Genetic Algorithm driven Clustering for Toxicity prediction

Dirk Devogelaere, Patrick Van Bael, Marcel Rijckaert

K.U.Leuven, Chemical Engineering Department
De Croylaan 46, B-3001 Heverlee,Belgium
Email: {dirk.devogelaere}@cit.kuleuven.ac.be
Tel: +16 32 27 07

SUMMARY

The pace of technological advancement in today's society has generated an enormous demand for methods facilitating the intelligent testing for the toxicity of new chemicals. Until now it is common use to make prediction based on 'real' tests. Recent investigations support the general assumption that macroscopic properties like toxicity and ecotoxicity strongly depend on microscopic features and the structure of the molecule.

This paper's authors have developed a computationally intelligent method for supervised training of regression systems, named GAdC (Devogelaere, 1999). Our method shall select those features needed to predict the toxicity and calculate the toxicity. The proposed methodology relies on supervised clustering with genetic algorithms and local learning.

The basis for our investigations is a set of 164 pesticides from seven different chemical classes with data on acute toxicity for rainbow trout, daphnia magna, etc …The concentrations for this aquatic toxicity are given in two representations, LC_{50} and $-\log_{10}(LC_{50}/(\mathrm{mmol/l}))$. 174 molecular descriptors such as constitutional and topological descriptors, electrostatic and quantum-chemical descriptors and others, which are partly continuous, partly discrete values, were calculated (Benfenati, 1999) for each of these 164 pesticides. The descriptors showing missing values are omitted in this first step of investigation. The use of 156 descriptors and only 164 data-points means that this is a difficult regression problem (high dimension with only few points). Therefore, we decided to use the leave-one-out crossvalidation for the prediction of the toxicity. This ensures the maximal possible statistical security by testing every output independently from all others.

In the GAdC, the first part of the chromosome contains the scalars, the second part the centers of the clusters. A scalar has a value between –1 and +1 depending on the importance of the descriptor for the prediction. Each center of a cluster contains a value (between minimum and maximum value of this descriptor) for each descriptor. One of the main advantages of this method is to be able to see which descriptors are important. This can easily be calculated by multiplying the positive value of the scalar with the positive mean value of the corresponding descriptor. The higher this Descriptor importance value (Div), the more important the corresponding descriptor to predict the toxicity. The Div of each descriptor is plotted in figure 1. From this figure its quite clear that this values decreases fast which means that you only need a part of the descriptors to build a good prediction model for the toxicity. This is illustrated by means of the cumulative curve.

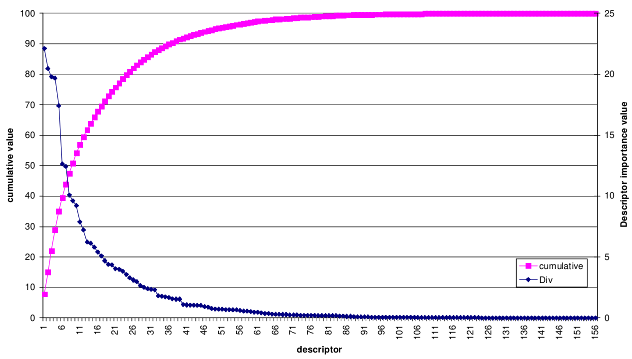

Figure 1: Importance of the descriptors

Acknowledgments

Part of this work has been supported by the Commission of the European Communities under the Program “Environment and Climate”, Project “COMET”, Contract No. ENV4-CT97-0508.

References

E. Benfenati, S. Pelagatti, P. Grasso, and G. Gini. COMET (1999), Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools, *AAAI 1999 Spring* Symposium *Series;* AAAI Press, Menlo Park, CA, 40-43.

D. Devogelaere, P. Van Bael, and M. Rijckaert (1999). Regression Through Genetic Algorithm driven Clustering, Proceedings *of the 1999 European Conference on Intelligent Techniques and Soft Computing (EUFIT),* Sept. 7-10, Aachen, Germany.