Hierarchical Clustering- An Efficient Technique of Data mining for Handling Voluminous Data

Shuhie Aggarwal
M.Tech (Computer Science & Engg.
KIET,Ghaziabad

Parul Phoghat
M.Tech (Computer Science & Engg.
KIET,Ghaziabad

Seema Maitrey
Department of CSE
KIET, Ghaziabad

ABSTRACT

The objective of data mining is to take out information from large amounts of data and convert it into form that can be used further. It comes with several functionalities, among which Clustering is worked upon in this paper. Clustering is basically an unsupervised learning where the categories in which the data to put is not known priorly. It is a process where we group set of abstract objects into similar objects such that objects in one cluster are highly similar in comparison to each and dissimilar to objects in other clusters. Clustering can be done by different number of methods such as-partitioning based methods, methods based on hierarchy, density based ,grid based ,model based methods and constraint based clustering. In this survey paper review of clustering and its different techniques is done with special focus on Hierarchical clustering. A number of hierarchical clustering methods that have recently been developed are described here, with a goal to provide useful references to fundamental concepts accessible to the broad community of clustering practitioners.

Keywords

Data Mining, Clustering Techniques, Hierarchical clustering, Agglomerative, Divisive

1. INTRODUCTION

Clustering is a process where the data divides into groups called as clusters such that objects in one cluster are very much similar to each other and objects in different clusters are very much dissimilar to each other[1][13].

Raw Data
Clustering
Data Cluster

Fig 1: Overview of Clustering

Clustering is useful in pattern analysis, decision-making, machine-learning situations, including data mining, pattern recognition, document retrieval, image segmentation. However in many cases we have a little knowledge about data given, it is under this situation clustering is particularly useful to find inter-relationship amongst data points[2].Clustering is an important task of data mining to divide the data into meaningful subsets and take out information from it[3].A cluster is henceforth a collection of objects which possess high similarity amongst each other and are very much dissimilar to objects belonging to different clusters[4] i.e. in other words inter cluster similarity is low and intra cluster similarity is high. There are several algorithms to do clustering, and the criteria of deciding a particular algorithm mainly depends on three factors which are- data set size, data dimensionality and time complexity.

2. NEED OF CLUSTERING

Clustering is a very important tool that involves analysis of large data of wide variety i.e. the data is multivariate as it comes from heterogeneous sources [5][7].Clustering technique has been employed in wide scientific areas.

Data clustering is being used in following three major areas [6]-

a) Underlying structure- to gain insight into the data, detect anomalies, identify salient features of the data

b) Natural classification- to identify degree of similarity among forms

c) compression- as a method for organizing the data and summarizing it through cluster prototypes

3. REQUIREMENTS OF CLUSTERING ALGORITHMS

Clustering is in itself a challenging field of research in which its potential applications pose their own special requirements. Main requirements of clustering algorithm are [5][10]-

(i) Scalability

(ii) Algorithm's ability to deal with types of attributes

(iii) Discover clusters of arbitrary shape

(iv) Minimum number of requirements for domain knowledge to determine input parameters

