Scientific Journal of Impact Factor (SJIF): 3.134

ISSN(Online): 2348-4470
ISSN(Print) : 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1, Issue 11, November -2014

A SURVEY ON ENHANCING AGGLOMERATIVE HIERARCHICAL TECHNIQUES

Komal N. Makadia[1], Prof. Maulik V. Dhamecha[2]

[1]*M.TECH (Computer Engineering) student, RK University, kmakadia05@gmail.com*
[2]*M.TECH (Computer Engineering), RK University, maulik.dhamecha@rku.ac.in*

Abstract: *Clustering algorithms classify data points into meaningful groups based on similarity. Clustering is used in biological and medical applications, computer vision, robotics, and geographical data. A hierarchical clustering is working on grouping of data objects into tree of clusters. This paper is used on the hierarchical clustering for large numerical datasets. This approach is used to clustering starts with each observation as its own cluster and then continually groups the observations into increasing larger groups. Birch does not perform well because of radius or diameter to control the boundary of a cluster. Each node in a CF tree can hold only a limited number of entries because of its size. BIRCH suffers from identifying only convex or spherical shapes of uniform size.*

Keywords *– Data Mining, Hierarchical Clustering Algorithm, BIRCH Algorithm, CURE Algorithm, ROCK Algorithm*

I. INTRODUCTION

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data. Clustering is the unsupervised classification of data into groups. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering techniques apply when there is no class to predict but rather when the instances divide into natural groups. Clustering is used for data analysis, outlier detection, pattern reorganization, and Image processing and market research. Clustering methods focused on scalability, effectiveness, complex shapes and types of data, high dimensional data, numerical and categorical data in large databases. Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree of structure called a dendrogram. Hierarchical Clustering are Agglomerative and Divisive. In Agglomerative Hierarchical Clustering which one starts at the leaves and successively merges clusters together. In Divisive Hierarchical Clustering which one starts at the root and recursively splits the clusters. Hierarchical method relates to the fact that once a step (merge or split) is performed, this cannot be undone. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm is an integrated hierarchical clustering algorithm. . Clusters are to merge or split is determined by a linkage criteria. It uses the clustering features (Clustering Feature, CF) and cluster feature tree (CF Tree) two concepts for the cluster description. BIRCH is an incremental method that does not required the whole database in advance. The paper is organized as follows: Section II describes categories of clustering algorithms. Section III describes Clustering Algorithms. Section IV describes conclusion and Section IV describes future work.

II.CATEGORIES OF CLUSTERING ALGORITHMS

Different clustering algorithms can be broadly classified follows:

Partitioning-based: They divide data objects into a number of partitions, where each partition represents a cluster. They satisfy the following requirements: (a) each group must contain at least one object and (b) each object must belong to exactly one group. In the *k-means* algorithm each cluster is represented by the mean value of the objects in the cluster. In the *k-medoids* algorithm, where each cluster is represented by one of the objects located near the center of the cluster. There are many other partitioning algorithms such as K-modes, CLARA, CLARANS, PAM, and FCM.

Hierarchical-Based: A hierarchical method creates a hierarchical decomposition of the given set of data objects. Hierarchical clustering methods can be classified agglomerative or divisive based on how the hierarchical decomposition is formed. (I) Agglomerative Clustering executes in a bottom–top fashion, which initially treats each data point as a singleton cluster and then successively merges clusters until all points have been merged into a single remaining cluster. (ii)Divisive Clustering initially treats all the data points in one cluster and then split them gradually until the desired number of clusters is obtained. BIRCH, ROCK, Chameleon and CURE are some algorithms.

Density-Based: Data objects are separated based on their regions of density, connectivity and boundary. A cluster is defined as a connected dense component, grows in any direction. So, density-based algorithms are capable of discovering clusters of arbitrary shapes. That provides a natural protection against Outliers. Thus the overall density of a point is analyzed to determine the functions of datasets that influence a particular data point. DBSCAN, DENCLUE,

OPTICS, DBCLASD are algorithms that use such a method to filter out noise (outliers) and discover clusters of arbitrary shape.

Grid-based: Grid-based methods quantize the object space into a finite number of cells that form a grid structure. Grid based structure performed most of the clustering operations. These approach is used it's fast processing time. They are typically independent of the number of data objects and dependent only on the number of cells in each dimension in the Quantized space. The performance of a grid-based method depends on the size of the grid, which is usually much less than the size of the database. Wave-Cluster and STING are typical examples.

