N. R. Kondratenko, Cand. Sc. (Eng.), Assist. Prof.; O. O. Manayeva

FUZZY CLUSTERING OF THE INTRNET PROVIDER SUBSCRIBERS

The paper proposes a genetic algorithm of fuzzy clustering on the basis o heterogeneous chromosomes with prior initialization of the coordinates of the cluster centers. The algorithm was tested for convergence and its functioning was illustrated by a computer experiment.

Key words: *fuzzy clustering, provider, genetic algorithm, heterogeneous chromosome, test functions, weighted average deviation, membership degrees.*

Internet is a global informational network linking a large number of regional networks and at the same time millions of computers at all ends of the planet in order to exchange data and to access informational and technological resources [1]. Internet service providers (ISP) are engaged in rendering the Internet access services and other relevant services. Such organizations possess large amounts of information about their users. This information should be systemized in a definite way as well as structured, generalized etc. These tasks are closely related to the task of cluster analysis [2].

There is a large number of clustering methods that can be classified as crisp and fuzzy. Crisp clustering methods divide an initial set of objects X into several non-intersecting subsets. In this case any object from X belongs to one cluster only. Fuzzy clustering methods allow the same object to belong to several clusters (or even to all clusters) simultaneously but with different degree of membership. The only difference is that in the case of fuzzy division the degree of object membership to a cluster takes values from the interval of [0, 1] and in the case of crisp division – from a two-element set {0, 1}. In many cases fuzzy clustering is more "natural" than a crisp one, e.g. for objects located on the border of clusters [3, 4].

For solving the set problem we suggest an approach based on fuzzy division of the space of objects into clusters. For the problem of classifying ISP subscribers into groups it is the approach that has practical importance. This is connected with one of the main requirements to the organization tariff plans – their flexibility. To satisfy this requirement, when such division is constructed some uncertainty should be assumed as to a subscriber membership to a certain group.

Let each subscriber be an object characterized by definite values of the given indicators (access speed, the volume of the input and output traffic used etc.). Accordingly, they could be represented as points in the multidimensional space. Practically, such an understanding of similarity means that subscribers are considered to be the more similar, the smaller the difference between similar parameters by which they are described [4].

Under such conditions it is expedient to solve this problem on the scale of a single ISP. In our case the subscribers of the above-mentioned provider, represented by a set of parameters, are considered as objects. The task is to divide the set of subscribers, presented in this way, into homogeneous fuzzy sets.

The goal of the presented research is mathematical modeling of the behavior of subscribers in relation to the telecommunication service provider and their division into homogeneous groups with the possibility of further analysis of the obtained results.

Problem statement

Let us set the problem of dividing the set of IP providers into fuzzy inhomogeneous subsets in accordance with the given set of indicators. Each subscriber can belong to a certain cluster with a certain membership degree in the range from 0 to 1. It is necessary to determine all degrees μ_{ij} of subscriber *j* membership to cluster *i* as well as all locations of the centers of clusters $c_i, i = \overline{1, m}$.

For solving this problem we propose a genetic algorithm that performs fuzzy clustering of IP subscribers in accordance with the above indicators and investigate it for convergence using a number of test functions.

Mathematical model

Let there be a set of users $I=\{I_1, I_2, \ldots, I_n\}$ of a certain ISP. Each of n subscribers is characterized by a number of attributes (dimensions) $X_i=\{x_1, x_2, \ldots, x_p\}$, among which are data transfer speed and the volumes of the ingoing and outgoing traffic within the given time period. The task of fuzzy clustering is as follows: on the basis of data contained in set I, to divide the set of subscribers I into $1<m<n$ clusters, i.e. to determine degrees μ_{ij} of each subscriber membership to each of m clusters that are given by centers $c_i, i=\overline{1,m}$.

