Journal of mathematics and computer Science 10 (2014) 144 - 156

Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

Introduce a New Algorithm for Data Clustering by Genetic Algorithm

J. Vahidi

Department of Applied Mathematics, Iran University of Science and Technology, Behshahr, Iran,
jvahidi@iust.ac.ir

Saeed Mirpour

Sama Technical and vocational training college, Islamic Azad University, Babol Branch, Babol, Iran

Article history:
Received February 2014
Accepted March 2014
Available online April 2014

Abstract

Clustering of data into adequate categories is one of the most important issues in pattern recognition. What is important in clustering, doing so is no predetermined pattern, provided that the same data should be in a category. In this paper, first, a clustering method using a grouping genetic algorithm (GGA) to describe, then the proposed model we introduce and the proposed method are tested on several sets of data and finally we compare the proposed method with the GGA algorithm.

The results show that the proposed algorithm is well-GGA gives us the answer and in terms of time and space complexity are much better than GGA.

Keywords: grouping genetic algorithm, clustering, pattern recognition

1. Introduction

There are two popular methods in data mining to find existed hidden patterns in data which are clustering and Classification. Also, in most cases these two methods have been stated interchangeably, but they are two different analytical approaches.

There are numerous algorithms for clustering data such as K-Means, DBSCAN, etc. Classical clustering algorithms generally have some different disadvantages. For example, K-means algorithm and its family of algorithms which introduced as a basic method for clustering, have some disadvantages. The most important disadvantages of it are its dependency on selecting

initial centroids (center points), that is, it is possible in some cases that the algorithm doesn't find correct solution.

Since, the problem of clustering data is so important and is so useful in many scientific fields, the main goal of this paper is to propose a new method for clustering data which can overcome disadvantages of previous clustering methods such as K-means and its family of algorithms.

K-means algorithm is one of the clustering algorithms proposed by J.MacQueen in 1967 and then in 1975 proposed by J.A.Hartigan and M.A.Wang. K-means is also one of the simplest clustering algorithms.

The other algorithm which proposed by S.C.Johnson in 1967 is hierarchal clustering algorithm and in 1973 Fuzzy C-means which proposed by Dunn and improved by Bezdek in 1981.

The first clustering algorithm which is proposed for very large databases is BIRCH. This algorithm tries to form the best clusters using available main memory (which is less than dataset in size).By the way, this algorithm reduces the required time for I/O operations. In BIRCH, first, data points are stored as sub-Clusters in dataset, which is considered as Cluster-features, here. N ext, generated cluster-features are clustered in K groups using common hierarchal clustering procedure.

BIRCH uses a tree structure namely CF-Tree, to create and store Cluster-Features. In this tree, in each instance, one object is dynamically constructed.

Dimension and size of CF-tree is determined by B and ε parameters which B denotes maximum number of children which are non-leaf nodes and ε determines the defined threshold of cluster-feature. It should be noted that in BIRCH an initial scan suffices to reach a good clustering. Though, more scans improve the quality of clustering.

The other main method for clustering stream data is STREAM [3]. In this algorithm, streams of data enter as sliced streams $x_1 \ldots x_n$ where each slice is a fit of main memory. Actually data streams usually contain slices that many points repeat in each slice. Since, the repetitive process of some point is a time consuming operation, therefor if the slice data has been showed as compact and weighted data, the clustering algorithm will run more quickly. So, each x_i in STREAM is shown as a weighted data set. So that each distinct point appears only once in it but has a weight equal to the number of occurrences of that point in this slice. Each x_i clusters using local search and only K weighted centroids (centroids are weighted using the number of points that approaches to them) is stored for each slice x_i. Then, the LOCAL SEARCH algorithm applied on all of the weighted centroids which are obtained from $x_1 \ldots x_i$, to form a set of weighted centroids for whole stream $x_1 \cup x_2 \cup \ldots \cup x_i$.

