Fuzzy Predicting Models in "Structure – Property" Problem

Eugeny Prokhorov, Ludmila Ponomareva, Eugeny Permyakov, and Mikhail Kumskov

Department of Computational Mathematics, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University **

Abstract. A new approach for analyzing the moleculedescriptor matrix for the QSAR problem (Quantitative StructureActivity Relationship) based on a fuzzy cluster structure of the learning sample is presented. The ways for generating fast rules for refusing prediction and searching the spikes in the learning sample are described. For this purpose, a special space of descriptors, simple for calculation, is introduced. The ways for optimizing the discriminant function according to fuzzy clustering parameters are examined. Highly predictive models based on the presented approach have been generated. The models are compared, and the efficiency of the described methods is revealed.

1 Introduction

The solution of the QSAR problem consists of two stages: the stage of description and the stage of discriminant function generation [3]. Very often the learning sample is separated into clusters, and each cluster is processed separately. In fact the cluster analysis of the learning sample determines the discriminant function generation. The method presented makes it possible to optimize the discriminant function with respect to clustering parameters. For screening a large database of compounds, it is extremely important to generate the rules for refusing prediction, and the rules should be fast in terms of computation. Fuzzy clustering makes it possible to remove the main disadvantages intrinsic to classical methods and to choose the discriminant function in wider, generic classes.

2 Problem statement

Detailed problem statement is given in [3]. It will specify the problem of constructing fast rejection rules. Let the alphabet of descriptors consists of M elements [2]. Feature vectors of the molecular graph G is called a vector $x = (x_1, \dots x_M) \in R^M$, where x_i – the value of the i-th descriptor computed for G. Describing the mapping $D : G \to R^M$ is called a map that assigns M–graph to his feature vector. Space R in this case is called the space of descriptors. We

** The work is supported by the grant RFFI 10-07-00694

call classifying function $F : R^M \to \{C_i\}_{i=1}^H$ that receives as an argument to a feature vector $x = (x_1, \dots x_M)$ of an arbitrary molecular graph G, and assigns the corresponding M-graph to the to one of the classes of activity C_i. Sometimes it is convenient to classifying function was defined on the set of molecular graphs. When an argument F to specify an M-graph, one should realize that F is computed on the corresponding feature vector. We set by definition $F(G_i) = F(D(G_i))$, where D - describing the mapping from the set of M-graphs in the space of descriptors.

Let the fixed algorithm Alg for constructing classifying function F on the training sample $\{(G_i, C_i)\}_{i=1}^N$, predictive model is called the set of training set $\{(G_i, C_i)\}_{i=1}^N$ and algorithm Alg construct classifying function $F = Alg(\{(G_i, C_i)\}_{i=1}^N)$.

To assess the predictive ability of models used the coefficient of cross-validation [1, 3]. Now we formulate the problem of constructing rejection rules: Rejection rule is called one or more functions $g : R^M \to \{0, 1\}$ with the following interpretation: $g(G_i) = 1$ will constitute a reject of prediction activity of this molecular graph, otherwise the prognosis can be made. Let $g(G_i) := g(D(G_i))$, where D - describing the mapping from the set of M-graphs in the space of descriptors R^M. Is called a molecular graph G_i is admissible if in accordance with accepted rejection rules, it belongs to the range of admissible argument for the classifying function F. I.e. $g(G_i) = 0$. Let $O = \{(G_i, C_i)\}_{i=1}^N$ – training sample, denoted

$$\tilde{O} = \{(G_i, C_i)\}_{i=1}^N \setminus \{(G_j, C_j) | g(G_j) = 1, j = 1, \dots N\}$$

– a sample composed only of admissible M-graphs learning sample O. We call rejection rule strong if it satisfies the inequality $R_{cv}^2(O, Alg) < R_{cv}^2(\tilde{O}, Alg)$.

