Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
SURVEY REPORT
Available Online at www.ijarcs.info

An Overview of Some Data Mining Clustering Techniques

Mr.Anup D.Sonawane
M. Tech IV Sem
SVCE Indore M.P. India
anup_sonawane@hotmail.com

Mr. Vijay Birchha
Asst Prof.
Computer Science and Engineering,
SVCE Indore M.P. India
vijaybirchha@gmail.com

Mr. Preetesh Purohit
Asso. Prof
Computer Science and Engineering
SVCE Indore M.P. India
preeteshpurohit@svceindore.ac.in

Abstract: Data mining clustering techniques are importance and used widely nowadays. Artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing are some common clustering. Several algorithms and methods have been developed in the recent year to solve clustering problem. But improving accuracy and efficiency are to issue are always arises for finding a new algorithm and process for extracting knowledge for. These issues motivated us to develop new algorithm and process for clustering problems. Clustering can be used to partition data set into a number of "interesting" clusters. Cluster analysis is applied to the data set and the resulting clusters are characterized by the features of the patterns that belong to these clusters. In this paper we presented a study of some data mining clustering techniques

Keywords: cluster, accuracy, efficiency, partition, features

I. INTRODUCTION

A Cluster is a set of entities which are alike, and entities from different clusters are not alike. Clusters may be described as connected regions of a multidimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points. Clustering is based on some properties like density, variance, dimension, shape, and separation. The cluster should be a tight and compact high-density region of data points when compared to the other areas of space. The shape of the cluster is not known a priori. It is determined by the used algorithm and clustering criteria [1,2].

Clustering is unsupervised learning because it doesn't use predefined category labels. A clustering algorithm attempts to find natural groups of components based on some similarity. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster [3,11,19].

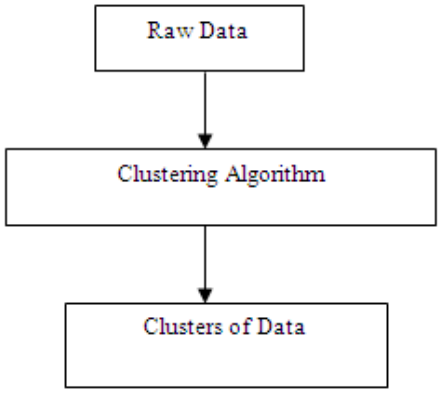

Figure 1 Clustering of raw data

II. PROPERTIES OF A CLUSTER

Clustering of object is a difficult task. To construct a cluster there should be some properties that have to be considered. Some of the properties are size, depth breath, etc [12,20].

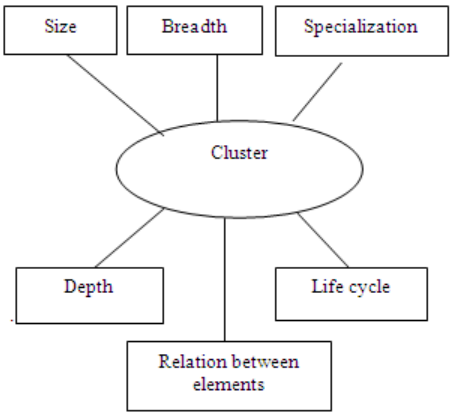

Figure 2 properties of a cluster

III. KEY ELEMENTS OF CLUSTER ANALYSIS

There are some key elements are considered during cluster analysis before the final results can be attained [13, 14, 18].

1. Data presentation.
2. Choice of objects.
3. Choice of variables.
4. What are data units or variables?
5. Normalization of variables.
6. Choice of similarity or dissimilarity
7. Choice of clustering criterion
8. Choice of missing data strategy.
9. Algorithms and computer implementation
10. Number of clusters.
11. Interpretation of results

IV. CLUSTERING METHODS

There are several clustering methods have been developed in past year. Each of which uses a different techniques and process induction principle. Farley and Raftery suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into additional three main categories:
Density-based methods, model-based clustering and grid based methods. An alternative categorization based on the induction principle of the various clustering methods is presented in (Estivill-Castro, 2000). We discuss some of them here [15,16,17].

