

International Journal of Statistics and Applied Mathematics



ISSN: 2456-1452

Maths 2025; 10(3): 93-96

© 2025 Stats & Maths

<https://www.mathsjournal.com>

Received: 17-01-2025

Accepted: 21-02-2025

Hemant Poonia

Department of Mathematics and
Statistics, CCS HAU, Hisar,
Haryana, India

Ramavtar

Oilseeds Section, Department of
Genetics & Plant Breeding, CCS
HAU, Hisar, Haryana, India

BK Hooda

Department of Mathematics and
Statistics, CCS HAU, Hisar,
Haryana, India

Ramniwas

Computer Section, COBS & H,
CCS HAU, Hisar, Haryana,
India

A hybrid approach using genetic algorithm with k-means in clustering of Indian mustard genotypes

Hemant Poonia, Ramavtar, BK Hooda and Ramniwas

DOI: <https://www.doi.org/10.22271/maths.2025.v10.i3b.2010>

Abstract

The study was carried out to check the applicability of a hybrid approach using a Genetic Algorithm with K-means for clustering of Indian mustard genotypes. The secondary data on growth and yield attributes of 80 Indian mustard genotypes was used for the identification of patterns and best genotypes for plant breeders. A hybrid clustering method was used for getting improved clusters by combining two clustering methods genetic algorithm and k-means. Initially, cluster centres were obtained using the genetic algorithm clustering method, and then cluster centres were used as input for the k-means procedure. The cluster's quality was measured in terms of the total within the sum of squares and the sum of squares ratio. It was concluded that an improved cluster can be obtained if the hybrid clustering method is applied to the subset of selected variables.

Keywords: Indian mustard genotypes, K-means and Genetic Algorithm, hybrid clustering, variables

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 partitioning clustering techniques such as K-means and Fuzzy c-means employ a greedy search method over the search space. It proceeds as follows: Firstly, K objects are randomly selected, where each of which initially represents a cluster mean or centre. Secondly, for each of the remaining objects, based on the distance between the object and the centre, an object is assigned to the cluster to which it is the most similar. Finally, the new centre for each cluster is computed, and the process iterates until the criterion function converges. Typically, the squared-error criterion is used.

K-means works well when the clusters are compact clouds that are rather well separated from one another by adopting the squared-error criterion. On the other hand, when clusters have widely different sizes, densities, or non-spherical shapes, it is difficult to detect natural clusters (Jiang and Xu, 2004) [5]. It will divide the objects in one cluster into two or more clusters to attempt to minimize the square-error criterion. In addition, the optimal cluster corresponds to the extremum when applying this criterion to evaluate the clustering results. Moreover, the algorithm will terminate at the local optimum if the results of initialization are exactly near the local minimal point. So, this paper intends to present a hybrid GA with a K-means cluster algorithm that efficiently eliminates this drawback of K-means.

Genetic algorithm-based clustering methods present in the literature include the Genetic K-means algorithm (GKA) given by Krishna & Murty (1999) [6].

Corresponding Author:

Hemant Poonia

Department of Mathematics and
Statistics, CCS HAU, Hisar,
Haryana, India

Hybrid GA-based clustering (HGACCLUS) was introduced by Liu *et al.* in 2004 ^[7], and the Efficient GA-based Clustering Technique was proposed by Lin *et al.* in 2005 ^[7]. Genetic algorithm for clustering gene expression data (GenClust) introduced by Gesu *et al.*, 2005. An improved hybrid genetic clustering algorithm given by Liu *et al.*, 2006 ^[8] and a Hybrid clustering technique that combines a novel genetic algorithm with K-Means was proposed by Rahman & Islam in 2014 ^[12]. Malki *et al.* (2016) ^[10] presented a hybrid version of the k-means algorithm with GAs that efficiently eliminates this empty cluster problem. Results of simulation experiments using several data sets prove our claim. Sumbherwal and Hooda (2021) ^[13] compared the performance of the genetic algorithm-based clustering method with conventional clustering methods, i.e. K-means and Ward's clustering methods, and observed that the genetic algorithm-based clustering method performed better than other clustering methods under all three cluster validity measures. Hooda *et al.* (2021) ^[13] considered the cluster analysis based on fuzzy sets for grouping of wheat genotypes using data on morphological characters. The performance of fuzzy clustering has also been examined in relation to the commonly used K-means clustering method for identifying clusters of wheat genotypes. Poonia *et al.* (2022) ^[11] compared the performance of the two clustering techniques,

k-means (KM) and Fuzzy c-means (FCM). These two methods were applied to secondary data on growth and yield attributes of wheat genotypes and examined the performance of both clustering approaches.

