A Comparative Study Of Data Clustering Techniques

Nishu Sharma[1], Atul Pratap Singh[2]
[1, 2] M. Tech Scholars, School of Computer Science & Engineering,
Galgotias University, Greater Noida

Abstract

As we know that clustering is a process for discovering groups and identifying interesting patterns. Data mining refers to extract knowledge from large database. We can say that data mining is a knowledge mining process. Size and need of dataset are increasing day by day. That's why we need data mining techniques for managing huge dataset. It is a process of knowledge discovery in database (KDD). Data mining techniques are useful in many more fields today such as biology , libraries, GIS, satellite images, marketing, medical diagnostics and many more field .in this paper this study tells us about the comparison between data mining techniques on the basis of size, model, application areas and others features. This study tells us when and which data mining techniques are used.

1. Introduction-

Data clustering[1] is the process of putting similar data objects into a group. Data objects of one group are dissimilar from the data objects of other group. Clustering algorithms are not used for only organize data. These are used also data compression and model construction. Cluster center is the heart of the cluster. Below we define the process of making data clusters [2].

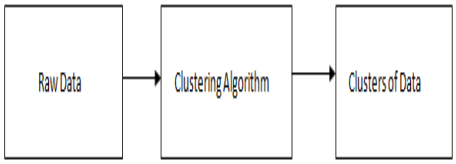

Fig.1 Clustering Process

Firstly, we take raw data, then apply clustering algorithm on the raw data and after that we will get the clusters of data. This is the process of making data clusters with the help of Clustering algorithm.

Clustering techniques [10] satisfy two main criteria:-

- Each cluster is homogeneous because cluster makes with the help of similar objects.
- Objects of one cluster are dissimilar from the other cluster's objects. Each

Cluster should be different from other clusters.

2. Related work

2.1. Idea of Clustering:-

We take the example of the library system[3]. In a library, there are a lot of books available. Books have some similar qualities and as well as dissimilar qualities. We manage or organize the books with the help of clustering easily. We keep the books same place which have similar qualities and keep different locations or place which have dissimilar qualities. It means we make the clusters of the books. With the help of clustering searching option for specific book is so much easy. We can easily search specific book and it take lesser time for searching specific book.

2.2. Applications of Clustering:-

These are some application of clustering [4]:

- Data Reduction
- Hypothesis Testing
- Business
- Biology
- Spatial data analysis
- Web mining

2.3. Requirements of good clustering algorithm [10]:-

2.3.1. Scalability: - Clustering algorithm works well with small datasets and if it will work well with huge datasets. That is called

scalability. It means clustering algorithm should be highly scalable.

2.3.2. Dealing with different types of attribute: - The ability to analyze with any type of attributes such as binary, categorical and ordinal data or mixtures of data types.

2.3.3. Discovery of cluster with arbitrary shape: - Clusters could be any type of shape. That's why algorithm should be detected any type of clustered shape.

2.3.4. Ability to deal with noise & outliers: - Databases contain noisy data (missing values) and outliers that's why algorithm should ability how to deal with these types of databases.

2.3.5. Interpretability & usability: - Users expect clustering results to be interpretable, and usable. That's why clustering may need to be tied up with specific semantic interpretations.

2.3.6. High dimensionality: - A database and data warehouse can contain several dimensions or attributes. Clustering algorithm should be managed high dimensional data.

2.3.7. Data order dependency: - Algorithm should be insensitive to the order of input.

3. Clustering Techniques

Clustering [5] Techniques which are apply on the raw data and after that we get the clusters. That is called clustering techniques. Clustering techniques are of many types but some of them are so much important. These are as follows:-

3.1. Partitioning Algorithm: - This method is based on the partitioning. This method is to partition the data into k groups. The general behavior is that objects in the same clusters are close to each other and objects in the different clusters are far to each other.

There are two types of partitioning algorithm. These are as follows:-

3.1.1. K-means algorithm- This algorithm[6] is based on the mean value or centroid of the objects into the clusters. These steps are as follows: -

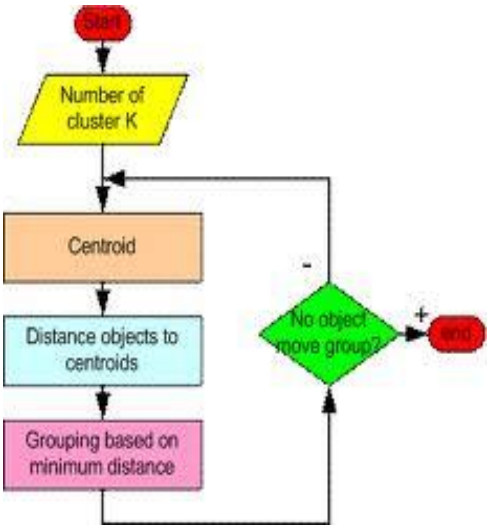

Fig.2 K-means

We arbitrarily [4] choose three objects as three initial cluster centers, where we mark the cluster center by "+". Make the clusters with the help of objects, which objects are nearest to the cluster center then make cluster. After making clusters we will find out the mean value of the cluster. According to the new mean value we will make or generate new clusters. This process is continuing until we find out similar mean values like this figure [7].