3 Solution method

The idea is to use the cluster structure of the original training set for building rejection rules for this compound. This is not only the ejections that simply do not fall into one cluster, but also about the compounds, predict the activity of which should not be on the more complex reasons based on the cluster structure. For example, molecules that belong equally to two clusters, the models which predict its activity in different ways. In addition, the important point is the need to determine the admissibility of the molecular graph with minimal computational cost. It is therefore proposed to compute rejection rules on special space of descriptors, much smaller dimension than the original, for example, only topological. Thus, we construct 2 of the space of descriptors, one – to build a rejection rules, another – to do the classification and predict of activity. It arises naturally reduced (special) the matrix molecule - descriptor, whose rows are the vectors in a special space of descriptors. Regarding the fuzzy classifying function, the approach is as follows. Apply a fuzzy clustering algorithm (c-means fuzzy, or any other) [2]. Fuzzy clustering techniques, in contrast to the clear methods allow the same object simultaneously belong to multiple clusters, but with varying degrees [2, 7]. Fuzzy clustering in many situations, the more "natural" than the clear,

for example, for facilities located on the border of the clusters. Fuzzy clusters describe the following matrix of the fuzzy partition:

$$S = [\mu_{ij}], \quad \mu_{ij} \in [0,1], \quad i \in \{1,\dots N\}, \quad j \in \{1,\dots k\}$$

in which the i-th row contains the degree of membership of an object to clusters $S_1,\dots S_K$.

The only difference between the matrix of the fuzzy partition and the corresponding matrix of clearly partition that, when the fuzzy partition the degree of membership of the object to the cluster takes values from the interval [0, 1], and when the clearly - from the two-element set 0, 1. Now, with the partition of the original space on the fuzzy clusters, within each construct local predictive model (we assume that this is linear regression, but can be used any other algorithm) [3, 4]. Suppose, for simplicity, we have 2 possible values of activity: active / inactive, denote their respective numerical values 1 and -1. For the new compound $\tilde{x} = (\tilde{x}_1,\dots \tilde{x}_M)$ we have k predictions of activity in accordance with the number of clusters (models). Let the i-th model gave a prediction R_i, then we can calculate the resulting prediction from the formula:

$$\tilde{y} = \frac{\sum_{i=1}^{k} R_i \mu_i}{k},$$

where μ_i – the degree of membership of the molecule to the i-th cluster. You can specify the scope of the normalization of response $\tilde{y}$, for example

$$\tilde{y} < -0.5 \Rightarrow \tilde{y} = -1,$$

$$\tilde{y} > 0.5 \Rightarrow \tilde{y} = 1,$$

otherwise $\tilde{y} = 0$ – the rejection of predict.

We now consider the optimization of fuzzy classifying function in the parameters of fuzzy clustering. In [5] described several methods for constructing the cluster structure of the training sample. We are interested in "fuzzy" generalization of the cluster structure for the application of this approach. Let discovered cluster structure of the original training set taking into account the removal of ejections. Suppose, as before, the number of clusters k, and is known for a clear partition of the matrix: $S = [\nu_{ij}], \quad \nu_{ij} \in \{0,1\}, \quad i \in \{1,\dots N\}, \quad j \in \{1,\dots k\}, \quad \sum_{j=1}^{k} \nu_{ij} = 1, i \in \{1,\dots N\}, \quad 0 < \sum_{i=1}^{N} \nu_{ij} < N, \quad j \in \{1,\dots k\}.$ in which the i-th row contains information about an object $x_i = (x_{i1},\dots x_{iM})$ belonging to one of the clusters $S_1,\dots S_k$. Assume also that each cluster is given by its center $Z_i = \{c_i 1,\dots c_{iq}\}$ – a subset of the points of the cluster S_i, center point is called the cores of the cluster, and the radius is called $r_i = max_{x_j \in S_i} \rho(x_j, Z_i)$. We construct a matrix of fuzzy partition $\tilde{S} = [\mu_{ij}]$, in which the i-th row contains the degree of membership of an object $(x_{i1},\dots x_{iM})$ to the clusters $\tilde{S}_1,\dots \tilde{S}_k$. Optimization parameters will be $\lambda_1, \lambda_2 \in R^M, \quad \lambda_1 \le 1 \le \lambda_2$. We define small and large cluster $\tilde{S}_i$ radius as $r_i^1 = \lambda_1 r_i$ and $r_i^2 = \lambda_2 r_i$ respectively. Then the elements of the matrix $\tilde{S} = [\mu_{ij}]$ calculated by the formula:

