Fuzzy Classification and Fast Rules for Refusal in the QSAR Problem

E. I. Prokhorov, L. A. Ponomareva, E. A. Permyakov, and M. I. Kumskov

Faculty of Mechanics and Mathematics, Moscow State University, Moscow, 119992 Russia
e-mail: qsar_msu@mail.ru, eugeny.prokhorov@gmail.com

Abstract—A new approach for analyzing the "molecule–descriptor" matrix for the QSAR problem (Quantitative Structure–Activity 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.

DOI: 10.1134/S105466181102091X

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.

PROBLEM DEFINITION

The problem is defined in detail in [3]. Here we define more exactly how to generate the fast rules for refusing prediction. Let the descriptors' alphabet consist of M elements. The *feature vector* of molecular graph G is the $x = (x_1, ..., x_M) \in R^M$ vector, where x_i is the value of the i-th descriptor calculated for G. *Describing mapping* D: $\{G\} \longrightarrow R^M$ is the mapping that associates the M-graph with its feature vector. In this case R^M space is *the space of descriptors*. Let *the discriminant function* be the function F: $R^M \longrightarrow \{C_i\}_{i=1}^H$, the argument of which is the feature vector $x = (x_1, ..., x_M) \in R^M$ of the arbitrary molecular graph G, and which refers the compound corresponding to the M-graph to a certain class of activity C_i. Sometimes it is convenient to specify the discriminant function on the set of molecular graphs. If graph F is designated as an argument, it means that F is calculated on the respective feature vector. Let us define $F(G_i) := F(D(G_i))$, where D is the describing mapping from the set of M-graphs to the space of descriptors R^M. Let the algorithm *Alg* for generating the discriminant function F according to learning sample $\{(G_i, C_i)\}_{i=1}^N$ be fixed; *the predictive model* is the complex consisting of learning sample $\{(G_i, C_i)\}_{i=1}^N$ and algorithm *Alg* for generating the discriminant function $F = Alg(\{(G_i, C_i)\}_{i=1}^N)$.

To estimate the predictive ability of the models, the cross validation coefficient R_{cv}^2 is used [1, 3].

Let us formulate *the problem for generating the rules for refusing prediction:*

The rule for refusing is one or several functions g: $R^M \longrightarrow \{0, 1\}$ with the following interpretation: $g(G_i) = 1$ means the refusal to predict the activity of the given molecular graph; otherwise, the prediction is possible.

Let $g(G_i) := g(D(G_i))$, where D is the describing mapping from M-graphs to the space of descriptors R^M.

Let the molecular graph G_i be *permissible,* if according to accepted rules for refusal, it belongs to the region of acceptable arguments of the discriminant function F, i.e., $g(G_i) = 0$.

Let $O = \{(G_i, C_i)\}_{i=1}^N$ be the learning sample; in this case

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

is the sample consisting only of permissible M-graphs of learning sample O.

Received December 26, 2010

Let the rule for refusal g: $R^M \longrightarrow \{0, 1\}$ be *strong*, if the following inequality $R^2_{cv}(O, Alg) < R^2_{cv}(\tilde{O}, Alg)$ is true.

METHOD OF SOLUTION

The idea is as follows: to use the cluster structure of the initial learning sample for generating the rules for refusal prediction for the given chemical compound. Here we speak not only about spikes, which simply do not hit into any cluster, but about compounds the activity of which should not be predicted due to finer reasons dealing with cluster structure. For example, for molecules belonging equally to two clusters, the models predict its activity in different ways.

In addition, it is necessary to determine the molecular graph validity by minimal computational burden. For this reason, we suggest to calculate the rules for refusal in the special space of descriptors with a significantly lower dimension than the initial one, for example, only topological.

Hereby, two spaces of descriptors are generated: one space is for generating the rules for refusal, and the second is for activity classification and prediction. Here the *reduced (special) matrix "molecule–descriptor,"* the rows of which are the vectors in the special space of descriptors, appear in a natural way.

As for the fuzzy discriminant function, the approach is as follows. We use a certain fuzzy classification algorithm (c-mean fuzzy or any other one) [2]. The fuzzy methods for clustering, in contrast to the well-defined method, allow the same object be the membership of several clusters simultaneously, but with different degrees. In many situations the fuzzy clustering is more "natural" than the accurate one, for example, for the objects placed at cluster edges.

The fuzzy clusters are described by the following *matrix of a fuzzy partition*

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

in which the i-th row contains the grade of membership of $(x_{i1}, \ldots, x_{iM})$ object to clusters $S_1, \ldots, S_k$. The only difference of the fuzzy partition matrix from the respective accurate partition one is as follows: under fuzzy partition the grade of object membership to the cluster possesses values from the interval [0,1], and under accurate partition, from the set {0, 1}.

Now the initial space is partitioned into fuzzy clusters, and inside each cluster we generate the local classifying model (hereinafter we accept that it is the linear regression, but any other algorithm is also possible) [3, 4]. Let us for simplicity have two possible values of activity, i.e., active and inactive, and let us designate them by the respective numbers 1 and –1.

For the new compound $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_M)$, we have k predictions for activity in accordance with the number of clusters (models). Let the i-th model give R_i predictions. In this case it is possible to calculate the resulting prediction as follows:

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

where μ_i is the coefficient of the given molecule membership to the i-th cluster. It is possible to set the frames of the response normalizing $\tilde{y}$, for example, $\tilde{y} < -0.5 \Rightarrow \tilde{y} = -1$, $\tilde{y} > 0.5 \Rightarrow \tilde{y} = 1$, otherwise $\tilde{y} = 0$: refuse prediction.