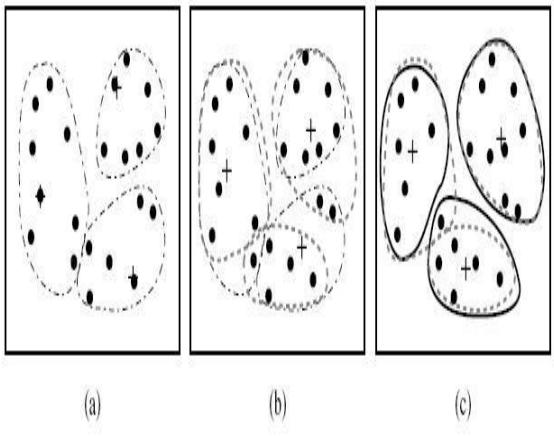

Fig.3 K-means process

Advantages: -

- Simplicity
- Effectiveness
- Easily understandable

Disadvantages: -

- It is not suitable for discovering clusters with nonconvex shapes or different size of clusters.
- It is sensitive to noise and outlier data points.

3.1.2. K-medoids algorithm- In this algorithm, each cluster [7] is represented by one of the objects which are located near the center of the cluster. This algorithm is based on the medoids.

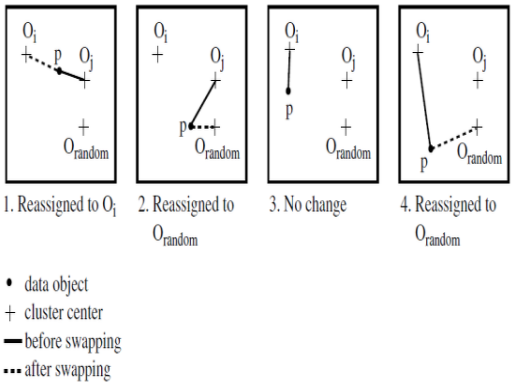

Fig.4 K-medoids

Arbitrarily choose k objects as the initial medoids[8], after that repeat, then assign each remaining objects to the cluster with nearest medoids. Randomly pick a non-medoid object O_{random}. Find out the total cost S of swapping O_j with O_{random}.

If S<0 then swap O_j with Orandom to make new set of k medoids until no change.

Advantages: -

- It can handle noise and outlier data points.
- This method is more robust than K-means algorithm.

Disadvantages: -

- It is more costly than K-means algorithm.
- Slow

3.2. Hierarchical Clustering: - This algorithm is based on the hierarchical decomposition. In this algorithm [9] combine or divide the hierarchical structure. In other words we can say that it is a tree of clusters also known as dendrogram. It contains the hierarchy of clusters. Hierarchical Clustering is of two types.

3.2.1. Agglomerative approach: - This approach is also called bottom up approach. This is based on the combined approach[10]. In this method firstly we start with bottom clusters and take clusters which have minimum distance between clusters. Then combined or merged these two clusters and make single cluster. This process is continued until remained single cluster.

3.2.2. Divisive clustering: - This approach is known as top down approach. It is based on division of clusters approach. We start from top of the hierarchy and take single cluster and divide or split them into clusters. That is called divisive or top down approach.

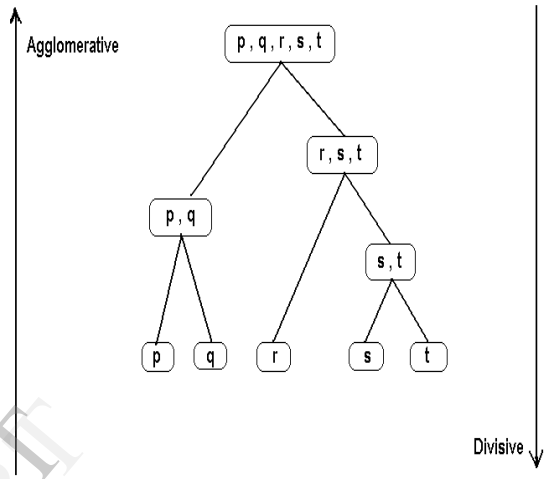

Fig.5 Divisive clustering

Advantages: -

- It has a logical structure.
- Easy to read and interpret.

Disadvantages: -

- After merging and splitting the objects it will neither undo.
- Unstable & unreliable.

