Development of Nonlinear Quantitative Structure-Activity Relationships using RBF Networks and Evolutionary Computing

Panagiotis Patrinos, Alex Alexandridis, Andreas Afantitis, Haralambos Sarimveis* and Olga Igglesi-Markopoulou
National Technical University of Athens, School of Chemical Engineering 9, Heroon Polytechniou Str. Zografou Campus, Athens 15780, Greece

Abstract

Quantitative Structure Activity Relationships (QSARs) are mathematical models that correlate structural or property descriptions of compounds (hydrophobicity, topology, electronic properties etc.) with activities, such as chemical measurements and biological assays. In this paper we propose a modeling methodology suitable for QSAR studies which selects the proper descriptors based on evolutionary computing and finally produces Radial Basis Function (RBF) neural network models. The method is successfully applied to the benchmark Selwood data set.

Keywords: Radial Basis Functions, QSAR, Neural Networks, Evolutionary Computing, Genetic Algorithms, Simulating Annealing

1. Introduction

One of the future challenges in Process Systems Engineering is the discovery and design of new products, by studying and analyzing the molecular level (Grossmann, 2003). Quantitative Structure Activity Relationships (QSARs) constitute an effort towards this direction by relating biological activities of compounds with their molecular structure. QSAR models are particularly useful in computer assisted molecular design, since they provide tools for predicting important activities of new molecules, before they are actually synthesized. In this manner valuable experimental time is saved and the procedure for designing new materials, especially drugs, is accelerated.

Given that for a set of compounds a number of experimental physicochemical parameters are provided and the respective biological activities have been determined, QSAR development is mathematically transformed into a modeling problem based on input-output examples. In choosing the proper method for solving that problem, two important issues must be taken into account: the first is that the relationships to be discovered are often nonlinear and the respective functional transformations are not known *a priori*. The second issue is that the available samples may be limited, so that in many cases the number of descriptors is greater compared to the number of compounds

* Author to whom correspondence should be addressed : hsarimv@central.ntua.gr

in the data set. Moreover, some of the descriptors that are considered as inputs may not have a significant influence on the activities that need to be estimated.

In the early stages, QSARs were simple regression models containing only few descriptors (mostly electronic or thermodynamic parameters). In the past few years the continuous increase in the performance of computer processors motivated the researchers to introduce a wide range of new descriptors in QSAR models that are more indicative of molecular structure and topology. Additionally, new modeling techniques were employed such as principal component analysis (PCA), partial least squares (PLS) and neural networks (So & Richards, 1992). As far as the selection of meaningful descriptors is concerned, a variety of optimization methods have been applied. Due to the inherent combinatorial nature of the problem, the majority of these methods are based on stochastic search techniques, such as evolutionary programming (Luke, 1994), genetic function approximation (Rogers & Hopefinger, 1993), simulated annealing (Sutter et al., 1995), etc.

In this work, we present a novel methodology for QSAR development, which combines the advantages of several advanced modeling technologies. More specifically, the Radial Basis Function (RBF) neural network architecture serves as the nonlinear modeling tool, by exploiting the simplicity of its topology and the fast fuzzy means training algorithm (Sarimveis et al., 2002). The proper descriptors are selected in two stages: In the first stage a specially designed genetic algorithm minimizes the cross validation error regardless of the number of descriptors, while in the second stage a simulated annealing technique aims at the reduction of the number of descriptors.

The performance of the method is illustrated through the application to the benchmark Selwood data set (Selwood et al., 1990). We obtained very successful results which are due to the ability of RBF networks to approximate any nonlinear relation and the successful selection of variables, which is achieved by the two evolutionary computation techniques.

2. The proposed methodology

Variable selection problems can be classified according to their size. Most of the problems that appear in QSAR analysis are medium or large-scale (Kudo & Sklansky, 2000), since the total number of available descriptors is usually more than 20. The GASA – RBF algorithm (acronym of the words Genetic Algorithm Simulated Annealing Radial Basis Function) that is presented in this work is specifically designed in order to take care of the large number of descriptors that appear in this special class of variable selection problems. To be more specific, the proposed approach decomposes the variable selection procedure into two stages:

1st stage: Minimization of the prediction error using a GA

Firstly, it is assumed that a set of input – output examples [**X**, **y**] is available. The dimensionality of the input matrix **X** is K x N, where K denotes the number of the compounds in the data set and N is the total number of descriptors, while **y** is a K x 1 vector containing the activities of the compounds.

