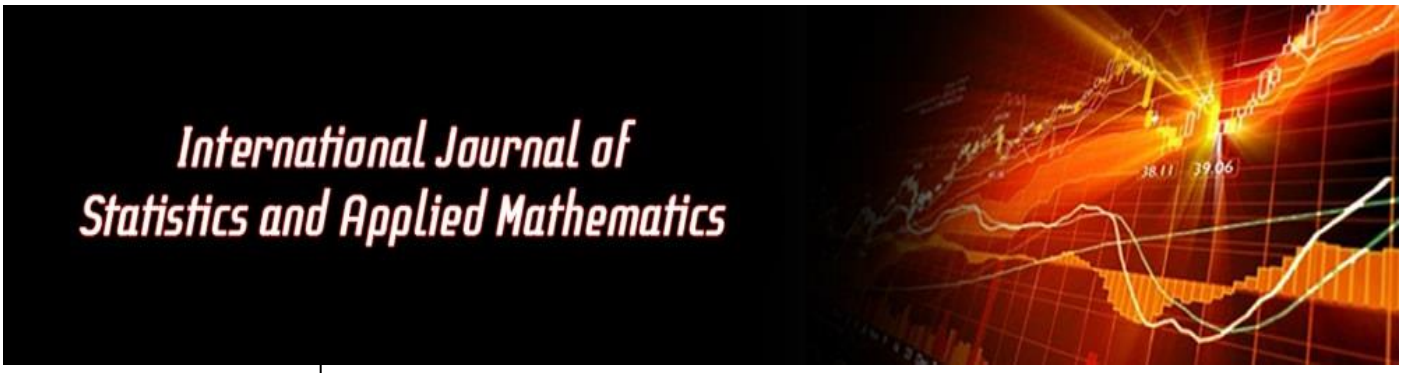
International Journal of Statistics and Applied Mathematics 2021; 6(5): 147-150



ISSN: 2456-1452
Maths 2021; 6(5): 147-150

www.mathsjournal.com
Received: 19-07-2021
Accepted: 21-08-2021

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 approach to cluster analysis

Nisha Sumbherwal and BK Hooda

Abstract
In this paper, performance of Genetic Algorithm based clustering method has been compared with conventional clustering methods that are K-means and Ward's clustering methods. The cluster quality has been compared using three cluster validity indices that are Calinski-Harabasz, Dunn and Average Silhouette Width. The results showed that genetic algorithm based clustering method performed better than other clustering methods under all the three cluster validity measures.

Keywords: genetic algorithm, average silhouette width, Calinski-harabasz index, Dunn index and mixed variables

1. Introduction
Genetic Algorithm (GA) was introduced by Holland (1975) [3] and later it was described by Goldberg (1989) [2]. 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. This is an optimization approach using GA to find a near global or global optimal solution for a given fitness function. Genetic algorithms are based on the principle of natural genetics and follow the process of evolution as stated by Charles Darwin. To optimize the compactness of the clusters, conventional partitional clustering techniques such as K-means and Fuzzy c-means employ a greedy search method over the search space. While computationally efficient, these algorithms suffer from the following drawbacks:

1. They get stuck at some local optimum solution, depending on the choice of the initial cluster centers.
2. They optimize a single cluster validity index and may not therefore cover the various characteristics of the datasets.
3. There is a need to define the number of clusters a priori.

The global optimization tool, GAs, can be used to overcome the problem of local optima to obtain the global optimum solution of the selected cluster validity measure.
Genetic algorithm based clustering methods present in the literature include Genetic K-means algorithm (GKA) given by Krishna & Murty in 1999 [4], GA-clustering suggested by Maulik & Bandyopadhyay in 2000, Fast Genetic K-means Algorithm (FGKA) proposed by Lu *et al*. (2004a) [8], Incremental Genetic K-means Algorithm (IGKA) introduced by Lu *et al*. (2004b) [9], Hybrid GA-based clustering (HGACLUS) introduced by Liu *et al*. in 2004 [7], Efficient GA based Clustering Technique proposed by Lin *et al*. in 2005 [5], Genetic Weighted K-means Algorithm (GWKMA) developed by Wu *et al*. in 2005 [12], Genetic algorithm for clustering gene expression data (GenClust) introduced by Gesu *et al*., 2005 [1], Improved hybrid genetic clustering algorithm given by Liu *et al*., 2006 [6] and Hybrid clustering technique that combines a novel genetic algorithm with K-Means was proposed by Rahman & Islam in 2014 [11].

