Cluster Analysis and Genetic Algorithms

Petr Dostál[1], *Pavel Pokorný*[2]

Abstract

The paper deals with the cluster analysis and genetic algorithms and describes their basis. The application of genetic algorithms is focused on a cluster analysis as an optimization task. The case studies present the way of solution of two and three dimensional cluster analysis in MATLAB program with use of the Genetic Algorithm and Direct Search Toolbox. The way of its possible use in business is mentioned as well.

1. Introduction

The cluster analysis represents a group of methods whose aim is to classify the investigated objects into clusters. The founders of cluster analysis were Tryon, Ward and James. There have been suggested many new algorithms recently. Some methods represent a modification of classical methods of cluster analysis; other ones use advanced methods such as neural networks, e.g. represented by Kohonen self-organizing maps, or genetic algorithms.

The aim of **cluster analysis** is to classify the objects into clusters, especially in such a way that two objects of the same cluster are more similar than the objects of other clusters. The objects can be of various characteristics. It is possible to cluster animals, plants, text documents, economic data etc.

The **genetic algorithms** simulate the evolution of human population. During the calculation by means of genetic algorithms we use such operators as selection, crossover and mutation. The selection means the choice of the best individuals. The crossover represents the exchange of so-called chromosomes among single individuals of the population. The mutation means the modification of a part of a particular chromosome if a random change happens. These operators are presented in Tab.1.

Selection			Crossover		Mutation	
1010	>	0101	Parents	Offspring	Before	After
10	>	5	11|**00**	11*10*	1101	0111
			00|*10*	00**00**		

Tab.1 Genetic operators

Genetic algorithms operate such that the initial population of chromosomes is created first; this population is changed by means of genetic operators until the process is finished. One cycle of the reproduction process is called the epoch of evaluation of a population (generation) and it is represented by the above mentioned steps.

2. Cluster analysis as an optimization task

The aim is to divide the set of N existing objects into M groups. Each object is characterized by the values of K variables of a K-dimensional vector. The aim is to divide the objects into clusters so that the variability inside clusters is minimized.

[1] *Petr Dostál, MSc PhD, Associate Professor of Economy and Management, Department of Informatics, Faculty of Business and Management, Brno University of Technology, Kolejní 4, Brno, + 420 54114 3714, dostal@fbm.vutbr.cz*

[2] *Pavel Pokorný, MSc., Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2, Brno, + 420 54114 1111, xmpokor03@std.fbm.vutbr.cz*

Let $\{\mathbf{x}_i; i = 1,2,\ldots,N\}$ be a set of N objects. Let x_{il} denote the value of l-th variable for i-th object. Let us define for $i = 1,2,\ldots,N$ and $j = 1,2,\ldots,M$ the weights

$$w_{ij} = \begin{cases} 1 & \text{if the i - th object is a part of j - th cluster ,} \\ 0 & \text{otherwise .} \end{cases} \tag{1}$$

The matrix **W** = [w$_{ij}$] has the following properties

$$w_{ij} \in \{0;1\} \text{ and } \sum_{j=1}^{M} w_{ij} = 1 . \tag{2}$$

Let centroid of j-th cluster $\mathbf{c}_j = [c_{j1}, c_{j2}, \ldots, c_{jK}]$ be calculated in such a way, that each of its elements is a weighted arithmetic mean of relevant values, i.e.

$$c_{jl} = \frac{\sum_{i=1}^{N} w_{ij} x_{il}}{\sum_{i=1}^{N} w_{ij}} . \tag{3}$$

The inner stability of j-th cluster is defined as

$$S^{(j)}(W) = \sum_{i=1}^{N} w_{ij} \sum_{l=1}^{K} (x_{il} - c_{jl})^2 \tag{4}$$

and its total inner cluster variance as

$$S(W) = \sum_{j=1}^{M} S^{(j)} = \sum_{j=1}^{M} \sum_{i=1}^{N} w_{ij} \sum_{l=1}^{K} (x_{il} - c_{jl})^2 . \tag{5}$$

