International Scientific Symposium «Intelligent Solutions»

Models and Evolutionary Methods for Objects and Systems Clustering

Maryna Antonevych
Nataliia Tmienova
Vitaliy Snytyuk

Kyiv-2021

RELEVANCE

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 globalization of the processes of the modern world, the various challenges facing humanity, make us think about its future.

However, thinking about the future is not enough, we need to predict it and use intelligent decision-making methods to reduce future risks

CLUSTERING PROBLEM

Natural evolution

Genetic Algorithm, Evolutionary Strategy, **MODS**, etc.

Neural networks

Kohonen neural network with the implementation of the principles of WTA.

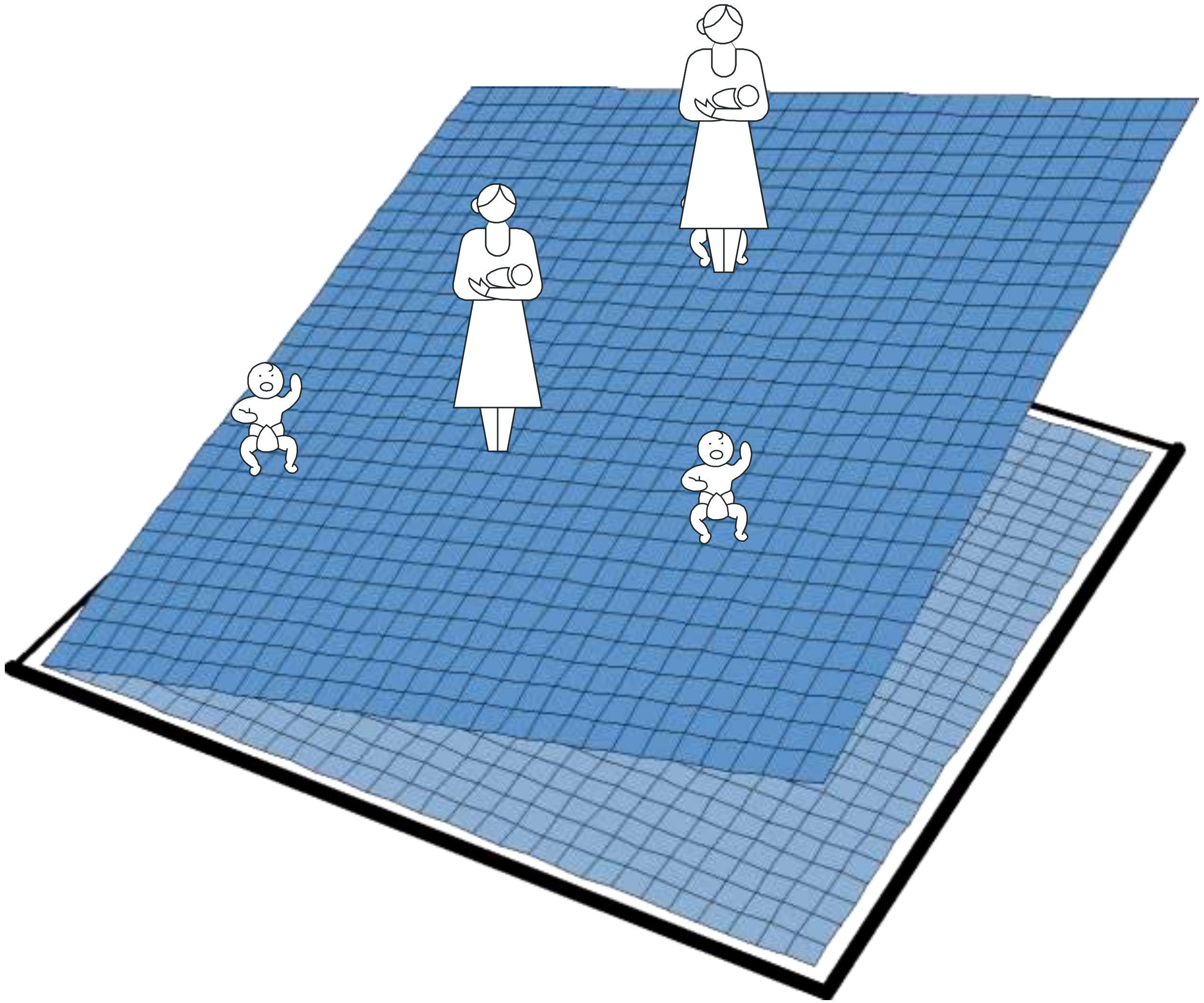
Next large group of clustering methods

Forel family of the Novosibirsk school of data analysis, the method of K-means and K-medians, EM-algorithm and others.

Hierarchical methods

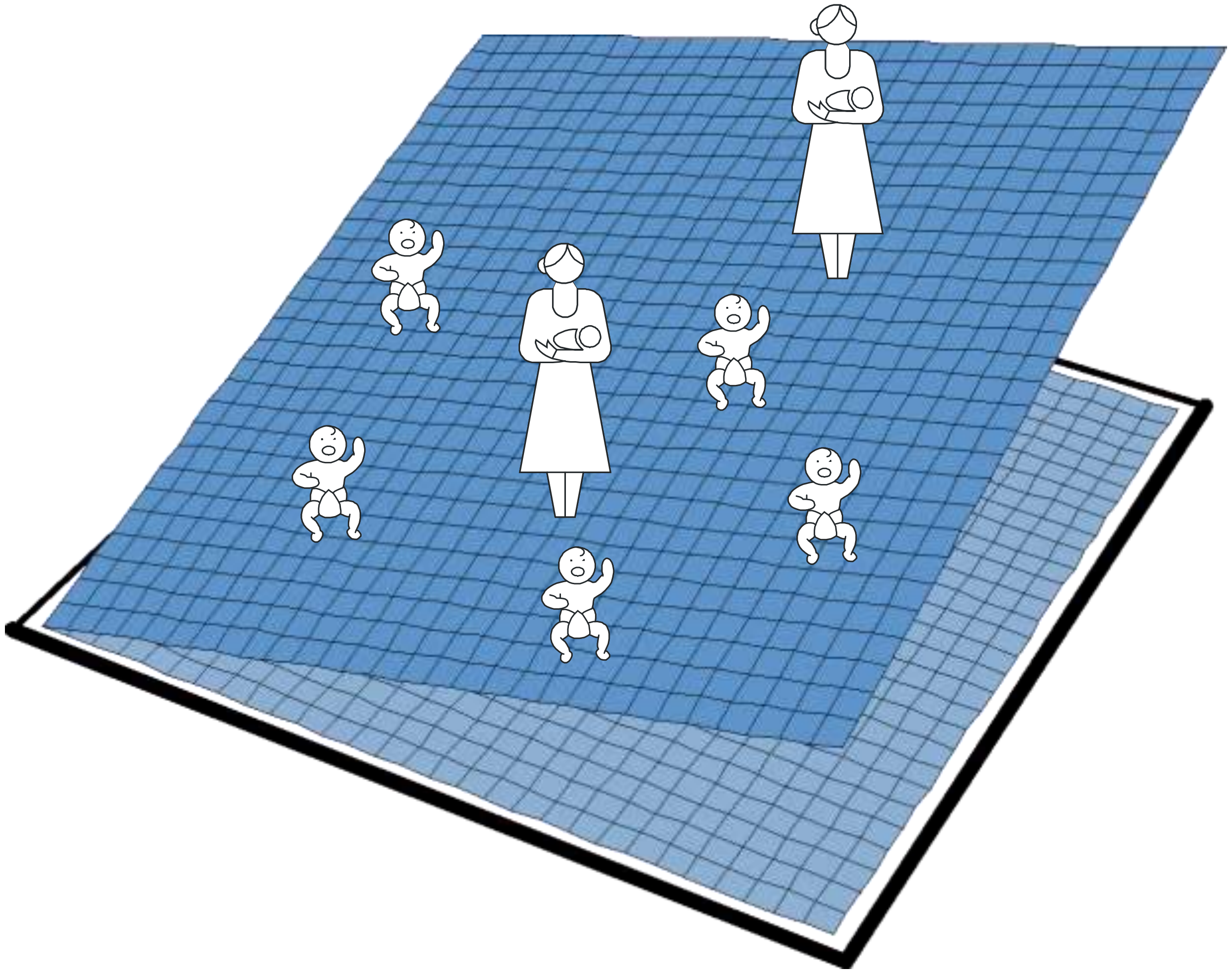
Ward's method, single linkage, pair-group method using arithmetic mean and complete linkage, and the pair-group method using centroid average.

ES & GA

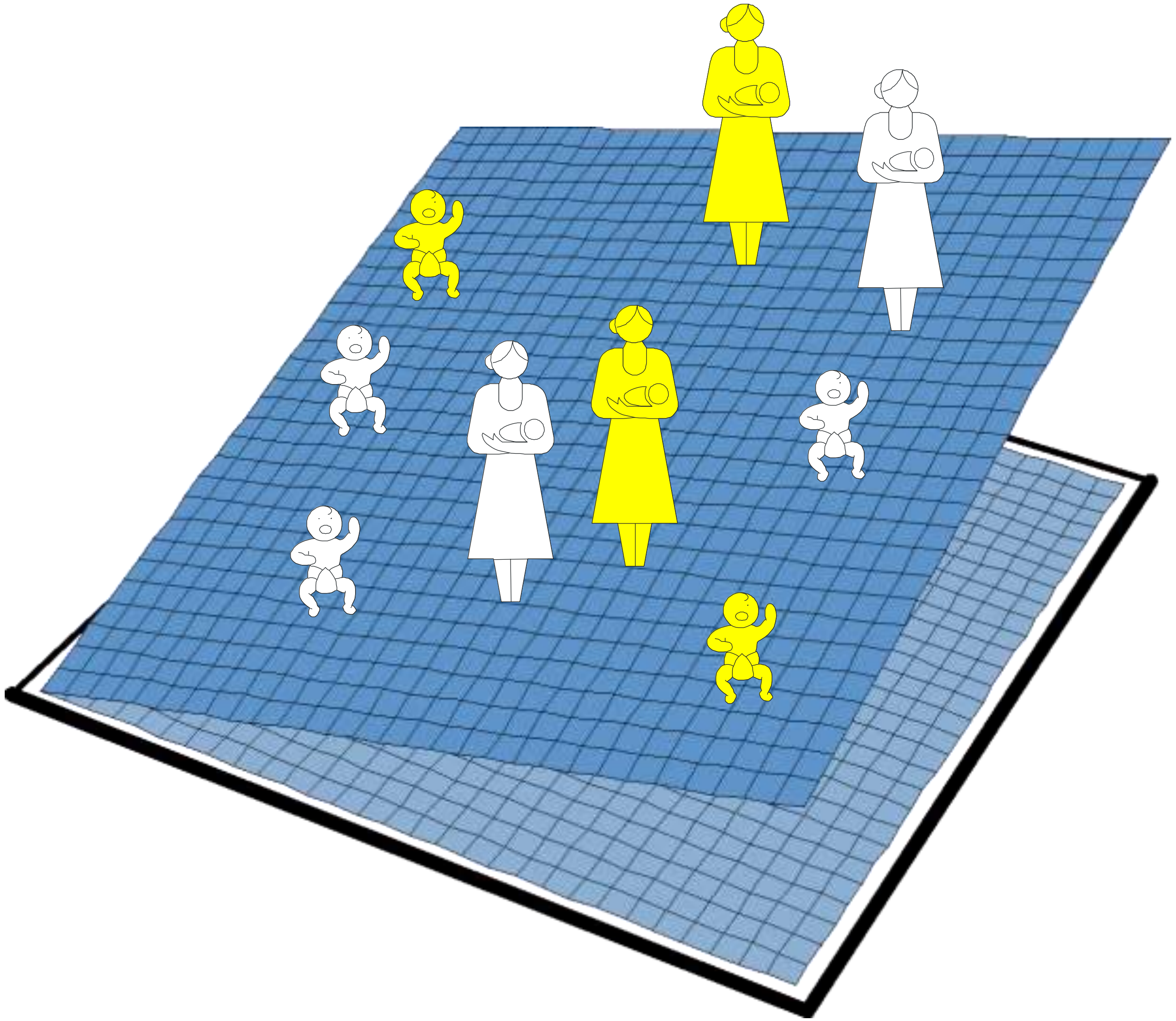


(1 + 1)-ES; Feature – one parent

EVOLUTIONARY STRATEGY



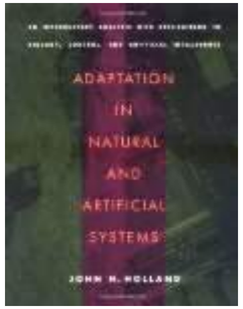
(1 , λ)-ES; ES with more than one offspring



