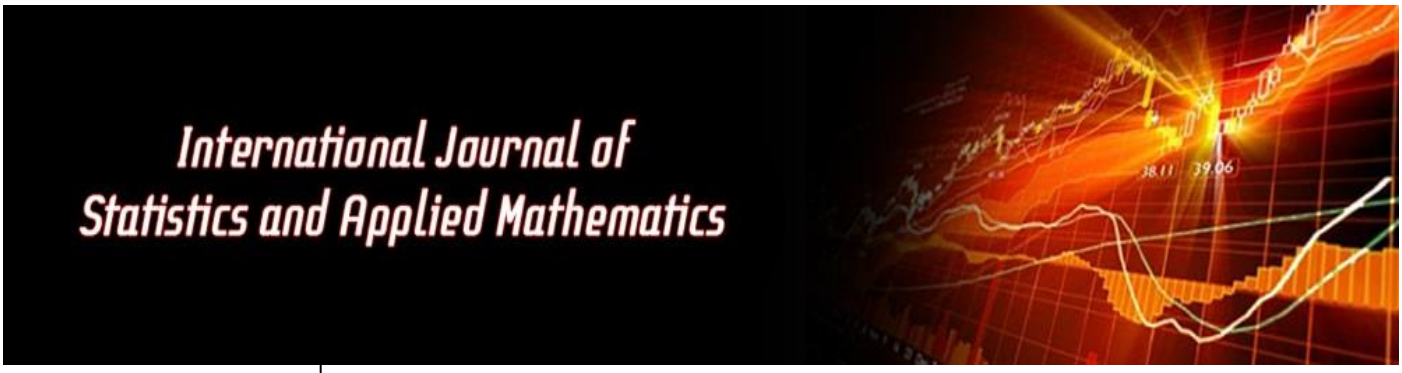
International Journal of Statistics and Applied Mathematics 2021; 6(5): 151-154



ISSN: 2456-1452
Maths 2021; 6(5): 151-154
© 2021 Stats & Maths
www.mathsjournal.com
Received: 22-07-2021
Accepted: 24-08-2021

Nisha Sumbherwal
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, CCS Haryana Agricultural University, Hisar, Haryana, India

BK Hooda
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, CCS Haryana Agricultural University, Hisar, Haryana, India

Corresponding Author:
Nisha Sumbherwal
Department of Mathematics and Statistics, College of Basic Sciences and Humanities, CCS Haryana Agricultural University, Hisar, Haryana, India

Genetic algorithm based clustering methods: A review

Nisha Sumbherwal and BK Hooda

Abstract
The paper presents the review on GA based clustering techniques. Clustering is widely used in different field such as biology, engineering, text mining, bioinformatics, and agriculture. To enhance the performance of clustering algorithms, Genetic Algorithms (GAs) is applied to the clustering algorithm. Genetic Algorithms are the best known evolutionary techniques which gives near global or global optimal solution for a given fitness function. The capability of GAs is applied to evolve the proper number of clusters and to provide appropriate clustering.

Keywords: clustering, genetic algorithm, fitness function, crossover, mutation, selection

1. Introduction
Genetic Algorithm(s) was introduced by Holland (1975) [4] and later it was described by Goldberg (1989) [3]. Goldberg was inspired by the theory of evolution as given by Darwin, which stated that the survival of an organism is affected by rule "survival of the fittest". Darwin also stated that the survival of an organism can be sustained through the process of reproduction, mutation and crossover. Darwin's theory of evolution was then adapted to the computational algorithm to find a solution to a problem. A solution which is generated by a genetic algorithm is called a chromosome and the collection of chromosomes is known as a population. A chromosome is made up of genes and its value can be either quantitative, binary, characters or symbols depending on the problem desired to be solved. The fitness is evaluated of the encoded chromosomes to measure the suitability of solution generated by genetic algorithm with the problem of investigation. The fitness function, also known as objective function represents the goodness of a chromosome. The fitness function is selected in such a way that a chromosome close to the optimal solution has a higher fitness value.
Few chromosomes in the population will go through the process called crossover, thus producing new chromosomes called offspring whose gene composition is the combination of their parent. In a generation, some chromosomes will also undergo the mutation process in their gene. The number of chromosomes undergoing crossover and mutation is controlled by the value of the crossover rate and the mutation rate. The chromosomes in the population that are taken for the next generation are selected on the basis of Darwin's evolution rule i.e. the chromosome with higher fitness value will have greater probability of being selected again in the next generation. The chromosome value will converge to a certain value after several generations, which is the best solution to the problem. This is an optimization approach using GA to find a near global or global optimal solution for a given fitness function. Various Genetic Algorithm based clustering methods that have been reviewed in detail are given in next section.

