ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

Clustering Dynamic Data Streams Using Enhanced Birch Technique

Shahini N[1], Arthinivasini R.B[2]

Asst. Professor, Dept. of Computer Science & Engineering, Panimalar Engineering College, Chennai, India[1]

Asst. Professor, Dept. of Computer Science & Engineering, Panimalar Engineering College, Chennai, India[2]

ABSTRACT:Conventional clustering algorithms which are been used in the Data mining concept, using k-means algorithm doesn't estimate well for very large datasets or data streams. The data are too huge such that the data doesn't fit into the main memory and have to be stored in a distributed manner. Hence conventional clustering algorithms typically requires repeated access to all the data through more than one passes which would be very expensive when large data are to be used. BIRCH is an approach, where it has the ability to incrementally as well as dynamically cluster the incoming and multi-dimensional metric data based on the traversal of complete binary tree, where the degree of leaf node is same. The subclusters which are the child of leaf nodes can have more than two nodes which are called as clusters in an attempt to produce the best quality clustering for a given set of resources such as memory and time constrains. By using this hierarchical clustering concept, the algorithm makes the entire usage of the available memory to derive the finest possible sub-clusters with minimized I/O costs.

KEYWORDS: Dynamically cluster, single pass, binary tree traversal

I. INTRODUCTION

Data Stream Mining is that the method of extracting data from continuous flow of information or streams. A knowledge stream is an ordered sequence of instances used in several applications of information stream mining where, there exist restricted computing and abuse storage capabilities. In several knowledge stream mining applications, the intention is to predict the category or cost of latest instances within the knowledge stream given some data regarding the category membership or values of previous instances within the knowledge stream. In various applications, the principles behind their labelling may have some modification over time, i.e. the expected value and the target achieved may be differed. This downside is cited as thought drift

A. Various Software for Data Stream Mining:

Rapid Miner is a free open-source software for knowledge discovery, machine learning, data stream mining and data mining, in learning time-varying concepts, and even in tracking drifting concepts MOA (Massive Online Analysis) is another free open-source software which is specific for mining data streams with concept drift of these data. It has various machine learning algorithms such as Classification, Clustering, Outlier detection and Regression. It also contains a prequential evaluation method. MOA supports bi-directional interaction with Weka, (Machine learning approach).

B. Streaming Algorithms

For finding out the most frequent elements in a data stream, some notable algorithms are used such as: Karp-Papadimitriou-Shenker algorithm, Count-Min sketch, Sticky sampling, Lossy counting, Sample and Hold, Sketch-guided sampling, Multi-stage Bloom filters.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

1) Applications

Streaming algorithms have many applications in networking like monitoring network links for elephant flows, estimating the flow size of distribution, counting the total number of distinct flows and so on. They even have applications in databases like estimating the size of a join.

C. Clustering

Cluster is a group of objects that belongs to similar classes. In other words the similar objects are grouped in a single cluster and dissimilar objects are grouped in another cluster. Clustering is the process of making group of abstract objects into classes of similar objects. A cluster of data objects can be treated as a one group. While doing the cluster analysis, the set of data should be partitioned into groups based on data similarity and then assign the label to the groups. The main advantage of Clustering over classification is that, clustering is adaptable to changes and identifies unique features that is differed from other groups.

Typical cluster models includes (1) Connectivity models: for example hierarchical clustering builds models based on distance connectivity. (2) Centroid models: for example the k-means algorithm represents each cluster by a single mean vector. (3) Distribution models: clusters are modelled using statistical distributions, like varying normal distribution used by the Expectation maximum algorithm. Density models: for example DBSCAN and OPTICS defines clusters as connected dense regions in the data space. Subspace models: in Biclustering which is also known as Co-clustering, the clusters are modelled with both the cluster members and relevant attributes. (4) Group models: some algorithms do not provide a refined model for their results and just provide the grouping information. (5) Graph-based models: a clique is a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. A "clustering" is actually a group of such clusters, sometimes containing all objects within the data set. In addition, it should specify the relationship between each of the clusters. For example, a hierarchy of clusters embedded in one another. Clustering may be roughly distinguished as hard clustering and soft clustering where each object belongs to a single cluster and cluster to a certain degree simultaneously.