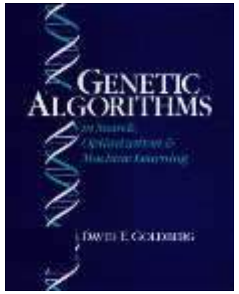
(μ, λ)-ES; ES with many parents and many offsprings

GENETIC ALGORITHM

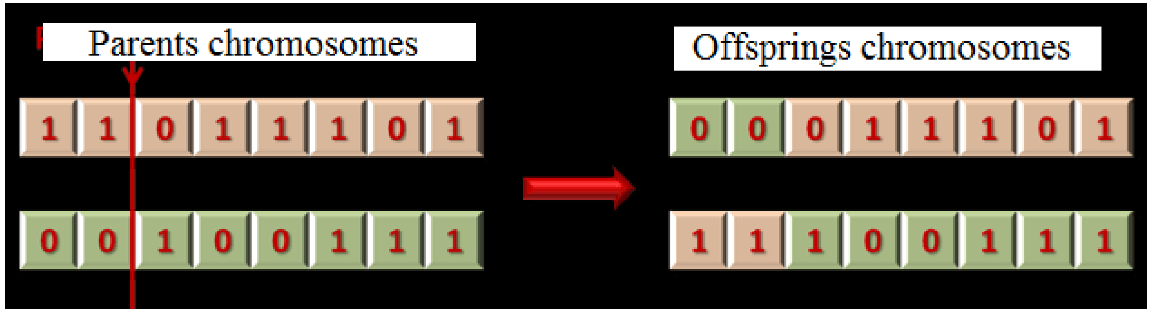
John Holland

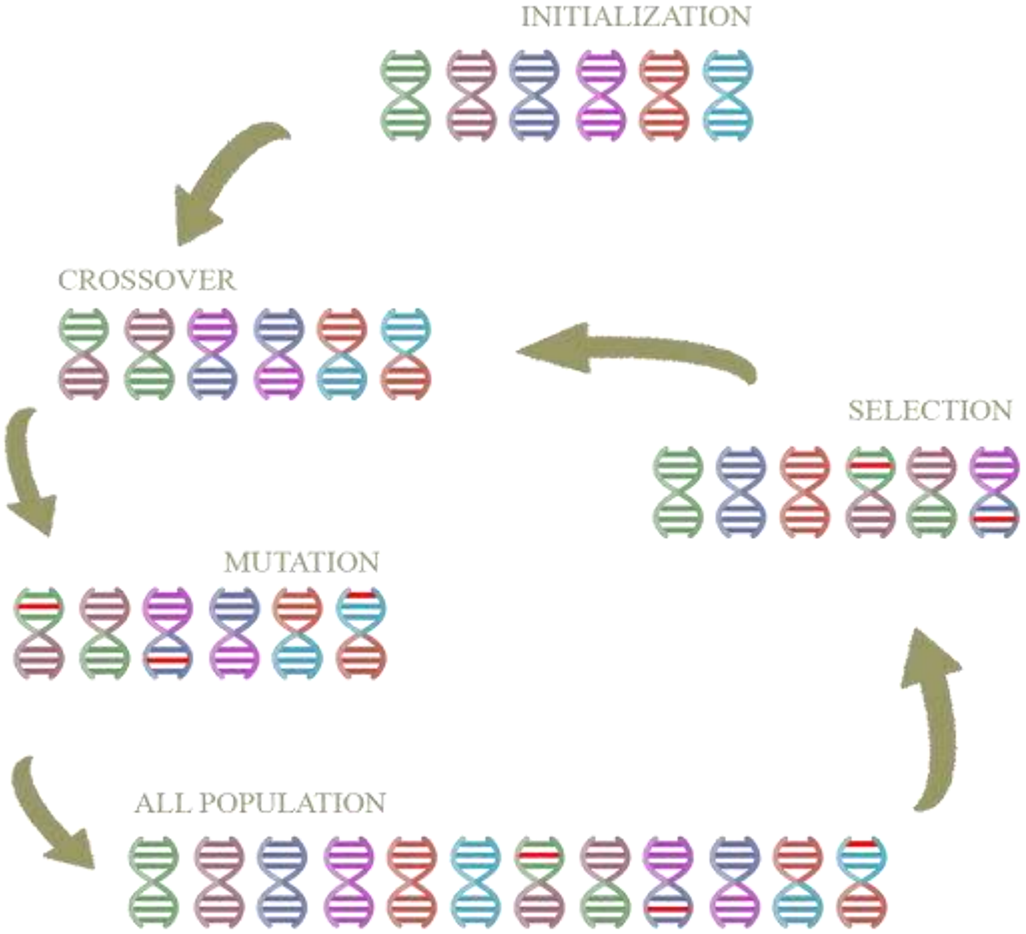


David Goldberg



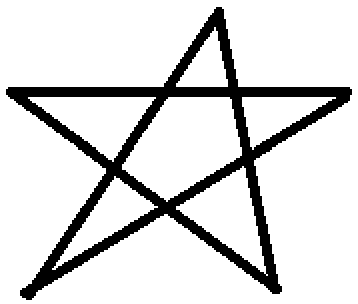
- Potential solution - chromosome
  - Feature – two parents



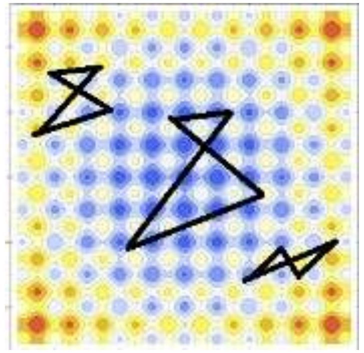


THE IDEA OF MODS

In order to deeply explore the search solution domain, it is necessary to use information from several solutions.



In order to accelerate convergence, it is necessary to comprehensively explore the neighborhood area of the best solutions



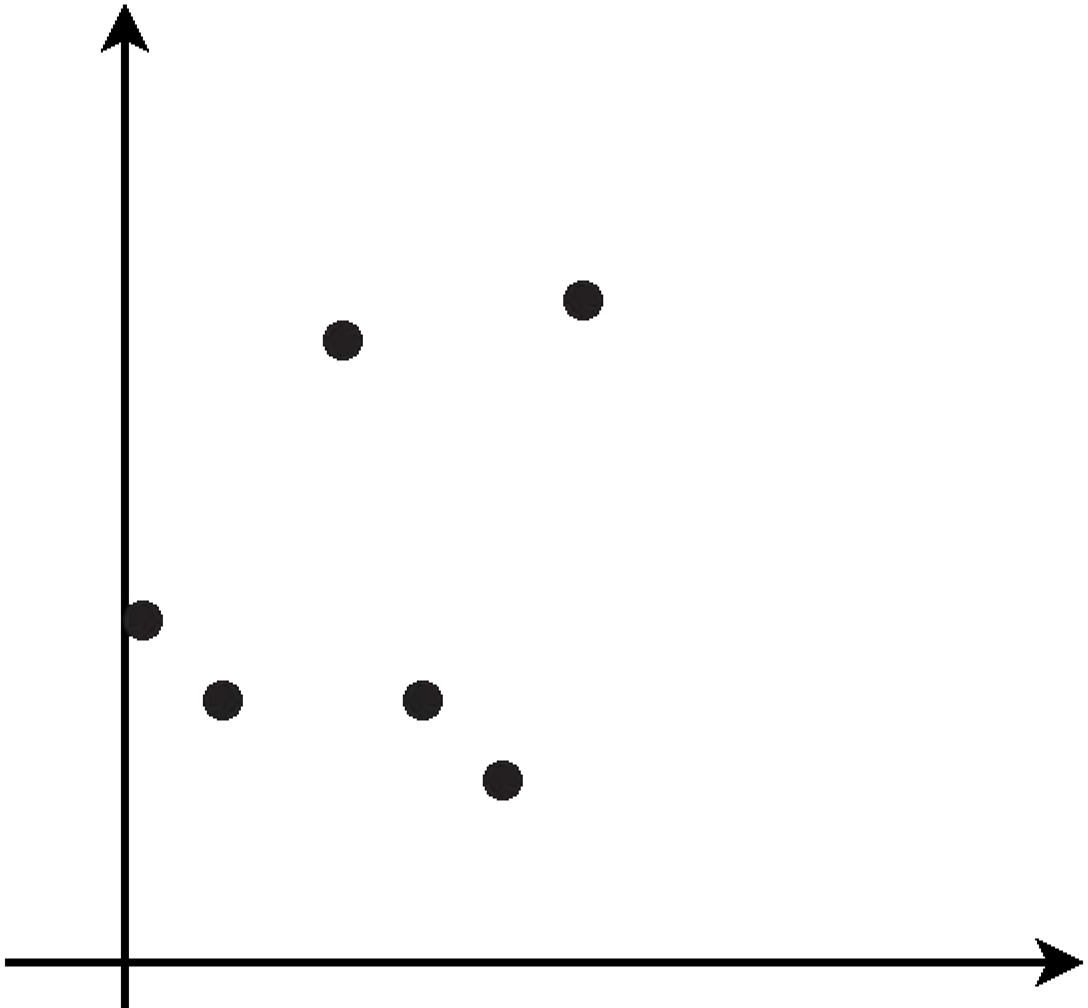
No Free Lunch Theorem

Method of deformed stars (MODS)

