Preetishree Patnaik el at. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 1, March 2017, pp. 61-65

An Implementation of Grid Density Based Clustering Algorithm in Big Data

Preetishree Patnaik[1], Dr. Vivek Jaglan[2]

[1]Research Scholar, Computer science, Amity University, Gurgaon, Haryana, India
[2]Associate professor, Computer Science, Amity University, Gurgaon, Haryana India

***Abstract*-Big Data and Big Data mining have evolved as powerful tool to analyze massive data sets for modern applications. With such evolution of information and technology today the usability of data has grownup exponential rate and therefore it is need to process out the big data in order to collect all such relevant information. The clustering algorithms act as a Meta-Learning tool with unsupervised learning method that will help to check the accurate analysis of massive or large data sets (Big Data), exploring and studying hidden patterns which are generated by modern applications. In this paper we have proposed a novel Grid Based clustering algorithm named GCAB, in a most efficient and effective manner to handle massive data sets and study different cluster patterns generated.**

Keywords: **Big Data, clustering, clusters criteria, clustering algorithms.**

I. INTRODUCTION

The term Big Data encompasses of all forms of data, including Web logs, data from social networking sites, sensor data, tweets, blogs, user reviews, and SMS messages. Big data and big data analytics are in the recent study of information technology and business intelligence. These data are generated from various social networking sites like Facebooks, twitter, etc ,online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, science data, sensors Smart phones and their applications. These data are in different format, hence required for databases to store and analyze the data sets and visualize using different software tools. In comparison to past decades the primary IT Industry has changes a lot, with more fast transaction people are accessing huge amount of data in various pattern e.g. Internet mails, video, images ,audio messages ,sensors data streams and etc with such huge accessibility of data makes a revolutionary change in analysis of data streams patterns .[1]

Thus the Data Scientists has announced that we are now in the "Era of Big data" or we are sinking to deep water of big data every day.

Today, we present in the era of digitalization with gigantic progress, development of technologies, web media, social networking sites, online world technologies through internet, Smartphone.etc where every user are accessing enormous, massive quantities of data from various data sources. Such enormous data sets having massive, diverse and complex structure of data is term as "Big Data". These massive data creates a lot of difficulties in storing, analyzing, searching and visualization process. But we know that this massive volume of data sets can be useful to user in various aspects and creates lots of confusion in its storing and analyzing. Therefore a big massive of datasets (BIG DATA) are need to be store in effective and efficient manner that helps in various type of operations (i.e. analytical operation, process operations, retrieval, reliability of data & etc)

Thus it is most important to execution of these massive data sets into secrete –correlation pattern and cluster models that makes easy of its utilization through implementation various types of clustering techniques, Data mining methods.

The utilities of massive, large volume of Big Data has grown up exponential due to which several issues and challenges are been observed by Data scientist. In order to store massive or huge volume of data it required huge volume of big data and creates a lot of issues related to data management and transaction. One of the efficient ways overcomes the issues and challenges by implementing Big Data Mining techniques such as cluster analysis. The following criteria need to be adopts as to overcome the above mentioned difficulties such as:

- Design an appropriate system which can handle Big Data efficiently i.e. Volume, velocity, veracity, variety.
- Analysis it to extract all relevant information that helps to study hidden data patterns and supports for decision support system.
- Identify the key issue associated with data storage, management and processing

II. CATEGORIZATION OF BIG DATA

In the big data is characterized by variety; velocity and volume are as follows:

- Variety: Big data come from a great range of sources and a further volume of data source is categorized into three types as Structured, Semi-structured and Unstructured.[1]

i) Structured data inserts a data warehouse already tagged. The structures data are organized manner easily sorted to store in database .these variety Data include the abstract data type, web links, pointers etc.

ii) Whereas the unstructured data are random and difficult to analyze. These are Heterogeneous and raw/incomplete data that are generated from multiple users in different sources. (e.g.: Bitmap images, objects, text, etc).

iii)Semi-structured data these are the combination of structure and un-structured data and doesn't conforms to a fixed set of tags or others semantics structure of data.

