# EVOLUTIONARY CLUSTERING OF COMPLEX SYSTEMS AND PROCESSES

Vitaliy Snytyuk

*Abstract: In a paper the method of complex systems and processes clustering based use of genetic algorithm is offered. The aspects of its realization and shaping of fitness-function are considered. The solution of clustering task of Ukraine areas on socio-economic indexes is represented and comparative analysis with outcomes of classical methods is realized.*

*Keywords: Clustering, Genetic algorithm.*

*ACM Classification Keywords: I.5.3. Clustering*

## Introduction

The process of a translational movement to creation of an information community is accompanied with problems connected to storage and handling of large scale arrays of data. Their solution is connected to intellectual analysis of data, which process engineering are formed on intersection of artificial intelligence, statistics, theory of data bases. There are KDD (knowledge discovery in databases) - detection of knowledge in data bases, data mining, OLAP (on-line analytical processing) - extraction of an information from many-dimensional data bases and others. The elements of indicated process engineering become the integral part of electronic storages of data (warehouses). A significant part of information make data, being socio-economic indexes of operation of complicated systems.

The presence of noise effects is peculiar to large scale arrays of information, their handling reduces in accumulation of a cumulative error. For overcoming an indicated problem it is necessary to determine the significant factors and to realize their analysis. The diminution of information entropy can be also reached by a grouping of systems and extraction of knowledge in smaller and functionally connected populations. Such procedures are directed on sequential overcoming of indeterminacy. The first step in this direction is a solution of the clustering task.

## The Analysis of Models and Methods of Clustering

The clustering task consists in the definition of systems or processes groups, which are closest one to another on some criterion. Thus of any suppositions about their structure, as a rule, is not done [Mandel, 1988], [Gorban, 2002]. The majority of clustering methods is founded on analysis of a factors matrix of a likeness, as which appear a distance, contingency, correlation etc. If by a criterion or metric the distance appears, as a cluster name group of points $\Omega$, such, that the average square of inside group distance up to a centre of group is less than an average distance up to a common centre in an initial set of systems, i.e. $\bar{d}_{\Omega}^{2} < \sigma^{2}$, where $\bar{d}_{\Omega}^{2} = \frac{1}{N}\sum_{X_i \in \Omega}(X_i - \bar{X}_{\Omega})^2$, $\bar{X}_{\Omega} = \frac{1}{N}\sum_{X_i \in \Omega} X_i$. Generally, criterions are:

1. Euclid's distance $d(X_k, X_l) = \left(\frac{1}{m}\sum_{j=1}^{m}(X_{kj} - X_{lj})^2\right)^{\frac{1}{2}}$.
2. Maximum distance on indications $d(X_k, X_l) = \max_{1 \le j \le m}\left|X_{kj} - X_{lj}\right|$.

3. Mahalonobis distance $d(X_k, X_l) = [(X_k - X_l) \cdot R^{-1} \cdot (X_k - X_l)^T]^{\frac{1}{2}}$.
4. Hamming's distance $d(X_k, X_l) = \frac{1}{m} \sum_{j=1}^{m} |X_{kj} - X_{lj}|$.

The solution of minimization task of a distance between systems is equivalent to a solution of minimization task of a distance up to system having averaged performances, as, for example, for a Hamming's distance

$$\sum_{\substack{j=1 \\ k<l}}^{m} |X_{kj} - X_{lj}| = \sum_{\substack{j=1 \\ k<l}}^{m} |X_{kj} - \overline{X} + \overline{X} + X_{lj}| \le \sum_{\substack{j=1 \\ k<l}}^{m} |X_{kj} - \overline{X}| + \sum_{\substack{j=1 \\ k<l}}^{m} |X_{lj} - \overline{X}| \le \sum_{j=1}^{m} |X_{kj} - \overline{X}| + \sum_{j=1}^{m} |X_{lj} - \overline{X}| = 2\sum_{j=1}^{m} |X_{kj} - \overline{X}|.$$

