The Importance of Scaling in Data Mining for Toxicity Prediction

Paolo Mazzatorta and Emilio Benfenati*

Istituto di Ricerche Farmacologiche "Mario Negri" Milano, Via Eritrea, 62, 20157 Milano, Italy

Daniel Neagu and Giuseppina Gini

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy

Received March 26, 2002

While mining a data set of 554 chemicals in order to extract information on their toxicity value, we faced the problem of scaling all the data. There are numerous different approaches to this procedure, and in most cases the choice greatly influences the results. The aim of this paper is 2-fold. First, we propose a universal scaling procedure for acute toxicity in fish according to the Directive 92/32/EEC. Second, we look at how expert preprocessing of the data effects the performance of qualitative structure−activity relationship (QSAR) approach to toxicity prediction.

INTRODUCTION

Preprocessing is the first step in predictive data mining (i.e., data mining with the goal of constructing a predictive model). It includes several operations on the data set before a statistical model is developed. Preprocessing can be based on a priori knowledge about the data, on assumptions underlying the statistical model, or on practical experience. Such manipulations can take many forms reflecting different modeling objectives and levels of knowledge about the properties of the data.[9,14,24]

In modeling, a proper selection of the transformation of a response variable implies a number of benefits. Notably, it may (i) simplify the response function by linearizing a nonlinear response-factor relationship; (ii) stabilize the variance; and (iii) make the distribution more normal.

Many quantitative structure−activity relationship (QSAR) methods require scaling of the original data to extract significant and useful information and remove unimportant, not interesting features. We used a QSAR approach based on artificial neural networks (ANN), to predict acute toxicity, i.e., the lethal concentration for 50% of the test animals (LC_{50}), for the fathead minnow (*Pimephales promelas*).

This approach employs a powerful pattern recognition paradigm, able to analyze various types of data,[1,6,15] but all the data must be scaled between zero and one. Scaling the descriptors is a very delicate procedure because we do not know the underlying relationship between the descriptor and the toxicity for most of them,[3,4] and therefore cannot foresee the influence of these manipulations. We therefore maintained the original distribution, using a range scaling in order to conserve it.

For the LC_{50}, we based ourselves on the EU Directive 92/32/EEC annex VI point 5.1,[10] which classifies the chemicals as shown in Table 1:

Table 1. EC Classification for Fish (Directive 92/32EEC Annex VI Point 5.1)

class	LC_{50}	dangerous for 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

The classification is clearly based on a logarithmic scale. This directive gives useful guideline for scaling acute toxicity. Furthermore, to have a useful general instrument, the scaling procedure must go beyond the limits of the data set mined. Because there is a natural lower limit, 0 mg/L, but not a upper one, the only solution was to have a function between 0 and 1 with an asymptote to 1. Thus every possible real value for the LC_{50} is represented. The inevitable loss of knowledge about the highest values is acceptable because they are in the less toxic class and because, according to the EU directive, less precision is required on high values.

This paper presents a scaling procedure for LC_{50} that takes account of the EU regulation, and the need for a universal approach, useful for aquatic toxicity, is presented in this paper.

MATERIALS AND METHODS

Data Set. We mined a data set of 554 organic compounds, commonly used in industrial processes. The U.S. Environmental Protection Agency[11−13,28] helped build up this data set, starting from a review of experimental data in the literature, referring to acute toxicity 96-h LC_{50}, for the fathead minnow (*Pimephales promelas*). Close analysis of a large amount of experimental information led to the association of a mechanism of action (MOA) for each compound. This toxicological data in this set is one of the biggest available and very reliable,[28] and therefore the information extracted from it has a general validity.

* Corresponding author phone: +392 39014420; fax: +392 39001916; e-mail: benfenati@marionegri.it.

