INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Hierarchical Clustering: A Data Mining Method for Efficiently Handling Large Data Sets

RAMESH BOLLI, Research Scholar, Department of CSE, J.S University, Shikohabad, U.P.

Dr. SURIBABU POTNURI, Professor, Supervisor, Department of CSE, J.S. University, Shikohabad, U.P.

Abstract: The purpose of data mining is to extract meaningful information from large databases and then turn that information into a format that can be used. Among its many qualities is clustering, which is the subject of this article. It is one of the features that it has. Clustering, in its most fundamental form, belongs to the category of unsupervised learning techniques. In this method, the categories into which the data should be classified are not known in advance. In it, we cluster sets of abstract objects according to the similarities between them, with items in one cluster being quite similar to one other and things in other clusters being very dissimilar from one another. Clustering may be accomplished by a number of different methods, such as grid-based, model-based, partitioning-based, hierarchy-based, density-based, and constraint-based strategies. In this overview study, the evaluation of clustering and its different approaches focuses mostly on hierarchical clustering as the primary focus of the investigation. Several different hierarchical clustering algorithms that have recently been created are discussed in this article. The purpose of this article is to make fundamental ideas accessible to the vast community of clustering practitioners.

Keywords: **Data Mining, Clustering Techniques, Hierarchical clustering, Agglomerative, Divisive**

1. INTRODUCTION

During the clustering process, the items that are included within one cluster have a high degree of similarity with one another, whilst the things that are contained inside other clusters are quite distinct to one another [1][13]. The following are

ISSN: 2057-5688

examples of applications that make use of clustering in machine learning: data mining, pattern identification, image segmentation, document retrieval, and pattern analysis. When we have a limited amount of knowledge about the data, clustering is a very useful tool for determining the interrelationships that exist between the data points [2].The process of extracting usable information from big datasets by breaking them into relevant subgroups is referred to as data mining, and it is dependent on clustering [3].In other words, the degree of similarity across clusters is low, while the degree of similarity within clusters is high. This indicates that the things that are included within a single cluster are quite similar to one another, but they are significantly different from the objects that are contained inside other clusters. Clustering may be accomplished by a variety of methods; however, while selecting an algorithm, the three most significant factors to take into account are the size of the data set, the dimensionality of the data, and the temporal complexity of the issue.

Fig1: Overview of Clustering

Clustering is an effective technique for evaluating large datasets that include a multitude of distinct variables (multivariate) owing to the fact that these datasets come from a variety of diverse sources. (5) [5]It is [7].There are several scientific subfields that have made use of the clustering approach.

At its core, data clustering may be used to three different applications.[6] It is a the process of determining the underlying structure of the data in order to get a better understanding of it, identify outliers, and highlight important components b) Utilizing natural categorization to determine the degree to which various forms are comparable to one another c) compressing the data by utilizing cluster prototypes as a technique of organizing and summarizing the information.

2. REQUIREMENTS OF CLUSTERING ALGORITHMS

ISSN: 2057-5688

Clustering presents unique challenges due to the nature of the area and the specific needs of its prospective applications. Clustering algorithms mostly need [5][10]

(i) The capacity to scale

(ii) An algorithm's capacity to handle different kinds of characteristics

(iii) Find groups with any form you can imagine

iv) The bare minimum of domain knowledge needed to calculate input parameters

(v) Competence in handling anomalies and unexpected events. The method is not sensitive to the sequence in which the input records are provided

