Environment and Ecology 39 (4A) : 1052—1060, October—December 2021
ISSN 0970-0420

Morphological Characters Based Clustering of Wheat Genotypes : A Fuzzy Set Approach

B. K. Hooda, Vikram Singh, Hemant Poonia, Manoj Kumar

Received 16 September 2021, Accepted 20 October 2021, Published on 14 November 2021

ABSTRACT

Cluster analysis is a method of grouping data with similar characteristics into larger units of analysis. Developments in fuzzy set theory gave rise to the concept of partial membership and have received increasing attention during recent years. Fuzzy set approach is based on the premise that key elements in human thinking are not just numbers but can be approximated to classes of objects in which the transition from membership to non membership is gradual rather than abrupt. Cluster analysis using fuzzy sets provides a powerful clustering method but it has not been much used for grouping genotypes by plant breeders. Therefore, in the present study, cluster analysis based on fuzzy sets has been considered 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. It has been observed that fuzzy C-Means clustering method provides more uniform distribution of the wheat genotype among various clusters as compared to the K-Means method. Also, the average inter cluster distance was observed to be more in fuzzy C-Means method, indicating better group separation for wheat genotypes.

Keywords Clustering, Wheat genotypes, Fuzzy C-Means, K-Means, Morphological characters.

INTRODUCTION

In the field of software data analysis is considered as a very useful and important tool as the task of processing large volume of data is rather tough and it has accelerated the interest of application of such analysis. It also makes data description possible by means of clustering visualization, association and sequential analysis. Data clustering is primarily a method of data description which is used as a common technique for data analysis in various fields like machine learning, data mining, pattern reorganization, image analysis and bio-informatics. Cluster analysis is also recognized as an important technique for classifying data, finding clusters of a dataset based on similarities in the same cluster and dissimilarities between different

B. K. Hooda, Hemant Poonia*, Manoj Kumar
Department of Mathematics and Statistics, CCS HAU-Hisar 125004, Haryana, India

Vikram Singh
Assistant Plant Breeder, Dept. of Genetics & Plant Breeding, CCS HAU-Hisar 125004, Haryana, India
Email:pooniahemant80@gmail.com
*Corresponding author

clusters (Rao and Vidyavathi 2010). Putting each point of the dataset to exactly one cluster is the basic of the conventional clustering method whereas clustering algorithm actually partitions unlabeled set of data into different groups according to the similarity. As compare to data classification, data clustering is considered as an unsupervised learning process which does not require any labelled dataset as training data and the performance of data clustering algorithm is generally considered as much poorer. Although data classification is better performance oriented but it requires a labelled dataset as training data and practically classification of labelled data is generally very difficult as well as expensive. As such there are many algorithms that are proposed to improve the clustering performance. Clustering is basically considered as classification of similar objects or in other words, it is precisely partitioning of datasets into clusters so that data in each cluster shares some common trait. The hierarchical, partitioning and mixture model methods are the three major types of clustering processes that are applied for organizing data. The choice of application of a particular method generally depends on the type of output desired, the known performance of the method with particular type of data, available hardware and software facilities and size of the dataset (Rao and Vidyavathi 2010).

K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points. Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters. Each cluster is characterized by its center point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. But use of several replicates with random starting point leads to a solution i.e. a global solution (Jain *et al.* 1999, Han and Kamber 2006, Hui *et al*. 2009). In a dataset, a desired number of clusters K and a set of K initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co-ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters.

Bezdek (1981) introduced Fuzzy C-Means clustering method in 1981, extend from Hard C-Mean clustering method. FCM is an unsupervised clustering algorithm that is applied to wide range of problems connected with feature analysis, clustering and classifier design. FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition (Yong *et al*. 2004). With the development of the fuzzy theory, the FCM clustering algorithm which is actually based on Ruspini Fuzzy Clustering theory was proposed in 1980's. This algorithm is used for analysis based on distance between various input data points. The clusters are formed according to the distance between data points and the cluster centers are formed for each cluster. Infact, FCM is a data clustering technique (Chen and Zhang 1998) in which a data set is grouped into n clusters with every data point in the dataset related to every cluster and it will have a high degree of belonging (connection) to that cluster and another data point that lies far away from the center of a cluster which will have a low degree of belonging to that cluster.