2. Genetic algorithm based clustering methods
2.1 Murthy & Chowdhury genetic algorithm
Murthy & Chowdhury in 1996 used genetic algorithm to search for an optimal number of clusters for the dataset without searching all possible partitions. The partition is encoded as a string of length t. The i^{th} element of the string denotes cluster number assigned to point i. An initial population of size P is selected randomly. Several strings of length t are generated randomly and the value of each element of the string are allowed to lie between 1 and K. The distance of the points from their respective cluster centers is used as the clustering metric.

Mathematically, the clustering metric μ (fitness function) for K clusters $C_1, C_1, \ldots, C_K$ is given as:

$$\mu = \sum_{l=1}^{K} \sum_{x_i \in C_l} \|x_i - c_l\| \quad (1)$$

Where, C_l denotes the l^{th} cluster and c_l is the l^{th} cluster center.

The selection operator mimics the survival of the fittest concept of natural genetic systems. Since it is a minimization problem therefore the probability of selecting a particular string in the population is taken inversely proportional to the fitness value.

Crossover operation on the mating pool of size P is performed in the following way:

1. Select $P/2$ pairs of strings randomly from the mating pool so that every string in the mating pool belongs to exactly one pair of strings.
2. For each pair of strings, generate a random number from $[0, t-1]$. If the value of random number is less than or equal to μ_c, perform crossover otherwise no crossover is performed, where μ_c denotes the crossover probability.

The Elitist strategy is implemented in the following way:

1. Copy the best string, S_o of the initial population in a separate location.
2. Perform selection, crossover and mutation operations to obtain a new population, P_1.
3. Compare S_1, the worst string in $P_1$with S_o in terms of their fitness values. If S_1 is found to be worse than S_o, replace S_1 by S_o.
4. Find the best string, S_2 in P_1 and replace S_o by S_2.

The process is executed for a fixed number of iterations and the best string obtained is selected as the optimal one.

2.2 Cowgill, Harvey & Watson genetic algorithm

Cowgill *et al*. (1999) [2] proposed a genetic algorithm for cluster analysis. They introduced a clustering algorithm that notably seeks to optimize a function which is defined in terms of within cluster cohesion and between cluster isolation. The algorithm attempts to maximize the variance ratio criterion, also called as Calinski-Harabasz index (CH) given by Calinski & Harabasz (1974) [1]. CH is defined as:

$$CH = \frac{\sum_{l=1}^{K} n_l \|c_l - c\| / K-1}{\sum_{l=1}^{K} \sum_{i=1}^{n_l} \|x_i - c_l\|^2 / n-K} \quad (2)$$

Where n denotes the total number of objects in the dataset and K represents number of clusters, n_l and c_l are the number of objects in l^{th} cluster and center of l^{th} cluster respectively and c denotes the global centroid.

The genetic algorithm (GA) begins with a population of size P. The population of chromosomes then undergo selection process in which chromosomes with higher value of CH has more probability of being selected. Then these selected chromosomes undergo crossover and mutation process. The process continues for many generations and the chromosome with the best fitness value in the end is selected.