(v) Ability to deal with noise and outliers

(vi) Insensitivity of the algorithm ie input records can be fed in any order

(vii) High dimensionality

(viii) Constraint based clustering

(ix) Interpretability and usability

(x) Incremental clustering

3.1 Steps of Clustering Process [12]

(i)Data cleaning and preparing data set for analysis

(ii) Creating new relevant variables

(iii) Selection of variables

(iv) Variable treatment: outlier and missing values

(v) Variable standardization

(vi) Getting cluster solution

(vii) Checking optimality of solution

(viii) Cluster profiling and labeling

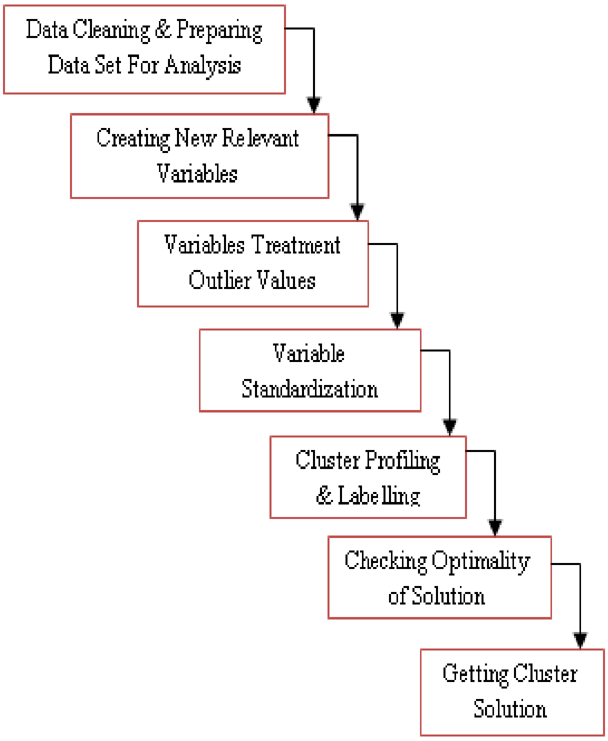
Data Cleaning & Preparing Data Set For Analysis
Creating New Relevant Variables
Variables Treatment Outlier Values
Variable Standardization
Cluster Profiling & Labelling
Checking Optimality of Solution
Getting Cluster Solution

Fig 2: Details of Clustering

Clustering process includes several steps required to be accomplished. First of all data is in raw form as it is unlabeled and coming from heterogeneous sources so it needs to be cleaned meaning all the noisy data is removed ,redundancies are removed .Then from the data we create new relevant variables, now these variables so formed are undergone a transformation and outlier values are removed i.e. all the values that are different from the entire data are removed then variable standardization is performed and from this we get cluster solution now from this we check the optimality of the solution and finally we get the desired clusters and now we do cluster profiling and labeling. Goal of clustering is to determine intrinsic grouping of an unlabeled data. There is as such no good criterion of clustering, this criteria is specified by user that the result of clustering will suit their needs [12].

3.2 Characteristics of good clustering

Good clustering will produce set of clusters with two important properties [8]-

(i) High intra class similarity- this means that similarity between objects in a cluster is very high or objects are very similar to each other

(ii) low inter class similarity-which simply means that objects that belong to different clusters are dissimilar to each other.

4. TYPES OF CLUSTERING

Clustering algorithms are classified as under the following categories [9]:

(i) Hierarchical algorithms
(ii) Agglomerative algorithms
(iii) Divisive algorithms
(iv) Partitioning methods
(v) Relocation algorithms
(vi) Probabilistic clustering
(vii) k-means
(viii) k-medoids
(ix) density based algorithms
(x) density based connectivity clustering
(xi) density functions clustering
(xii) grid based methods
(xiii) constraint based clustering
(xiv) method based on occurrence of categorical data
(xv) clustering algo that are used in machine learning
(xvi) evolutionary methods
(xvii) scalable clustering algorithm
(xviii) algorithm for high dimensional data
(xix) sub space clustering
(xx) projection techniques
(xxi) co-clustering techniques

4.1 Detail of major clustering process

Major clustering algorithms are described as under [9][14]:

a) Hierarchical approach- It works by grouping data objects into a tree of clusters i.e. it performs Hierarchical decomposition, by some particular criteria. It uses several popular methods like Diana, Agnes, BIRCH, ROCK, CURE.

b) Partitioning Approach - Divide the data set into various groups or partitions and evaluate them according to some criteria .E.g. - k-means, k- Medoids, CLARANS

c) Density based approach- Based on connectivity and density functions

d) Grid based approach- It uses a multi-resolution grid data structure. E.g. - STING, CLIQUE

e) Model based- It attempts to optimize fit between given data and some mathematical model. E.g. - expectation minimization, COBWEB

f) Frequent pattern based- It analyses frequent patterns.

g) Constraint based clustering- Real world applications may need to perform clustering under various constraints these are specified by users. So we need to do constraint based clustering.

