# Evolutionary Method of Ranking and Classification of Biological Objects

I.A. Tsygankova

St. Petersburg Institute for
Informatics and Automation RAS
St. Petersburg, Russia
e-mail: itsygankova88@yandex.ru

## ABSTRACT

Method of classification of biological objects is suggested in this article. The method is based on the evolutionary approach to the solution of the extremal problem of multivariable function. Method is aimed at processing multidimensional data arrays which features are high dimensionality and small sample size of objects. The method is based on the ranking of the objects in multidimensional space relative to some base element. Search of this base element is carried out by a modified genetic algorithm. The method implements dual ranking of objects relative to the base element: the ordering of objects into classes, and the ordering of objects in ascending distance from the base element within classes. Belonging of the new object to one of the classes is determined by its rank in an ordered series of objects of learning sample. The proposed classification method does not require reducing the dimensionality of the feature space. This eliminates the loss of important information and allows considering internal communications in these information arrays.

## Keywords

Data processing, genomic information, classification, ranking, base element.

## 1. INTRODUCTION

One of the important tasks to be solved in the process of development of new biologically active substances and methods of diagnostics is the task of classification of objects by genome information. The characteristic features of genome data are high dimensionality of the feature space and small sample size of objects.

At present there are numerous known methods of classification [1-7], with efficiency depending strongly on the specificity of the subject area in which this problem was formulated and the characteristics of the original information. The analysis of the existing methods showed that their use for classification of objects by genome information requires prior reduction of feature space dimension - or the factual reduction of dimension takes place already in the process of formation of the classifying rule. The specificity of genome information makes the reduction of feature space dimension principally unacceptable, as this might lead to loss of important information about the unknown mutual bonds of genes, which is extremely important in developing new methods of diagnostics and medications. Therefore the development of classification methods oriented at processing of genome information and not requiring reduction of feature space dimension is still relevant.

## 2. FORMULATION OF THE PROBLEM

Let's assume that there is a finite set of objects

$$= \{s_1, s_2, \dots, s_i, \dots, s_n, s_{n+1}, s_{n+2}, \dots, s_{n+m}\}.$$

This set is divided into two disjoint subsets (classes) $K_0$ and $K_1$:

$$K_0 = \{s_1, s_2, \dots, s_n\},$$

$$K_1 = \{s_{n+1}, s_{n+2}, \dots, s_{n+m}\},$$

where $n$ is the number of class $K_0$ objects; $m$ is the number of class $K_1$ objects; $n + m = N$ is the total number of objects.

Each object is described by the set of parameters (objects – points of the $p$ – dimensional space)

$$s_i = \{X_i, y_i\},\ i = 1, \dots, n + m.$$

$$X_i = (x_1, x_2, \dots, x_j, \dots, x_p),\ j = 1, \dots, p,$$

$$y_i = \begin{cases} 0, & if\ s_i \in K_0 \\ 1, & if\ s_i \in K_1 \end{cases},$$

where $X_i$ is the vector of input parameters of $s_i$; $y_i$ is the classificatory (objective) parameter, which determines the class membership of the object $s_i$.

Parameters of $X_i$ can take values from any set of valid values of real numbers. Values of certain parameters $x_j$ for some objects may not be defined, i.e. there are missing values in the data table. Dimension of feature space $R^p$ significantly exceeds the sample size, i. e. $p \gg N$.

Is it necessary to suggest a method which enables to classify the object $q$, predetermined by vector $X_q = (x_1, x_2, \dots, x_p)$ with acceptable accuracy and without reducing the feature dimension?

## 3. CLASSIFICATION METHOD BASED ON RANKING

The considered problem of classification is poorly formalized because all information about objects is represented only by the set of input and output parameters, which cannot be definitely called full, consistent and unbiased. In such a situation the most effective are the methods [8-11] which are based on evolutionary approach to solution of extreme problems of function of many variables which in contradistinction to traditional methods of optimal

solution search are focused on obtainment of admissible solution, which is better than that obtained before or preset as initial.

