Proceedings of the Federated Conference on Computer Science and Information Systems pp. 137–140
DOI: 10.15439/2016F232
ACSIS, Vol. 8. ISSN 2300-5963

Hybrid Fuzzy-Genetic Algorithm Applied to Clustering Problem

Krzysztof Pytel
Faculty of Physics and Applied Informatics
University of Lodz, Poland
Email: kpytel@uni.lodz.pl

***Abstract*—Clustering is a task of grouping a set of objects in such a way that objects in the same group (called a cluster) are similar to each other and dissimilar to objects belonging to other groups (clusters). The article presents the idea of the hybrid Fuzzy Logic-Genetic Algorithm (FLGA) system that supports solving clustering problems. The Genetic Algorithm (GA) realizes the process of multi-objective optimization - it aims at optimal distribution of clusters and correctly assigns each object to a cluster. The Fuzzy Logic Controller (FLC) is used for setting the number of clusters. The FLC uses additional fuzzy logic criteria obtained from experts. Experiments show that the proposed algorithm is an efficient tool for the clustering problem. The algorithm can be also used for solving similar optimization problems.**

I. Introduction

CLUSTERING (or cluster analysis) is the problem of classifying an unlabeled set of objects into groups of similar objects, called clusters. Each cluster consists of objects that are similar to one another and dissimilar to objects belonging to other clusters. Clustering is often based on similarity or dissimilarity measure. This measure is problem-dependent. The similarity or dissimilarity between the objects is usually computed, based on the distance between objects. The most popular distance measure is the Euclidean distance, but other measures, such as the Manhattan or Minkowski distances, could also be used. Clustering can be formally considered as a kind of NP-hard grouping problem. The main difficulty in a clustering problem is that it is an unsupervised task, so usually we do not know the number and the distribution of clusters, the shape of clusters or association of objects to clusters. Clustering is not one specific algorithm, but a general task. A classical clustering method is the k-means [1]. A k-means algorithm is sensitive to the choice of an initial partition, and this step can have a significant impact on the performance of algorithm. The algorithm could converge to a local minimum. A k-means algorithm needs determining a number of clusters in all data sets (parameter k), an inappropriate parameter k may yield poor results.

The Genetic Algorithm is an optimization method that simulates the process of natural evolution. They usually search for approximate solutions for composite optimization problems in a large search space. A characteristic feature of genetic algorithms is that in the process of evolution they do not use the knowledge specific for a given problem, except for the fitness function assigned to all individuals. Genetic algorithms can be used for solving wide range of optimization problems.

Clustering can be formulated as a multi-objective optimization problem. It consists of three different objective functions: looking for an appropriate number of centroids, optimal placeing of centroids in a given area, and assigning each object to a cluster, represented by the centroid, to minimize the distance between the objects in the same cluster.

II. Problem formulation

A clustering problem is one of practical examples of multi-objective optimization problems. The clustering problem can be defined as: let us consider a data set $X = \{x_1, x_2, ..., x_N\}$ be a set of N objects. Each object $x_i = [x_{i1}, x_{i2}..., x_{id}]$ has d dimensions. The goal of the clustering algorithm is to find k clusters $C_1, C_2, ..., C_k$ so that objects belonging to the same cluster are more similar to each other than to the objects belonging to other clusters. The Euclidean distance between each pair of objects is typically used as a similarity measure. Clustering must comply assigning constrains: each cluster must contain at least one object, and the object can be assigned to one cluster only.

$$\begin{cases} min \quad f_1(x_1, x_2, ..., x_N) \\ opt \quad f_2(n) \\ min \quad f_3(n, x_1, x_2, ..., x_N) \\ \text{subject to: } assigning\ constraints \end{cases} \tag{1}$$

where: f_1 - is the function representing the distance between the objects assigned to the same cluster,

f_2 - is the function representing the number of clusters,

f_3 - is the function assigning an object to a cluster,

$(x_1, , x_N)$ - are objects in d dimentional space,

$1 \leq n \leq n_{max}$ - is the number of clusters.