10.1021/ci025520n CCC: $22.00
Published on Web 08/20/2002

Table 2. Statistical Information about the Toxicity Values in the Data Set

	value (mg/L)
maximum	75200.00
minimum	0.00019
range	7.5200e+004
standard deviation	5.7249e+003
variance	3.2774e+007
geometric mean	24.1313
arithmetic average	1.0600e+003

Table 3. Descriptors Selected for the Models

descriptor	code
total energy (kcal/mol)	QM1
heat of formation (kcal/mol)	QM3
LUMO (eV)	QM6
relative number of N atoms	C9
relative number of single bonds	C24
molecular weight (amu)	C35
Kier&Hall index (order 0)	T6
average information content (order 1)	T22
moment of inertia B	G2
molecular volume	G10
molecular surface area	G12
TMSA total molecular surface area [Zefirov's PC]	E13
FPSA-2 fractional PPSA (PPSA-2/TMSA) [Zefirov's PC]	E24
PPSA-3 atomic charge weighted PPSA [Zefirov's PC]	E28
FPSA-3 fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]	E31
LogD pH9	pH9
LogP	LogP

The data set was randomly partitioned 70−30% between 388 training cases, used to develop the models and 166 testing cases, and used to evaluate the prediction ability of the models. Statistical information about the data set mined is summarized in Table 2.

Descriptors. The descriptors were calculated using different software: Hyperchem 5.0 (Hypercube Inc., Gainesville, FL, U.S.A.), CODESSA 2.2.1 (SemiChem Inc., Shawnee, KS, U.S.A.), and Pallas 2.1 (CompuDrug; Budapest, Hungary). Out of the hundreds of descriptors proposed by these software just 153 gave a nonconstant or nonmissing value for all the objects. The set of descriptors resulting can be split into six categories according to the classification present in the software CODESSA:[20] constitutional (34), geometrical (14), topological (38), electrostatic (57), quantum-chemical (6), and physicochemical descriptors (4).

To ensure a good model it is useful to select the variables that describe the molecules best. Some of these descriptors add no information and just increase the noise, making it more difficult to analyze the results. Furthermore, with a limited number of variables the risk of overfitting is reduced.[17,29] The descriptor selection was done through principal components analysis (PCA) with the principal components eigenvalue (Scree) plot method. Descriptors with the highest scores on the first four components of PCA were chosen and reduced, eliminating those most closely correlated. A final criterion was to keep a pool of descriptors representing the different aspects of the molecule considered (physicochemical, electronic, and topological, etc.). The selected descriptors are listed in Table 3.

NIKE. ANN are powerful tools, used to develop a wide range of real-world applications, especially when traditional solving methods fail.[2,5,25] They offer advantages such as ideal learning ability from data, classification capabilities and generalization, computational speed once trained, with parallel processing, and noise tolerance. The major shortcoming of neural networks is their low transparency.

In developing ANN models we used the hybrid intelligent system shell NIKE,[22,23] in order to automate the processes involved, from data representation for toxicity measurements, to the prediction of toxicity for a given new input. We used the first NIKE module, called IKM-CNN (implicit knowledge module-based on crisp neural networks), which takes charge of models the data set as a multilayer perceptron (MLP).[27] For a chosen IKM-CNN, the descriptors are considered the inputs for the neural nets. IKM-CNN are described as MISO (multi-input single output) structures, having as output the normalized value of toxicity (LC_{50}). The neural nets are trained on the specific data and adjusted using the number of the hidden layer neurons and momentum term.

Scaling. We scaled all descriptors between zero and one, using a range scaling. However toxicity was scaled with the following six scaling procedures:

- Range Scaling (RS):

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where y_i = scaled value, x_i = original value, $min(x)$ = minimum of the collection of x objects, and $max(x)$ = maximum of the collection of x objects.

This is the most common approach because it is simple and keeps the linear distribution of the data.

- Range-Logarithmic Scaling (RLS):

$$y_i = \frac{\log_{10}(x_i) - \min(\log_{10}(x))}{\max(\log_{10}(x)) - \min(\log_{10}(x))} \tag{2}$$

Logarithmic transformation is generally used when the coefficient of variance is constant.

- Range-Logarithmic Scaling Modified (RLSM):

$$y_i = \frac{\log_{10}(x_i + 1) - \min(\log_{10}(x + 1))}{\max(\log_{10}(x + 1)) - \min(\log_{10}(x + 1))} \tag{3}$$

This is a modification of the previous transformation, it uses $\log_{10}(x_i + 1)$ to overcome to limit of $\log_{10}(x_i)$ when $x_i = 0$.

