Models and Evolutionary Methods for Objects and Systems Clustering

Maryna Antonevych, Nataliia Tmienova and Vitaliy Snytyuk

Taras Shevchenko National University of Kyiv, B. Havrylyshyna Str., 24, Kyiv, 04116, Ukraine

Abstract

As we know, clustering refers to the problems of Data Mining. The clustering problem is usually solved without a supervisor with using the features of the object or system. It is necessary to carry out clustering in various fields: agriculture, medicine, biology, etc. In this paper propose to carry out clustering using two objective functions and constructed evolutionary population methods based on genetic algorithms, evolutionary strategies and the method of deformed stars.

New elements of such methods for calculation of objective function are developed, advantages of the population approach before known classical methods of clustering are shown. A comparative analysis of the proposed clustering methods effectiveness using a known sample of iris data and a sample of randomly generated data is fulfilled. The peculiarities of each proposed method application, its advantages and disadvantages are determined.

Keywords [1]

Objects, systems, clustering, models, evolutionary methods

1. Introduction

The diversity of the modern world, the need and desire to know it are the most important reasons for the systematization, clustering and classification of objects and systems. The dynamics of the amount of information and the limitations of human perception require the creation of data filtering systems, their organization, determining the importance and so on. Obviously, these factors emphasize the importance of developing and using modern clustering methods based on new data technologies.

The globalization of the processes of the modern world, the various challenges facing humanity, make us think about its future. The famous writer John Galsworthy wrote: "If you do not think about the future, you may not have it." However, thinking about the future is not enough, we need to predict it and use intelligent decision-making methods to reduce future risks.

It is known that effective prediction is based on the results of solving the problem of identification, both structural and parametric. Today, most often, the basis of prediction methods are time series, or the dependence of unknown characteristics on known factors, one of which may be time. Prediction using time series is based on the analysis of retrospective data and the construction of autoregressive models, usually linear. Otherwise, there are multifactorial dependencies, the values of the factors of which are in a certain area, taking into account the possible values.

II International Scientific Symposium «Intelligent Solutions» IntSol-2021, September 28–30, 2021, Kyiv-Uzhhorod, Ukraine
EMAIL: marina.antonevich@gmail.com (A. 1); tmyenovox@gmail.com (A. 2); snytyuk@knu.ua (A. 3)
ORCID: 0000-0003-3640-7630 (A. 1); 0000-0003-1088-9547 (A. 2); 0000-0002-9954-8767 (A. 3)


It is important to note that the economy of the modern world is unstable, the processes in it are heterogeneous and non-stationary. The use of data from the whole possible spectrum for their identification is incorrect, as their logical-dynamic nature can lead to averaging and displacement of results. That is why the output data should be divided into clusters that correspond to homogeneous processes and allow adequate identification and high-precision prediction.

2. Clustering Problem and Analysis of Methods its Solving

The urgency of solving the clustering problem has led to the development of a large number of clustering methods. Since this problem has an optimization character, the No Free Lunch Theorem (NFL) is applicable to it [1]. The meaning of this theorem is that there is no single method that can best find a solution to any optimization problem. That is why both the development of clustering methods and the study of their effectiveness continue.

Historically, hierarchical methods were developed first. They were divided into agglomerative and divisive methods of clustering [2,3]. For the first methods, clusters were formed by merging smaller groups into larger ones, for the second, on the contrary, clusters were formed by dividing larger groups into smaller ones. The class of similar methods includes dendrograms, as well as decision trees. Typical representatives of the agglomerative paradigm are Ward's method, single linkage, pair-group method using arithmetic mean and complete linkage, and the pair-group method using centroid average.

The next large group of clustering methods includes algorithms of the Forel family of the Novosibirsk school of data analysis [4], the method of K-means and K-medians [5], EM-algorithm [6], and others. Their feature is the iterative nature of the search and refinement of cluster centers.

In recent decades, clustering methods based on new paradigms, usually inspired by the structure and functioning of biological and social systems have been developed. In this case, the affiliation of objects to clusters could be ambiguous, with a certain degree of correspondence, the object could belong to a particular cluster. This is the method of C-means. The method of C-means is a method of fuzzy clustering, which is based on the idea of calculating subjective conclusions [7]. Much of the clustering methods were developed based on the use of neural networks. In the foreground here is the Kohonen neural network with the implementation of the principles of WTA (winner take all) and unambiguous assignment of the object to a particular cluster and SoftMax with the assignment of the object to several clusters with certain probabilities [8].

Another approach to developing clustering methods is to use technologies that imitate elements of natural evolution. Such technologies are genetic algorithms, evolutionary strategies, swarm algorithms, imitation of metal annealing, and others. Like the above methods, they belong to the Soft Computing paradigm [9,10]. Evolutionary methods have a certain feature in comparison with other methods: under certain conditions with a certain probability, they guarantee finding the global optimum of the objective function. That is why the development and study of the effectiveness of such methods is an urgent task.

