Solving the Structure-Property Problem Using k-NN Classification

Aleksandr Perevoznikov, Alexey Shestov,
Evgenii Permiakov, and Mikhail Kumskov

Lomonosov Moscow State University, Faculty of Mechanics and Mathematics,
Department of Computational Mathematics
GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation
qsar_msu@mail.ru, sperev@yahoo.com
<http://www.math.msu.su/departement/vychmath/index.html>

Abstract. The solution of the "structure-property" based on the molecular graphs descriptors selection with k-NN classifier is proposed. The results of comparing the construction of predictive models using the search and without it are given. The stability of the classifier function construction quality is tested using the test sample.

Keywords: Pattern Recognition, QSAR, QSPR, k-NN.

1 Introduction

The task of the "structure-property" problem is to estimate the relationship between the structure of chemical compounds and their properties. Solution of this problem can be divided into two stages: choosing the description of molecular graphs and the construction of the classifying function. structural descriptors (features) [1] - pairs and triples of singular points [2,3] defined on a triangulated molecular surface of the compound were selected for description and singular points, defined on molecular graphs. The structural character spectrum of the molecular graph represents the number of the molecular fragments repetitions in a molecular graph by complete enumeration of all pairs, triples, quadruples of the singular points [1,4,5]. As a result, the number of descriptors obtained is very large (about 1000 - 10000) and we must choose how to use them. There are several approaches: the use of descriptors, principal component analysis [8], the selection of features. Then the classifying function on the base of the k-NN is constructed.

2 Problem Statement

The task is to develop a method of descriptors selection of based on the k-NN classifier and compare it with two other approaches, as well as the construction of the classifier function, we denote it by F . $\phi(F)$ - the quality functional, it allows us to determine which function is the best. In our calculations we used 2

types of the $\phi(F)$ - the percentage of correctly classified objects for the molecular sets which have discrete activity and the following formula

$$\phi(F) = 1 - \frac{\sum_{i=1}^n (F(M_i) - A_i)^2}{\sum_{i=1}^n A_i^2}. \quad (1)$$

where $F(M_i)$ is the value of the classifying function on the i -th molecule, A_i is activity of the i -th molecule (numeric expression of the property), N is a number of molecules, for molecular sets which have a numeric activity value.

3 Building of a Classifying Function Based on k-NN

Let each molecule is compared a point in Euclidean space R^M (coordinates are the descriptors values) and the activity value, such as class number, boiling point. Then there are two possible cases:

1. The activity is discrete. Assume for simplicity that should classify molecules into two classes (C_1, C_2), and we have a training set C . In order to determine a class of a new molecule x we need to find the k nearest points to it from C (denote this set as C_x). Let m_1 be a number of points from $C_x \cap C_1$, and m_2 be a number of points from $C_x \cap C_2$. If $m_1 > m_2$, then the new molecule belongs to the first class, if $m_2 > m_1$ the new molecule belongs to the second class.

2. Activity is indiscrete. Suppose we have a training set C . In order to estimate the activity of a new molecule x we must find the k nearest points to it from C and take the average of their activities.

More about k-NN method can be found in [7,8].

Our classification method is based on the assumption that closely related molecules have a similar value of the activity, so distant molecules should not be allowed to affect the result. Given this, we found it necessary to use the limited radius(not to take into account molecules which are located outside the circle of some radius). To determine it we construct a minimal spanning tree for the molecules of the set represented as points in R^M and take an average of its edges length.

4 Algorithm of Descriptors Selection of Based on the k-NN

The input is a matrix of descriptors. Denote it A . The output is a set of descriptors, where $\phi(F)$ has its greatest value. The algorithm consists of four steps:

1. Using the k-NN method we classify molecules on the base of each descriptor of the matrix A . The best n descriptors are selected by the $\phi(F)$ value.

2. For each set of descriptors obtained in the previous step, in turn, we add another column of the matrix A , classify molecules on the base of this new set, and the best n^2 sets of descriptors are selected.

3. Values of ϕ , obtained at the second step, compared with the values on the first. For those sets that do not improve the quality of the forecast, we assume that $\phi = 0$. We select the N best sets.

4. If $\phi = 0$ for all sets or the number of descriptors in the set corresponds to the maximum m, then we end the calculation, otherwise proceed to the second step.

This algorithm can handle a matrix with a huge number of descriptors M. We cant use exhaustive search algorithm due to the the large computational complexity 2^M . When using the algorithm described above leading term in the evaluation of the complexity is nmN^2M , where N is the number of molecules.

5 Results

5.1 The Results on a Set of Amber Odorants

Comparison of the quality of the k-NN classification using the algorithm of selection of descriptors (method 1) , based on all features (method 2) and based on factors (method 3) was conducted on a set of amber odorants (low molecular weight compounds with amber scent), consisting of 129 molecules [2].

8 descriptors matrixes were used. The comparison of the classification quality is given in the table below.

