International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

Extensive Survey on Hierarchical Clustering Methods in Data Mining

Dipak P Dabhi[1], Mihir R Patel[2]

[1]*Dipak P Dabhi Assistant Professor, Computer and Information Technology (C.E & I.T), C.G.P.I.T, Gujarat, India*
[2] *Mihir R Patel Assistant Professor, Computer and Information Technology (C.E & I.T), C.G.P.I.T, Gujarat, India*

Abstract - *Clustering is the task of grouping the object based on similarity and dissimilarity. In the Datamining, Hierarchical Clustering algorithm is one of the most important and useful method. The hierarchical methods group training data into a tree of clusters. This tree also called dendrogram, with at the top all-inclusive point in single cluster and at bottom all point is individual cluster. The tree /dendrogram can be formed in agglomerative (bottom-up) or divisive (top-down) manner. The goal of this survey is to provide a comprehensive review of different Hierarchical clustering techniques in data mining and also provide comparative study of different hierarchical clustering algorithms.*

Key Words: Clustering, Hierarchical Clustering algorithm, Agglomerative, Divisive.

1. INTRODUCTION

Data mining is the extraction of useful knowledge and interesting patterns from a large amount of available information. In this paper, data clustering is examined. Data clustering is an important technique for exploratory Spartial data analysis, and has been studied for many years. It is very useful in many practical domains such as image processing, classification, Pattern recognition, Economic Science, WWW, etc [1].

There is number of clustering methods are available like Partitioning method, Hierarchical method, density based method, model based method, grid based method etc [2]. Each method have own advantages and disadvantages [3]. There is not any clustering algorithm that can used to solve all problems. In general all algorithms are designed with certain assumptions and favor some type of application, Specific data and biases. In this paper all Hierarchical Clustering algorithm is examined. As the name suggest, the hierarchical methods, in general tries to decompose the dataset of n objects into a hierarchy of a groups. This hierarchical decomposition represented by a tree structure [1]. This tree structure called dendrogram; whose root node represents one cluster, containing all data points and at the leaves there are n clusters, each containing one data point. There are two general approaches for the hierarchical method: agglomerative (bottom-up) and divisive (top down) [4].

1.1 Agglomerative Hierarchical Algorithm:

An hierarchical agglomerative clustering(HAC) or agglomerative method (bottom-up strategy) starts with n leaf nodes(n clusters) that is by considering each object in the dataset as a single node(cluster) and in successive steps apply merge operation to reach to root node, which is a cluster containing all data objects [5].

The merge operation is based on the distance between two clusters. There are three different notions of distance: single link, average link, complete link [3]. These three notions also consider as an individual algorithm.

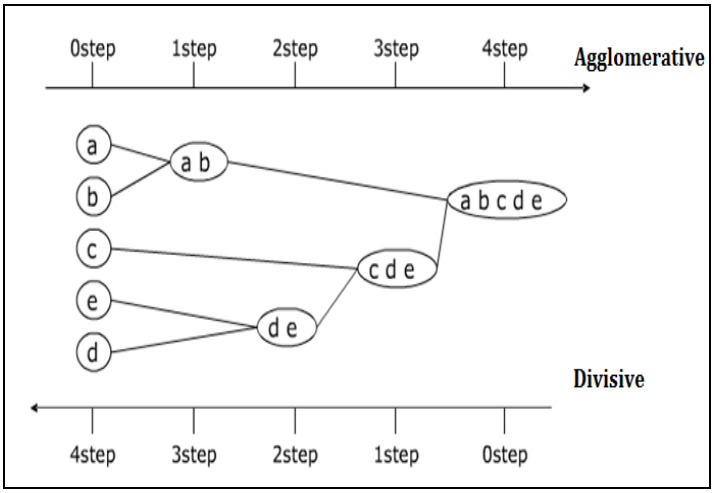


Fig.1: Application of agglomerative and divisive to a data set of five objects, {a, b, c, d, e} [1].