- Volume: Volume or the size of data has been larger than terabytes and petabytes. The grand scale and rise of data outstrips fixed store and analysis technique. As the Big data size is massive and huge in nature, so it's a biggest challenge for the data scientist to design the large database for its effective storage and visualization. [1]
- Velocity: The range of data used is in max range, Velocity is a necessary parameter not only for big data, but also all processes. For time limited processes to be executed, big data used should be in organization streams to have a maximize value. [1]

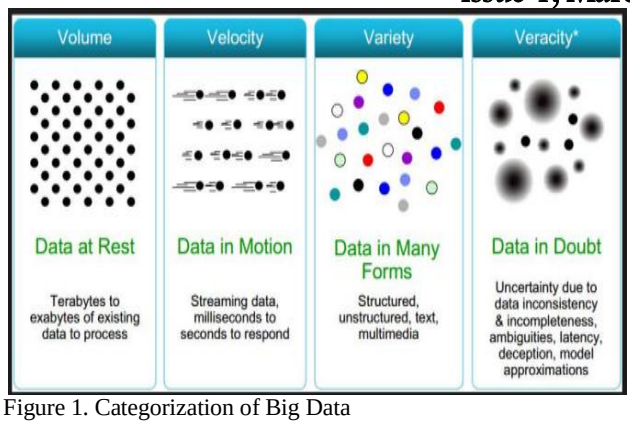

Figure 1. Categorization of Big Data

III. TAXONOMY OF CLUSTERING METHODS

The term "clustering" or cluster analysis was first coined by "Driver and Kroebar" which is famous for unsupervised learning method of Data Mining. However different scientist developed different types of clustering algorithms that varies in their properties, clustering models and etc.

In general clustering can be defined as is a process of grouping a set object into a class of similar objects. Or "Clustering is a process of division of DATA into a group of similar objects". [13]

The shape and size of cluster creation and visualization varies from one another with their respective properties of the algorithm. Despite from huge number of survey for clustering algorithms available for various domains i.e. machine learning, information retrieval, pattern recognition, bio-informatics, semantic medical sciences it makes difficult to user to decide which algorithm is appropriate to analysis the massive data sets. Therefore we have implements the taxonomy of clustering algorithms and propose these classifications to develop a frame work that covers major factors in selecting suitable algorithms for massive data sets.

The clustering Algorithms are broadly classified into four categories which are as follows.

A. Partitioned Based Clustering

This is a traditional based clustering method and the partitioning algorithms divide data objects into a number of partitions, where each partition represents a cluster. The main objective as each of the clusters must have atleast one object and object belongs to the cluster must be similarity class among themselves e.g. K-means, K-modes, CLARA, PAM, FCM. [13]

B. Hierarchical Based Clustering

This method of clustering is done by using a dendrogram (Tree representation of datasets) were the data are organized in hierarchical manner and it also includes the medium of proximity of datasets. The proximity of cluster is calculated by intermediate nodes. Hierarchical Clustering methods are of two types as Agglomerative (bottom up) and Divisive (Top down). An Agglomerative clustering method starts with one object for each cluster and recursively merges two or more cluster followed by bottom up approach where as a Divisive Clustering method start with the datasets as one cluster and recursively splits the most appropriate cluster e.g. BIRCH.[13]

C. Density based Clustering

In density-based clustering, clusters are defined areas of higher density than the remainder of the data set. Objects in these sparse areas that are required to separate clusters are usually considered to be noise and border points. Here, data objects are separated based on their regions of density, connectivity and boundary. They are closely related to point-nearest neighbors. A cluster defined as a connected dense component grows in any direction that density leads to. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes. Also, this provides a natural protection against outliers. Thus the overall density of a point is analyzed to determine the functions of datasets that influences a particular data point. DBSCAN, OPTICS, DBCLASD and DENCLUE are algorithms that use such a method to filter out noise and discover clusters of arbitrary shape. [13]

D. Grid Based Clustering

The space of the data objects is divided into grids (cells). The main advantage of this method is its fast processing time, because it scans the DB once to calculate the statistical values to generate a grid structure. The grid-based clustering techniques independent of the number of data objects that occupy a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. 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. However, for highly irregular data distributions, using a single uniform grid may not be sufficient to obtain the required clustering quality of the time requirement. Wave-Cluster and STING are typical examples of this category. [12] [13]

IV. GRID DESNITY BASED CLUSTERING IN DETAILS

In this section of paper we have given the actual idea behind the implementation of the grid based clustering techniques. [12]

E. Basic Structure for Grid Outline

In the various types of clustering Algorithms explained in the [Section 2] have many types of drawbacks with handling with massive data sets. In this section we have proposed a grid density based clustering algorithm in big data which resolves the drawbacks of other conventional algorithms and can be applicable for massive as well as to multidimensional data sets. [17]