The objective of the first stage is to optimize the prediction error, regardless of the number of selected descriptors. This is accomplished by employing a specifically designed GA, which uses a hybrid coding of genes containing both binary and integer values. The length of each chromosome is equal to the total number of descriptors N plus one. Binary coding is used to denote whether a descriptor is present in the model (the gene has the value 1) or not (the gene has the value 0), while the integer coding of the last gene of the chromosome denotes the number of fuzzy sets used by the fuzzy means training algorithm.

The GA starts by randomly generating an initial population of P chromosomes and proceeds with the following cross validation procedure for computing the fitness of each chromosome: Using only the descriptors that are represented by 1s in the binary genes and the number of fuzzy sets contained in the last gene of the chromosome, RBF models are developed, by excluding one data point $[\mathbf{x_i}\, y_i]$ at a time from the training data set. Each of the above models is employed to predict the corresponding y_i-th output of the removed compound. When this procedure is completed the cross-validated root mean square error (RMSECV) for each chromosome is calculated by:

$$\mathrm{RMSECV}_j^{\mathrm{GA}} = \sqrt{\frac{\sum_{i=1}^{K} (\hat{y}_{(-i),j} - y_i)^2}{K}}\ , j=1,2,\ldots,P \qquad (1)$$

where $\hat{y}_{(-i),j}$ is the prediction of the j-th network for the i-th example that has been removed from the training data.

Based on the calculated fitness of the chromosomes, the algorithm proceeds with the natural selection process, which must assure the reproduction of the fittest chromosomes in the next generation. The reproduction is implemented as a linear search through a roulette wheel (Michalewitz, 1996). Then, two genetic operators are applied: the typical one-point crossover operator and two different types of mutation. Uniform flip bit mutation is applied to the binary genes with probability equal to p_{um}, while non-uniform mutation (Michalewitz,1996) with probability equal to p_{num} is used for the genes that represent the number of fuzzy sets. The genetic algorithm is iteratively applied until a specific number of iterations has been completed or a desired minimum error has been achieved.

2nd stage: minimization of the number of input variables using GSA

As shown in the descriptive flowchart of the second stage (figure 1), the outcome of the first stage is a chromosome $\mathbf{s}_b^{\mathrm{GA}}$ which defines the RBF model that achieves the minimum prediction error $\mathrm{RMSECV}_b^{\mathrm{GA}}$. The chromosome $\mathbf{s}_b^{\mathrm{GA}}$ consists of a sequence of N binary variables $\mathbf{v}_b^{\mathrm{GA}}$ that shows the descriptors that are used by the produced model and an integer gene l_b^{GA}, which is the optimal number of fuzzy sets in the fuzzy means algorithm. However it is possible that the same or an almost equal prediction error can be achieved using a reduced number of descriptors. The objective of the second stage is to investigate this possibility, using a simulated annealing algorithm.