(vii) Lots of dimensions

(viii) Clustering with constraints

(ix) Practicality and ease of understanding

(x) Clustering with increments

A. Steps of Clustering Process

(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

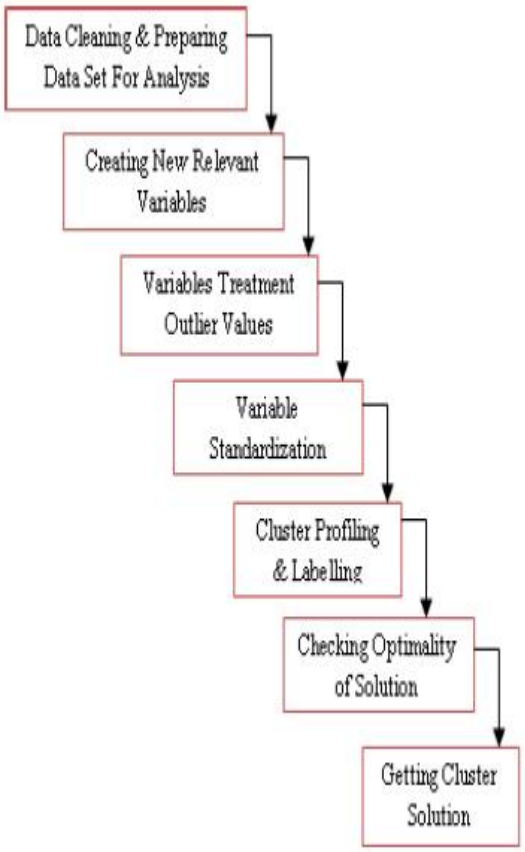

Fig2: Details of Clustering

The clustering procedure is multi-stage and must be completed in a certain order. The data has to be cleaned, in the sense that any redundant or noisy information is

ISSN: 2057-5688

eliminated, because it is currently in raw form, unlabeled, and sourced from a variety of sources. The next step is to extract new relevant variables from the data. After that, the variables are transformed to remove outlier values. Then, the variables are standardised. This yields a cluster solution. From there, we check if the solution is optimal. Finally, we obtain the desired clusters. The last step is to profile and label them. The basic idea behind clustering is to find natural ways to organize unlabeled data. The customer specifies that the clustering outcome should meet their demands, however there is currently no appropriate clustering criterion [12].

B. 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.

C. Details of major clustering process

Major clustering algorithm*s* 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,

ISSN: 2057-5688

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.

3. HIERARCHIAL CLUSTERING

There are two further ways to categorize hierarchical algorithms: agglomerative, which works from the bottom up, and divisive, which works from the top down [11].

a) Agglomerative algorithms start by grouping things into smaller clusters called "atomic clusters." As the algorithm progresses, it merges these clusters into bigger ones until either all objects are in one cluster or the termination condition is met.[12]

b) Divisive algorithms — in contrast to the agglomerative technique, they begin with all objects in a single cluster and proceed by dividing them until they form independent entities or the termination condition is satisfied [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) Agglomerative Granger Neural Networks: Features: In this case, we're using the single link approach; we're using a dissimilarity matrix; we're merging nodes with the lowest dissimilarity; and we're going in a non-descending way, eventually, until all of the nodes are in the same cluster.

b) DIANA Features: The inverse of AGNES, with each node forming its own cluster.

ISSN: 2057-5688

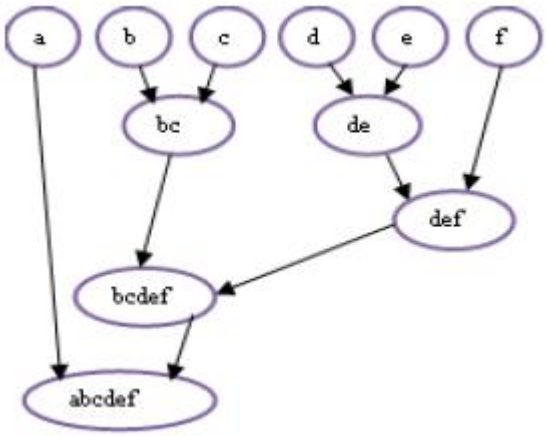

Fig 3: Representation of Hierarchical Clustering

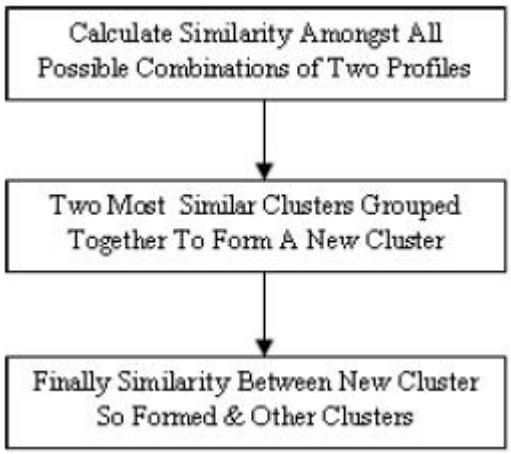

Fig4: Flowchart of hierarchical clustering

A. Some Other Hierarchical Clustering Methods

(a) BIRCH —complete with CF tree — Hierarchical clustering is carried out on massive data sets using Balanced Iterative Reducing and Clustering utilizing Hierarchies. Two problems with agglomerative clustering approaches are eliminated. We learn about clustering features and clustering feature trees in this article. Given n-dimensional data points in a cluster, Xi, the cluster's characteristic feature (CF) vector is defined as a triplet of N, linear sum of data points (LS), and square sum of data points (SS).[15] [16]CF Tree—a two-parameter, height-balanced tree—using branching factor B and threshold T.