To solve the above problem, we propose a method [12] based on assumption that there is some base element in the multidimensional feature space relative to which a ranged sequence of objects is formed, separating the learning sample into two classes. The method realizes a dual ranging of objects relative to the base element: ordering of objects by class and ordering of objects by increase of the distance from the base element within classes. Then the class membership of object $q$ is determined by the rank $t_q$ of classified object in ordered series of learning sample objects.

Suppose the base element $\mathbf{X}^* = (x_1^*, x_2^*, \dots, x_p^*)$ is a point in $p$–dimensional space. Let us rank the objects of learning sample relative to the base element in increasing the distance $\rho_i$ and each object is assigned rank $t_i, (i = 1,2, \dots, n, n+1, n+2, \dots, n+m)$ . We take the order of the classes in the object sequence: $K_0 \prec K_1$. Then, as the boundary between the classes will be considered an object whose rank $t_n$ is equal to the number of class $K_0$ objects.
The decision that the object $q$ belongs to one of the classes will take from the condition

$$q \in \begin{cases} K_0, & if\ t_q \le t_n \\ K_1, & if\ t_q > t_n \end{cases}.$$

Degree of membership of object class will be determined by the formula

$$\eta = \begin{cases} 1 - \frac{t_q}{t_n}, & if\ q \in K_0 \\ \frac{t_q - t_n}{t_{n+m} - t_n}, & if\ q \in K_1 \end{cases}.$$

The search of base element is made using the evolutionary approach realized through the use of modified genetic algorithm, where the actual objects of learning sample are considered as an initial population. The population size is fixed and equal to the size of learning sample. The fitness of each individual (object) is evaluated using the fitness function

$$F = \max(f_0, f_1),$$

where $f_0$ and $f_1$ are error estimates for the classification of class $K_0$ objects and class $K_1$ objects, respectively. To calculate $f_0$ and $f_1$ are counted incorrectly classified objects in the intervals $[t_1, t_n]$ and $[t_{n+1}, t_{n+m}]$ taking into account the rank of these objects in an ordered sequence.

The lower the value of function $F$ is, the higher the "quality" of the individual. As a result of artificial evolution, including selection, crossing and mutation of individuals, quality of solutions in the population improves gradually. Work of the genetic algorithm is completed when: function $F$ reaches the expected value; or execution of preset number of iterations (generations) does not improve already reached value $F$; or upon expiration of preset period of time allotted for the problem solution. Premature stop of work of the genetic algorithm may occur in the case of population degeneracy.

## 4. NUMERICAL EXPERIMENT

To evaluate the effectiveness of the proposed method, a numerical experiment was carried out with the use of real experimental data, regarding the Atlantic salmon's genes expression level. A sample of 100 fishes was used as a learning sample. Each fish was described by a vector of numerical parameters which were the results of instrumental measurement of the expression level for various genes in the Atlantic salmon. The parameter vector dimension was 967. The learning sample was divided into two classes. One class included 50 fishes infected with salmon anaemia virus, and the other class - 50 fishes not contaminated with this virus.

Numerical experiment was conducted with the following key parameters of genetic algorithm:

- population size $N = 100$
- probability of crossing $P_c = 0{,}99$
- probability of mutation $P_m = 0{,}001$
- number of elitist chromosomes $N_{ch_e} = 1$
- number of the least fitted chromosomes which underwent replacement in population with the most fitted chromosomes $N_{ch_l} = 1$

Some results of the numerical experiment are shown in Fig.1 and Fig.2.

Distribution histogram of objects depending on the distance from the base element is presented in Fig.1. Objects of class $K_1$ are shown above the x-axis, and objects of class $K_0$ – under this axis.