2.3 Maulik & Bandyopadhyay genetic algorithm

Maulik & Bandyopadhyay in 2000 [6] suggested a genetic algorithm based clustering technique called GA-clustering. The searching capability of genetic algorithms is exploited in the GA-clustering method in order to search for appropriate cluster centers in the feature space that optimize the similarity metric of the resulting clusters. The distance of the points from their respective cluster centers is used as the clustering metric. Mathematically, the clustering metric μ for K clusters $C_1, C_1, \ldots, C_K$ is given as:

$$\mu = \sum_{l=1}^{K} \sum_{x_i \in C_l} \|x_i - c_l\| \quad (3)$$

The task of the GA is to search for appropriate cluster centers $c_1, c_2, \ldots, c_k$ such that μ is minimized. The GA-clustering algorithm follows the basic steps of GA which are described as follows:

Each string is considered as a sequence of real numbers (floating point representation) which represent the K cluster centers. The length of a chromosome for p dimensional space is $p \times K$ in which the first p positions represent the p dimensions of the first cluster center, the next p positions represent p dimensions of the second cluster center, and so on. The cluster centers encoded in each chromosome are initialized to K randomly selected points from the data set and this process is repeated for each of the P chromosomes in the population, where P is the size of the population.

The fitness is computed in two different phases. The clusters in first phase are formed according to the centers encoded in the chromosome under consideration. For this, each point $i(i = 1, 2, \ldots, n)$ is assigned to one of the clusters C_l with center c_l such that

$$\|x_i - c_l\| < \|x_i - c_m\|, l = 1, 2, \ldots, K \; and \; l \neq m \quad (4)$$

After grouping the objects in K clusters, the cluster centers encoded in the chromosome are replaced by the mean points of the respective clusters. Thus for cluster C_l, the new center c_l^* is computed as

$$c_l^* = \frac{1}{n_l} \sum_{x_i \in C_l} x_l \quad (5)$$

where, n_l is the number of data points in the l^{th} cluster. The c_l^*s replace the previous c_l s in the chromosome.

A single point crossover with a fixed crossover probability, μ_c is used. In range $[1, t-1]$, a random integer, called the crossover point is generated for chromosomes of length t. To produce two offspring, the portion of the chromosomes lying to the right of the crossover point is exchanged.

With a fixed probability of μ_m, each chromosome undergoes mutation. A number η in the range $[0, 1]$ with a uniform distribution is generated for mutation. If the value at a gene position is v then after mutation it becomes

$$v \pm 2 \times \eta \times v, v \neq 0$$
$$v \pm 2 \times \eta, v = 0$$

The + or - sign occurs with equal probability.

For a maximum number of iterations, the process of fitness computation, selection, crossover, and mutation processes are performed. The best string seen upto the last generation provides the solution to the clustering problem.

2.4 Roy & Sharma genetic algorithm

Roy & Sharma in 2010 [8] introduced Genetic K-means clustering algorithm for mixed variables data. Let X denotes a $n \times p$ data matrix consisting of n observations in p dimensional space. Each partionining is represented by a string which is a sequence of numbers $b_1, b_2, \ldots, b_n$ where b_i represents the cluster number to which the object i belongs

and it takes value from $\{1,2,\dots,K\}$. The cost function for mixed data clustering is defined as:

$$\psi = \sum_{i=1}^{n} d(x_i - c_l) \quad (6)$$

$$d(x_i - c_l) = d_{num}(x_i, c_l) + d_{cat}(x_i, c_l) \quad (7)$$

where, $d_{num}(x_i, c_l)$ denotes the distance of object i from its closest cluster center c_l, for numeric attributes, $d_{cat}(x_i, c_l)$ denotes the distance of object i from its closest cluster center c_l, for categorical attributes.

$$d_{num}(x_i, c_l) = \sum_{k=1}^{p_n} \left(w_k (x_{ik}^r - c_{lk}^r)\right)^2$$

$$d_{num}(x_i, c_l) = \sum_{k=1}^{p_c} \delta(x_{ik}^c; c_{lk}^c)^2$$

Thus,

$$d(x_i, c_l) = \sum_{k=1}^{p_n} \left(w_k (x_{ik}^r - c_{lk}^r)\right)^2 + \sum_{k=1}^{p_c} \delta(x_{ik}^c; c_{lk}^c)^2 \quad (8)$$

$\sum_{k=1}^{p_n} \left(w_k (x_{ik}^r - c_{lk}^r)\right)^2$ denotes the distance between object i and its closest cluster center c_l, for numeric attributes, w_k denotes the significance of the k^{th} numeric attribute, which is computed from the data set and $\sum_{k=1}^{p_c} \delta(x_{ik}^c; c_{lk}^c)^2$ denotes the simple matching distance between data object i from its closest cluster center c_l for categorical attributes.