Many agglomerative clustering algorithms have been proposed, such as, single-link, complete-link, average-link, AGNES, CURE, BIRCH, ROCK, CHAMELEON [1][3].

1.2 Divisive Hierarchical Algorithm

The Divisive clustering (top-down strategy) is also known as DIANA (Divisive analysis). This algorithm initially treats all the data points in one cluster and then split them gradually until the desired number of clusters is obtained. To be specific, two major steps are in order. The first one is to choose a suitable cluster to split and the second one is to determine how to split the selected cluster into two new clusters. For a dataset having n objects there is 2n-1 – 1

possible two-subset divisions, which is very expensive in computation [4].
Some Divisive clustering algorithms have been proposed, such as MONA and etc. In this paper, we focus on different hierarchical algorithms. The rest of the paper is organized as follows: Section II defines the different Hierarchical algorithm with its cons and pro.

2. HIERARCHICAL CLUSTERING METHODS:

There is difference between clustering method and clustering algorithm. A clustering method is a general strategy applied to solve a clustering problem, whereas a clustering algorithm is simply an instance of a method [6].

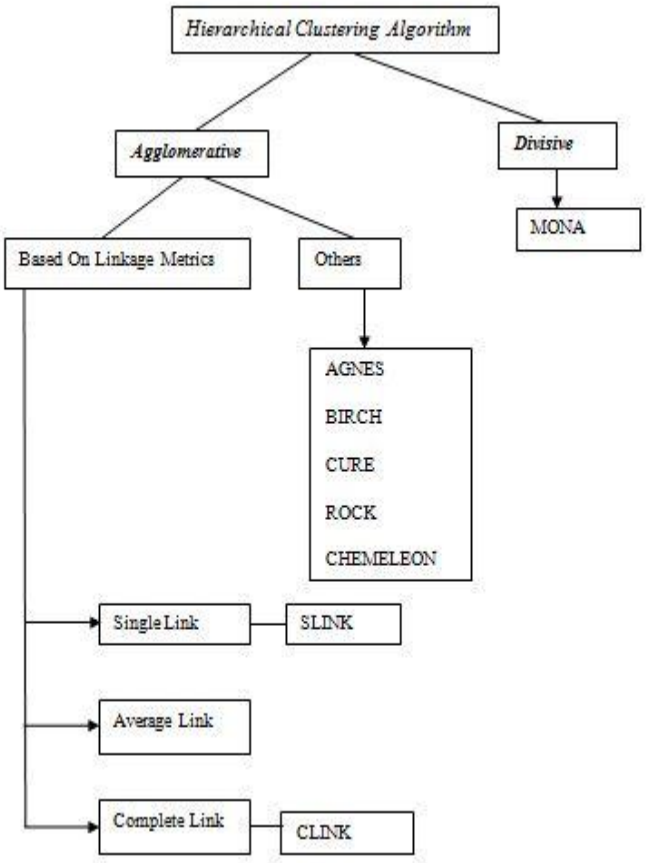


Fig.2: Categorization of HCA

As mentioned earlier no algorithm exist to satisfy all the Requirements of clustering and therefore large numbers of clustering methods proposed till date, each with a particular intension like application or data types or to fulfill a specific requirement.

All Hierarchical clustering algorithms basically can be categorized into two broad categories: Agglomerative and Divisive. The detail categorization of the clustering algorithm is given in figure 2. Though we had tried to provide as much clarity as possible, there is still a scope of variation. The overview of each categorization is discussed below.

2.1. Single Link Algorithm:

This algorithm is type of agglomerative HCA. The single-linkage algorithm use minimum distance to compute the space between clusters; this also known as nearest-neighbor clustering algorithm and SLINK [8]. The algorithm start with every point as a individual cluster and based on distance function the two minimum distance function are merge in single cluster also call strongest links first, then these single links join the points into clusters.