3.3. Density based Clustering: - This technique is based on the density of the objects. It is used for arbitrary shapes. Irregular shapes are managed by density based clustering. These steps are as follows [11]: -

3.3.1. DBSCAN Algorithm: - Select an arbitrary point p. Find out all the points density reachable from p. If p is a core point, cluster is formed. If p is a border point, no points are density reachable from p. And DBSCAN[12] visits the next point of the databases. Continue the process until all the points have been processed.

Advantages: -

- Find clusters of arbitrary shape not just convex.
- It can filter out noise.

Disadvantages: -

- DBSCAN does not deal with high dimensional data.

3.4. Grid based Clustering: - This is used for spatial data[13,14]. Spatial data includes structure of objects in space. We quantize data into cells. Then we with only with those objects that are belong to cells

3.4.1. CLIQUE Algorithm: - This is the combination of both grid based and density based clustering. It is useful for clustering high dimensional [15] data in large databases. Steps are as follows: -

- Bottom-up to find out dense units or crowded areas in units.
- Generating minimal number of regions, each region cover one cluster.

Advantages: -

- Processing time is very fast because it depends on the number of cells.
- Effective in large image database.

Disadvantages: -

- Quality of clustering is dependent on the granularity [16, 17] of cells.

4. Comparison

Comparison [18, 19] of all these protocols are given in table below:-

S. No		K-means	K-medoids	Hierarchical	DBSCAN	CLIQUE
1	Application area	Neural network,AI, market segmentation ,pattern recognition	With KD tree for s/w fault prediction	Applied science psychology AI Social science	Satellites images, XRAY crystallography	Social n/w Biological n/w
2	Shape and Size	Spherical	Spherical	Arbitrary& non convex shape	Spherical & arbitrary shape	Find projected clusters in subspace of dimensional space
3	Cluster model	Centroid	Centroid	Connectivity model	Density model	Graph based
4	Scalability	Yes	Scale well only for small data set	No	To large dataset	Good scalability
5	Type of attribute	Numerical	Numerical/categorical	Symbolic attribute	numerical	numerical
6	Outlier handling	No	no	yes	yes	yes

Table.1 Comparison of clustering protocols

5. Conclusion

In this paper we described the process of clustering in a short term from the data mining point of view. We discussed the properties of a "good" clustering methods used to find meaningful partitioning. Clustering lies at the heart of data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data size is increasing day by day and their properties and data interrelationships change, managing and organizing that huge data set is very challenging. We have provided a brief introduction to cluster analysis. All of these approaches have been successfully applied in a number of areas, although there is a need for more extensive study to compare these different techniques and better understand their strengths and limitations. In particular, there is no reason to expect that one type of clustering approach will be suitable for all types of data.

6. References

[1] A. Gupta, "Comparisons among data mining algorithms", ICRITO'2013.

[2] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice-Hall, 1988

[3] D.L.Boley,. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 1998.

[4]M.R. Anderberg, M. R. 1973. Cluster Analysis for applications. Academic Press

[5] L. Wanner, "Introduction to Clustering Techniques", International Union of Local Authorities, July, 2004.

[6]A.K. Jain, "Data Clustering: 50 Years Beyond K-Means" , Pattern Recognition Letters, Vol 31 Issue 8 : pp.651-666 , June, 2010

[7]T. Velmurugan, and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach" An experimental approach. Information. Technology. Journal, Vol, 10, No .3, pp478-484, 2011.

[8] J. Han and M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, August 2000.

[9] J. Han , M. Kamber, "Data Mining Techniques", Morgan Kaufmann Publishers, 2000.

[10] B. S. Everitt, "Cluster Analysis",3rd Edition, Edward Arnold Publishers ,1993.

[11] P. Rai ,S. Singh , "A Survey of Clustering Techniques" , International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010

[12] M. Steinbach, G.Karypis, V.Kumar, "A Comparison of Document Clustering Techniques," University of Minnesota, Technical Report #00-034 (2000).

[13] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques", Intelligent Information Systems Journal, Kluwer Pulishers, 17(2-3): 107-145

[14] R. Ali, U. Ghani, A. Saeed, **"**Data Clustering and Its Applications", Rudjer Boskovic Institute, 2001

[15] P.Arabie, L.J. Hubert, 1996. An overview of combinatorial data analysis, in: Arabie ,P., Hubert, L.J., and Soete, G.D. (Eds) Clustering and Classification, 5-63, World Scientific Publishing Co ., NJ

[16] J.Hartigan 1975 "*Clustering Algorithms"*. John Wiley & Sons, New York, NY

[17] G. Fung, "A Comprehensive Overview of Basic Clustering Algorithms"., June 22, 2001

[18] K. Hammouda, F. Karray , "A Comparative Study of Data Clustering Techniques" University of Waterloo, Ontario, Canada N2L 3G1

[19] O.A. Abbas, Department of computer Science, Yarmouk University, Jordan, "Comparison Between Data Clustering Algorithm" ,The International Arab Journal Of Information Technology, vol.5, No.3, July 2008