In CURE algorithm, firstly, a number of C data points (C is constant) which are scattered in cluster, are selected these points shows the shape and size of cluster. Then, the selected points

are an approached to centroid of cluster with coefficient α. The scattered points after approaching are considered as a representative of clusters. In each step of hierarchal clustering algorithm CURE, the clusters which have the least distance to each other, are merged. CURE will operate well in presence of outliers and can detect variety of clusters with no spherical phases for large data bases; the CURE uses a combination of randomized sampling and partitioning. First, the selected random samples from dataset are partitioned and each partition individually be clustered. Individual clusters in second pass are clustered again to generate desired clusters [9].

The rest of the paper has been structured as follows: next section summarizes some important definitions on clustering and clustering measures and distances. Section 3 presents the grouping genetic algorithm. Section 4 contains the experimental part of the paper, where the performance of the proposed clustering algorithm is evaluated. Section 5 closes the paper by giving some suggestions and conclusion.

2. Some Definitions and Measures

Essentially, a collection of similar data is called Cluster and the operation of grouping data points in subsets of main groups, i.e. Clusters, is called Clustering. In Clustering we try to assign data points to some clusters so that within cluster similarity is (be) maximized and between cluster similarity (of data points located in two different clusters) is (be) minimized.

The similarity measure used in this paper, is distance measure that is, which places less distant (near to each other) objects in the same cluster.

Computing distance between two data points is so important. Distance, which is also called Congruence, helps us to move along (in) data space and form clusters. Computing distance between two data points enables us to understand the nearness of these data and based on it, we can decide whether place them in the same cluster or not. There are various mathematical functions to compute distance between data points including: Euclidean distance, Hamming distance, etc.

The most practical distance which is defined using a definite and symmetric matrix namely A, is given below:

$$d^2(x_i - x_j) = \left\| x_i - x_j \right\|_A = (x_i - x_j).A.(x_i - x_j)^T \qquad (1)$$

Where, T denotes Transpose operation and matrix A determines shape and size of set of vectors which are located in a distance to the specified vector X_i.

The most popular form of a distance with norm A, is the simplest case, that is when A=I, where I is identity matrix. In this case, the generated distance is Euclidean distance and is defined as follows:

$$d_E^2(x_i, x_j) = \|x_i - x_j\|^2 = (x_i - x_j).(x_i - x_j)^T \qquad (2)$$

There are other norms which are used in cases that all clusters have ellipsoid shape but direction and size of every ellipsoid is different to others. Mahalanobis distance which is particularly used in such cases is defined as follows:

$$d_M^2(x_i, x_j) = \|x_i - x_j\|_{\Sigma^{-1}} = (x_i - x_j).\Sigma^{-1}.(x_i - x_j)^T \qquad (3)$$

Where $\sum$ denotes covariance matrix.

3. GGA Algorithm

The Grouping Genetic Algorithm[2](GGA)[10] , is one class of evolutionary algorithms which is modified specifically to cope with grouping problems i.e. the problems that in them some items should be assigned to a set of predefined groups. So, in GGA, coding scheme, mutation and crossover operators are modified in comparison to traditional Genetic Algorithm in order to obtain a more compact and efficient algorithm that can be applied to grouping based problems.

In GGA, encoding is carried out by separating each individual in the algorithm into two parts:

C= [l|g], the first part is the elements section, whereas the second part is called the group section of the individual. As an example following our notation, in a solution for a clustering problem with N elements (observation) and K clusters, the individuals will have the following aspects:

$$l_1, l_2, ..., l_n \mid g_1, g_2, ..., g_k$$

Note that l_j represents the cluster to which jth observation is assigned, whereas group section keeps a list of tags associated to each of the clusters of the solution. In a formal way:

$$l_j = g_i \leftrightarrow x_j \in C_i$$

Note also that the length of the element section is fixed for a given problem (equals N), but the group section's length is not fixed, it varies from one individual to another. Thus the GGA does not need as input parameter the number of clusters, but it searches for the best k in terms of the objective function.
The crossover operator implemented in the grouping genetic algorithm used in this paper is a modified version of the one initially proposed by Falkenauer (1992) to adapt it to the clustering problem. The process follows a two parent's one offspring schema.
In GGA algorithm, two different mutation and crossover operators namely Mutation by cluster splitting and Mutation by cluster merging are described.
For more details, refer to [2].