Model-based: A method optimizes the fit between the given data and some (predefined) mathematical model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. Also, it leads to a way of automatically determining the number of clusters based on standard statistics, taking "noise" or outliers into account. Emmis an algorithm that performs expectation-maximization analysis based on statistical modeling. COBWEB is a conceptual learning algorithm that performs probability analysis and takes *concepts* as a model for clusters.

Linkage functions:

Linkage functions are hierarchical methods that merging of clusters is based on distance between clusters. There are **S**ingle-link, **Ave**rage-link and Complete-link .They are also agglomerative hierarchical algorithms.

A. Single-linkage clustering: The distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster. It is denoted S-link.

Min (d(x,y):x ∈ A,y ∈ B)

B. Complete-linkage clustering: The distance between one cluster and another cluster to be equal to the greatest distance from any member of one cluster to any member of the other cluster. It is denoted Com-link.

Max (d(x,y):x ∈ A,y ∈ B

C. Average-linkage clustering: The distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. It is denoted Ave-link.

III. CLUSTERING ALGORITHM

3.1 BIRCH:

BIRCH is designed for clustering a large amount of numerical data. It is an integration of agglomerative hierarchical clustering. It overcomes the two difficulties scalability and the inability to undo what was done in the previous step. BIRCH introduces two concepts, (I) *clustering feature* and (ii) *clustering feature tree* (*CF tree*), which are used to summarize cluster representations. The structures are helping the clustering method to achieve good speed and scalability in large databases. It also make effective for incremental and dynamic clustering.

(I)A clustering feature (CF): A clustering feature is a three-dimensional vector, which summarizes the information off a cluster. Given *n d*-dimensional objects or points in cluster, {***xi***}, then the CF of the cluster is defined as

CF = (*n*, ***LS***, ***SS***)

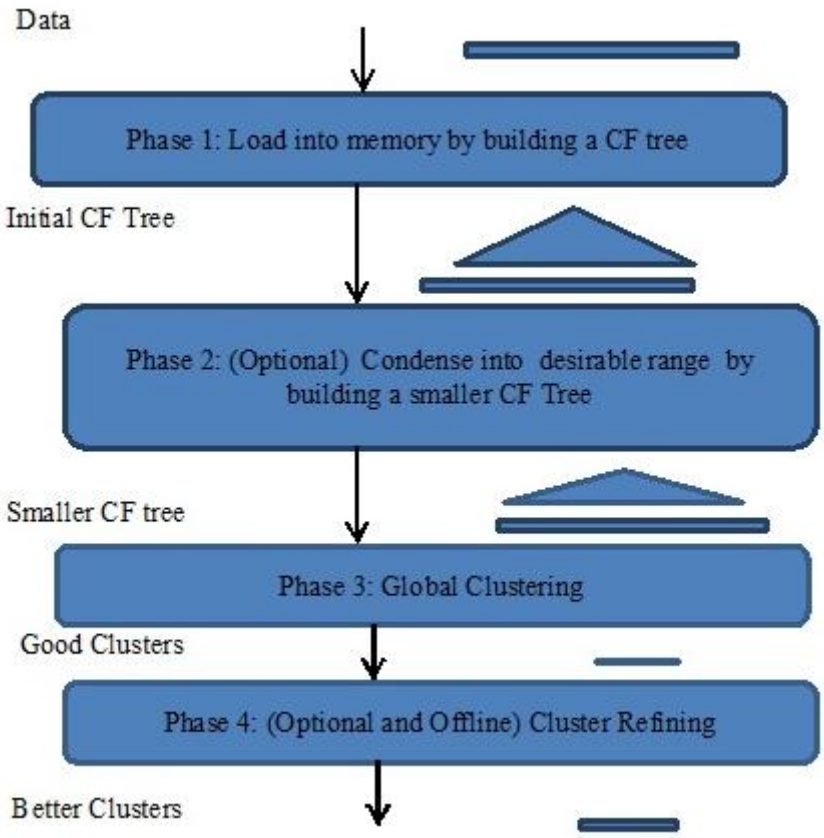
Where n=number of points in the cluster

LS=Linear sum of the n points= $\sum_{i=1}^{n} X_i$

SS=Square sum of the n points= $\sum_{i=1}^{n} X_i^2$

(ii)CF tree: A CF tree is a height-balanced tree. It stores the clustering features for a hierarchical clustering. A CF tree has two parameters: *branching factor*, *B*, and *threshold*, *T*. The branching factor defines the maximum number of children per nonlife node. The threshold parameter defines the maximum diameter of sub clusters stored at the leaf nodes of the tree. These two parameters influence the size of the resulting tree.

