CONPREHENSTAIVE ALGORITHMS OF CLUSTERING TECHNOLOGY : A REVIEW

[1]Hemant Rathore, [2]Manoj kumar, [3] Pooja Singh
[1]Assistant professor, [2]Assistant professor,[3]B.Tech.Scholar
[1]Department of Computer Science Engineering
[1]Rajasthan Insitute Of Engineering and Technology,Bhankrota,Jaipur,India

Abstract - Data analysis is used as a common method in modern science research, which is through communication science, computer science and biology science. Clustering, as the basic composition of data analysis, plays a significant role. On one hand, many tools for cluster analysis have been designed, along with the information increase and subject intersection. On the other hand, each clustering algorithm has its own strengths and weaknesses, due to the complexity of information. In this review paper, we initiate at the explanation of clustering, take the primary elements involved in the clustering process, such as the distance or similarity measurement and evaluation indicators, into consideration, and analyse the clustering algorithms from two perspectives, the traditional ones and the modern ones. All the discussed clustering algorithms will be compared in detail and comprehensively

Index Terms – Clustering, k-means, Gaussian (EM), Fuzzy, Quality Threshold (QT) , Agglomerative Hierarchical, DBSCAN, DIANA, AGNES .

I. INTRODUCTION

Clustering is a process which partition a given data set into homogeneous group based on given feature such that similar object are kept in a group whereas dissimilar object are in different group. It is the most important unsupervised learning problem.it deal with finding structure in a collection of unlabeled data. For better understanding please refer to fig. (i). Clustering analysis has been an emerging research issue in data mining due its variation of applications. Clustering, considered as the most important question of unsupervised learning, deals with the data structure partition in unknown area and is the basis for further learning. Clustering is an unsupervised learning method that groups a set of given data points into well separated subsets.

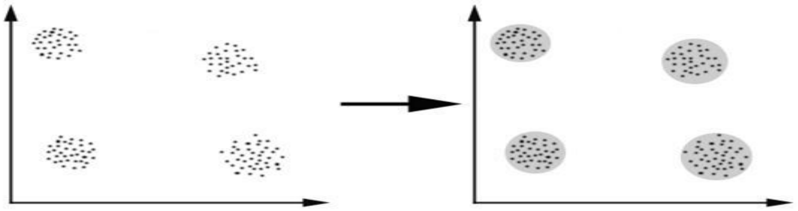

fig. (i) : showing four clusters formed from the set of unlabeled data

In this review paper we discuss different clustering algorithms [1].

II. DIFFERENT TYPES OF CLUSTERING ALGORITHMS

1. k-means clustering algorithm

K-means is the one of the simplest unsupervised learning algorithm that solve clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main concept is to explain k centers, one for single cluster These centers should located in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other .This algorithm aims at minimizing an objective function know as squared error function.

This algorithm give best result when data set are distinct or well separated from each other[2].

2. Gaussian (EM) clustering algorithm

This algorithm consider apriori that there are 'n' Gaussian and then algorithm try to fits the data into the 'n' Gaussian by expecting the classes of all data point and then maximizing the maximum likelihood of Gaussian centres

This algorithm gives externally useful result for the real world data set.

3. Fuzzy c-means clustering algorithm

This algorithm works by assigning participation to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one.

This algorithm gives best result for overlapped data set and comparatively better then k-means algorithm. Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more then one cluster center[3].

4. Quality Threshold (QT) clustering algorithm

This algorithm requires the apriori particular of the threshold distance within the cluster and the minimum number of elements in each cluster. Now from each data point we find all its candidate data points. And Which are within the range of the threshold distance from the given data point. This way we find the candidate data points for all data point and select the one with large number of candidate data points to form cluster. Now data points which belongs to this cluster is removed and the same procedure is repeated with the reduced set of data points until no more cluster can be formed satisfying the minimum size criteria

This algorithm clusters that pass a user-defined quality threshold will be returned. All possible clusters are considered candidate cluster is generated with respect to every data points and tested in order of size against quality criteria.

5. Hierarchical clustering algorithm

Hierarchical clustering algorithm is of two types:
i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting)
ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).
Both this algorithm are exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.

Agglomerative Hierarchical clustering -This algorithm works by combination the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed.

6. MST based clustering algorithm

The planned algorithm searches for that optimum value of the threshold for which the Intra-Inter distance ratio is minimum. It needs not to mention that this optimum value of the threshold must lie between these two extreme values of the threshold. However, in order to reduce the number of iteration we never set the initial threshold value to zero.

This algorithm gives comparatively better performance then k-means algorithm

7. Kernel k-means clustering algorithm

This algorithm exploit the same trick as k-means but with one difference that here in the calculation of distance, kernel method is used instead of the Euclidean distance.

This algorithm is able to identify the non-linear structures and give best suited for real life data set[4].

8. Density based clustering algorithm

Density based clustering algorithm has played a important role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability.

Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbours which are within distance.

Density Connectivity - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbours and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbour of "r", "r" is neighbour of "s", "s" is neighbour of "t" which in turn is neighbour of "p" implies that "q" is neighbour of "p"

Does not require a-priori specification of number of clusters in this algorithm. In this algorithm Able to identify noise data while clustering. DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.

III. PERFORMANCE COMPARISON

The result analysis shows that K-means algorithm performs well without inserting the principle component analysis filter as compared to the Hierarchical clustering algorithm.Hierarchical clustering as compared to Farthest fast clustering gives better performance. K-means give best result when data set are distinct or well separated from each other. Fuzzy c-means gives best result for overlapped data set and comparatively better then k-means algorithm. MST gives comparatively better performance then k-means algorithm but less performance Fuzzy c-means**..** kernel k-means identify the non-linear structures and give best suited for real life data set[5].

IV. CONCLUSION

We have described clustering algorithm and this is basic composition of data analysis, plays a significant role. Clustering analysis has been an emerging research issue in data mining due its variety of applications. K-means has aims at minimizing an objective function know as squared error function. K-means algorithm give best result when data set are distinct or well separated from each other. Gaussian (EM) clustering algorithm gives externally useful result for the real world data set. Fuzzy c-means gives best result for overlapped data set and comparatively better then k-means algorithm. Quality Threshold (QT) clustering algorithm clusters that pass a user-defined quality threshold will be returned. MST gives comparatively better performance than k-means algorithm**.** Kernel k-means is able to identify the non-linear structures and give best suited for real life data set. Does not require a-priori specification of number of clusters in Density based clustering algorithm. Density based clustering algorithm able to identify noise data while clustering. DBSCAN algorithm is easy to find arbitrarily size and arbitrarily shaped clusters.

V. REFERENCES

[1] Dongkuan Xu, Yingjie Tian “A Comprehensive Survey of Clustering Algorithms”, Ann. Data. Sci. (2015) 2(2):165–193.

[2] Prakash singh,Aarohi surya“Performance analysis of clustering algorithms in data mining in weka” ,IJAET,ISSN : 22311963 ,Jan,2015,

[3] Sunila Godara and Ritu Yadav “Performance analysis of clustering algorithms for Character recognition using weka tool” ISSN 2230-9624. Vol 4, Issue 1, 2013, pp119-123.

[4] Pallavi , Sunila Godara “A Comparative Performance Analysis of Clustering Algorithms” IJERA,ISSN: 2248-9622 ,Vol. 1, Issue 3, pp.441-445

[5] Anjana Gosain,Sonika Dahiya”,Performance Analysis of Various Fuzzy Clustering Algorithms: A Review”, Procedia Computer Science 79 (2016) 100 – 111