2. Methodology
The data on pearl millet crop during kharif season 2018-19 respectively was obtained from the Department of Genetics and Plant breeding at CCS Haryana Agriculture University, Hisar,

Haryana. The dataset contained 60 genotypes each with 8 attributes. Among them 5 variables are numeric and 3 are categorical. The numeric variables were days to fifty percent flowering**,** plant height, head length**,** head diameter and effective tillers per plant and the categorical variables were agronomic score, blast score and compactness score.

2.1 Genetic Algorithm

The Genetic Algorithm begins by initializing a population of potential solutions encoded into strings called chromosomes. Each solution has some fitness value based on which the fittest parents that would be used for reproduction are found (survival of the fittest). The new generation is created by applying genetic operators such as selection (based on natural selection to create the mating pool), crossover (exchange of information among parents) and mutation (sudden small change in a parent) on selected parents. Thus the quality of the population is improved as the number of generation increases. The process continues until some specific criterion is met or the solution converges to some optimized value.

The workflow of Genetic algorithm is presented in form of flowchart (Maulik *et al*., 2011) [10] in Fig.1.

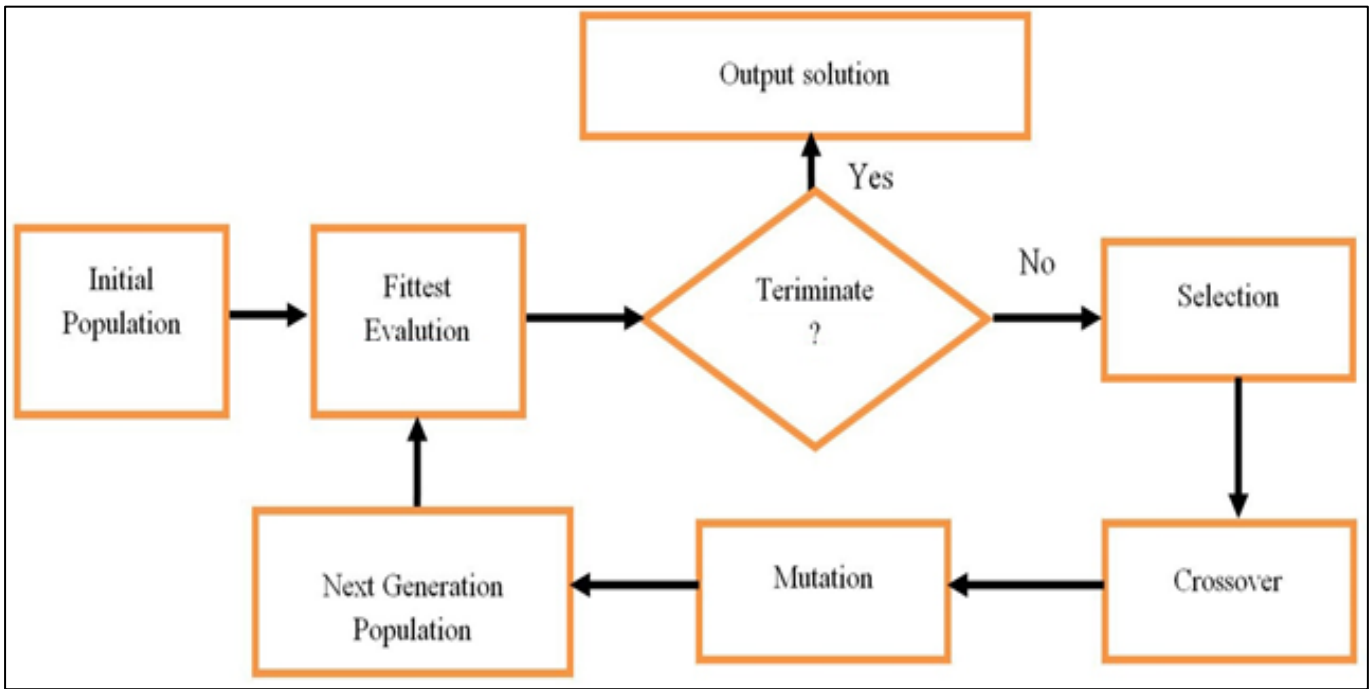


Fig 1: Flowchart of Genetic Algorithm

3. Results and Discussion

Table 1: Comparison of GA based clustering method with K-means and Ward's clustering methods

Indices	Clustering methods		
	GA	K-means	Ward's
Average Silhouette Width	0.428	0.207	0.190
Dunn	0.340	0.144	0.175
Calinski-Harabasz	18.482	18.462	17.158

The result of the performances of different clustering methods on numeric variables data are presented in Table 1. For clustering numeric data two clustering methods i.e. K-means and Ward's methods are used and their performances have been compared with GA. The above table reveled that values of Average Silhouette Width, Dunn index and Calinski-Harabasz index are 0.428, 0.340 and 18.482 respectively for GA. The values obtained from K-means clustering method are 0.207, 0.144 and 18.462 respectively and for Ward's method the values are 0.190, 0.175 and 17.158 respectively. Thus it can be clearly seen that genetic algorithm based clustering method outperformed than K-means and Ward's methods.

Figure 2, 3 and 4 depict the cluster plots obtained from GA, K-means and Ward's clustering methods respectively.