The conventional clustering algorithm techniques calculate the distance among the cluster centre using dissimilarity Metrics (E.g. Euclidean distance) between the different patterns to estimate the nearest cluster centre with index value. The conceptual idea of the proposed algorithm was given by [Warnekar1979] to organize the data space having a multidimensional data structure which is represented in form grid structure. [Erich Schikuta 1996] The pattern generated with grid structure are treated as different points in d- dimensional structures were this pattern are stored according to random fashion for a topological distribution .Finally the grid structure partitioned the data space into rectangular shape blocks called as grid blocks.[17]

F. Creation of Grid Blocks

A grid Block is constructed as d-dimensional hype-rectangle (rectangle shaped cube) containing maximum of $\mathbf{B_s}$ pattern called as block size. [17]

Let X = {x_1, x_2….x_n} be the set of n-patterns. $\mathbf{X_i}$ = i^{th} pattern consisting of a tuple with describing features as (a_{i1},a_{i2},a_{i3}……..a_{id}) i.e. where d= number of dimensions.
The following mathematical properties satisfied for the grid structure:
Let say:
Ø= empty set, for all X_i, $\mathbf{X_i \in B_j}$ ..... eq(1)
$\mathbf{B_j \cap B_k} = Ø$; if $j \neq K$ …….. eq(2)
$\mathbf{B_j} \neq Ø$ and
$\cup \mathbf{B_j} = X$
The above eq(1)(2) depicts that the GCAB algorithm makes the blocks(B_i) and pattern(X) into a nested sequence of non-empty and disjoint clustering where
[C_{lu1},C_{lu2},C_{lu3},…….C_{luwn}]

Where Clu= no of cluster generated and w_n = number of cluster generated by i^{th} process.
At the initial stage, 0^{th} clustering each block of cluster represented as

$C_{luj}=\mathbf{B_j}$ Such that j =1, 2……b and $\mathbf{W_0} = b$.

G. Calculation of Density Index

4.1. Calculation of Density Indices

In the grid density based algorithm (GrdClu-Bd) the density index can be calculated for each block by using following clause:

- The number of patterns represents as point in a block
- Calculate the spatial volume of each block i.e. **(V_b).** The spatial volume defined as the block(B) with Cartesian product of extent (e) in each dimension of the block (B), i= 1,2,……..d;-
  i.e.

$$\mathbf{V_b} = \pi\, e_{bi}$$

Now the density index D_b of block (B) defined as the ratio of the total number of actual pattern (P_b) points contained in block (B whose spatial volume is (V_b).
∴ $\mathbf{D_b} = \mathbf{P_b}/\mathbf{V_b}$……eq(3)

After the density index is calculated then the blocks are sorted according to their density index value .The block with highest density index value is sorted first and followed by the pattern correlation became cluster centre and then the remaining cluster iteratively developed into a new cluster centre with the density index value . [17]

V. PROPOSED GCAB ALGORITHM

ALGORITHM: Grid Based Clustering Algorithm in Big Data (GCAB)

Input:
[U] = The number of cluster iteration is stored in u[].
W[u] = after successful iteration, the number of cluster formed stored in w[u].
C [u,v] = It is a set value variable containing the clustered blocks of iteration u[] and cluster c[v].
Output: Grid Cluster patterns are generated in 2D and 3D view.

[Initialization]: set u[] = 0 w[] = 0, c[][] = 0;
[Grid Structure and Density Block indices]: construct the grid structure and then calculate the Density Block indices (Db) using the formula in eq(3)
$D_b=\frac{P_b}{V_b}$ where P_b = Number of data points in the Block B and V_b= spatial volume of the block i.e. $V_b = \pi\, e_{bi}$.

[Sorting]:
Sorting process is executed according to the density block indices value such that a sorted block sequence B1',B2',.....Bb' and at Initial stage mark all generated Block as "Not Active" and "Not Clustered ".

```
while {
        A "Not Active" block exists }

        do;

u = u+1;
Mark Density block indices as " active" for the first sequence of block
        Generated as B1', B2',.....Bj' ;

for each "Not clustered " Block
    {Bi' = Bi' ; Bi',Bi',.....Bj' ;}
```

[Cluster creation]:

```
{
    Set c[u];
  W[u] <= w[u]+1;
 C [u,w[u]] <- B1' ;
}
mark block B1' = clustered ;
```

VI. EXPERIMENTAL ANALYSIS

The above proposed GCAB algorithm has implemented in MatLab.7.01 version. The main advantages of using MatLab (Matrix Laboratory) as it execute integrates numerical computation, visualization and data analysis. Beside these factors MatLab also support high –level language programming environment for solving mathematical analysis effectively and efficiently. The algorithm analysis has done with applying on number of data sets and the different pattern of cluster is found.