There are also finer distinctions possible, for example: strict partitioning clustering: here each object belongs to exactly one cluster strict partitioning clustering with outliers: objects can also belong to no cluster, and are considered outliers overlapping clustering, while in a hard clustering, objects belong to more than one cluster. Hierarchical clustering: objects that belong to a child cluster also belong to the parent cluster.

II. RELATED WORK

A. Evaluation of Data Stream Mining

The performance of an algorithm that operates on data streams is measured by three root factors. The number of passes the algorithm must make over the stream, the available memory and the running time of the algorithm. These algorithms have many similarities with online algorithms since they both require decisions to be made before all the data are available and are not alike. Data stream algorithms have limited memory only but they may be able to defer action until a group of points arrive. If this is an approximation algorithm, then the accuracy of the result is another important issue. The accuracy is often stated as an (ϵ, δ)approximation meaning that the algorithm achieves an error of less than $\epsilon$with probability $1 - \delta$

B. Cluster Analysis

Cluster analysis or clusteris that the task of clustering a collection of objects in such the simplest way that objects within the same group (called a cluster) area unitadditional similar (in some sense or another) to everyapart from to those in differentteams (clusters). It's a main task of exploratory data processing, and a standard technique for applied mathknowledge analysis, employed inseveral fields, together with machine learning, pattern recognition, image analysis, info retrieval, and bioinformatics. Cluster analysis itself isn't one specific rule, howeverthe overall task to be resolved. It may be achieved bynumerous algorithms that take issueconsiderably in their notion of what constitutes a cluster and the way to with efficiencyrealize them. Standard notions of clusters embracegroup with tiny distances

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

among the cluster members, dense areas of the informationarea, intervals or explicitapplied math distributions. Clusterwillthus be developed as a multi-objective improvementdrawback. The suitableclusterrule and parameter settings rely upon the individual knowledge set and meant use of the results. Cluster analysis isn't an automatic task, however a repetitivemethodof informationdiscovery or interactive multi-objective improvement that involves trial and failure. It'llusually be necessary to switchknowledge preprocessing and model parameters till the result achieves the required properties.

C. Clustering Methods

1) Hierarchical Method

This methodology produce the hierarchical decomposition of the given set of information. The Hierarchical method can be classified on basis of how the hierarchical decomposition is formed as follows

a) Agglomerative Approach

This approach is additionally referred to as bottom-up approach. During this we tend to begin with every object forming a separate cluster. It carry on doing therefore till all of the groups keeps on merging the objects that are close to each other. It carry on doing so until all of the groups are merged into one or until the termination condition holds.

b) Divisive Approach

This approach is additionally referred to as top-down approach. During this we tend to begin with all of the objects within the same cluster. Within the continuous iteration, a cluster is separate into smaller clusters. It's down till every object in one cluster or the termination condition holds. Disadvantage of this methodology is rigid i.e. once merge or split is finished, it will never be undone

2) Approaches to Improve Quality of Hierarchical Clustering

Here are two approaches that are used to improve quality of stratified clustering: (1) Perform careful analysis of object linkages at every stratified partitioning. (2)Integrate stratified agglomeration by first employing astratifiedcollectiverule to cluster objects into microclusters, soacting macroclustering on the microclusters. The foremost advantage of this technique is quicktime interval. It's dependent solely on the quantity of cells in every dimension within thequantityarea.

3) Applications of Cluster Analysis

Clustering Analysis is looselyemployed inseveral applications likeresearch, pattern recognition, information analysis, and image process. Cluster mayfacilitate marketers discover distinct teams in their client basis. And that theywill characterize their clientteamssupportedgetting patterns. In field of biology it may bewont to derive plant and animal taxonomies, categorize genes with similar practicality and gain insight into structures inherent in populations. Cluster conjointly helps in identification of areas of comparable land use in Associate in nursing earth observation information. It conjointly helps within the identification of teamsof homesduring atown according house sort, price and geographic location. Cluster conjointly helps in classifying documents on the online for info discovery. Cluster is additionallyemployed in outlier detection applications like detection of MasterCard fraud. As an information mining operate Cluster Analysis function a tool to achieve insight into the distribution of informationto watch characteristics of every cluster.