Note that in the article we did not aim to study methods based on other hypotheses about the shape of clusters, their functional nature, etc. [11-13]. The developed methods are designed to solve the classical problem of clustering, where objects are points in space of a certain dimension, because the solutions of such a problem are most often used by decision makers in solving practical problems of today [14].

When solving the clustering problem, the proposed metric which determines the distance between object plays an important role, s. It is necessary to consider what kind of similarity is being considered. In addition to distance, it can be an existing analytical relationship between objects or a form of their condensation. The problems of studying possible metrics were studied in [15-18].

3. Clustering Method on the Base of Evolutionary Strategy

Suppose that objects or systems of research are characterized by a certain range of properties. Then

each object can be represented as a point in the n-dimensional space of the following properties, that is:

$$Xi = (x_{i1}, x_{i2}, ..., x_{in}), i = \overline{1,m},$$

where n is the number of properties, m is the number of objects.

Data about objects are in the matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}.$$

It is necessary, based on the data of the matrix X, to divide the set of studied objects into clusters. As a rule, the number of clusters is known a priori, but there are problems where the optimal number of clusters also needs to be determined.

Consider the first problem and assume that the number of clusters is K, and $K < m$. At the first stage we normalize the elements of the matrix X and get the matrix $X' = \{x'_{ij}\}, i = \overline{1,m}, j = \overline{1,n}$, its elements are

$$x'_{ij} = \frac{x_{ij} - \min\limits_i x_{ij}}{\max\limits_i x_{ij} - \min\limits_i x_{ij}}, x_{ij} \in [0,1].$$

As a result of normalization, the points from the matrix X will lie in a single hypercube $\Omega = [0,1]^n$. Remind that the distance between two points in n-dimensional space is equal to $d_{ij} = d(X_i, X_j) = (\sum\limits_{k=1}^{n} (x_{ik} - x_{jk})^2)^{1/2}, i, j = \overline{1,m}$.

We write the objective function for the problem of clustering the objects described above:

$$F(C) = \sum_{i=1}^{K} \sum_{j} d(C_i, X_j), \quad (1)$$

where C is the set of point-centers of the clusters, C_i is the point-center of the i-th cluster, j is the index that indicates the affiliation of X_j to the i-th cluster. There are problems in which the essential requirement is that the centers of the clusters must be as far apart as possible. Therefore, in this case, function (1) is modified to the form:

$$F(C) = (\sum_{i=1}^{K} \sum_{j} d(C_i, X_j)) / (\sum_{i=1}^{K} \sum_{j>i}^{K} d(C_i, C_j)) \quad (2)$$

or

$$F(C) = (\sum_{i=1}^{K} \sum_{j} d(C_i, X_j)) - (\sum_{i=1}^{K} \sum_{j>i}^{K} d(C_i, C_j)). \quad (3)$$

Let us solve the problem of finding cluster centers

$$C^* = \arg\min_{c \in [0,1]^n} F(c). \quad (4)$$

To do this, we use an evolutionary strategy $(\mu + \lambda)$ with an elite selection of elements of the new population, where μ is the number of potential parental solutions, λ is the number of potential descendant solutions. The authors of the evolutionary strategy consider that the necessary condition to ensure the diversity of the population of solutions is $\lambda \geq 7\mu$.

Step 1. Initialization of evolutionary strategy parameters. $t = 0$.

Step 2. Generation of a population of potential solutions of problem (1-4), $C^t = \{C_{ij}^{tl}\}, i = \overline{1,k}, j = \overline{1,n}, l = \overline{1,\mu}$. Thus, the number of potential solutions is equal μ (in most cases 10-30).

Step 3. Build μ matrices of distances from the centers of clusters to objects:

$$A_e^t = \begin{pmatrix} d_{11}^{tl} & d_{12}^{tl} & \dots & d_{1m}^{tl} \\ d_{21}^{tl} & d_{22}^{tl} & \dots & d_{2m}^{tl} \\ \dots & \dots & \dots & \dots \\ d_{k1}^{tl} & d_{k2}^{tl} & \dots & d_{km}^{tl} \end{pmatrix},$$

where $l = \overline{1,\mu}, d_{ij}^{tl} = d(C_i^{tl}, X_j)$, C_i^{tl} is the center of the i-th cluster in the l-th potential solution.

Step 4. Find the value of the analog of the objective function (1) for each of the μ potential solutions,

$$F(A_e^t) = \sum_{j=1}^{m} \min_i d_{ij}^{tl}, l = \overline{1,\mu}. \quad (5)$$

Thus, the value of the objective function is matched to each potential solution.