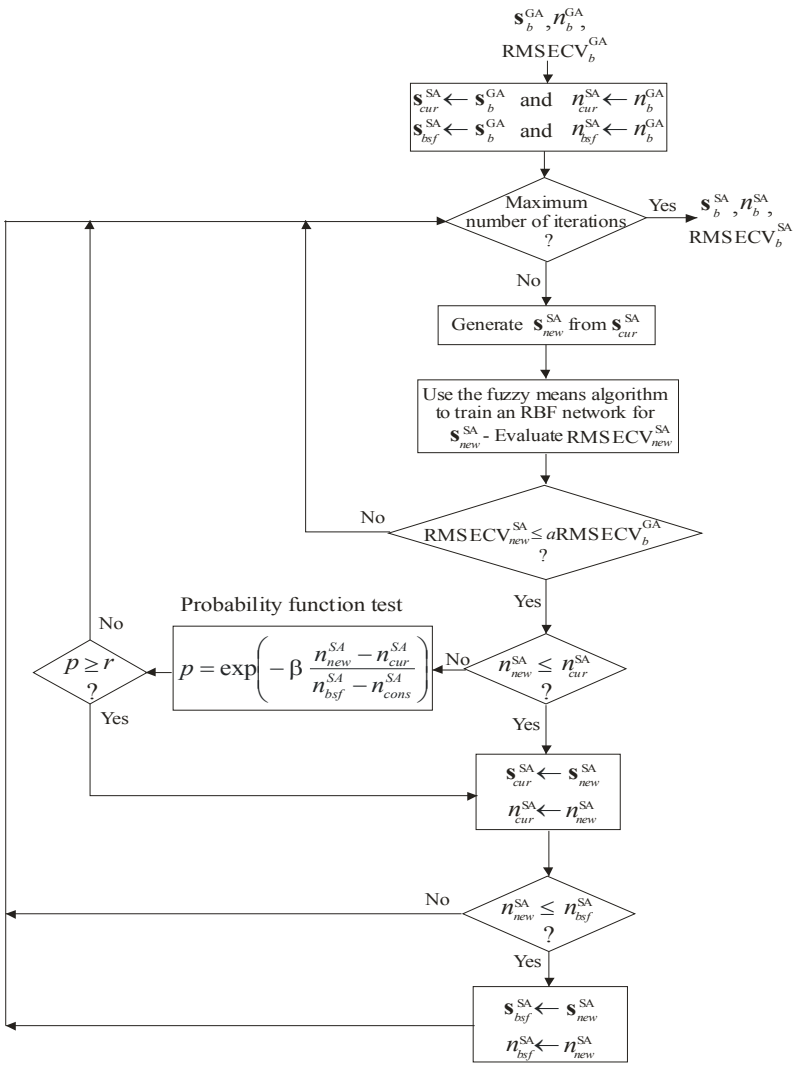

Figure 1. A flowchart of the simulated annealing algorithm

During each iteration the algorithm executes a perturbation in the current subset of variables $\mathbf{v}_{cur}^{SA}$, by flipping some of the binary values in the current solution. Thus, a new subset of variables $\mathbf{v}_{new}^{SA}$ is created. During the first iterations the maximum number d_c of digits that can change is given relatively large values, but as the algorithm progresses this number reduces, so that in the last iterations only one flip is allowed. The last gene of the chromosome l_{cur}^{SA} can also be modified to l_{new}^{SA} using non-uniform mutation with probability p_{num}, thus producing a new solution $\mathbf{s}_{new}^{SA}$.

In order to evaluate the new solution, the algorithm checks whether its RMSECV is equal or smaller than a threshold value $RMSECV_t$, which is produced by multiplying

the output of the first stage $\text{RMSECV}_b^{\text{GA}}$ by a constant a that is greater than unity. In other words, some of the accuracy of the network predictions is sacrificed, in order to achieve a considerable reduction in the number of selected descriptors. In the case that $\text{RMSECV}_{new}^{\text{SA}}$ is equal or smaller than the threshold value, the new solution $\mathbf{s}_{new}^{\text{SA}}$ is forwarded for further processing and is finally adopted if it reduces the number of selected descriptors or passes the probability function test (see figure 1). Otherwise it is rejected and a different random perturbation of the initial solution $\mathbf{s}_{cur}^{\text{SA}}$ is performed. In figure 1 the symbol n represents the total number of selected descriptors in the respective solution (for example n_{new}^{SA} is the total number of descriptors in solution $\mathbf{s}_{new}^{\text{SA}}$).

The final outcome of the second stage is the solution $\mathbf{s}_b^{\text{SA}}$ that is close to the global optimum as far as the prediction error is concerned, while at the same time it contains the minimum possible number of input variables.

3. Results and discussion

The Selwood data set (Selwood et al., 1990) was employed for the evaluation of the GASA-RBF algorithm. The set serves as a benchmark problem which has been used extensively to evaluate variable selection algorithms in QSAR studies (So & Karplus, 1996; Waller & Bradley, 1999; Nicolotti et al., 2002; Liu et al., 2003). It comprises of 31 compounds, for which the values of 53 descriptors are available. The objective is to predict the antifilarial antimycin activity, expressed as $\log(IC_{50})$.

The results of the GASA-RBF algorithm are presented in table 1.The four descriptors chosen are: partial atomic charge for atoms 4 (ATCH4) and 5 (ATCH5), dipole vector coordinate Y (DIPV_Y) and the logarithm of the partition coefficient for octanol/water (LOGP). A comparison of the descriptors selected and the performance of the resulting models for various methods found in the literature is presented in table 2.