4. Proposed Algorithm

In proposed algorithm, we have illustrated a new clustering algorithm and then applied it to some experimental datasets. In this algorithm, we have used a different encoding scheme in

comparison to that used in GGA. In addition, some modifications have been made in mutation and crossover operators which result in improving the efficiency of GGA algorithm.

4-1- Problem Formulating

Data encoding in this proposed method is different to that in GGA algorithm.

Due to above definitions, if suppose there are N object and we want to assign them to K clusters, individuals will be as follows:

$$G_1 = l_1, l_2, ..., l_i$$
$$G_2 = l_{i+1}, l_{i+2}, ..., l_j$$
$$.$$
$$.$$
$$.$$
$$G_K = l_r, l_{r+1}, ..., l_n$$

Note that, data are assigned randomly to each cluster and l_i denotes the number of cluster which i^{th} object is assigned to it. Indeed, in this algorithm, the number of individuals is equal to the number of clusters. Since the solution of problem is in these K clusters, so, we consider initial population as K cluster and will have tried to obtain optimal solution of clustering problem using genetic algorithm operators. After that, we created these individuals randomly, then find average point for each individual, then, sort genomes of individuals in terms of the distance between each object to these average points. The distance measure is one of the fundamental components in dealing with clustering problems, because the similarity between two different vectors as X_i and X_j, is usually related to a distance measure in feature space S. To calculate distance, one can use the most popular and simplest measure that is Euclidean distance measure which is defined as follows:

$$d_E^2(x_i, x_j) = \|x_i - x_j\|^2 = (x_i - x_j).(x_i - x_j)^T$$

Where T denotes transpose operator.

4-2- Selection Operator

In this algorithm, we can use competitive selection mechanism. This method selects some members of population and then if a specific condition is satisfied, selects the best one or some of the more better of them as parents. If the condition is not satisfied, the worst member or a number of more worse individuals are considered as parents to form future population. The process which is used in proposed algorithm is regarded as a replacement of crossover operator and because of the size of initial population is not so many and indeed, is equal to number of clusters, we try using all of individual in crossover operation. However, one can simply select two parents successively from population.

4-3- Crossover Operator

The crossover which has used in this paper in proposed method is single point crossover operator. This new operator combines two individuals by selecting one random position such as P. The position of point P is less or equal to length of individuals. If the length of genomes in individuals is N, then by combining two parent individuals, two offspring forms as follows:

On offspring by copying genomes 1 to (P-1) of first parent and genomes P to N of second parent and likewise the other offspring by copying genomes 1 to (P-1) of second parent and P to N from first parent are created. In this kind of crossing over of two parents, two offspring are created. This kind of crossover is shown is figure 1 which we supposed that P=4.

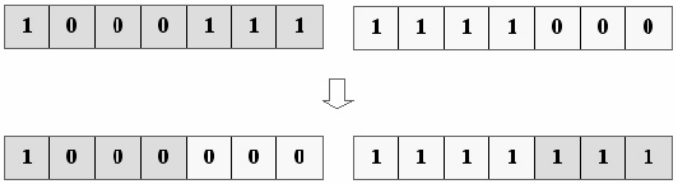

Figure 1: single point crossover

Since, genomes of each individual are ordered in terms of distance of themselves to average point of that individual, therefore to improve the efficiency of this method and speeding up the algorithm to obtain optimal solution, in most cases the point P is selected in range of 1 to half of the length of individual. As a result, the genomes which located in more far distance to average point, more likely be selected. Indeed, genomes which are located in far distance to average point, doesn't belong to considered cluster and should have find their desire location in other clusters.