Step 5. Generate a population of potential descendant solutions as follows. For every l-th potential solution:

Perform λ times.

For each i-th center of the cluster:

Generate a random number $r \in \{1,2,...,n\}$ and modify the r-th coordinate:

$C_{ir}^{tl} = C_{ir}^{tl} + \xi(N(0, \frac{2}{9\mu}))$, where $\xi(N(0, \frac{2}{9\mu}))$ is a normally distributed random number with zero mathematical expectation and variance $\frac{2}{9\mu}$.

If $C_{ir}^{tl} < 0$, then $C_{ir}^{tl} = 1 + C_{ir}^{tl}$.

If $C_{ir}^{tl} > 1$, then $C_{ir}^{tl} = C_{ir}^{tl} - 1$.

This means that the descendant solution will differ from the parent solution by one coordinate. Write the descendant solution in the intermediate population C^P.

Step 6. Calculate function (4) for all descendant solutions from C^P.

Step 7. Combine $C^t \cup C^p$ and arrange potential solutions in descending order of the objective function (4).

Step 8. $t = t + 1$. Build a population C^t from the best μ population solutions $C^{t-1} \cup C^p$.

Step 9. If the stop condition is not carry out, go to step 3.

Step 10. Output of the result (the centers of clusters).

As is known, evolutionary strategy is a method of global optimization, implemented, as a rule, by a population algorithm. Unlike other evolutionary modeling algorithms, the evolutionary strategy does not use binary coding or other types of coding. At the beginning of the algorithm, a population of potential solutions (cluster centers) is generated.

If there are three clusters, then we generate, for example, 20 triplets of cluster centers. For each such triple, we find the points that belong to the corresponding cluster and calculate the values of the objective function. Then we store these triplets and the values of the function in the intermediate population. Next, for each triple, we generate triplets-offsprings using the normal distribution on the principle that the point of the initial triangle is subject to modification, and the modified points form new triplets.

According to the authors of the evolutionary strategy [10], to ensure population diversity and the optimal number of computational procedures, the number of solutions-offsprings must refer to the number of solutions-parents at least 7 to 1.

Thus, for example, for 20 initial triplets should be at least than 140 modified triplets-offsprings. For these 140 triplets, the values of the objective function are also calculated and the corresponding data are also entered into the intermediate population. Thus, in the intermediate population there are 160 triplets and values of the objective function. Arrange these 160 triplets according to the values of the objective function. The first 20 triplets will be written in another intermediate population and will be considered as 60 separate points.

Next, we randomly form 20 triplets and the computational process is iteratively repeated until the stop criterion is met. Note that in the proposed method we use the well-known 3-sigma rule, which

determines the value of the standard deviation of the points-offsprings from the parent points, which significantly affects the depth of the study neighborhood the parent solution and the rate of convergence of the algorithm.

4. Clustering on the Base of Genetic Algorithm

Genetic algorithm as well as evolutionary strategy under certain conditions is a method of global optimization. That is why its elements can be used to solve the problem of clustering. Consider the appropriate method. Assume that the points to be broken lie in a rectangular hyperparallelepiped of n -dimensional space. Find the rationing, all points will be in a single hypercube $\Omega = [0,1]^n$. The coordinates of the points form a matrix X (see above), their number is m. Assume that the number of clusters is K. The solution of problem (1) - (4) will be the vector of cluster centers $C^* = (C_1^*, C_2^*, ..., C_K^*)$. The search for the vector C^* is organized according to the following algorithm.

Step 1. $t = 0$ (iteration number). Initialization of method parameters n, m, K.

Step 2. Generate uniform distributed in $\Omega = [0,1]^n$ points (potential centers of clusters)

$$C = (C_1, C_2, ..., C_l),\ K < l << m,$$

$$C_j = (C_{1_j}, C_{2_j}, ..., C_{n_j}),\ j = \overline{1,l}.$$

Step 3. From the elements of the matrix C form matrix

$$P_t = \begin{pmatrix} C_{i_1}^1 & C_{i_2}^1 & \dots & C_{i_k}^1 \\ C_{i_1}^2 & C_{i_2}^2 & \dots & C_{i_k}^2 \\ \dots & \dots & \dots & \dots \\ C_{i_1}^q & C_{i_2}^q & \dots & C_{i_k}^q \end{pmatrix},$$

where q is the number of K -gons, and none of the K -gons coincides with the other, that is

$$\forall i_j\ \exists! v, w: C_{i_j}^v = C_{i_j}^w.$$

Step 4. For each K -gon find the distance from it to each point to be assigned to the cluster, that is calculate the distance

$$d(C_{i_j}^v, x_l) = d((C_{i_j}^{v_1}, C_{i_j}^{v_2}, \dots, C_{i_j}^{v_n}), (x_{l_1}, x_{l_2} \dots, x_{l_n}))\ \forall j = \overline{1,k},\ \forall l = \overline{1,m}\ .$$

$$F_q = 0.$$

Крок 5. For each object to be clustered, we find the center of the cluster, the distance to which is the smallest:

$$C_{i_j}^{v^*} = \min_j d(C_{i_j}^v, x_l),$$

$$F_q = F_q + d(C_{i_j}^{v^*}, x_l).$$

Step 6. Arrange K -gons by the value of the function F_q in ascending order. The vertices of the best K -gons are entered in the intermediate population G_t.