The following table gives a sample similarity matrix for five items (I1 – I5) and the dendogram shows the series of merges that result from using the single link technique.

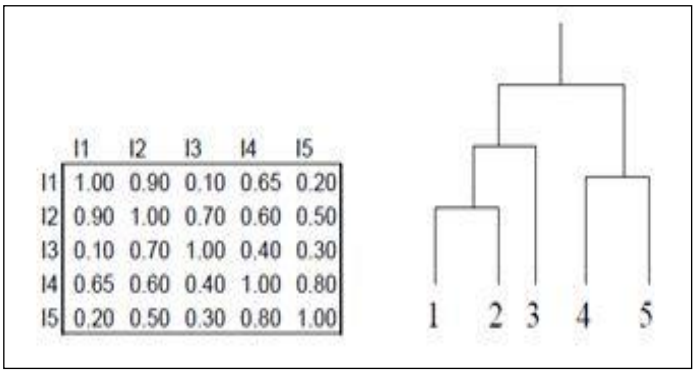


Fig.3: Single Linkage Example

2.1.1 Advantages:

-SLINK is excellent to handling non-elliptical shapes

2.1.2 Disadvantages:

-Sensitive to outliers

2.2. Complete Link Algorithm:

This algorithm is type of agglomerative HCA. This algorithm uses the maximum distance to measure the distance between clusters, it is sometimes called a farthest-neighbor clustering algorithm and CLINK.

The clustering process is ended with maximum distance between nearest cluster exceeds a user-defined threshold; it is called a complete-linkage algorithm.

The following table gives a sample similarity matrix and the dendogram shows the series of merges that result from using the complete link technique.

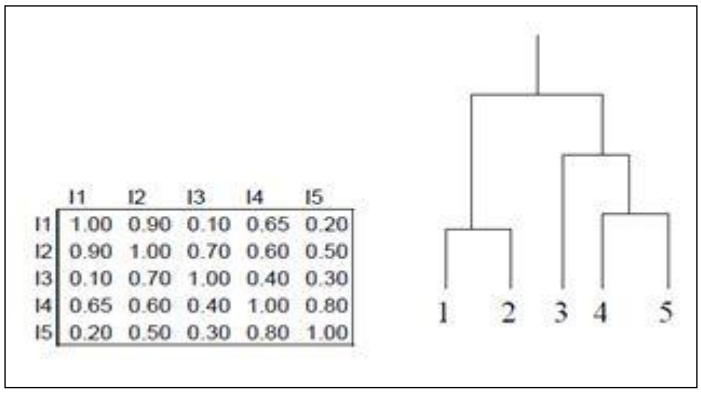


Fig.4: Complete Linkage Example

2.2.1 Advantages:

- Complete-linkage is not strongly affected by outliers
-Handle a large dataset

2.2.2 Disadvantages:

-Trouble with convex shapes

2.3. Average Link Algorithm:

This algorithm is type of agglomerative HCA. In this algorithm the distance between two clusters is defined as the average of distances among all pairs of objects, where each pair is made up of one object from each group. This approach is intermediate approach between Single Linkage and Complete Linkage approach.

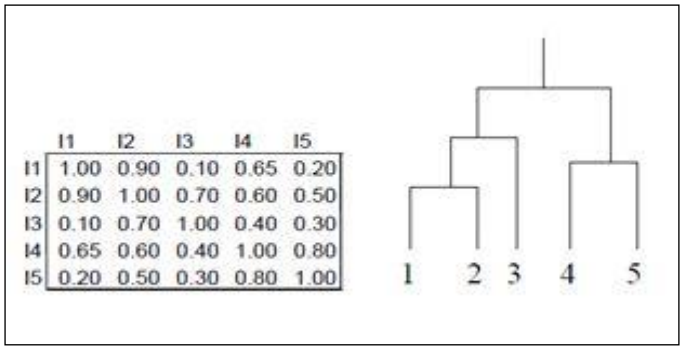


Fig.5: Average Linkage Example

The below table gives a sample similarity matrix and the dendogram shows the series of merges that result from using the group average approach. The hierarchical clustering in this simple case is the same as produced by MIN.