4) Data Clustering Algorithms

Clustering algorithms may beclassifiedsupported their cluster model. There are presumably over one hundredprinted cluster algorithms however not all give models for his or her clusters and mayso not simply be classified. There's no objectively "correct" cluster algorithmic rule. The most applicable cluster algorithmic rule for a selecteddrawbacktypicallymust be chosen through an experiment, unless there's a mathematical reason to like one cluster model over another. It ought to be noted that an algorithmic rulethat's designed for one quite model has no discovera knowledge set that contains a radically totally differentquite model.

III. PROBLEM DEFINITION AND METHODOLOGY

A. Problem With Previous Methods

Conventional clustering approaches like k-means and k- medoids doesn't scale well for a really massive datasets. Previous cluster algorithms performed less effectively over terriblymassive databases whichthey are doing not match into main memory and ought to be hold onduring a distributed manner. As a result, there was plenty of overhead

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

maintaining high cluster quality whereas minimizing the price of addition IO (input/output) operations. Moreovermost of BIRCH's predecessors examine all information points, i.e it checks on all presently existing clusters equally

B. Solution for the Problems

Hence the effectively scalable BIRCH clustering technique for very large datasets is used, where efficiency including less storage space, increased speed is improved.

1). BIRCH Algorithm: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an unsupervised datamining rule to perform hierarchical clustering over significantly large data-sets. The main advantage of BIRCH algorithm is the ability to incrementally and dynamically cluster the incoming, multi-dimensional metric data points in an endeavorto provide the most effective quality clustering for a given set of resources (memory and time constraints). In most cases, BIRCH solely needs one scan of the information. Additionally, BIRCH even claims to be the "first clustering rule projected within the info space to handle 'noise'.

2). Advantages with BIRCH: It is natural in every cluster that, the cluster decision is made while not scanning all knowledge points and presently existing clusters. It exploits the observation that knowledge area isn't sometimes uniformly occupied and each information is not equally vital. It makes full use of obtainable memory to derive the best potential sub-clusters whereas minimizing I/O costs. In addition, it is a progressive methodology that doesn't need the entire knowledge set ahead.

C. BIRCH Clustering Algorithm

Given a collection of N dimensional information points, the cluster feature of the set is outlined because the triple *CF=(N,LS,SS),*where LS is a the linear addition and SS is the square sum of information points. Cluster options are organized in a CF tree, that could be a height balanced tree with two parameters such as branching factor B and threshold T. Every non-leaf node contains at the most B entries of the shape$[CF_i, child_i]$, where $child_i$ could be a pointer to its *i*th child node and CF_i is the clustering feature which represents the associated subcluster.

A leaf node contains at the most L entries each of the form $[CF_i]$. In addition, it has two pointers such as previous and next thatare used to chain all the leaf nodes along. The tree size depends on the parameter T. A node is neededto suitin a page of size P. B and L area unit determined by P. therefore P is varied for performance standardization. It isa really compact illustration of the dataset as a result ofevery entry in the leaf node is notone data but a subcluster.

In this algorithm the first step is to scan all data and to build an initial memory CF tree using the given amount of memory. In the second step, all the leaf entries in the initial CF tree is scanned to rebuild a smaller CF tree, where outliers are removed and the crowded subclusters are grouped into larger clusters. In the third step, an existing clustering algorithm is used to cluster all leaf entries, where the agglomerative hierarchical clustering algorithm is applied directly to the subclusters represented by their CF vectors.

In addition it also provides the flexibility of allowing the user to specify either the desired number of clusters or the desired diameter threshold for clusters. After the completion of the above step, a set of clusters is obtained that captures major distribution pattern in the data. But there may exist minor or localized inaccuracies that could be handled by this step 4, which is optional. The centroids of the clusters produced in the third step are used as seeds in the step 4 and the data points are redistributed to its closest seeds to obtain a new set of clusters. Step 4 provides with an option of discarding outliers which is a point that is too far from its closest seed can be treated as an outlier

IV. ARCHITECTURE FOR CLUSTERING

A. Overall Clustering Framework

The stream framework consists of two main components: 1. Data Stream Data (DSD) that manages or creates a data or information stream, and2. Data Stream Task (DST) that performs stream mining task on data. Figure 4.4 shows a high level read of the interaction of the parts. We have a tendency to begin by creating a Data Stream Data object and a Data Stream Task object. Then the Data Stream Task object starts receiving data from the DSDobject. At any time, we are able to get the present results from the DST object. DSTs will implement any kind of data stream mining task such as classification or clustering.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

