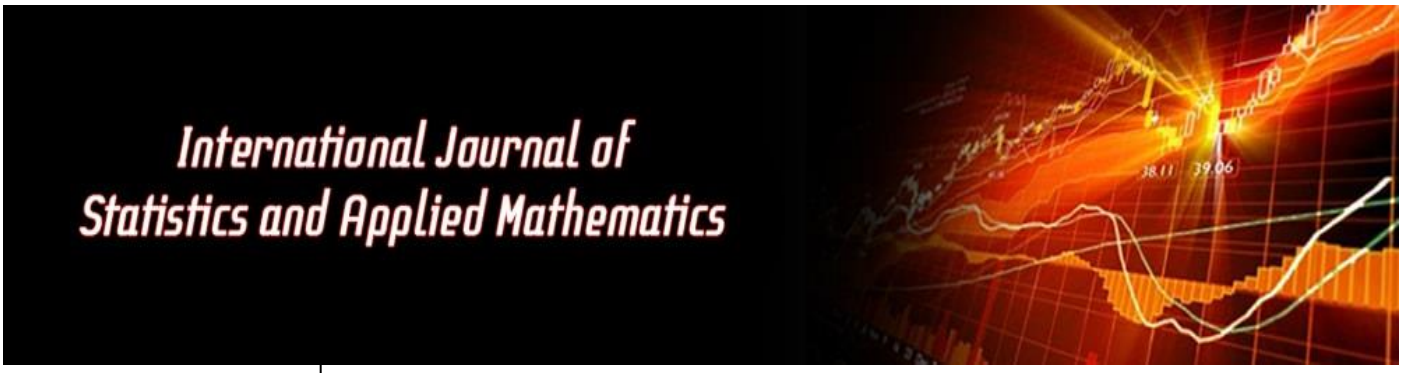
International Journal of Statistics and Applied Mathematics 2020; 5(4): 19-21



ISSN: 2456-1452
Maths 2020; 5(4): 19-21

www.mathsjournal.com
Received: 07-05-2020
Accepted: 09-06-2020

Adarsh VS
M.Sc. Scholar, Department of Agricultural Statistics, College of Agriculture Vellayani, KAU, Kerala, India

Dr. Brigit Joseph
Associate Professor and Head, Department of Agricultural Statistics, College of Agriculture Vellayani, KAU, Kerala, India

Pratheesh P Gopinath
Assistant Professor, Department of Agricultural Statistics, College of Agriculture Vellayani, KAU, Kerala, India

Corresponding Author:
Adarsh VS
M.Sc. Scholar, Department of Agricultural Statistics, College of Agriculture Vellayani, KAU, Kerala, India

Comparison of clustering techniques used in divergence analysis

Adarsh VS, Dr. Brigit Joseph and Pratheesh P Gopinath

Abstract
Cluster analysis has been generally used as an efficient tool in the quantitative estimation of genetic diversity for a breeding programme (divergence analysis), grouping of different genotypes of a particular crop and it is more useful in choosing suitable parents for heterosis breeding. Clustering techniques aims to group data according to common properties. This grouping is often based on the distance between the data. Clustering techniques are divided into hierarchical and non-hierarchical methods according to the fragmentation technique of clusters. There are various clustering methods which come under these techniques viz., single linkage, complete linkage, average linkage, Ward's method and Tocher method. The efficiency of the techniques and methods are given less importance. In this context an attempt was made in this paper for the comparative study of the different clustering techniques and methods in small samples. The cluster validation techniques are used for the efficiency measurements of these methods. The results of the analysis are presented comparatively at the end of the study and which methods are more convenient for data set is explained.

Keywords: Clustering methods, Kmeans, average linkage, Tocher

Introduction
Divergence analysis of genotypes involves evaluation of germplasm, which is of great importance for genetic improvement of the crop. Furthermore, evaluation of germplasm is imperative, in order to understand the genetic background and breeding value of the available germplasm (Anderson, 1984). Success of crop improvement programme depends on the extent of variability, choice of parents for hybridization and selection procedure. Cluster analysis is a potential tool for measuring divergence among a set of populations based on multiple characters. It is also used as an efficient tool in the quantitative estimation of genetic diversity for a breeding programme Moreover, grouping of different genotypes of a particular crop will be more useful in choosing suitable parents for heterosis breeding. Cluster analysis is a multivariate technique is used to classify objects or cases into homogenous groups called clusters. There are various clustering methods and techniques making their classification a difficult task (Hartigan, 1975) [7]. In fact, each clustering method works in a different way and very often yields results different from the others. The validation of these clustering techniques and methods helps for the effective clustering of the objects. Hence an attempt was made in this paper to compare the various clustering techniques and methods, to study their efficiency and to propose the best out of it for small samples.
Peixoto *et al.*, 2019 conducted a study on 19 genotypes of tomato to compare methods of multivariate analysis for the evaluation of genetic diversity in tomato genotypes with different levels of resistance to pests. The results of the study indicated that the Tocher method result was difficult to provide the divergence between genotypes since most of the genotypes was concentrated on a single group (94.74%).