.

Bora and Gupta (2014) evaluated the performance between K-Means and Fuzzy C-Means algorithms based on time complexity. Velmurugun (2012) has compared the clustering performance of K-Means and Fuzzy C-Means algorithms using different shapes of arbitrary distributed data points and reported that the K-Means performs better than FCM.Simhachalam and Ganesan (2016) presented a comparative study of partition algorithms such as K-Means (KM), Fuzzy C-Means (FCM), Gustafson–Kessel (GK) with different famous real world data sets, liver disorder and wine from the UCI repository. The performance of the three algorithms was analyzed based on the clustering output criteria. Ghosh and Dubey (2013) compared the two important clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms. These algorithms are applied and performance was evaluated on the basis of the efficiency of clustering output.

Table 1a. The detail of the genotypes and variables considered is given below: Observations recorded: 1. Days to heading (50%) 2. Days to maturity 3. Plant height (cm) 4. Flag leaf length (cm) 5. Flag leaf breath (cm) 6. Tillers/meter 7. Spike length 8. Grains/spike 9. 1000 grain weight. 10.Grain yield/plot (7.2 sqm) 11. Grain yield (q/ha).

Sl. No		Genotype	Sl. No		Genotype	Sl. No.		Genotype
1	G1	WH 1105(101)	18	G18	WH 1105(118)	35	G35	WH 1105(135)
2	G2	WH 1105(102)	19	G19	WH 1105(119)	36	G36	WH 1105(136)
3	G3	WH 1105(103)	20	G20	WH 1105(120)	37	G37	WH 1105(137)
4	G4	WH 1105(104)	21	G21	WH 1105(121)	38	G38	WH 1105(138)
5	G5	WH 1105(105)	22	G22	WH 1105(122)	39	G39	WH 1105(139)
6	G6	WH 1105(106)	23	G23	WH 1105(123)	40	G40	WH 1105(140)
7	G7	WH 1105(107)	24	G24	WH 1105(124)	41	G41	WH 1105(141)
8	G8	WH 1105(108)	25	G25	WH 1105(125)	42	G42	WH 1105(142)
9	G9	WH 1105(109)	26	G26	WH 1105(126)	43	G43	WH 1105(143)
10	G10	WH 1105(110)	27	G27	WH 1105(127)	44	G44	WH 1105(144)
11	G11	WH 1105(111)	28	G28	WH 1105(128)	45	G45	WH 1105(145)
12	G12	WH 1105(112)	29	G29	WH 1105(129)	46	G46	WH 1105(146)
13	G13	WH 1105(113)	30	G30	WH 1105(130)	47	G47	WH 1105(147)
14	G14	WH 1105(114)	31	G31	WH 1105(131)	48	G48	WH 1105(148)
15	G15	WH 1105(115)	32	G32	WH 1105(132)	49	G49	WH 1105(149)
16	G16	WH 1105(116)	33	G33	WH 1105(133)	50	G50	WH 1105(150)
17	G17	WH 1105(117)	34	G34	WH 1105(134)	51	G51	WH 1124

MATERIALS AND MATHODS

Clustering is an unsupervised data analysis which is used to partition a set of records or objects into clusters or classes with similar characteristics. The partition is done in such a fashion that most similar (or related) objects are placed together, while dissimilar (or unrelated) objects are placed in different classes or groups. The desired characteristics of clustering methods are ability to deal with different types of

Table 1b. Grouping of wheat genotypes into three clusters.