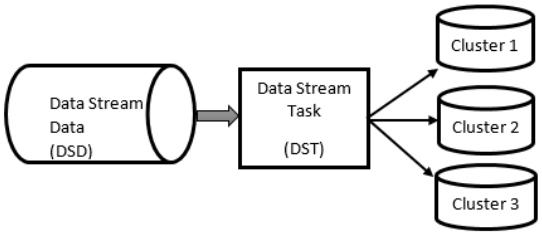

Figure 1: Overall framework of Data Stream Architecture

B. Clustering Architecture with respect to BIRCH Algorithm

The clustering architecture based on the BIRCH algorithm is implemented is shown in the figure

In Phase 1, the data is loaded into the memory. Then an initial in-memory CF-Tree is built with the data in one scan. The subsequent phases become fast, accurate, less order sensitive

In Phase 2, the data is condensed. The CF-Tree is rebuilt with a larger tree. But this condensing is optional

In Phase 3, we do Global clustering (Modified BIRCH). Now the clustering algorithm is used for the existing CF leafs

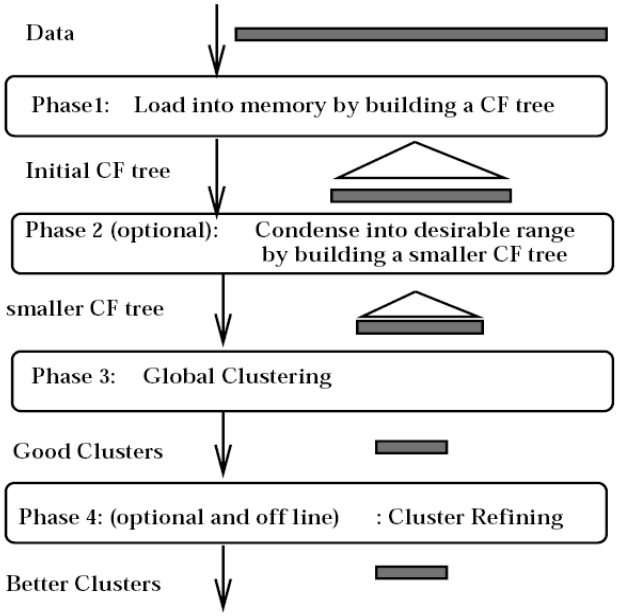

Figure 2: Birch Overview

V. IMPLEMENTATION OF DATA STREAM CLUSTERING

BIRCH is an approach, where it has the ability to incrementally as well as dynamically cluster the incoming and multi-dimensional metric data based on the traversal of complete binary tree, where the degree of leaf node is same. The subclusters which are the child of leaf nodes can have more than two nodes which are called as clusters in an attempt to produce the best quality clustering for a given set of resources such as memory and time constrains. By using this hierarchical clustering concept, the algorithm makes the entire usage of the available memory to derive the finest possible sub-clusters with minimized I/O costs. The clustering is done on the basis of top down approach, where it has the ability to incrementally as well as dynamically cluster the incoming and multi-dimensional metric data by the use of BIRCH algorithm. This Birch algorithm has been enhanced based on the complete binary tree strategy for traversal, where the child node can be in any level of degree, containing that all the leaf node are in the same order of degree by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

maintaining the concept of complete binary tree. The subclusters which are the child of leaf nodes can have more than two nodes which are called as clusters. This is an attempt which is going to produce the best quality clustering for a given set of resources such as memory and time constrains. By using this hierarchical clustering concept, the algorithm makes the entire usage of the available memory to derive the finest possible sub-clusters with minimized I/O costs.

A. Software Required for this Work

R Data Mining Tool(R Studio), RScript Coding Language, Windows 8 Operating System, Packages used are cluster, cluster.datasets, Rattle, Clusteval, NbClust, Plot3D, Stats and even some more.

1) Coding Language

An **R** script is simply a text file containing (almost) the same commands that you would enter on the command line of **R**. It refers to the fact that if sink()is used to send the output to a file, then some commands have to be enclosed in print() to get the same output as on the command line.