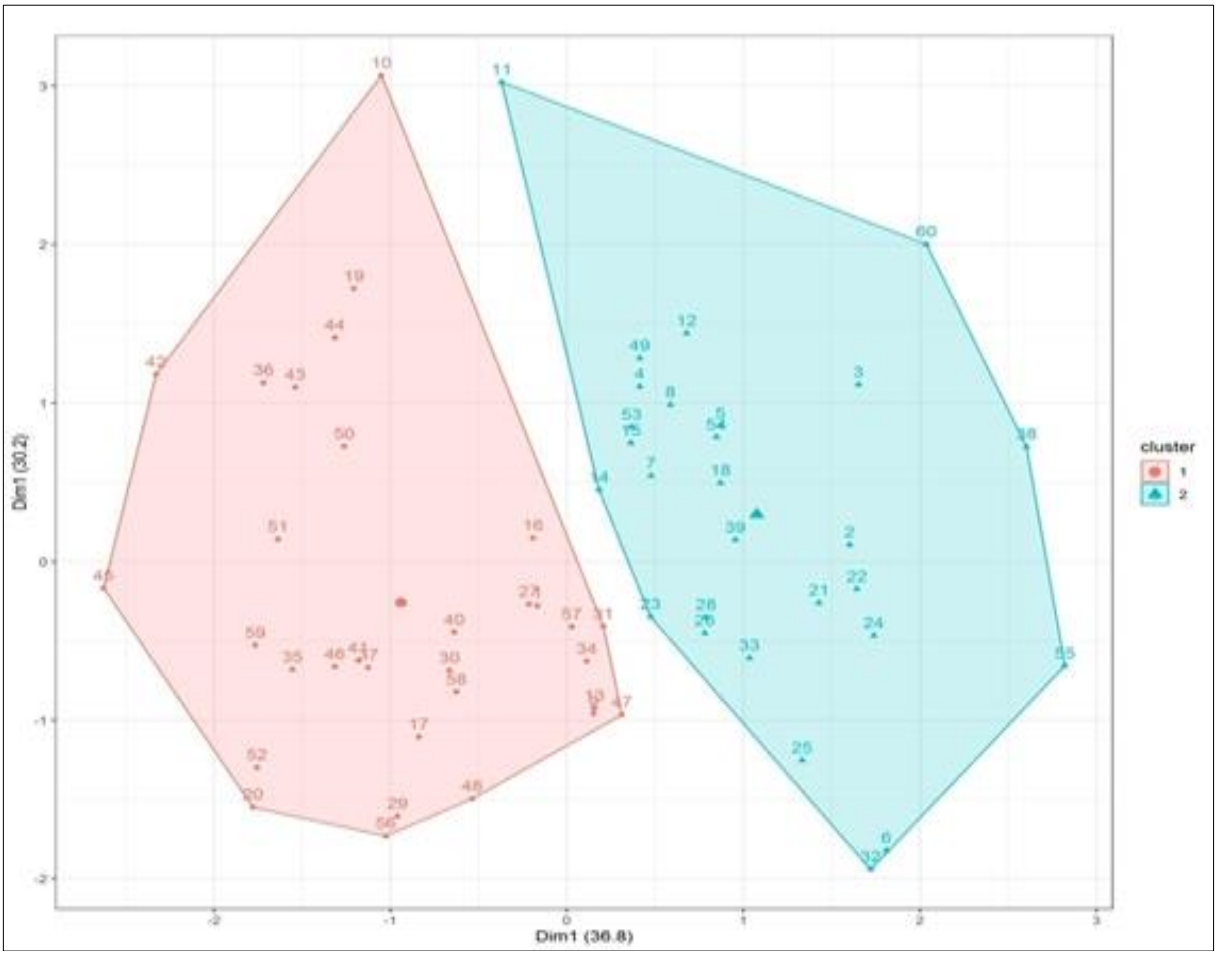


Fig 2: Cluster plot of GA clustering method

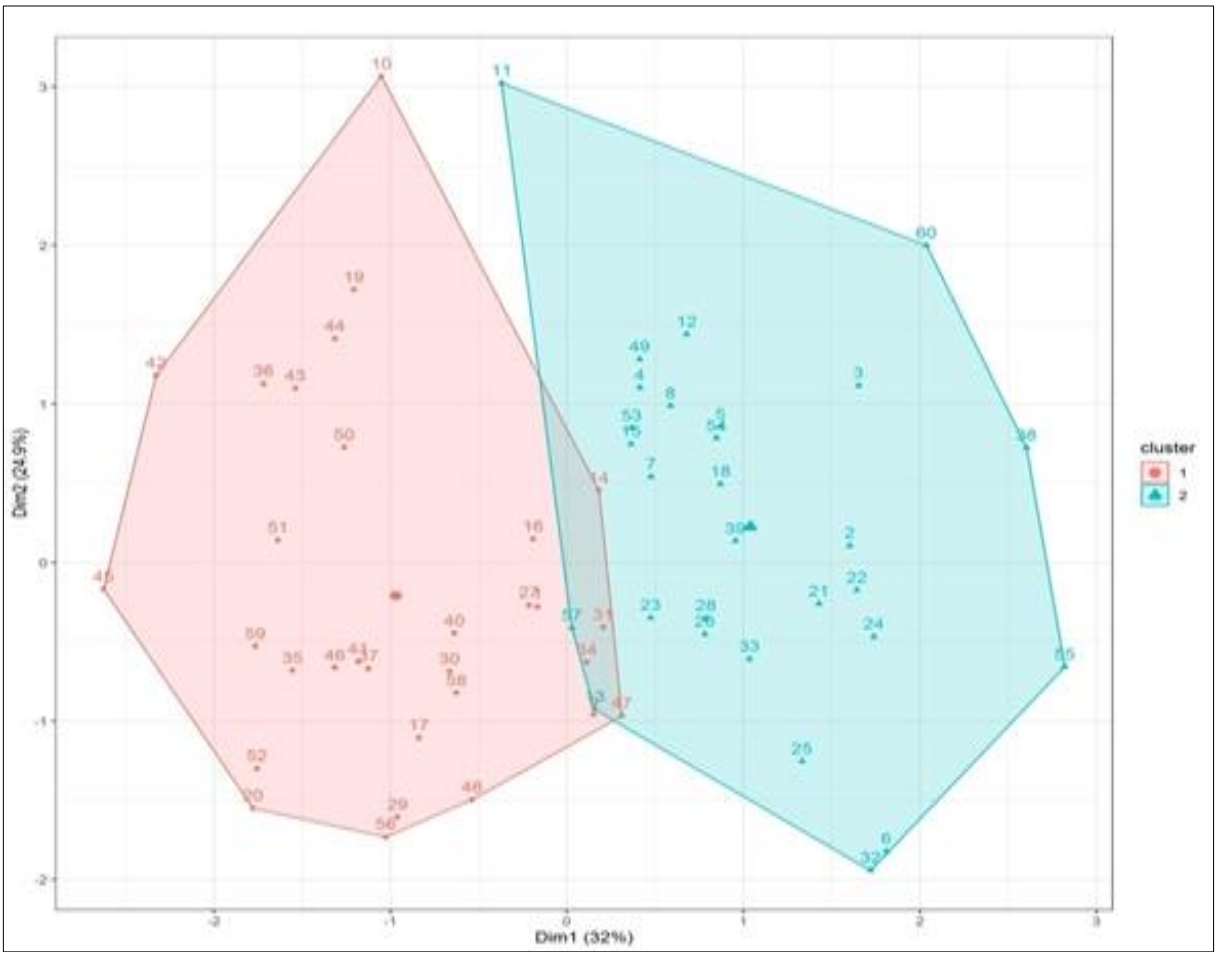


Fig 3: Cluster plot of K-means clustering method

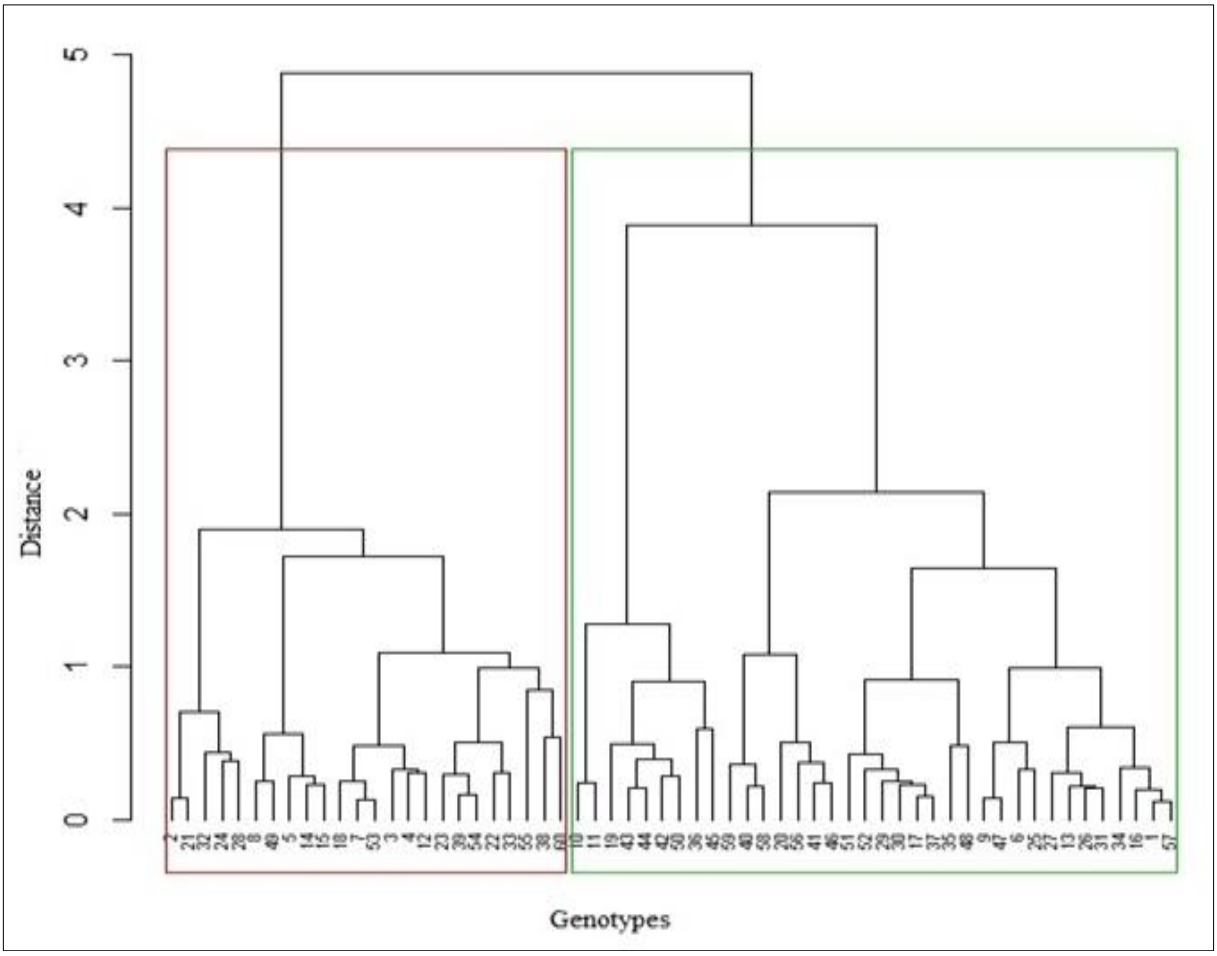


Fig 4: Cluster dendrogram of Ward's method

4. Conclusions

Genetic algorithms are based on the principle of natural genetics and follow the process of evolution as stated by Charles Darwin. To optimize the compactness of the clusters, conventional partitional clustering techniques such as K-means employs a greedy search method over the search space. Also these methods get stuck at some local optimum solution, depending on the choice of the initial cluster centers. Thus, global optimization