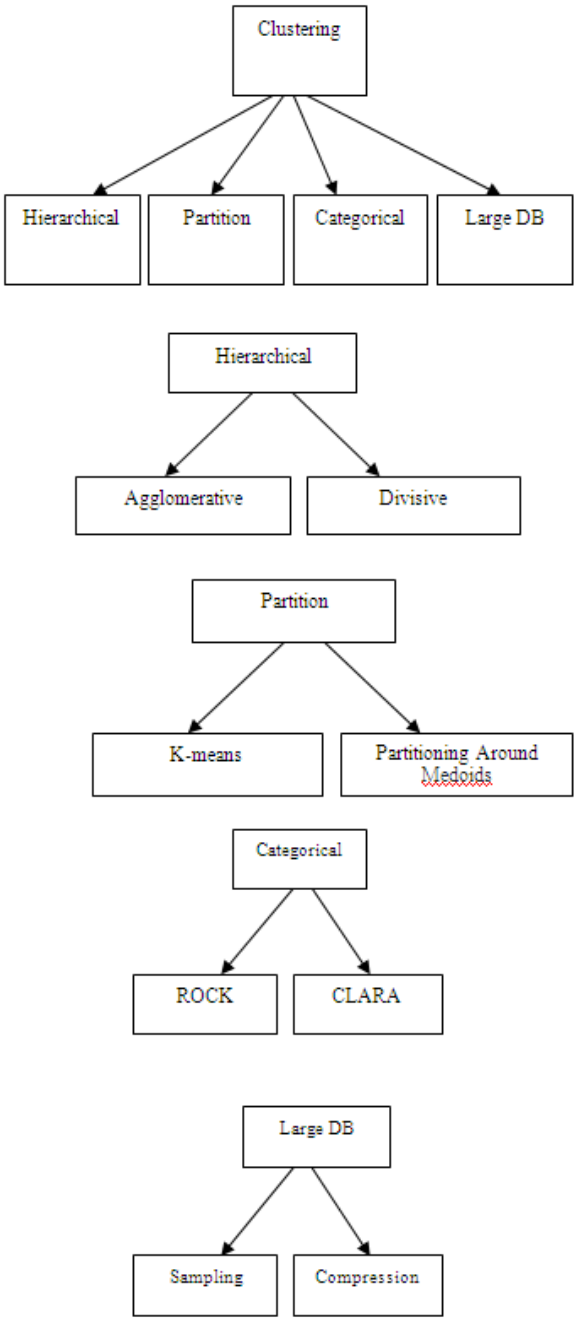

Figure 3 Categories of clustering methods

V. LITERATURE REVIEW

In 2011 K. Ranjini proposed "Performance Analysis of Hierarchical Clustering Algorithm" They explain the implementation of agglomerative and divisive clustering algorithms by using various types of data. They implements and analysis running time of the algorithms using different linkages (agglomerative) to different types of data are taken for analysis[4].
In 2012 Akshay Krishnamurthy proposed "Efficient Active Algorithms for Hierarchical Clustering". They show that a family of active hierarchical clustering algorithms has strong performance. They show that clustering can be improved by using statistical properties. We propose a general framework for active hierarchical clustering [5].
In 2013 M. Emre Celebi et al. proposed "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm" .They proposed an overview of K-Means method with computational efficiency. They compared eight commonly used linear time initialization method for a large and diverse collection of real and synthetic data sets[21].
In 2013 K. Sasirekha, P. Baby proposed "Agglomerative Hierarchical Clustering Algorithm- A Review". They showed that data mining hierarchical clustering method are used to build a hierarchy of clusters. They also show that hierarchical clustering generally fall into two types: Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy [6].
In 2013 Rupanka Bhuyan proposed "A Survey of Some Density Based Clustering Techniques". They proposed a study of various density based clustering methods such as DBSCAN, OPTICS, DENCLUE, VDBSCAN, DVBSCAN, DBCLASD and ST-DBSCAN along with their characteristics, advantages and disadvantages. They also showed how these methods are important and r applicable to different types of data sets to mine useful and appropriate patterns[25].
In 2014 Archana Singh and Avantika Yadav proposed "Hybrid Approach of Hierarchical Clustering". They proposed a hybrid approach of clustering based on AGNES and DIANA clustering algorithms, an extension to the standard hierarchical clustering algorithm. In the proposed algorithm, they used single linkage as a similarity measure. The proposed clustering algorithm provides more consistent clustered results from various sets of cluster centroids with tremendous efficiency [7].
In 2015 Y. S. Thakare et al. proposed "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics". They check the performance of basic k means algorithm using various distance metrics for real life dataset. By the experimental analysis and result. They showed that the performance of k means algorithm is varying on the distance metrics for selected database. The proposed work help to select suitable distance metric for particular application[22].

In 2015 Olga Tanaseichuk "An Efficient Hierarchical Clustering Algorithm for Large Datasets". They show that Hierarchical clustering is a widely adopted unsupervised

learning algorithm. Standard implementations of the exact algorithm for hierarchical clustering require O(n)2 time and O(n)2 memory and thus are unsuitable for processing datasets with large object. They present a hybrid hierarchical clustering algorithm requiring less time and memory [8].