- Tangent Hyperbolic Scaling (THS):

$$y_i = \tanh(x_i) \tag{4}$$

This transformation is introduced to extrapolate the model beyond the data set limits, because it tends asymptotically to 1.

Tangent Hyperbolic-Logarithmic Scaling (THLS):

$$y_i = \tanh(\log_{10}(x_i + 1)) \tag{5}$$

This is intended to combine the efficiency of stabilizing the variance and generalizing the data set, given by the previous transformations.

Tangent Hyperbolic − Logarithmic Scaling Modified (THLSM):

$$y_i = \tanh(0.4903\log_{10}(x_i + 1) + 0.0562) - 0.0095 \tag{6}$$

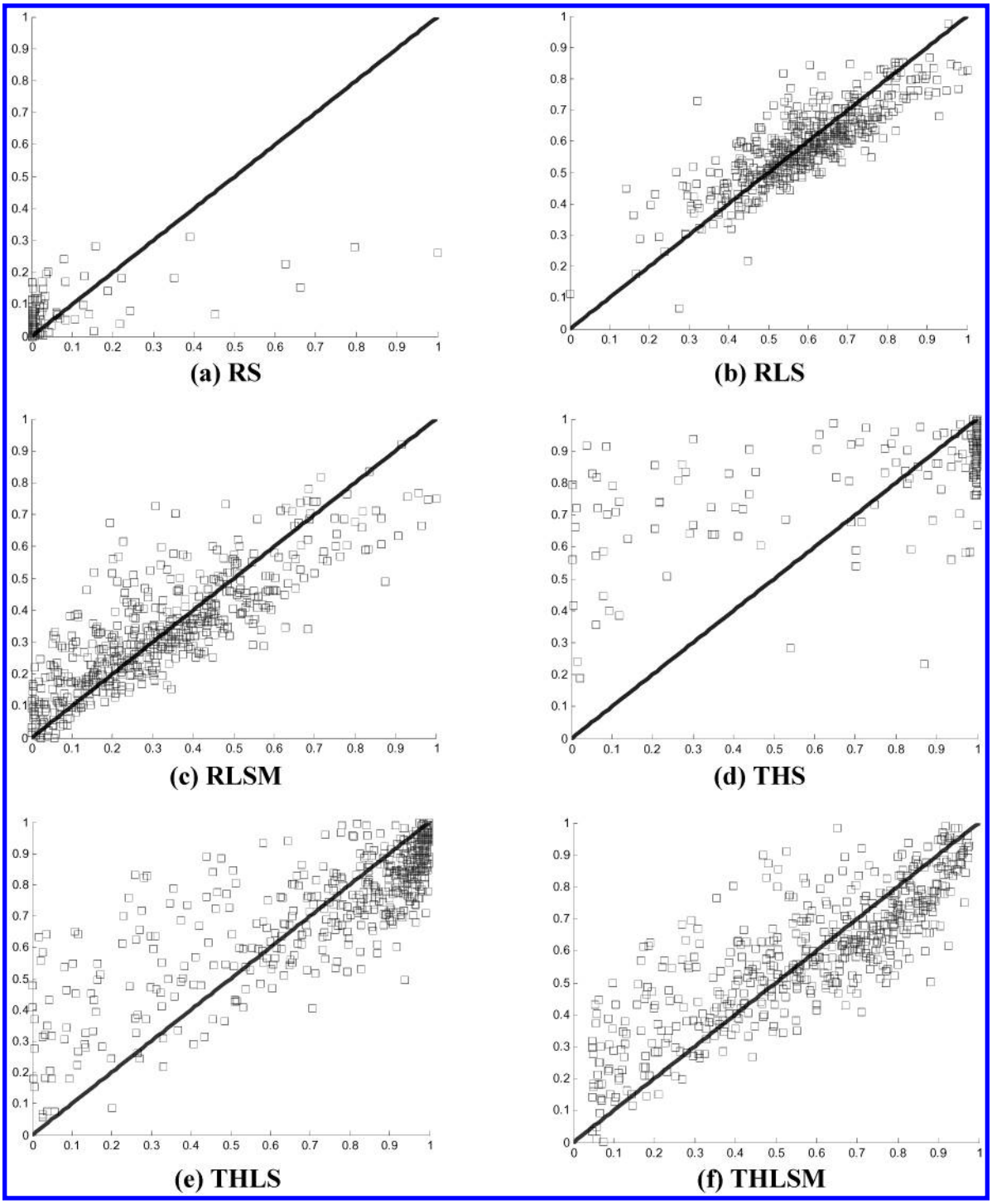

Figure 1. Performance validation of the 554 chemicals for: (a) range scaling; (b) range-logarithmic scaling; (c) range-logarithmic scaling modified; (d) tangent hyperbolic scaling; (e) tangent hyperbolic-logarithmic scaling; and (f) tangent hyperbolic-logarithmic scaling modified. On the x-axes there are the real values of the scaled LC_{50} and on the y-axes there are the predicted value.

This is a modification of the previous algorithm in order to achieve the best application to the EC classification.[4]

RESULTS AND DISCUSSION

We used a fully connected three-layered crisp neural network with 25 hidden neurons, according to the general rules on the maximum and minimum number of hidden neurons.[16,18,21] The back-propagation algorithm with a momentum term of 0.9 was used for training (a high momentum term prevents too many oscillations of the error function,[27] and small variations around 0.9 did not suggest in our case high differences). The networks were trained up to 5000 epochs.

Figure 1 shows the predictive ability of models developed with the six different scaling procedures. The preprocessing operations have very strong influence. Besides the real performance of the model, uncritical scaling makes it extremely difficult to extract information from the data set.

Table 1 shows statistical indices of the models developed. R^2 is the explained variance by the model and represents the ability of the model in correlating the descriptors and the experimental answer. Mean squared error (MSE) gives