Fig 1 shows the overview of BIRCH algorithm. It consists of four phases. Phase 1 starts to scan all the data with initial threshold value and inserting points into the tree. If it runs out of memory before it finishing to scanning the data, it increases the threshold value .It rebuilds an initial memory CF tree using the given amount of memory and recycling space on disk. CF tree is trying to reflect the clustering information of the dataset under the memory limit. Phase 2 is a bridge between Phase 1 and Phase3. It scans the leaf entries in the Initial CF tree to rebuild a smaller CF tree, while removing more outlier and grouping crowded sub clusters into larger ones. Phase 3 is applying global clustering algorithm to the sub-clusters given by leaf entries of the CF tree. This phase improves clustering quality. Phase 4 is a scan the entire dataset to label the data points. This phase is Outlier handling. BIRCH is appropriate for very large databases by making the memory constraints and time explicit. BIRCH can be used to help solve real-world problems. BIRCH performs on some real datasets.

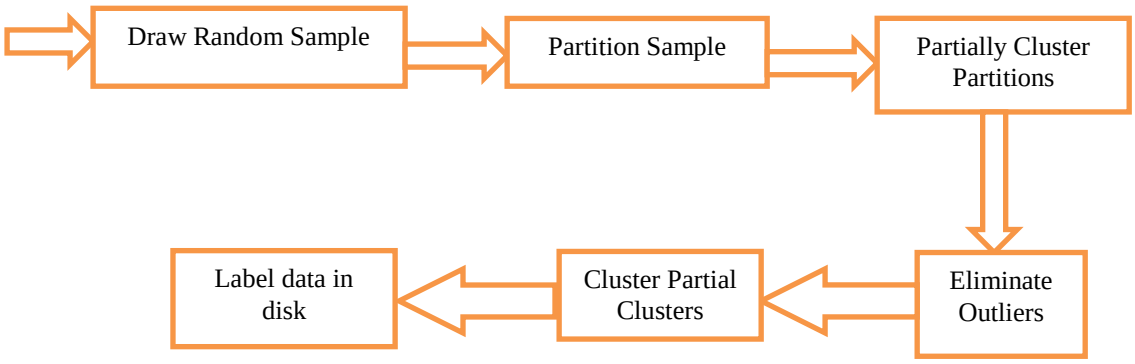


[Fig1.An Overview of BIRCH algorithm]

If the cluster shapes are not spherical then BIRCH does not perform well. Each node in a CF tree can hold only a limited number of entries because of its size. BIRCH has a stability problem. It suffers from high computational time requirement.

3.2 CURE:

CURE can identify spherical and non -spherical clusters .It is an agglomerative hierarchical clustering algorithm where disjoint clusters are successively merged until the number of clusters reduces to the desired number of clusters.

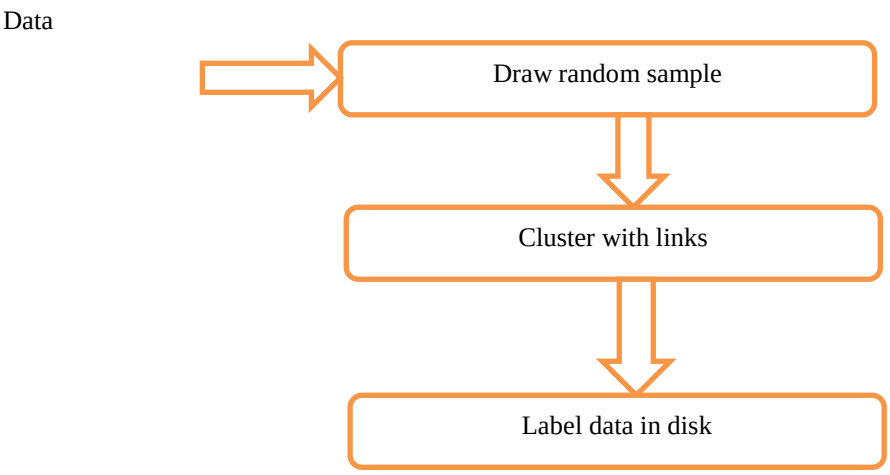


[Fig 2. Six steps in CURE]

Each step is important in achieving Scalability and efficiency. It is also improving concurrency. CURE uses random sampling and partitioning to speed up clustering. It can handle larger databases efficiently. It is wide variances in size. It is less sensitive to outliers.

