Drug design by machine learning: Modelling drug activity

Dunja Mladenić, Aram Karalič
Jožef Stefan Institute
Artificial Intelligence Laboratory
Jamova 39, 61000 Ljubljana, Slovenia
Phone: (+38)(61) 159 199, Fax: (+38)(61) 161 029
E-mail: dunja.mladenic@ijs.yu, aram.karalic@ijs.yu

Abstract

This paper describes an approach to modelling drug activity using machine learning tools. Some experiments in modelling the quantitative structure-activity relationship (QSAR) using a standard, Hansch, method and a machine learning system GOLEM were already reported in the literature. The paper describes the results of applying two other machine learning systems, MAGNUS ASSISTANT and RETIS, on the same data. The results achieved by the machine learning systems, are better then the results of the Hansch method; therefore, machine learning tools can be considered as very promising for solving that kind of problems. The given results also illustrate the variations of performance of the different machine learning systems applied to this drug design problem.

1 Introduction

It is well known that biological drug activity is related to its molecular structure. An important issue in the process of drug design is how to determine the quantitative relationship between the structure and the activity of drugs [6]. The standard method for determining the structure-activity relationship is the Hansch method [3]. On the other hand, many machine learning methods offer a possibility to automatically synthesize the quantitative structure-activity relationship (QSAR).

King et al. in [6] used the machine learning system GOLEM for the modelling of QSAR. They report that machine learning can yield rules that model a QSAR of the specified class of drugs better than the Hansch method [3].

The aim of our work is to test two machine learning systems, MAGNUS ASSISTANT and RETIS, on the same problem of modelling the structure-activity relationship, and to compare these results with the results of the Hansch method and the results obtained with GOLEM.

The problem domain is described in the first section. Second section briefly describes each of the two machine learning systems used in the experiments. Section 3 describes some mechanisms for the evaluation of results. Experiments and results are described in Section 4, while the comparison of the results are given in Section 5. The final section concludes with the ideas for further work.

2 Domain description

Experimental data were obtained from the authors of the experiments with the system GOLEM [8]. The data consists of the chemical structure, chemical properties and biological activity for 55 thrimethoprim analogues.

The chemical structure of a drug is depicted on Figure 1. On the basic constant part of the drug there are three terminals (labelled 3, 4, 5) to which three substituents (R_3, R_4, and R_5) are connected.

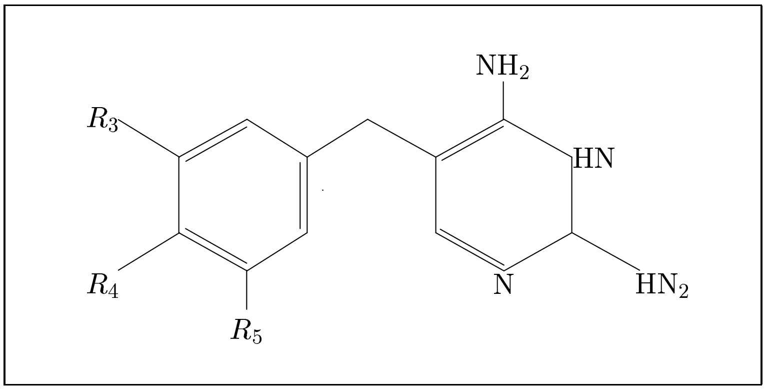

Figure 1: The structure of thrimethoprim analogues.

In our data set, there are 25 possible substituents. In the most drugs substituents R_3 and R_5 are either the same or R_5 is the hydrogen group (i.e. there is no substitution at position 5). Each of the 24 non-hydrogen substituents is described with 9 chemical properties: polarity, size, flexibility, hydrogen-bond donor, hydrogen-bond acceptor, π donor, π acceptor, polarisability and σ effect. The detailed description of substituents together with their chemical properties can be found in [6].

Each drug can therefore be described with 22 attributes; 9 chemical properties of R_3, 9 chemical properties of R_4, 4 additional boolean value attributes (which test whether $R_3 = H$, $R_4 = H$, $R_5 = R_3$, $R_5 = H$) and a real number (drug's biological activity) which represents the class of the drug.