In 2015 Adrian E. Raftery proposed "Bayesian Model Averaging in Model-Based Clustering and Density Estimation" They proposed Bayesian model averaging method for post processing the results of model-based clustering. They summaries and averaged, the posterior model probabilities, instead of being taken from a single best model. They used BMA in model-based clustering for a number of datasets. They show that BMA provides a good clustering for taking data set [23].

In 2016 Jitendra Pal Singh Parmar et al "A Survey on K-Means Clustering Algorithms for Large Datasets" They used two methods global k-means (GKM) and the fast global k-means (FGKM) algorithms for analysis. They iteratively append one cluster center at a time. By using numerical experiments they show the performance of these methods. They implement both algorithms and compare their time and memory for clustering creation[24].

In 2016 K.Jeyalakshmi, S.Shanmugapriya "An Efficient Hierarchical Clustering Algorithms Approach Based on Various- Widths Algometric Clustering". They proposed method by initially assigning each point to its own cluster and then repeatedly selecting and merging pairs of clusters, to obtain a single all inclusive cluster. The key parameter in agglomerative algorithms is the method used to determine the pair of clusters to be merged at each step[9]

In 2017 Shaoning Li , Wenjing Li proposed "A Novel Divisive Hierarchical Clustering Algorithm for Geospatial Analysis". They proposed a new method, cell-dividing hierarchical clustering (CDHC), based on convex hull retraction. They used following steps a convex hull structure is constructed to describe the global spatial context of geospatial objects. Then, the retracting structure of each borderline is established in sequence by setting the initial parameter. The objects are split into two clusters with the borderlines. Finally, clusters are repeatedly split and the initial parameter is updated until the terminate condition is satisfied [10].

VI. PROBLEM WITH CLUSTERING

The important problems with cluster analysis that this work have identified are as follows:

The identification of distance measure: For numerical attributes, distance measures can be used. But identification of measure for categorical attributes in strength association is difficult.

The number of clusters: Identifying the number of clusters & its proximity value is a difficult task if the number of class labels is not known in advance. A careful analysis of inter & intra cluster proximity through number of clusters is necessary to produce correct results.

Types of attributes in a database: The databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other types like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.

Merging decision in not given: Hierarchical clustering tends to make good local decisions about combining two clusters since it has the entire proximity matrix available. However, once a decision is made to merge two clusters, the hierarchical scheme does not allow for that decision to be changed. This prevents a local optimization criterion from becoming a global optimization criterion [17,18].

VII. ADVATAGE AND DISATVATGE

There are several clustering algorithm. Each and every clustering approach has some advantage and disadvantage. Selection of a particular clustering method is depend on the data set other factors. We give some advantage and disadvantage of these methods.

Table 1 advantage and disadvantage

Clustering Techniques	Advantage	Disadvantage
K mean	Produce clusters with relatively uniform size	K value not known
Divisive	Produce more accurate hierarchies	Decisions based on local patterns
DBSCAN	Does not require one to specify the number of clusters	The quality of DBSCAN depends on the distance measure

VIII. COMPARISON OF HIERARCHICAL METHODS

Table 2 some hierarchical clustering method

Name of methods	Basic Techniques
BIRCH	The algorithm depends on a threshold value
ROCK	Assume a static value supplied interconnectivity model
CURE	Used shrinking factor and have an optimum value for effective
CHAMELEON	Based on optimum threshold value

IX. CONCLUSION

We represent a study of some clustering method. We give a study over some cluttering techniques on the basis of their properties.ome method has good scalability and some has good performance. Selection of a method depends on the nature of the data object and other factors.

Hierarchical clustering tends to make good local decisions about combining two clusters since it has the entire proximity matrix available. However, once a decision is made to merge two clusters, the hierarchical scheme does not allow for that decision to be changed. This prevents a local optimization criterion from becoming a global optimization criterion.

Agglomerative algorithms is a type of hierarchical clustering it places each object in its own cluster and then it merges these atomic cluster into larger and larger clusters until all objects are in a single cluster or until termination condition

holds. Most commonly used hierarchical agglomerative clustering methods are Single linkage and complete linkage. It is very difficult to decide to select a method for a given objects because each method has its own advantage and disadvantage. In our proposed work our main focus is this problem. We use Dendogram distance between two object and correlate with original distance matrix. This correlation between two matrixes gives a value between 0 and 1. The value near to give more accurate cluster.

