Performance Review of Clustering Algorithms

Garima Sehgal
Research Scholar, DCSA
Kurukshetra University
Kurukshetra, India

Dr. Kanwal Garg
Assistant Professor, DCSA
Kurukshetra University
Kurukshetra, India

***Abstract*- The objective of this paper is to review the performance of clustering algorithms. This paper discusses about the categories of clustering algorithms, reviews from various articles and journals concerning clustering algorithms and issues and challenges of clustering algorithms. The author will make an attempt to evaluate the performance of various clustering algorithms and to select an optimum clustering algorithm for the prediction of future.**

Keywords- Clustering Algorithms, Tools.

I. INTRODUCTION

Clustering is division of data into groups of similar objects. Each group called cluster, consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups[12].

Clustering methods are basically classified into three categories i.e partitioning methods, hierarchical methods, density based methods. Partitioning clustering algorithms have been popular clustering algorithms. Given a set D of n objects in a d dimensional space an input parameter k, a portioning algorithm organizes the objects into k clusters such that the total deviation of each object from is cluster center or from a cluster distribution is minimized. The deviation of a point can be computed differently in different algorithms and is more commonly called a similarity function[4].

Hierarchical method creates a hierarchical decomposition of the given set of data objects forming a dendrogram- a tree which splits the database recursively into smaller subsets. The dendrogram can be formed in two ways bottom up or top down. The bottom up approach, also called the "agglomerative" approach, starts with each object forming a separate group. It successively merges the objects or groups according to some measures like the distance between two centers of two groups and this is done until all of the groups are merged into one, or until a termination condition holds. The top down also called the "divisive" approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters accordingly to some measures until eventually each object is in one cluster, or until a termination condition holds.[4]

Density based method typically regard clusters as dense regions of objects in the data space which are separated by the regions of same density. Density based methods can be use to filter out noise and, and discover clusters of arbitrary shapes.[4]

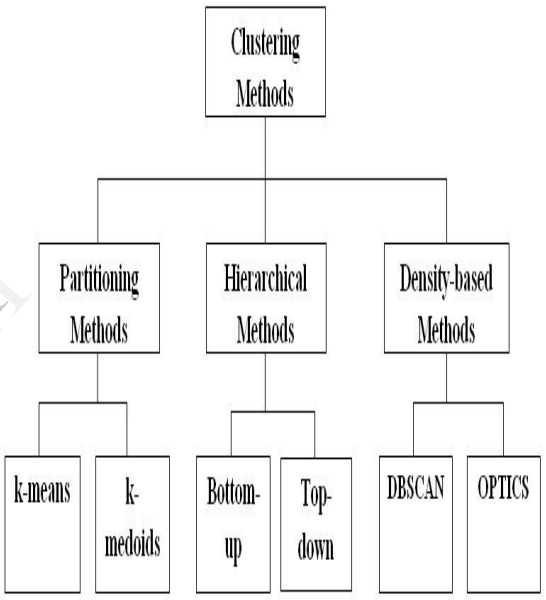

Figure 1 Clustering Algorithms[2]

In this paper there is brief overview of the clustering algorithms and other related facts have been described. Section 1 gives the introduction, section 2 explores the reviews of various researchers, section 3 describes performance review of clustering algorithms, in section 4 is concerned with issues and challenges, section 5 concludes the paper.

II. LITERATURE REVIEW

Various journals and articles concerning clustering algorithms were studied from year 2008 to 2013. Some compared clustering algorithms while some modified the existing algorithms to improve the performance.

Osama Abu Abbas[12] compared different data clustering algorithms. The algorithms which was under investigation were k means algorithm, hierarchical clustering algorithm, self organizing maps algorithm and expectation maximization clustering algorithm. All these algorithms

were compared according to the following factors size of dataset, number of clusters, type of dataset and type of software used. Some conclusions were extracted belong to the performance, quality and accuracy of clustering algorithms. A similar work was done by S. Revati and T.Nalini.[14] described the comparative study of clustering algorithms across two different data items. The performance of various clustering algorithms was compared based on the time taken to form the estimated clusters. The results were depicted as a graph. It was concluded that as the time taken to form the cluster increases as the number of cluster increases. Soumi Ghosh and Sanjay Kumar Dubey.[16] compared two important clustering algorithms namely centeroid based k means and representative based FCM(fuzzy c means). These algorithms are applied and performance is evaluated on the basis of efficiency of clustering output. The number of data points as well as number of clusters are the factors upon which the behaviour patterns of both the algorithms are analyzed.