Clus-ter	K-Means		Fuzzy K-Means		Ward linkage		Ward linkage City block distance		Chebychev distance	
	Cluster size	Wheat genotypes	Cluster size	Wheat genotypes	Cluster size	Wheat genotypes	Cluster size	Wheat genotype	Cluster size	Wheat genotype
I	3	G3, G15, G43	19	G4, G13, G17, G20 G21, G22, G25, G26, G27, G28, G31, G33, G34, G35, G40, G41, G45, G48, G51	11	G4, G17, G20, G21, G22, G25, G26, G28, G31, G35, G51	21	G1, G2, G3, G5, G9, G10, G11, G12, G18,` G23, G24, G27, G29, G32, G33, G34, G36, G37, G41, G42, G50	13	G4, G11, G17, G18 G20, G22, G26, G31, G34, G35, G41, G45, G48
II	17	G2, G10, G11, G13, G16, G18, G23, G24, G27, G29, G32, G33, G34, G38, G39, G41, G46	14	G6, G7, G8,G14, G15, G19, G30, G38, G39, G43, G44, G46, G47, G49	16	G6, G7, G8, G13, G14, G15, G16, G19, G30, G38, G39, G40, G44, G46, G47, G49	6	G8, G15, G38, G39, G43, G49	27	G1, G2, G3, G5, G8, G9, G10, G12, G13, G15, G16, G23, G24, G25, G27, G29, G30, G32, G36, G37, G38, G39, G42, G43, G44, G47, G50
III	31	G1, G4, G5, G6, G7, G8, G9, G12, G14, G17, G19, G20, G21. G22, G25, G26, G28, G30, G31, G35, G36, G37, G40, G42, G44, G45, G47, G48, G49 G50, G51	18	G1, G2, G3, G5, G9, G10, G11, G12, G16, G18, G23, G24, G29 G32, G36, G37, G42, G50	24	G1, G2, G3, G5, G9, 24 G1, G2, G3, G5, G9, G23, G24, G27, G29, G32, G33, G34, G36, G37, G41, G42, G43, G45, G48, G50		G4, G6, G7, G13, G14, G16, G17, G19 G20, G21, G22, G25, G26, G28, G30, G31, G35, G40, G44, G45, G46, G47, G48, G51	11	G6, G7, G14, G19, G21, G28 G33, G40, G46 G49, G51

Table 2a. Cluster centers for 3–Means and Fuzzy 3-Means.

Sl. No	Character	3-Means			Fuzzy 3-Means		
		I	II	III	I	II	III
1	Days to heading (50%)	93	90	91	88.95	91.36	91.78
2	Days to maturity	146	144	144	142.74	144.43	144.78
3	Plant height(cm)	112	112	111	112.32	113.36	109.17
4	Flag leaf length(cm)	30.67	30.22	30.50	29.82	33.36	28.75
5	Flag leaf breath(cm)	2.48	2.25	2.31	2.20	2.50	2.26
6	Tillers/meter	102	115	131	127.00	124.64	119.61
7	Spike length	13.0	13.3	13.2	13.16	14.04	12.59
8	Grains/spike	74.5	62.4	60.7	57.68	64.36	64.92
9	1000 grain weight	37.4	37.9	41.1	41.71	39.49	38.09

Table 2b. Cluster centers for Ward methods.

Sl. No.	Character	Euclidean			City block			Chebychev		
		I	II	III	I	II	III	I	II	III
1	Days to heading (50%)	89.09	91.06	91.00	90.95	92.33	89.88	90.00	91.70	88.64
2	Days to maturity	142.73	144.06	144.38	144.24	145.83	143.17	143.69	144.74	142.18
3	Plant height(cm)	112.18	112.56	110.46	110.19	109.83	113.04	113.69	109.52	113.73
4	Flag leaf length(cm)	29.90	32.52	29.25	29.20	33.33	30.75	28.68	30.12	33.19
5	Flag leaf breath(cm)	2.17	2.44	2.27	2.23	2.57	2.30	2.20	2.31	2.40
6	Tillers/meter	130.64	125.63	119.33	118.38	116.00	130.38	125.69	120.41	129.64
7	Spike length	13.15	14.21	12.55	12.58	14.27	13.48	12.69	13.27	13.64
8	Grains/spike	55.68	62.94	64.42	64.40	66.67	58.88	56.81	63.89	63.82
9	1000 grain weight	42.94	40.28	38.10	37.88	38.33	41.90	41.50	38.27	41.65

Table 3a. Distance matrix for K-Means and Fuzzy clustering methods.

Cluster	Distances between 3-means clusters			Distances between Fuzzy 3-means clusters		
	I	II	III	I	II	III
I	2.48	18.45	32.58	3.24	2.04	2.01
II	18.45	3.05	16.12	2.04	3.83	2.09
III	32.58	16.12	3.02	2.01	2.09	3.04

Table 3b. Distance matrix for Ward method's clustering method.

