ISSN No:-2456-2165

Evaluation of BIRCH Clustering Algorithm for Big Data

Akshatha S R
Master of Technology
Department of Computer Science and Engineering
B.N.M Institute of Technology
Bengaluru, Karnataka 560050

Dr. Niharika Kumar
Associate Professor
Department of Computer Science and Engineering
B.N.M Institute of Technology
Bengaluru, Karnataka 560070

Abstract:- **Clustering algorithm are as of late recapturing consideration with the accessibility of vast datasets. Many clustering algorithm don't scale well with the expanding dataset and requires the cluster count parameter which is difficult to assess. In a BIRCH Clustering algorithm, computing the threshold parameter of BIRCH from the data is evaluated and the issue of Scalability with increasing dataset sizes and the cluster count is solved.**

Keywords:- *BIRCH Clustering Algorithm, Tree-BIRCH, Flat Tree.*

I. INTRODUCTION

Cluster is a subset of objects which are same or similar. Data clustering is an unsupervised learning where the set of data points are grouped in such a way that the data points in a group are similar to each other than to those in other groups. The clustering algorithms are k-means clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), etc. Many clustering algorithms don't scale well with the dynamic (increasing) dataset and the algorithms requires the number of clusters to be formed as an input. The prediction of the cluster count is not so easy where it needs the information, such as data exploration, feature engineering and cluster document.

Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) is an unsupervised and data mining clustering algorithm which is introduced by Zhang et al [8]. This is the high speed clustering algorithm available. It is a data mining algorithm which is used to form a cluster over categorically immensely colossal data-sets. Birch cluster algorithm works on only numerical dataset. Favorable position of BIRCH is its ability of clustering with the increasing records in a dataset, multi-dimensional metric information indicates in an endeavor create the good quality clustering for a given arrangement of assets (memory and time imperatives). Much of the time, BIRCH just requires a solitary output of the database. It is neighborhood in that each clustering choice is made without examining all information focuses and right now existing groups. It misuses the perception that information space isn't normally consistently possessed and only one out of every odd information point is similarly imperative. It makes full utilization of accessible memory to infer the finest conceivable sub-clusters while limiting I/O costs. It is additionally an incremental strategy that does not require the entire dataset ahead of time.

BIRCH already solved the scalability problem for large dataset. To improve the quality of the cluster, the number of cluster is to be given as an input. The main aim is to achieve the quality and the speed of the cluster without giving the cluster count. This can be achieved by deleting the global clustering phase which is done at the end of the BIRCH clustering algorithm. It has four phases: the first phase of the algorithm is loading the data into the memory, second is condensing tree, third is global clustering and the last is cluster refinement. The cluster count is needed at global clustering step so it is removed and yet preserve the cluster quality and speed. The other part of the BIRCH clustering algorithm is called as tree-BIRCH

The tree-BIRCH may go wrong in dividing the parent node, splitting of clusters and cluster merging. This is solved by computing optimal threshold parameter. The threshold parameter is computed by minimum distance and the maximum cluster radius. This is calculated automatically which is done in Bach and Jordan [9]. Tree-BIRCH is quick and it is used for the online clustering which is the elimination if global clustering step. To solve the supercluster splitting, the flat tree is introduced. The flat tree is the tree where it has a parent node and all the child nodes are attached to that parent node. This is achieved by keeping the branching factor infinite. The parent node will be having the information about all the child nodes that is, the parent node maintain the triples total number of data point, linear summation and the square summation of the child node. When new data point is entered, the triples are updated which is built in a clustering feature tree. Tree-BIRCH clusters faster than the k-means clustering algorithm for the same size of the dataset [2].

Dataset is also called as informational index which gathers the information. Each rows in a dataset is seems to be one record. There are many rows and columns it consists. It is a collection of the records. The informational index records esteems for every one of the factors. Each esteem is known as a datum. The informational index may involve information for at least one individuals, relating to the quantity of columns. The term informational collection may likewise be utilized all the more freely, to allude to the information in a gathering of firmly related tables, comparing to a specific investigation or occasion. Here, it is approach to aim at datasets from two-dimensional. The

geospatial data is being used for BIRCH clustering algorithm. Geospatial data includes the location, size, shape of the object on the earth like mountain, lake, etc.