Comparison of various clustering algorithms using weka tool was performed by following researchers Narendra Sharma et al.[9] provides detailed introduction to weka clustering algorithms. With the help of figures they have shown the working of various algorithms , advantages and disadvantages of algorithms and concluded with the result that that k means clustering algorithm is smallest algorithm compared to other algorithms. Bharat Chaudhari and Manan Parikh.[2] analyzed the performance of three major clustering algorithms k means, hierarchical clustering and density based clustering algorithms and compared the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. Performance of three techniques was presented compared using a clustering tool weka. Sharmila and R.C Mishra[15] compared different clustering algorithms using weka tool. It also presented advantages and disadvantages of different clustering algorithms.

Some modifications were made by various researchers to improve the performance of existing clustering algorithms. Malay k. Pakhira [8] presented a modified version of the k means algorithm that efficiently eliminates the empty cluster problem. He has shown that the modified algorithm is semantically equivalent to the originally k means and there is no performance degradation due to modifications. J. Hencil Peter[6] proposed a new algorithm to improve the quality of the DBSCAN algorithm. The existing algorithm Fast DBSCAN and memory effect in DBSCAN has been combined to improve the quality of the output as well as to speed up the performance. O.A Mohamed Jafar and R.Shivkumar[11] presented a brief survey of an ANT based clustering algorithm and overview of some of its applications. The algorithm has the number of features that make it interesting. It has the ability to automatically discovering number of clusters.

III. PERFORMANCE REVIEW

The advantages and disadvantages of clustering algorithms are discussed in tabular form:-

Table 1: ADVANTAGES AND DISADVANTAGES

CLUSTERING ALGORITHM	ADVANTAGES	DISADVANTAGES
Partitioning Based	1.For large number of variables, partitioning algorithm may be faster than hierarchical method. 2. Partitioning method may produce constricted clusters than hierarchical clusters.	1.Difficulty in comparing quality of clusters formed. 2. Knowing number of clusters in advance is difficult. 3. Does not give good result with non globular clusters. Different primary partitions can result in different final clusters.
Hierarchical Based	1.Does not require the number of clusters to be identified in advance. 2. calculates the whole hierarchy of clusters. 3.Good result visualization.	1.May not scale well 2. No obvious clusters " flat partition" can be derived. 3. No automatic discovering of optimum clusters.
Density Based	1.Does not require one to specify the number of clusters in advance. 2.can find arbitrarily shaped clusters 3.Density based methods is designed for use with databases that can accelerate region queries.	1.Cannot cluster datasets with large large diffrence in densities 2. Quality of density based methods depends upon the distance measure.

Partitiomg based methods are the best if the number of clusters is known in advance and the clusters to be formed are circular in shape.

Hierarchical based methods are good in result visvalization, but the major problem is of flat partitioning.

Density based methods can accelerate region queries, and can produce arbitraily shaped clusters but cannot cluster datasets with large difference in densities.

IV. ISSUES AND CHALLENGES

The various issues and challenges concerning clustering algorithms are discussed:-

A.Effect of Normalization :- Comparing between results of algorithms using normalized data and non normalized data will give different results.

B.Time Complexity :- As the complexity of data set increases like audios, videos, pictures and other multimedia

database, this is turn create a time complexity for clustering algorithms.

C.Result Interpretation :- Output of clustering algorithms can be interpreted in different ways which may create confusion for understanding the results by users.

D.Selection of Algorithm :- The selection of a clustering algorithm may based on the type of dataset, time requirement, efficiency needed, accuracy required, error tolerance etc. so the main challenge is to choose the correct type of clustering algorithm for the data set which are based on user requirements among many known clustering algorithms so that user can get the desired results which helps in further research for data mining process.