Materials and Methods

Secondary data on growth and yield attributes of 80 Indian Mustard genotypes recorded during 2020-21 were used to explore a hybrid clustering procedure that can combine k-means and a Genetic algorithm-based clustering method to improve clusters. The data used in this experiment were obtained from the Oilseeds Section, Department of Genetics and Plant Breeding CCS HAU, Hisar.

To explore the hybrid clustering method, which combines k-means and Genetic Algorithm based clusters for getting improved clusters, the following procedure was adopted:

- Initially, a Genetic Algorithm was used to obtain cluster centres. These cluster centres were then used as input for the k-means procedure, and the results were compared with those of the usual k-means method.
- A Genetic Algorithm was applied to select variables based on the Maximum RV-Coefficient Criterion, and then clusters were obtained and compared.

Results and Discussion

The detail of the genotypes is given below

Sr. No	Genotype	Sr. No	Genotype	Sr. No	Genotype	Sr. No	Genotype
1	M-9	21	M-52	41	NPJ-231	61	DRMRIJ-20-13
2	M-14	22	M-54	42	NPJ-232	62	DRMRIJ-20-19
3	M-15	23	M-56	43	NPJ-233	63	DRMRIJ-20-23
4	M-18	24	Sel-1	44	NPJ-234	64	DRMRIJ-20-26
5	M-24	25	T 6342	45	PM-25	65	DRMRIJ-20-35
6	M-25	26	RH-0023	46	NPJ-241	66	DRMRIJ-20-56
7	Giriraj	27	PM-25	47	NPJ-242	67	DRMRIJ-20-64
8	M-26	28	RH-9811	48	NPJ-243	68	DRMRIJ-20-65
9	M-29	29	RH-0204	49	NPJ-244	69	DRMRIJ-20-72
10	M-32	30	M-0305	50	NPJ-196	70	Giriraj
11	PM-25	31	RH-345	51	IC-277699	71	DRMRIJ-20-75
12	M-35	32	RH-0504	52	IC-355399	72	DRMRIJ-20-77
13	M-38	33	RH-1303	53	EC-597318	73	DRMRIJ-20-85
14	M-39	34	RH-1378	54	LES-57	74	DRMRIJ-20-91
15	M-45	35	RH-1224	55	Giriraj	75	DRMRIJ-20-109
16	M-46	36	RH-1368	56	PDZ-6	76	PM-25
17	M-48	37	RH-1371	57	Pusa Kishan	77	DRMRIJ-20-114
18	M-50	38	Giriraj	58	DRMRIJ-20-6	78	DRMRIJ-20-137
19	M-51	39	RH(00)-7003	59	DRMRIJ-20-9	79	DRMRIJ-20-145
20	Giriraj	40	RH-19-7	60	PM-25	80	DRMRIJ-20-160

Variables on which observations were recorded

1.	Days to Flowering (DF)
2.	Days to Maturity(DM)
3.	Plant Stand(PS)
4.	Plant Height(PH)
5.	No. Primary Branches(NPB)
6.	No Secondary Branches (NSB)
7.	Main Shoot Length (MSL)
8.	Siliqua on Main Shoot (SOMS)
9.	Siliqua Length (SL)
10.	Seeds per Siliqua (SPS)
11.	Seed Yield per Plant (SYPP)
12.	1000-seed Weight (TSWT)

The k-means clustering and genetic algorithm-based methods were used to find clusters of the 80 Indian mustard genotypes based on all 12 variables and the best subsets of 3 variables. The k-means () and Gama () functions in the r package were used to obtain clusters. Cluster centres obtained by genetic

algorithm-based clustering were used as input for the hybrid method, thereafter called the GAK-means method. The resulting clusters and their sizes are given in Table 1, along with the within-cluster sum of squares (WSS) for each cluster.

Table 1: Cluster sizes and WSS obtained using k-means, genetic algorithm based on clustering and hybrid methods

Clustering Method	Subset size	Cluster	Cluster Size	WSS
K-means	12(ALL)	C1	25	235.60
		C2	26	238.23
		C3	29	242.06
GA	12(ALL)	C1	78	-
		C2	1	-
		C3	1	-
GAK-means	12(ALL)	C1	26	238.23
		C2	25	235.60
		C3	29	242.06
Variables Subset-Based Clustering				
K-means	3 (1,5, 11)	C1	36	61.86
		C2	31	50.36
		C3	13	19.09
GA	3 (1,5, 11)	C1	23	40.83
		C2	45	77.74
		C3	12	15.58
GAK-means	3 (1,5, 11)	C1	16	27.15
		C2	36	54.90
		C3	28	43.92

Cluster memberships of different genotypes and means are given in Tables 2 & 3 below:

Table 2: Genotype in clusters based on best subset 3 variables and all the variables allocated using k-means, genetic algorithm and GAK-means method