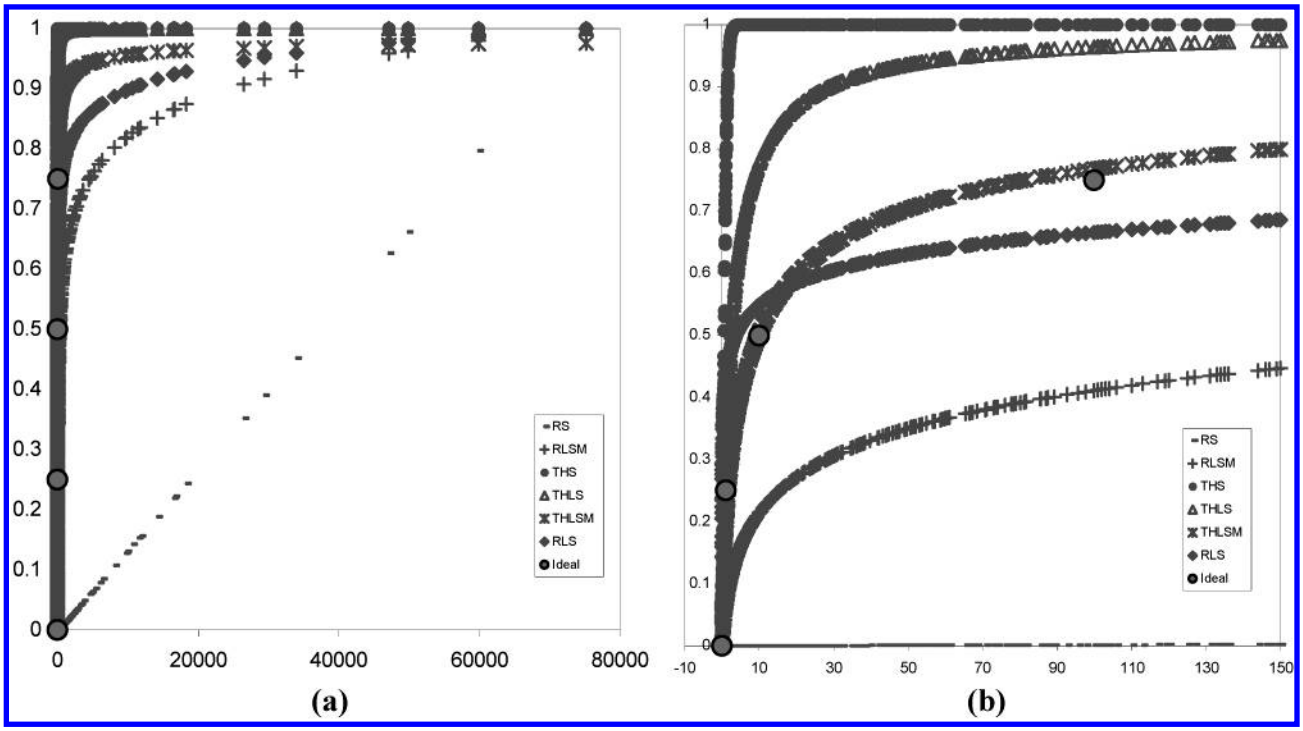

Figure 2. Scaling algorithms in the whole range (a) and in the significant interval [0−150 mg/L] (b). The black dots in (b) are landmarks for the ideal transformation. On the x-axes there are the original values and on the y-axes there are the scaled values.

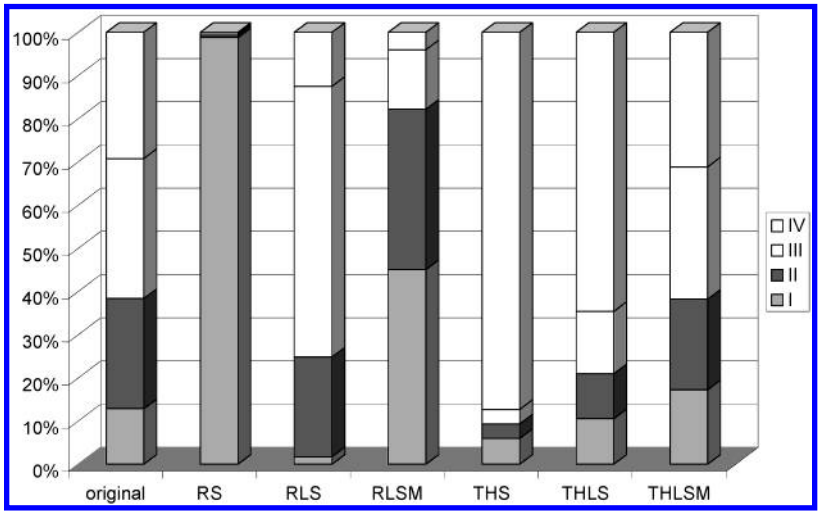

Figure 3. Percentage distribution of total number of compounds (554) within the EC toxicological classes for the original data set (first column) and in four classes of the same width after each scaling procedure.

a general information of the performances of the model on the data set mined.

RS (Figure 1a) is not able to manage the high variance of the data. In view of the fact that very few data have high values, it concentrates most of the other data in a small interval and loses important information about them. Worse, it loses information about the most toxic class of compounds, which is, of course, the most important.

RLS (Figure 1b) is a good example of preprocessing. The objects are well distributed in the whole interval and the model takes advantage of this. The weakness of this transformation lies in the presence of limits which restrict its application on the data set considered. In fact, it needs a min and max value to be computed so it is limited between these values.

RLSM (Figure 1c) has a better distribution in the whole interval, and it results in a better ability of the model in describing the relationship between the inputs and the output (R^2) but has worst mean error (Table 4).

Table 4. Statistical Indices of the Models Developed

	RS	RLS	RLSM	THS	THLS	THLSM
R^2	0.3165	0.6807	0.6813	0.4962	0.6615	0.6875
MSE	0.003954	0.006884	0.014482	0.028526	0.028659	0.021617

To overcome the problem of the limits of the data set we used a new approach. THS (Figure 1d) responds to our requirement for a generalizable manipulation but has a strong negative effect on the data distribution, with obvious consequences on the model performances; indeed, once again most of the data are compressed in one area.

Using logarithmic transformation first, to keep to the EU guidelines, and then using tangent hyperbolic scaling in order to make it generalizable, we tried to overcome both these

hurdles. There is a useful improvement, but it is not yet the best distribution (Figure 1e).