4-4- Mutation operator

Mutation cause to search unknown areas of problem space. One can deduce that the most important task of mutation is escaping from local optimum. In mutation one genome can be added (removed) to (from) an individual. In this algorithm, to avoid placing in a local optimum, in each mutation, one genome is removed from an individual. In some situations there are some outliers than can effect on clustering results and therefore the algorithm cannot reach the optimal solution. So, mutation operator by removing one genome from individuals can improve clustering result.

4-5- Fitness function

After creating initial population, we should assign to each individual a value. The function which determines value of each individual, first calculated the average of each genome in an individual as follows:

$$AVG_{c_j} = \frac{\sum_{i=0}^{n} N_i}{n}$$

Where N_i denotes i'th object (genomes) and AVG_{cj} denotes average point of that individual and n is the number of genomes in it.

Next, calculate the distance of each genome in an individual to average point of it.

$$D_{N_i} = N_i - AVG_{c_j}$$

Now, calculate the average of these distances and consider it as value of the individual. We have:

$$V_{C_j} = \frac{\sum_{i=0}^{n} D_{N_i}}{n}$$

Where V_{cj} is considered as the value of individual. Note that less the value of individual the value of Vcj, more the value of individual.

After that we calculate value of each individual, we can apply existed genetic algorithm operators to population and therefor form the next generation.

4-6- Experimental results

In this section, we apply proposed algorithm on some artificial dataset and the results are shown in figures. First, we apply the algorithm on 50 randomly generate data point.

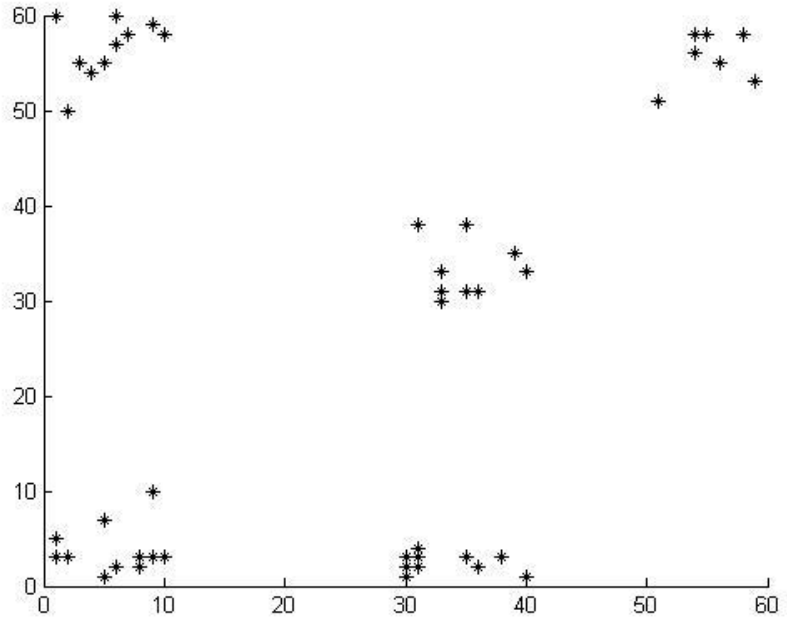

Figure 2: the first randomly generated artificial dataset

Proposed algorithm, has divided these data point in 5 clusters and the result of clustering is showing in figure 3:

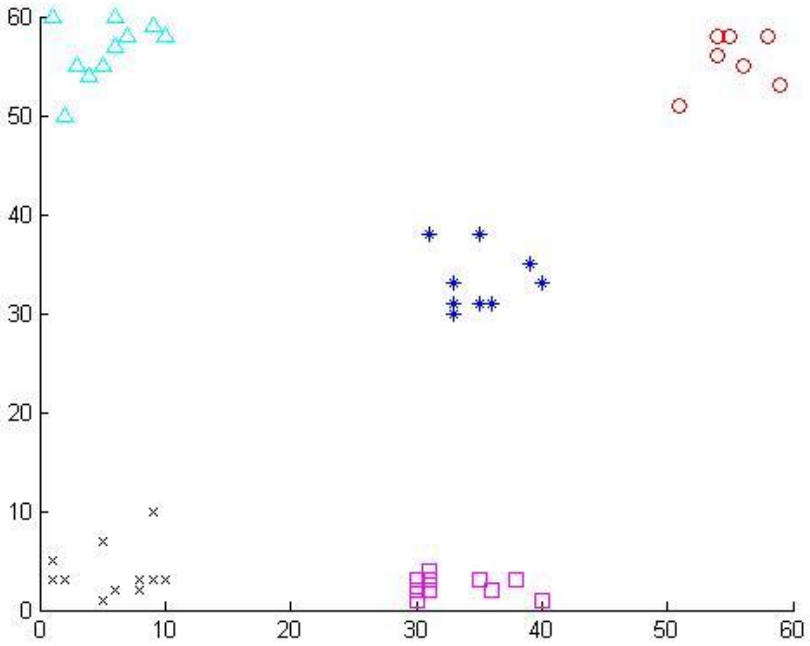

Figure 3: the best clustering result for the first artificial dataset

In the next experiment with artificial dataset, we have tried to create data points with respect to averages and covariance matrices and cluster this data set whit our proposed algorithm. We evaluate the performance of the proposed algorithm in 2D clustering problems using 200 data points which have been generated randomly by a Gaussian distribution of 9 clusters with equal probability of occurrence and average points of each class and covariance matrices were as follows, respectively:

$\mu_1 = (1,-1)$, $\mu_2 = (1.5,0)$, $\mu_3 = (0,1)$, $\mu_4 = (-1,1)$, $\mu_5 = (2,-1)$, $\mu_6 = (-2,-1)$, $\mu_7 = (-0.5,2)$, $\mu_8 = (-1,-1)$, $\mu_9 = (1.5,0)$

$$\Sigma_1 = \ldots = \Sigma_8 = \begin{bmatrix} 0.2^2 & 0 \\ 0 & 0.2^2 \end{bmatrix}$$

Fig 4 shows the observations which are randomly generated under mentioned statistical distribution.