Now let us examine the problem for optimizing the fuzzy discriminant function over fuzzy clustering parameters.

Several methods for generating the cluster structure of the learning sample are described in detail in [5]. We are interested in fuzzy generalization of this cluster structure for applying the described approach.

Let us have already determined the cluster structure of the initial learning sample by considering the fact that the spikes have been deleted. Let the number of clusters be k and the *accurate partition matrix* be known:

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

$$\sum_{j=1}^{k} \nu_{ij} = 1, \quad i \in \{1, \ldots, N\},$$

$$0 < \sum_{i=1}^{N} \nu_{ij} < N, \quad j \in \{1, \ldots, k\}.$$

In this matrix the i-th row contains information whether the object $(x_{i1}, \ldots, x_{iM})$ has the membership in one of the $S_1, \ldots, S_k$ clusters.

Let each cluster be set by its center $Z_i = \{c_{i1}, \ldots, c_{iq}\}$, i.e., by a certain subset of points of cluster S_i. We call the center points the cores of the given cluster, and $r_i = \max_{x_j \in S_i} \rho(x_j, Z_i)$ is its radius.

Let us generate the fuzzy partition matrix $\tilde{S} = [\mu_{ij}]$, in which the i-th row contains the grade of object $(x_{i1}, \ldots, x_{iM})$ membership to clusters $\tilde{S}_1, \ldots, \tilde{S}_k$. The optimization parameters are $\lambda_1, \lambda_2 \in R$, $\lambda_1 \le 1$, $\lambda_2 \ge 1$. Let us determine the minor and major radii of the $\tilde{S}_i$ cluster as $r_i^1 = \lambda_1 r_i$ and $r_i^2 = \lambda_2 r_i$, respectively. In this case we calculate the elements of matrix $\tilde{S} = [\mu_{ij}]$ as follows:

$$\mu_{ij} = \begin{cases} 1, \text{ if } \rho(x_i, Z_j) < r_i^1 \\ 0, \text{ if } \rho(x_i, Z_j) > r_i^2 \\ \dfrac{r_j^2 - \rho(x_i, Z_j)}{r_j^2 - r_j^1}, \text{ otherwise.} \end{cases}$$

The function of point membership in the cluster can also be nonlinear and contains additional parameters of optimization. Such an approach makes it possible to use the cluster structure of the sample properly, and we do not limit ourselves to only one cluster.

RESULTS

The presented algorithm was implemented and applied to three samples: amber odorants, glycosides, and toxic compounds. During model generation the evolution selection of descriptors [4] was performed, and for clustering the set of standard algorithms, such as hierarchical clustering and k-mean, were used. The presented method makes it possible to increase significantly the prediction quality under generating the simple local models. If the process of fuzzy discriminant function generation is optimized, the prediction quality increases approximately by 5%. For all three samples, models with high (for QSAR problem) predictive ability are generated. Both the fast rules for refusal and the models can be used for screening the databases of chemical compounds [5] to reveal the compounds characterized by the examining feature.

CONCLUSIONS

The presented results demonstrate the practical significance of the described approach. The new methods make it possible to generate highly predictive models. In the majority of cases, the prediction quality is increased significantly in comparison with classical methods.

Other parameterizations of fuzzy cluster structure of the learning sample and fuzzy discriminant function optimization according to new parameters are of interest.

This work must be continued as follows: to test the fast rules for refusal and the fuzzy discriminant function under screening a large databases of compounds with unknown activity.

ACKNOWLEDGMENTS

This project was supported by the Russian Foundation for Basic Research, project no. 10-07-00694.

REFERENCS

1. D. A. Devet'yarov, S. S. Grigor'eva, E. A. Permyakov, M. I. Kumskov, L. A. Ponomareva, and I. V. Svitanko, "QSAR Problem Solution for Molecules with Spatial Conformation Set," in *The System for Predicting Chemical Compounds Properties: Algorithms and Models: Collection of Scientific Papers*, Ed. by M. I. Kumskov (MAKS Press, Moscow, 2008) [in Russian].
2. S. D. Shtovba, *Introduction into the Theory of Fuzzy Sets and Fuzzy Logics* (Izd. Vinnitskogo Gos. Tekhn. Univ., Vinnitsa, 2001) [in Russian].
3. E. I. Prokhorov, A. V. Perevoznikov, I. D. Voropaev, M. I. Kumskov, and L. A. Ponomareva, "Search for Molecular Representation and Methods for Predicting the Activity in QSAR Problem," in *Proc. 14th All-Russian Conf. "Mathematical Methods for Image Recognition" MMPO–2009* (MAKS Press Moscow, 2009), pp. 589–591.
4. D. A. Devet'yarov, M. I. Kumskov, G. N. Apryshko, F. M. Nosevich, et al., "Comparative Analysis of How to Use the Fuzzy Descriptors for Solving the "Structure–Activity" Problem for Glycosides Production," in *Proc. 14th All-Russian Conf. MMPO-14* (MAKS Press Moscow, 2009), pp. 575–578.
5. E. I. Prokhorov, A. V. Perevoznikov, L. A. Ponomareva, and M. I. Kumskov, "Neural Network As an Instrument for Implementing the Piecewise–Linear Classifier under Mass Molecules Screening in QSAR Problem," Neirokomput.: Razrab., Prom., No. 3, 39–45 (2010).