Comparison of three Different Approaches to the Property Prediction Problem

Damijana Keržič* and Borka Jerman Blažič

Jožef Stefan Institute, Ljubljana, Slovenia

Vladimir Batagelj*

Department of Mathematics, University of Ljubljana, Ljubljana, Slovenia

Received May 18, 1993®

Neighborhood subspace approximation method has been developed for solving the property prediction problem. In this paper the performance of this method is evaluated on three groups of compounds. The molecular structure was encoded as sequences of well-known structural (topological) indices. Euclidean distance has been used for determining the similarities between compounds. Property prediction results of the new method are compared with the results of the neighborhood based approximation (clustering) method and with the results obtained by an artificial intelligence program using machine learning tools.

1. INTRODUCTION

The property prediction problem is one of the basic problems of QSAR (quantitative structure–activity relationship) studies. The interest in quantification of the similarity between chemical structures arises from the expectation that molecules with similar structures also have similar physicochemical properties and biological activities. The basic assumption is that the molecular properties are determined primarily by the structure and that the compounds with similar structure lead to similar biological activity or similar physicochemical properties. Therefore, it can be assumed that the property values in general change smoothly over structurally similar compounds.

In order to define structural similarity, one needs to represent chemical compounds in mathematical therms. In the near past graph-theoretical methods have been largely applied in the QSAR. The standard approach in operationalizing the prediction problem in QSAR is by the use of structural (topological) indices (numerical graph invariants) based on molecular graphs in which hydrogen atoms are omitted, because they normally do not play a major role in determining the structure of a molecule. Structural indices express in numerical form the structure of the chemical compound represented and were developed for the purpose of obtaining correlations with the physicochemical properties of chemical compounds and for expressing the molecular dissimilarity (or similarity).

In the paper we compare the prediction power of our new method to two alternative techniques, i.e., with a neighborhood based method and with an attribute-value machine learning system for construction of regression trees, named RETIS. The structural descriptors used in our approach are some well-known structural indices.[1,11,12] The first two methods are based on the metric approach and the third one is from the field of artificial intelligence:

(1) Neighborhood Subspace Prediction Method. Every compound is represented as a vector of selected structural indices (point in a metric space). The same way of representation was used in all three methods. We compute a selected distance between compounds.[3] Then we look for the most suitable subspace (such as a line, a plane, etc.) in the induced metric space where the point with the unknown property value is located. A (generalized) inter-/extrapolation over this subspace is applied to approximate the unkown property value.[4]

(2) Neighborhood Based Prediction Method. The prediction procedure assumes that the predicted property value depends on the values of some elements in the cluster—the neighborhood of compounds which are similar enough to it. The clustering of the compounds in the compact groups is based on searching for the most highly correlated compounds (kernels) for the future clusters. The procedure is explained in details in ref 6.

(3) Machine Learning Systems RETIS. RETIS is a system for automatic knowledge acquisition from a given set of examples (compounds with known property values). The knowledge, induced from examples, is represented in a form of a regression tree which is interpreted in a similar manner as a classification tree. The main difference is that here a leaf prescribes a value to the function, approximated by the regression tree. For details about RETIS see ref 13.

2. DISTANCES, LINES, PLANES, AND PROPERTY PREDICTIONS

In this section we present in mathematical terms the basic features of the *neighborhood subspace prediction* method.

Let $\mathcal{E}$ be a *set of units* and $\mathcal{L}$ a *learning set* (set of examples), $\mathcal{L} \subseteq \mathcal{E}$. On the learning set $\mathcal{L}$ the (property) function p: $\mathcal{L} \rightarrow \mathrm{R}$ (a measured physical quantity or biomedical, toxicological, or environmental activity) is known. The *property prediction problem* can be expressed as follows:

For the unit $Z \in \mathcal{E} \setminus \mathcal{L}$ predict the function value $p(Z)$.