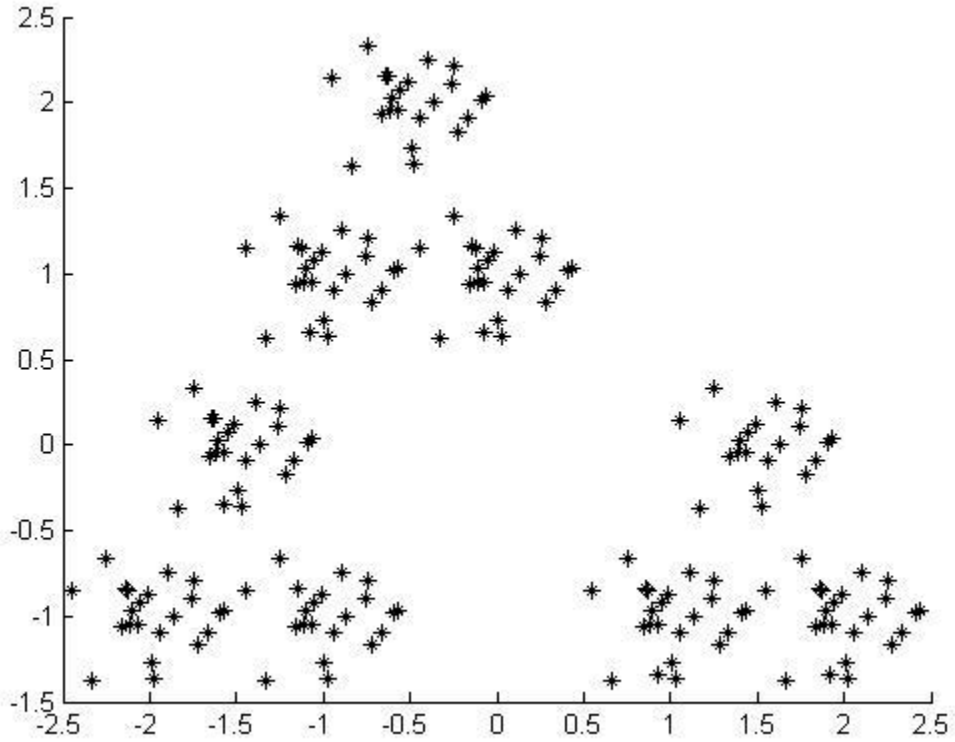

Fig 4: the second generated artificial dataset using statistical distribution

Fig 5 shows the clustering results of randomly generated data points using proposed algorithm. As we observe, 3 clusters are created.

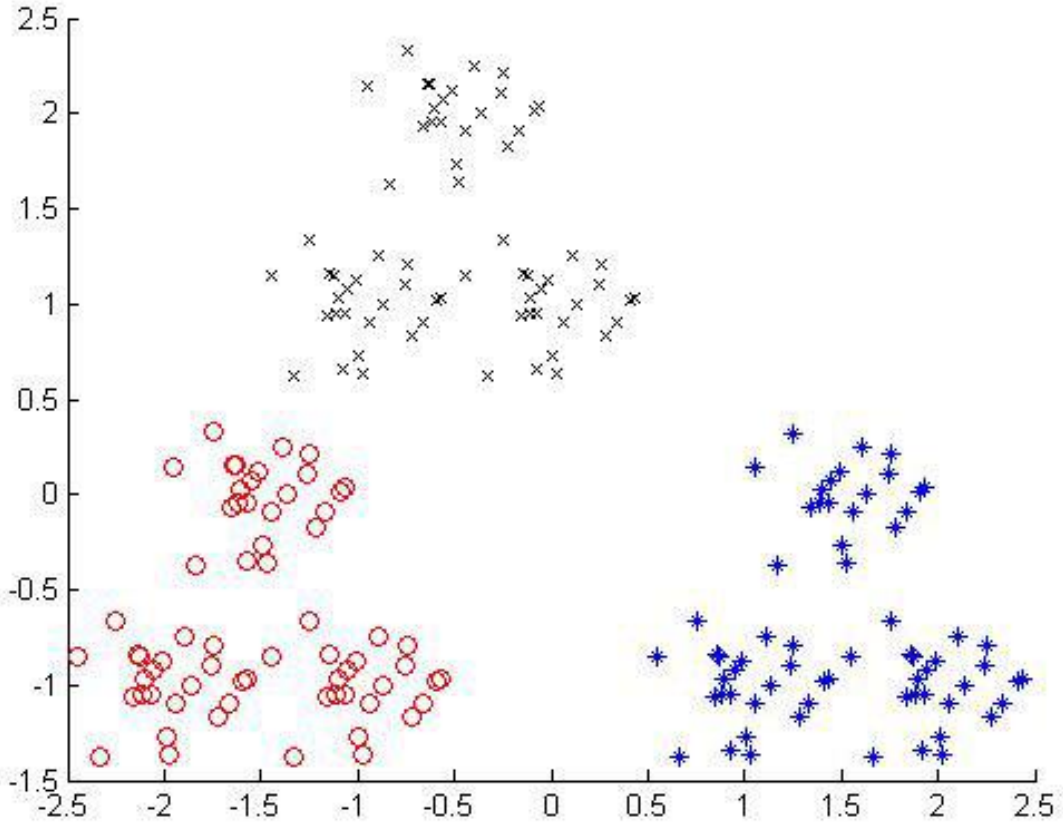

Fig 5: the best clustering result for the second generated artificial dataset using statistical distribution

In the last experiment, the algorithm based on 1000 data randomly in two dimensions and in the interval [0,500; 0,500] has been developed, tested. The result of this experiment are shown in figure 6-can be seen.