The task of clustering is accompanied with two problems: the definition of an optimum amount of clusters and deriving of their centres. Input data for the clustering task are the values of parameters of researched systems. It is obvious, that the definition of an optimum amount of clusters is a prerogative of contributor. Let's assume, that the number of clusters $K$ is given and $k << m$, where $m$ - amount of plants. Let's receive the task

$$\sum_{i=1}^{K} \sum_{j=1}^{m_i} \left\| X_j - \overline{X}_i \right\| \to \min, \qquad (1)$$

where $\overline{X}_i$, $i = \overline{1,K}$ - average value in a cluster, $\left\| X_j - \overline{X}_i \right\|$ - distance between systems. A solution of the task (1) are the centers of clusters $\overline{X}_i$, which can contain among considered systems, that is a rather strict condition, and can be represented by any points of researched area.

To traditional methods of cluster analysis refer tree-like clustering, two-way joining, K-means clustering, method of dendrites, method correlative populations and method of full-spheres [Pluta, 1989]. Advantages of indicated methods is their simplicity, invariance of their engineering concerning a type of input data and used metrics. To shortages refer weak formalizing, that hampers application of computers, and also low exactitude, a corollary that is tentative estimation of a space structures of the factors and their selfdescriptiveness. One more method of a solution of the clustering task is the use of a self-organized Kohonen's map [Kohonen, 1988]. By a problem of use such neural networks is a choice of initial weight factors, continuous type of operation and effectiveness, which evaluation for today remains by a problem.

As an alternate method we offer to use genetic algorithm.

## Genetic Algorithms – Nonclassical Technique of Optimization Task Solution

The first variants of genetic algorithm and reviewing of its application aspects have appeared in [Fraser, 1962], [Fraser, 1968], [Bremermann, 1965], [Holland, 1969], [Holland, 1975]. The further researches have shown it effectiveness in a solution of engineering, economic ecological and other problems. Principal idea underlying a construction of genetic algorithm, is the use of ideas of a natural selection, selection and mutation. Its canonical variant contains such stages:

1. Definition of a general population of individuals $\Theta$, being potential solutions of the optimization task of fitness-function.
2. Realization of preliminary steps of algorithm consisting in quantifying of the elements $K$ of a sample population $\Xi$, where $k << |\Theta|$; a choice of a normalization mode for input data; a choice of recombination, mutation and inversion variant and also appropriate probabilities.
3. For each element $\theta_i \in \Xi$, $i = \overline{1,k}$ computed values of fitness-function $f_i = F(\theta_i)$.

4. With probabilities $P_i^k$, proportional by a values $f_i$, to select two individuals and to realize recombination, owing to which realization we shall receive two new individuals.
5. With probability $\frac{1}{2}$ to select one of obtained individuals and with probability $P^m$ to realize mutation.
6. The obtained individual is putted in a new population $\Xi^n$.
7. Repeat stage 3-6 $\left[\frac{k}{2}\right]$ times.
8. Rewrite the elements $\Xi^n$ in a population $\Xi$, deleting old individuals.

By criterion of a termination of genetic algorithm the following conditions can appear: convergence of the elements of a population $\Xi$ to one element; the maximum absolute deviation between elements of a population $\Xi$ will be less some positive number $\delta$; maximum absolute deviations between values of fitness-function will be less some small positive number $\varepsilon$.

Table 1

Values of researched factors

| | | | | |
|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | … | $x_{1n}$ |
| 2 | $x_{21}$ | $x_{22}$ | … | $x_{2n}$ |
| … | … | … | … | … |
| m | $x_{m1}$ | $x_{m2}$ | … | $x_{mn}$ |

## Shaping of Fitness-function for the Clustering Task

Input data of the clustering task are factors values (tab. 1). Beforehand, we execute their normalization, for example, under the formula $x_{ij}^{'} = \frac{x_{ij} - x_{j\min}}{x_{j\max} - x_{j\min}}$. Owing to such transformation of all factors values will lay in a single hypercube $[0,1]^n$. The fitness-function is realized by the following algorithm:

Step 1. A value of fitness-function to put equal to zero ( $F = 0.$ )

Step 2. To set a number of clusters $K$ and to specify a value $m$.