More information about cluster analysis can be found in publications [1][2].

III. Proposed Fuzzy Logic-Genetic Algorithm

The clustering problem is discussed in literature, eg. [8][9]. There are a lot of publications concerning different methods of solving this problem, for example k-means or genetic algorithms, but these methods work well if a number of clusters is known before running an algorithm. The proposed Fuzzy Logic-Genetic Algorithm (FLGA) consists of two modules: the Genetic Algorithm (GA) and the Fuzzy Logic Controller (FLC). The GA seeks for an optimal placement of centroids

and assigns objects to clusters. It delivers information concerning a current state of optimization to the FLC after a fixed number of generations. The FLC looks for an optimal number of clusters. The FLC is engaged between generations of the GA in fixed intervals of generations. The FLC modifies the number of clusters in dependence on information delivered from the GA. If the FLC changes the number of clusters, it sends information to the GA and modifies the individuals' genes to comply with the new constraints. The system is able to find the optimal number of clusters simultaneously with an optimization executed by the GA, so we do not need to know the number of clusters before running the algorithm.

In the proposed FLGA the individuals' genes (potential solutions) are encoded by the means of a composite data structure consisting of:

- the table describing the position of centroids by geographical coordinates - coordinates (x, y) of centroids are represented by real numbers, the number of genes in a table is equal to the number of clusters,
- the table describing an association of objects to clusters - association of objects to clusters are coded by integer numbers, eg. number i in position k means the association of object k to cluster i, the number of genes in a table is equal to the number of objects. This method of gene coding ensures assigning every object to one cluster only.

The value of the fitness function of an individual is calculated as a total distance between objects and centroids. Because clustering is a problem of minimization of the distance, we introduced additional constant C to transform the increasing fitness function into the decreasing function optimized. The value of constant C was chosen experimentally for every solved task. In our experiment different types of crossing-over of chromosomes were used. The standard one-point crossing is used when the number of clusters does not change. We introduce two new crossing-over operators, used when the number of clusters changes:

- CR1 - is used when the descendant's length of genotype is greater than the genotype's length of its parents. The descendant's genotype is obtained by copying all the genes from its first parent and lacking the genes of its second parent beginning from the end of the genotype.
- CR2 - is used when the descendant's length of genotype is smaller than the genotype's length of its parents. The number of the genes copied from every parent is diminished in proportion to the genotype's length to obtain the required length of the descendant's genotype.

Genetic Algorithms can be used for solving multi-objective optimization problems. They can be used in hybrid systems with other methods inspired by observation of nature. For example, the Fuzzy Logic Controller can effectively direct the process of evolution in the Genetic Algorithm toward a desired area of the search space [4][5].

A basic task of the FLC in the proposed system is evaluation of the solutions found till now. The FLC uses experts' knowledge and the knowledge collected by the GA and transferred to the FLC in fixed intervals of GA's generations. The FLC optimizes the number of clusters and is engaged in fixed intervals of GA's generations - it makes decisions about the diminution or the enlargement of the clusters' number. The FLC calculates the change of the clusters' number, based on two parameters:

- the relation of the distance between the centroids to the distance between the centroids and the objects assigned to these clusters,

$$rd_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d(i,j)}{\sum_{i=1}^{n}\sum_{k=1}^{m} d(i,k)} \tag{2}$$

  where:
  - rd_1 - the relation of the distance between the centroids to the distance between the centroids and the objects assigned to these clusters,
  - $\sum_{i=1}^{n}\sum_{j=1}^{n} d(i,j)$ - the distance between the centroids,
  - $\sum_{i=1}^{n}\sum_{k=1}^{m} d(i,k)$ - the distance between the centroids and the objects assigned to these clusters,
  - n - the number of clusters (centroids),
  - m - the number of objects.