Cluster	Euclidean distance			City block distance			Chebychev distance		
	I	II	III	I	II	III	I	II	III
I	7.8	9.9	15.4	11.9	6.0	14.3	10.6	10.6	9.5
II	9.9	12.5	8.0	6.0	15.9	17.6	10.6	13.0	11.8
III	15.4	8.0	13.7	14.3	17.6	10.4	9.5	11.8	11.8

attributes with high dimensionality, effective handling of outliers and noise with minimum knowledge, , ability to discover the underlying shapes and structures of the data, scalability, usability and interpretability. Clustering methods are categorized into five different methods: Partitioning method, hierarchical method, data density based method, grid based method and model based or soft computing methods. Among these five methods partition based methods, K-Means (KM), Fuzzy C-Means (FCM) and Ward linkage

Table 4. ANOVA for K-Means, Fuzzy C-Means and Ward's method.

Var	DF v1	DF v2	K-Means		Fuzzy C-Means		Ward's method					
							Euclidean distance		City block distance		Chebychev distance	
			F	Sig	F	Sig	F	Sig	F	Sig	F	Sig
DTH	2	48	2.053	.139	9.203	.000	2.767	.073	2.398	.102	8.197	.001
DTM	2	48	2.706	.077	7.625	.001	3.210	.049	4.294	.019	9.501	.000
PH	2	48	.037	.964	2.953	.062	.843	.437	1.631	.206	4.360	.018
FLL	2	48	.057	.944	16.200	.000	7.557	.001	9.011	.000	10.085	.000
FLB	2	48	2.821	.069	24.652	.000	10.969	.000	13.172	.000	4.864	.012
TPM	2	48	80.775	.000	2.499	.093	5.617	.006	7.859	.001	3.672	.033
SP	2	48	.206	.814	5.758	.006	12.815	.000	9.298	.000	2.659	.080
GPS	2	48	6.610	.003	7.545	.001	7.524	.001	5.110	.010	5.941	.005
TW	2	48	10.376	.000	8.758	.001	15.572	.000	13.041	.000	10.588	.000

Table 5. Membership degree matrix for Fuzzy 3-Means.

Genotype	Membership degree Cluster number			Genotype	Membership degree Cluster number		
	I	II	III		I	II	III
G1	0.16	0.09	0.75	G27	0.5	0.26	0.23
G2	0.15	0.11	0.74	G28	0.63	0.28	0.09
G3	0.18	0.22	0.6	G29	0.18	0.26	0.56
G4	0.68	0.09	0.23	G30	0.2	0.68	0.12
G5	0.05	0.05	0.9	G31	0.78	0.05	0.17
G6	0.19	0.68	0.13	G32	0.26	0.25	0.49
G7	0.34	0.5	0.16	G33	0.43	0.28	0.29
G8	0.17	0.45	0.38	G34	0.69	0.13	0.18
G9	0.08	0.27	0.64	G35	0.64	0.27	0.09
G10	0.19	0.18	0.63	G36	0.15	0.06	0.79
G11	0.37	0.19	0.44	G37	0.26	0.31	0.43
G12	0.07	0.09	0.84	G38	0.08	0.8	0.12
G13	0.66	0.19	0.16	G39	0.1	0.79	0.11
G14	0.12	0.82	0.07	G40	0.49	0.34	0.17
G15	0.14	0.47	0.39	G41	0.55	0.15	0.29
G16	0.34	0.24	0.42	G42	0.35	0.15	0.5
G17	0.83	0.08	0.09	G43	0.16	0.45	0.4
G18	0.31	0.18	0.51	G44	0.33	0.53	0.13
G19	0.21	0.64	0.15	G45	0.34	0.32	0.33
G20	0.77	0.13	0.1	G46	0.15	0.74	0.11
G21	0.6	0.3	0.1	G47	0.07	0.85	0.08
G22	0.82	0.1	0.08	G48	0.45	0.2	0.35
G23	0.09	0.1	0.81	G49	0.16	0.58	0.26
G24	0.18	0.1	0.72	G50	0.11	0.2	0.69
G25	0.49	0.2	0.32	G51	0.72	0.15	0.13
G26	0.77	0.06	0.17				

clustering algorithms are implemented on data sets of wheat genotypes to generate three clusters.