Table 1. The comparison of the methods quality

	Method 1	Method 2	Method 3
1	75.9689	59.6899	59.6899
2	74.4186	59.6899	60.4651
3	75.1937	59.6899	63.5658
4	76.7441	59.6899	59.6899
5	75.9689	60.1562	61.2403
6	77.5193	59.6899	60.4651
7	76.7441	59.6899	61.2403
8	76.7441	59.6899	61.2403

It can be seen from the table that the quality of the algorithm of descriptors selection is better then the quality of the algorithm on all features and all factors. The difference is about 15 percents. But the quality of the algorithm based on all factors is better then the quality of the algorithm based on all features. So the selection of principal components can improve the quality of the forecast, but some pre-selection of descriptors must be done.

5.2 The Results on a Set of Toxic Compounds

First, let us say a few words about the set. Chemical toxicity can cause a lot of biologically hazardous effects such as damage to genes, or even result in death

of man or animals. About 120000 chemical compounds are to be tested in the next 10 years. This will require up to 45 million laboratory animals. An alternative is to provide toxicity prediction based on models built on the previously tested compounds. We studied a set which is divided into two parts. A training set consists of 644 molecules and is used for building models and test set(449 compounds) - for their review. In total there are 5 different feature spaces. On this set the quality of classification was checked, both of them will predict new (not involved in constructing the model) of the molecules. Below is a table with the results.

Table 2. Comparison of the classification quality on the training and the test sets

	Training set	Test set
1	0,6114	0,6180
2	0,7441	0,6873
3	0,7297	0,6886
4	0,7809	0,8090
5	0,7852	0,7582

The table shows that in almost all the feature spaces results on the training set is better than on the test. This is due to the fact that the features used for construction of models are not universal. But, nevertheless, the result did not deteriorated greatly, and so the model can be used for prediction.

6 Conclusion

An algorithm of molecular graphs descriptors selection based on the k-NN classifier solving the structure-property problem was developed and implemented. Computational experiments have shown that the application of this algorithm can significantly improve the quality of classification, and the decreasing of the quality on the learning set is not significant. In the future, we plan to improve the quality of the prediction through the implementation of connection with the construction of descriptors. The algorithm of molecular graphs descriptors selection based on the k-NN classifier is fixed, and we vary the parameters of the construction of the alphabet descriptors. Structural descriptors are taken, and the initial parameters (denoted by T_0) are set. The quality functional $\phi(T_0)$ is computed. Then consider the set $M = \{T | |T - T_0| < \varepsilon, \max_{T \in M} \phi(F(T))\}$ is found. T , where the maximum value is reached, is denoted T_m and the process is repeated until $T_m \neq T_0$. So we can find the best description of molecular graphs for the algorithm.

References

1. Kohov, V.A.: The Method of Quantifying the Similarity of Graphs Based on Structural Spectrs. *Izvestiya RAN, Technical Cybernetics* 5, 143–159 (1994)
2. Svitanko, I.V., Devetyarov, D.A., Tcheboukov, D.E., Dolmat, M.S., Zakharov, A.M., Grigoryeva, S.S., Chichua, V.T., Ponomareva, L.A., Kumskov, M.I.: QSAR Modeling on the Basis of 3D Descriptors Representing the Electrostatic Molecular Surface (Ambergris Fragrances). *Mendeleev Communications* 17(2), 90–91 (2007)
3. Svitanko, I.V., Kumskov, M.I., Tcheboukov, D.E., Dolmat, M.S., Zakharov, A.M., Ponomareva, L.A., Grigoryeva, S.S., Chichua, V.T.: QSAR Modeling on the Base of Electrostatic Molecular Surface (Amber Fragrances). In: 16th European Symposium on Quantum Structure-Activity Relationships and Molecular Modelling, EuroQSAR, Italy (2007)
4. Kumskov, M.I., Smolenskiy, E.A., Ponomareva, L.A., Mitushev, D.F., Zefirov, N.S.: System of structural descriptors for solving structure-property problem. *Reports of the Academy of Sciences* 336(1), 64–66 (1994)
5. Hurst, T., Heritage, T.: HQSAR - A Highly Predictive QSAR Technique Based on Molecular Holograms. In: 213th ACS National Meeting, San Francisco, CA, CINF 019 (1997)
6. Hoffman, B., Cho, S.J., Zheng, W., Wyrick, S., Nicols, D.E., Mailman, R.B., Tropsha, A.: Quantitative structure-activity relationship modelling of dopamine D[1] antagonists using comparative molecular field analysis, generic algorithms-partial least-squares, and K nearest neighbour methods. *J. Med. Chem.* 42, 3217–3226 (1999)
7. Zheng, W., Tropsha, A.: Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbour principle. *J. Chem. Inf. Comput. Sci.* 40, 185–194 (2000)
8. Clark, M., Cramer III, R.D., Jones, D.M., Patterson, D.E., Simeroth, P.E.: Comparative Molecular Field Analysis(CoMFA) Toward Its Use with 3DStructural Databases. *Tetrahedron Comput. Methodol.* 3, 47–59 (1990)