Amongst all these algorithms we will mainly focus on Hierarchical Clustering. It works by grouping data into tree of clusters, i.e. it performs hierarchical decomposition based on some criteria. It uses distance matrix as a clustering criteria this method requires a specification of termination condition. This type is further divided into Agglomerative clustering which is a bottom up strategy and Divisive clustering which is a top down strategy.

5. HIERARCHIAL CLUSTERING

Hierarchical algorithms can further be classified into agglomerative (which is a bottom up approach) and divisive (which is a top down approach)[11].

a) Agglomerative algorithms – 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.[12]

b) Divisive algorithms – it is the reverse of agglomerative strategy by starting all objects in one cluster and further ahead it simply works on principle of division i.e. it divides until each object forms a separate of its own or till termination condition holds true[12].

Table 1: Agglomerative Vs. Divisive Clustering

Agglomerative Clustering	Divisive clustering
Starts with a single data point	Starts with a big cluster
Add two or more clusters recursively	divide into smaller clusters recursively

a) **AGNES (Agglomerative): Features**: Here single link method is being used, Uses dissimilarity matrix, Merge nodes that have least dissimilarity, Goes in a non descending fashion and Eventually all nodes belong to the same cluster.

b) DIANA (Divisive)

Features: Inverse approach of AGNES and Each node forms its own cluster.

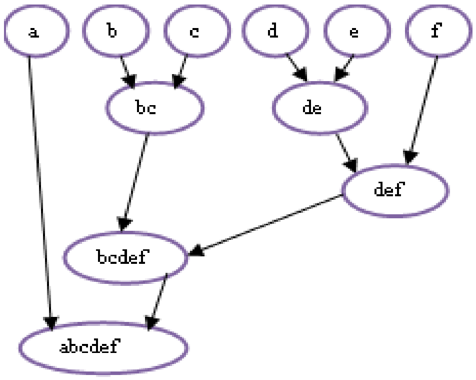

Fig 3: Representation of Hierarchical Clustering

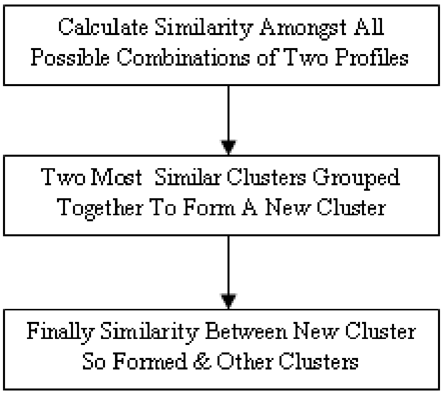
Calculate Similarity Amongst All Possible Combinations of Two Profiles
Two Most Similar Clusters Grouped Together To Form A New Cluster
Finally Similarity Between New Cluster So Formed & Other Clusters

Fig 4: Flow chart of hierarchical clustering

5.1 Commonly used Hierarchical Agglomerative clustering approach

A set of N items is to be clustered and a N*N distance matrix then the basic process of agglomerative clustering is as described [15]

a) Assign each object to a cluster let distance between clusters is same as distance between objects they possess.

b) Find most similar pair of clusters and merge them into a similar cluster

c) Find distance between new clusters and all old clusters

d) repeat steps (b) & (c) until all new clusters are merged into a single cluster of size N[15Hierarchical clustering can be represented by a two dimensional Dendogram, which shows division or fusion made at each step of analysis[15].

Most commonly used hierarchical agglomerative clustering methods and their features are-

(i) **Single linkage**- it will at each step merge the most similar pair of objects that are not yet in the same cluster [16].

(ii) **Double linkage**- uses the least similar pair between each of the two clusters to determine inter cluster similarity [16].

(iii) **Average linkage clustering**- distance between two clusters is average between all pair of objects, each pair has object from each group[16].