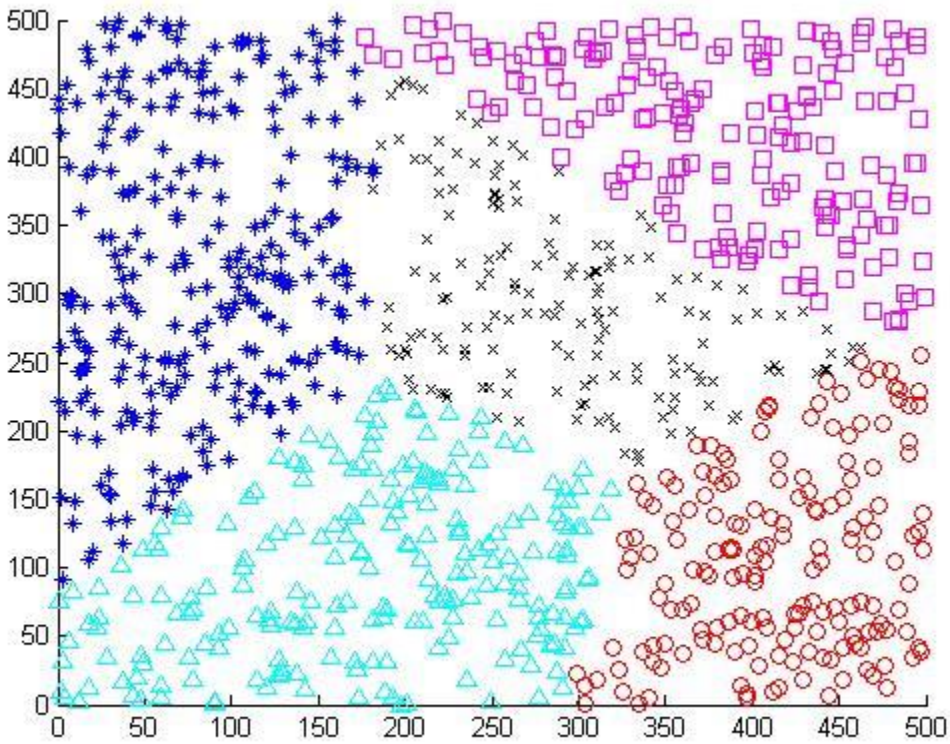

Fig 6: the best clustering result for 1000 data randomly

5- Conclusions

The proposed algorithm in the paper can be caparisoned to GGA algorithm which mentioned in section 3 with respect to time and space complexity.

First, we comparison these algorithms in terms of space complexity:

In GGA algorithm the required space of problem for each individual is equal to the number of data plus number of clusters and this space is deferred on the size of initial population. In fact, the required space for GGA is equal to P*(N+K), where P is the number of initial population, N is the number of data for clustering and K is the number of clusters. In the other hand, the proposed algorithm which illustrated in section 4 has a constant space complexity equal to the number of data for clustering. This constant space complexity is due to changing the encoding scheme of problem.

Now, we evaluate the time complexity of GGA and proposed algorithm. To do this, firstly, operators and functions which have used in these algorithms should be caparisoned.

In GGA, for each crossover operation, each individual should be traversed 5 times, so that crossover operation can be done. In fact, if the number of data in given problem is n, this

operation will be completed in 5n time, while in proposed algorithm, before crossover operation, a sorting algorithm based on average point of data has been run and after it the crossover operation is done simply with n traversing of individuals. Since, for sorting each individual in proposed algorithm based on average point, two steps is required:

First step for obtaining average point and second step for sorting the data based on average point which in general the time complexity of this algorithm will be (n+nlogn). The second existed operator in these algorithms is mutation operator. GGA algorithm to do this operation requires one traversing which in fact needs n time unit to do it, but in proposed algorithm again due to the mentioned sorting algorithm, time complexity of this operation is (n+nlogn). Time unit the third operation which has been used in these algorithms is fitness function. GGA algorithm for calculating the value of fitness function requires n^2 time unit but the required time for calculating this value in proposed algorithm is order of n.

In general it can be concluded that the proposed algorithm is more efficient than GGA algorithm with respect to time complexity.

