Prediction of properties of molecular graphs based on RBF neural networks

Mikhail I. Kumskov (*IEEE member)*, Varvara O. Vasilyeva and Ekaterina O. Rybakova

Abstract—A method for forming the architecture of an RBF neural network for solving the problem of predicting the properties of molecular graphs is proposed and investigated. The method is based on the identification of the cluster structure of the training dataset in the
"structure-property" problem. The results of the method study for predicting the properties of molecules for various variants of the attribute description of molecular graphs are presented. The RBF neurons of the network cover the identified clusters, the placement of neurons is based on the k-means algorithm for each identified cluster.

A technique for forming the architecture of a neural network is substantiated based on a preliminary analysis of the training data set and the assignment of RBF neurons to the elements of the training set. The obtained results are comparable with the previously implemented methods, which were based on an evolutionary algorithm - Group Method of Data Handling (GMDH).

Index Terms—Quantitative Structure-Activity Relationship, RBF neural network, Group Method of Data Handling, k-means, Cluster Analysis

1 Introduction

1. Most of the organic synthesis takes place without prior modeling and is based on the experience and knowl-edge of a chemist, as a result, only about one out of 5,000 drugs reach clinical trials and even fewer reach the final goal. This paper presents a way to accelerate the described process by constructing models linking the structure and properties of substances.

There is a scientific d iscipline c alled chemoinformatics, which studies the application of computer science methods to solve chemical problems. G. Paris from Novartis gave it the following definition: c hemoinformatics i s a scientific d iscipline c overing t he d esign, c reation, organization, management, search, analysis, dissemination, visualization and use of chemical information. The fields o f application of chemoinformatics are as follows: prediction of physicochemical properties of chemical compounds (in particular, lipophilicity, water solubility), properties of materials, toxicological and biological activity, ADME/T, ecotoxicological properties, development of new drugs and materials.

In this paper, we will consider the problem of classification of biological activity, this task is called Quantitative Structure Activity Relationships (QSAR). In the computer prediction of the properties of chemical compounds problem there is also a regression problem, that is, in this class of problems it is necessary to predict the value of a physical or chemical property, these tasks are called Quantitative Structure Properties Relationships (QSPR).

To build QSAR models, the chemist determines the features focused on the analysis of a specific p roperty in a number of chemical compounds. These signs describe some structural features of molecules.

2 Problem statement

First, let's introduce the necessary definitions.
For an arbitrary set V, we denote by V_k the set of all subsets of k-elements of the set V. For example, the set V_2 coincides with the set of disordered pairs of different elements of V.

Definition The graph G is a triple $G = (V, E, \delta)$ consisting of the sets V, E and the mapping $\delta : E\ toV_1 \cup V_2$. Elements from V are called vertices of the graph G, elements from E are edges of the graph G and δ is a boundary or incidence mapping.

Consider a labeled graph whose vertices are interpreted as atoms of a molecule, and whose edges are interpreted as valence bonds between them.

Vertex and edge labels encode atoms (their properties) and bond types, respectively. For example, the vertices can store information about three-dimensional coordinates, the symbol of a chemical element, the charge of the nucleus,

- *M. Kumskov is with the Department of Computational Mathematics, Lomonosov Moscow State University, Moscow, 119992. E-mail:* ***mikhail.kumskov@math.msu.ru***
- *V. Vasilyeva is with the Department of Computational Mathematics, Lomonosov Moscow State University, Moscow, 119992. E-mail:* ***varvaravas@gmail.com***
- *E. Rybakova is with the Department of Computational Mathematics, Lomonosov Moscow State University, Moscow, 119992. E-mail:* ***catrybakova@gmail.com***

Manuscript received November 11, 2022;

polarizability, atomic weight, atomic radius. And in the edge labels – multiplicity, lengths, orders of connections.

There is also an activity label that we will predict. Then the answer space $\mathbf{Y}$ consists of two possible elements: -1 (compound is inactive) and +1 (compound is active).

Definition Training set X in "structure-activity" problem is a finite set of $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ pairs from $\mathbb{X} \times \mathbb{Y}$. For a given sample, the vector $(y_1, \ldots, y_n)$ is called the target vector.