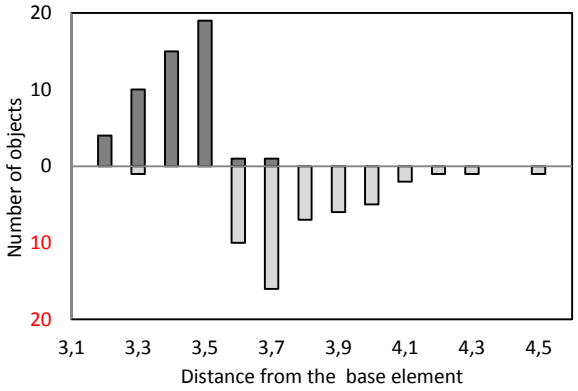


Fig.1 Distribution histogram of objects of two classes according to distance from the base element

Rank sequence of objects of the learning sample obtained in the learning process is presented in Fig.2. Objects of class $K_1$ are shown above the x-axis and the objects of class $K_0$ are shown under the x-axis.

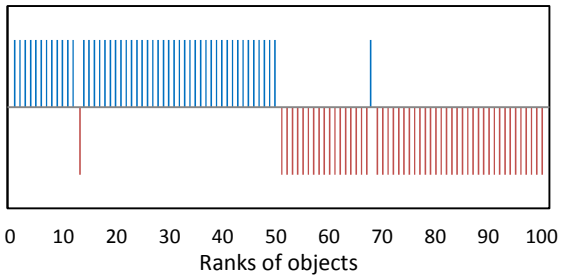


Fig.2 Rank sequence of the learning sample objects relative to the base element

As it can be seen from Fig.1 and Fig.2 the sequence of objects ranked according to increase of distance relative to

the base element is divided into two classes with minimum error.

The estimation of the method effectiveness was performed on the control sample of 10 objects not included in the learning sample. The control sample included five objects of each class. All control objects were classified correctly. Thus, it is believed that the proposed classification method based on the ranking of objects relative to the base element, is effective in solution of problems of classification of objects with high dimension of feature space without prior reduction of dimension.

## 5. CONCLUSION

The work proposes the evolutionary classification method based on the ranking of objects in multidimensional space relative to the base element, which is searched using modified genetic algorithm. The method is focused on the processing of multi-dimensional arrays of information, features of which are the high dimension of the feature space and the small size of object sample. The proposed classification method enables non-conduction of preliminary reduction of the feature space dimension, which, in its turn, eliminates the loss of important information and takes into account the interconnections in information arrays under consideration.

## REFERENCES


[1] S.A. Aivazyan, V.M. Buchstaber, I.S.Yenyukov, L.D. Meshalkin. Applied Statistics: Classification and Reduction of Dimensionality. – Moskva: Finansy i statistika, 1989 [in Russian]

[2] M.A. Ayzerman, Je.M. Bravermann, L.I. Rozonoyer. Method of potential functions in the theory of machine learning. – Moskva: Nauka, 1970 [in Russian]

[3] V.N. Vapnik, A.Y. Chervonenkis. Recognition theory (statistical learning problems). – Moskva, Nauka, 1974 [in Russian]

[4] Y.I. Zhuravlev. Favorites scientific works. – Moskva: Magistr, 1998 [in Russian]

[5] N.G. Zagoruyko. Applied methods of data analysis and knowledge. – Novosibirsk: Izd-vo In-ta matematiki, 1999 [in Russian]

[6] V.D. Mazurov. Committee method in optimization and classification. – Moskva: Nauka, 1990 [in Russian]

[7] L. A. Rastrigin, R. H. Erenshteyn. Method of collective recognition. – Moskva: Jenergoizdat, 1981 [in Russian]

[8] D. Rutkovskya, M. Pilinsky, L. Rutkovsky. Neural networks, genetic algorithms and fuzzy systems. – Moskva: Gorjachaja linija, 2008 [in Russian]

[9] A.A. Freitas Data Mining and Knowledge Discovery with Evolutionary Algorithms. – Berlin: Springer, 2002

[10] Z. Michalewicz Genetic algorithms + data structures = evolution programs. – Berlin etc.: Springer, 1996.

[11] Evolutionary optimization / Ed. by R. Sarker, M. Mohammadian, X. Yao. – Boston etc.: Kluwer acad. publ., 2002.

[12] I.A. Tsygankova, "Evolutionary method for classification of biological objects", *Information Systems and Technologies*, No. 5, pp.50-58, 2012 [in Russian]