2. Materials and Methods
The present study is based on secondary data on replicated yield trial data of 20 cultures of paddy from Regional Agricultural Research Station (RARS), Pattambi under Kerala Agricultural University, Kerala. The main items of observations taken for the study were plant height (cm), tillers per plant, good grain per panicle, chaff per panicle, grain yield (kg/ha) and straw yield (kg/ha). As the initial step of cluster analysis, the similarity measure selection was

done. Some of these measures are: Euclidean Distance, Manhattan Distance and Minkowski Distance (Deza and Deza., 2009) [4]. First aim of usage of distance methods is to obtain similarity according to distance between data which is not grouped. Thus, similar data can be included in the same cluster. To imply clustering analysis, it is assumed that data should be normal distribution. However, this is just theoretical assumption and is ignored in practice. Only the suitability of the calculated distance values to normal distribution is considered (Gulagiz and Sahin., 2017) [6]. Secondary data on replicated yield trial data of 20 cultures of paddy were obtained from RARS Annual Report 2017-2018. clusterCrit, fpc, factoextra and clValid packages of Rstudio software were used for the analysis (Das and Augustine., 2017) [1].
The hierarchical and non hierarchical clustering techniques were performed using Euclidean distance as the measure of similarity. There were seven different clustering methods used for the comparative study.

2.1 Single linkage method
It is a method of agglomerative clustering in which two objects having minimum distance (nearest neighbour) form the cluster.

2.2 Complete linkage method
In this method, the object having maximum distance (farthest distance) between them constitutes two groups.

2.3 Average linkage method
The similarity between two clusters depends on the average distance between the similar members (average distance).

2.4 Centroid method
In a cluster of points, the centroid is the point that has the average coordinates of all the objects of the cluster. The similarity between two clusters depends on the distance b/w the centroids of the clusters.

2.5 Ward's method
This method is distinct from all other methods as it uses an analysis of variance approach to evaluate the distance between the clusters (Ward, 1963). This procedure is based on minimizing the loss of information from joining two groups.

2.6 Tocher method
This method is extensively used in clustering quantitative data based on their distances calculated as D^2 values using Mahalanobis Distance (Rao, 1952) [8].

2.7 K means approach
This is a non-hierarchical clustering technique. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum (Davidson, 2002) [2].
The different clustering method results thus obtained was compared using different cluster validity index. The cluster validation provides a way of validating the quality of clustering algorithms. The commonly used cluster validity indices are Dunn Index and Davies-Bouldin (DB) Index. Conditions for a good clustering method includes high intra- cluster homogeneity within the clusters and high inter-cluster separation among the clusters. The different clustering method results thus obtained were compared using different cluster validity indices.

2.8 Dunn Index
The Dunn index (DI) a metric for evaluating clustering algorithms, is an internal evaluation scheme, where the result is based on the clustered data itself (Dunn, 1974) [5].

$$U = \min_{1\le j\le c}\left\{\min_{1\le j\le c, j\ne i}\left\{\frac{\delta(Xi, Xj)}{\max\{\Delta(Xk)\}}\right\}\right\}$$

where, $\delta(Xi, Xj)$ is the inter cluster distance and $\Delta(Xk)$ is the intra cluster distance of Xk.

2.9 Davies-Bouldin Index
The Davies–Bouldin (DB) index is a metric for evaluating clustering algorithms, is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset (Davies and Bouldin., 1979) [3].

$$U = \frac{1}{k\sum_{i=1}^{k}\left\{\max_{i\ne j}\left\{\frac{\Delta(Xi)+\Delta(Xj)}{\delta(Xi,Xj)}\right\}\right\}}$$

where, $\delta(Xi, Xj)$ is the inter cluster distance between the clusters Xi and Xj.