Definition The descriptor is an invariant characteristic of the molecular graph G, which has a numerical value. The descriptor alphabet is a finite set of all the different descriptors used to analyze the training sample. Vector of m-dimensional features $(x_1, \ldots, x_m) \in R^m$ is assigned to the molecular graph G, where x_j is the value of the j-th descriptor from the descriptor alphabet calculated for G.

Definition The "molecule-descriptor" matrix is a matrix of size $n \times m$, the i-th row of which is the feature vector of the i-th molecular graph from the training set X.

Let's give a training sample $\mathbb{X}$, that is, a database of n chemical compounds, each of which is represented as a molecular graph.

Thus, it is required:

- to construct a Molecule-Descriptor matrix of size $n \times m$, where m – number of descriptors for a given compound, n – number of compounds;
- to construct a function $F(x_1, \ldots, x_M)$ that receives a new compound (as a matrix string) and relates it to one of the activity classes or predicts the numerical value of the property (for the "structure-property" problem). Which of the classifying functions is "better", allows you to determine the quality functional $\varphi(F)$ [3].

Our specific formulation of the problem will be the choice of a feature description and the construction of a dependency, that is, the choice of an algorithm that predicts the value of the studied property based on the vector of the properties of the molecule.

2.1 Feature description and labeling

To date, the theory of constructing and using a set of descriptors has been described. At the same time, further deepening of ideas about the molecular structure makes it possible to create new descriptors and models reflecting these ideas. Consider the hierarchy of descriptors used to describe the chemical structure.[4]

TABLE 1: Example of a descriptor hierarchy

Descriptor class	Descriptor types
Elementary level descriptors	1.The number of atoms of the same grade 2.Atomic weights of structure fragments
Descriptors of the structural formula	1.Topological indexes 2.Structural fragments
Electronic level descriptors	1.Partial charges on atoms 2.Molecular refraction 3.The energies of the highest occupied and lowest unoccupied orbitals
Descriptors of intermolecular interactions	1.Gamete constants 2.Steric constants

The construction of "structure-activity" dependencies using the spatial representation of the molecules of the training sample was called 3D-QSAR. The main stages of 3D-QSAR modeling are as follows:

1) Select molecules in the training database, each of which has an activity experimentally measured for this biological system;
2) To construct spatial representations of these molecules and to carry out their "alignment" according to the given rules for choosing orientations;
3) Calculate a set of spatially dependent features for all molecules;
4) Construct a function expressing the dependence of the calculated features and the studied biological activity;
5) To determine the stability and predictive ability of the found functional dependence;
6) If necessary, modify the model by repeating steps 1-5.

The choice of the method of describing the structure is determined by the nature of the problem being solved and the existing limitations on obtaining experimental and calculated data. However, it is already clear that the number of descriptors for a specific task can be huge (much more than a sample of substances), since at this stage it is not known which specific features are best used to solve the problem. Therefore, in addition to the two tasks described in paragraph ¡2¿, it is necessary to determine the most significant features for this task.

One of the basic techniques for reducing the dimension of the feature vector is labeling. Let the vertices of the graph have signs $(a_1, \ldots, a_k)$. Then we can build clusters by some properties (for example, (a_1, a_2)) using cluster analysis methods and get one mark instead of two, three or more vertices, meaning belonging to a particular cluster.

2.1.1 Understructural molecular descriptors

A large group of approaches to the description of substances in the form of a feature vector is associated with the fragmentation of the molecular structure, that is, with its decomposition into some fragments-descriptors selected from the condition of simplicity of isolation or substantive

considerations. Substructural analysis is based on the assumption that the biological effect of the substance is due to the presence of some structural elements (substructures) in its composition. The connection structure is represented by a fragmentary code. Among the types of fragments used , the following can be noted:

1) atoms and pairs of bonded atoms;
2) attached atomic fragments having a central atom characterized by its bonds and atoms attached to it;
3) cyclic fragments characterizing the shape of cycles and the position of heteroatoms in them;
4) fragments of ″heteropaths″ describing chains in a molecule that begin and end with heteroatoms;
5) fragments in Viswesser linear notation;
6) fragments of an expanded language of ″gross bond formulas″ with the introduction of microfragment modifications;
7) substructures characteristic of the studied SBD.

