A Relative Examination on Clustering Techniques: Agglomerative, K-Means, Affinity Propagation and DBSCAN

J.Kiran Kumar, M.Seshashayee

Abstract: ***Clustering is a procedure of grouping a collection of certain objects into a relevant sub-group. Each sub-group is called as a cluster, which guides users to comprehend the collections in a data set. It is an unsupervised learning technique where each dispute of this type deals with discovering a structure during the accumulation of unlabeled data. Statistics, Pattern Recognition, Machine learning are some of the active research in the theme of Clustering techniques. A Large and Multivariate database is built upon excellent data mining tools in the analysis of clustering. Many types of clustering techniques are—Hierarchical, Partitioning, Density–based, Model based, Grid–based, and Soft-Computing techniques. In this paper a comparative study is done on Agglomerative Hierarchical, K-Means, Affinity Propagation and DBSCAN Clustering and its Techniques.***

Keywords : Agglomerative Hierarchical Clustering, K-Means Clustering, Affinity Propagation, DBSCAN.

I. INTRODUCTION

Clustering is categorizing sets of knowledgeable data into various numbers of clusters so that the data within the cluster are similar, but are perfectly disparate to the data in the new clusters. Dissimilarity and similarity are determined by the attribute values describing the objects. The objective of clustering is to provide measures and criteria that are utilized for determining whether two objects are similar or dissimilar. Aim of clustering technique is descriptive, and classification is predictive.

All clustering techniques are to find the cluster's center that represents each cluster this is a common approach. Cluster center represents an input vector, to identify among clusters the vector to which it is owned. An equivalent metric is utilized to measure the closeness of data points in a cluster. Cluster analysis is often utilized as standalone data processing device.

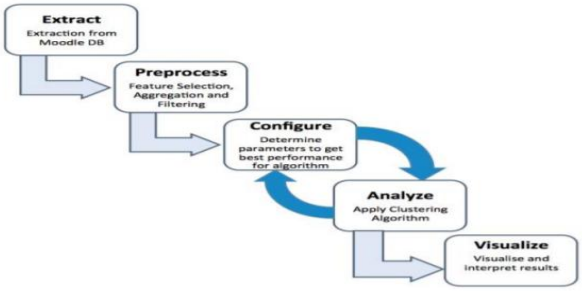

Fig. 1.Steps for performing clustering evaluation.

II. DATA CLUSTERING TECHNIQUES

A. Agglomerative Hierarchical Clustering

An Agglomerative hierarchical clustering procedure utilizes a bottom-up master-plan. It consistently initiates by allowing each object from its retained cluster and repetitively amalgamates clusters into bigger and bigger clusters, until all the objects are in a unique cluster or certain concluded conditions are persuaded. The sole cluster enhances the hierarchy's root. For the integrating step it notices two clusters that are adjacent to one another and unites the two to form one cluster. Because two clusters are combined per recapitulation, where each cluster holds at the minimum one object, an agglomerative technique needs at most n repetitions. It is less responsive to noise in the dataset. Agglomerative Hierarchical algorithm implementation is superior and exhibit more standards.

B. K-Means Clustering

K-Means Clustering is a division technique. It gives rise to a particular number of disjoint, flat (non-hierarchical) clusters. It is well befit to producing globular Cluster; the execution of K-Means technique is more effective than Hierarchical Clustering technique. It can also be utilized in unreserved data; it is firstly transformed into numeric data by allocating ranks. K-Means is too responsive to noise in the dataset. K-Means algorithm also rises its time of execution. K-Means algorithms show fewer standards.