(iv) **Average group linkage clustering**- groups once

formed are represented by mean values for each variable, and henceforth mean vector and inter-group distance is defined in terms of distance between two such mean vectors[15]

(v) **Centroid method**- each cluster is represented by co ordinates of group centroid, and at each stage pair of clusters with most similar mean centroid is merged [15]

(vi) **Median** method- similar to centroid method.

In the year 1966, Lance and Williams proposed formula to calculate dissimilarity between new clusters and existing points.Suppose there is an object C_i and other object C_j and both have been combined to form new cluster C_{ij}, the dissimilarity d between the new cluster and currently existing cluster C_k is given by[16]:

$$d_{Ci,j}c_k = \alpha_i dc_ic_k + a_j dc_jc_k + \beta dc_ic_j + \gamma \mid dc_ic_k - dc_jc_k \mid$$

Weaknesses of HACM

- They can never undo what was done previously
- They do not scale well [12]

5.2 Some Other Hierarchical Clustering Methods-

(a)BIRCH (with CF Tree)- Full form is Balanced Iterative Reducing and Clustering using Hierarchies, performs hierarchical clustering over large data sets. It overcomes two difficulties of agglomerative clustering methods. It introduces two things clustering feature and clustering feature tree.

Clustering feature- given n dimensional data points in a cluster, Xi, CF vector of the cluster is defined as a triple CF= (N,LS,SS), where LS is linear sum of data points and SS is square sum of data points[15] [16]

CF Tree- height balanced tree with 2 parameters, branching factor B and threshold T.

i) Each non-leaf has atmost B entries of form [CFi, child i].

ii) Child which is pointer to ith node

iii) CFi a subcluster represented by this child.

A leaf node contains atmost L entries each of the form [CFi], it also has 2 pointers prev and next. Tree size is a function of T.B and L are determined by p where p is page size[15].

This algorithm is implemented in four phases- in first phase it scans all data and builds an initial CF tree and then in second phase it scans all the leaf nodes in the initial CF tree to build a smaller CF tree and removing outliers, after this step, obtain a set of clusters that obtain major distribution pattern . Step 4 provides us an option of discarding outliers[15]

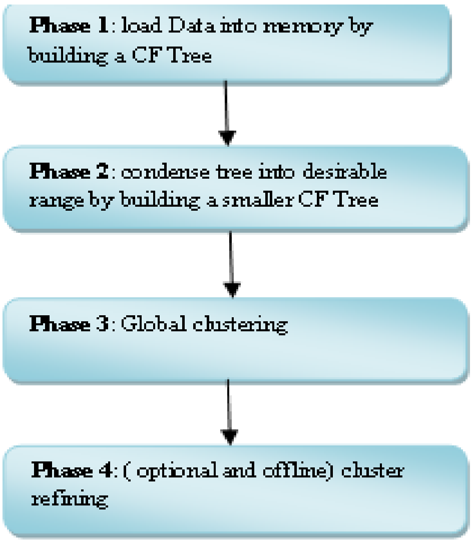
Phase 1: load Data into memory by building a CF Tree
Phase 2: condense tree into desirable range by building a smaller CF Tree
Phase 3: Global clustering
Phase 4: (optional and offline) cluster refining

Fig 5: Overview of Birch

(b) ROCK Clustering Technique- it explores the concept of links i.e. the number of common neighbors between two objects and is meant for data with categorical attributes ROCK algorithm works as follows[16] –

1) Obtain a sample of points from the data set.
2) Compute link value for each set of points.
3) Perform agglomerative hierarchical clustering using maximum number of shared neighbors.
4) Assign remaining points to clusters that have been found[12].