The distances between an object and a centroid can be calculated in this case by means of common Euclidean distances

$$D_E(\mathbf{x}_p, \mathbf{x}_q) = \sqrt{\sum_{l=1}^{K} (x_{pl} - x_{ql})^2} = \left\| \mathbf{x}_p - \mathbf{x}_q \right\| . \tag{6}$$

The aim is to find such matrix W* = [w^*_{ij}], that minimizes the sum of squares of distances in clusters from their centroids (over all M centroids), i.e.

$$S\left(W^*\right) = \min_W \{S(W)\} \tag{7}$$

3. Case study

The input data are represented by coordinates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$ that characterize the objects. It is possible to define any number of clusters. The fitness function represents the sum of squares of distances between the objects and centroids. The coordinates of centroids c_{j1}, c_{j2}, …, c_{jK} ($j=1,2,\ldots,M$) are changed. The calculation assigns the objects to their centroids. The whole process is repeated until the condition of optimum (minimum) of fitness function is reached. The process of optimization ensures that the defined coordinates x_{i1}, x_{i2}, …, x_{iK} ($i=1,2,\ldots,N$) of objects and assigned coordinates c_{j1}, c_{j2}, …, c_{jK} of clusters have the minimum distances. The fitness function is expressed by following formula

$$f_{\min} = \sum_{i=1}^{N} \min_{j\in(1,2,\ldots,M)} \left(\sqrt{\sum_{l=1}^{K}(x_{il}-c_{jl})^2}\right), \qquad (8)$$

where N is the number of objects, M the number of clusters and K dimension.

The calculation can be performed with help of *gatool* command or by creation of an M-file in MATLAB.

3.1 Two dimensional task solved by *gatool* command

Input data are represented by 14 objects with $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ coordinates (see Tab.2).

Object	Coordinates of objects	
Number	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
1	0,00	0,16
2	0,34	0,00
3	0,39	0,26
4	0,35	0,49
5	0,50	0,36
6	0,46	0,48
7	0,51	0,83
8	0,52	0,99
9	0,66	0,36
10	0,81	0,61
11	0,64	0,95
12	0,85	1,00
13	0,93	0,98
14	1,00	0,56

Tab.2 Coordinates of objects

The input data are in an MS Excel format file *Clust2.xls*. It is necessary to program the fitness function defined by formula (8). See File 1 named S*h2.m*.

```
function z=Sh2(x)
global LOCATION
z=0;
for i=1:size(LOCATION,1)
    for j=1:(size(x,2)/2)
        distances(j)=sqrt((LOCATION(i,1)-x(j))^2+(LOCATION(i,2)-
x(size(x,2)/2+j))^2);
    end
    min_distance=min(distances);
    z=z+min_distance;
end
```