i)A maximum of B entries of the type [CFi, kid i] may be found in each non-leaf.
ii) A new node that points to the ith parent node iii) This kid represents CFi, a sub cluster.
A leaf node has two pointers, prior and next, and no more than L entries of type [CFi]. T determines the size of the tree, whereas p determines the size of B and L, where p is the size of the page [15]. Step one involves scanning all data and constructing an initial CF tree. Step two involves scanning all leaf nodes in the initial CF tree to construct a smaller CF tree and eliminate outliers. Step three involves obtaining a set of clusters

ISSN: 2057-5688

that exhibit a major distribution pattern. This algorithm is implemented in four stages. We may choose to ignore extreme cases in Step 4 [15].

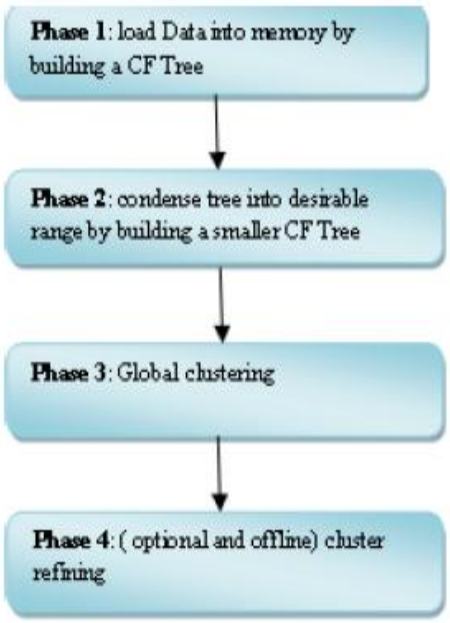

Fig5: Overview of Birch

(a) The ROCK Clustering Technique– It is designed for data with categorical attributes and it delves into the notion of links, which are the number of common neighbors between two items. The ROCK algorithm goes like this:[16] - A 1) Take a small subset of the data points.

2. Determine the link value for every pair of points.

3) Use the largest number of common neighbors to do agglomerative hierarchical clustering.

4) Plot the remaining points according to the discovered groupings [12].

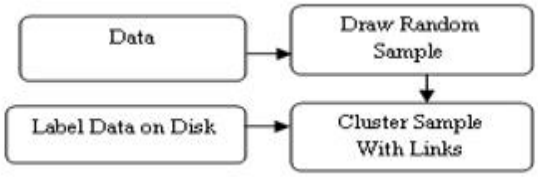

Fig6: Overview of ROCK

(c) Chameleon is a dynamic modelling-based hierarchical clustering system. In this clustering method, the closeness of clusters and the degree of connectivity between items inside them are used to determine how similar they are.High interconnection and proximity between two clusters allow them to be combined. The data pieces in Chameleon are represented by nodes in a sparse network, and the commonalities among the data are shown by weighted edges. Each data item in a metric space dataset has a defined amount of properties. There are two steps that Chameleon uses to locate clusters in the dataset [16]. Initially, Chameleon groups the data items into distinct, smaller clusters using a graph-partitioning technique. Step two involves repeatedly

ISSN: 2057-5688

merging these sub-clusters using an algorithm to identify the real clusters. Clustering only merges two clusters if their respective levels of inter-connectivity and proximity are higher than the internal levels of connectivity and proximity of objects inside the clusters. All data formats may be utilized by CHAMELEON's dynamic cluster modelling technique, provided a similarity matrix can be built. By comparing the relative interconnection RI (Ci, Cj) and relative closeness RC (Ci, Cj), CHAMELEON finds the degree to which two clusters, Ci and Cj, are similar. Two clusters are combined by CHAMELEON when their RI (Ci, Cj) and RC (Ci, Cj) values are high [16].
(d) CURE CLUSTERING—CURE doesn't get thrown off by extreme data points and can spot clusters with irregular forms and sizes.
Methods used by CURE clustering schemes -

In its middleware, CURE bridges the gap between two competing approaches: one that relies on centroid-based clustering (which uses a single point to describe a cluster) and another that utilizes all points within a cluster to represent it (which is very sensitive to outliers and data point positions). A representative sample is defined as a constant c of evenly distributed points inside a cluster [15].
1) At every stage, the cluster with the closest representatives is combined.
2) Carry out random sampling now; the method will run faster, and the overhead to obtain a random sample is modest.
3) Divide the total number of data points, n, into p subsets, then divide each subset by p.Cluster each partition partially, beginning with the first n/q clusters.
4) To capture every potential shape that the cluster may take, a constant c is selected. At each phase, the cluster is merged based on the proximity of its representatives.
5) The likelihood of outliers merging with other points is low because of how far apart they are.