Let $d: \mathcal{E} \times \mathcal{E} \rightarrow \mathrm{R}_0^+$ be a *semidistance* on $\mathcal{E}$, i.e. for all $X,Y,Z \in \mathcal{E}$ hold

$$d(X,X) = 0 \tag{1}$$

$$d(X,Y) = d(Y,X) \tag{2}$$

$$d(X,Z) + d(Z,Y) \geq d(X,Y) \tag{3}$$

d is a *dissimilarity* iff eqs 1 and 2 hold. d is said to be *Euclidean if* units from $\mathcal{E}$ can be embedded in an Euclidean space, φ: $\mathcal{E} \rightarrow \mathrm{R}^k$, such that for all $X,Y \in \mathcal{E}$: $d(X,Y) = \delta(\varphi(X),\varphi(Y))$, where δ is the Euclidean distance. It can be shown that each dissimilarity can be transformed into an Euclidean semidistance.[5] From now on we shall assume that d is an Euclidean semidistance.

* Abstract published in *Advance ACS Abstracts,* March 1, 1994.

0095-2338/94/1634-0391$04.50/0

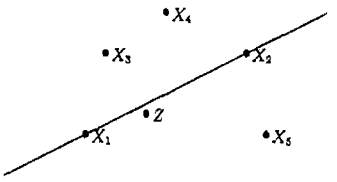


Figure 1. Line.

```
K = 0; S = {X ∈ L \ {Z}, d(X,Z) < δ}; found = false
search : while S ≠ ∅ do begin
        X = argmin_{T∈S} d(Z,T);
        for Y ∈ K do if Z is close enough to ⟨X,Y⟩ then begin
            X' = X; Y' = Y; found = true; exit search
        end
        K = K ∪ {X}; S = S \ {X}
    end;
    if found then p_Z = p'(Z,X',Y') else p_Z = undefined
```

Figure 2. Prediction procedure.

In our space $(\mathcal{E},d)$ we can define a *ray* $[X,Y\rangle$ from unit X through unit Y as

$$[X,Y\rangle = \{Z : |d(X,Y) - d(X,Z)| = d(X,Z)\}$$

and a *line* as a union of two rays, $\langle X,Y\rangle = [X,Y\rangle \cup [Y,X\rangle$.

Suppose that for a unit $Z \in \mathcal{E}\setminus\mathcal{L}$ with unknown property value there exists a ray $[X,Y\rangle, X,Y \in \mathcal{L}$, such that unit Z lies on it, $Z \in [X,Y\rangle$. Then we can use a linear inter-/extrapolation of the property p along this ray to approximate $p(Z)$:

$$p'(Z,X,Y) = p(X) + \frac{p(Y) - p(X)}{d(Y,X)}\, d(Z,X)$$

where p' denotes predicted value of the property p.

The main question to be answered in order to apply this approach is: given a unit $Z \in \mathcal{E}\setminus\mathcal{L}$, is there a line passing through it?

Usually there is no such line, but we can often find lines to which Z is "close". In such cases we choose one of these lines for property approximation. There is a nontrivial problem of which line to choose. We solve this problem with some conditions on neighborhoods of units. In ref 2 we proposed the following procedure (see Figure 2) for the approximation of $p(Z), Z \in \mathcal{E}\setminus\mathcal{L}$. Let $\mathcal{H}$ denotes the set of units which have already been examined. At every step we choose the closest new candidate in set $\mathcal{S}$ and then look for the appropriate line passing through the chosen candidate, unit Z and the units in $\mathcal{H}$. In the last sentence of the procedure we predict the unknown value. p_Z = *undefined* means that we cannot predict the value with the proposed method. One possible solution to avoid undefinedness is the use, in such a case, of some simple alternative prediction method, such as taking the median value, the average value, or the value of the nearest neighbor. We express the condition "Z is close enough to $\langle X,Y\rangle$" by

$$(|d(X,Y) - d(X,Z)| - d(Y,Z)| < \epsilon) \wedge (cd(X,Y) > d(Y,Z))$$

The first term expresses the relaxed line condition, and the second term expresses the requirement that we do not extrapolate on the basis of two units which are too close.

If there is no appropriate line, we can generalize the method to the neighborhood subspaces of higher dimensions nearly containing the unit Z. The first step in the generalization is a *plain*.