tool, GAs, can be used to overcome the problem of local optima to obtain the global optimum solution of the selected cluster validity measure. To evaluate the performances of various clustering methods, three fitness functions i.e. Average Silhouette Width, Dunn and Calanski-Harbasz indices were used. It was found that GA performed better than K-means and Ward's clustering methods under all three cluster validity measures.

5. References

1. Gesú VD, Giancarlo R, Bosco GL, Raimondi A, Scaturro D. GenClust: A genetic algorithm for clustering gene expression data. BMC bioinformatics 2005;6(1):1-11.
2. Goldberg DE. Genetic algorithms in search, optimization and machine learning. Addison Wesley Publishing Company, Boston 1989.
3. Holland JH. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor 1975.
4. Krishna K, Murty M. Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics 1999;29:433-439.
5. Lin HJ, Yang FW, Kao YT. An efficient GA-based clustering technique. Journal of Applied Science and Engineering 2005;8(2):113-122.
6. Liu Y, Peng J, Chen K, Zhang Y. An improved hybrid genetic clustering algorithm. In Hellenic Conference on Artificial Intelligence, Springer, Berlin, Heidelberg 2006;2:192-202.
7. Liu YG, Chen KF, Li XM. A hybrid genetic based clustering algorithm. In Proceedings of International Conference on Machine Learning and Cybernetics 2004;3:1677-1682.
8. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ. FGKA: A fast genetic k-means clustering algorithm. In Proceedings of ACM symposium on applied computing 2004a, 622-623.
9. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ. Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC bioinformatics 2004b;5(1):1-10.
10. Maulik U, Bandyopadhyay S, Mukhopadhyay A. Multiobjective genetic algorithms for clustering: applications in data mining and bioinformatics. Springer Science & Business Media 2011.
11. Rahman MA, Islam MZ. A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowledge-Based Systems 2014;71:345-365.
12. Wu FX, Kusalik AJ, Zhang WJ. Genetic Weighted K-means for Large-Scale Clustering Problems. In FLAIRS Conference 2005;5:864-865.