6) Points that were eliminated in the first phase are linked to the cluster whose representative point is closest; each cluster is represented by a percentage of randomly picked points.

ISSN: 2057-5688

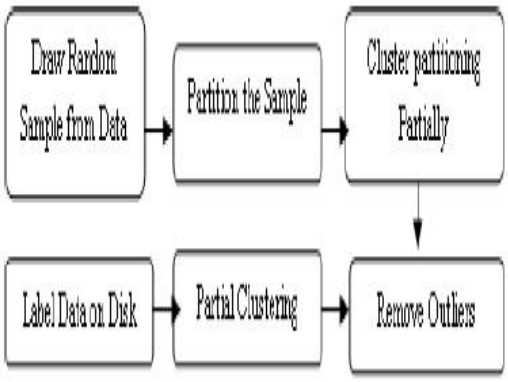

Fig7: Representation of CURE

4. COMPARING PERFORMANCE OF ALGORITHMS

By constructing CF trees for the datasets, the hierarchical clustering algorithm BIRCH groups the data items into clusters. When running the clustering process, a threshold value T is required. As a main memory based method, BIRCH does both global and local clustering. Because of this, it uses an optimal threshold value to handle outliers and has an inherent memory restriction. When it comes to clustering data items, their processing time is linear [7].

Clustering data with Boolean and categorical characteristics is done using the ROCK algorithm. If the degree of resemblance between two points is more than a certain threshold, we say that they are neighbours. Sort the information into groups using a hierarchical clustering algorithm. They take the user-supplied networking model as given, which is static. When the model underestimates or overestimates the data set's interconnectedness or when distinct clusters display differing interconnectivity properties, such rigid models are prone to making the wrong merger judgments [15].

The CHAMELEON algorithm is a hierarchical agglomerative algorithm. The method generates high-quality clusters by using inter-connectivity and relative cluster proximity as parameters, which are then compared to an ideal threshold value. If a similarity matrix can be built, then the dynamic cluster modelling approach employed in CHAMELEON may be used to any form of data [12].

The shrinking factor α is a parameter that determines the behaviour of hierarchical algorithms like CURE. When it comes to grouping data items, the shrinking factor should be set to its optimal value. Additionally, outliers have less of an impact due to the decreasing factor. When clusters exhibit non-spherical forms with non-uniform sizes, the BIRCH labelling phase is prone to splitting them, since the space described by a single centroid is a sphere. Since BIRCH scans the whole

ISSN: 2057-5688

dataset, CURE only has to account for a small fraction of the data when doing its sampling, which results in a cost savings of over 50% compared to BIRCH. In CURE, two pass clustering is used. Scalable to handle massive datasets, it employs efficient sampling methods. Since its first pass may be partitioned, it can use many processors at once. A higher number of partitions allow the execution time to remain linear as the dataset size increases. Achieving efficiency, scalability, and better concurrency requires careful attention to each of these steps. Consequently, the space complexity of CURE is O(n) [15][16].

Table2: 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)

CONCLUSION

Data sets with many records and/or dimensions provide significant challenges for most current clustering techniques in terms of scalability and accuracy. In this paper, we show that CURE, given a collection of sample points for each cluster, can identify clusters with non-spherical shapes and large size variances. Even with a big database, CURE's random sampling and partitioning techniques allow for fast execution times. When there are outliers in the database, CURE performs well. These are identified and removed.

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. ofthe VLDB Conference, Santiago, Chile, September 1994

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

[4] Marinova–Boncheva, Vera. "Using the agglomerative methodof hierarchical clusteringas a data miningtool in capital market." (2008).

[5] Jain, Anil K., M. Narasimha Murty,

ISSN: 2057-5688

and PatrickJ. Flynn. "Data clustering: a review."ACM computing surveys (CSUR) 31.3 (1999): 264-323.

[6] AnilK.Jain.DataClustering:50Years beyondK-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.