Neural Networks is that they are transparent in the definition of the molecular descriptors selected by the learning algorithm. They offer a simple relationship between descriptors and activity, proving the sign of the coefficients used in the equation and the importance of the parameters. The selected descriptors contain useful information, which can be related to the biomechanisms involved in the toxic effect. The interpretational and descriptive property of GMDH models is very important and of high

value for understanding and solving complex real-world problems in practice.

Interpreting model M_0, the molecular descriptor HOMO is the highest occupied molecular orbital energy. It indicates the easiness to ionize the chemical, and it is related to reactivity. In our equation it has a negative parameter, which means that the higher the energy required to ionize the molecule, the lower is its toxicity. This corresponds to what is expected from theory. HEAT of formation is a measure of the stability of the chemical for a given molecular weight. If the heat of formation is higher, the chemical's stability is lower. As expected, the sign in the generated equation is positive. POLARIZABILITY plays a complex role, because first it is involved in the transfer inside the cell, then it implicates a certain reactivity once inside the cell. This can explain why this descriptor is involved twice in the equation, and with opposite behavior. FLEXIBILITY can be seen as an index of the increased capability to interact with other molecules, including biomolecules, and thus it can be directly related to toxicity, as indicated by the sign in the equation.

However, justifying model M_0 against the noise sensitivity characteristic in a second level of model validation shows that this model reliably describes only 28% of the target data (Descriptive Power [10]), and that it was built on a poorly sized data set, i.e., a high variables per samples ratio (65/92) relative to the data mining algorithm used. From [4] follows that this implies less confident results and an increased risk to model also noncausal correlations, which is underlined by the fact that only 5 descriptors are jointly included in both linear models, M_0 and the earlier GMDHM1 model. So, model reliability needs improvement. There are two options: Either adding additional samples – actually measured samples or samples obtained by Jittering, for example – or using less input variables. The latter can be realised by finding a nucleus of inputs, first (dimension reduction). Two aspects are important here: (1) Identifying a nucleus from a given set of variables is an ill-posed task. To make it regular, the objective of the final model must be considered in this step, already, for choosing an adequate dimension reduction tool. For example, if one is seeking a classification model that is composed of a certain set of input variables, a successive correlation analysis of a single input with the target variable may not help very much for finding that nucleus, because both objectives do not really correspond very close. (2) In the context of a knowledge discovery, nucleus identification and data

mining should embed seamlessly into the overall process, which finally means automation and workflow processing.

A new approach in that direction is based on setting up a GMDH Neural Network that in turn employs GMDH Neural Networks on different input data subsets as its Active Neurons for selecting a nucleus. If the identified nucleus is then passed to the data mining algorithm (GMDH again) on the fly, a multileveled self-organisation is implemented. In this way preprocessing actively and appropriately supports the final modelling objective as it, in fact, tightly connects preprocessing and data mining tasks into a single process.

We used this technology to create models M_1 to M_5 based on these input data: $D_1=\{\underline{x}, \underline{u}\}$, $D_2=\{\underline{x}, \underline{v}\}$, $D_3=\{\underline{x}, \underline{w}\}$, $D_4=\{\underline{x}, \underline{u}, \underline{w}\}$, $D_5=\{\underline{x}, \underline{v}, \underline{w}\}$. In a final level of self-organisation, a combined model was synthesised using GMDH on the input data set $D_C=\{\underline{x}, \underline{M}\}$. Table 1 lists the obtained results, and table 2 gives a qualitative overview on which descriptors are actually used in the models.