2.3.1 Advantages:

- It can handle categorical and numeric data.

2.3.2 Disadvantages:

- It can fail easily when cluster in hyper spherical shape

2.4. AGNES (Agglomerative Nesting):

This algorithm is type of agglomerative HCA and they used Single Linkage method and the dissimilarity matrix. This is the basic step of AGNES work.

Step 1: Assign each object to a single individual cluster and find the distances among the clusters to the same as the distances (similarities) between the items they contain.

Step 2: Find most similar pair of clusters and merge so now one cluster are less.

Step 3: Calculate similarities (distance) between the new cluster and each of the old clusters.

Step 4: Repeat steps 2 and 3 until all items are clustered into a single Cluster of size N.

In this procedure step 3 can done by many ways like using single-link clustering algorithm, Average-link and also complete link clustering algorithm.

2.4.1 Advantages:

-It can produce an ordering of the objects, which may be informative for data display.
-Smaller clusters are generated, which may be helpful for discovery.

2.4.2 Disadvantages:

-Do not scale well: time complexity of at least O (n2), where n is the number of total objects
-Can never undo what was done previously

2.5. CURE (Clustering Using Representatives):

Clustering Using Representatives (CURE) is an agglomerative method introduced in 1998 by Sudipto Guha. CURE use a constant number of representative points to represent a cluster [9]. It takes a random sample to find clusters of arbitrary and sizes, as it represents each cluster via multiple representative points for small data sets. We summarize our description of CURE by explicitly listing the different steps:

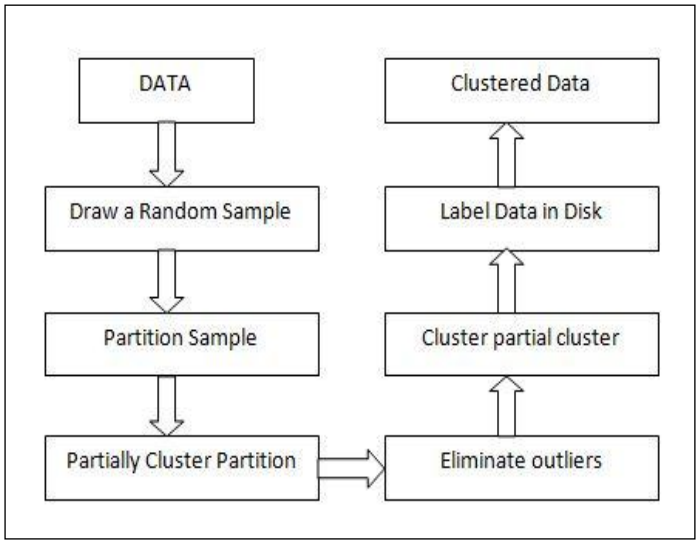


Fig.6: CURE process

1. Draw a random sample.
2. Partition the sample into p equal sized partitions.
3. Cluster the points in each cluster using the hierarchical clustering algorithm to obtain m/pq clusters in each partition and a total of m/q clusters. Some outlier elimination occurs during this process.
4. Eliminate outliers. This is the second phase of outlier elimination.
5. Assign all data to the nearest cluster to obtain a complete clustering.

2.5.1 Advantages:

-Able to handle large dataset

2.5.2 Disadvantages:

-CURE cannot handle differing densities.
-Time Complexity
-cannot handle noise effectively

2.6. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):

BIRCH is an agglomerative method introduced in 1996 by Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH deals with large datasets by first generating a more compact summary that retains as much distribution information as possible, and then clustering the data summary instead of the original dataset [10]. The I/O cost of BIRCH algorithm is linear with the size of dataset: a single scan of the dataset yields a good clustering, and additional passes can be used to improve the quality further but its optional phase [7].