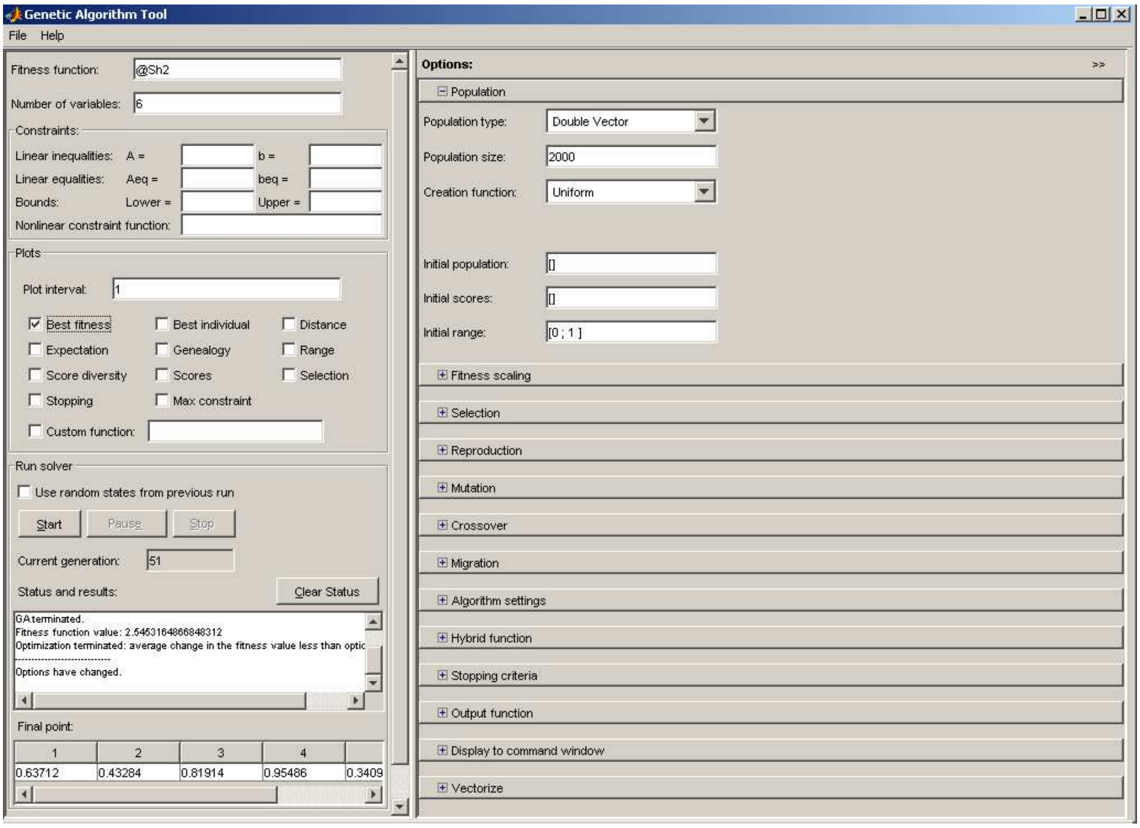
File 1 *Sh2.m*

Another File 2 called *Glob.m* changes the data to be global in the program.

```
global LOCATION;
LOCATION=(xlsread('Clust2','Location'))
```

File 2 *Glob.m*

When the command *Glob* and *gatool* are written in MATLAB, only two parameters are necessary to be filled in, i.e. the *Fitness function* in the form *@Sh2* and the number of clusters multiplied by two (three clusters correspond to *Number of variables* of 6). It is suitable to set up the *Population size* to be *2000*. The calculation starts by pressing the button *Start*.When the calculation is terminated, the final results are displayed in the area *Status and results* and *Final point*. See the Fig.1.



Fitness function: @Sh2

Number of variables: 6

Plot interval: 1

Current generation: 51

Status and results:

```
GAterminated.
Fitness function value: 2.5453164866848312
Optimization terminated: average change in the fitness value less than optic
---------------------------
Options have changed.
```

Final point:

1	2	3	4	
0.63712	0.43284	0.81914	0.95486	0.3409

Population type: Double Vector

Population size: 2000

Creation function: Uniform

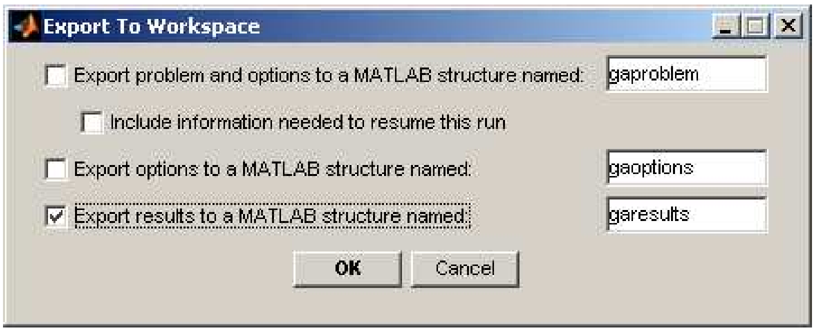
Initial population: []

Initial scores: []

Initial range: [0 ; 1]

Fig.1 Genetic Algorithm Tool

These areas inform us about the reason of termination of the calculation, values of the fitness function and about the coordinates of clusters. The coordinates are sorted for x_1 at first and then for x_2. The coordinates can be recorded into files by the menu *Export to Workplace*. See the Fig.2.



□ Export problem and options to a MATLAB structure named: gaproblem

□ Include information needed to resume this run

□ Export options to a MATLAB structure named: gaoptions

☑ Export results to a MATLAB structure named: garesults

Fig.2 Export to Workspace

The report of garesults variable is as follows
garesults = *x: [0.9169 0.4330 0.6357 0.5841 0.3384 0.9550]*.