II. RELATED WORK

The scalability problem is solved in k-mean clustering algorithm by parallelization [2]. Here the work is done on two types of k-means, first one is sequential k-means and another is the parallel k-means. In sequential k-means, the MacQueen[3] is used where the clusters are formed by performing the computations many times. At the starting stage the centroid is selected randomly from those data points. Once this is done, the selected centroid is calculated multiple times to select a perfect centroid where it includes all the data points. Euclidean distance equation is used to calculate the centroid. The second is parallel k-means where it make use of Modha [4] which is a distributed memory multiprocessors. In this method the data points are divided equally based upon the processors and then the calculations are done on the divided subset. Once this is done then the centroid is calculated on the divided subset. The parallel k-means runtime is much faster than the sequential k-means. Graphical processing unit is a memory used here which is a shared multiprocessor. Processors are also called as thread which results in master-slave relationship where the master will select the centroid. Labelling step in a Graphical processing unit seems to be more advantageous of the k-means for huge amount of dataset and more number of cluster counts. Parallel k-means through Compute Unified Device Architecture is also taken care. This is tested only for the limited available memory and the issue of scalability is solved by the parallel k-means.

The improved version BIRCH in the threshold value is studied from the improved multi threshold birch clustering algorithm [5]. BIRCH uses clustering feature (CF) in each clustering feature tree. It is used for the huge amount of dataset. The clustering feature consists of three attribute that is N, LS and SS. N represents the total number of data points in a cluster, LS represents the summation of the data points in a cluster and SS represents the square summation of the data points in a cluster. The clustering feature is maintained at each node with a static threshold value which becomes a problem. In paper [5], another element threshold value T is added to the clustering feature. This is done to make use of multiple threshold value in an algorithm. In multi threshold birch algorithm, the clustering feature has four attributes are N, LS, SS, T. The threshold value is given at the initial stage. When the new data point is entered, it has to check for the nearest node in a clustering feature tree. Once the near leaf node is found then it has to check for the threshold value. If the threshold value is not violated then the newly entered node is assigned to that nearest node and the threshold value is updated to the clustering feature and the updated threshold value is considered as the new threshold value. If the threshold value is violated by the entered node, then increase the threshold value by multiplying the threshold value with the modified factor of the threshold. Once it is multiplied then the threshold value becomes larger enough to enter the new node. If the threshold value is small then the cluster is divided and if the value of the threshold is increased then the clusters are merged. In multi threshold birch the value of the threshold is dynamic. It keeps changing as the cluster operator.

The clustering algorithm Density-Based spatial clustering of application with noise (DBSCAN) is proposed [6]. The DBSCAN clustering algorithm visits all the data points many times. The structure of the clustering is arbitrarily shaped because spatial databases are spherical. This clustering algorithm has three points. Core point, border point and noise point. Eps represents the radius of the cluster and minpts is the minimum number of point in the cluster. Consider if the minpts value is 4 and the point in a cluster is greater than the minpts then it is called as core point. The data point which are less than minpts and those point are within the radius Eps those are the neighbors of the core point are called border point. The noise points are those which are not a core point as well as the border point. This clustering algorithm need density as an input parameter. Density parameter gives the maximum number of points a cluster can have within a radius Eps. This clustering algorithm handles the noise and the value of Eps and minpts of the clusters should be known. If one point of the cluster is known then all points is checked either it density reachable or not. In this algorithm, the global values are used for the Eps and minpts parameter throughout the cluster formatting. The cluster with least density parameter is considered as a good cluster. These clusters are said to be a good candidates for the global values which has no noise in it. The clusters are combined only if the density of two clusters are near to each other. The evaluation is done on real data and synthetic data of SEQUOIA2 000 benchmark.