The use of substructures is a typical example of an approach to solving complex problems that depend on the representation of an object. There are various ways to classify fragments, when each type of substructure identifies a certain aspect of the molecule, and the use of one or another type of fragments depends on the nature of the problem being solved by QSAR.[5]

As you can see, there are a huge number of methods for selecting molecular descriptors. The method we have chosen to solve the problem will be described in Part ¡3.1¿.

2.2 Searching for functional dependence

Classical machine learning methods are now used to search for functional dependence. For example, linear regression, regression by main components (Analysis of main components, PCA), regression by ridge. We also use kernel methods (kernel image recognition), the Support Vector Machine (SVM) method. Recently, with a sharp increase in interest in neural networks, they are trying to solve problems using artificial neural networks (ANN), the k-nearest neighbor method (KNN), etc. In this paper, the GMBH method will be used, which is described in section 3.2

The general formulation of the QSAR modeling problem is considered, which consists in predicting the numerical value of a chemical property or its presence based on knowledge of these values for other compounds. One of the important tasks is to choose the right features (descriptors) to build a prediction algorithm. An overview of various methods of selecting descriptors is made, and a brief overview of the 3D-QSAR method using the spatial structure of the molecule is presented. Also an important component of the formulated task is the search for functional dependence. Popular approaches to solving the problem are described.

3 Description of the solution steps

3.1 Construction of the Molecule-Descriptor matrix

This section describes the approach proposed in [1], which we will follow when vectorizing graphs and forming the ″molecule-descriptor″ matrix. The columns of this matrix will correspond to structural fragments — symbolic names consisting of special codes of atoms. The exact definition is given later in this section.

We introduce symbolic markers designed to account for the topological and chemical features of atoms in molecular structures. In classical theory, in the simplest case, atoms are distinguished based on two characteristics: chemical individuality and valence. To distinguish the topological features of atoms in the graph, we will use the atom degree characteristic (the number of edges at the corresponding vertex of the molecular graph) and introduce the corresponding marker p (p = ″power of atom″), which can take (for classical organic compounds) seven values (0, 1, 2, 3, 4, 5, 6). In this case, $p = 0$ only for disconnected (″single″) atoms in the M-graph. One marker of degree p is not enough to distinguish all variants of the ″valence environment″ of a carbon atom. We introduce a marker of the chemical bond of atom b, which we define as follows:

- ″s″ (single) – all bonds of an atom are single,
- ″d″ (double) – an atom has a double bond,
- ″t″ (triple) – an atom has a triple bond,
- ″w″ – an atom has two double bonds,
- ″a″ (aromatic) – an atom has an aromatic bond.

In addition to the p and b markers, we will use another marker r (ring) - a marker of the position of the atom in the ring system.

A graph is called connected if any of its two vertices can be connected by a chain. We will call the connection (edge) of a molecular graph ring, if when it is removed the connectivity of the graph is not violated, and chain (acyclic) - otherwise. If an atom has ring bonds, we will call it a ″ring atom″. Among the ″ring″ atoms, we will distinguish between ″purely ring″ atoms, i.e., atoms whose edges are all annular, and ″ring with a substituent″ atoms that have an acyclic edge. Let's define the r-marker as follows:

- ″c″ (chain) – atom is acyclic (chain),
- ″r″ (ring) – ring atom without substituent,
- ″s″ (substitute) – ring atom with a substituent.

The atom label, which includes markers, will be written as a string of the following form:

Atom name Marker p Marker b Marker r,

or briefly – NNpbr.

The atom name is written with an uppercase character, the p marker – with a digit, the b and r markers – with lowercase letters. If the marker is not used (we say, ″off″), then the symbol ″$*$″ is used instead of its designation.

Let's give an example of marking the atoms of the caffeine molecule. For a more convenient representation, let us number its atoms.