3 Machine learning algorithms

This section describes machine learning algorithms ASSISTANT and RETIS applied to modelling of the drug activity.

3.1 Assistant

ASSISTANT [1] is an *attribute-value* learning algorithm for binary classification tree construction. The algorithm belongs to the TDIDT family of algorithms whose widely known representative is ID3 [10]. The basic ASSISTANT's algorithm is as follows:

Given:
set of training examples

if all examples belong to the same class **then**
this node becomes a leaf labelled with the class of examples
else
for the next node select the best attribute (the one which minimizes the expected entropy of the training set)
split the set of examples in the node into disjoint subsets according to binarized values of the attribute
recursively repeat the whole algorithm for each subset
endif
output constructed classification tree

Once built, a classification tree is interpreted in the following way: each internal node of the tree contains a test on a value of an attribute. According to the result of the test, the interpretation of the tree proceeds to the left or right subtree of the node. A leaf classifies an example into the appropriate class.

3.2 Retis

RETIS [4] is an *attribute-value* machine learning system for construction of regression trees. Regression trees are similar to binary classification trees. The most important distinction is that, while classification trees are used to classify objects into *discrete* classes, regression trees are used when the class is *continuous.* A regression tree actually implements a function $y(x_1, x_2, \ldots, x_n)$ of n continuous or discrete attributes $x_1, x_2, \ldots, x_n$.

The construction algorithm for regression trees is similar to the algorithm from the section 3.1. Since the class is continuous, an estimate of standard deviation of the class values is used as the measure of impurity of a node when searching for the best attribute. The best split of examples in a node is therefore the one which minimizes the expected standard deviation.

Once built, a regression tree is interpreted in a similar manner as a classification tree. The main difference is that a leaf prescribes a value to the function, approximated by the regression tree.

4 Mechanisms for the evaluation of results

Different methods for modelling QSAR were used on the same data. Some are resulting with the activity for each testing drug (HANSCH and RETIS), while others give the comparison of pairs of drugs (GOLEM and ASSISTANT). To enable the comparison of different methods, an uniform format for presenting results — ranking of drugs — was used.

Ranking from drug activity is trivial by sorting drugs according to the activity and by enumerating them, starting with one. If some drugs have an identical activity, the same rank is assigned to them all, calculated as the mean of the ranks that drugs would have had if their activity had been slightly different. Ranking from paired-comparison was done using the *David method* (see section 4.1) as in [6].

Comparison of results of different methods was done by calculating the *Spearman rank-order correlation coefficients* of the predicted and observed ranks (see section 4.2), for each method.

4.1 Ranking from paired-comparison

A translated problem results with drug pairs represented by a relation greater activity(X_i, X_j). A paired-comparison set is a set of pairs for which greater activity(X_i, X_j) is true. Ranking of drugs from paired-comparison was done by calculating the *David number* [2] for each drug X_i as DavidNum(X_i) = *total number of wins of drugs defeated by* X_i − *losses of drugs to whom* X_i *lost* + X_i*'s wins* − X_i*'s losses.* A pair (X_i, X_j) in the paired-comparison set denotes that X_i wins X_j i.e. X_j looses from X_i. A simple implementation of the above formula enables ranking according to the matrix of wins $A = [\alpha_{ij}]$, where α_{ij} is the number of times drug i wins over drug j.[1] David number is calculated from the above described matrix by the formula

$$\mathrm{DavidNum}(X_i) = \sum_j \alpha_{ij} w_j - \sum_j \alpha_{ji} l_i + \sum_j \alpha_{ij} - \sum_j \alpha_{ji}$$

where w_i is the i-th row of A (wins of drug i) and l_j is the j-th column of A (losses of drug j).

According to the David number, drugs were sorted and enumerated starting with number one. In the case of the tied David numbers, the ranks are also tied and calculated as the mean of the ranks if the David numbers hadn't been tied, Eg. in the case of David numbers: $(1, 2, 8, 8, 10, 15, 20, 20, 20, 40)$ the computed ranks are: $(1, 2, 3.5, 3.5, 5, 6, 8, 8, 8, 10)$, since $3.5 = \frac{3+4}{2}$ and $8 = \frac{7+8+9}{3}$.