Values μ_{ij} are restricted by the following constraints:

1. $0 \le \mu_{ij} \le 1$;
2. $\sum_{i=1}^{m} \mu_{ij} = 1$ for all j. (1)

To evaluate the quality of fuzzy clustering a weighted average deviation of the points-subscri from the centers of clusters are used:

$$E = \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij}^{m} \left\| x_j - c_i \right\|^2,$$

where $m \ge 1$ is an exponential weight that determines fuzziness and scattering of the clusters.

It is necessary to find such location of the cluster centers c_i and the value $\{\mu_{ij}\}$, for which the terion value would be minimal and at the same time the conditions of constraint (1) would be observed [5]. For solving this problem we propose a genetic optimization algorithm. The classical genetic algorithm is an inertial process, at each iteration of which the population is subjected to the operations of selection, crossbreeding and mutation. By stopping the iteration process at a definite moment and choosing the best individual in the population an acceptable problem solution can be obtained.

We suggest a heterogeneous chromosome as a formalized representation of the solution. A chromosome set consists from two qualitatively different parts. The first of them is a rectangular matrix containing degrees of membership of points-subscribers to corresponding clusters. Its elements determine the degree of connection of the corresponding subscriber with a definite cluster. The second defines the coordinates of the cluster centers in the space of attributes. For each of such chromosomes a certain value of the target function is calculated.

At the preparatory stage first initialization of the centers of clusters is performed. It is based on the geometrical representation of subscribers as points in the space. For this each axis is divided into $N = 2 + E(3{,}322 \lg n)$ intervals. Thus, the entire space of attributes is divided into N^d cubes having equal volumes. where d is a number of attributes (dimensions). Geometrical centers of the cubes, inside which the largest number of points get, are taken as initial centers of the clusters.

Then the initial population is subjected to the operations of crossbreeding and mutation in the following sequence:

– one-point crossbreeding: two chromosomes are cut in a randomly chosen point and exchange the obtained parts. The operation is performed with all components of the chromosome according to the same scheme;

– two-point crossbreeding: chromosomes are regarded as cycles formed by connecting the ends of a linear chromosome. To replace the segment of one cycle by a segment of another cycle, two points of the cut are chosen.

– uniform crossbreeding: each gene of the descendant is created by copying the corresponding gene from one or the other of the parents in accordance with a randomly generated mask. When in the position of the mask stands 1, the gene is copied from the first parent chromosome, if 0 – from the second. The process is repeated with the parents, who were exchanged, to create a second descendant. For each pair of parents a new mask is randomly generated;

– mutation: generating new membership degrees for a single, randomly chosen point as well as random change of the position of each cluster center according to one dimension in the space of

attributes.

Individuals completely identical in their chromosome sets are deleted from the population and replaced by mutants, formed according to the above scheme.

As a result of such actions, we obtain 7n genetically unique descendants, for each of which the target function is calculated and then the mechanism of natural selection, based on the strategy of elitism, is realized within a population. In this case, solutions with the lowest value of the objective function are guaranteed to pass to the population of the next generation, which contributes to faster convergence of the genetic algorithm. On the whole the most prospective solution variants are processed by means of crossbreeding while mutations implement the mechanism of the optimization process going beyond the local minimums. As a result, there is high probability that the algorithm will converge to a solution that is maximally close to the optimum.

Computer experiment

Let us solve the problem of clustering the subscribers of a telecommunication service provider using three attributes: data transfer speed and volumes of ingoing and outgoing traffics within a fixed period of time. The volume of the sample studied is 100 users.

The results of clustering the subscribers according to the above indicators are presented in tables 1 and 2.