If the option *Best fitness* is chosen, the process of calculation can be displayed. See Fig.3.

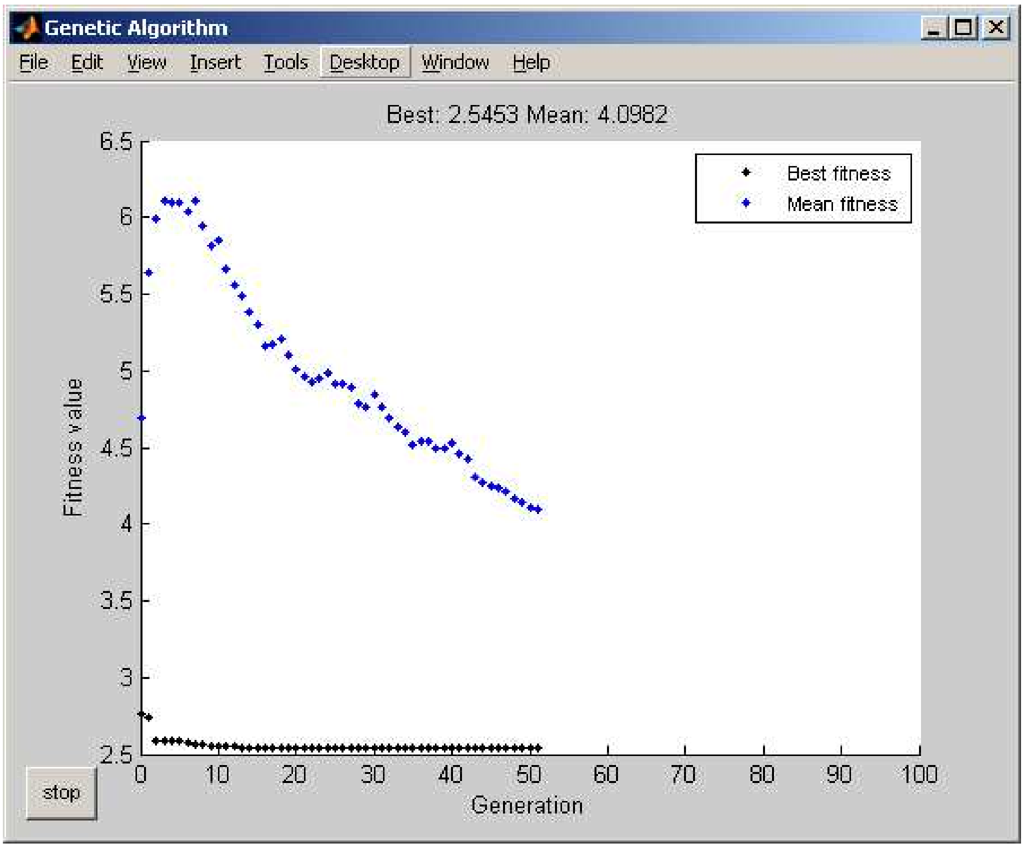


Fig.3 Best fitness diagram

3.2 Three dimensional task solved by program

The input data are represented by 14 objects with $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ coordinates. See the Tab.3.

Object	Coordinates of objects		
Number	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$
1	0,00	0,16	0,16
2	0,34	0,00	0,00
3	0,39	0,26	0,26
4	0,35	0,49	0,49
5	0,50	0,36	0,36
6	0,46	0,48	0,48
7	0,51	0,83	0,83
8	0,52	0,99	0,52
9	0,66	0,36	0,66
10	0,81	0,61	0,81
11	0,64	0,95	0,64
12	0,85	1,00	0,85
13	0,93	0,98	0,93
14	1,00	0,56	1,00