  This parameter lets us determine a suitable number of clusters. The low value of this parameter can be due to a small number of clusters with relation to the number of objects. The large value of this parameter can be due to a big number of clusters with relation to the number of objects.
- the relation of the distance between the centroids and the objects assigned to these clusters to the distance between the centroids and the objects not assigned to these clusters,

$$rd_2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m1} d(i,k)}{\sum_{i=1}^{n}\sum_{l=1}^{m2} d(i,l)} \tag{3}$$

  where:
  - rd_2 - the relation of the distance between the centroids and the objects assigned to these clusters to the distance between the centroids and the objects not assigned to these clusters,
  - $\sum_{i=1}^{n}\sum_{k=1}^{m1} d(i,k)$ - the distance between centroids and objects assigned to these clusters,
  - $\sum_{i=1}^{n}\sum_{l=1}^{m2} d(i,l)$ - the distance between the centroids and the objects not assigned to these clusters,
  - n - the number of clusters (centroids),
  - $m1$ - the number of the objects assigned to these clusters,
  - $m2$ - the number of the objects not assigned to these clusters,

  This parameter lets us determine a suitable density of objects in the clusters. The low value of this parameter can be due to a big number of clusters with relation to the number of objects. The large value of this parameter can be due to a small number of clusters with relation to the number of objects. The value of this parameter is 0

TABLE I
FUZZY VALUES OF CLUSTERS' NUMBER CHANGE

		rd_1		
		Small	**OK**	**Large**
rd_2	**Large**	Enlarge	Enlarge	Not change
	OK	Enlarge	Not change	Diminish
	Small	Not change	Diminish	Diminish

when every object belongs to its own cluster, the value 1 is when all objects belongs to one cluster only.

The knowledge of experts is expressed by the following rules:

- enlarge the number of clusters if the relation of the distance between the centroids to the distance between the centroids and the objects assigned to these clusters (rd1) is small and the relation of the distance between the centroids and the objects assigned to these clusters to the distance between the centroids and the objects not assigned to these clusters (rd2) is large,
- do not change the number of clusters if the relation of the distance between the centroids to the distance between the centroids and the objects assigned to these clusters (rd1) is suitable and the relation of the distance between the centroids and the objects assigned to these clusters to the distance between the centroids and the objects not assigned to these clusters (rd2) is suitable,
- diminish the number of clusters if the relation of the distance between the centroids to the distance between the centroids and the objects assigned to these clusters (rd1) is large and the relation of the distance between the centroids and the objects assigned to these clusters to the distance between the centroids and the objects not assigned to these clusters (rd2) is small.

As the result from the FLC we accepted:

- signal to enlarge the number of clusters (+1),
- signal to do not change the number of clusters (0),
- signal to diminish the number of clusters (-1).

The knowledge base (rule base) of FLC is shown in Table 1 (fuzzy values of clusters' number change).

Figure 1 show membership functions of the relation of the distance between centroids to the distance between the centroids and the objects assigned to these clusters, the relation of the distance between the centroids and the objects assigned to these clusters to the distance the between centroids and the objects not assigned to these clusters and the value of the clusters' number change respectively. The shape of the membership functions was established experimentally and user can adapt them to solved problem. The FLC uses the center of gravity as a defuzzyfication method. Similar systems were successfully applied to other multiobjective optimalization problems, such as the Connected Facility Location Problem (ConFLP) [6] or the Wireless Access Points Placement Problem (WAPP) [7].