C. Affinity Propagation

Bredan Frey and Delbert Dueck first presented Affinity propagation in the year 2007. It is built on the theory of message passing between points. Dissimilar to k-means and k-medoids, there is no demand in the calculation of the number of clusters by affinity propagation prior to the execution of the algorithm. In this type of clusters, affinity propagation algorithm, the data points can be viewed as a network where each data point dispatch pieces of information to each other data points. The main theme of these pieces of

Revised Manuscript Received on January 5, 2020
***** Correspondence Author
J. Kiran Kumar*, PG Student, Department of CS,GIS,GITAM(Deemed to be University), Visakhapatnam, India, kirankumarjami51@gmail.com
Dr. M. Seshashayee, Department of CS,GIS,GITAM (Deemed to be University), Visakhapatnam, India, mseshashayee@gmail.com

information is the readiness of the points being exemplars.

D. DBSCAN

Clustering investigation is primarily an unsupervised learning method that splits the data points into a certain number of batches or groups, such that the data points in the same groups have indistinguishable possessions and data points in non-identical groups have different possessions in little sense. It contains of numerous distinct procedures built on distinct evolution.

Basically, all clustering techniques utilize the identical perspective i.e. firstly we determinate uniformity and then we employ it to cluster the similar points into groups or batches. Clusters are opaque regions in the data space, parted by regions of the lower opaque of points. The DBSCAN algorithm is forecasted on this inherent conception of "clusters" and "noise".

III. RESULTS AND ANALYSIS

The examination was supervised on Cluster data set, which it was taken from Github for assessing the precision and presentation of clustering techniques. The exact dataset is taken as input to the Agglomerative Hierarchical, K-Means, Affinity Propagation and DBSCAN clustering.

A. Data Pre-Processing

It is a very necessary step and it should be adopted in clustering, as this procedure make use of concepts like constant, average, minimum, maximum, standard deviation to compute missing values in the tuples. These missing values required to be avoided for precise outputs. Preprocessing demands steps like data polishing, data combination, data alteration, data depletion and data distinction.

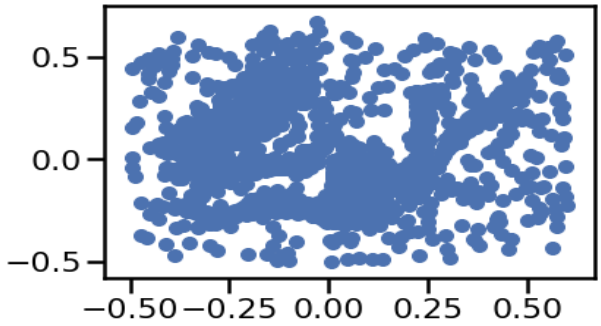

Fig. 2.Cluster representation of data

B. Figures and Tables

The results of the experiments will be shown below:

Table- I: Performance Comparison of Algorithms

Algorithm	No of clusters	Time taken
Agglomerative Hierarchical clustering	10	0.248s
K-Means clustering	10	0.214s
Affinity Propagation	-	18.896s
DBSCAN	-	0.030s