Step 7. Next, we perform the usual crossover operations on them by panmixing, inbreeding or outbreeding.

Step 8. Perform a mutation operation.

Step 9. If the stop condition is not met, then $t = t + 1$ and go to step 3. Otherwise, display the centers of the clusters and calculate to which cluster the objects belong.

Stop criteria can be one of the following conditions:

1. Achieving a certain predetermined number of iterations.
2. The increment of the objective function on adjacent iterations does not exceed a small predetermined number.

When applying the elements of the genetic algorithm, it is necessary to take into account the fact that its convergence to the global optimum is not guaranteed, not least because, in the general case, the condition of continuity in the transition from integers to binary numbers is not met. In particular, for example, the numbers 15 and 16 differ by one, and their binary counterparts 01111 and 10000 differ as much as possible. In order to ensure the continuity of calculations, it is proposed to use the Gray code. Its advantage is the ease of obtaining from the usual binary representation and ensuring the continuity of the computational process.

5. Clustering on the Base of Method of Deformed Stars. Type 1

Another evolutionary method that can be used to solve the clustering problem is the method of deformed stars, which was previously proposed by the author in [19, 20]. In this paper, the simplest case was considered when the solution space was one or two-dimensional, and the stars were in the form of a segment or a triangle. Further, the method was improved [19]; in the two-dimensional case, a generalization was obtained for quadrangles and pentagons. And, finally, in [20], a universal method of deformed stars was obtained for solving the global optimization problem in the n-dimensional case.

The main ideas of the method are that the analysis of the set of potential solutions provides more information about finding the global optimum of a certain functional dependence, and various deformations of polygons that form potential solutions allow exploring the area of probable global optimum. Similar ideas can be used to solve the clustering problem, because it has an objective function, the minimum of which must be found, and objects that are points in n-dimensional space.

The illustration of the method of deformed stars for the clustering problem is given in Fig.1 and Fig.2. The first figure shows two-point stars that can rotate, stretch, or compress in two-dimensional space, thus approaching the real centers of the two clusters. If it is assumed that there will be three clusters, then the stars will be triangles. The transformations over them will be shown below.

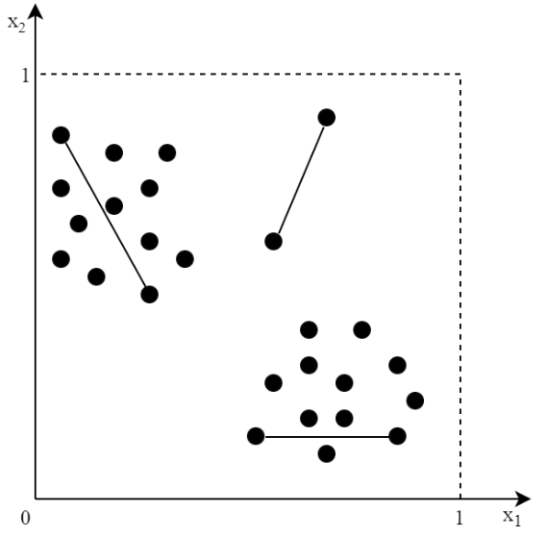

Figure 1: Two-point stars

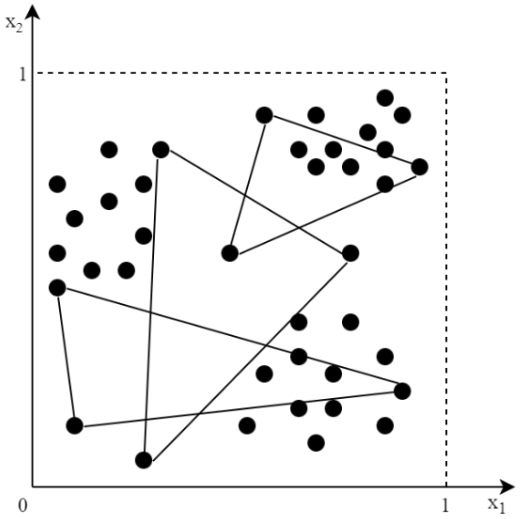

Figure 2: Three-point stars

Remind that the objects to be clustered are the points in the hypercube $\Omega = [0,1]^n$ in the n-dimensional space. Then the steps of the method will be as follows:

Step 1. Determine the size of the sample population of potential solutions (cluster centers) m. Usually, $m \in \{20, 21, ...50\}$. $t = 0$.