Step 3. To execute initialization of a membership matrix of the elements to clusters $T_k$.

Step 4. For all systems to execute the following steps. Let $n = 1$.

Step 5. To calculate a distance from $n$-th system up to centres all $K$ clusters, which are individuals from a sample population.

Step 6. Among all distances $d_j$, $j = \overline{1,K}$ to select minimum $d_q$ and to refer $n$-th system to $q$-th cluster. To bring an appropriate entry to a matrix $T_k$.

Step 7. $F = F + d_q$. $n = n + 1$.

Step 8. If steps 5-7 are carried out for all systems, the value of fitness-function $F$ is obtained, otherwise to pass to step 5.

It is obvious, that the algorithm of fitness-function deriving can be optimized. The possibility of improving is its absolute property. A variety of operations variants of genetic algorithm represent a set of exterior properties of the process of fitness-function deriving. The possibility of task solution of its optimization also assumes binary and

decimal representation of input data. And if in the first case in procedures of genetic algorithm dominating is the rectangular distribution, in second - at searching an optimum solution the preference is returned to values having normal distribution with mean value, conterminous with a centre of cluster. The definition of an optimum variance - one more task, which remains unsolved.

## Clustering of Ukraine Regions on Socio-economic Indications

For check of effectiveness of the offered clustering method regions of Ukraine were selected. The clustering should be realized, proceeding from values of socio-economic indexes. Such indexes are:

$X_1$ - total surplus value in account per one man (in actual values, UAH);

$X_2$ - territory (thousand sq. km);

$X_3$ - investment in a fixed capital per one man (in comparative values, UAH);

$X_4$ - direct foreign investments per one man (USD);

$X_5$ - employment of the population per 10 thousand;

$X_6$ - money incomes of the population per one man (UAH);

$X_7$ - credits submitted to subjects of managing per one man;

$X_8$ - amount of the obtained patents on the inventions on 10 thousand.

As classical techniques were selected tree-like clustering and K-means clustering. A priori two clusters are given. On K-means clustering techniques the following outcomes (tab. 2) are obtained. To the first cluster are referred Dnepropetrovsk, Donetsk, Zaporozhye, Nikolaev, Odessa, Poltava and Kharkov regions. According to tree-like clustering (fig. 1) to the first cluster the same regions, except for Donetsk region are referred, though it is close to elements of the first cluster.

The clustering was carried out also with use of evolutionary modelling. For a termination criterion of computing process was selected a maximum quantity of iterations equal 1000. For the same two clusters and eight factors the amount of variables (chromosome), for which the optimization of fitness-function was carried out, has made 16. Twenty elements have come in a sample population. As the fitness-function was polyextreme, the mutation probability has made 0,4. Such value has increased time of evaluations, but has considerably increased an exactitude of accounts for the score of beating out of function from local minima.

For process of evaluations monitoring in real-time mode the information about a value of fitness-function on each iteration (fig.2) was output; about an average distance to centre of clusters (fig. 3); values of centres of clusters (fig. 4). The value of fitness-function has decreased with $6 \cdot 10^9$ down to 11351587, and on the initial stages the diminution happened as hyperbola, and on last - linearly. The average distance to centre of clusters decreased linearly, with a constantly decreasing variance.

In an outcome of evaluations two centres of clusters are obtained. Coordinates first are $X_1 = 4553,\ X_2 = 0{,}01,\ X_3 = 915,\ X_4 = 99,\ X_5 = 4623,\ X_6 = 2554,\ X_7 = 791,\ X_8 = 1{,}34.$

Coordinates second are $X_1 = 2952,\ X_2 = 0{,}02,\ X_3 = 530,\ X_4 = 58,\ X_5 = 4288,\ X_6 = 1555,$ $X_7 = 297,\ \ X_8 = 0{,}59.$ To the first cluster concern Dnepropetrovsk, Donetsk, Nikolaev, Odessa, Poltava and Kharkov regions. The outcomes of three considered techniques are close, that testifies to an exactitude of evolutionary modelling. Its advantage is also indication of clusters centres and formalization of the computing process. As was indicated above, the offered process engineering can be advanced.