4.2 Evaluation of results

The results of experiments were the ranks of drugs and these were compared by the Spearman rank-order correlation coefficient to the actual (observed) ranks reported in [9]. Two variants of correlation were used, the first ignoring and the second considering ties in the data. The Spearman rank-order correlation coefficient is defined as a linear correlation coefficient of the ranks (R_i, S_i) and is given by the formula

$$r(R_i, S_i) = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}} \tag{1}$$

Considering ties in the data and the substitution $D = \sum_i (R_i - S_i)^2$, the Spearman rank-order correlation coefficient is given by the formula

$$r(R_i, S_i) = \frac{1 - \frac{6}{N^3 - N}[D + \frac{1}{2}\sum_k ({f_k}^3 - f_k) + \frac{1}{2}\sum_m ({g_m}^3 - g_m)]}{\left[1 - \frac{\sum_k ({f_k}^3 - f_k)}{N^3 - N}\right]\left[1 - \frac{\sum_m ({g_m}^3 - g_m)}{N^3 - N}\right]} \tag{2}$$

[1]Notice that $\alpha_{ii} = 0$ and $\alpha_{ij} + \alpha_{ji} = 1$. To satisfy this condition, the case $\alpha_{ij} = \alpha_{ji}$ (i.e. a tie) was substituted by $\alpha_{ij} = \alpha_{ji} = \frac{1}{2}$ and the case $\alpha_{ij} > \alpha_{ji}$ by $\alpha_{ij} = 1$, $\alpha_{ji} = 0$.

where N is the number of examples, f_k is the number of ties in the k-th group of ties among the R_i's and g_m is the number of ties in the m-th group of ties among the S_i's. If all the f_k's and all the g_m's are equal to one, meaning that there are no ties, then the given formulas are equal, so the latter formula can be said to be more general then the former.

5 Experiments and Results

Since our goal was to compare the results of two machine larning systems with the results of the Hansch method and GOLEM, example set was split into training example set (44 examples) and testing example set (11 examples) in the same way as in [6].

5.1 Magnus Assistant

MAGNUS ASSISTANT [7] — an implementation of the ASSISTANT algorithm — was used in our experiments. It enables experiments with several hundreds of attributes and thousands of examples.

Since ASSISTANT is designed to classify objects into *discrete* classes, the original structure-activity problem was translated to learning of the relation "greater activity" over the pairs of drugs (as in [6] for the program GOLEM). Examples are drug pairs, each example described by a boolean class and values of 44 attributes (22 for each drug) describing drug's structure and chemical properties. A value of the class indicate whether the first drug's activity is greater then the activity of the second. By translating the original problem, the original set of 44 training and 11 testing examples is transformed into the set of 1742 training and 1046 testing examples, which were obtained from the authors of experiments with the system GOLEM described in [6].

When building classification trees, various values and combinations of pre-pruning parameters were used [1], starting with no pruning (0, 100, 0) and gradually increasing the influence of pre-pruning.[2] Trees generated in this way were tested on 11 drugs producing ranking of drugs. These rankings are shown and compared with the obtained ranking in Table 1.

[2]Post-pruned tree produced the same ranks of 11 testing drugs as the original tree.

Magnus Assistant					
Experiment 1 pre-pruning: 0/2.5/3, 100, 0			Experiment 2 pre-pruning: 5, 100, 0		
observed rank	predicted rank	correlation	observed rank	predicted rank	correlation
3.5	2		3.5	2	
1	9		1	4	
9	8	considering	9	9	considering
7.5	6.5	ties	7.5	9	ties
6	4.5	0.486	6	9	0.481
7.5	11		7.5	9	
10.5	10	ignoring	10.5	9	ignoring
10.5	6.5	ties	10.5	5.5	ties
5	4.5	0.538	5	5.5	0.743
3.5	3		3.5	3	
2	1		2	1	
Experiment 3 pre-pruning: 0, 90, 0			Experiment 4 pre-pruning: 0, 100, 3/4/4.5		
observed rank	predicted rank	correlation	observed rank	predicted rank	correlation
3.5	3		3.5	2.5	
1	11		1	5.5	
9	8	considering	9	8.5	considering
7.5	7	ties	7.5	5.5	ties
6	6	0.310	6	5.5	0.502
7.5	9.5		7.5	5.5	
10.5	9.5	ignoring	10.5	8.5	ignoring
10.5	5	ties	10.5	10	ties
5	4	0.353	5	11	0.661
3.5	2		3.5	2.5	
2	1		2	1	

Table 1: Classification of testing examples with Magnus Assistant.