Tab.3 Coordinates of objects

The input data are in an MS Excel format file *Clust3.xls*. It is convenient to program the task. See the File 3 called *Sh3.m*. The File 4 *Draw3.m* draws the graph and File 5 *Cluster3.m* calculates the distances.

```
function Sh3
global LOCATION
num=input('Number of clusters:');
num=3*num;
PopSize=input('Population size:');
FitnessFcn = @Cluster3;
numberOfVariables = num;
LOCATION=(xlsread('Clust3','Location'))
my_plot = @(Options,state,flag)
Draw3(Options,state,flag,LOCATION,num);
Options =
gaoptimset('PlotFcns',my_plot,'PopInitRange',[0;1],'PopulationSize',Po
pSize);
[x,fval] = ga(FitnessFcn,numberOfVariables,Options);
assign=zeros(1,size(LOCATION,1));
for i=1:size(LOCATION,1)
    distances=zeros(num/3,1);
    for j=1:(size(x,2)/3)
        distances(j)=sqrt((LOCATION(i,1)-x(j))^2+(LOCATION(i,2)-
x(size(x,2)/3+j))^2+(LOCATION(i,3)-x(2*size(x,2)/3+j))^2);
    end
    [min_distance,assign(i)]=min(distances);
end
assign
fval
xyz=zeros(num/3,3);
for i=1:(num/3)
    xyz(i,1)=x(1,i);
    xyz(i,2)=x(1,num/3+i);
    xyz(i,3)=x(1,2*num/3+i);
end
xyz
```

File 3 *Sh3.m*

```
function state = Draw3(Options,state,flag,LOCATION,num)
[unused,i] = min(state.Score);
x=state.Population(i,:);
for i=1:size(LOCATION,1)
    for j=1:(size(x,2)/3)
        distances(j)=sqrt((LOCATION(i,1)-x(j))^2+(LOCATION(i,2)-
x(size(x,2)/3+j))^2+(LOCATION(i,3)-x(2*size(x,2)/3+j))^2);
    end
    [min_distance,assign(i)]=min(distances);
end
for i=1:size(LOCATION,1)
plot3(LOCATION(i,1),LOCATION(i,2),LOCATION(i,3),'sr','MarkerFaceColor'
,[3*(assign(i))/num,3*(assign(i))/num,3*(assign(i))/num],'MarkerSize',
10);
xlabel('x');ylabel('y');zlabel('z');
grid on;
hold on;
end
plot3(x(1:size(x,2)/3),x((size(x,2)/3+1):2*size(x,2)/3),x(2*size(x,2)/
3+1:size(x,2)),'sr','MarkerFaceColor','b','MarkerSize',10);
hold off;
```

File 4 *Draw3.m*

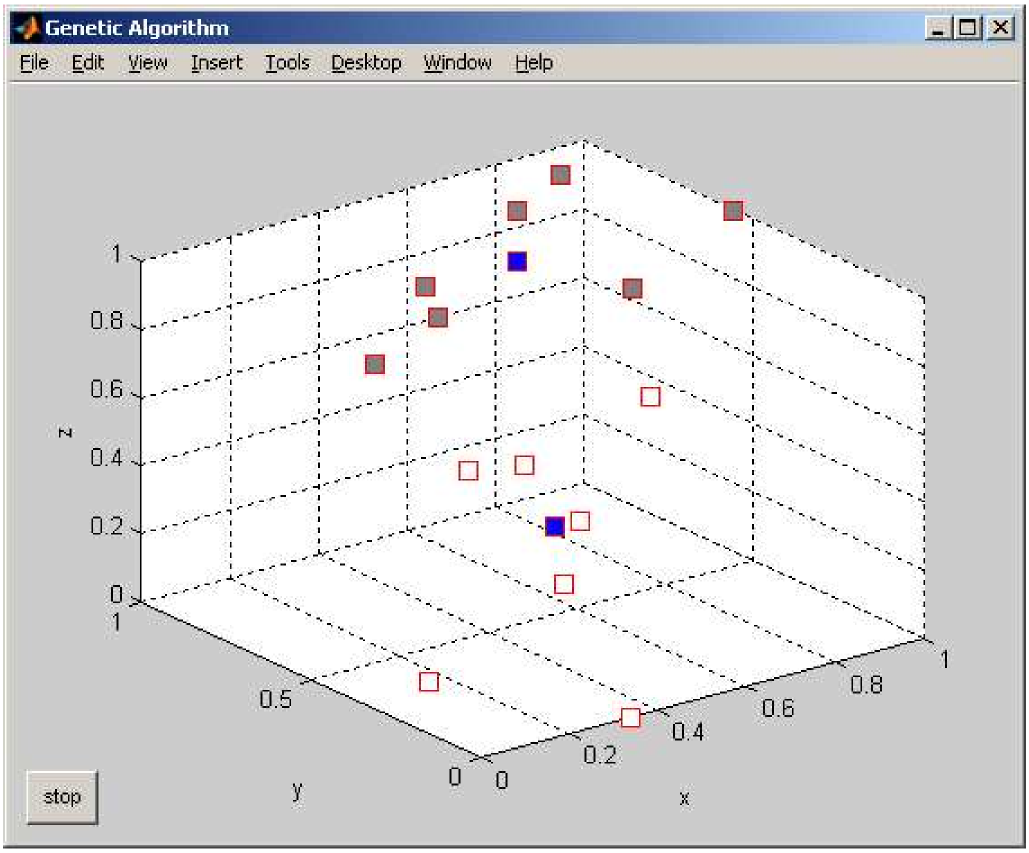
```
function z=Cluster3(x)
global LOCATION
z=0;
for i=1:size(LOCATION,1)
    for j=1:(size(x,2)/3)
        distances(j)=sqrt((LOCATION(i,1)-x(j))^2+(LOCATION(i,2)-
x(size(x,2)/3+j))^2+(LOCATION(i,3)-x(2*size(x,2)/3+j))^2);
    end
    min_distance=min(distances);
    z=z+min_distance;
end
```