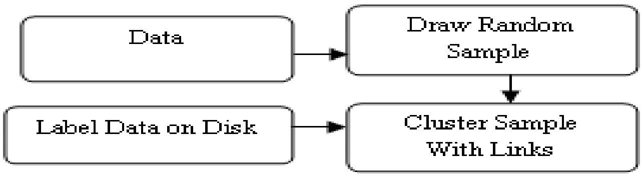

Fig 6: Overview of ROCK

(c) CHAMELEON – It is hierarchical clustering using dynamic modeling. In this cluster similarity is assessed based on how well connected objects are within a cluster and on the proximity of clusters .Two clusters are merged if their interconnectivity is high and they are close together. Chameleon uses a sparse graph in which nodes represent data items, and weighted edges represent similarities amongst the data. . Data sets in a metric space have fixed number of attributes for each data item in it. Chameleon finds the clusters in the data set with the help of a two-phase process [16]. In the first phase, Chameleon uses a graph-partitioning algorithm to cluster the data items into different and comparatively small sub clusters. In the second phase, it uses an algorithm to find the genuine clusters by repeatedly combining these sub clusters. In the clustering process, two clusters are merged only if the inter-connectivity and closeness (proximity) between two clusters are high relative to the internal inter-connectivity of the clusters and closeness of items within the clusters. The approach of dynamic modeling of clusters used in CHAMELEON is applicable to all types of data as long as a similarity matrix can be constructed. CHAMELEON identifies the similarity between a pair of clusters namely, Ci and Cj by evaluating their relative interconnectivity RI (Ci, Cj) and relative closeness RC (Ci, Cj). When the values of both RI (Ci, Cj) and RC (Ci, Cj) are high for any two clusters, CHAMELEON merges those two clusters [16].

(d) CURE CLUSTERING- CURE is more robust to outliers, and identifies clusters having non-spherical shapes and wide variety of size.

Approaches used by CURE clustering algorithms-

CURE employs a new approach middleware between Centroid based (which uses only one point as a representative of cluster) and all points approach (uses all points inside for cluster representation, it is very sensitive to outliers and position of data points). A constant c of well scattered points in a cluster is chosen to be representative [15].

1)cluster with closest representatives is merged at each step.

2)Now do random sampling, overhead to generate random sample is small and it also speed up the algorithm.

3)Partition n data points into p partition (n/p).Partially cluster each partition starting from n/q clusters.

4)A constant c is chosen to catch all possible form that could have the cluster. The cluster with closest pair of representatives is merged at each step.

5)Since outliers are very distant from other points so their possibility to merge with other points is very less.

6)Each cluster so formed is represented by a fraction of randomly selected representative points and points that were removed at the very first step are associated with cluster whose representative point is closer.

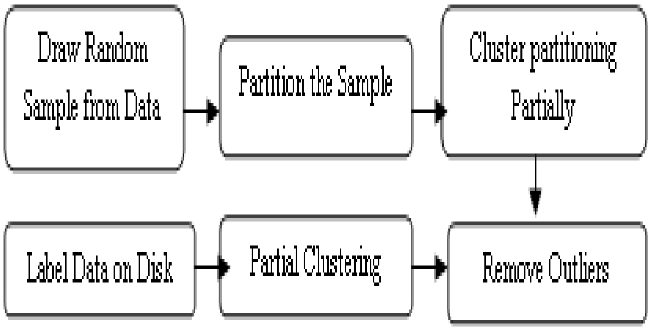

Fig 7: Representation of CURE

6. COMPARING PERFORMANCE OF ALGORITHMS

The hierarchical clustering algorithm BIRCH clusters the data objects by building CF trees for the data sets. The algorithm depends on a threshold value T with which the clustering is executed. Global as well as local clustering is carried out by BIRCH and is main memory based algorithm. Thus it has inherent memory constraint and is effective in handling outliers with an optimum threshold value. It has a linear processing time for clustering data objects [7].

ROCK is clustering algorithm for data with categorical and Boolean attributes. A pair of points is defined to be neighbors if their similarity is greater than some threshold. Use a hierarchical clustering scheme to cluster the data. They assume a static, user supplied interconnectivity model. Such models are inflexible and can easily lead to incorrect merging decisions when the model under- or overestimates the interconnectivity of the data set or when different clusters exhibit different interconnectivity characteristics [15].