Clustering analysis with different Cluster Pattern is shown in below figures and we have concluded that the Elapse time is increased with increase in the volume or size of data sets increases. Elapse time to run the real time datasets required 47.704368 seconds.

In this paper we have implemented the (GCAB) Grid density based clustering Algorithm using Big Data which work efficiently and effectively over multidimensional and large massive data sets .The advantages of GCAB algorithm as it easily estimate the density and helps in cluster analysis with different patterns of cluster in low processing time. The

Experimental analysis of the GCAB algorithm has implemented with different data sets and processing time of the algorithm depends upon the size of data sets. The above studied different pattern of cluster representation helps to solving the main issue in Big Data Analytics i.e Visualization, Sorting, Searching and Analysis –secrete correlation patterns in Big Data Mining.

As per the future work we suggest and investigate different types of grid based clustering algorithm and verify the efficiency of algorithm handling massive or large data sets.

The GCAB 3-dimensional algorithm Elapsed time is 1.272737 seconds and number of Clusters found is 249 and for 2-dmensional GCAB Elapsed time is 11.779441 seconds.

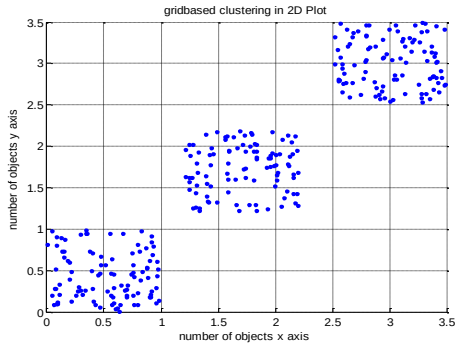

Figure 2. 2D reprsenetaion of GCAB Algorithm using Matlab version 7.01 with numerical datasets.

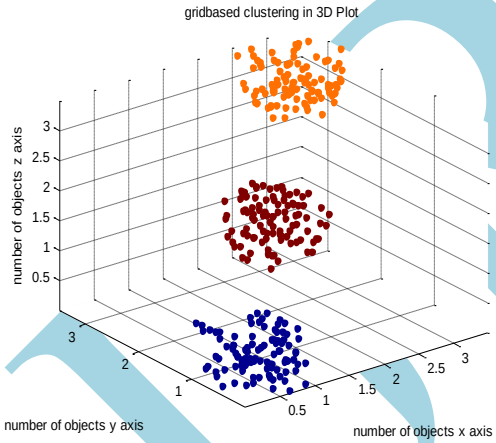

Figure 3. 3D Representation of GCAB Algorithm with numerical datasets.

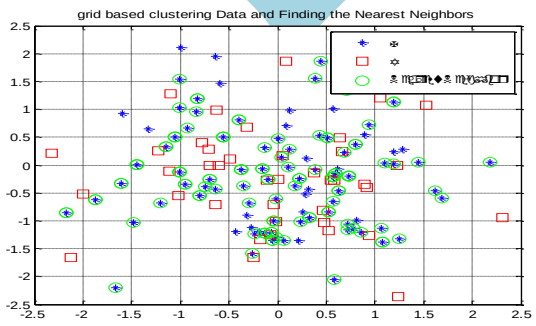

Figure 4. Nearest Neighbour Search in GCAB Algorithm

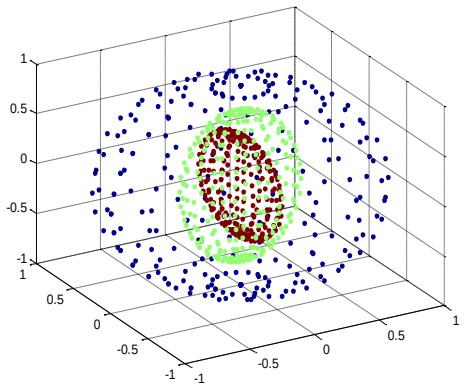

Figure 5. GCAB with Multidimensional represenation of Clusters.