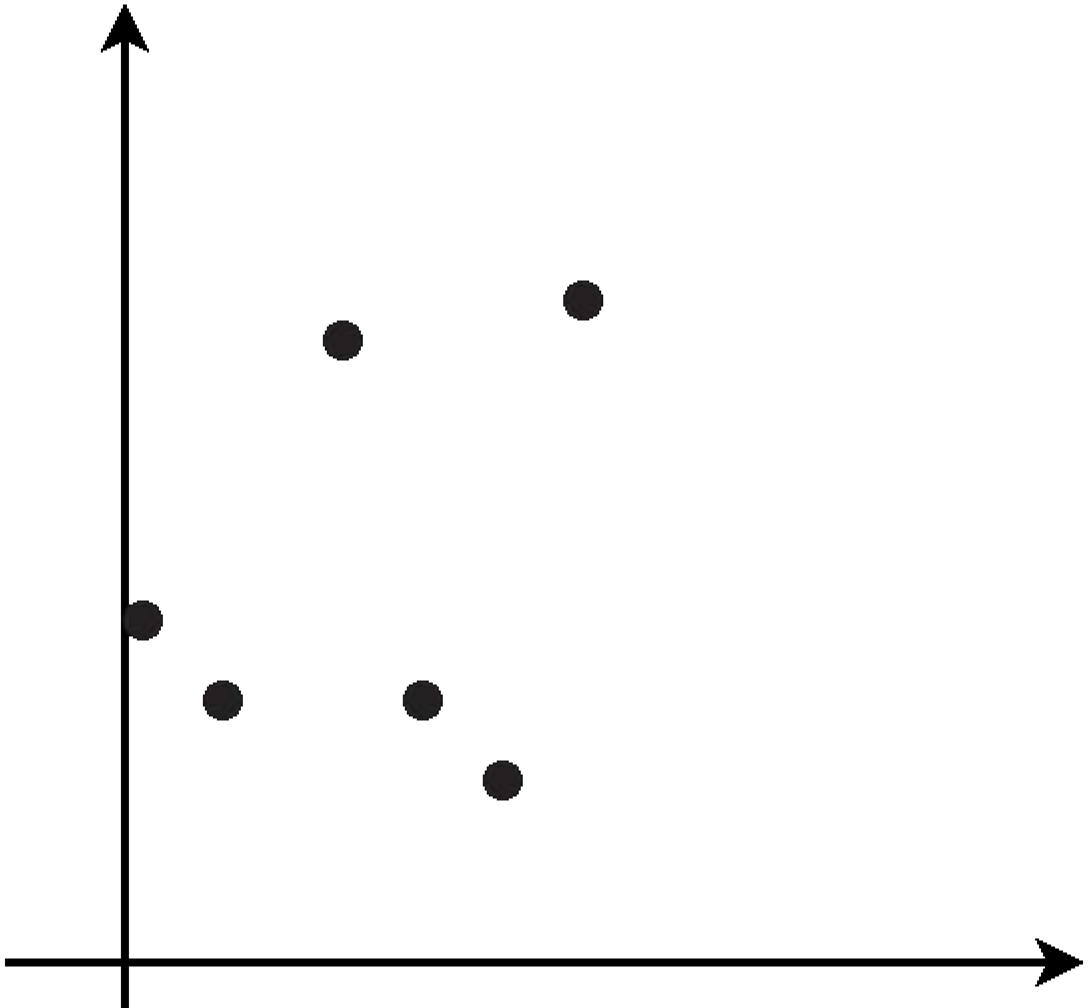
MODS-2

INTELLIGENT SOLUTIONS

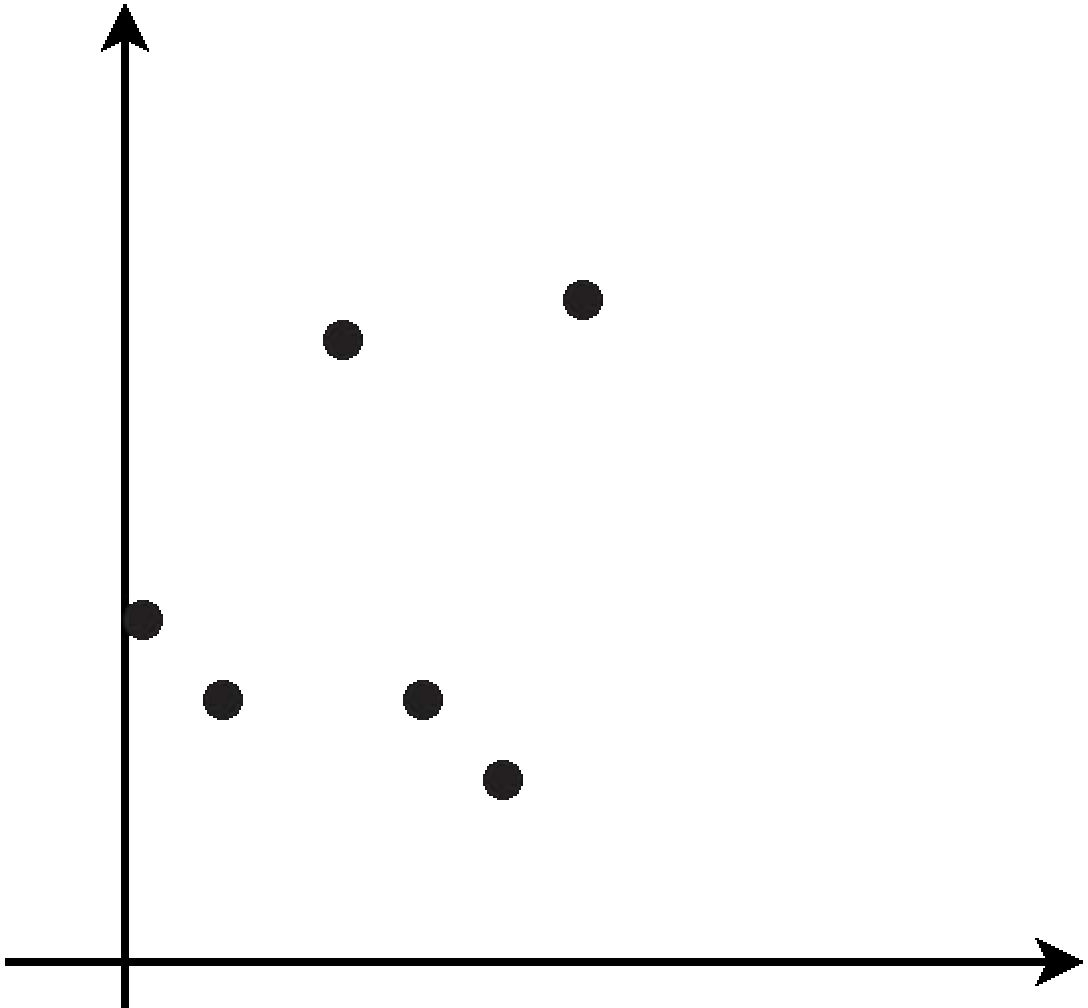
INITIALIZE THE POPULATION



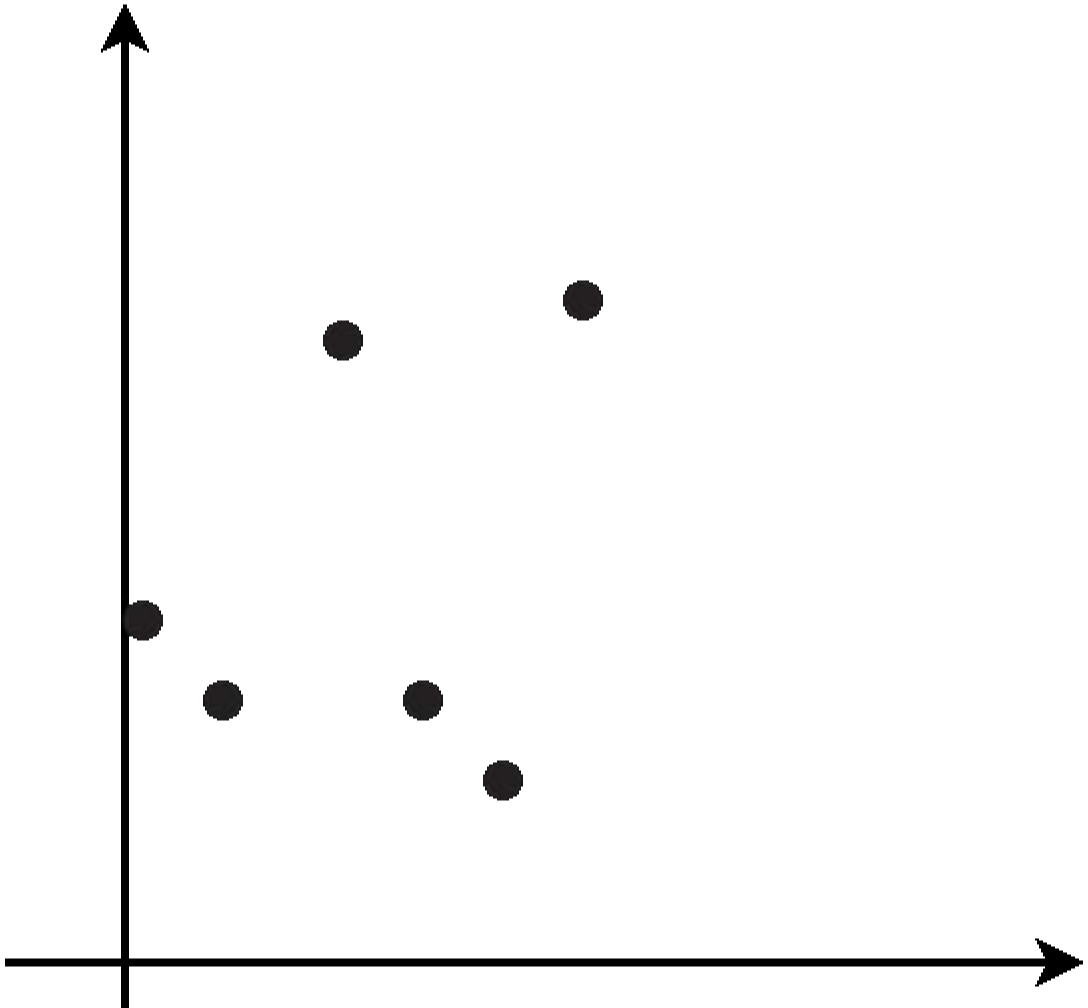
PARALLEL TRANSFER



TURN AROUND BEST POINT



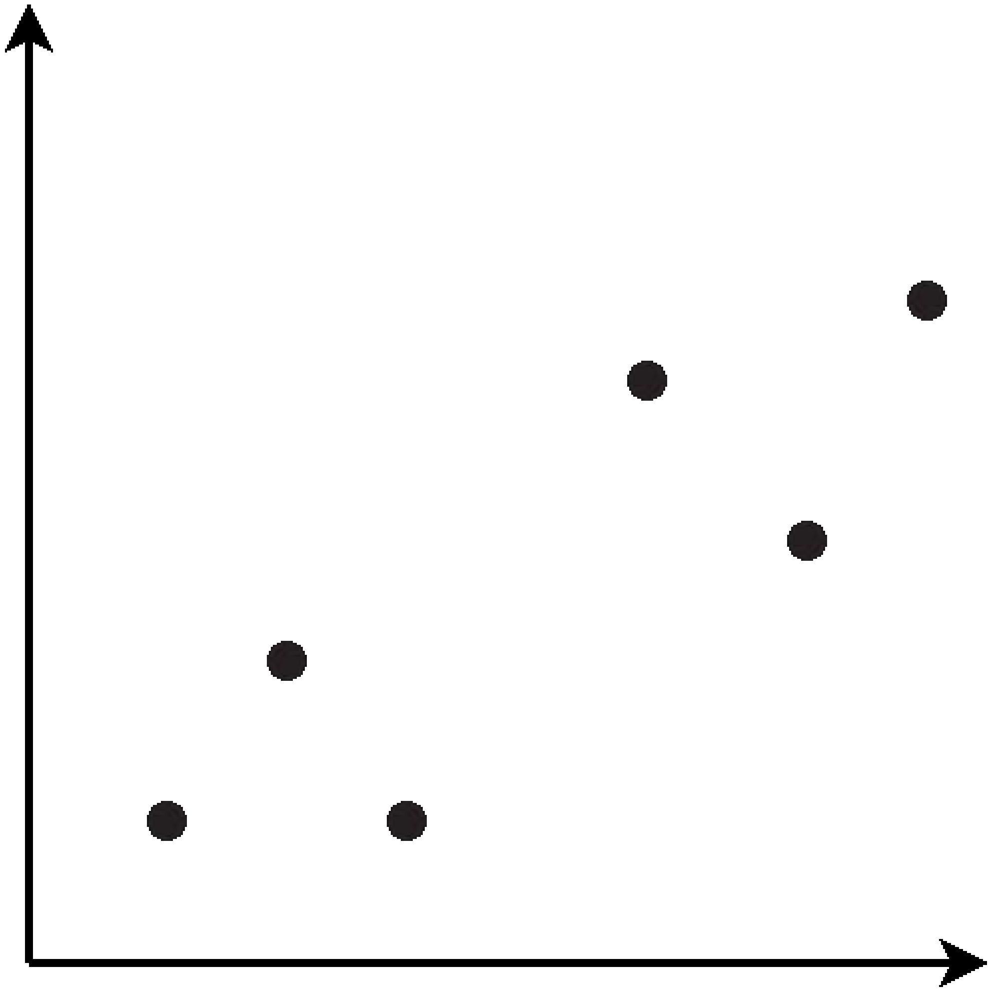
COMPRESSION



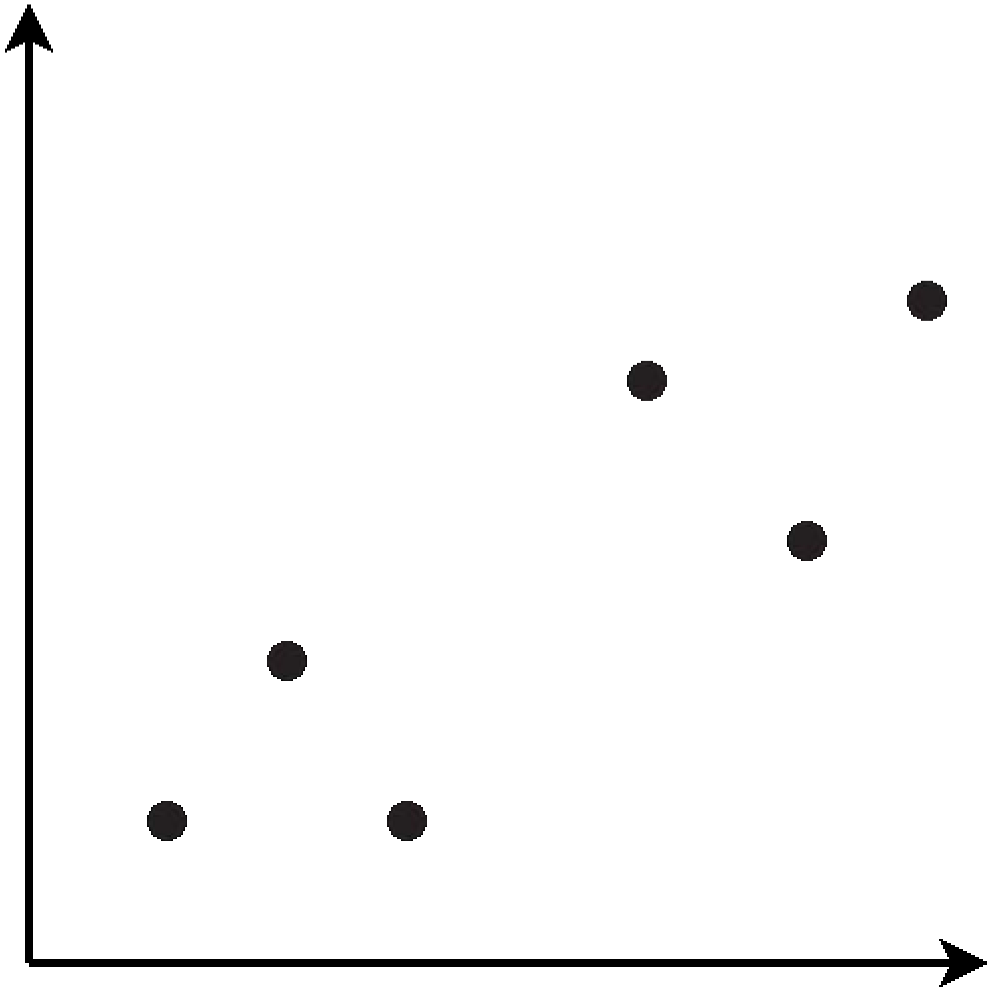
MODS-3

INTELLIGENT SOLUTIONS

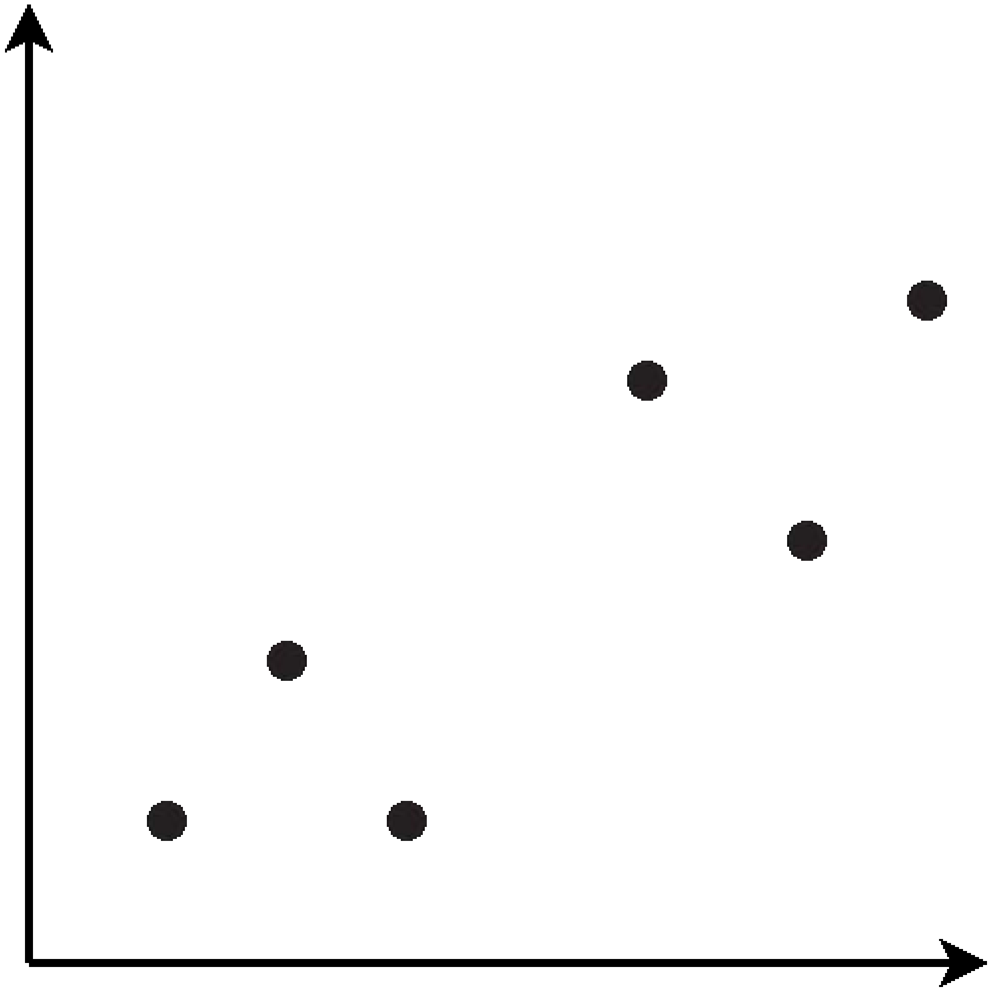
INITIALIZE THE POPULATION



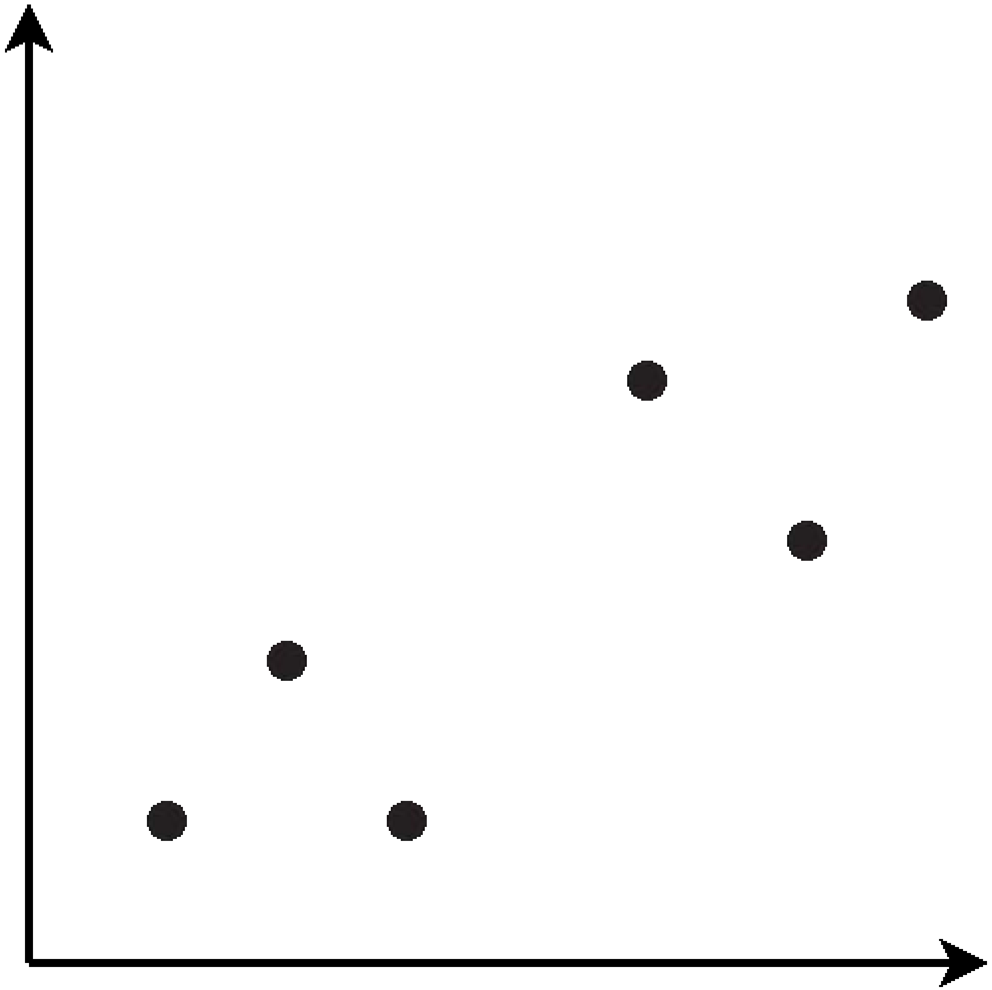