Table 2

Results of clustering for Ukraine regions

| Domain | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |

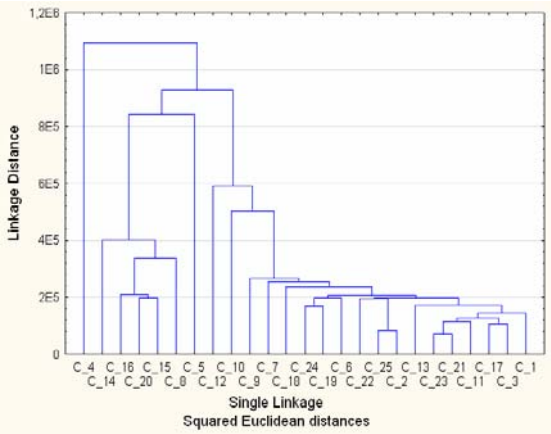


Fig.1 - Outcomes of tree-like clustering

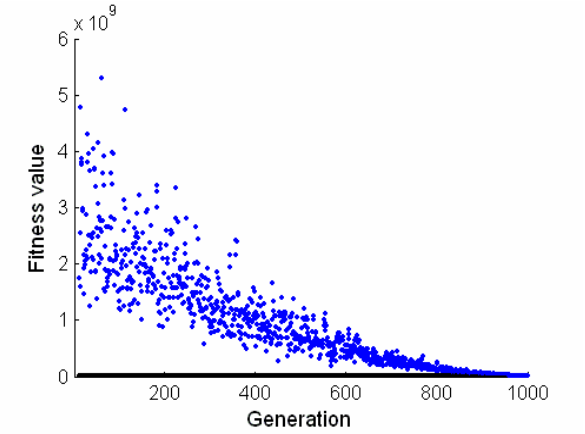


Fig.2 - Value of fitness-function

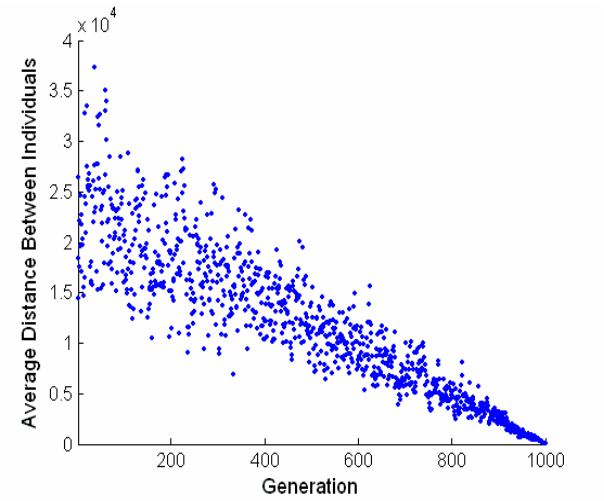


Fig.3 - Distance to centre of clusters

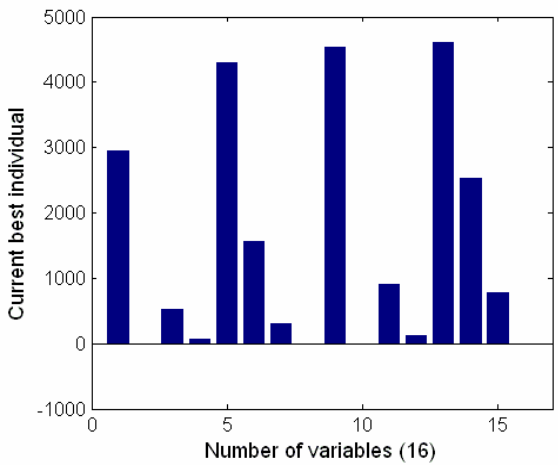


Fig.4 - Coordinate of clusters centres

## Inference

The offered method of evolutionary modelling based on using of genetic algorithm, effectively functions at handling arrays of large dimensionality, as in it are optimum combined purposeful searching and elements of chance directed on beating out to goal function from local minima. Any preliminary conditions for its use is not required. A principal condition of optimization of evaluations is right algoritmization of account of values of goal function. Multidirectedness of the process of improving of algorithm speed (for genetic algorithms especially is actual) and its exactitude (searching of a global minimum of fitness-function), and also it actuality testify to necessity of the task solution of the offered technique optimization.

## Bibliography

[ Mandel, 1988] I.D. Mandel. Cluster analysis. Moscow: Finance and statistics, 1988.

[ Gorban, 2002] A.N. Gorban, A.Yu. Zinovyev. Method of Elastic Maps and its Applications in Data Visualization and Data Modeling // Int. Journal of Computing Anticipatory Systems, CHAOS. - 2002. - Vol. 12. - P. 353-369.

[ Pluta, 1989] V. Pluta. The comparative many-dimensional analysis in econometric modelling. - Moscow: Finance and statistics, 1989.

[ Kohonen, 1988] T. Kohonen. Self-organization and associative memory. - New-York, 2d. Ed., Springer Verlag, 1988.

[ Fraser, 1962] A.S. Fraser. Simulation of genetic systems. J. Of Theor. Biol., vol. 2, pp. 329-346, 1962.

[ Fraser, 1968] A.S. Fraser. The evolution of purposive behavior. In Purposive Systems, H. Von Foerster, J.D. White, L.J. Peterson, and J.K. Russel, Eds. Washington, DC: Spartan Books, pp. 15-23, 1968.