$$\mu_{ij} = 1, \qquad if \quad \rho(x_i, Z_j) < r_i^1,$$

$$\mu_{ij} = 0, \qquad if \quad \rho(x_i, Z_j) > r_i^2,$$
$$\mu_{ij} = \frac{r_j^2 - \rho(x_i, Z_j)}{r_j^2 - r_j^1} \qquad otherwise.$$

Membership function of the point to cluster can also be nonlinear and contain additional optimization options. This approach allows us to meaningfully use the cluster structure of the sample is not limited in this case beyond a single cluster.

4 Results

This algorithm was implemented and applied to three samples – amber odorants, glycosides, and toxic compounds. In constructing the models used algorithm of evolutionary selection of descriptors [4], we used a set of standard clustering algorithms such as hierarchical clustering, k-means, etc Method showed significant improvement in prediction quality in the construction of simple local models. Optimization of the fuzzy classifier function by an average of 5 % improves the prognosis. In all three samples were obtained for models with high predictive ability. Along with the fast rejection rules our models can be used for subsequent screening of databases of chemical compounds [5] in order to identify compounds having the property under consideration.

5 Conclusion

These results demonstrate the practical significance of the proposed in paper approach. New methods have yielded predictive models of high quality. In most cases, a significant improvement in prediction quality as compared with classical methods. Interests are other parameterizations of fuzzy cluster structure of the training set and optimization of fuzzy classifying function for the new parameters. Continuations of the work are also testing fast rejection rules and a fuzzy classifying function in the screening of large databases of compounds with unknown activity.

References

[1] Devetyarov D.A., Grigorieva S.S., Permyakov E.A. Kumskov M.I., Ponamoreva L.A., Svitanko I.V. — Solution to the problem “structure – property” for molecules with multiple spatial conformations. // System of predicting the properties of chemical compounds: Algorithms and Models: Collected Works, Ed. MI Kumskov. Moscow: MAKS Press, 2008 (in Russian).

[2] Shtovba S.D. Introduction to the theory of fuzzy sets and fuzzy logic. Vinnitsa: Publishing the Vinnitsa State Technical University, 2001. - 198. (in Russian).

[3] Prokhorov E.I., Perevoznikov A.V., Voropaev I.D., Kumskov M.I., Ponomareva L.A. - Search representations of molecules and methods of prediction activity in the problem of “structure – property” // Reports of the 14th All-Russian Conference ”Mathematical Methods for Pattern Recognition” MMRO-2009. — Moscow: MAX Press. — 2009. — S. 589-591 (in Russian).

[4] Devetyarov D.A., Kumskov M.I. Apryshko G.N., Noseevich F.M. et al — Comparative analysis of fuzzy descriptors in solving the "structure-property" problem for a sample of glycosides // 14-th All–Russia. Conf. MMRO-14. — M.: MAKSPress, 2009. — C. 575-578 (in Russian).

[5] Prokhorov E.I., Perevoznikov A.V., Ponomarev L.A., Kumskov M.I. Neural network as a tool to implement a piecewise linear classifier for mass screening of molecules in "structure – property" problem. // Neurocomputers: development, application. — 3. — S. 39-45 (in Russian).

[6] M.I. Kumskov, E.A. Smolensky, L.A. Ponomareva, D.F. Mityushev, N.S. Zefirov. Systems of Structural Descriptors for QSAR Problem Solving// Proceedings of the Academy of Science – 1994. — Vol. 336, No. 1. — P. 64-66 (in Russian).

[7] L.A. Zadeh. Fuzzy Sets. Information and Control. — 1965. — P. 338–353.

[8] Sergei O. Kuznetsov, Mikhail V. Samokhin: Learning Closed Sets of Labeled Graphs for Chemical Applications. ILP 2005: 190-208