RESULTS AND DISCUSSION

Clustering was done with using characters for Tables 1-6. Grain yield and yellow rust were used for making categories of genotype in relation to yield performance and yellow rust resistance. The K-Means and Fuzzy C-Means clustering and Ward's Hierarchical methods were tired with clusters Figs.1-3. Euclidean distance, Manhatten distance and Chebysive distance functions were used for clustering genotypes in Ward's method. The clustering pattern of genotypes

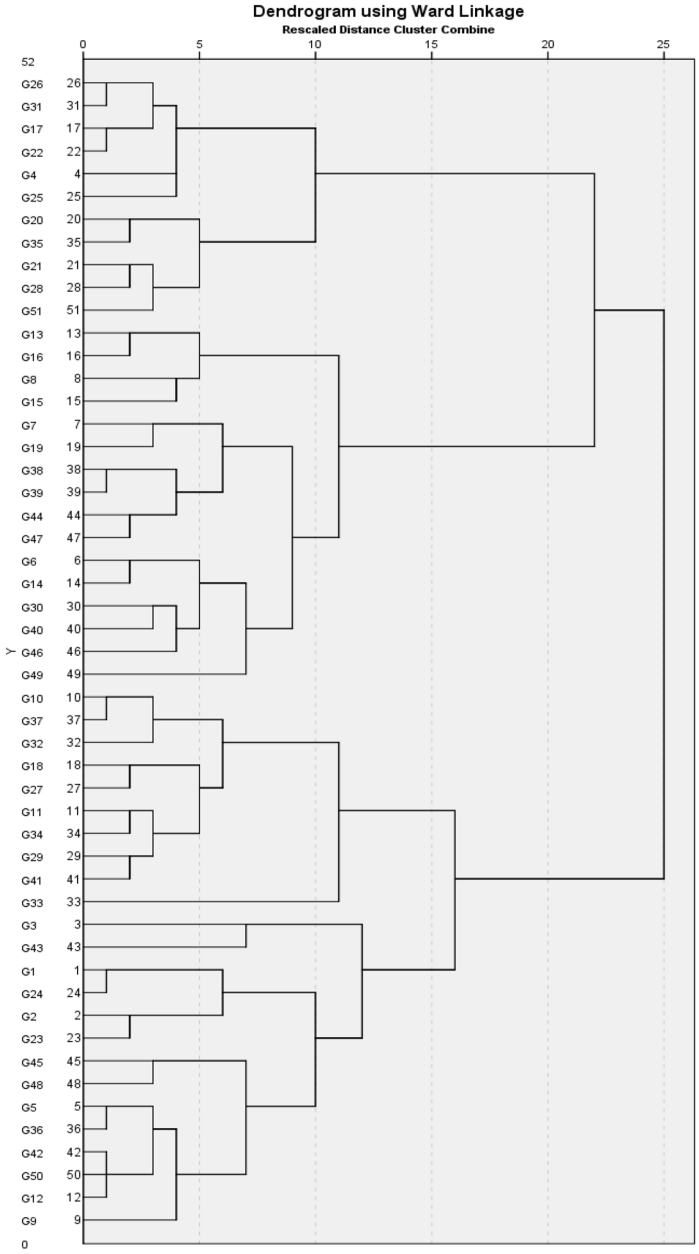

Fig. 1. Dendogram for Ward's method using Euclidean distance function.

using K-Means and Ward's is given in Table 1 and along with the nearest hard clustering solution obtained from the Fuzzy clustering approach.

CONCLUSION

Dendrograms were obtained for Ward clustering methods using 3 distance functions which indicated presence of 3 clusters of genotypes. Sizes and composition of clusters is different by different methods. Out of total 51 genotypes, 13 genotypes had no clear cut assignment in Fuzzy clustering method. The intra and inter-cluster distances were obtained