The idea to modify the previous transformation comes directly from the EU classification. We had a relatively good solution (THLS) which only needed to fit on the ideal distribution given by the directive. We used a nonlinear curve-fitting solver in the least squares sense (lsqcurvefit).[19] That is, given input data *xdata* and the observed output *ydata*, find coefficients *x* that "best fit" the equation *F(x, xdata)*. Lsqcurvefit uses the large-scale algorithm, a subspace trust region method based on the interior-reflective Newton method described in refs 7 and 8. Each iteration involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients (PCG)

$$\min_x \frac{1}{2}|F(x,xdata) - ydata|_2^2 = \frac{1}{2}\sum_i (F(x,xdata_i) - ydata)^2 \quad (7)$$

where *xdata* is the vector of the class limits given by the EC, *ydata* is the vector of the best ideal distribution, and *F(x, xdata)* is the vector valued function. In this case the algorithm finds the coefficients *x* that "best fit" the equation THLSM:

$$xdata = [0; 1; 10; 100; \text{inf}]$$

$$ydata = [0; 0.25; 0.5; 0.75; 1]$$

$$F(x,xdata) = \tanh(x_1 \log_{10}(xdata + 1) + x_2) + x_3 \quad (8)$$

CONCLUSIONS

The strong effect of scaling is easily understandable in the following figure (Figure 2a,b), which shows the distribution of the data after each transformation in the whole interval of the data set (Figure 2a) and just for the significant interval [0−150 mg/L], where all the most toxic classes are concentrated (Figure 2b). The *ideal* transformation is the one that succeeds in scaling the original toxic classes into classes of the same width so that each transformed class has the same accuracy and the same original variance.

Figure 2 shows that RS does not describe the data set reliably. This transformation forces almost every object (99%) into a very small interval (0−0.25), losing important information. THS behaves much the same way, but this scaling puts most of the data (87%) in the last interval [0.75−1]. RLS assigns too few elements (less than 2%) to the first class. RLSM and THLS are easily understood from Figure 2b, in the light of the previous reasoning. They have a better distribution than the previous transformations but are far from ideal. THLSM is the scaling that best fits the characteristics of the *ideal* transformation.

Further considerations can be drawn from Figure 3. The distribution of the objects into four classes of the same width after every transformation is compared with the original distribution of the data set according to the EU classification.

Once more THLSM gives the best approximation. Although this is not the most important characteristic—RLS has quite a different distribution but still gives good models—this is an indicator of the reliability of this scaling technique.

ACKNOWLEDGMENT

This work is partially funded by the EU under contract HPRN-CT-1999-00015. The authors wish to express their appreciation to the reviewers and to thank Prof. A. R. Katritzky (Gainesville, FL) and Prof. M. Karelson (Tartu, Estonia) for the use of CODESSA.