File 5 *Cluster3.m*

The program enables us to set up the number of required clusters and the population size. The higher number of individuals the more precise solution but the higher duration of the calculation. Futher, the program sets up the options for optimization and the optimization command *ga* is called. The program involves the calculation of fitness function and it fills the variables with data that inform us about the coordinates of centroids and the assignment of objects to clusters and displays them.

The two and three dimensional tasks can be drawn. The file *Draw3.m* performs this process. The program uses the optimized values and the command *plot3* makes the drawing. The graph distinguishes the assignment of objects to the clusters by different signs.

The program is started by command *Sh3* in MATLAB. Then it is necessary to set up the requested number of clusters, e.g. *Number of clusters* to be *2* and Population size to be *1000*. During the calculation the dynamical three-dimensional graph is presented. See Graph 1.

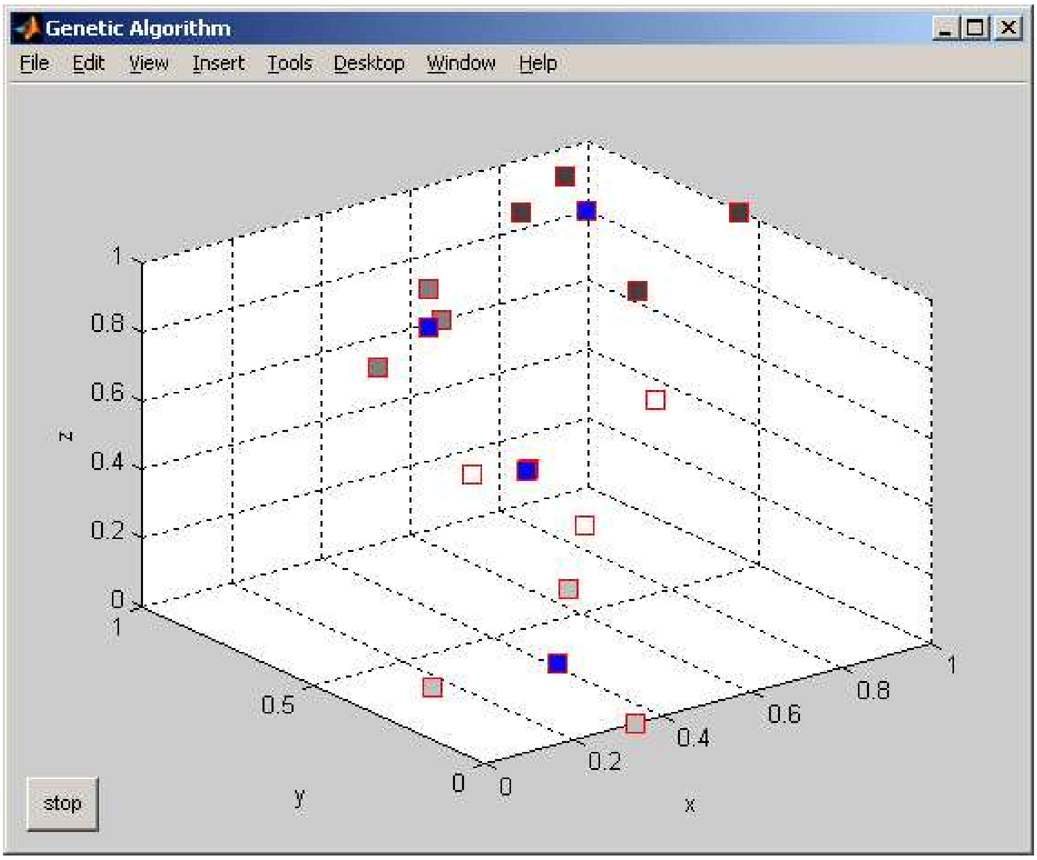


Graph 1 Three-dimensional graph – two clusters

When the calculation is terminated, the input parameters and results of calculation are displayed on the screen. The results are presented by coordinates of clusters and assignment of objects to clusters. The three-dimensional graph presents these fact.

```
Number of clusters: 2
Population size: 1000
Optimization terminated: average change in the fitness value less than options. TolFun.
assign =    2     2     2     2     2     2     1     1     2     1     1     1     1     1
fval =    3.9863
xyz =
 0.7510    0.8726    0.7944
 0.4319    0.3471    0.3683>>
```

The following case represents the same task if four clusters are required, i.e. the *Number of clusters* is *4*. The results are presented by coordinates of clusters and assignment of objects to clusters. The three-dimensional graph presents these facts. See Graph 2.



Graph 2 Three-dimensional graph – four clusters

```
Number of clusters: 4
Population size: 1000