Step 2. Generate a population of potential solutions $P_t = \{\bar{x}^1, \bar{x}^2 ..., \bar{x}^m\}$, $\bar{X}^i = \{x^{i1}, x^{i2}, ..., x^{in}\}, i = \overline{1,m}$.

Step 3. Randomly form k triangles whose vertices are potential solutions. Note that one vertex can belong to several triangles, $k \leq m$.

Step 4. For each j-triangle:

Step 4.1. Find the value of the objective function F_j.

Step 4.2. For each l-th vertex ($l=\overline{1,3}$) we rotate the l-th vertex, form new points:

$x_\theta^{jl} = x_\theta^l \cos\alpha^l - x_\eta^l \sin\alpha^l$,

$x_\eta^{jl} = x_\theta^l \sin\alpha^l + x_\eta^l \cos\alpha^l$,

$x_\nu^{jl} = x_\nu^l, \nu \neq \theta, \eta$

Step 4.3. For a triangle composed of new points, we find the value of the objective function G_j.

Step 4.4. For each l-th vertex of the initial triangle we play a uniformly distributed number $\xi \in (0,1)$. Also, let us play a number $\theta \in \{1,2,...,n\}$ and a random number $\eta \in \{-1;1\}$. The new vertex will be as follows:

$\overline{x}^l = (x^{l_1}, x^{l_2}, ..., x^{l_\theta} + \eta \cdot \xi, x^{l_{\theta+1}}, ..., x^{l_n})$.

Step 4.5. For a triangle formed from new vertices, we find the value of the objective function H_j.

Step 5. $t = t+1$. The points of the triangle to which corresponds $\max\{F_j, G_j, H_j\}$ are entered in the population P_t. If $j \neq n$, then $j = j+1$, and go to step 4.

Step 6. If the stop condition is not carry out, go to step 3.

Step 7. The end of the algorithm.

The peculiarity of this algorithm is that it is not necessary to look for the centers of polygons in n-dimensional spaces, in contrast to the classical algorithm MODS [16]. For simplification, the rotations of the points are made relative to the coordinates center in a randomly selected plane. The operations of stretching and compressing the vertices of the polygon to simplify the algorithm are also performed on one randomly determined spatial coordinate. The proposed algorithm is a global optimization algorithm, because in this case there is a probabilistic convergence with the number of iterations, which goes to infinity. A necessary condition for such convergence is the elite selection of elements of the population of potential solutions-offsprings.

6. Clustering on the Base of Method of Deformed Stars. Type 2

The method of deformed stars can be a basic algorithm for solving a wide range of optimization problems, one of which is clustering. In addition, the method of deformed stars is a parametric method that allows wide variational possibilities for creating new elements of the method and their optimization. In particular, in the proposed clustering method, triangles whose vertices are the centers of the clusters are randomly generated. For the n-dimensional case, n-triangles are considered. The operation of mutation is to create descendant triangles, the vertices of which are the midpoints of the segments that connect the nearest points of the three triangles. These points will form an intermediate population. The population of descendants is formed from the points of the initial population and the points of the intermediate population, but those that correspond to the triangles with the best values of the objective function.

Let us all preconditions of the method application are fulfilled. Steps 1, 2, 3 remain the same.

Step 4. For each triangle we find the values of the objective function: $F_1, F_2, ..., F_k$.

Step 5. Perform $k-1$ times:

Step 5.1. According to the principle of tournament selection (in proportion to the values $F_1, F_2, ..., F_k$) to determine randomly $i, j \in \{1,2,...,k\}, i \neq j$.

Step 5.2. Mark the vertices of j-th triangle A_j, B_j, C_j.

Step 5.3. Find the vertex closest to A_i, that is, the vertex that corresponds $\min\{d(A_i, A_j), d(A_i, B_j), d(A_i, C_j)\}$. Let us mark it as $D_j = A_j \vee B_j \vee C_j$.

Step 5.4. Connect by the segments points A_i and D_j, and find its middle:

$A_i = (\frac{a^{i_1} + d^{j_1}}{2}, \frac{a^{i_2} + d^{j_2}}{2}, ..., \frac{a^{i_n} + d^{j_n}}{2})$, where $A_i = (a^{i_1}, a^{i_2}, ..., a^{i_n})$ and $D_j = (d^{j_1}, d^{j_2}, ..., d^{j_n})$.