Table 1

The results of clustering: membership degrees

Subscriber code	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0	0,004049983	0,058683567	0,268992839	0,668273611
1	0,233707476	0,039341806	0,080751559	0,646199159
2	0,083366999	0,115781932	0,751763956	0,049087113
3	0,189849775	0,73390408	0,010152989	0,066093155
4	0,997684019	0,000579759	0,001224275	0,000511947
5	0,010728999	0,784860715	0,19198237	0,012427916
…	…			
99	0,059285226	0,903040624	0,002809315	0,034864835

Table 2

The results of clustering: location of cluster centers

Measuring the space of attributes	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Data transfer speed	3208,12	430,24	484,15	819,71
Ingoing traffic volume	4133,44	1613,21	887,96	701,37
Outgoing traffic volume	514,09	2503,17	105,46	88,01

From the location of cluster centers we see that subscribers having the highest access speed are in the first cluster, the second cluster represents a small segment of users with the volume of the ingoing traffic within the day under consideration approaching the outgoing traffic or exceeding it. The third cluster includes the subscribers for whom the data transfer speed is, mainly, low and the ingoing traffic exceeds the outgoing traffic. The fourth cluster differs from the third one by a higher access speed while the ingoing/outgoing traffic ratio is approximately the same as in the third cluster.

The genetic algorithm investigation for convergence

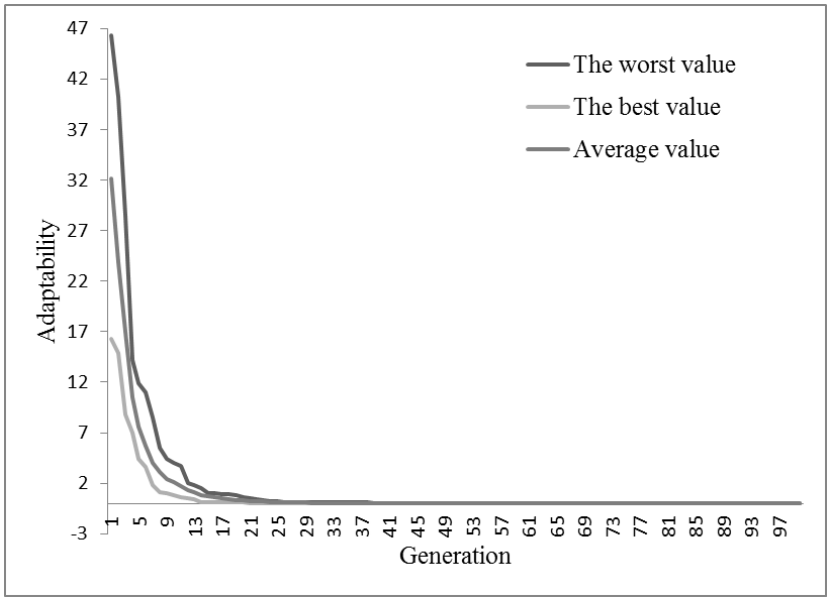
Let us evaluate the effectiveness of the proposed genetic algorithm using test functions. Its optimization capabilities are investigated for the number of variables n=10 and n=100 for limit number of generations – 100 and 1000 correspondingly. For this we perform a series of ten experiments for each of the test functions given below.

1. Spherical function (first function of de Jong) - a continuous convex unimodal test function, it is considered the easiest to optimize.

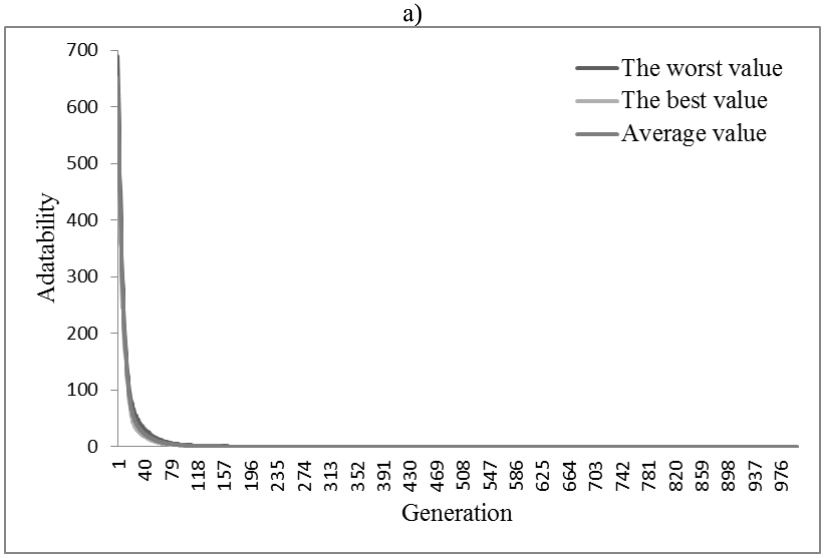
$$f_1(\mathbf{x}) = \sum_{i=1}^{n} x_i^2, \tag{3}$$