TABLE 2: An example of labeling atoms of a caffeine molecule

Atom number	0	1	2
Structure	$-CH_3$	$-\overset{\vert}{N}-$	$=CH-$
Marking	C_1sc	N_3as	C_2ar
Atom number	3	6	7
Structure	$-N=$	$-\overset{\Vert}{C}-$	$O=$
Marking	N_2ar	C_3as	O_1dc

Definition A structural fragment of length k is a chain of k of labeled atoms, and each subsequent atom is adjacent to the previous one, i.e. each pair of consecutive atoms in the chain corresponds to a chemical bond in the molecule. Strings of length 2, for example, are encoded as follows: NNpbrNNpbr.

Table 3 shows examples of coding of structural fragments of length 2. The r marker is off, since the fragments shown do not give a complete picture of the structure of the molecule (the presence of aromatic rings).

TABLE 3: An example of encoding fragments

Fragment	$-\overset{\vert}{\underset{\vert}{C}}-C\equiv$	$=C=\overset{\vert}{\underset{\vert}{C}}$
Code of fragment	C_4s*C_2t*	C_2w*C_3d*
Fragment	$=C=C=$	$\overset{\vert}{\underset{\vert}{C}}H-\overset{\vert}{\underset{\vert}{C}}-$
Code of fragment	C_2w*C_2w*	C_3s*C_4s*

3.1.1 Topological indexes

The topological index is an invariant of the molecular graph, a certain numerical value that characterizes the structure of the molecule as a whole. Usually, topological indices do not reflect the multiplicity of chemical bonds and types of atoms (C, N, O, etc.), hydrogen atoms are not taken into account.

We will use topological indices to conduct a preliminary visual development of the training sample, search for interesting dependencies and identify a possible cluster structure.

For an arbitrary molecular graph G, we introduce the concepts of adjacency and distance matrices (atoms are numbered arbitrarily).

Definition The adjacency matrix of a molecular graph G of n atoms is a matrix of size $n \times n$, on (i, j)-th place of which is 1, if between i-th and j-th atoms there is a chemical bond, or 0, otherwise.

Definition The matrix of distances of the molecular graph G of n atoms is a matrix of size $n \times n$, at the (i, j) place of which there is a number equal to the topological distance (the number of edges along the shortest path) between the i-th and j-th atoms.

The distant matrix is filled by running the wave algorithm proposed in [5]. A step-by-step description and pseudo-code of the algorithm are presented in [1].

Algorithm .1. Wave for an individual vertice

```
f ← 1
mark_num ← 1
mark[ii] ← mark_num                          ▷ burning
while at least 1 vertice is not burning do
    if f = 0 then
        break
    end if
    f ← 0
    for i = 0, ..., n − 1 do
        if mark[i] = mark_num then
            for j = 0, ..., n − 1 do
                if a[i][j] = 1 and mark[j] = 0 then
                    mark[j] ← mark_num + 1
                end if
            end for
        end if
    end for

    for i = 0, ..., n − 1 do
        if mark[i] == 0 then
            f ← 1
        end if
    end for
    mark_num ← mark_num + 1
end while
```

The following is a description of some well-known molecular graph invariants that can be calculated through the adjacency and distance matrices.

1) Wiener index[6]

$$W(G) = \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} d_{i,j},$$

where $d_{i,j}$ is the topological distance between the i-th and j-th atoms in the molecule, or (i, j)-th element of the distance matrix.

2) Randic index[7]

$$R(G) = \sum_{(v_i, v_j) \in V} \frac{1}{\sqrt{d(v_i)\, d(v_j)}},$$

where v_i and v_j are adjacent vertices forming an edge (v_i, v_j), $d(v_k)$ is a vertex degree v_k.

3) Balaban Index[8]

$$J(G) = \frac{q}{\mu + 1} \sum_{(v_i, v_j) \in V} \frac{1}{\sqrt{s_i\, s_j}},$$

where v_i and v_j are adjacent vertices forming an edge (v_i, v_j), s_i is the sum of the elements of the i-th row or i-th column of the distance matrix, μ is a cyclomatic number calculated by the formula $\mu = q - n + 1$ (this is the least number of edges, the removal of which leads to graph without cycles).

4) Diameter and radius of a molecular graph

$$diam(G) = \max_i \max_j d_{i,j}, \quad rad(G) = \min_i \max_j d_{i,j},$$

where $d_{i,j}$ is the (i,j)-th element of the distance matrix.