Table 1: Model errors of models M_1-M_5 and of the combined model

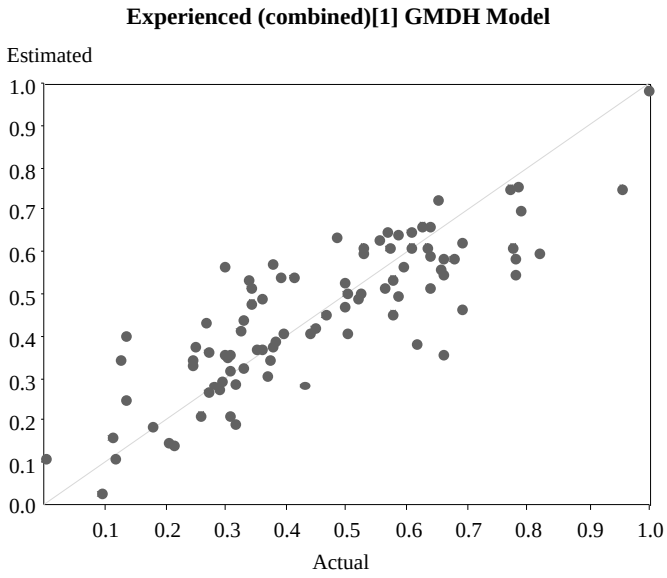
Model	Number of inputs: potential	selected	AEV	MSE	Descriptive Power [%]
M0	65	10	0.453	0.014	28
M1	65	7	0.471	0.014	28
M2	65	5	0.353	0.01	37
M3	68	4	0.419	0.012	32
M4	99	7	0.393	0.011	33
M5	99	7	0.317	0.009	39
Combined (M6)	39	8	0.296	0.008	43

Table 2: Descriptors actually used in a model

	homo	heat	polariz	volume	flex
M1		1	1	1	1
M2		1	1		1
M3	1	1	1		1
M4	1	1	1		1
M5	1	1	1		1
Combined (M6)	1	1	1		1
% used	**67**	**100**	**100**	**17**	**100**

	pH=2	Kappa2	Kappa3	kA2	kA3
M1	1			1	1
M2	1			1	
M3					
M4	1	1		1	
M5	1	1	1		
Combined (M6)	1	1	1	1	
% used	**83**	**50**	**33**	**67**	**17**

A plot of the combined model is shown in figure 1. The uncertainty expressed by different estimations of models M_1-M_6 is displayed in figure 2 for each of the 92 compounds of the



training data. The prediction interval for 12 aromatic compounds having an unknown carcinogenicity value is shown in figure 3. The predictions are obtained using models M_1-M_6.

Figure 1: Experienced vs. estimated carcinogenicity values of the final linear model M_6

