NEURAL NETWORKS FOR STRUCTURE-ACTIVITY RELATIONSHIP PROBLEMS.

Rafał Adamczak and Włodzisław Duch[1]
Department of Computer Methods, Nicolaus Copernicus University
ul. Grudziądzka 5, 87-100 Toruń, Poland

Abstract

The problem of finding relations between structure of large molecules and their chemical and biological activity is known as the structure-activity relation problem (SAR). Two neural networks developed in our group were applied to this problem: the Feature Space Mapping neurofuzzy system and the constrained MLP network used to extract logical rules. Two SAR data sets were analyzed: antibiotic activity of pyrimidine compounds, and carcinogenicity data from the Predictive-Toxicology Evaluation project of the US National Institute of Environmental Health Science (NIEHS).

1. INTRODUCTION

Understanding the quantitative relationships between chemical structure and biological and chemical activity of molecular compounds (SAR – structure activity relationships) is very important because it can speed up the process of drug design and perhaps help to find better drugs. In most cases such relationships cannot be derived from theoretical models since molecular compounds are very complex and models of biological receptors or chemical reactions are much too difficult for reliable computer simulations. Expensive and time-consuming experiments are done to determine the structure-activity relations. In SAR problems the goal is to construct predictive theory starting from a given set of chemical compounds of a known structure and activity. The chemicals can be described by a large number of attributes, such as topological indices describing the three-dimensional molecular structures, quantum mechanical descriptors, molecular field parameters etc. Finding most informative attributes is very important for the success of SAR models.

There are many methods used for analyzing SAR data and usually the results of these methods are quite similar. If several theories with statistically equivalent predictive accuracy exist the theory which has better explanatory power is preferred, otherwise the one with higher accuracy is better. Most neural networks and statistical methods have small explanatory power, therefore to be interesting for chemists their classification accuracy should be significantly better than that of the rule-based or inductive machine learning methods. Many SAR problems are concerned with prediction of some continuous properties (approximation problem); other problems may be reduced to classification.

SAR problems have been treated so far using statistical methods, such as linear regression, and with machine learning algorithms such as decision trees or inductive logic pro-

[1] e-mails: {raad,duch}@phys.uni.torun.pl, WWW: http://www.phys.uni.torun.pl/kmk

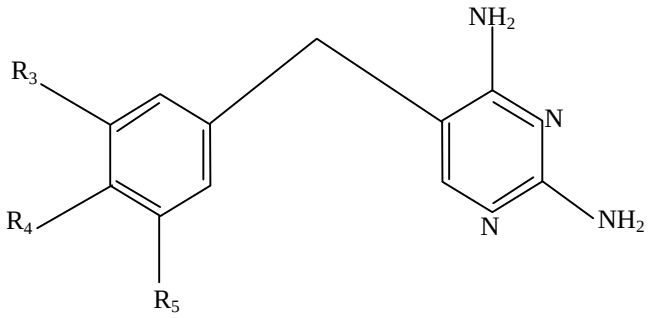
gramming. In this paper we report first application of neural networks to two SAR datasets, prediction of the antibiotic activity of pyrimidine compounds and prediction of carcinogenicity of organic chemicals.

In the next section the datasets that were analyzed are briefly described. The third section contains a brief description of neural models used. Comparison of results obtained and short discussion finishes the paper.

II. DATA DESCRIPTION

Pyrimidines

Pyrimidine class of chemical compounds [1] exhibit antibiotic activity. They inhibit the activity of bacterial forms of some enzymes in a stronger way than the human forms and therefore kill bacteria. All pyrimidines have a common template (see figure below).



At three possible substitution positions R_3, R_4, R_5 chemical groups can be added. 9 features describe every chemical group: group name, polarity, size, hydrogen-bond donor, hydrogen bond acceptor, pi-donor, pi-acceptor, polarizability, and the sigma effect. Each feature has few symbolic attributes that are coded by integer numbers. Because there are 3 substitution positions and each of them has 9 features the pyrimidine template is described by 27 integer-valued features. Some pyrimidines do not have any chemical groups at one or two substitution positions, therefore all corresponding features have missing values. Lack of appearance of a chemical group in some substitution position is very informative and therefore should be treated as additional feature value.

Each example in the pyrimidine data is a pair of chemical compounds represented by 54 features. There are two classes: either the first compound has bigger activity then the second, or vice versa. There are 2788 examples in the data and 5 fold cross-validation tests are done on this data to compare the results.