X. REFERENCES

[1] Data Mining: Concepts and Techniques Jiawei Han and Micheline Kamber Simon Fraser University Note: This manuscript is based on a forthcoming book by Jiawei Han and Micheline Kamber, c 2000 (c) Morgan Kaufmann Publishers.

[2] Arun K. Pujari "Data Mining Techniques" Universities Press, 2001

[3] Margaret H. Dunham "Data Mining: Introductory and Advanced Topics" Publisher : Pearson 2002-09-01 ISBN-13 : 9780130888921

[4] K.Ranjini Performance Analysis of Hierarchical Clustering Algorithm Performance Analysis of Hierarchical Clustering Algorithm" Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011 (2011)

[5] Akshay Krishnamurthy "Efficient Active Algorithms for Hierarchical Clustering" Appearing in Proceedings of the 29 th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012.

[6] K.Sasirekha, P.Baby Agglomerative Hierarchical Clustering Algorithm- A Review "International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013 1 ISSN 2250-3153

[7] Archana Singh and Avantika Yadav "Hybrid Approach of Hierarchical Clustering"World Applied Sciences Journal 32 (7): 1181-1191, 2014 ISSN 1818-4952 © IDOSI Publications, 2014

[8] Olga Tanaseichuk, Alireza Hadj "An Efficient Hierarchical Clustering Algorithm for Large Datasets" Austin J Proteomics Bioinform & Genomics - Volume 2 Issue 1 - 2015 ISSN : 2471-0423

[9] K. Jeyalakshmi, S. Shanmugapriya "An Efficient Hierarchical Clustering Algorithms Approach Based on Various-Widths Algometric Clustering" International Journal of Innovative Research in Computer and Communication Engineering (An ISO3297: 2007 Certified Organization) Vol. 4, Issue 7, July 2016".

[10] Shaoning Li , Wenjing Li "A Novel Divisive Hierarchical Clustering Algorithm for Geospatial Analysis" ISPRS Int. J. Geo-Inf. 2017, 6, 30; doi:10.3390/ijgi6010030

[11] Parul Agarwal et al. "Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes" International Journal of Innovation, Management and Technology, Vol. 1, No. 2, June 2010 ISSN: 2010-0248

[12] Ashish Jaiswal et al. " Hierarchical Document Clustering: A Review" 2nd National Conference on Information and Communication Technology (NCICT) 2011 Proceedings published in International Journal of Computer Applications® (IJCA).

[13] Sudesh Kumar "Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 10 3161 – 3166

[14] Shraddha Shukla A Review ON K-means DATA Clustering APPROACH" International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 (2014), pp. 1847-1860 © International Research Publications House http://www. irphouse.com

[15] Shuhie Aggarwal et al. " Hierarchical Clustering- An Efficient Technique of Data mining for Handling Voluminous Data" International Journal of Computer Applications (0975 – 8887) Volume 129 – No.13, November2015

[16] Alessandro Farinelli "A hierarchical clustering approach to large-scale near-optimal coalition formation with uality guarantees "Engineering Applications of Artificial Intelligence Received 30 December 2015; Received in revised form 28 September 2016; Accepted 20 December 2016.

[17] Arpit Bansal et al. " Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining" International Journal of Computer Applications (0975 – 8887) Volume 157 – No 6, January 2017

[18] Sukhvir Kaur et al. "Survey Of Different Data Clustering Algorithms" International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016.

[19] Jasmine Irani "Clustering Techniques and the Similarity Measures used in Clustering: A Survey" International Journal of Computer Applications (0975 – 8887) Volume 134 – No.7, January 2016.

[20] Anurag kumar et al. "A Study on Web Content Mining" International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 1 Jan. 2017, Page No. 20003-20006 Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.29

[21] M. Emre Celebi et al. "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm" Expert Systems with Applications, 40(1): 200–210, 2013

[22] Y. S. Thakare et al. " Performance Evaluation of K-means Clustering Algorithm with Various Distance MetricsY" International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 11, January 2015

[23] Adrian E. Raftery "Bayesian Model Averaging in Model-Based Clustering and Density Estimation" University of Washington Technical Report no. 635
Department of Statistics University of Washington July 2, 2015

[24] Jitendra Pal Singh Parmar " A Survey on K-Means Clustering Algorithms for Large Datasets" IJARCCE ISSN (Online) 2278-1021 ISSN (Print) 2319 5940 International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 9, September 2016.

[25] Rupanka Bhuyan A Survey of Some Density Based Clustering Techniques Department of IT & Mathematics Icfai University Nagaland, 6th Mile, Sovima, Dimapur–797112 Dept. of Computer Science & Engineering, Sikkim Manipal Institute of Technology, Majitar, Rangpo, East Sikkim–737132