where $-5{,}12 \le x_i \le 5{,}12$, $i=1\ldots n$. It has one global minimum equal to one in the point where $x_i=0$, $i=1\ldots n$.

Changes in the best, worst and average adaptability of the best individual in the population for this function in a series of ten experiments are shown in Fig. 1.



a)



b)

Fig. 1 Changes in the best, worst and average adaptability of the best individual in the population for a spherical function for n=10 (a) и n=100 (b)

The proposed algorithm was tested in a similar way on other common test functions. Table 3 shows best, worst and average values of adaptability of the best individual in the last generation of the population obtained for different test functions.

Table 3

Results of the algorithm investigation for convergence

Test function	The value of adaptability of the best individual in the population					
	Best		Worst		Average	
	n=10	*n*=100	*n*=10	*n*=100	*n*=10	*n*=100
Spherical	0,0005884	0,000238	0,002541	0,000478	0,001472	0,000324
Step	0	0	0	0	0	0
Rastrigin's	0,088847	0,031292	0,775811	0,053048	0,335579	0,044519
Schwefel's	0,417962	0,20015783	3,743808	0,468031	0,914933	0,338687
Griewank's	0,3411336	0,0493624	1,001547	0,141237	0,65774	0,092017

From table 3 it s evident that average error of finding the global extremum for any of the test functions do not exceed the order of the first decimal place. This confirms the high efficiency of the proposed genetic algorithm, including the case with a large number of variables.

Conclusions

This paper proposes an approach for solving the problem of fuzzy clustering of ISP subscribers and develops a genetic algorithm with the application of heterogeneous chromosomes.

Study of the proposed genetic algorithm for convergence has shown that it is a powerful optimization algorithm and, therefore, can be used in the clustering task which is characterized by the presence of a large number of parameters and, as a rule, of a significant number of local extrema.

With the help of the proposed algorithm clustering of ISP subscribers was carried out according to the indicators that characterize their use of services of the organization. As a result, a set of users was divided into compact groups, between which there are significant differences according to these indicators. The conducted cluster analysis led to the conclusion that the application of fuzziness, especially in the problems of cluster analysis, makes it possible to work and get results in the conditions of a considerable noisiness of data. Therefore, further research in this direction is promising.

REFERENCES

1. Олифер В. Г. Компьютерные сети: принципы, технологии, протоколы / В. Г. Олифер, Н. А. Олифер. – СПб.: Питер, 2006. – 958 с.

2. Муссель К. Предоставление и биллинг услуг связи. Системная интеграция / К. Муссель. – М.: Эко-Трендз, 2003. – 319 с.

3. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл; Пер. с англ. Е.З. Демиденко. – М.: Статистика, 1977. – 128 с.

4. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Статистика, 1988. – 176 с.

5. Зайченко Ю. П. Нечеткие модели и методы в интеллектуальных системах / Ю. П. Зайченко. – К.: Издательский дом «Слово», 2008. – 344 с.

Kondratenko Nataliya – Cand. Sc(Eng)., Assoc. Prof., Prof. of the Information Protection Department.

Manayeva Olga – Master's course student.
Vinnytsia National Technical University.