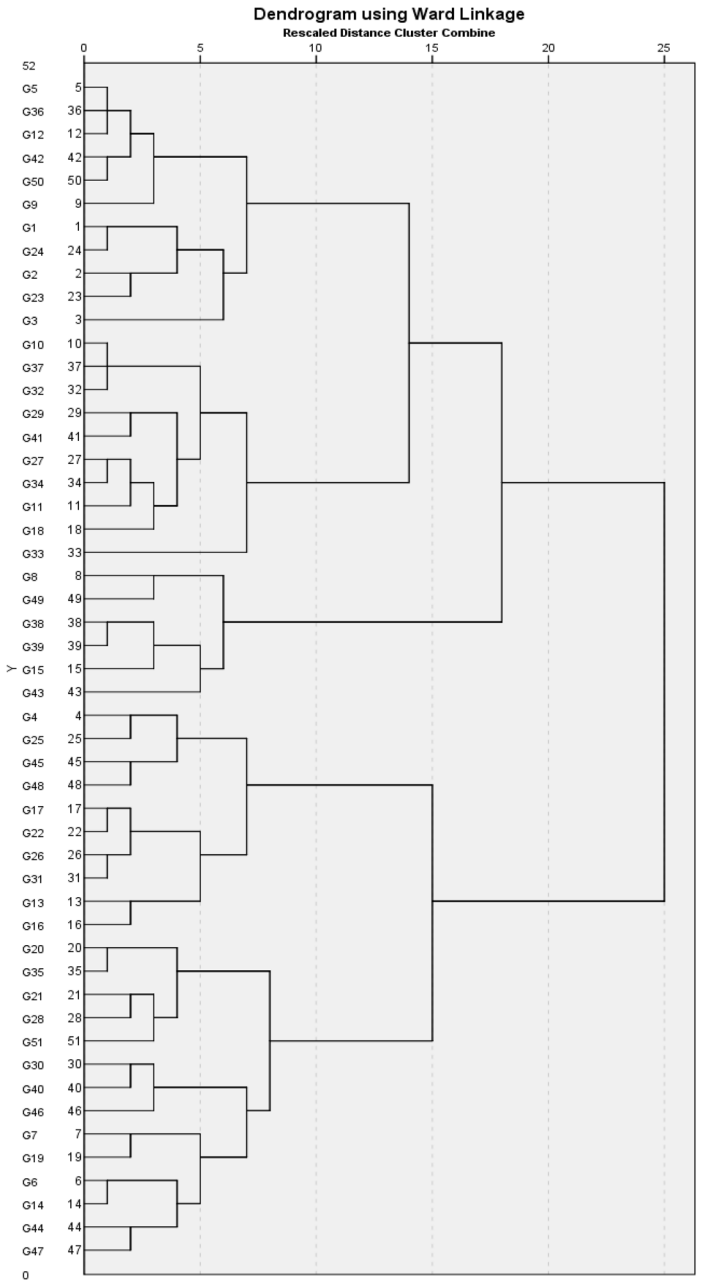

Fig. 2. Dendogram for Ward's method using cityblock distance function.

for all the 3 methods. Intra-cluster distances have found to be smaller in case of K-Means and Fuzzy clustering method giving relatively compact clusters. Inter-cluster distances have been found maximum in case of K-Means clustering indicating most distinct clusters. In Ward clustering with different distance