COMPRESSION



TURN AROUND BEST POINT

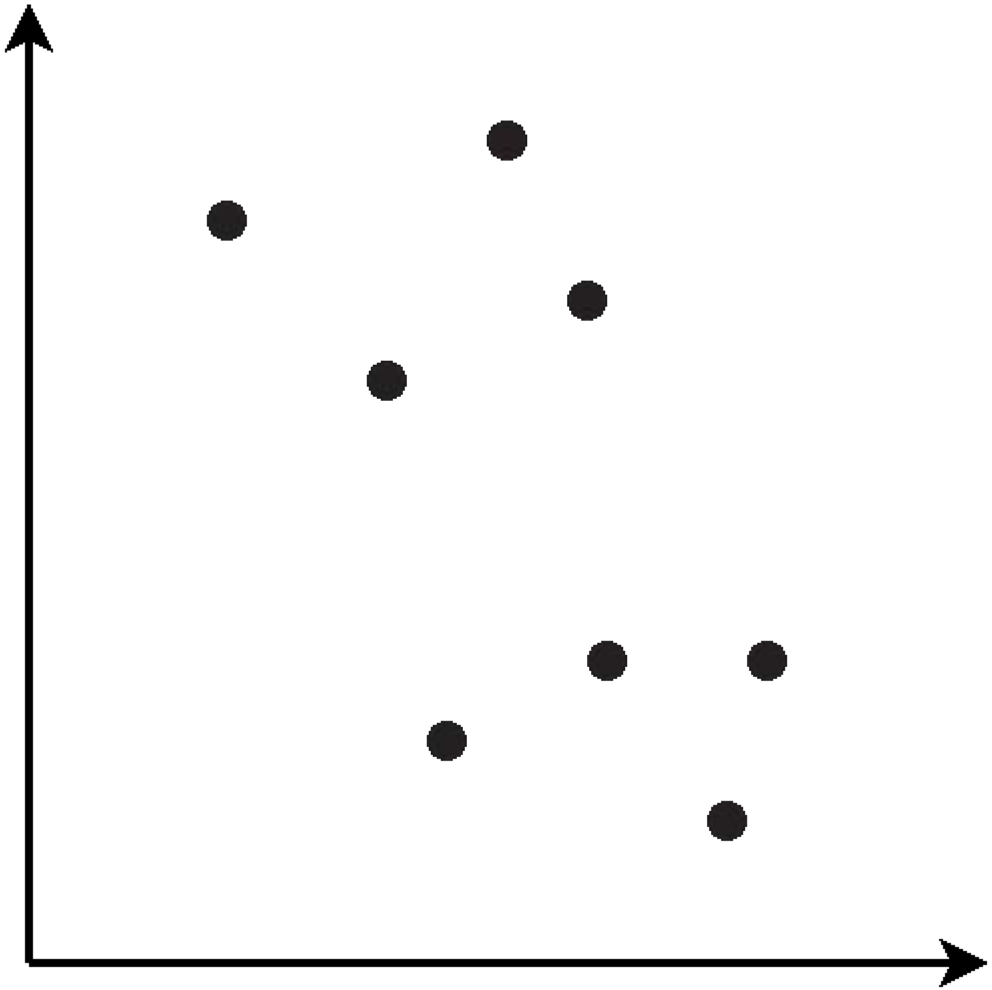


TURN AROUND CENTRE

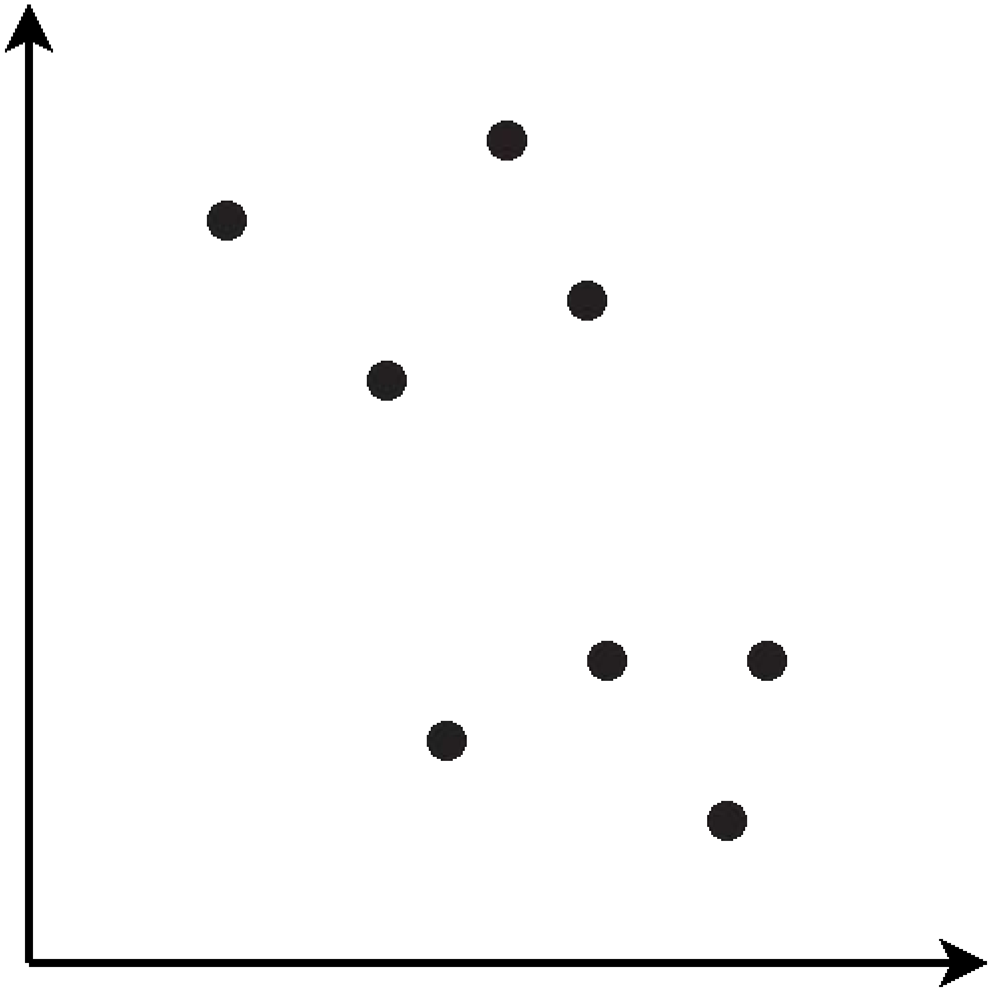


MODS-4

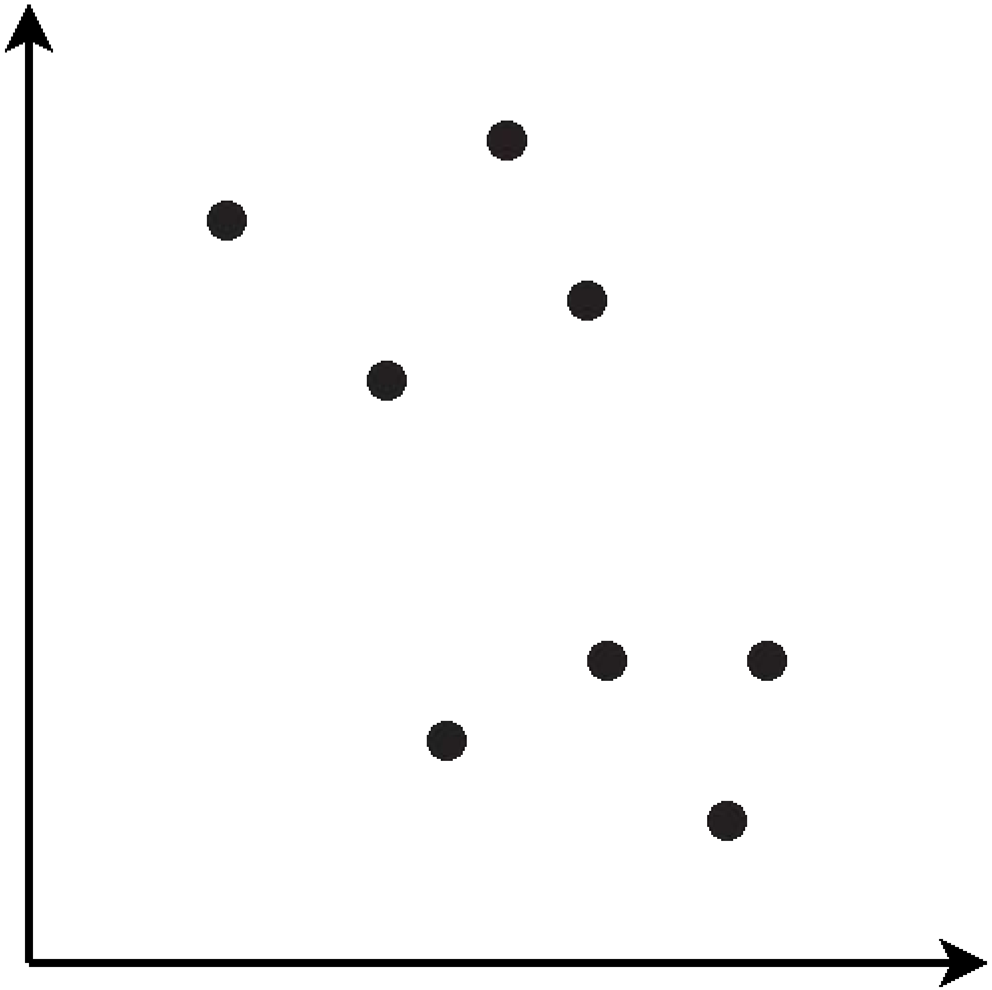
INITIALIZE THE POPULATION



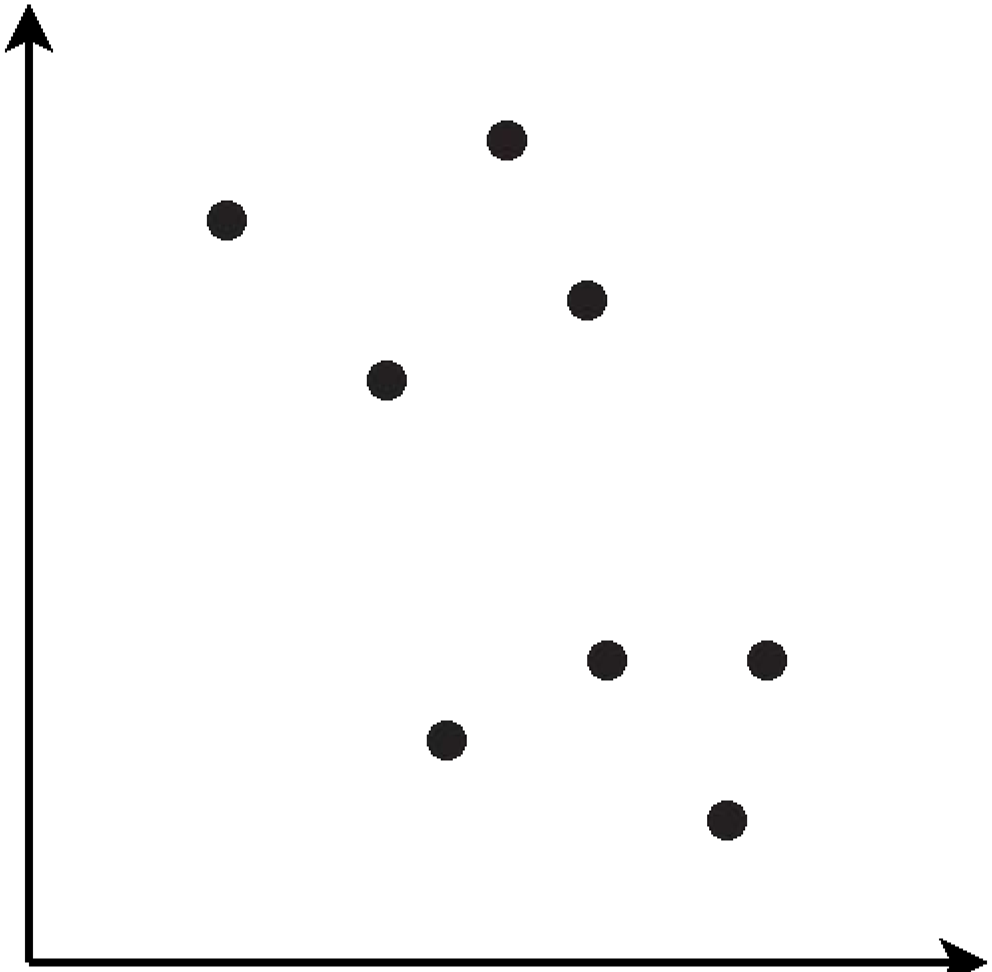
COMPRESSION



TURN AROUND BEST POINT

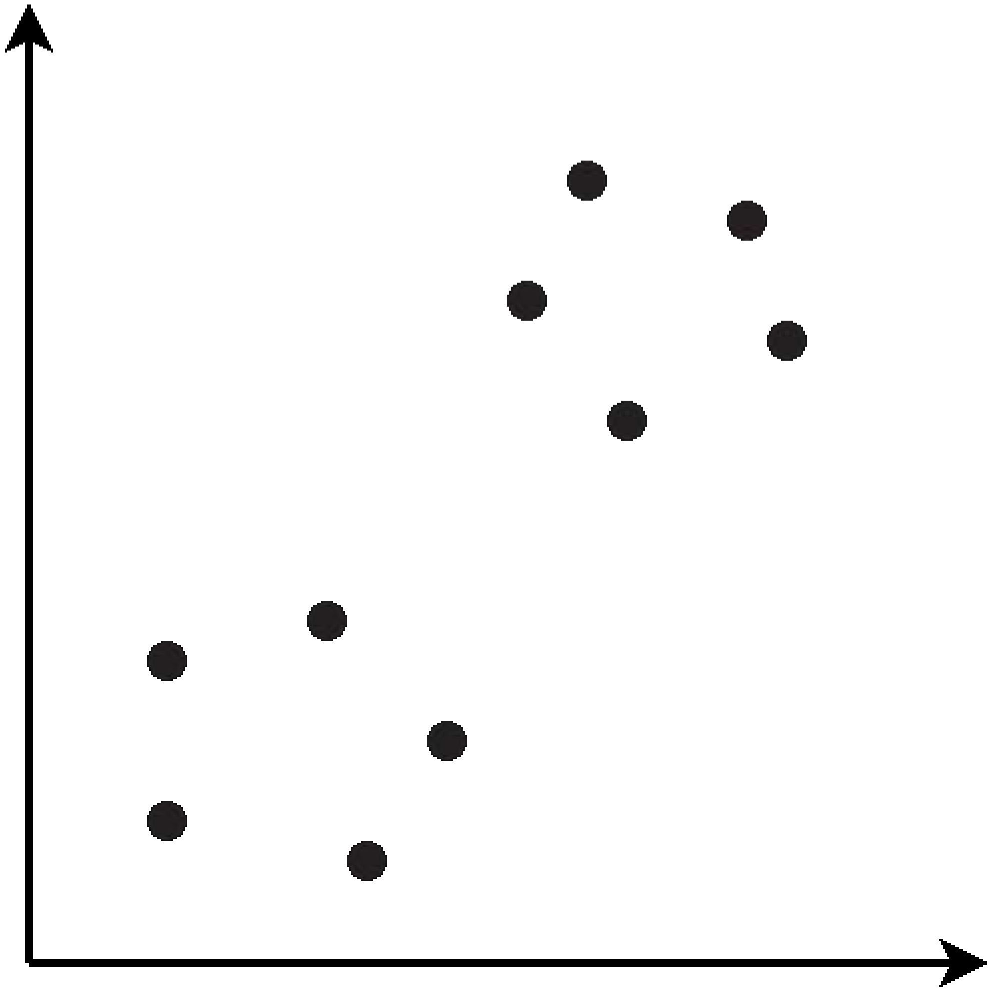


TURN AROUND CENTRE

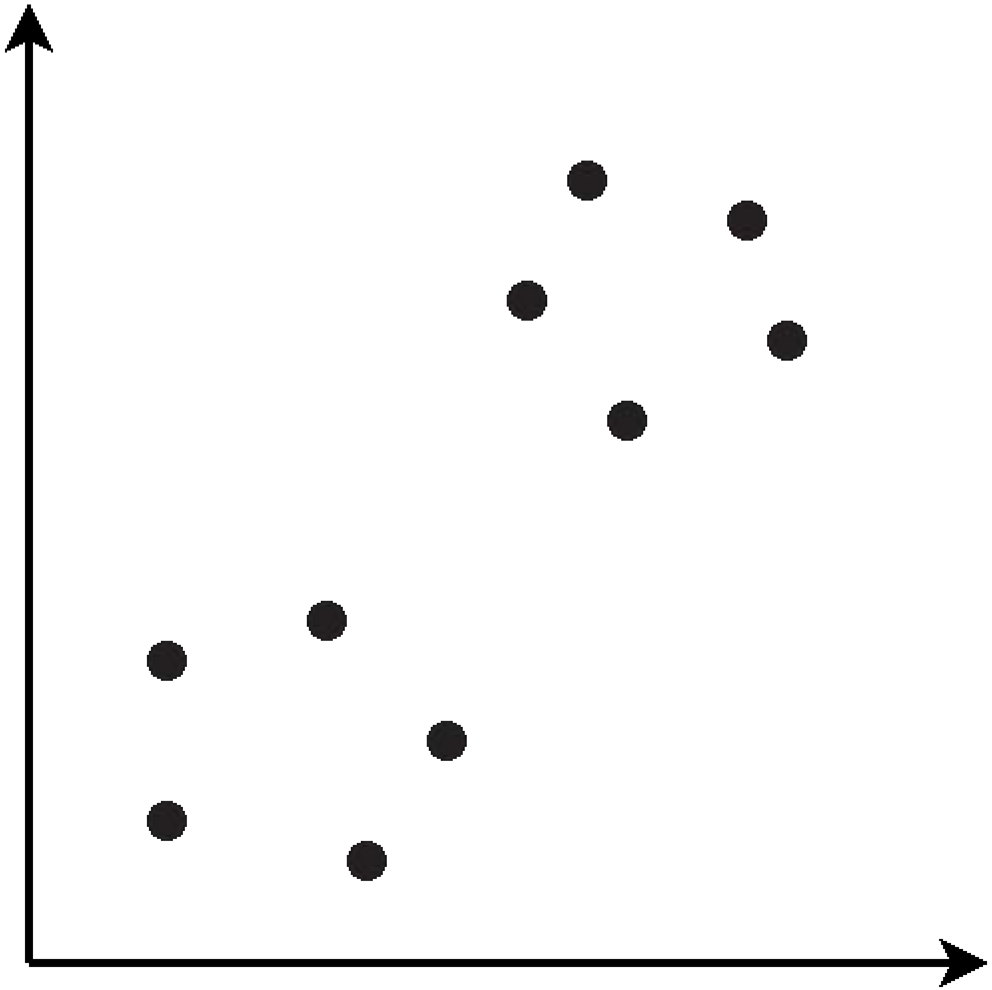


MODS-5

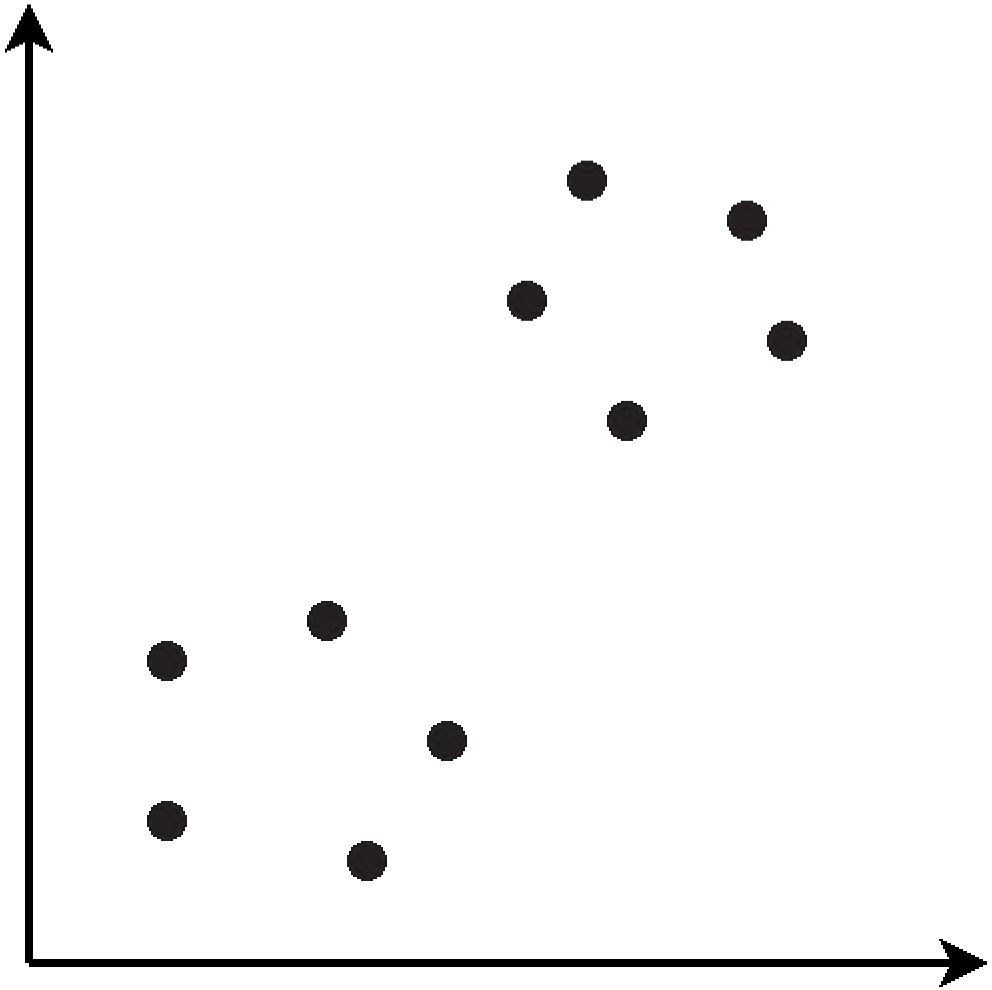