Refrences

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," Proceedings of the 29th VLDB Conference, 2003.

[2] L.E. Agustin-Blas, S. Salcedo-Sanz, S. Jimenez-Fernandez, L. Carro-Calvo, J. Del Ser “A new grouping genetic algorithm for clustering problems” Expert Systems with Applications 39 (2012) 96959703.

[3] Daniel Barbard, “Requirements for Clustering Data Streams, ” ACM SIGKDD Explorations Newsletter, vol. 3, pp. 23-27, 2002.

[4] Jurgen Beringer, Eyke Hullermerier, "Fuzzy Clustering of Parallel Data streams," Data & Knowledge Engineering , pp. 180-204, 2006.

[5] Albert Bifet1, Geo_ Holmes1, Bernhard Pfahringer1, Philipp Kranen2, Hardy Kremer2, Timm Jansen2, and Thomas Seidl2,” MOA: Massive Online Analysis, for Stream Classi_cation and Clustering,” 2010.

[6] Lior Cohen, Gil Avrahami, Mark Last, Abraham Kandel, "Info-fuzzy algorithms for mining dynamic data streams," Applied Soft Computing, pp. 1283-1294, 2008.

[7] M.M Gaber, Arkady Zaslavsky and Shonali Krishnaswamy, “Mining Data Streams: A Review,” SIGMOD Record, vol. 34, no. 2, June 2005.

[8] Mohammad GhasemiGol, Hadi Sadoghi Yazdi, Reza Monsefi “A New Hierarchical Clustering Algorithm on Fuzzy Data (FHCA),” International Journal of Computer and Electrical Engineering, vol. 2, no. 1, February 2010.

[9] S. Guha, R. Rastogi, and K. Shim, “CURE: An Efficient Clustering Algorithm for Large Databases,” Proc. SIGMOD, pp. 73-84, 1998.

[10] Richard J. Hathaway, James C. Bezdek, “Extending Fuzzy and Probabilistic Clustering to Very Large Data Sets,” Journal of Computational Statistics and Data Analysis. vol.51, no.1, pp. 215-234, 2006.

[11] Madjid Khalilian, Norwati Mustapha, “Data Stream Clustering: Challenges and Issues,” 2010.

[12] Raghu Krishnapuram,James M.Keller, “A Possibilistic Approach to Clustering,” IEEE TRANSACTIONS ON FUZZY SYSTEMS, vol. 1, no. 2, MAY 1993.

[13] Raghu Krishnapuram and James M. Keller, “Correspondence the Possibilistic C-Means Algorithm: Insights and Recommendations,” IEEE TRANSACTIONS ON FUZZY SYSTEMS, vol. 4, no. 3, AUGUST 1996.
[14] Alireza Rezaei Mahdiraji, "Clustering data streams: A survey of algorithms," International Journal of Knowledge-based and Intelligent Engineering Systems, pp. 39-44, 2009.
[15] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy,”Mining Data Streams: A Review,” SIGMOD Record, vol. 34, no. 2, June 2005.
[16] D. P. Mercer, Linacre College,” Clustering large datasets,” 2003.
[17] Liadan O’Chalaghan, Nina Mishra, Adam Meyerson,Sudipto Guha and Rajeev Motwani, “Streaming data algorithms for high quality clustering, ” Proc. of IEEE International Conference on Data Engineering, pp. 685– 694, 2002.
[18] Nikhil R. Pal, Kuhu Pal, James M. Keller, and James C. Bezdek, “A Possibilistic Fuzzy c-Means Clustering Algorithm,” 2005.
[19] Renxia Wan, Xiaoya Yan, Xiaoke Su, “A Weighted Fuzzy Clustering Algorithm for Data Stream,” presented at ISECS International Colloquium on Computing, Communication, Control, and Management.CCCM'08, 2008.
[20] Xuanli Lisa Xie,Gerardo Beni, ”A Validity Measure for Fuzzy Clustering, ” 1990.
[21] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,”