E.Initial Clusters :- Identifying the number of clusters in advance in partitioning based algorithms is difficult.

V. CONCLUSION

This paper deals with study of different kinds of clustering algorithms. It first defines the clustering which is procedure of assemblage of objects in groups whose members contain some kind of resemblance. After that a detailed study about performance of clustering algorithms in different perceptions are examined. The paper highlights the concern issues and challenges which may be helpful for the upcoming researchers to carry on their work.

VI. REFERENCES

1. Ashwani Gulhane, Prashant L.Paikrao and D.S Chaudhari " A Review of Image Data Clustering Techniques". International Journal of Soft Computing and Engineering, ISSN:2231-2307, Volume 2, Issue 1, March 2012.
2. Bharat Chaudhari and Manan Parikh " A Comparitive Study of Clustering Algorithms Using Weka Tool". International Journal of Applications or Innovation in Engineering and Management ISSN:2319-4817, Volume 1, Issue 2, October 2012.
3. Brijesh Kumar Bhardwaj and Sourabh Pal " Mining Educational Data to Analyze Students Performance". International Journal of Advance Computer Science and Applications Volume 2, 2011.
4. J.Han, M.Kamber and A.K.H Tung "Spatial Clustering Methods in Data Mining : A Survey"
5. Ilango Murugappan and Dr Mohan Vasudev " Projected Clustering Algorithm-A Review" International Journal of Engineering Research and Technology ISSN:2278-0081, Volume 1, Issue 8, October 2012.
6. J.Hencil Peter and A. Antonysamy " An Optimised Density Based Clustering Algorithm" International Journal of Computer Applications ISSN:0975-8887, Volume 6, September 2010.
7. Kumar Dhiraj and Shantanu Kumar Rath "Comparison of SGA and RGA based Clustering Algorithm for Pattern Recognition". International Journal of Recent Trends in Engineering, Volume 1, May 2009.
8. Malay K. Pakira " A Modified K means Algorithm to Avoid Empty Clusters" International Journal of Recent Trends in Engineering. Volume 1, May 2009.
9. Narendra Sharma , Aman Bajpai and Ratnesh Litoria "Comparison the various Clustering Algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering,ISSN:2250-2459,Volume 2, Issue 5, May 2012.
10. 10.Neeraj Kumar Mishra, Vikram Jain, Sandeep Sahu " Survey on Recent Clustering Algorithms in Wireless Sensor Networks" International Journal of Scientific and Research Publications, ISSN:2250-3193, Volume 3, Issue 4, April 2013.
11. O.A Mohamed Jafar and R. Shivkumar " Ant Based Clustering Algorithm: A brief Survey" International Journal of Computer Theory and Engineering ISSN:1793-8201, Volume 2, October 2010.
12. Osama Abu Abbas " Comaprison between Data Clustering Algorithms" The International Arab Journal of Information Technology, Volume 5, July 2008.
13. Richa Loohach and Dr. Kanwal Garg " An Insight Overview of Issues and Challenges Associated with Clustering Algorithms" International Journal of Research in IT and Management, ISSN:2231-4334, Volume 2, Issue 2, Febuary 2012.
14. S. Revathi and T.Nalini " Performance comparison of various clustering algorithms" International journal of advanced research in Computer Science and Software Engineering Volume 3, Issue 2, February 2013.
15. Sharmila and R.C Mishra "Performance Evaluation of Clustering Algorithms" International Journal of Engineering Trends and Technologies, Volume 4, Issue 7, July 2013.
16. Soumi Ghosh and Sanjay Kumar Dubey " Comparative analysis of k-means and fuzzy c meansalgorithms"International Journal of Advanced Computer Science and Applications,Volume 4,2013.
17. V. Illango , R. Subaramanian and V.Vasudevan "Cluster Analysis Research Design Model, Problems ,Issues, Challenges, Trends and Tools" International Journal on Computer Science and Engineering, Volume 3, August 2011.