Clustering Method	Variables Used	Cluster	Sr. Nos. of Genotypes
K-means	12 (All)	C1	1, 2, 3, 6, 8, 9, 10, 15, 16, 21, 25, 28, 33, 40, 41, 52, 53, 54, 57, 58, 59, 62, 65, 69, 73
		C2	11, 20, 22, 24, 27, 29, 31, 36, 37, 42, 43, 45, 46, 48, 49, 50, 51, 60, 64, 66, 71, 72, 74, 75, 76, 78
		C3	4, 5, 7, 12, 13, 14, 17, 18, 19, 23, 26, 30, 32, 34, 35, 38, 39, 44, 47, 55, 56, 61, 63, 67, 68, 70, 77, 79, 80
GA	12 (All)	C1	1 to 51, 53 to 80 (excluding 52 and 24)
		C2	52
		C3	24
GAK-means	12 (All)	C1	11, 20, 22, 24, 27, 29, 31, 36, 37, 42, 43, 45, 46, 48, 49, 50, 51, 60, 64, 66, 71, 72, 74, 75, 76, 78
		C2	1, 2, 3, 6, 8, 9, 10, 15, 16, 21, 25, 28, 33, 40, 41, 52, 53, 54, 57, 58, 59, 62, 65, 69, 73
		C3	4, 5, 7, 12, 13, 14, 17, 18, 19, 23, 26, 30, 32, 34, 35, 38, 39, 44, 47, 55, 56, 61, 63, 67, 68, 70, 77, 79, 80
Variables Subset Based Clustering			
K-means (1,5,11)	3	C1	4, 7, 13, 14, 17, 18, 20, 22, 25, 26, 29, 30, 31, 32, 34, 38, 42, 43, 44, 46, 47, 51, 55, 56, 61, 63, 67, 68, 70, 71, 72, 74, 75, 77, 79, 80
		C2	1, 2, 3, 6, 8, 9, 10, 12, 15, 16, 19, 21, 23, 28, 33, 35, 39, 40, 41, 52, 53, 54, 57, 58, 59, 62, 64, 65, 66, 69, 73
		C3	5, 11, 24, 27, 36, 37, 45, 48, 49, 50, 60, 76, 78
GA (1,5,11)	3	C1	4, 5, 7, 9, 12, 14, 17, 18, 19, 26, 30, 32, 34, 39, 44, 53, 62, 65, 67, 68, 77, 79, 80
		C2	1, 2, 3, 8, 10, 13, 15, 16, 20, 21, 22, 23, 25, 28, 29, 31, 33, 35, 38, 40, 41, 42, 43, 46, 47, 48, 51, 52, 54, 55, 56, 57, 58, 59, 61, 63, 64, 66, 69, 70, 71, 72, 73, 74, 75
		C3	6, 11, 24, 27, 36, 37, 45, 49, 50, 60, 76, 78
GAK-means (1,5,11)	3	C1	5, 6, 11, 24, 27, 36, 37, 45, 48, 49, 50, 60, 62, 72, 76, 78
		C2	1, 2, 3, 8, 10, 15, 16, 21, 22, 23, 25, 28, 29, 31, 33, 35, 38, 40, 41, 42, 46, 51, 52, 53, 54, 56, 57, 58, 59, 64, 66, 69, 71, 73, 74, 75
		C3	4, 7, 9, 12, 13, 14, 17, 18, 19, 20, 26, 30, 32, 34, 39, 43, 44, 47, 55, 61, 63, 65, 67, 68, 70, 77, 79, 80

Table 3: Cluster means obtained using k-means, genetic algorithm, and genetic algorithm based k-means method

Method	Cluster	Characters											
		DF	DM	PS	PH	NPB	NSB	MSL	SOMS	SL	SPS	SYPP	TSWT
k-means	Clusters Based on all 12 variables												
	C1	51.1	139.0	20.1	192.4	6.1	14.9	64.0	49.2	3.6	11.2	15.9	4.4
	C2	49.1	138.4	19.7	170.9	4.7	12.2	71.5	47.2	4.2	13.4	16.2	4.9
	C3	50.7	139.4	19.6	193.2	5.3	13.6	75.0	54.4	3.9	13.6	21.0	5.1
GA	C1	50.4	138.9	19.8	187.1	5.4	13.6	71.0	50.4	3.9	12.8	17.9	4.8