3.1.2 Structural 3D descriptors

The approach, called 3D-QSAR, is based on the use of a spatial representation of molecules. Descriptors are constructed not from the initial molecular graphs, but from the three-dimensional graphs derived from them, at the vertices of which are not atoms, but the so-called singular points. Only then, using the obtained three-dimensional labeled molecular graph, the values of structural 3D-descriptors are calculated[2].

At the first stage, it is necessary to construct a molecular surface using this graph. For each atom, the van der Waals radius is calculated, and then a ball of the given radius is built around each atom. The union of the constructed balls is the basis of the molecular surface. To obtain the surface itself, a ball of a certain radius is rolled over this union (usually taken equal to the radius of the hydrogen atom), i.e. "smoothing" is performed. You can use the Discovery Studio Visualizer package or the PyMol system to automatically carry out this process.

At the second stage, it is necessary to carry out a cartographic projection of the molecular surface onto a plane, in order to then work with the projection as a flat image. The image is now prompted to search for special points. There are many ways to do this. Algorithms that are directly designed for finding special points, such as FAST, ORB, SURF, etc., can be used. You can search for contours in the image using edge detectors, convolving the image with various filters (Gauss, Laplace, Sobel, Pruitt, etc.) and take special points on the breaks of the detected contours. You can apply segmentation algorithms to the image (threshold segmentation, methods of building up areas, segmentation using cluster analysis methods, etc.) and take either the joints of the segments or the centers of gravity of the segments as singular points. Then the neighborhoods (5×5, or 7×7, or 9×9 pixels) of the singular points are vectorized, on the basis of which clustering is carried out. As a result, each special point receives a symbolic label of the cluster it fell into, or it is specially marked as "cluster dust". Thus, the image turns into a labeled planar graph, which then needs to be transferred back to the molecular surface, the result will be the vertices of the new labeled spatial graph.

3.2 Construction of the "molecule-feature" matrix

Suppose we are working with a sample of n chemical compounds and k is the complexity of the fragments that will be listed in molecular graphs. Scheme for constructing the "molecule-feature" matrix can be broken down into the following steps:

1) For each of the n molecular graphs, atoms are labeled.
2) For each of the n molecular graphs, a complete list of all its k-fragments is constructed.
3) The resulting lists for each of the n molecular graphs are combined into one list, from which all repetitions are removed. This is how the alphabet of k-fragments is formed for the given sample.
4) Let the cardinality of the alphabet be m. Molecule-feature matrix will have the size $n \times m$, and on (i,j)-th place will be the number of repetitions of the j-th k-fragment in the i-th molecular graph from the training set.

The difficulty in solving this problem is that in practice the number of columns is much greater than the number of rows ($m >> n$). This issue is solved by preliminary filtering of features using special algorithms. One such algorithm is described in detail in ¡3.4¿.

3.3 Clustering the training sample

Clustering was carried out in the space "Wiener index – Randich index" by the k-means algorithm. This is the most popular clustering method. He was preferred because of the high speed of work. The operation of the algorithm is such that it seeks to minimize the total square deviation of cluster points from the centers of these clusters, i.e.

$$\sum_{i=1}^{k} \sum_{p \in cluster_i} ||p - c_i||^2 \to min,$$

where k is the number of clusters, c_i is the center of the i-th cluster.

The steps of the algorithm are as follows:

1) Fix the parameter k
2) Initialize k centers (e.g. randomly)
3) Distribute points into clusters: assign each point to a cluster with the center closest to this point
4) Move the centers so that they really are the centers of the resulting clusters
5) If at least one center has changed in step 4, go to step 3

A more detailed description, features and comparison with other cluster analysis algorithms are given in [4].

3.4 Evolutionary algorithm of GMDH

As mentioned earlier, one of the problems that arise when working with the "molecule-feature" matrix, is the problem of the "information explosion": the matrix is very wide. A possible solution is to pre-select the "most significant" signs, which allows you to make the group method of data handling (GMDH). This method was described in [10], [11], here we consider its modification for the "structure-property" problem, proposed in [1].