3. Results and Discussion
3.1 Clustering methods
The dissimilarity distance matrix for the 20 cultures were calculated using Euclidean distance measure. The comparison of clustering techniques as well as different types of agglomerative clustering methods was done using Rstudio programming software. The cluster analysis results using different clustering methods is given in Table 1. The number of clusters selected for

the study was four using the gap statistics method. None of the clustering results showed any similarity in the clustering pattern. In Tocher method of clustering the first cluster had maximum number of accessions (14) as compared with other clusters. This result was difficult to provide the divergence between accessions since most of the genotypes was concentrated on a single group (70%). The other methods clustering pattern had quite better clustering pattern. The divergence study in the case of small number of samples Tocher method was found be less efficient as with that of the other clustering methods.

Table 1: Different clustering techniques along with the clusters obtained

Clustering Techniques	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Single linkage method	17, 20, 5, 7	8, 9	4, 13, 14, 18, 1, 3, 2, 11, 15	16, 19, 12, 6, 10
Complete linkage method	5, 7	8, 9, 16, 19, 12, 6, 10	17, 20	14, 18, 4, 13, 1, 3, 2, 11, 15
Average linkage method	5, 7	17, 20	4, 13, 14, 18, 1, 3, 2, 11, 15	8, 9, 16, 19, 12, 6, 10
Centroid method	17, 20	5, 7	8, 9, 16, 19, 12, 6, 10	4, 13, 14, 18, 1, 3, 2, 11, 15
Ward's method	17, 20	14, 18, 4, 13, 1, 3, 2, 11, 15	16, 19, 12, 6,10	8, 9, 5 ,7
Kmeans approach	6, 8, 9, 10, 12 19	17, 20	5, 7	1, 2, 3, 4, 11, 13, 14, 15, 16, 18
Tocher method	11, 15, 18, 14, 13, 3, 1, 2, 4, 19, 16, 6, 10, 12	8, 9, 7	17, 20	5

Where 1, 2, 3,…20 are the twenty cultures of paddy taken for the study.

3.2 Cluster validation

The different clustering method results were subjected to cluster validation using the two important cluster validity indices viz., Dunn Index and DB index. The criteria for deciding the best clustering technique and method in the case of Dunn Index is higher the Dunn index value, better is the clustering and in DB index, lower the DB index value, better is the clustering. The Dunn index and DB index values for every clustering methods were obtained (Table 2). The Kmeans approach had the highest Dunn index value followed by average linkage method and ward's method respectively and Tocher method had the least. So, the better clustering was done by Kmeans, the non hierarchical clustering technique. In hierarchical clustering average linkage method was found to be best. The DB index value ranged from 0.47 to 0.55. The least value was for the Kmeans approach and highest was for the Tocher method. This result also insisted that Kmeans approach clustered well while Tocher method clustering was not so efficient.

Table 2: Clustering methods with their respective Dunn index and DB index value

Clustering Methods	Dunn Index	DB Index
Single linkage method	0.398	0.520
Complete linkage method	0.409	0.513
Average linkage method	0.445	0.503
Centroid method	0.405	0.521
Wards method	0.419	0.514
Kmeans approach	0.515	0.477
Tocher method	0.383	0.559

4. Conclusion

Comparative analysis of the clustering techniques suggested that non-hierarchical clustering techniques ie, Kmeans approach was found to be the most efficient technique. Among the hierarchical clustering techniques average linkage method was found to be the most efficient and Tocher method was found to be the least for small samples. Further investigations are required to confirm the statistical significance of different clustering techniques. The significance as well as the relevance of Tocher method in divergence analysis should be given importance in the future studies with different similarity measures.

5. References

1. Das D, Augustine DP. Comparative study of clustering algorithms using R. Int. J. Res. Eng. IT Social Sci. 2017; 07(2):14-20.
2. Davidson I. Understanding K-Means Non-Hierarchical Clustering. SUNY Albany Technical Report, State University of New York, Albany, 2002, 101.
3. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intelli. 1979; 1(2):224-227.
4. Deza MM, Deza E. Encyclopedia of distances. Springer, Berlin, Heidelberg, 2009, 583.
5. Dunn JC. Well-separated clusters and optimal fuzzy partitions. J Cybernetics. 1974; 4(1):95-104.
6. Gulagiz FK, Sahin S. Comparison of hierarchical and non-hierarchical clustering algorithms. Int. J Comput. Eng. and Inf. Technol. 2017; 9(1): 1-6.
7. Hartigan JA. Clustering Algorithms (2nd Ed.). John Wiley and Sons, Canada, 1975, 420.
8. Rao CR. Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York, 1952, 390.
9. Ward JH. Hierarchical grouping to optimize an objective function. J Am. Statist. Assoc. 1963; 58:69-78.