Step 5.5. Find the vertex closest to B_i, that is, the vertex that corresponds $\min\{d(B_i, A_j), d(B_i, B_j), d(B_i, C_j)\}$, provided that this vertex is not a vertex D_j. Let it be $E_j = A_j \vee B_j \vee C_j$, $E_j \neq D_j$. Similarly, we find the middle of the segment that connects the points B_i and E_j, let it be a point B_l.

Step 5.6. Two points left: C_i and G_j. The middle of the corresponding segment will be C_l.

Step 5.7. Points A_l, B_l, C_l are recorded in the population of the new generation.

Step 6. Perform the generation of random three points (A_k, B_k, C_k) in $\Omega = [0,1]^n$. Mutation.

Step 7. $t = t + 1$. $P_t = (A_1, B_1, C_1, ..., A_k, B_k, C_k)$.

Step 8. If the stop condition is not carry out, go to step 3.

Step 9. The end of the algorithm.

The difference between the second method of clustering and the previous methods is that only one modification of the method of deformed stars is used. Such a modification is a special type of mutation operation. Mutation allows to simultaneously avoid hitting the local optimum and more deeply explore the environment of finding a potential optimum. This algorithm is easier to implement, its execution time is relatively short.

7. Analysis of experimental results

The proposed methods belong to the class of clustering methods based on the evolutionary paradigm. Its main idea is to gradual approximation the desired solution based on the development and movement of the population of potential solutions. Population development under certain conditions provides a variety of search and depth of research in a particular area. This search is slower than in other methods, but allows you to work with different, not necessarily differentiable objective functions, which will allow you to generalize them later to determine clusters of objects based on metrics other than Euclidean.

We will conduct experiments and test the effectiveness of the proposed methods. To do this, consider a well-known sample of data on irises, which contains fields such as Petal Length and Widht, Sepal Length and Width, Classes Types. To illustrate the operation of clustering methods without general restrictions, we use only the data of the fields Petal Length and Widht. Another sample was randomly generated based on a uniform distribution. The number of elements in both samples is 150 points. It is assumed that these data represent three clusters.

The experiments were performed using the following clustering methods: K-means, Forel, CGA (clustering with use of the genetic algorithms elements), CES (clustering with use of the evolutionary strategy elements), and CMODS (clustering with use of the method of deformed stars elements). The algorithm stop criterion is a fixed number of iterations. The results of the calculations are recorded in Table 1 and Table 2. Data of Table 1 correspond to the objective function (1), Data of Table 2 correspond to the objective function (3).

Table 1

Test results for the objective function of type (1)

	Random points			Irises	
	300 iter	1000 iter	2000 iter	1000 iter	2000 iter
K-means	34.77	32.83	34.2	29.44	29.43
Forel	46.02	47.17	44.6	33.81	33.33
CGA	35.92	35.75	34.72	30.85	31.4
CES	34.83	33.34	34.24	35.66	34.23
CMODS	34.49	32.52	33.92	29.36	29.32

The first experiment concerned randomly generated points (Fig. 3). The Forel method showed consistently the worst results. The best results are shown by CMODS, and these results are on average better by 0.86% than in the method of K-means, the results of which are in the second place. The number of iterations has a negligible effect on the accuracy of clustering, except for the CGA method,

where the value of the objective function decreases monotonically with the increasing number of iterations. For the Petal Length and Widht iris clustering experiment, there is a reverse trend for CGA. The results of the CMODS method remain the best.

In another experiment (objective function (3)) it was necessary to minimize intracluster distances and maximize distances between cluster centers. For random points, as the number of iterations increases, the value of the objective function increases. The CMODS method demonstrates the best accuracy again. Methods based on other population algorithms show good results, but they are worse than the results of the K-mean method. It is worth noting that for problems with higher data dimensions, the results of evolutionary methods will show better results.

Table 2
Test results for the objective function of type (2)

	Random points		Irises	
	1000 iter	2000 iter	1000 iter	2000 iter
K-means	32.11	33.34	27.11	27.11
Forel	38.71	43.12	36.39	36.39
CGA	32.49	35.07	28.89	28.98
CES	31.26	33.64	32.36	31.15
CMODS	30.92	33.07	27.06	26.79