5.2 Retis

When learning regression trees, each training example consisted of a continuous class, representing the biological activity of a drug, and 22 values of attributes describing drug's structure and chemical properties.

When building the regression tree, no pre-pruning was used. The obtained tree was then m-pruned [5] for various values of the parameter m. Such pruned trees were used to predict ranks of 11 testing drugs. Predicted activity and ranks are shown and compared with the observed values in Table 2.

Retis							
		Experiment 1 post-pruning: 0.0/0.01			Experiment 2 post-pruning: 0.05/0.1/0.25		
observed activity	observed rank	predicted activity	predicted rank	Spearman correlation	predicted activity	predicted rank	Spearman correlation
7.87	3.5	6.70	3		6.68	2	
7.56	1	7.69	6.5		7.69	6.5	
8.85	9	8.35	9	considering	8.35	9	considering
8.82	7.5	8.35	9	ties	8.35	9	ties
8.57	6	8.35	9	0.415	8.35	9	0.377
8.82	7.5	8.87	11		8.87	11	
8.87	10.5	7.69	6.5	ignoring	7.69	6.5	ignoring
8.87	10.5	7.23	5	ties	7.23	5	ties
8.42	5	6.96	4	0.510	6.94	4	0.507
7.87	3.5	6.65	1.5		6.68	2	
7.74	2	6.65	1.5		6.68	2	
		Experiment 3 post-pruning: 0.5			Experiment 4 post-pruning: 1/4/5		
observed activity	observed rank	predicted activity	predicted rank	Spearman correlation	predicted activity	predicted rank	Spearman correlation
7.87	3.5	6.92	3		6.92	3	
7.56	1	7.69	6.5		7.69	8.5	
8.85	9	8.35	9	considering	8.35	8.5	considering
8.82	7.5	8.35	9	ties	8.35	8.5	ties
8.57	6	8.35	9	0.023	8.35	8.5	-0.520
8.82	7.5	8.87	11		8.87	8.5	
8.87	10.5	7.69	6.5	ignoring	7.69 5	8.5	ignoring
8.87	10.5	6.92	3	ties	6.92 5	3	ties
8.42	5	6.92	3	0.362	6.92	3	0.320
7.87	3.5	6.92	3		6.92	3	
7.74	2	6.92	3		6.92	3	

Table 2: Classification of testing examples with Retis.

6 Comparison of results

Machine learning systems MAGNUS ASSISTANT and RETIS were tested on the same data. MAGNUS ASSISTANT learned the "greater activity" relation over pairs of drugs and 11 testing drugs were classified according to the learned relation. On the other hand, RETIS learned to predict the activity of individual drugs. The result of classification are real numbers, one for each testing drug. To enable the comparison of the results, both types of results were translated to ranks.

The results — predicted ranks of testing drugs obtained without pruning, given in Table 3 — were compared by the Spearman rank-order correlation coefficient with the observed ranks. Spearman rank-order correlation coefficient for the two machine learning systems and the two additional ones, the Hansch method and the machine learning system GOLEM, are given in Table 4. For each method/system, two real numbers are given, one presenting the Spearman rank-order correlation coefficient ignoring and the other considering eventual ties in ranks. Since the ties are produced when a method/system can not distinguish between ranks, it is suggested to consider ties in the calculation of a correlation coefficient using Equation (2).

observed rank	rank predicted with method/system			
	HANSCH	GOLEM	MAGNUS ASSISTANT	RETIS
3.5	1	2	2	3
1	4.5	9	9	6.5
9	7	9	8	9
7.5	11	5.5	6.5	9
6	9	9	4.5	9
7.5	2	9	11	11
10.5	8	9	10	6.5
10.5	6	5.5	6.5	5
5	10	4	4.5	4
3.5	4.5	2	3	1.5
2	3	2	1	1.5

Table 3: The rank ordering of 11 testing drugs.