Based on these results two important observation can be made: First, the partition coefficient LOGP is selected by all the algorithms, which indicates the significance of this descriptor. In fact, LOGP is commonly used in QSAR analysis and rational drug design as a measure of molecular hydrophobicity, that is a measure of the movement of the drug through the membranes. Furthermore, table 2 indicates the superiority of the GASA-RBF algorithm both in modeling (higher correlation coefficient-r^2 values) and predictive ability (higher cross-validated correlation coefficient-q^2 and lower RMSECV values).

Table 1. Results of the GASA-RBF algorithm for the Selwood data set

Parameters	1st stage	2nd stage
Number of total descriptors	53	24
Number of selected descriptors	24	4
Number of fuzzy sets	9	9
Number of RBF network centers	19	16
RMSECV	0.3906	0.4261

Table 2. Comparison of the results of various methods for the Selwood data set

Method	Descriptors	r^2	q^2	RMSECV
GASA-RBF	ATCH4, ATCH5, DIPV_Y, LOGP	0.949	0.778	0.426
GNN (So & Karplus, 1996)	NSDL3, MOFIY, LOGP	0.845	0.750	
VSMP (Liu et al., 2003)	ATCH4, DIPV_Z, DIPMOM, MOFI_Y, LOGP, SUM_F	0.805	0.718	0.436
FRED (Waller &Bradley,1999)	ATCH4, ESDL3, VDWVOL, LOGP, SUM_F	0.829	0.683	
GPQSAR (Nicolotti et al., 2002)	SURF_A, LOGP, SUM_F, SUM_R	0.807	0.751	
GAS (Cho & Hermsmeier, 2002)	MOFI_Y, LOGP, SUM_F1	0.721	0.647	0.483

4. Conclusions

This work presents a novel variable selection method for QSAR analysis. The method decomposes the optimization problem into two sub-problems: In the first sub-problem the objective is to minimize the modeling error, while in the second the objective is to minimize the number of descriptors. In both stages, stochastic search techniques are utilized for performing the optimization task. The correlation between the input and output data is modeled based on the fuzzy means algorithm, which is used to train RBF neural network models. The GASA-RBF modeling framework exhibits small computational times and excellent prediction accuracy.

References

Grossmann, I.E., 2003, PSE2003, Kunming (China).

Kudo, M., Sklansky, J., 2000, Patt. Recogn., 33, 25-41.

Liu, S.S., Liu, H.-L., Yin, C-S., Wang, L.-S., 2003, J. Chem. Inf. Comput. Sci., 43, 964-969.

Luke, B.T., 1994, J. Chem.Inf. Comp. Sc., 34, 1279-1287.

Michalewitz, Z., 1996, Genetic Algorithms + Data Structures = evolution programs, Berlin, Springer Verlag, 3rd ed.

Nicolotti, O., Gillet, V.J., Fleming, P.J., Green, D.V.S., 2002, J. Med. Chem., 45, 5069-5080.

Rogers, D.R., Hopfinger, A.J., 1994, J. of Chem. Inf. Comput. Sci., 34, 854-866.

Sarimveis, H., Alexandridis, A., Tsekouras, G., Bafas, G., 2002, Ind. Eng. Chem. Res., 41, 751-759.

Selwood, D.L., Livingstone, D.J., Comley, J.C., O'Dowd, B.A., Hudson, A.T., Jackson, P., Jandu, K.S., Rose, V.S., Stables, J.N., 1990, J. Med. Chem., 33, 136-142.

So, S.S, Karplus, M., 1996, J. Med. Chem., 39, 1521-1530.

So, S.S., Richards, W.Q., 1992, J. of Med. Chem., 36, 3565-3571.

Sutter, D.L., Dixon, S.L. Jurs, P.C., J. of Chem. Inf. Comput. Sci., 35, 77-84.

Waller, C.L., Bradley, M.P., 1999, J. Chem. Inf. Comput. Sci., 39, 345-355.

Cho, J.C., Hermsmeier, M.A., 2002, J. Chem. Inf. Comput. Sci., 42, 927-936.