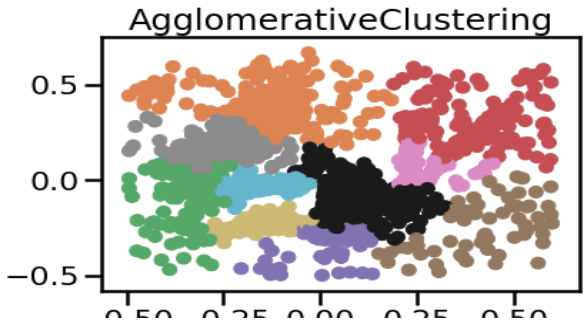

Fig. 3.Clusters found by Agglomerative Clustering

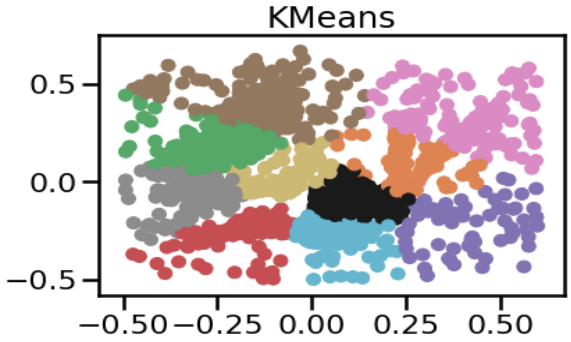

Fig. 4.Clusters found by K-Means Clustering

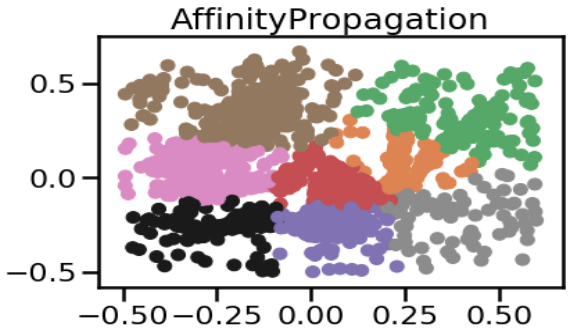

Fig. 5.Clusters found by Affinity Propagation

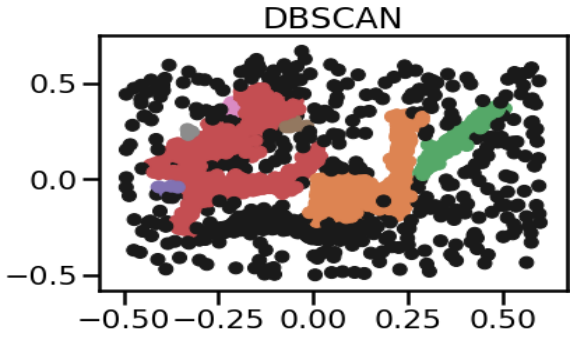

Fig. 6.Clusters found by DBSCAN

IV. CONCLUSION

By studying several algorithms on the given dataset as input in Table- I. The Time taken for all the algorithms is different. When comparing between various clustering algorithms such as Agglomerative Hierarchical, K-Means, Affinity Propagation and DBSCAN.

DBSCAN gives accurate and perfect results which are required by the clustering

algorithms. DBSCAN Results are calculated in the span of 0.030s which is the minimum when compared with the other algorithms.

FUTURE SCOPE

Implementation using HDBSCAN can be done to gain more speedy results. Since HDBSCAN is the next version of DBSCAN as its converts DBSCAN into hierarchical clustering by using a technique which extracts a flat clustering based on the stability of cluster.

REFERENCES

1. Data Mining concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, Elsevier Publications, 3rd Edition, 2013.
2. S.Vijayalaksmi and M Punithavalli (2012).A Fast Approach to Clustering Datasets using DBSCAN and Applications (0975 – 8887) Vol 60– No.14, pp. 1-7.
3. "Intelligent Air Pollution Prediction System using Internet of Things (Iot)", International Journal of Engineering and Advanced Technology, 2019.
4. Pradeep Rai and Shubha Singh (2010) A Survey of Clustering Techniques, International Journal of Computer Applications (0975 – 8887) Vol 7– No.12, pp. 1-5.
5. Guha, Meyerson, A. Mishra, N. Motwani, and O. C. . ."Clustering data streams: Theory and practice." IEEETransactions on Knowledge and Data Engineering, vol. 15,pp. 515-528, 2003.
6. DBSCAN (Density based clustering in machine learning). Available: https://www.geeksforgeeks.org/dbscan-Clustering-in-ml-density-based-clustering./
7. Study of K-Means and Hierarchical Clustering Techniques. Available: https://www.researchgate.net/publication/293061584_Comparative_Study_of_KMeans_and_Hierarchical_Clustering_Techniques.

AUTHORS PROFILE

J.Kiran Kumar pursuing Master of Computer Applications, Department of CS, GIS, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh. His area of interest in Data Mining, Parallel Processing and Machine learning.

M.Seshashayee has been awarded Ph.D. in Computer Science and Technology and presently working as Assistant Professor in the Department of CS, GIS, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh. Her Research specialization is in Image Segmentation Methods using Data Mining Techniques. She has published more than 18 publications in International Journals. She is also the member of CSI and IAENG.