CHAMELEON is an agglomerative hierarchical algorithm. The algorithm with inter-connectivity and relative closeness between clusters as parameters, which are checked against an optimum threshold value, produces good quality clusters. The methodology of dynamic modeling of clusters used in CHAMELEON is applicable to all types of data as long as a similarity matrix can be constructed [12].

CURE is one such hierarchical algorithm that depends upon a parameter, the shrinking factor α. The shrinking factor should have an optimum value for effective clustering of the data objects. Further, the shrinking factor reduces the ill effects of outliers considerably. Since the space defined by a single centroid is a sphere, BIRCH labeling phase has a tendency to split clusters when they have non-spherical shapes of non-uniform sizes. Cure is more than 50% less expensive because BIRCH scans the entire data set where CURE sample count must count for a very little contribution of sampling from a large data set. CURE involves two pass clustering. It uses efficient sampling algorithms and scalable for large datasets. Its first pass is partition able, hence it can run concurrently on multiple processors (Higher number of partitions help keeping execution time linear as size of dataset increase).Each step is important in achieving scalability and efficiency as well as improving concurrency. CURE hence has an O (n) space complexity [15][16].

Table 2: Complexities of all Clustering Algorithms

Algorithm	Complexity
Birch	O(n)
Chameleon	O(n^2)
Rock	$O(\max(n^2 m_a, n^2 \log n))$
CURE	O(n)

7. CONCLUSION

The majority of existing clustering algorithms encounter serious scalability and/or accuracy related problems when used on data sets with a large number of records and/or dimensions.

This Paper Demonstrates that CURE can detect cluster with non-spherical shape and wide variance in size using a set of representative points for each cluster. CURE can also have a good execution time in presence of large database using random sampling and partitioning methods. CURE works well when the database contains outliers. These are detected and eliminated.

8. FUTURE SCOPE

CURE uses one of the techniques called sampling, which is very simple to use but produces the biased result which creates confusion in taking the strategic decisions and hence deviates from the main task. So it is our future plan to replace the sampling techniques with one of the best possible technique so that it can provide the correct result and hence accurate decisions can be taken properly.

9. REFERENCES

[1] Mohanraj, M., and A. Savithamani. "A Review of Various Clustering Techniques in Data Mining."

[2] Raymond T. Ng and Jiawei Han. Efficient and effective clustering methods for spatial data mining. In Proc. of the VLDB Conference, Santiago, Chile, September 1994

[3] Soni, Neha, and Amit Ganatra. "Comparative study of several Clustering Algorithms." International Journal of Advanced Computer Research (IJACR)(2012): 37-42.

[4] Marinova–Boncheva, Vera. "Using the agglomerative method of hierarchical clustering as a data mining tool in capital market." (2008).

[5] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.

[6] Anil K. Jain. Data Clustering: 50 Years beyond K-Means 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, December 8, 2008

[7] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[8] Hinneburg, and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pages 58–65, 1998.

[9] Rokach, Lior. "A survey of clustering algorithms." Data mining and knowledge discovery handbook. Springer US, 2010. 269-298

[10] F. Farnstrom, J. Lewis, and C. Elkan. Scalability for clustering algorithms revisited. SIGKDD Explorations, 2: 51–57, 2000.

[11] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. Computer Journal, 26:354-359, 1983

[12] Wikipedia.org

[13] Osmar R. Zaïane: Principles of Knowledge Discovery in Databases - Chapter 8: Data Clusterin P. Smyth, “Clustering using Monte Carlo cross-validation,” in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, 1996, pp. 126–133.g.

[14] P.Indirapriya Dr D.K Ghosh A survey on different clustering algorithms in data mining technique ,IJMER,Vol 3 Issue 1, 2013 pp267-274

[15] Data Mining Concepts and Techniques Jiawei Han and Micheline Kamber

[16] Data Mining Margaret H Dunham