INITIALIZE THE POPULATION



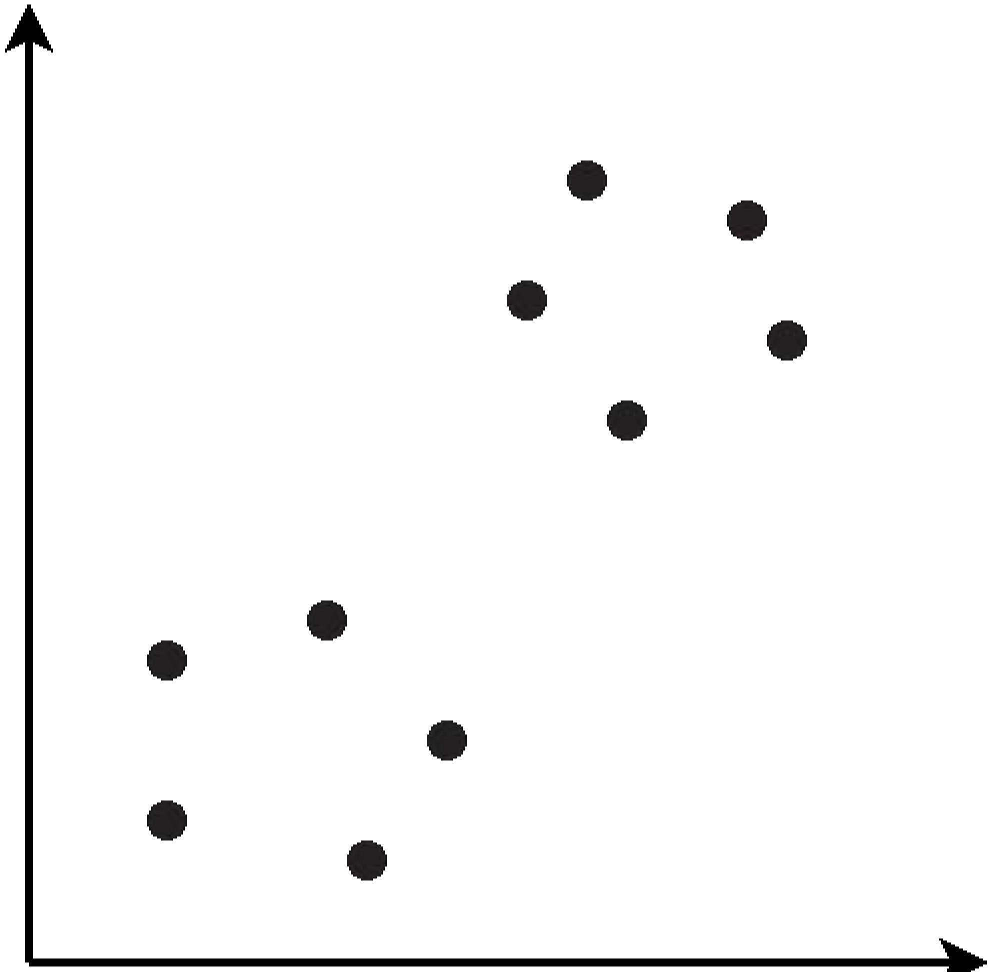
COMPRESSION



TURN AROUND BEST POINT



TURN AROUND CENTRE



DISADVANTAGES OF EA

the heuristic nature of evolutionary computations **does not guarantee the optimality of resulting solution** (in practice, it is often important to obtain one or more suboptimal alternative solutions in a given time, especially since the initial data can be dynamically modified, inaccurate or incomplete);

relatively high **computational complexity**, which is overcome by parallelization at the level of evolutionary computations organization and at the level of their direct implementation in the computational systems;

relatively low efficiency in the final phases of evolution modeling (search operators in evolutionary algorithms are not focused on getting to the local optimum quickly); unresolved problem of self-adaptation.