In hierarchical clustering [7], the cluster is formed in a tree manner. It has two strategy: The first is Agglomerative clustering and the second is divisive clustering. The hierarchical agglomerative clustering is an algorithm which is also called as Single-linkage (SLINK) and complete linkage (CLINK) in rare cases.it is a leave to root node concept. The agglomerative clustering algorithm is a bottom up approach where the clusters are merged and has time complexity O(n3) and needs memory O(n2). The maximum distance between the data points of the clusters are known as CLINK and the minimum distance between the data point of the clusters are known as SLINK. The mean distance between the data points of the clusters are known as average linkage clustering. The variance is calculated and wen it is increased, the two sub-clusters are combined. The process of clustering is ended with the small number of clusters. The other strategy is Divisive clustering algorithm which is the top down approach where the clusters are splits as moves down $O(2^n)$. It is a root to leaves concept. The merging and the splitting of the cluster are done in greedy manner where the result is presented in dendrogram. A dendogram is a tree diagram which is used to arrange the clusters from the output of hierarchical agglomerative clustering and divisive clustering. The BIRCH clustering algorithm and its application is studied [1].

III. PROPOSED MODEL

The dataset is termed as a geospatial data of zones. Each sample in the database is said to be as an instance. Computations takes place at each instance of the dataset. Fig. 1 shows the architecture of the BIRCH where the dataset is the first phase and the BIRCH takes only numerical values. Some of the geographic information system of geospatial data are Natural Earth-Vector, Global map, Land and Ocean Boundaries, Elevation, Weather and Climate, Hydrology, Snow, Natural disaster, Land cover, Forest geographic information system and etc. The second phase is the clustering phase where the tree-Birch and clustering feature are exercised. Tree-BIRCH clusters 200,000 data points into 2000 clusters in a couple of seconds which leading to faster convergence. The output of this step gives the cluster data. Tree-BIRCH is the fastest clustering algorithm available.

The characteristics defined from all the mechanisms focuses on a keyword called tree-BIRCH, which is the proposed technique. BIRCH takes three parameter that is, branching factor, threshold, and cluster count. When the data point is entered, the clustering feature tree and hierarchical tree is built. In a hierarchical tree, the each node seem to be a cluster. Intermediate nodes are the parent clusters and the leaf nodes are the actual clusters. Branching factor says how many child nodes a parent can have and if it exceeds then the parent node has to split up. The dataset used is spatial dataset which is given as an input to the clustering phase. In clustering phase, the BIRCH clustering algorithm will scan the dataset and process to form a clusters. the advantages of BIRCH clustering algorithm is it takes only single scan of whole data and saves the time of scanning repeatedly as done in other clustering algorithms. Once the data is loaded, then next step is the formation of clustering feature where it has the information about the sub-clusters. Once all the data points are evaluated, the cluster formation is done. The tree-BIRCH can be used for the online clustering algorithm. The radius of the clusters should not be more than the threshold value.

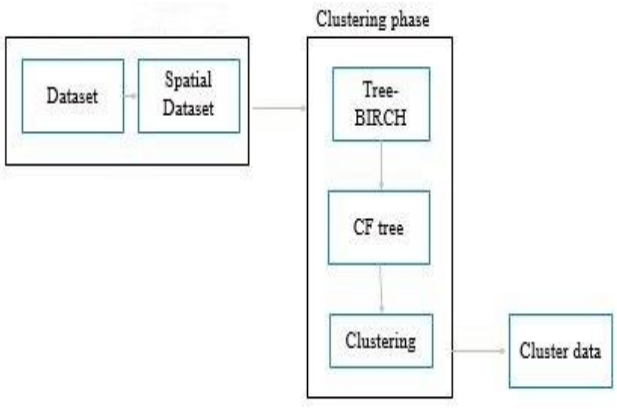

Fig 1:- Architecture

The Clustering Feature tree formation is done depend on the entry of the data points. If the data point of a cluster entered in increasing the distance from the center, tree-BIRCH will probably return only one group. If two data point are of opposite side then the cluster is split into two. If two points are from different cluster but near to each other form a single cluster when the given threshold is large. Decreased in threshold parameter will also decrease in cluster combining. Increase in threshold parameter will reduces the clustering splitting. The inadequate clustering can't be affected by the decision of the threshold value. This happens if a cluster covers with two areas having a place with two non-leaf node in the CF-tree is a height-balanced tree. This sort of mistake is supercluster splitting. To avoid this supercluster splitting, the flat tree is introduced where all the non-leaf are the children of the root node. The dataset with huge amount of data with large ratio of a cluster and distance to the cluster radius produces less error. Fig. 2. Shows cluster radius and the cluster distance.