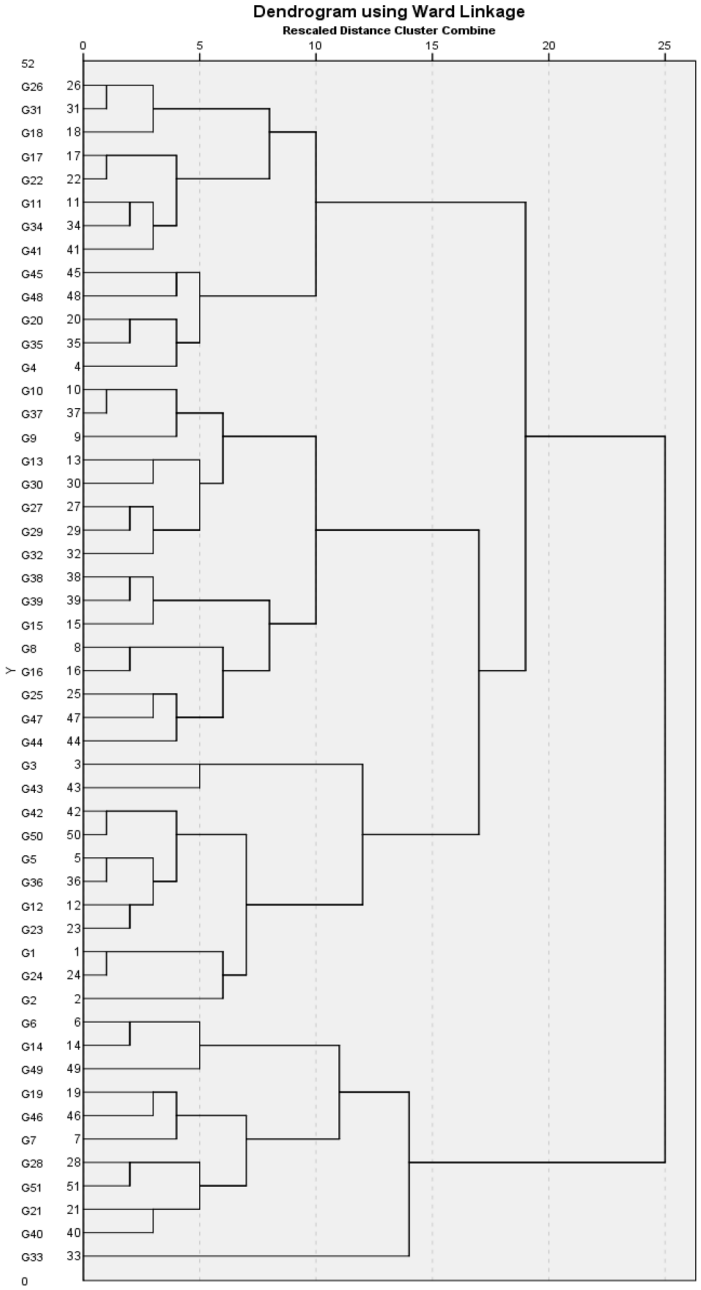

Fig. 3. Dendogram for Ward's method using Chebychev distance function.

functions, there is not much difference in intra and inter-cluster distances. Out of total 51 genotypes, 25, 17, 9 genotypes are giving high, moderate and low yield respectively.

REFERENCES

Bezdek JC (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press,pp 1981.

Table 6. Characterization of wheat genotypes based on yield.

Yield	No. of genotype	Genotype
Low	9	G1, G8, G10, G18, G33, G34 , G38, G41, G43
Moderate	17	G3, G9, G11, G13, G15, G16, G23, G24, G26, G27, G29, G31, G32, G36, G37, G39, G51
High	25	G1, G4, G5, G6, G7, G12, G14, G17, G19, G20, G21, G22, G25, G28, G30, G35, G40, G42, G44, G45, G46, G47, G48, G49, G50

Bora DJ, Gupta AK (2014) A comparative study between Fuzzy clustering algorithm and hard clustering algorithm. *Int J Comput Trends Technol* 10 (2): 108–113.

Chen S, Zhang D (1998) Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. *IEEE Transac Syst Man Cybernetics 3*4: 1907-1916.

Ghosh S, Dubey SK (2013) Comparative analysis of K-Means and Fuzzy C-Means algorithms. *Int J Adv Computer Sci Appl* 4 (4): 35-39.

Han J, Kamber M (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. 2nd edn. New Delhi.

Hui L (2008) Method of image segmentation on high-resolution image and classification for land covers. 4th *Int Conf Natural Comput* 5: 563-566.

Hui X, Wu J, Jian C (2009) K-Means clustering versus validation measures: A data distribution perspective. *IEEE Transac Syst Man Cybernetics* 39 (2): 319-331.

Jain K, Murty MN, Flynn PJ (1999)Data clustering: A review. *ACM Computing Surveys* 31(3): 264-323.

Rao VS,Vidyavathi S (2010) Comparative investigations and per formance analysis of FCM and MFPCM algorithms on Iris data. *Ind J Computer Sci Engg* 1(2): 145-151.

Simhachalam,Ganesan G (2016) Performance comparison of Fuzzy and non-fuzzy classification methods. *Egypt Informatics J* 17 (2): 183-188.

Velmurugun T (2012) Performance comparison between K-Means and Fuzzy C-Means algorithms using arbitrary data points. *Wulfenia J* 9 (8): 234–241.

Yong Y, Chongxun Z, Pan L (2004) A novel Fuzzy C-Means clustering algorithm for image thresholding. *Measurement Sci Rev* 4(1): 11-19.