3.4.1 Classic GMDH scheme

Let X be a "molecule-feature" matrix size $n \times m$, pre-centered and normalized. Centering means calculating the average over the entire matrix and then subtracting the resulting value from all elements. Normalization means dividing each column by its norm, for example, Euclidean.

It is necessary to construct a linear function

$$a(x^1, x^2, ..., x^k) = w_0 + w_1 x^1 + w_2 x^2 + \ldots + w_k x^k$$

from k feature columns $x^1, x^2, \ldots, x^k$, which are selected among the columns of the matrix X. The construction of the function $a(x^{j_1}, x^{j_2}, \ldots, x^{j_k})$ is carried out in steps called selections. At each selection, no more than Q columns are selected according to the principle of maximizing some quality criterion. The columns x^j of the matrix X can strongly correlate with each other, which can have a bad effect on the result of the algorithm – lead to uninformativeness, the selection of similar columns. To avoid this, a pairwise correlation threshold is introduced – C.

(1) 1-st selection
We iterate over all kinds of regression equations with two variables in the class of linear functions:

$$buf^1 = a(x^i, x^j) = w_0 + w_1\, x^i + w_2\, x^j, \quad i, j = 1, \ldots, m,$$

where x^i and x^j are columns of the matrix X. The total number of such equations is $C_m^2 = m(m-1)/2$. According to the optimization criterion, Q of the best equations are selected whose pairwise correlations do not exceed C. They will take part in the next selection.

(2) 2-nd selection
The system adds the variables x^i to the equations selected in the first step using a linear function of two variables:

$$buf^2 = a(x^i, buf_j^1) = w_0 + w_1\, x^i + w_2\, buf_j^1, \quad i = 1, \ldots, m,\ j = 1, \ldots, Q,$$

where x^i are the columns of the matrix X, buf_j^1 are the columns selected during the 1-st selection. According to the optimization criterion, the best Q equations are selected whose pairwise correlations do not exceed C. They will take part in the next selection.

$\vdots$

(k) k-th selection
The system adds the variables x^i to the equations selected at the $(k-1)$-th step using a linear function of two variables:

$$buf^k = a(x^i, buf_j^{k-1}) = w_0 + w_1\, x^i + w_2\, buf_j^{k-1}, \quad i = 1, \ldots, m,\ j = 1, \ldots, Q,$$

where x^i are the columns of the matrix X, buf_j^{k-1} are the columns selected during $(k-1)$-th selection. The total number of such equations is $m\,Q$. According to the optimization criterion, the best Q equations are selected whose pairwise correlations do not exceed C. They will take part in the next selection.
The stopping criterion is to carry out a given number of I selections.

If we now consider the list of columns x^i, included in buf^i in the course of at least one selection, we get a list of the most "significant" columns, on the basis of which we will build the equation of the final model.

As optimization criteria, we will use the R^2 criterion if the regression problem is solved, and accuracy if the classification problem is solved (see ¡2.1.3¿).

3.4.2 *Introducing a nonlinear transformation*

The use of GMDH allows not only to select significant features in the course of building models, but also makes it possible to perform functional transformations on feature columns in the course of calculations. One of the possible transformations is the introduction of nonlinearity based on the idea of fuzzy logic. We will call this transformation the feature fuzzification.

We fix $k \in \mathbb{N}$. For column x^i put $a = \min\limits_{j=1,\ldots,n} x_j^i$, $b = \max\limits_{j=1,\ldots,n} x_j^i$. On the segment $[a; b]$ we introduce a uniform grid with a step h :

$$\omega_h = \{a + th \mid t = 0, \ldots, T\}, \text{ where } a + Th = b.$$

We will consider all possible pairs of points of the form $(z_1, 0)$, $(z_2, b-a)$, where $z_1, z_2 \in \omega_h,\ z_1 < z_2$. We introduce a mapping φ that constructs from the vector x^i the vector $\varphi(x^i)$ with components

$$\varphi(x^i)_j = \begin{cases} 0, & x_j^i \in [a;\, z_1), \\ \dfrac{b-a}{z_2 - z_1}\,(x_j^i - z_1), & x_j^i \in [z_1;\, z_2], \\ b-a, & x_j^i \in (z_2;\, b]. \end{cases}$$

Figure 3 illustrates this approach. The abscissa shows the present values of the vector components, and the ordinate shows the changed values after applying the fuzzification. The bisector would correspond to the absence of transformations over the vector x^i.