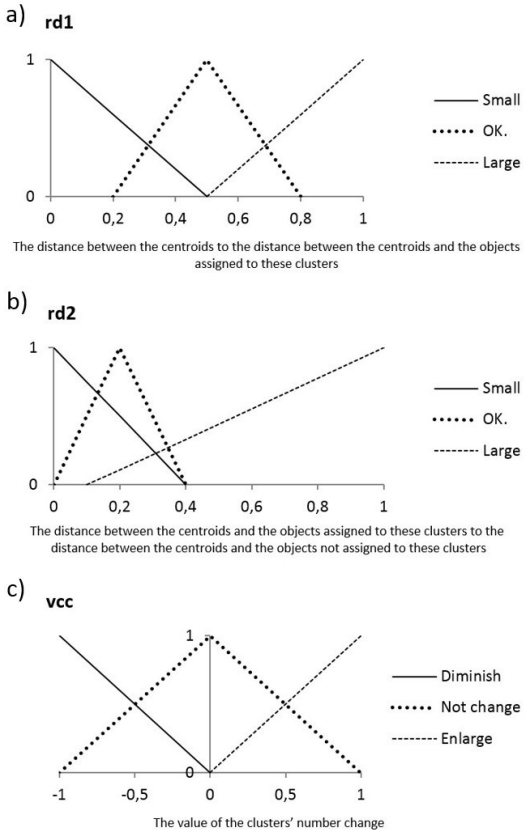

Fig. 1. Membership functions: a) the relation of the distance between the centroids to the distance between the centroids and the objects assigned to these clusters, b) the relation of the distance between the centroids and the objects assigned to these clusters to the distance between the centroids and the objects not assigned to these clusters, c) the value of the clusters' number change

IV. COMPUTATIONAL EXPERIMENT

The goal of our experiments is verification of the idea of the hybrid fuzzy-genetic algorithm to solving a clustering problem. In experiments we verify the ability of the FLC to optimize of the number of clusters, basing on experts' knowledge and data originated in the GA. Optimization of centroids' positions and optimal assigning of objects to clusters is realized by a genetic algorithm. For tests we used a set of data from "The Fundamental Clustering Problems Suite" (FCPS) [12] as a benchmark. We have chosen 4 two-dimensional problems from 400 to 4096 objects and from 2 to 3 clusters. Figure 2 show the distribution of objects in space and known a priori classifications in problems selected for our tests.

All tasks were solved by a k-means algorithm (we used the k-means method from the "rattle" library in R programming language [11]), proposed a hybrid genetic algorithm with the fuzzy logic (FLGA) and the simple genetic algorithm (SGA) - an algorithm proposed in [3], and modified by me to solve a clustering problem. The correct value clusters' number was used in a k-means and the SGA algorithms, the

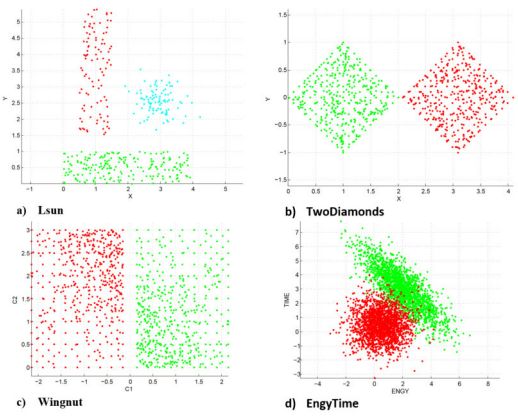

Fig. 2. The distribution of objects in space and known a priori classifications in problems selected for tests: a) Lsun, b) TwoDiamonds, c) Wingnut, d) EngyTime

TABLE II
THE DISTANCE BETWEEN OBJECTS AND CENTROIDS

		The distance between objects and centroids		
The problem Name	**The number of objects**	**k-means**	**SGA**	**FLGA**
Lsun	400	838	924	1112
TwoDiamonds	800	1645	1653	1654
Wingnut	1070	1815	1855	1821
EngyTime	4096	7912	9454	10093

FLGA was started from an incorrect number of clusters. Each algorithm was executed 10 times. In Table 2 and 3 there is the best distance between objects and centroids and the number of correctly assigned objects to clusters obtained by all algorithms. Figure 3 shows the distribution of objects in space obtained by the k-means algorithm and the FLGA algorithm.

V. CONCLUSIONS

The proposed Fuzzy Logic-Genetic Algorithm was able to find a solution near the optimum. However, looking at charts in Figure 3 presenting the distribution of objects after an optimization, it is easy to notice that improvement of this result is still possible. In Lsun task, the assignment of objects to clusters in the FLGA algorithm is more similar to known a priori classification, than in the k-means algorithm.

In all tasks proposed, the FLC correctly qualified the number of clusters.

The time operation of the FLGA on a PC computer did not exceed 10 minutes for the task of optimization of 4096 objects.