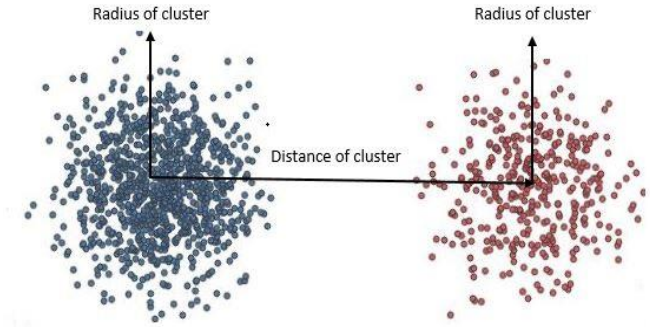

Fig 2:- Cluster radius and Cluster distance

To deduce the formula for the ideal limit given two conditions: first assume that the branching factor is picked sufficiently extensive for the BIRCH tree to be level, and second, that every one of the clusters have roughly a similar number of components. All the cluster has the same number of elements will be the flat tree.

The objective is to acquire the ideal limit parameter as a function of the maximum radius Rmax of the cluster and the minimum cluster distance Dmin, where both are known. The underlying presumption is that those two qualities are either definitely known or simple to get.

IV. IMPLEMENTATION

The radius of the cluster and the distance of the cluster is calculated based on the clustering feature which is structured at each node of CF tree.

CF = (N, LS, SS).

Where, *N* represents the number of data points in a cluster, *LS* represents the linear summation of the data points and *SS* represents the square summation of the data points. Based upon this clustering feature the centroid of the cluster, radius of the cluster and distance between the clusters are calculated.

$$LS = \sum_{j=1}^{N} \overrightarrow{Y_j} . \quad (1)$$

Where Y_j indicates the data points from the dataset. The calculation of LS an SS is shown in the equation 1 and equation 2.

$$SS = \sum_{J=1}^{N} (\overrightarrow{Y_j})^2 . \quad (2)$$

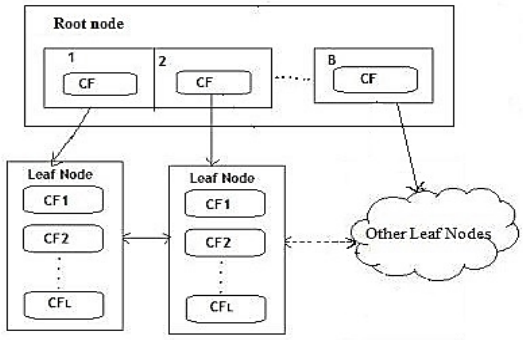

Fig 3:- CF tree

The structure of the clustering feature for flat tree is shown ion the Figure 5.3. It shows the flat tree where all the child nodes (sub-clusters) are assigned to the single parent node. The following equations are used for the calculation cluster centroid C, radius R and distance between the clusters D.

$$\text{Centroid: } C = \frac{LS}{N} . \quad (3)$$

The centroid of the cluster is calculated by dividing the summation of the linear sum of data point with the total number of data points in a cluster. Equation 3 is used for the calculation of cluster centroid. . Equation 4 is to calculate the radius of the cluster.

$$\text{Radius: } R = \sqrt{\frac{\sum_{j=1}^{N} (\overrightarrow{Y_j} - C)^2}{N}} =$$

$$R = \sqrt{\frac{N*C^2 + SS - 2*C*LS}{N}} . \quad (4)$$

To calculate the distance between the clusters, consider the clustering feature of two clusters as shown in equation 5.