When can we say that four different units lie on the same plane in a metric space? Let $X_1, X_2, X_3 \in \mathcal{E}$ be the units not lying on the same line, $X_1 \notin \langle X_2,X_3\rangle$. We say that these three units determine a plane which contains $Z \in \mathcal{E}$ if the following condition holds

$$V(X_1,X_2,X_3,Z) = 0$$

where V denotes the volume of a parallelotop determined with these units.[10] Our information about the units and their mutual relations is based only on the semidistance d. Therefore the above condition can be expressed as[10]

$$\mathbf{A} = \frac{1}{8}\begin{bmatrix} 2\Delta_{12} & \Delta_{12}+\Delta_{13}-\Delta_{23} & \Delta_{10}+\Delta_{12}-\Delta_{02} \\ \Delta_{12}+\Delta_{13}-\Delta_{23} & 2\Delta_{13} & \Delta_{10}+\Delta_{13}-\Delta_{03} \\ \Delta_{10}+\Delta_{12}-\Delta_{02} & \Delta_{10}+\Delta_{13}-\Delta_{03} & 2\Delta_{10} \end{bmatrix}$$

$$V^2(Z,X_1,X_2,X_3) = \det \mathbf{A}$$

where $\Delta_{ij} = d^2(X_i,X_j)$; $i,j = 0, 1, 2, 3$; and $X_0 = Z$.

If Z lies inside the triangle (X_1,X_2,X_3), the approximation of the unknown property value is calculated by generalized linear interpolation in the following way:

$$p'(Z,X_1,X_2,X_3) = \frac{p(X_1)\,V(Z,X_2,X_3) + p(X_2)\,V(X_1,Z,X_3) + p(X_3)\,V(X_1,X_2,Z)}{V(X_1,X_2,X_3)}$$

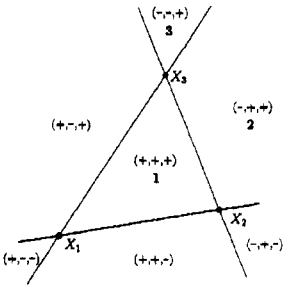
In general, as we can see from Figure 3, we have to determine the type of location of unit Z. We have seven types of regions. The plus/minus combination in the linear inter-/extrapolation depends on the region where Z lies. This approach can be easily generalized to higher dimension subspaces:

$$p'(Z,X_1,X_2,\ldots,X_k) = \frac{1}{V(X_1,X_2,\ldots,X_k)} \sum_{i=1}^{k} \sigma_i p(X_i)\, V(X_1,\ldots,X_{i-1},Z,X_{i+1},\ldots,X_k)$$

where σ_i is -1 or $+1$.

We do not yet know an efficient method for determining from metric data the type of the region. Since the dimension of the subspace is small, we are using a brute force approach, searching through all possible sign combinations $\sigma = (\sigma_1,\ldots,\sigma_k)$ and selecting one which minimizes the expression

$$|V(X_1,X_2,\ldots,X_k) - \sum_{i=1}^{k} \sigma_i V(X_1,\ldots,X_{i-1},Z,X_{i+1},\ldots,X_k)|$$



$$p'(Z) = \begin{cases} \dfrac{p(X_1)V(Z,X_2,X_3) + p(X_2)V(X_1,Z,X_3) + p(X_3)V(X_1,X_2,Z)}{V(X_1,X_2,X_3)} & \text{unit } Z \text{ lies in 1} \\ \dfrac{-p(X_1)V(Z,X_2,X_3) + p(X_2)V(X_1,Z,X_3) + p(X_3)V(X_1,X_2,Z)}{V(X_1,X_2,X_3)} & \text{unit } Z \text{ lies in 2} \\ \dfrac{-p(X_1)V(Z,X_2,X_3) - p(X_2)V(X_1,Z,X_3) + p(X_3)V(X_1,X_2,Z)}{V(X_1,X_2,X_3)} & \text{unit } Z \text{ lies in 3} \end{cases}$$

Figure 3. Generalized linear inter-/extrapolation in plane.