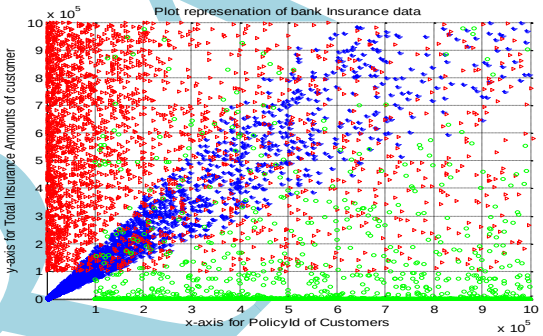

Figure 6. GCAB Alogrithm with real-time data represented with 3000 cluster points

VII. CONCLUSION

In this paper we have implemented the (GCAB) Grid density based clustering Algorithm using Big Data which work efficiently and effectively over multidimensional and large massive data sets .The advantages of GCAB algorithm are that it easily estimates the density and helps in cluster analysis with different patterns of cluster in low processing time. The Experimental analysis of the GCAB algorithm has been implemented with real time data sets and processing time of the algorithm depends upon the size of data sets. The cluster patterns generated help in resolving the main issue in Big Data Analytics i.e. Visualization, Sorting, Searching and Analysis –secrete correlation patterns in Big Data Mining.

In future, we will analyze the different types of grid based clustering algorithms and verify their efficiency for handling massive data sets.

VIII. ACKNOWELDGEMENT

I am thankful to my co-author and Computer Science Engineering Department, Amity University Haryana (Gurgaon) for support in fulfilling the data mining laboratory for implementation of the GCAB algorithm and evaluation of results.

IX. REFERENCES

[1] Sere S. and Duyug S. "Big Data: A Review."Ieee Explore, 2013.

[2] Rakesh A., Johannes G., Dimitrios G., Prabhakar R, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", IBM Almaden Research

Center 650 Harry Road, San Jose, CA 95120,ACM Digital Library .

[3] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan," Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications". In International Conference Management of Data (SIGMOD'98), 1998

[4] Hanh Le ,Mbogho A. Takizawa, "A Survey on Clustering Algorithms for Wireless Sensor Networks", Network-Based Information Systems (NBiS), 2010 13th International Conference on 14-16 Sept. 2010.

[5] Charu C. Aggarwal and ChengXiang Zhai , "A survey of Text Clustering Algorithm", SpringerLink , Mining Text Data 2012, pp 77-128.

[6] Ruhana and Mahamud , "big data clustering using grid computing and ant based algorithm" . Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 2013.28-30 August, 2013 Sarawak, Malaysia. University Utara Malaysia .

[7] Hrishav B., Barua,Saurav j. Sarmah, " An Extended Density based clustering Algorithm for large spatial 3D data using Polyhedron Approach".International Journal of Computer Applications (0975 – 8887) Volume 58– No.2, November 2012.

[8] Amandeep K. and Navneet K. "Grid Density Clustering Algorithm ",International Journal of Innovative Research in Science, Engineering and Technology , Vol. 2, Issue 10, October 2013.

[9] Hrishav B., Barua,Saurav j. Sarmah, " An Extended Density based clustering Algorithm for large spatial 3D data using Polyhedron Approach". International Journal of Computer Applications (0975 – 8887) Volume 58– No.2, November 2012.

[10] Oded M., Lior R., "data mining and knowledge discovery handbook", Springer Science+Business Media.Inc, pp.321-352, 2005.

[11] GuiBin H. and RuiXia Y., "Irregular Grid-based Clustering Over High-Dimensional Data Streams", IEEE 1st International Conference on Pervasive Computing, Signal Processing and Applications, pp.783-786, 2010.

[12] ilango and mohan, "A Survey of Grid Based Clustering Algorithms", International Journal of Engineering Science and Technology, pp 3441-3446. 2010.

[13] Fahad, Alshatri, Tari, et.all Member, "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis ". IEEE Transactions on Emerging Topics in Computing transactions, 2014.

[14] Zheng H., Wang Z., et.all, "Clustering Algorithm Based on Characteristics of Density Distribution", IEEE 2nd International Conference on Advanced Computer Control, vol.2, pp.431-435, 2010.

[15]. Amineh A., Teh Y., Wah, et.all, "A Study of Density-Grid based Clustering Algorithms on Data Streams",IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery, vol.3, 2011.

[16] Guohua L., Xiang Y., et.all, "An Incremental Clustering Algorithm Based on Grid",IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.

[17]Erich Schikuta, "Grid-Clustering: An Efficient Hierarchical Clustering Method for Very Large Data Sets", IEEE Proceedings of ICPR '96, 1015-4651/1996.