2.5 Selection Operator

For selection of chromosome, Proportional selection technique is used. In this technique population of next generation is determined by Z independent random experiments. Each experiment randomly selects a solution from the current population $S_1, S_2, \dots, S_Z$ according to the probability distribution $p_{(1)}, p_{(2)}, \dots, p_{(Z)}$ given as:

$$p_{(z)} = \frac{F(S_z)}{\sum_{z=1}^{Z} F(S_z)} \; z = 1,2,\dots,Z \quad (9)$$

$F(S_z)$ denotes the fitness value of solution S_z. The objective is to minimize ψ. Therefore, solutions which have smaller distance from their closest cluster center have higher probabilities of been selected hence should be assigned higher fitness value while illegal strings should be assigned with lower fitness values as they are less desirable and have lower probabilities for survival.

The value of $F(S_z)$ is $1.5 \times d_{max} - d(S_z)\; if\; S_z$ is legal string otherwise its value is $e(S_z) \times F_{min}$ where d_{max} is the maximum value that has been encountered till the present generation, F_{min} is the smallest fitness value of the legal strings in the current population if they exist, otherwise F_{min} is defined as 1.

2.6 Mutation Operator

During mutation, b_i are replaced with b_i^* for $i = 1,2,\dots,n$ simultaneously where b_i^* is a cluster number which is randomly selected from {1,2,…,K} with probability distribution $\{p_{(1)}, p_{(2)}, \dots, p_{(K)}\}$ given as:

$$p_{(l)} = \frac{1.5 * d_{max}(x_i) - d_{euc}(x_i, c_l) + 0.5}{\sum_{l=1}^{K} (1.5 * d_{max}(x_i) - d_{euc}(x_i, c_l) + 0.5)} \quad (10)$$

where $d_{euc}(x_i, c_l)$ is the Euclidean distance between object i and l^{th} cluster center and

$$d_{max}(x_i) = max\{d_{euc}(x_i, c_l)\}$$

The value $d_{euc}(x_i, c_l)$ is defined as 0 if l^{th} cluster is empty. The bias 0.5 is introduced to avoid divide by zero error in the case that all objects are equal and are assigned to the same cluster in the given solution.

2.7 K-means Operator

K-means operator (KMO) is introduced to speed up the convergence process. In a solution that is encoded by $b_1, b_2, \dots, b_n$, b_i is replaced by b_i^* for $i = 1,2,\dots,n$ simultaneously, where b_i^* is the cluster number whose centroid is closest to point i. To account for illegal strings, $d(x_i, c_l) = \infty$ if l^{th} cluster is empty. This avoids reassigning all objects to empty clusters. Therefore after the application of KMO, illegal string will remain illegal.

2.8 Singh & Misra genetic algorithm

Singh and Misra (2014) [9] investigated the use of Genetic Algorithms to determine the optimal classification as well as the effectiveness of initial parameters. For this, the K cluster centers encoded in each chromosome are initialized at randomly selected points from the data. This process is repeated for each of the P chromosomes in the population, where P is the size of the population. In the fitness computation process, a fixed number K is generated randomly and is considered as the number of clusters to be formed. The clusters are then formed according to the centers encoded in the chromosome. This is performed by calculating the sum of the distance of each data point from each cluster center in a cluster. After this, by comparing the sum of each individual, the individuals having minimum sum is identified this means that the sum of all clusters in that individual is minimum and this give the fitness function. The Roulette wheel selection and single point crossover with a fixed crossover probability i.e. μ_c are used.