[ Bremermann, 1965] H.J. Bremermann, M. Rogson, S. Salaff. Search by Evolution. In Biophysics and Cybernetic Systems. M. Maxfield, A. Callahan, and L. J. Fogel, Eds. Washington DC: Spartan Books, pp. 157-167, 1965.

[ Holland, 1969] J.H. Holland. Adaptive plans optimal for payoff-only environments. Proc. Of the 2nd Hawaii Int. Conf. On System Sciences, pp. 917-920, 1969.

[ Holland, 1975] J.H. Holland. Adaptation in Natural and Artificial Systems. Ann Arbor: Univ. Of Michigan Press, 1975.

[ Skurikhin, 1993] A.N. Skurikhin, A.J. Surkan. Identification of parallelism in neural networks by simulation with language J. Proc. of the intern. conf. On KPL, APL Quote Quad, Vol.24, No. I, pp.230-237, Toronto, Canada, August 1993.

## Author's Information

Vitaliy Snytyuk – Taras Shevchenko national university of Kyiv, postdoctorate researcher of cybernetics faculty; Glushkov's av. 2, building 6, Kiev, Ukraine; e-mail: svit@majar.com

# ALGORITHM OF CONSECUTIVE DEFINITION OF RANKING OF THE OBJECTS NEAREST TO THE SET CYCLIC RELATION BETWEEN OBJECTS

Grigoriy Gnatienko

*Abstract: The problem of a finding of ranging of the objects nearest to the cyclic relation set by the expert between objects is considered. Formalization of the problem arising at it is resulted. The algorithm based on a method of the consecutive analysis of variants and the analysis of conditions of acyclicity is offered.*

*Keywords: ranking, the binary relation, acyclicity, basic variant, consecutive analysis of variants*

*ACM Classification Keywords: K.3.2 Computer and Information Science Education.*

## Introduction

The problem of ordering of set of objects in degrees of display of some properties is one of the primary goals of expert reception of estimations [1]. The essence of a problem will consist in definition of the full order on set of compared objects under the set partial order.

Among problems of decision-making, the problem of linear ordering of objects is allocated with a plenty of concrete applications and a unconditional urgency of a theme. This problem traditionally is in the center of