3. APPLICATIONS

The neighborhood subspace prediction method was tested on three groups of compounds: *decane isomers*,[8] *benzamidine derivatives*,[9] and *trimethoprim analogues*.[7] In the first test

Table 1. Prediction Power Coefficients

	R	R_A/R	R_B/R	R_C/R	n	N
decane isomers	5.75	0.59	0.99	0.62	39	74
benzamidines	0.57	0.47	0.49	0.57	79	90
trimethoprimes	0.77	0.78	0.96	0.86	37	55

group we tried to predict the boiling points; in the other two groups the prediction values were some biological activities. In the case of benzamidine derivatives the activity is expressed as log $(1/C)$, where C is the molar concentration causing 50% inhibition of complement,[9] and for the trimethoprim analogues biological activity is measured as log $(1/K_i)$, where K_i is the equilibrium constant for the association of the drug to DHFR (dihydrofolate reductase).[7]

In the case of decanes we selected as the structural descriptors the number of paths of lengths 2 and 3 (p_2,p_3), because it was shown that for several thermodynamical molecular properties these two numbers have strong description powers.[8] It is interesting that an inter-/extrapolation line exists for every decane isomer in the set.

For the other two groups of compounds we used selected structural indices which compose the composite index.[1,11,12] Every compound was represented with a vector of the values of the indices, $i(G) = (i_1(G),i_2(G),...,i_k(G))$. An index was selected on the basis of its correlation with other indices and the property we intended to predict. For benzamidines optimal results have been found with six indices: polarity number p, Balaban index J, expanded Wiener number $\tilde{W}$,[11] graph distance index GDI, Balaban information index U,[12] and radius of a graph r. For the group of trimethoprim analogues the following eight indices were selected: Wiener index W, polarity number p, Balaban index J, mean square distance index $D^{(2)}$, graph distance index GDI, information index for the equality of distances T_D^E, Zagreb group index M_2, and Balaban information index U. We standardized the obtained vectors to reduce the influence of indices of higher values.

A serious problem in comparison was a big number of isostructural compounds with different property values in the original data sets (see Table 1: N = number of all compounds, n = number of different (nonisostructural) compounds). Compounds $X,Y \in \mathcal{E}$ are isostructural iff $i(X) = i(Y)$. We decided to use in our comparison the reduced subsets of compounds containing only one representative from each class of isostructural compounds.

As a semidistance d, we used the Euclidean distance

$$\delta(X,Y) = \left(\sum_{i=1}^{k} (x_i - y_i)^2\right)^{1/2}$$

We have compared the prediction power of our method with the neighborhood based method[6] and with an attribute-value machine learning system for construction of regression trees named RETIS.[13]

We used the *leave-one-out* method—for each $X \in \mathcal{E}$ we use all other units as a learning set $\mathcal{L}_X = \mathcal{E} \setminus \{X\}$. We measured the prediction power of the method M by the coefficient R_M/R, where

$$R = \frac{1}{n} \sum_{X \in \mathcal{E}} |p(X) - p_{\text{med}}|$$

is the average prediction error of the trivial *median method*

Table 2. Methods Orderings

	decanes		benzamidines		trimethoprimes	
ABC	4	13	18.5	32.5	6.5	16
ACB	9		14		9.5	
BAC	1	9	7.5	23	5.5	10.5
BCA	8		15.5		5	
CAB	14	17	7	23.5	7.5	10.5
CBA	3		16.5		3	

$p(X) = p_{\text{med}}$ = median of property values; and

$$R_M = \frac{1}{n} \sum_{x \in \mathcal{E}} |p(X) - p_M(X)|$$

is the average prediction error of the method M.

In Table 1 the values of the prediction power coefficient for all three methods, A = neighborhood subspace prediction method, B = neighborhood based prediction methods, and C = RETIS, for the selected data sets are presented.

On all three data sets our method is the best among the compared methods. The improvements of predictions upon the trivial (median) method are relatively poor. This can be partially explained by inadequate description (selection of indices and weights in distance) of compounds. In the case of trimethoprime analogues there are several compounds which have isomorphic structural graphs—they differ only in one atom. In such cases the isostructurality problem can only be resolved by also including in the compound description additional information about the atoms (e.g. atomic mass).