$CF_n = [N_n, LS_n, SS_n]$ and $CF_m = [N_m, LS_m, SS_m]$:

$$D = \sqrt{\frac{\sum_{j=1}^{N_n} \sum_{k=1}^{N_m} (\overrightarrow{Y_j} - \overrightarrow{Z_k})^2}{N_n * N_m}} =$$

$$\sqrt{\frac{N_n * SS_m + N_m * SS_n - SS - 2 * LS_n * LS_m}{N_n * N_m}} . \quad (5)$$

If *Dmin* is clearly larger than *Rmax*, it would be beneficial to increase the threshold beyond. While the lower bound on the threshold is nearly the same for all cluster distances, the upper bound increases linearly, roughly with half the increase of the distance. The following expression is used for choosing the threshold.

$T = D_{\min} + R_{\max}$ Where, *Dmin* ≥*Rmax*.

Flat tree for the cluster feature tree is achieved by making the branching factor value infinite or maximum number which should be more than the number of entries.

V. RESULTS

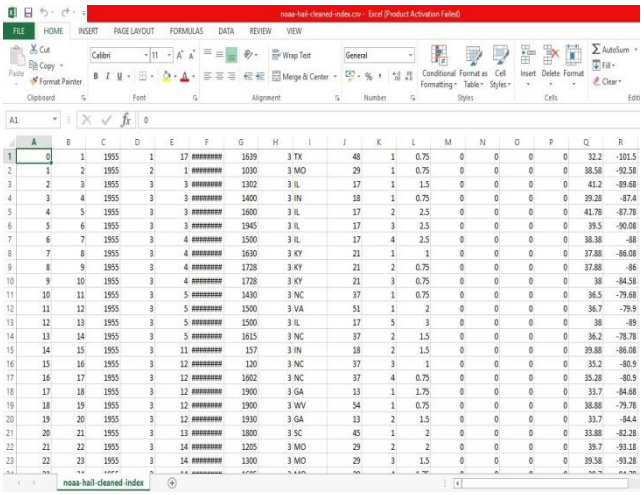

Fig 4:- Dataset

The snapshot of the dataset is shown in the Fig. 3. This dataset is preprocessed and applied for an algorithm. As the threshold increases, the cluster splitting will be decreased. When the threshold decreases the cluster combining is avoided. The variation of the threshold of tree birch and the levelled tree is shown in the Fig. 4 and Fig. 5.

When the supercluster splitting is avoided in a flat tree then the cluster count should be less than the tree-BIRCH as shown in Fig. 6. The flat tree runtime is less than the tree-BIRCH as shown in the Fig. 7.