CLUSTERING. PROBLEM FORMULATION

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:

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

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 $K \ll M$.

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

$$d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^{n} (x_{ik} - x_{jk})^2\right)^{1/2}, i, j = \overline{1,m} \ .$$

INTELLIGENT SOLUTIONS

CLUSTERING WITH EA

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

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

where C is the set of point-centers of the clusters, Ci is the point-center of the i-th cluster, j is the index that indicates the affiliation of Xj 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:

CLUSTERING WITH EA

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

or

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

Let us solve the problem of finding cluster centers

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

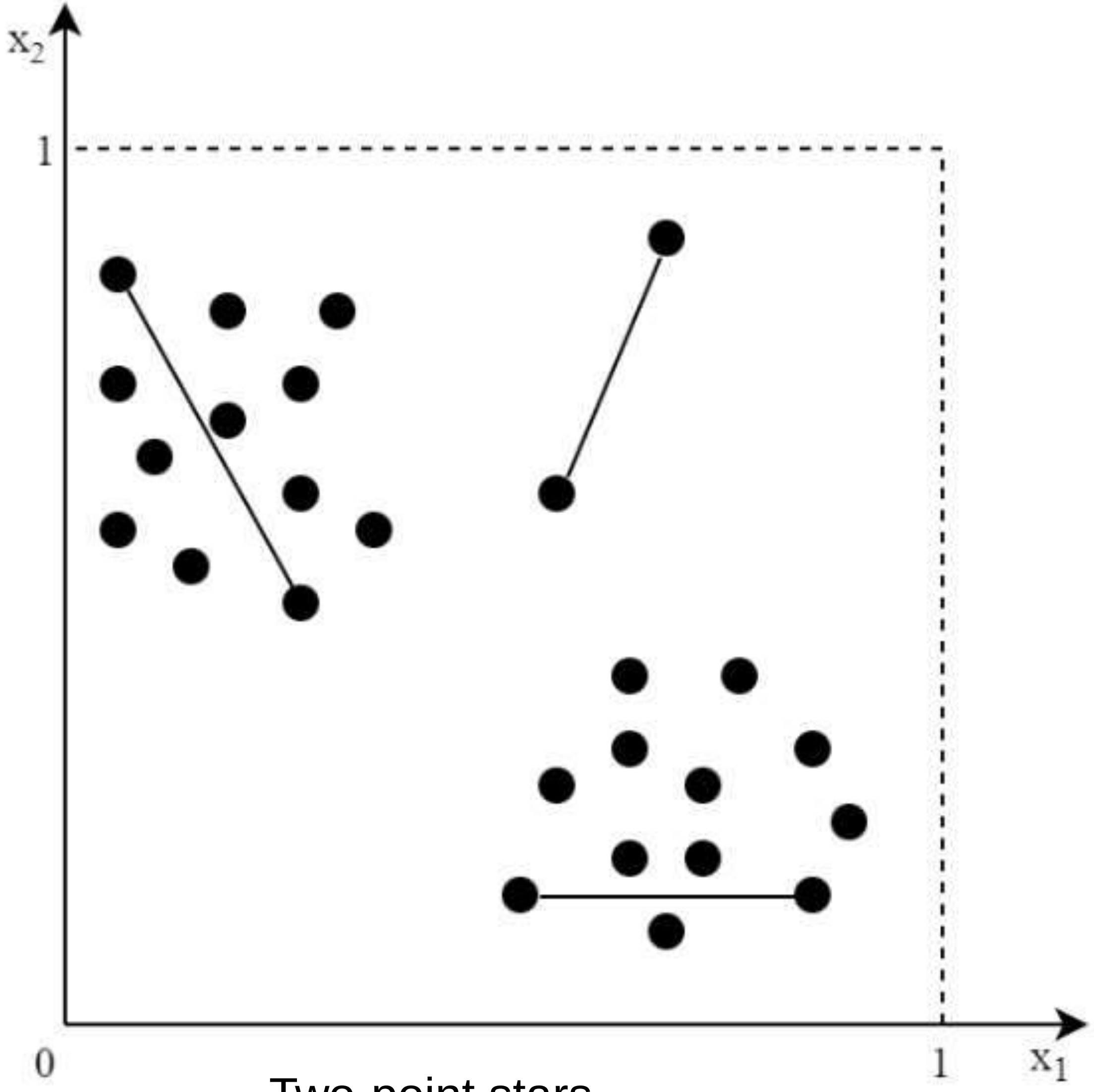
CMODS. ALGORITHM 1

CMODS ALGORITHM

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.

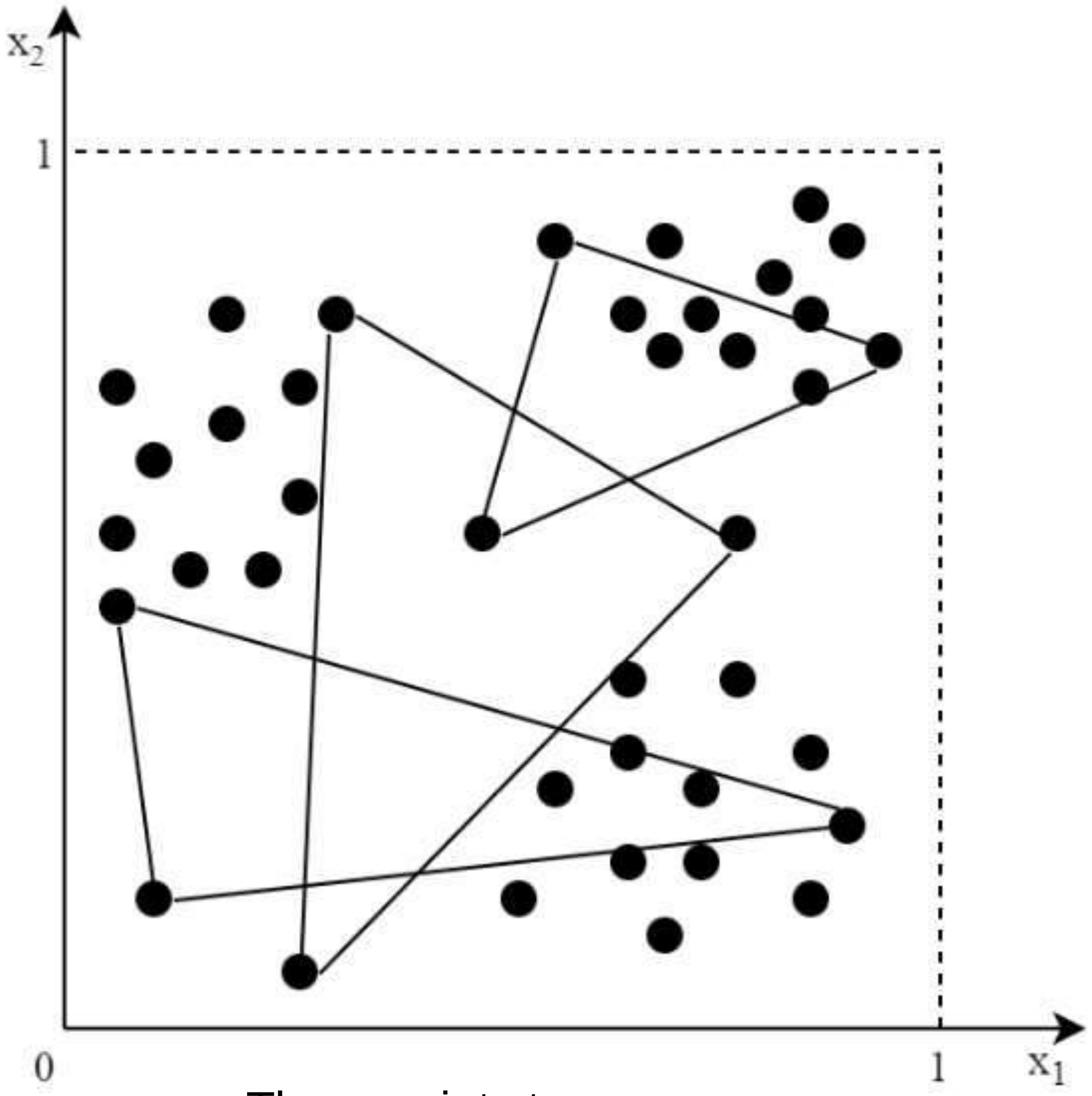
Similar ideas can be used to solve the clustering problem, because it has an objective function, the minimum of which must be found, and objects that are points in *n*-dimensional space.

CMODS ALGORITHM



Two-point stars

CMODS ALGORITHM



Three-point stars

CMODS ALGORITHM 1

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

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

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

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:

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

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

$$x_v^{jl} = x_v^l, v \neq \theta, \eta$$

CMODS ALGORITHM 1

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.

CMODS ALGORITHM 1

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.

CMODS. ALGORITHM 2

INTELLIGENT SOLUTIONS

CMODS ALGORITHM 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.

CMODS ALGORITHM 2

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:

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

CMODS ALGORITHM 2

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.

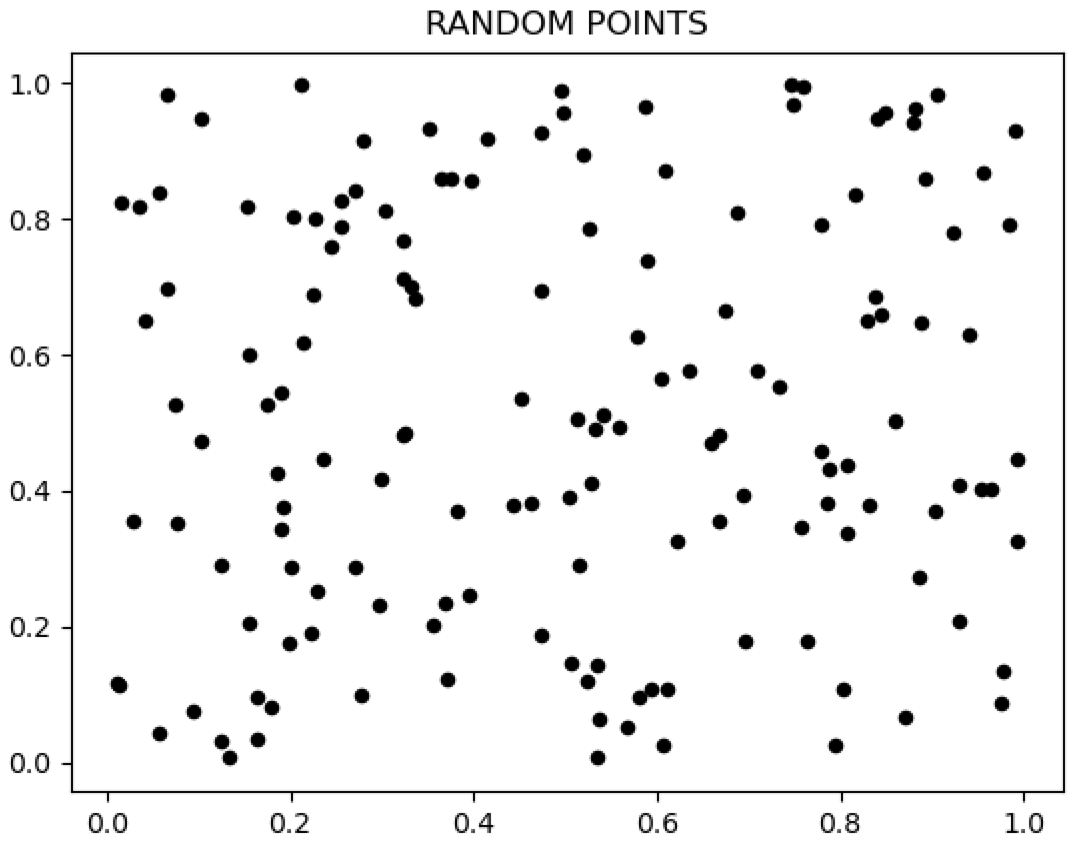
The difference between the second method of clustering and the previous methods is that only one modification of the method of deformed stars is used. Such a modification is a special type of mutation operation. Mutation allows to simultaneously avoid hitting the local optimum and more deeply explore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