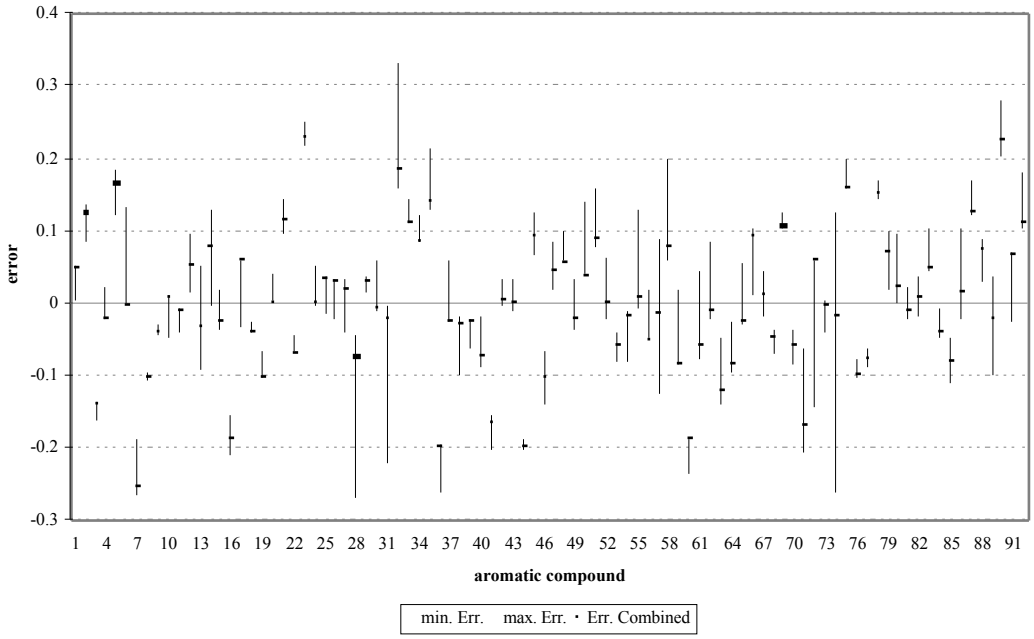


Figure 2: Predictive uncertainty based on models M_1 to M_6

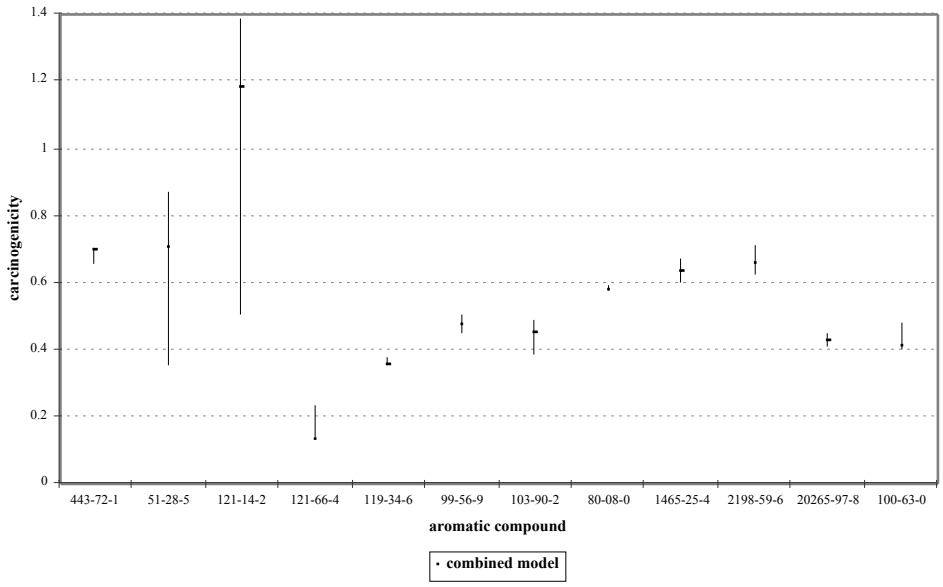


Figure 3: Prediction interval for 12 compounds obtained from applying models M_1 to M_6

3. Summary

We applied self-organising data mining technology to model, describe, and predict carcinogenicity of aromatic compounds. Using a new algorithm of a state space dimension reduction to identify a nucleus of relevant descriptors and, for the first time, a noise sensitivity characteristic for final model evaluation, it was possible to increase both descriptive and predictive power of models. A comprehensive set of 6 descriptors is used in more than 50% of the models. However, only up to 43% of the carcinogenicity is described by the models, reliably, which is a not very satisfactory rate. Generating nonlinear models may help increasing the power of models even more, but, with respect to noise filtering, it is open to which extent. This is up to the future once second-level model evaluation for nonlinear models is implemented into KnowledgeMiner. Another possible source of existing models' uncertainty is missing true relevant descriptors. Adding new descriptors to the vector of potential inputs could result in a more sufficient description of the system's input-output behaviour and thus in more powerful predictive models.

References

[1] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery: An Overview. In: U.M. Fayyad (ed.): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press. Menlo Park, California, 1996, pp. 1-36

[2] J.-A. Mueller, F. Lemke, Self-Organising Data Mining. Extracting Knowledge From Data. Libri/BoD, Hamburg, 2000

[3] J.-A. Mueller, Self-organising Data Mining, Proceedings ICIM 2002, Lviv, 2002

[4] F. Lemke, J.-A. Mueller, Validation in Self-organising Data Mining, Proceedings ICIM 2002, Lviv, 2002

[5] G. Gini, M. Lorenzini, E. Benfenati, P. Grasso, M. Bruschi, Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. In: Journal of Chemical Information and Computer Sciences, 39(1999)6, pp. 1076-1080

[6] Data set generated within the COMET project, ENV4-CT97-0508, funded by the European Comission.

[7] http://www.knowledgeminer.net

[8] http://www.knowledgeminer.net/pdf/carcino.pdf

[9] J.-A. Mueller, Knowledge extraction from data based on GMDH. USIM, ...

[10] F. Lemke, Does my model reflect a causal relationship? http://www.knowledgeminer.net/isvalid.htm