.

Spearman rank-order correlation coefficient	Method/System			
	HANSCH	GOLEM	MAGNUS ASSISTANT	RETIS
considering ties	0.373	0.125	0.486	0.415
ignoring ties	0.415	0.457	0.538	0.510

Table 4: The correlation of predicted and observed rank.

.

One should be aware of the fact that the number of training examples as well as their format (i.e. the number of attributes or relations considered) influence the results of classification. It can be seen that for MAGNUS ASSISTANT, the correlation between the observed and predicted ranks are the greatest, despite the fact that RETIS is designed for learning regression trees and the given QSAR problem is a regression problem. Namely, RETIS requested less preparation of data and learned from 44 examples. On the other hand, GOLEM and MAGNUS ASSISTANT were tested on the translated problem on 1742 training examples of pairs of drugs (produced from the original 44 examples).

The low correlation GOLEM achieved (considering ties) can be explained by ties in the predicted ranks, specialy rank 9 (see Table 3). Since the actual spliting of example set into training and testing examples has a significant influence on the results, some experiments with a different spliting, carried out by King et al., result with better correlations.

It must be noted that some of the machine learning systems used background knowledge, as GOLEM, while for the others, such as RETIS, similar information was contained in the examples. The use of different methods for noise handling, incorporated in different machine learning systems additionaly influences the classification results. It is hard to say which system is the most appropriate for the described sort of problems but it can be concluded, that machine learning tools are promising for modelling biological information.

7 Conclusion

The work described in this paper was done in order to compare the results of two machine learning systems, MAGNUS ASSISTANT and RETIS to the results of GOLEM and the Hansch method described in [6]. All the systems were applied to the same data. The predicted values were compared by the correlation coefficient to the observed values, resulting with a real number (correlation coefficient) for each method/system.

The idea of comparing the correlation coefficients is borrowed from the paper [6] and can be considered useful for comparing of such kind of testing results, produced with different machine learning systems. The same idea can be used to illustrate the variations of performance of different machine learning systems.

The comparison of the results shows that machine learning tools are very promising for modelling biological information. Therefore, machine learning tools can be used in further experiments on the same problem with more then 55 drugs and also on some similar problems.

References

[1] Cestnik, B., Kononenko, I. and Bratko, I., (1987), *ASSISTANT 86: A Knowledge-Elicitation Tool for Sophisticated Users*, in Progress in Machine Learning, ed. by I.Bratko and N.Lavrač, Sigma Press, Wilmslow.

[2] David, H.A. (1987) *Ranking from unbalanced paired-comparison data.* Biometrica 74, 432-436.

[3] Hansch, C., Maloney, P.P., Fujita, T. and Muri, R.M. (1962) *Nature* 194, 178-180.

[4] Karalič, A. (1991) *Machine learning of regression trees from incomplete data*(in Slovene). MSc Thesis, University of Ljubljana, Faculty of Electrical Engineering and Computer Science.

[5] Karalič, A., Cestnik, B. (1991), *The Bayesian Approach to Tree-Structured Regression.* Proceedings of ITI-91, Cavtat, Croatia.

[6] King, D.S., Muggleton, S., Lewis, R.A. and Sternberg, M.J.E. (1991) *Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationship of trimethoprim analogues binding to dihydrofolate reducates.* Submitted to PNAS.

[7] Mladenić, D. (1990) *Machine learning system Magnus Assistant* (in Slovene). BSc Thesis, University of Ljubljana, Faculty of Electrical Engineering and Computer Science.

[8] Muggelton, S.H. and Feng, C. (1990) *Efficient induction of logic programs.* First Conference on Algorithmic Learning Theory. Tokyo: Ohmsha.

[9] Press, W.H., Flannery, B.P, Teukolsky, S.A. and Vetterling, W.T. (1987) *Numerical Recepies* 484-494.

[10] Quinlan, J.R. (1986) *Induction of Decision Trees.* Machine Learning, Vol. 1, pp. 81-106.