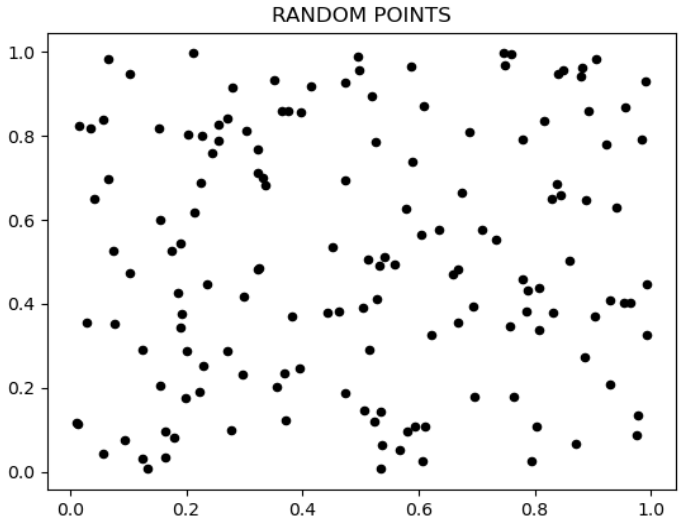

Figure 3: Random points distribution

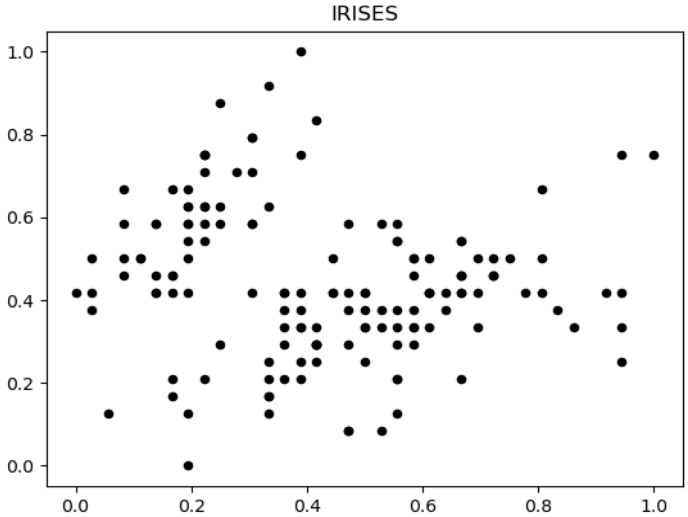

Figure 4: Irises data

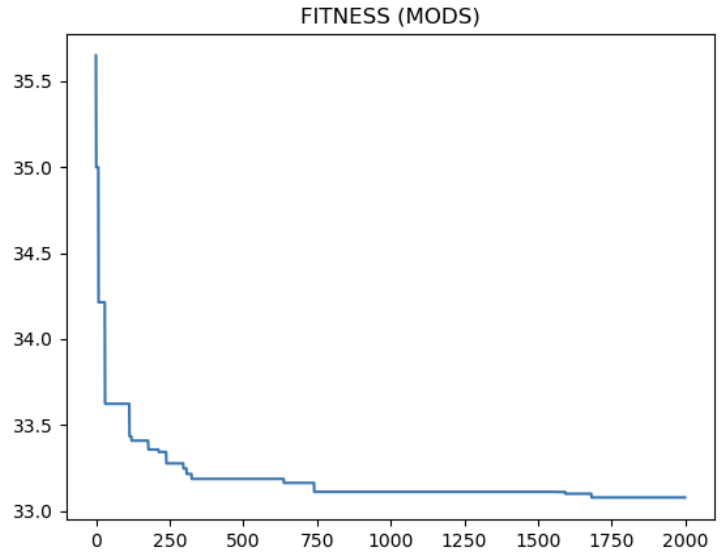

Figure 5: Fitness-function for random points

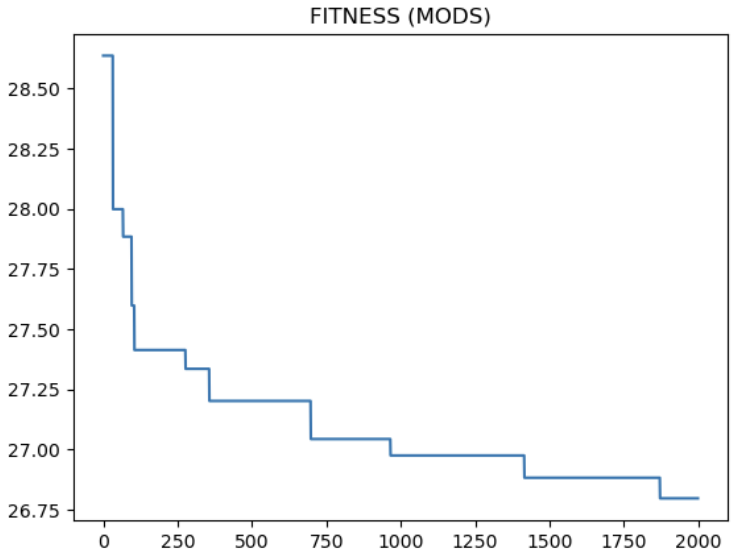

Figure 6: Fitness-function for irises

8. Conclusions and prospects

The obtained results testify to the prospects of development, research, and progress of evolutionary optimization technologies in general and their application to solve the clustering problem. Such methods are of particular value when the objective clustering function is complex, possibly with a polyextreme dependence, and the characteristic space of objects or systems is multidimensional.