3.3 ROCK:

ROCK is Robust Clustering using links. It is a Hierarchical Clustering Algorithm for Categorical Attributes. It explores the concept of links for data with categorical attributes. ROCK is more robust than other clustering algorithm that focused only on point similarity.



[Fig 3. ROCK Algorithm]

ROCK is described in the figure. A random sample is drawn from database. A Hierarchical Clustering algorithm uses links instead of distances. ROCK algorithm employs links is applied to the sampled points. It means iteratively merge clusters C*i and* Cj that maximize the goodness function

G (p1, p2) = total number of crosslink's / expected number of crosslink's

And stop merging once there are no more links between clusters or the required number of clusters has been reached. Finally, the clusters involved only sampled points. They are used to assign the remaining data points on disk to the appropriate clusters. ROCK runs on real and synthetic datasets such as voting. Real data used for comparison to traditional algorithms. Synthetic data used to demonstrate scalability. ROCK performs time series data.

IV. CONCLUSION

In this paper, a various agglomerative algorithms are defined. BIRCH is an incremental model. A combination of random sampling and partitioning allows to handle large dataset more effectively. CURE adjusts well to clusters having non-spherical shapes and wide variances in size. CURE can handle large databases efficiently. New algorithm is generated fast and stable increment approach. The major advantage of BIRCH is capable to perform fast execution of clustering process.

V. FUTURE WORK

We used on Hybrid approach for new algorithm. New algorithm is generated by covered disadvantages of BIRCH algorithm. The future algorithm works with following steps: 1. Creating Random Sample from large dataset. Random Sampling is used to reduce the size of the input to algorithm for handle datasets.. 2. Partitioning sample and Construct CF tree. First cluster in partitions then merge partitions and then constructing of tree makes data traverse easily.. 3. Partially clustering of partitions. 4. Eliminate Outliers from non-constructed nodes. Outliers are partially eliminated and spread out by random sampling, are identified because they belong to small clusters that grow slowly. 5. Global clustering (Re clustering of partitions). 6. Labeling Output. This will benefits a lot in clustering as it divides the data in partitions so that small division can be clustered accurate and fast. New algorithm producing is more accurate and efficient.

REFERENCES

[1] A. Fahad, N. Alshatri, Z. Tari, A. Alamari, I. Khalil A. Zomaya, Foufou, and A. Bouras, "A Survey of Clustering Algorithms for Bigdata: Texonomy & Empirical Analysis", IEEE Transaction on Emergining Topics in Computing 2014.

[2] Deepak Gupta, Vinay Kr. Goyal, Harish Mittal, "Estimating of Software Quality with Clustering Techniques" Third International Conference on Advanced Computing Communication Technologies 2013.

[3] Er. Armpit Gupta, Er.Ankit Gupta, Er. Amit Mishra , RESEARCHPAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS Publication: International Journal of Advance Technology and Engineering Research (IJATER),2011.

[4] Tian Zhang, Raghu RamKrishnan, "BIRCH: An Efficient Data Clustering Method for Very large databases",TechnicalReport, Computer Science Dept., Univ. of Wisconsin-Madison,1995.

[5] Prof. Mrs. J.R. Prasad, Prof. R.S.Prasad, Dr. U.V.Kulkarni, "Impact of Feature Selection Methods in Hierarchical Clustering Technique: A Review", Proceeding of the International MultiConference of Engineers and Computer Scientists 2008 Vol I IMECS 2008 ,19-21 March,2008,Hong Kong.

[6] Sudipto Guha, Rajeev Rastogi and Kyuseok Shim, "CURE: An Efficient Clustering Algorithm For Large Databases", Elsevier Science Ltd, Information System Vol. 26,No. 1,pp. 35-58,2001.
[7] Rahmat Widia Sembring, Jasni Mohmad Zain, Abdullah Embong, "A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course" in Journal of Computing, Voume 2, Issue 12, December 2010.
[8] S.Guha, R.Rastogi and K. Shim, "ROCK : A Robust Clustering Algorithm for Categorical Attributes", School of informatics.
[9] I.K. Ravichandra Rao (2003), "Data Mining and clustering Techniques", DRTC Workshop on Semantic Web,Banglore.
[10] Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh ,Nasibeh Emami Chukanlo, "A survey of hierarchical clustering Algorithms",The Journal of Mathematics and Computer Science 20
[11] H. Han and Kamber,"Data Mining : Concepts and Techniques", Morgan Kaufmann Publishers,2001.