BIRCH consists of a number of phases beyond the initial creation of the CF tree. The phases of BIRCH are as follows:

1. Load the data into memory by creating a CF tree that "summarizes" the data.

2. Build a smaller CF tree if it is necessary for phase 3. T is increased, and then the leaf node entries (clusters) are reinserted. Since T has increased, some clusters will be merged.

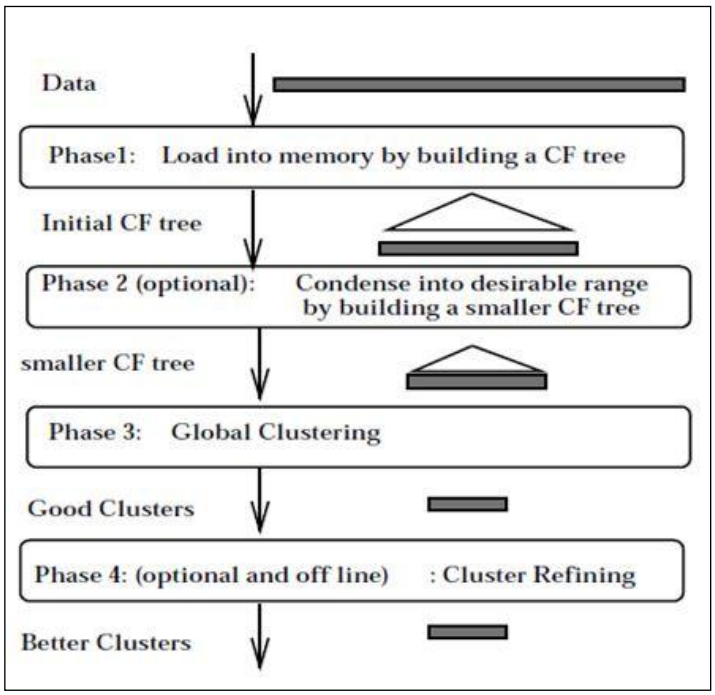


Fig.7: BIRCH process

3. Perform global clustering. Different forms of global clustering (clustering which uses the pair wise distances between all the clusters) can be used.

4. Redistribute the data points using the centroids of clusters discovered in step 3 and thus, discover a new set of clusters. By repeating this phase, multiple times, the process converges to a local minimum.

2.6.1 Advantages:

- Effectively Handle a Outlier
-Incremental Clustering (Dynamic Model)
-Linearly scale when dataset is increase

2.6.2 Disadvantages:

-Handles only numerical data.
-Sensitive to order of data records.
-Favors only clusters with spherical shape.

2.7. ROCK (RObust Clustering using linKs):

ROCK is an agglomerative method introduced in 1999 by S Guha, R Rastogi, and K Shim. ROCK (Robust Clustering using Links) is a hierarchal clustering algorithm to handle the data with categorical and Boolean attributes. ROCK combines, from a Conceptual point of view, nearest neighbor, relocation, and hierarchical agglomerative methods [11].A pair of points is defined to be neighbors if their similarity is greater than some threshold. It handles large number of data and it reduces complexity.