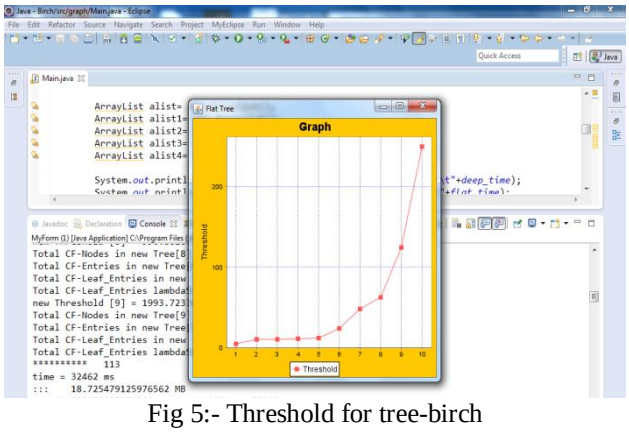

Fig 5:- Threshold for tree-birch

ISSN No:-2456-2165

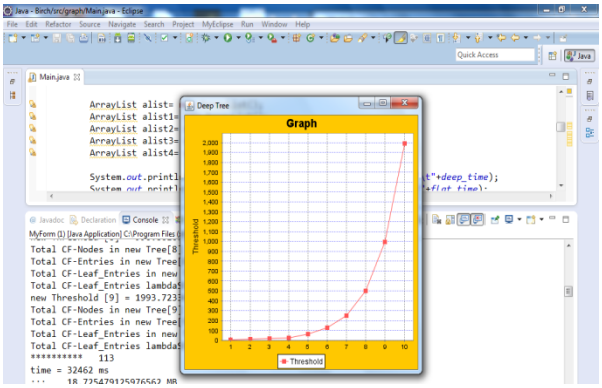

Fig 6:- Threshold for flat tree

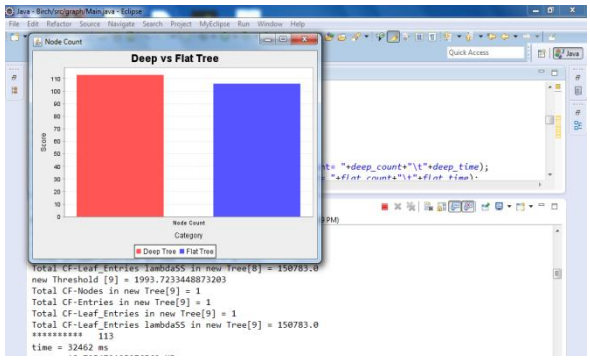

Fig 7:- Cluster count for the deep tree and leveled tree

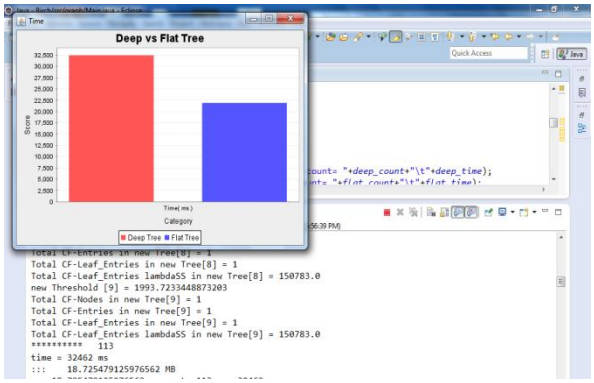

Fig 8:- Runtime

VI. CONCLUSION AND FUTURE WORK

Choosing the cluster count for any clustering algorithm is uneasy and the information about the dataset is required. The aim of this project is to cluster the data points without looking at the dataset. This is done by eliminating the global clustering phase from the BIRCH algorithm which is the third phase of the algorithm where the cluster count is required. To do this the levelled tree mechanism is introduced where in a tree there will be a single level. It is also called as flat tree. Here in a levelled tree, all the children in a tree will have a single parent node. This avoids the supercluster splitting and the runtime of the proposed tree is less than the deep tree.

The algorithm is worked for the two dimensional dataset. In future work, it is suggested to work on multi-dimension and extended dataset.

REFERENCES

[1]. T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: a new data clustering algorithm and its applications, Data Min. Knowl. Discov. 1(2) (1997) 141–182.
[2]. M. Zechner, M. Granitzer, Accelerating k-means on the graphics processor via cuda, in: First International Conference on Intensive Applications and Services, 2009, pp.7–15.
[3]. J.B. Macqueen, Some methods for classification and analysis of multivariate ob-servations, in: Proceedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability, vol. 1, University of California Press, 1967, pp.281–297.
[4]. Inderjit S. Dhillon and Dharmendra S. Modha. A dataclustering algorithm on distributed memory multiprocessors. In Large-Scale Parallel Data Mining, Lecture Notes in Artificial Intelligence, pp 245–260, 2000.
[5]. N. Ismael, M. Alzaalan, W. Ashour, Improved multi threshold birch clustering algorithm, Int. J. Artif. Intell. Appl. Smart Devices 2(1) (2014) 1–10.
[6]. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discover-ing clusters in large spatial databases with noise, in: E. Simoudis, J. Han, U.M. Fayyad (Eds.), Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp.226–231.
[7]. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "14.3.12 Hierarchical clustering". The Elements of Statistical Learning (2nd ed.). New York: Springer. pp. 520–528. ISBN 0-387-84857-6. Archived from the original (PDF) on 2009-11-10. Retrieved 2009-10-20.
[8]. Zhang, T.; Ramakrishnan, R.; Livny, M. (1996). "BIRCH: an efficient data clustering method for very large databases". Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96. pp. 103–114. doi:10.1145/233269.233324.
[9]. F.R. Bach, M.I. Jordan, Learning spectral clustering, in: Advances in Neural In-formation Processing Systems, 2004, pp.305–312.