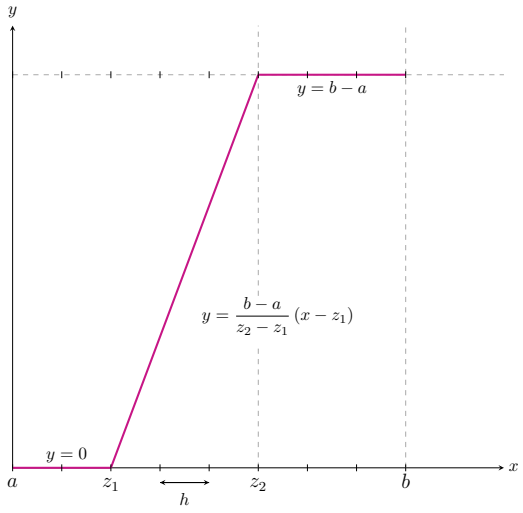


Fig. 1: Visualization of the φ function

The difference between this approach and the classical GMDH, described in ¡3.4.1¿, is that at each k-th selection, the x^i column participates in the equation after applying the φ fuzzification to it:

$$buf^k = w_0 + w_1\, \varphi(x^i) + w_2\, buf_j^{k-1}, \quad i = 1, \ldots, m,\ j = 1, \ldots, Q.$$

So. we provide a detailed description of each stage of solving the "structure-property" problem from in 2.3. Several methods of vectorization of molecular graphs areproposed, illustrations and examples of counting the intro-duced descriptors are given.

The method of the "molecule-feature" matrix construction is described. The GMDH algo-rithm for the QSAR modeling problem is considered: in the classical version and modified, with the introduction of the concept of feature fuzzification. P seudocodes a re g iven for the algorithms used

4 EXPERIMENTS

The experiments were carried out on samples with the code Np** and Npb*, on three vertices. Each sample has a threshold value of activity, that is, if the value in the vector y is less than it, then the molecule is active, otherwise— inactive. As a model, MGUA with ridge regression and regularization coefficient = 0.1 was used for more competent and stable work with outliers. Scaling or standardization was applied to the matrix of topological indexes — this is such a data preprocessing, after which each feature has an average of 0 and a variance of 1. Accordingly, after it, each column will have a normal distribution with the same mean and variance. A threshold value is subtracted from the target vector to classify the sign (+1, -1)

For the samples, a cluster search was performed using the k-means and DBSCAN methods. Additionally, a DB-SCAN cluster search was performed on the main compo-nents of the sample (PCA) found using SVD decomposition.

For the k-means method, the hyperparameter k=3 is chosen, since for a larger k, there are too few objects in some clusters for at least some stable solution. The method itself is unstable, since different launches result in clusters of different sizes. This is probably due to random initialization at the beginning of the algorithm.

	Cluster-analysis algorythm			precisio
bzr NNdb*	k-means	CLUSTER # 0	-1	0.91
			1	0.00
			accuracy	0.91
		CLUSTER # 1	-1	0.71
			1	0.50
			accuracy	0.69
		CLUSTER # 2	-1	0.78
			1	0.71
			accuracy	0.74
	DBSCAN	CLUSTER # -1	-1	0.89
			1	0.86
			accuracy	0.88
		CLUSTER # 0	-1	0.66
			1	0.65
			accuracy	0.65
		CLUSTER # 1	-1	0.95
			1	0.00
			accuracy	0.95
		CLUSTER # 2	-1	0.62
			1	0.75
			accuracy	0.69
	PCA and DBSCAN	CLUSTER # -1	-1	0.67
			1	0.89
			accuracy	0.76
		CLUSTER # 0	-1	0.67
			1	0.62
			accuracy	0.66
		CLUSTER # 1	-1	0.81
			1	0.00
			accuracy	0.81
		CLUSTER # 2	-1	0.57
			1	1.00
			accuracy	0.75

5. CONCLUSION

In the DBSCAN method, the parameters min samples = 30 were selected, that is, the minimum cluster size is 30, and eps=5.5, which was selected in such a way that there were as few points of failure as possible (CLUSTER # -1) and as many clusters as possible (with eps=7, almost all points belonged to cluster 0, the rest to -1).