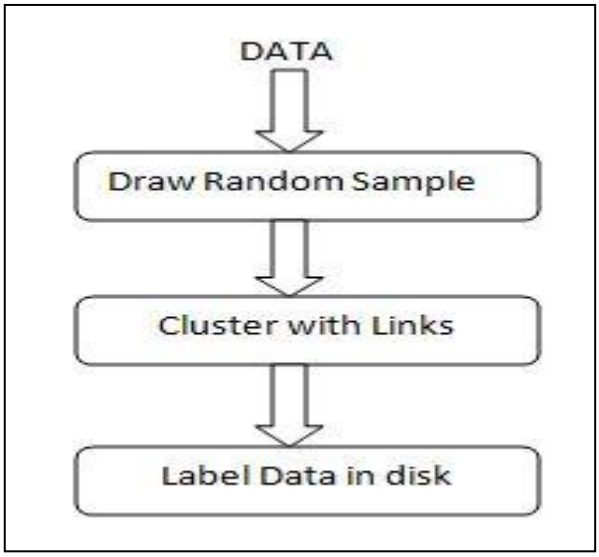


Fig.8: ROCK process

Clusters are generated by the sample points. With appropriate sample size; the quality of clustering is not affected. ROCK performs well on real categorical data, and respectably on time-series data.

2.7.1 Advantages:

-Handle a large dataset.
-Run on real & synthetic data sets
-Effectively handle a Categorical dataset

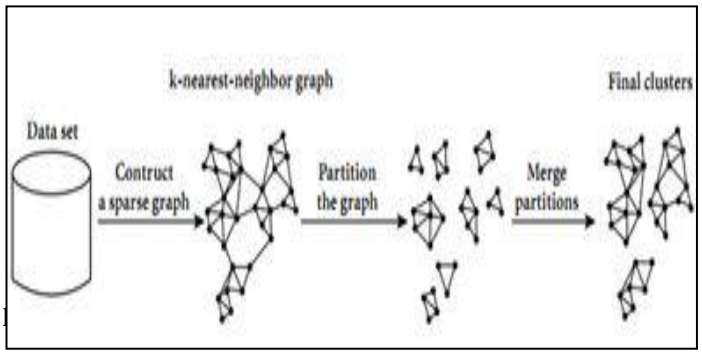
2.7.2 Disadvantages:

-Not support incremental dataset: Static Model

2.8. CHAMELEON (Clustering Using Dynamic Modeling):

CHAMELEON is an agglomerative method introduced in 1999 by George Karypis, Eui-Hong Han, and Vipin Kumar. Chameleon is a clustering algorithm that combines an initial partitioning of the data using an efficient graph partitioning algorithm with a novel hierarchical clustering scheme that dynamically models clusters. Application is spatial data.

Chameleon is a Two-phase clustering algorithm. In first phase it generates a k-nearest neighbor graph (using graph-partitioning algorithm) that contains links only between a point and its k-nearest neighbors [12]. All through the second phase use an agglomerative hierarchical clustering algorithm to find the real clusters by commonly combine together to sub-clusters.



Phase 1: It uses a graph partitioning algorithm to divide the data set into a set of individual clusters.

Phase 2: it uses an agglomerative hierarchical mining algorithm to merge the clusters.

2.8.1 Advantages:

-incremental algorithm

2.8.2 Disadvantages:

-Time complexity of CHAMELEON algorithm in high dimensions is O (n2).

2.9. MONA (Monothetic Analysis):

Mona is a divisive hierarchical clustering method, but it differ from DIANA which can process a dissimilarity matrix and n'_p data matrix of interval scale variables, Mona handle data matrix with binary variables. Each separation is carried out, using a well selected single variable– that is why the algorithm is called monothetic. Many other HCA use all the variables simuntaneously, therefor called polythetic.
The algorithm constructs a clustering hierarchy, starting with one large cluster. After each separation it select one variable and divide the single cluster into two cluster. this process continued until each cluster having only one value. Such clusters cannot be split any more. A final cluster is then a singleton or an indivisible cluster.