The predictive toxicology evaluation (PTE)

The PTE data was obtained from the Oxford University Computing Laboratory [2]. The database contains description of 330 compounds that represent all organic chemicals for

which US National Toxicology Program (NTP) experts have completed reports. 182 of these compounds (55%) are classified as carcinogenic, and 148 as non-carcinogenic. Every compound is described by 417 features. The feature set consist of 8 subsets: features from 1 to 69 describe atom type, feature no. 70 mutagenecity alert, features from 71 to 285 WARMR alert, no. 286-313 are counts of generic chemical groups found in the molecule, 314-376 are NTP bulk properties, 377-404 are alerts that were used in [3], 405-416 genotoxity test results, no. 417 is the *ames test*, and the last feature, no. 418, is the class value, 0 if the chemical is non-carcinogenic, 1 if it is carcinogenic, and 2 if its carcinogenicity is not known.

10 cases for which carcinogenicity is not yet known are found in the test set (called PTE-2), which contains also 20 other compounds of known carcinogenicity. Large number of features and small number of test cases makes it a difficult problem for any classification method.

III. DESCRIPTION OF NEURAL MODELS USED

In SAR problems it is very important to obtain a theory for predicting structure relation activity with strong explanatory power. From our point of view the best explanatory power have crisp logical rules which usually can be obtained by decision trees, or inductive logic programming. It is also possible to obtain crisp logical rules from neural networks. Here we briefly describe two such methods developed in our group: FSM and MLP2LN.

FSM

Feature Space Mapping (FSM) [4],[5] has been introduced as a universal adaptive system based on multidimensional separable functions. Viewed from various perspectives FSM is a neurofuzzy system, a density estimation network, a memory based system or an example of a self-organizing system. The main idea is simple: components of the input and output vector define features, and combinations of these features define objects in the feature space, described by the joint density probability of the input/output data vectors using a network of properly parameterized transfer functions.

FSM is a constructive neural network that estimates probability density of input-output pairs in each class. The architecture of the FSM network consists of three layers: an input, one hidden, and an output layer. Number of nodes in the hidden layer depends on the problem and is found automatically during the training phase. In the output layer there is only one node, used to estimate the confidence of the network in its classification. The search for the node in the hidden layer with the biggest activation determines the class and other properties of the given vector.

Generally there are no restriction on the type of transfer functions in the FSM model but so far only localized functions were used. The simplest functions with suitable properties for probability density modeling are of the Gaussian type or an approximate Gaussian type, for example the bicentral functions [6]:

$$Bi(\mathbf{x};\mathbf{D},\mathbf{b},\mathbf{s}) = \prod_{i=1}^{N} \sigma(e^{s_i}(x_i - D_i) + e^{b_i})(1 - \sigma(e^{s_i}(x_i - D_i) - e^{b_i}))$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$, and **x** is the input vector. Shape adaptation of the density $Bi(\mathbf{x};\mathbf{D},\mathbf{b},\mathbf{s})$ is possible by shifting centers *D*, changing spreads *b* and rescaling slopes *s*. It is possible also to use other localized functions like: triangular, trapezoidal, rectangular. Rectangular functions:

$$G(\mathbf{x};\mathbf{D},\sigma) = \begin{cases} 1 \ \ if \ \ | D_i - x_i | < \sigma_i \\ 0 \end{cases}$$

are very useful for extraction of crisp logical rules in the FSM network. Further details about this network are available in [5].

MLP2LN

MLP2LN [6] is a smooth transformation from an MLP network into a network performing logical operations (Logical Network, LN). This transformation is achieved during network training by: gradually increasing the slope of sigmoidal functions to obtain crisp decision regions simplifying the network structure by inducing the weight decay through a penalty term enforcing the integer weights values 0 and ±1, interpreted as 0 irrelevant input, +1 positive and –1 negative evidence. These objects are achieved by adding two additional terms to the standard mean square error function

$$E(W) = \frac{1}{2}\sum_p \sum_k (Y_k^{(p)} - f_k(\mathbf{x}^{(p)};\mathbf{W}))^2 + \frac{\lambda_1}{2}\sum_{i,j} W_{ij}^2 + \frac{\lambda_2}{2}\sum_{i,j} W_{ij}^2 (W_{ij}-1)^2 (W_{ij}^2+1)^2$$

After training we obtain the network with value of weights +1,-1,0 and sigmoid functions width very high slopes. Thanks to this it is possible to extract logical rules from the network.

IV. RESULTS

Result for pyrimidines obtained with the FSM network are compared below with three other methods [1]. First of these algorithms is the ILP (inductive logic programming) system called Golem [8], the second is a standard linear regression, and the last one is a decision tree called CART. To compare our algorithm with the results from the reference [1] as a classification accuracy we had to use the mean Spearman's rank correlation coefficient (which is an approximation to the exact correlation coefficient) defined as:

$$r_S = 1 - \frac{6}{n^3 - n}\sum_i d_i^2$$

where ***n*** is the number of pairs and ***d*** is the distance in rank of pairs (in this case the data comes in pairs). r_s has value between +1 and –1, where +1 means perfect positive correlation and –1 perfect negative correlation.