For mutation, a number η in the range $[0, 1]$ is generated with uniform distribution. If the value at a gene position is v and $\eta < 0.5$ then after mutation it becomes $v - 1$ otherwise it becomes $v + 1$.

The processes of fitness computation, selection, crossover and mutation are executed for a maximum number of iterations. The best string seen up to the last generation provides the solution to the clustering problem.

2.9 Malki, Rizk, El-Shorbagy & Mousa genetic algorithm

Malki *et al.* in 2016 presented a Hybrid Genetic Algorithm with K-means clustering algorithm in order to overcome the drawbacks of K-means clustering method such as it may produce empty clusters depending on the initial cluster centers, converges to local optimum value and may also not able to give global solution to large problems with reasonable amount of computational effort.

The Hybrid Genetic Algorithm with K-means clustering algorithm comprise of two phases:

Phase I: K-means Algorithm

Phase II: Genetic algorithm

In phase I, K initial cluster centers $c_1, c_2, \dots, c_K$ are randomly selected from the dataset. Then a point i is assigned to cluster C_l if and only if $\|x_i - c_l\| < \|x_i - c_m\|$ for all $l =$

$1, 2, \ldots, K$ and $m \neq l$. The new clusters centers $c_1^*, c_2^*, \ldots, c_K^*$ are computed as follows:

$$c_l^* = \frac{1}{n_l} \sum_{x_i \in C_l} x_i \quad (11)$$

where, n_l denotes the number of objects in l^{th} cluster. The phase I is terminated when $c_l^* = c_l$ for all $l = 1, 2, \ldots, K$. Thus in this phase an initial center for all pre-determined clusters are obtained.

In phase II, GA is used for clustering in which chromosomes are encoded using point based technique. The fitness function is given as:

$$\mu = \sum_{l=1}^{K} \sum_{x_i \in C_l} \|x_i - c_l\| \quad (12)$$

This fitness function is minimized.

After then chromosomes are selected using Roullete wheel selection technique. The probability of selecting individuals is directly proportional to the fitness of the individual. The selected chromosomes undergo crossover and mutation and the best solution is found with minimum fitness value.

3. Conclusion

The problems related to grouping of objects into two or more segments occur in many areas such as engineering, text mining, bioinformatics, agriculture, pattern recognition, voice mining, mechanical engineering, spatial data analysis, image segmentation, textual document collection and artificial intelligence. Genetic Algorithms are the best known evolutionary techniques which gives near global or global optimal solution for a given fitness function. The capability of GAs is applied to evolve the proper number of clusters and to provide appropriate clustering. The paper presents the review of GA based clustering techniques. Various clustering techniques reviewed in detail were Murthy & Chowdhury genetic algorithm, Cowgill, Harvey & Watson genetic algorithm, Maulik & Bandyopadhyay genetic algorithm, Roy & Sharma genetic algorithm and Malki, Rizk, El-Shorbagy & Mousa genetic algorithm.

4. References

1. Calinski RB, Harabasz J. A dendrite method of cluster analysis. Communication in Statistics 1974;3(1):1-27.
2. Cowgill MC, Harvey RJ, Watson LT. A genetic algorithm approach to cluster analysis. Computers & Mathematics with Applications 1999;37(7):99-108.
3. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Addison Wesley Publishing Company, Boston 1989.
4. Holland JH. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor 1975.
5. Malki AA, Rizk MM, El-Shorbagy MA, Mousa AA. Hybrid genetic algorithm with k-means for clustering problems. Open Journal of Optimization 2016;5:71-83.
6. Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. Pattern Recognition 2000;33:1455-1465.
7. Murthy CA, Chowdhury N. In search of optimal clusters using genetic algorithms. Pattern Recognition Letters 1996;17(8):825-832.
8. Roy DK, Sharma LK. Genetic k-means clustering algorithm for mixed numeric and categorical data sets. International Journal of Artificial Intelligence & Applications 2010;1(2):23-28.
9. Singh V, Misra AK. A genetic algorithm for k-means clustering. International Journal of Emerging Technologies in Computational and Applied Sciences 2014;7:359-364.