3. COMPARATIVE STUDY OF DIFFERENT ALGORITHM:

NAME	PROPOSED BY & YEAR	HIERARCHICAL	FOR LARGE DATASET	SENSITIVE TO OUTLIER	MODEL	TYPE OF DATA	TIME COMPLEXITY
S-LINK	R. SIBSON 1973	AGGLOMARATIVE	NO	SENSITIVE TO OUTLIER	STATIC	NUMERIC	O (N^2)
C-LINK	DEFAYS 1977	AGGLOMARATIVE	NO	NOT STRONGLLY AFFECTED BY OUTLIER	STATIC	NUMERIC	O (N^2)
AVE-LINK	---	AGGLOMARATIVE	NO	---	STATIC	CATEGORICAL, NUMERIC	O (N^2)
AGNES	KAUFMANN, ROUSSEEUW 1990	AGGLOMARATIVE	NO	---	STATIC	NUMERIC	O (N^2)
CURE	GUHA,RASTOGI,SHIM 1998	AGGLOMARATIVE	YES	LESS SENSITIVE TO NOISE	STATIC	NUMERIC	O (N^2 LOGN)
BIRCH	ZHANG, Raghu,LINVY, 1997	AGGLOMARATIVE	YES	Handle noise Effectively	DYNAMIC	NUMERIC	O (N)
ROCK	GUHA,RASTOGI,SHIM 1999	AGGLOMARATIVE	YES	---	STATIC	CATEGORICAL	O (N^2)
CHAMELEON	KARYPIS ,HAN, KUMAR 1999	AGGLOMARATIVE	YES	---	DYNAMIC	DISCRETE	O (N^2)
MONA	---	DIVISIVE	NO	---	STATIC	NUMERIC	O (N^2 LOGN)

3. CONCLUSIONS

In this Survey we study the different kind of Hierarchical clustering techniques in details and summarized it. we included definition, procedure to work of clustering techniques. Paper also gives detail about classification of Hierarchical clustering techniques and their respective algorithms with the advantages, disadvantages and comparative study of all algorithms. So this paper provides a quick review of the different Hierarchical clustering techniques in data mining.

ACKNOWLEDGEMENT

The authors would like to express their thanks to Dr. Amit Ganatra for their comments and suggestions which help to improve the content of paper.

REFERENCES

[1] Jiawei Han, Micheline Kamber, "Data Mining – Concepts and Techniques", Morgan Kaufman Publication, 2001. (Book style)

[2] Rui Xu, Donald Wunsch," Survey of Clustering Algorithms", IEEE, Year: 2005.

[3] Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo**,** A survey of Hierarchical clustering algorithms", The Journal of Mathematics and Computer Science, Year: 2012.

[4] G.Thilagavathi, D.Srivaishnavi, N.Aparna, "A Survey on Efficient Hierarchical Algorithm used in Clustering", IJERT, Year: 2013.

[5] Prof. Neha Soni, Prof. Amit Ganatra, "Categorization of Several Clustering Algorithms from Different Perspective: A Review", ijarcsse, Year: 2012.

[6] Prof. Neha Soni, Prof. Amit Ganatra, "Comparative study of several Clustering Algorithms", International Journal of Advanced Computer Research, Year: 2012.

[7] Nagesh Shetty, Prof. Rudresh Shirwaikar,"A Comparative Study: BIRCH and CLIQUE", IJERT 2013.

[8] R. Sibson," "SLINK: an optimally efficient algorithm for the single-link cluster method", The Computer Journal, Year: 1973.

[9] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim," CURE: an efficient clustering algorithm for large databases", ACM, Year: 1998.

[10] Tian Zhang, Raghu Ramakrishnan, Miron Livny**,** "BIRCH: A New Data Clustering Algorithm and Its Applications", ACM, Springer, Year: 1997.

[11] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim," ROCK: A Robust Clustering Algorithm for Categorical Attributes", Year: 1999.

[12] George Karypis, Eui-Hong Han,Vipin Kumar," CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", IEEE,Year: 1999.

BIOGRAPHIES

"Dipak P Dabhi is an Assistant Professor of Computer Engineering and Information Technology Department at the CGPIT,Uka Tarsadiya University. His research interest include data mining,big data and semantic web".

"Mihir R Patel is an Assistant Professor of Computer Engineering and Information Technology Department at the CGPIT,Uka Tarsadiya University. His research interest include data mining,big data".