Method	R_S
Golem (ILP)	0.684
Regression	0.654
CART	0.499
FSM	0.78

Result given in the table above for the FSM network was obtained without any feature extraction and by using Gauss functions as the activation functions. Although FSM gave significantly better result then other methods only fuzzy rules with Gaussian membership functions may be derived in this case, while ILP gives explanations in form of crisp rules. Unfortunately attempts to produce such rules using either MLP2LN or rectangular functions in the FSM proved to be difficult so far, resulting in a large number of rules.

The PTE data are much more difficult then the pyrimidine data because the number of the training vectors is rather small and the dimensionality of the problem is very large. To decrease the dimensionality of the problem, which is very important for such large number of dimensions, FSM with rectangle transfer function was done first [5]. This allows to find logical rules and neglect dimensions that have no influence on the FSM network nodes activities. Out of 417 features only 60 features were left and used also for clauclation with Gaussian transfer functions. Results from FSM are compared with 5 algorithms. Distill-Light is a stochastic algorithm, STEPS is an evolutionary programming system that evolves programs trees using constructs from the Escher programming language, GloBo is a stochastic system, OFAI is a combination of two systems, C4.5 rules and naive Bayes classifier. More information about these algorithms can be found at the WWW address: http://oldwww.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/pte2-summary.html

Method	Accuracy %
Distill-Light	90
STEPS	85
GloBo	85
FSM rules	80
FSM gauss	75
OFAI	75
Default	70

The differences between the results in the table above may seem large but since only 20 test vectors are given each vector correctly classified changes the result by 5%. Results shown here say nothing about the merits of the methods used. The FSM results with the Gaussian transfer functions were obtained after choosing the best network parameters using cross-validation on the training set. Since there are stochastic elements in the algorithm every time the FSM network is running a different solution may be obtained, therefore results on the test set can be different too. In this case classification accuracy between 45%-75% is obtained on the test set but we may of course select the best network repeating the runs on the training set.

Similar results were obtained by using FSM for rule extraction. The best network was equivalent to 11 rules with 53 premises which used only 24 features. These rules classified 16 vectors from the test set correctly (80%), leaving 3 vectors as unclassified and 1 wrongly classified. Most of the time results are much worse and there is no correlation between results on the training set and the test set even when the number of rules is the same.

V. DISCUSSION

The two problems described in this paper are quite difficult and require further investigation. For pyrimidines the results are very good but the neural description of the data has little explanatory power. For toxicology data good results may be obtained by chance since very few test vectors are provided. Analysis of the rules generated by FSM show be done by the domain experts. Although FSM made only 1 error it left some vectors unclassified – assuming uncertainty of the input measurements [7] it should be possible to obtain non-zero classification probabilities for some of them.

REFERENCES

[1] R.D. King, A. Srinivasan, M.J.E. Sternberg (1995), Relating chemical activity to structure: an examination of ILP successes, New Generation Computing, 13: 411-433

[2] A. Srinivasan, S.H. Muggleton, R.D. King, M.J.E. Sternberg (1997), The predictive toxicology evaluation challenge, In: Proceedings of 15th International Conference on Artificial Intelligence (IJCAI-97). Morgan Kaufman, Los Angeles, CA; the data is available at:

http://oldwww.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/

[3] J. Ashby and R.W. Tennant (1991), Definitive relationships among chemical structure, carcinogenicity and mutagenecity for 301 chemicals tested by the U.S. NTP. Mutation Research, 27: 229-306.

[4] W. Duch, G.H.F. Diercksen (1995), Feature Space Mapping as a universal adaptive system, Computer Physics Communication 87: 341-371

[5] W. Duch, R. Adamczak (1997), New developments in Feature Space Mapping model, Third Conference on Neural Networks and Their Applications, Kule, pp. 65-70

[6] W. Duch N. Jankowski (1999), Survey of neural transfer functions. Neural Computing Surveys 2: 163-213

[7] W. Duch, R. Adamczak and K. Grąbczewski (2000), Methodology of extraction, optimization and application of crisp and fuzzy logical rules. IEEE Transactions on Neural Networks (in print)

[8] S.H. Muggleton, C. Feng (1990), Efficient induction of logic programs, Proceedings of the First Conference on Algorithmic Learning Theory, Tokyo 1990