EXPERIMENTAL RESULTS

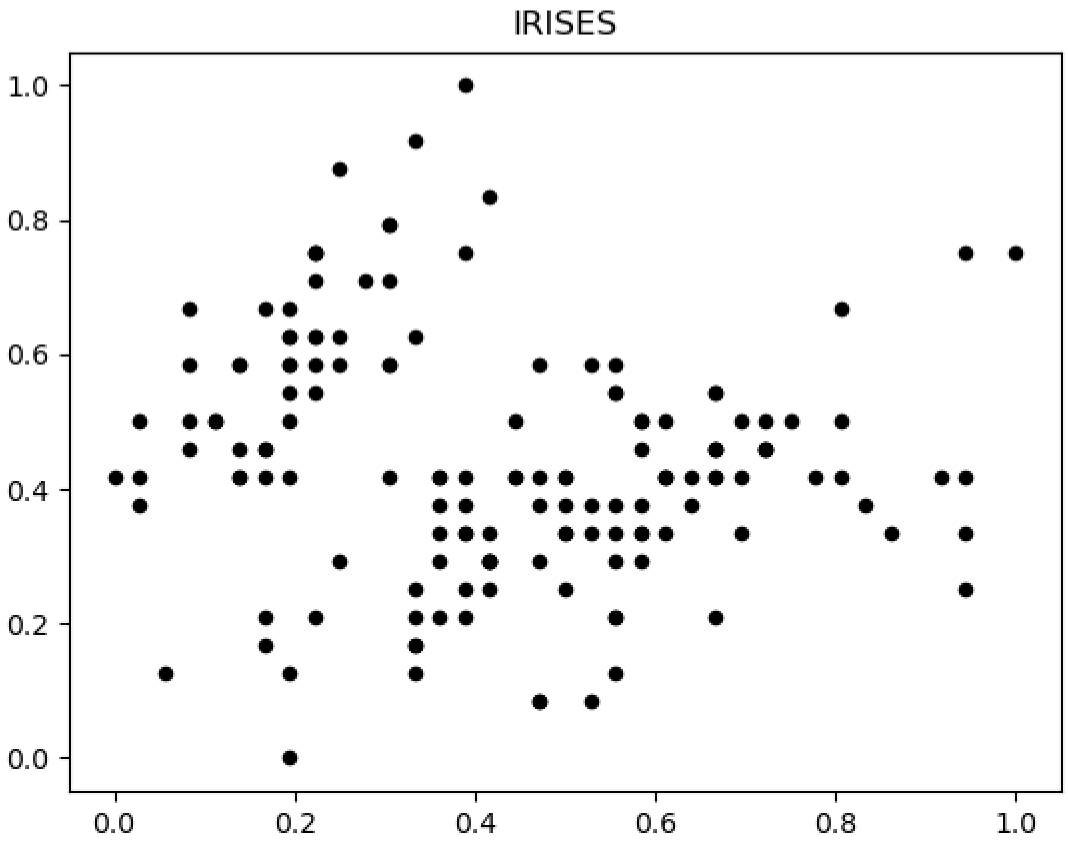
EXPERIMENTAL RESULTS

Consider a well-known sample of **data on irises**, which contains fields such as Petal Length and Widht, Sepal Length and Width, Classes Types.

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.



Random points distribution



Irises data

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.

EXPERIMENTAL RESULTS

The first experiment concerned **randomly generated points**. 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.

For **iris clustering** experiment the results of the CMODS method remain the best.

EXPERIMENTAL RESULTS

	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

Table 1. Test results for the objective function of type (1)

EXPERIMENTAL RESULTS

In experiment with **another objective function** 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.

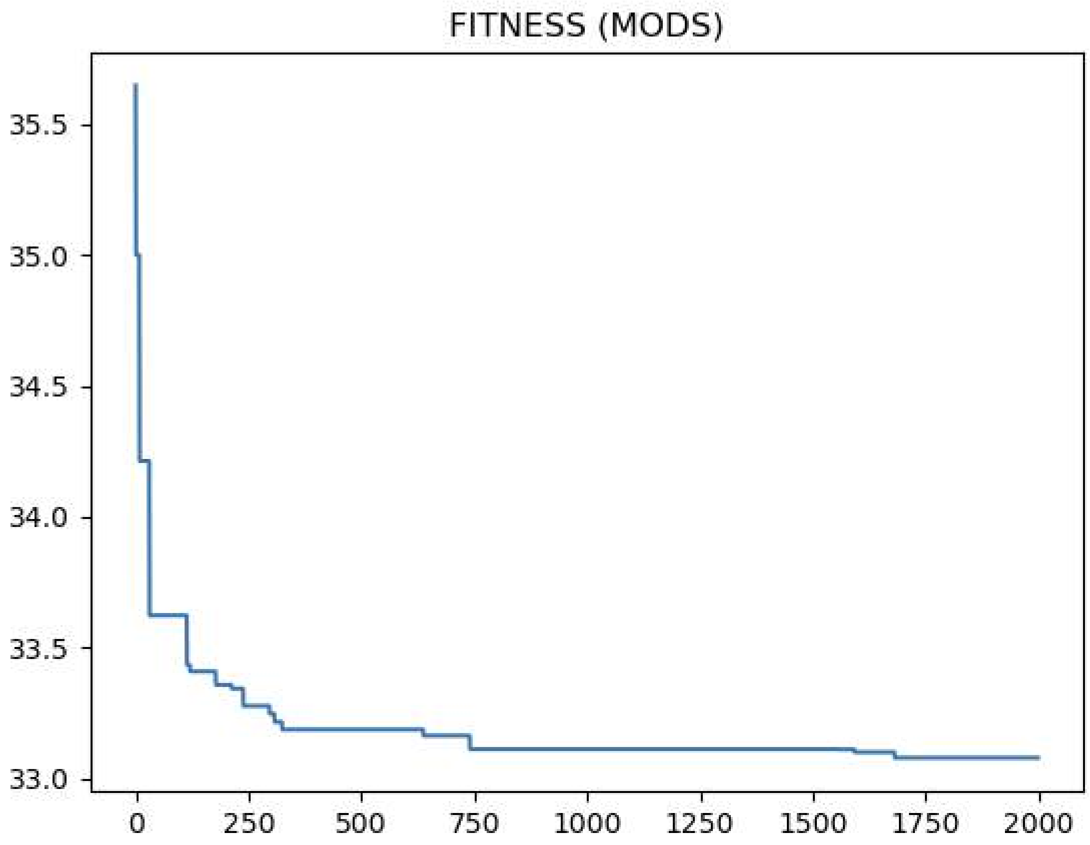
Methods based on other population algorithms show good results, but they are worse than the results of the K-means method.

EXPERIMENTAL RESULTS

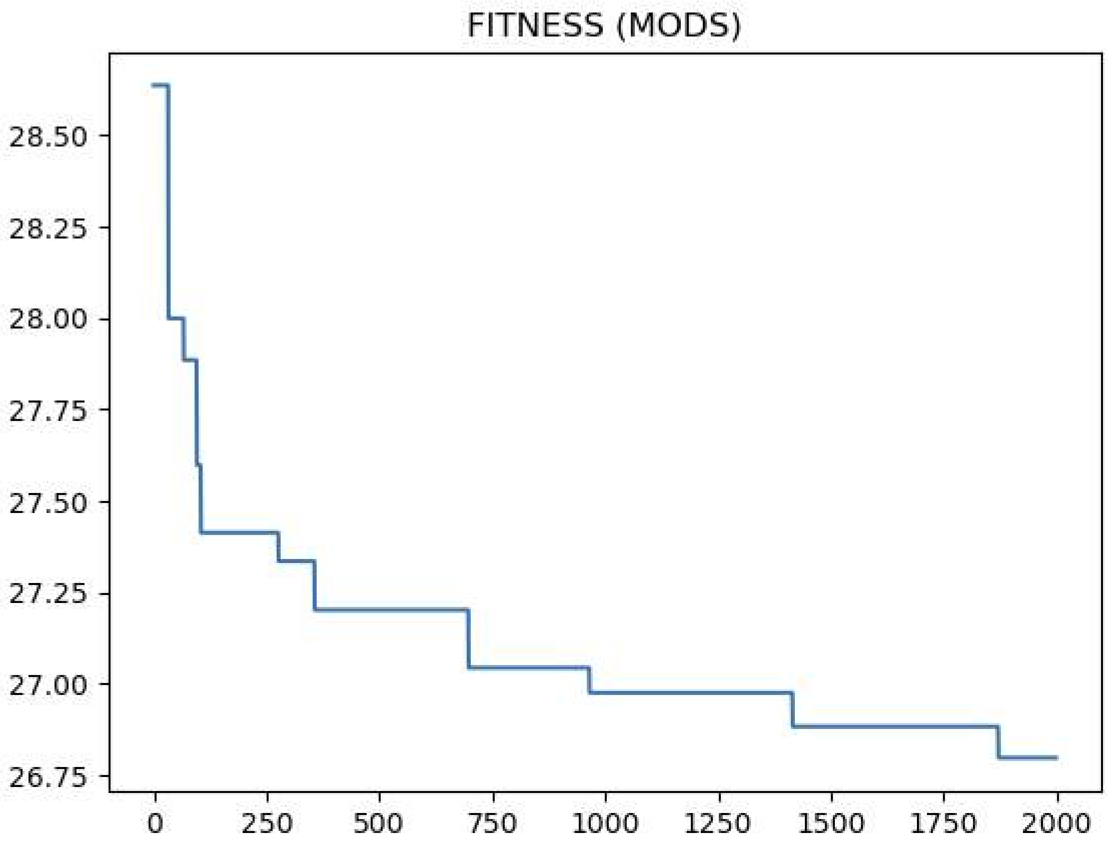
	Random points		Irises	
	1000 iter	2000 iter	1000 iter	2000 iter
K-means	32.11	33.34	27.11	27.11
Forel	38.71	43.12	36.39	36.39
CGA	32.49	35.07	28.89	28.98
CES	31.26	33.64	32.36	31.15
CMODS	30.92	33.07	27.06	26.79

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

EXPERIMENTAL RESULTS



Fitness-function for random points



Fitness-function for irises

CONCLUSION

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.

The method of deformed stars has demonstrated its **advantages** in most of 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.

CONCLUSION

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.

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.

Thank you for attention!

Maryna Antonevych	marina.antonevich@gmail.com	Ukraine	Taras Shevchenko National University of Kyiv
Nataliia Tmienova	tmyenovox@gmail.com	Ukraine	Taras Shevchenko National University of Kyiv
Vitaliy Snytyuk	snytyuk@knu.ua	Ukraine	Taras Shevchenko National University of Kyiv