In the PCA method, 10 main components were taken, with a different value of the hyperparameter, somewhat similar, unremarkable results are obtained. The advantage of this method is its relatively high speed due to the reduction in the dimension of the matrix.

ACKNOWLEDGMENTS

The paper was published with the financial support of the Ministry of Education and Science of the Russian Federation as part of the program of the Moscow Center for Fundamental and Applied Mathematics under the agree-ment 075-15-2022-284, this research has been supported by the Interdisciplinary Scientific and Educational School of Moscow University "Brain, Cognitive Systems, Artificial Intelligence"

REFERENCES

[1] M. Kumskov, A. Shestov, K. Bellonin, "Representation of the spatial structure of a molecular surface taking into account various local physicochemical properties in the "structure-property" problem", in *Models and architectures of neural networks in the "structure-property" problem*, Moscow, Russia: MAX Press, 2020, pp. 27–34. 2

[2] M. Kumskov, E. Rybakova, N. Ermolaev, "Analysis and preprocessing of molecular graphs training samples for the formation of RBF neural network architectures" in *Collection "Prediction of properties of molecular graphs based on RBF neural networks"*, Moscow, Russia: Publishing solutions, 2020, pp. 26-37. 3

[3] M. Kumskov, E.Khritova, *Clusters in the "structure-property" problem*, Ekaterinburg, Russia: Ridero, 2019, pp. 27–34 4

[4] M. Kumskov, D. Mityushev, "Solving the Inverse Problem of Molecular Graph Recognition by Constructing Minimal Paths in a Labeled Graph Space" in *Pattern Recognition and Image Analysis*, vol.6, n.2, 1996, pp. 277–278. 5

[5] H. Wiener, "Structural determination of paraffin boiling points" in *Journal of the American Chemical Society*, vol.69(1), 1947, pp. 17-—20. 6

[6] M. Randić, "Characterization of molecular branching" in *Journal of the American Chemical Society*, vol.97(23), 1975, pp. 6609-–6615. 7

[7] A. Balaban, "Highly discriminating distance-based topological index" in *Chemical Physics Letters*, vol.89, 1982, pp. 399–404. 8

[8] R. King, *Chemical applications of topology and graph theory*, Moscow, Russia: Mir, 1987 9

[9] A. Ivakhnenko, Yu. Zaichenko, V. Dimitrov *Decision-making based on self-organization"*, Moscow, Russia: Soviet radio, 1976 10

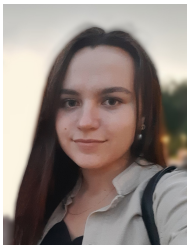
[10] A. Ivakhnenko, Yu. Yurachkovsky, *Modeling of complex systems based on experimental data*, Moscow, Russia: Radio and communications, 1987 11

[11] O. Zefirova, "Creating of QSAR" in *Chemistry*, vol.11, 2008

Mikhail I. Kumskov (Member, IEEE) is Doctor of Physical and Mathematical Sciences, Profes-sor of the Faculty of Mechanics and Mathematics, Lomonosov Moscow State University. Graduated from the Faculty of Computational Mathematics and Cybernetics of Lomonosov MSU in 1978, defended his candidate's dissertation in 1981, and his doctoral dissertation in 1997. Author of over 90 publications.

Research interests - machine learning, structural object recognition, image analysis, predicting the properties of molecular graphs, QSAR, QSPR.

Varvara O. Vasilyeva is 6th year student of the Department of Computational Mathematics, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University.
Research interests - machine learning, structural object recognition, predicting the properties of chemical compounds, QSAR, QSPR.



Ekaterina O. Rybakova is 6th year student of the Department of Computational Mathematics, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University.
Research interests - machine learning, predicting the properties of chemical compounds, QSAR, QSPR, Fuzzy Systems