	C2	52.0	143.0	18.0	129.7	4.8	11.3	22.3	64.3	3.3	11.1	13.3	5.6
	C3	45.0	136.0	19.0	134.0	5.2	14.3	75.7	44.3	4.1	15.1	14.7	3.9
GAK-means	C1	49.1	138.4	19.7	170.9	4.7	12.2	71.5	47.2	4.2	13.4	16.2	4.9
	C2	51.1	139.0	20.1	192.4	6.1	14.9	64.0	49.2	3.6	11.2	15.9	4.4
	C3	50.7	139.4	19.6	193.2	5.3	13.6	75.0	54.4	3.9	13.6	21.0	5.1
Clusters Based on subset of 3 variables (1, 5, 11)													
k-means (1,5,11)	C1	50.7	139.6	19.5	186.5	4.7	12.6	74.5	52.0	4.0	13.6	19.6	5.1
	C2	51.1	139.2	20.0	190.9	6.1	14.9	64.9	49.7	3.7	11.7	16.5	4.5
	C3	47.4	136.7	20.1	171.2	5.2	12.9	72.4	48.2	4.1	13.2	16.3	4.7
GA (1,5,11)	C1	50.6	139.0	19.6	190.7	5.8	15.1	73.0	53.1	3.6	13.5	21.8	4.9
	C2	51.0	139.6	19.7	186.8	5.1	12.9	68.5	49.8	4.0	12.3	16.3	4.8
	C3	47.3	136.4	20.3	172.1	5.3	13.1	72.6	47.9	4.0	13.1	16.2	4.6
GAK-means (1, 5, 11)	C1	47.7	137.3	20.1	172.6	5.3	13.0	72.0	48.1	4.0	13.1	16.6	4.7
	C2	51.2	139.4	19.9	186.8	5.4	13.6	67.2	49.4	3.9	11.9	15.6	4.7
	C3	50.7	139.4	19.5	191.8	5.3	13.9	73.7	53.2	3.8	13.8	21.5	5.1

The quality of the clusters was measured in terms of total WSS and in terms of the percent sum of squares ratio (BSS/TSS), where, higher value indicates better clusters. The

values of these measures along with other sums of squares are given below in Table 4.

Table 4: Within sum of squares and between sum of squares

Clustering Methods	BSS	Total_WSS	TSS	(BSS/TSS)*100
K-means	232.11	715.89	948	24.48
GAK-means	232.11	715.89	948	24.48
K-means (1, 5, 11)	105.69	131.31	237	44.6
GA (1,5,11)	102.87	134.13	237	43.4
GAK-means (1, 5, 11)	111.04	125.96	237	46.9

Table 4 shows that the quality of clusters based on the sum of squares and their ratio is at par in case of all variables as input. However, clusters based on subset of variables are of better quality for the GAK-means or the hybrid method. Thus, we conclude that improved cluster can be obtained if GAK-means method is applied on the subset of selected variables.

Conclusion

A hybrid clustering method was used for getting improved clusters by combining two clustering methods (genetic algorithm and k-means) called GAK-means. Initially, cluster centres were obtained using the genetic algorithm clustering method. These cluster centres were then used as input for the k-means procedure, and the results were compared with those of the usual k-means method. Cluster's quality was measured in terms of the total within the sum of squares and the sum of squares ratio (BSS/WSS), and concluded that an improved cluster could be obtained if the GAK-means method is applied to the subset of selected variables.

References

1. Gesù 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. Hooda BK, Singh V, Poonia H, Kumar M. Morphological characters based clustering of wheat genotypes: A fuzzy set approach. Environment and Ecology. 2021;39(4A):1052-1060.
5. Jiang SY, Xu YM. An efficient clustering algorithm. In: Proceedings of the International Conference on Machine Learning and Cybernetics. 2004;3:1513-1518.
6. Krishna K, Murty M. Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics. 1999;29:433-439.
7. Lin HJ, Yang FW, Kao YT. An efficient GA-based clustering technique. Journal of Applied Science and Engineering. 2005;8(2):113-122.
8. Liu Y, Peng J, Chen K, Zhang Y. An improved hybrid genetic clustering algorithm. In: Proceedings of the Hellenic Conference on Artificial Intelligence. 2006;2:192-202.
9. Liu YG, Chen KF, Li XM. A hybrid genetic based clustering algorithm. In: Proceedings of the International Conference on Machine Learning and Cybernetics. 2004;3:1677-1682.
10. Malki AA, Rizk MM, Shorbagy MAE, Mousa AA. Hybrid genetic algorithm with k-means for clustering problems. Open Journal of Optimization. 2016;5:71-83.
11. Poonia H, Singh V, Hooda BK, Kumar M. Performance comparison of fuzzy clustering and hard clustering (k-means) methods for grouping of wheat genotypes. Environment and Ecology. 2022;40(3D):1886-1892.
12. Rahman MA, Islam MZ. A hybrid clustering technique combining a novel genetic algorithm with k-means. Knowledge-Based Systems. 2014;71:345-365.
13. Sumbherwal N, Hooda BK. Genetic algorithm approach to cluster analysis. International Journal of Statistics and Applied Mathematics. 2021;6(5):147-150.