2) Advantage of using RScript

R provides a wide variety of graphical and statistical techniques, which includes linear and nonlinear model, time-series analysis, classical statistical tests, clustering, classification, and other mining concepts. R is easily extensible through functions and extensions. R community is highlighted for its contributions in terms of various packages. Many of the standard functions in R are written in R language itself, which seems simple for the users to follow the algorithmic choices made. Another strength of R is its static graphics, which can produce high-quality publication graphs which includes mathematical symbols. Dynamic graphs and interactive graphics are also available through additional packages.

VI. CONCLUSION

K-means, k-medoids are the conventional clustering algorithms which clusters the data which are static. Hence the data streams are clustered dynamically in a single pass within the limited memory space and efficiency is improved using BIRCH technique by the concept of complete binary tree. Given a limited amount of main memory, this enhanced BIRCH can minimize the time required for I/O.BIRCH is a scalable clustering algorithm with respect to the number of objects, and good quality of clustering of the data.

REFERENCES

[1]Matthew Bola, John Forrest, and Michael Hahsler, "Clustering large dataset using datastream clustering techniques", Springer Journal, pp 135-143, 2014
[2]Amineh Amini, Teh Ying Wah, Hadi Saboohi, "On Density-Based Data Streams Clustering Algorithms", Springer Journal, Vol 29, Issue 1, pp 116-141, 2014
[3]Shi Yu, Léon-Charle Tranchevent, Bart De Moor, Yves Moreau, "Optimized Data Fusion for *k*-means Laplacian Clustering", Springer Journal, Vol 345,2011
[4]Wenjun Xiao, Wenhong Wei, Weidong Chen, Yong Qin, "Extended clustering coefficients: Generalization of clustering coefficients in small-world networks", Springer Journal, Vol 16, Issue 3, pp 370-382, 2007
[5]Xiaohui Huang, Shenzhen, Yunming Ye; Haijun Zhang, "A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation", IEEE Transactions Vol 25, Issue 8, pp 1433 - 1446, 2013
[6]Jie Lian, Naik, K Agnew, "A Framework for Evaluating the Performance of Cluster Algorithms for Hierarchical Networks", ACM Transactions Vol 15, pp 1478-1489 Issue 6, 2007
[7]Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Sebti Foufou, and Abdelaziz Bouras, IEEE Transactions, "A Survey of Clustering Algorithms for Big Data", Vol 2, Issue 3, pp 267-279, 2014
[8]Rajasekaran. S, "Efficient parallel hierarchical clustering algorithms Parallel and Distributed Systems", IEEE Transactions, Vol 16, Issue 6, pp 497-502, 2005
[9]Das. S, Abraham. A, Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions, Vol 178, pp 137-174, 2009
[10]Maulik, U.Bandyopadhyay, S. "Performance evaluation of some clustering algorithms and validity indices", IEEE Transactions, Vol 24, Issue 12, pp 1650 – 1654, 2003
[11]Cattinelli, I, Valentini. G, Paulesu. E, Borghese, "A Novel Approach to the Problem of Non-uniqueness of the Solution in Hierarchical Clustering", IEEE Transactions, Vol 24, Issue 7, pp 1166 – 1173, 2013

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 9, September 2019

[12]Praditwong, K. Harman M, Xin Yao, "Software Module Clustering as a Multi-Objective Search Problem", IEEE Transactions, Vol 37, Issue 2, pp: 264-282, 2011
[13]ZHANG Jian-Pen, CHEN Fu-Cai, LI Shao-Mei, LIU Li-Xiong, "Data Stream Clustering Algorithm Based on Density and Affinity Propagation Techniques", IEEE Transactions, Vol. 40, Issue 2, pp 277-288, 2014
[14]Weizhong Zhao, Huifang Ma, Qing He, "Parallel *K*-Means Clustering Based on MapReduce", IEEE Transactions, Vol 5931, pp 674-679, 2009
[15]Xiangliang Zhang, Cyril Furtlehner, Cecile Germain-Renaud, Michele Sebag, "Data Stream Clustering with Affinity Propagation", IEEE Transactions, Vol 26, Issue 7, pp 1, 2013
[16]Evangelos E. Papalexakis, *Student Member, IEEE*, Nicholas D., and Rasmus Bro From K-Means to Higher-Way Co-Clustering", IEEE Transactions, Vol 61, Issue 2, pp 493 – 506, 2012