It is worth paying attention to the method of deformed stars, which has demonstrated its advantages in most experiments. Such results can be explained by the fact that when searching for the optimal location of clusters, information from a certain number of points at the same time is taken into account. In methods based on elements of a genetic algorithm or evolutionary strategy, such information is insufficient. In particular, in a genetic algorithm, each potential solution contains information about two parental solutions, and in an evolutionary strategy, each solution contains information about only one parental solution. Note that in MODS the amount of information in the system is determined by the shape of the star, and the more vertices it has, the more information there is. At the same time, computation time increases. That is why the qualification of the researcher plays

a decisive role in solving the problems of clustering using the ideas of evolutionary population algorithms. The proposed methods require modification and can be used in clustering problems, when the degree of reliability of objects or systems is not only the distance between them but also, for example, functional dependence.

7. References

[1] D.H. Wolpert, W.G. Macready, "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67 (1997).

[2] M.J. Zaki, W. Meira, Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, New York, 2014.

[3] B. Everitt, Cluster analysis, Chichester, West Sussex, U.K: Wiley, 2011.

[4] N.G. Zagoruiko, Applied methods for data analysis and knowledge, Nowosibirsk,IM SB RAS, 1999 [5] E.M. Mirkes, K-means and K-medoids applet, University of Leicester, 2011.

[6] D. Kosiur, Understanding Policy-Based Networking, 2nd. ed., Wiley, New York, NY, 2001.

[7] H.-c Liu, J.-m Yih, D.-b Wu, and S.-w Liu, Fuzzy possibility c-mean clustering algorithms based on complete mahalanobis distances: in 2008 international conference on Wavelet Analysis and Pattern Recognition, IEEE, pp. 50–55, 2008.

[8] T. Kohonen, Self-Organizing Maps, Berlin – New York: Springer-Verlag, 2001.

[9] R.˜H Sheikh, M˜Raghuwanshi, and A.˜N Jaiswal. Genetic Algorithm Based Clustering: A Survey: in proceedings of 2008 first international conference on Emerging Trends in Engineering and Technology, 2(6):314–319 (2008).

[10] D. Simon, Evolutionary optimization algorithms, John Wiley @ Sons, Inc., Hobohen, New Jersey, 2013.

[11] H.-P. Kriegel, P. Kröger, J. Sander, A. Zimek, "Density-based Clustering". WIREs Data Mining and Knowledge Discovery. 1 (3) (2011) : 231–240.

[12] M. Meilă, "Comparing Clusterings by the Variation of Information". Learning Theory and Kernel Machines. Lecture Notes in Computer Science. **2777** (2003), pp. 173–187.

[13] E. Achtert, C. Bohm, H.P. Kriegel, P. Kröger, A. Zimek, "On Exploring Complex Relationships of Correlation Clusters": in procediings 19th international conference on Scientific and Statistical Database Management (SSDBM 2007). p.7.

[14] V.V. Tsyganok, S.V. Kadenko, O.V. Andriichuk, Simulation of Expert Judgements for Testing the Methods of Information Processing in Decision-Making Support Systems, Journal of Automation and Information Sciences 43(12), (2011): 21–32.

[15] Z.-Z. Liu and Y. Wang, Handling Constrained Multiobjective Optimization Problems With Constraints in Both the Decision and Objective Spaces, IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 870-884 (2019).

[16] A.F. Voloshin, G.N. Gnatienko, E.V. Drobot, A Method of Indirect Determination of Intervals of Weight Coefficients of Parameters for Metricized Relations Between Objects, Journal of Automation and Information Sciences, 35, (2003): 1–4.

[17] J.W. Kelly, A.D. Degenhart, D.P. Siewiorek, A. Smailagic and W. Wang, “Sparse linear regression with elastic net regularization for brain-computer interfaces” : in: IEEE Annual International Conference of the Engineering in Medicine and Biology Society, pp. 4275-4278, SanDiego, CA (2012).

[18] M.A. Rubeo and T.L. Schmitz, “Mechanistic force model coefficients: A comparison of linear regression and nonlinear optimization”, Precision Engineering, vol. 45 (2016).

[19] M. Antonevych, A. Didyk, V. Snytyuk, Optimization of Function of Two Variables by Deformed Stars Method, in: Proceedings of the 2019 IEEE international conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine, 2019, pp. 475-480, doi: 10.1109/ATIT49449.2019.9030453.

[20] V. Snytyuk, N. Tmienova, Method of Deformed Stars for Global Optimization, in: Proceedings of the IEEE 2nd international conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, 2020, pp. 1-4, doi: 10.1109/SAIC51296.2020.9239208.