TABLE III
THE NUMBER OF CORRECTLY ASSIGNED OBJECTS

		The number of correctly assigned objects		
The problem Name	**The number of objects**	**k-means**	**SGA**	**FLGA**
Lsun	400	391	187	324
TwoDiamonds	800	800	567	767
Wingnut	1070	981	676	913
EngyTime	4096	4010	2128	2945

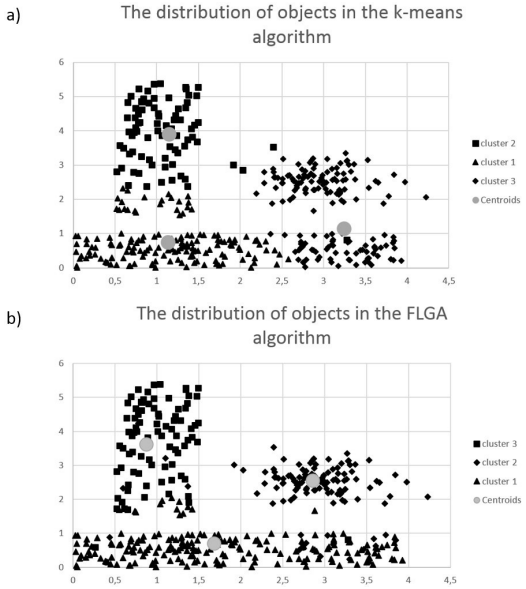

Fig. 3. The distribution of objects in space obtained by: a) the k-means algorithm, b) the FLGA algorithm

In tasks with a large number of objects, the time of calculations can be considerably longer. The parameters of an algorithm, eg. the number of generations, can be changed to fulfil the users' needs and reach a required accuracy of calculations.

The proposed algorithm is an efficient tool for solving clustering problems, where the number of clusters cannot be pre-determined. The proposed algorithm can be used for solving similar problems of multi-objective optimization.

REFERENCES

[1] Berkhin, P., "Survey of clustering data mining techniques." *Technical report*, Accrue Software, San Jose, CA, 2002.
[2] Maimon, O., Rokach, L., "Data Mining and Knowledge Discovery Handbook", Springer. DOI: 10.1007/978-0-387-09823-4
[3] Michalewicz Z., "Genetic Algorithms + Data Structures = Evolution Programs", *Springer Verlag*, Berlin (1992).
[4] Pytel K., Nawarycz T., "Analysis of the Distribution of Individuals in Modified Genetic Algorithms" *[in] Rutkowski L., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J., Artificial Intelligence and Soft Computing*, Springer-Verlag Berlin Heidelberg (2010).
[5] Pytel K., "The Fuzzy Genetic Strategy for Multiobjective Optimization", *Proceedings of the Federated Conference on Computer Science and Information Systems*, Szczecin, (2011).
[6] Pytel, K, Nawarycz, T., "A Fuzzy-Genetic System for ConFLP Problem", *Advances in Decision Sciences and Future Studies, Vol. 2*, Progress & Business Publishers, Krakow 2013.
[7] Pytel, K., Nawarycz, T. "The Fuzzy-Genetic System for Multiobjective Optimization", *[in] Rutkowski L., Korytkowski M, Scherer R., Tadeusiewicz R., Zadeh L., Zurada J., Swarm and Evolutionary Computation*, Springer-Verlag Berlin Heidelberg 2012.
[8] Tan, P. N., Steinbach, M., Kumar, V., "Introduction to Data Mining" *Parson*, 2006
[9] Jiang, D., Tang, C., Zhang, A., "Cluster analysis for gene expression data: a survey", *IEEE Transactions on Knowledge and Data Engineering (Volume:16 , Issue: 11. pp. 1370 - 1386*, 2004
[10] Zitzler E., "Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications", Zurich (1999).
[11] Rattle: A Graphical User Interface for Data Mining using R http://rattle.togaware.com/
[12] The Fundamental Clustering Problems Suite https://www.uni-marburg.de/fb12/datenbionik/data/