An alternative approach to comparison of prediction methods is based on the orderings (permutations) induced by absolute error. We define the contribution of unit X to the permutations by the rules that can be induced from the following examples.

$$|p(X) - p_A(X)| < |p(X) - p_B(X)| < |p(X) - p_C(X)| \Rightarrow ABC = 1$$

$$|p(X) - p_A(X)| = |p(X) - p_B(X)| < |p(X) - p_C(X)| \Rightarrow ABC = BAC = {}^1/_2$$

$$|p(X) - p_A(X)| = |p(X) - p_B(X)| = |p(X) - p_C(X)| \Rightarrow ABC = ... = CBA = {}^1/_6$$

The value of a permutation is the sum of contributions of all units to it. In Table 2 values of permutations for all three methods for selected data sets are presented. Again, our method leads.

4. CONCLUSION

We proposed a new approach to the property prediction problem—the neighborhood subspace prediction method. Although the comparisons show that it gives better results, there is still a lot of space for improvements. For example: the linear inter-/extrapolation could be replaced with quadratic inter-/extrapolation, if there are three units with known property values on the same line. In the prediction used in the comparisons we considered in our method only lines and planes. We expect to improve the accuracy of the results by considering the neighborhood spaces of higher dimension. Another question which needs further study is the appropriateness of a given prediction model/method for a given data set.

ACKNOWLEDGMENT

We would like to thank Artificial Intelligence Laboratory, Computer Science Departement, Jožef Stefan Institute, for making it possible to use their program RETIS.

REFERENCES AND NOTES

(1) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Topological Indices for Structure-Activity Correlations*, Topics in Current Chemistry 114; Springer: Berlin, Heidelberg, 1983.
(2) Batagelj, V.; Keržič, D. Distances, Lines and Property Prediction. International meeting on distance analysis, Distancia' 92, Rennes, France, June 22–26, 1992.
(3) Batagelj, V. Similarity Measures Between Structured Objects. In *Proceedings of MATH/CHEM/COMP 88*; Graovac, A.; Ed.; Studies in Physical and Theoretical Chemistry; Elsevier: Amsterdam, 1989; Vol. 63, pp 25–40.
(4) Batagelj, V.; Keržič, D. Neighborhood Subspace Method for Property Prediction. Paper presented at MATH/CHEM/COMP'93, Rovinj, Croatia, June 1993.
(5) Gower, J. C.; Legendre, P. Metric and Euclidean Properties of Dissimilarity Coefficients. *J. Classif.* **1986**, *3*, 5–48.
(6) Jerman-Blažič, B.; Fabič-Petrač, I. Evaluation of the Molecular Similarity and Property Prediction for QSAR Purposes. *Chemom. Intell. Lab. Syst.* **1989**, *6*, 49–63.
(7) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. E. Drug design by machine learning: The use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. Submitted for publication in *PNAS*.
(8) Randić, M.; Wilkins, C. L. Graph Theoretical Ordering of Structures as a Basis for Systematic Searches for Regularities in Molecular Data. *J. Phys. Chem.* **1979**, *83*, 1525–1540.
(9) Hansch, C.; Yoshimoto, M. Structure-Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidines. *J. Med. Chem.* **1974**, *17*, 1160–1167.
(10) Sommerville, D. M. Y. An Introduction to the Geometry of *n*-Dimensions; Dover Publications, Inc.: New York, 1958.
(11) Tratch, S. S.; Stankevitch, M. I.; Zefirov, N. S. Combinatorial Models and Algorithms in Chemistry. The Expanded Wiener Number-A Novel Topological Index. *J. Comput. Chem.* **1990**, *11*, 899–908.
(12) Balaban, A. T.; Balaban, T.-S. New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8*, 383–397.
(13) Karalič, A. RETIS A Knowledge Acquisition System, User's manual, Software version 2.07/ nf; Institut Jožef Stefan: Ljubljana, Slovenia, 1992.