REFERENCES AND NOTES

(1) Bate, A.; Lindquist, M.; Edwards, I. R.; Olsson, S.; Orre, R.; Lansner, A.; De Freitas, R. M. *Eur. J. Clin. Pharmacol.* **1998**, *54*, 315−321.
(2) Becraft, R.; Lee, P. L.; Newell, R. B. Integration of Neural Networks and Expert Systems for Process Fault Diagnosis. In *Proceedings of the International Joint Conference on Artificial Intelligence*; 1991; pp 832−837.
(3) Benfenati, E.; Grasso, P.; Pelagatti, S.; Gini, G. *On Variables and Variability in Predictive Toxicology*; IV Girona Seminar on Molecular Similarity, 5−7 July 1999, Girona, Spain.
(4) Benfenati, E.; Piclin, N.; Roncaglioni, A.; Varì, M. R. Factors influencing predictive models for toxicology. *SAR QSAR Environ. Res.* **2001**, *12*, 593−603.
(5) Bishop, C. M. *Neural networks for pattern recognition*; Clarendon Press: Oxford, 1995.
(6) Chastrette, M.; Cretin, D.; Aidi, C. E. *J. Chem. Inf. Comput. Sci.* **1995**, *36*, 108−113.
(7) Coleman, T. F.; Li, Y. An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM J. Optimization* **1996**, *6*, 418−445.
(8) Coleman, T. F.; Li, Y. On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds. *Mathematical Programming* **1994**, *67*(2), 189−224.
(9) Cuesta Sánchez, F.; Lewi, P. J.; Massart, D. L. Effect of different preprocessing methods for principal component analysis applied to the composition of mixtures: Detection of impurities in HPLC—DAD. *Chemometrics Intelligent Lab. Systems* **1994**, *25*, 157−177.
(10) Directive 92/32/ECC, the seventh amendment to Directive 67/548/ECC, OJL 154 of 5.VI.92; 1992; p 1. www.eurunion.org/legislat/chemical.htm.
(11) ECOTOX, *ECOTOXicology Database System, Code List*; prepared for U.S. Environmental Protection Agency, Office of Research, Laboratory Mid-Continent Division (MED), Duluth, MN, by OAO Corporation, Duluth, MN, February 2000.
(12) ECOTOX, *ECOTOXicology Database System, Data Field Definition*; prepared for U.S. Environmental Protection Agency, Office of Research, Laboratory Mid-Continent Division (MED), Duluth, MN, by OAO Corporation, Duluth, MN, February 2000.
(13) ECOTOX, *ECOTOXicology Database System, User Guide*; prepared for U.S. Environmental Protection Agency, Office of Research, Laboratory Mid-Continent Division (MED), Duluth, MN, by OAO Corporation, Duluth, MN, February 2000.
(14) Famili, A.; Wei-Min, S.; Weber, R.; Simoudis, E. Data Pre-processing and Intelligent Data Analysis. *Intelligent Data Analysis* **1997**, *1*, 3−23.
(15) Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; Bruschi, M. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1076−1080.
(16) Gori, M.; Scarselli, F. Are multilayer perceptrons adequate for pattern recognition and verification? *IEEE Trans. Pattern Anal. Machine Intelligence* **1998**, *20/11*, 1121−1131.
(17) Hasegawa, K.; Kimura, T.; Funatsu, K. GA strategy for variable selection in QSAR studies: application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 112−120.
(18) Hecht-Neilson, R. *Neurocomputing*; Addison-Wesley Pub. Co.: 1990.
(19) http://www.mathworks.com/access/helpdesk/help/toolbox/optim/lsqcurvefit.shtml.
(20) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Comprehensive Descriptors for Structural and Statistical Analysis*; Reference Manual, version 2.0, Gainesville, FL, 1994.
(21) Kurkovà, V. Kolmogorov's theorem is relevant. *Neural Computational* **1991**, *2/4*, 617−622.
(22) Neagu, C.-D.; Palade, V. Neural Explicit and Implicit Knowledge Representation. In Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies KES2000, 30th & 31st Aug, 1st Sept 2000; University of Brighton, Sussex, U.K, 2000; pp 213-216.
(23) Neagu, C.-D.; Palade, V. Modular neuro-fuzzy networks used in explicit and implicit knowledge integration. The 15th International

FLAIRS-02 Conference, Special Track on Integrated Intelligent Systems; AAAI Press: Pensacola, FL, 2002; in press.
(24) de Noord, O. E. The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics Intelligent Laboratory Systems* **1994**, *23*, 65−70.
(25) Pal, S. K.; Mitra, A. Multilayer Perceptron, Fuzzy Sets, and Classification. *IEEE Trans. Neural Networks* **1992**, *3*(5), 683−697.
(26) Piclin, N.; Pintore, M.; Ros, F.; Benfenati, E.; Chrétien, J. R. *Data Base Mining and Prediction of Pesticide Toxicity.* Presented at Chimiométrie 2000, 6−7 December, 2000, Paris, France.
(27) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing, Explanations in the Microstructure of Cognition*; MIT Press: 1986.
(28) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister D. E.; Drummond S. J. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (pimephales promelas). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(29) Xu, L.; Zhang, W.-J. Comparison of different methods for variable selection. *Anal. Chim. Acta* **2001**, *446*, 475−481.

CI025520N