Optimization terminated: average change in the fitness value less than options. TolFun.
assign = 3     3     3     4     4     4     2     2     4     1     2     1     1     1
fval =    2.6107
```

```
xyz =
   0.8937    0.8656    0.9031
   0.5983    0.9413    0.6327
   0.2646    0.1319    0.1377
   0.4552    0.4683    0.4783>>
```

5. Conclusion

The cluster analysis has a wide range of use in various branches. One of the branches is economy and business. We can mention for example the search of best location of a market, bank or firm. The term cluster in business (according to the definition by Porter) means the geographical collection of mutually linked firms, specialized suppliers, providers of services, firms of similar branches and associated institutions, such as universities, agencies and business associations of different directions which contest, but also cooperate. The advantage of the use of genetic algorithms is their applicability in various types of optimization problems with a high speed of calculation and found solution very close to the optimal one. The article describes the way how to perform it at best in entrepreneurial and business area.

Literature

[1] DAVIS, L. *Handbook of Genetic Algorithms*, Int. Thomson Com. Press, USA, 1991, 385 P., ISBN 1-850-32825-0.

[2] DOSTÁL, P. *Moderní metody ekonomických analýz – Finanční kybernetika*, UTB Zlín, 2002, 110 p., ISBN 80-7318-075-8.

[3] DOSTÁL, P., RAIS, K. Genetické algoritmy a jejich využití v modelování, In *Odborná konference Firemní management v praxi úspěšných*, EPI s.r.o, 2002, pp. 41-44, ISBN 80-7314-004-7.

[4] DOSTÁL, P., RAIS, K. *Operační a systémová analýza II*. VUT – FP - Brno 2005, 160 p., ISBN 80-214-2803-1.

[5] DOSTÁL, P., RAIS, K., SOJKA, Z. *Pokročilé metody manažerského rozhodování*, Grada, 2005,168 p, ISBN 80-247-1338-1.

[6] DOSTÁL P. Využití metody klastrování v problematice klastrů firem, Zlín 2007, *In Finance a účetnictví ve vědě, výuce a praxi*, Conference, pp.51, 5p., ISBN 80-7318-536-7.

[7] ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V. *Shluková analýza dat*, Professional publishing, 2007, 196 p., ISBN 80-86946-26-9.