International Journal For Technological Research In Engineering
Volume 3, Issue 10, June-2016 *ISSN (Online): 2347 - 4718*

AN IMPLEMENTATION OF GRID DENSITY BASED CLUSTERING ALGORITHM IN BIG DATA

Neeraj[1], Mrs. Chhavi Rana[2]
[1]M.Tech (CSE), [2]Coordinator, Dept. of CSE)
University Institute of Engineering and Technology MDU Rohtak

ABSTRACT: The era of "Big Data" is upon us. From big consumer stores mining shopper data to Google using online search to predict incidence of the flu, companies and organizations are using troves of information to spot trends, combat crime, and prevent disease. Online and offline actions are being tracked, aggregated, and analyzed at dizzying rates. For example, questions like, how many calories we consumed for breakfast, how many we burned on our last run, and how long we spend using various applications on our computer, can be recorded and analyzed. We can lose weight by realizing we tend to splurge on Thursdays. We can be more efficient at work by realizing we spend time more than we thought on Facebook. Data warehousing and data mining are related terms, as is NoSQL. With data firmly in hand and with the ability given by Big Data Technologies to effectively store and analyze this data, we can find answers to these questions and work to optimize every aspect of our behavior. Amazon can know every book you ever bought or viewed by analyzing big data gathered over the years. The NSA (National Security Agency) can know every phone number you ever dialed. Facebook can and will analyze big data and tell you the birthdays of people that you did not know you knew. With the advent of many digital modalities all this data has grown to BIG data and is still on the rise. Ultimately Big Data technologies can exist to improve decision-making and to provide greater insights...faster when needed but with the downside of loss of data privacy.
Keyword: Big Data, NSA, Hadoop, Clustering, MATLAB.

I. INTRODUCTION

1.1 Motivation:

The term "Big Data" defined as enormous data sets having a large more assorted and complex configuration of representation that creates difficulty in storing, analyzing searching and visualization process. This process of execution of the massive data sets into a secrete correlation mold called as "Big Data Mining" which implies the same concept of discovery the hidden and relevant data through various principles of Data Mining. Clustering algorithms have emerged as an alternative powerful and meta-learning tool helps to analyze the massive volume of data (Big Data) generated by many applications. In general, we studied that Big Data creates a lot of confusion while categorization of Big Data. Therefore the most relevant clustering algorithm must be used to classify the Big Data. Big Data has become one of the buzzwords in IT during the last couple of years. Initially it was shaped by organizations which had to handle fast growth rates of data like web data, data resulting from scientific or business simulations or other data sources. Some of those companies' business models are fundamentally based on indexing and using this large amount of data. The pressure to handle the growing data amount on the web e.g. leads Google to develop the Google File System and MapReduce. Efforts were made to rebuild those technologies as open source software. This resulted in Apache Hadoop and the Hadoop File System and laid the foundation for technologies summarized today as 'Big Data'. 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 Facebook, twitter, etc ,online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, science data, sensors Smart phones and their applications [1]. These data are in different format, hence required for databases to store and analyze the data sets and visualize via typical database 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 . 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.

1.2 Goals:

The First goal of this project is to study, analysis of Grid Based clustering techniques implementation to Big Data as optimization time efficiency. This result to verify and analysis different types of cluster pattern formed with very large and multi dimensional datasets. The second goal is summarize with associated Grid Based Clustering algorithms for the 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 data sets(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 a secrete correlation and pattern Analysis of the

cluster models that makes easy of its utilization through implementation various types of clustering techniques and Data mining methods.

1.3 Problem Statement:

This work deeply focus on the clustering Algorithms and clustering techniques. Clustering is the most significant unsupervised-learning problem. The usability of cluster analysis has been used broadly in data recovery, pattern recognition, text and web mining, software reverse engineering and image segmentation. The problem deal with implementing of the algorithms to handle the Big Data analytics and using the different data sets to handle the Big Data with multiple resolutions and checks the efficiency of the algorithm and its CPU utilization time. The major function of clustering is to finding a construction of similar data items and the clustering involves partitioning a given dataset into a number of groups of data whose members are similar in some way. The first category focuses on adding 'Big Data' functionality to operational applications to handle huge amounts of very fast incoming transactions. This can be as diverse as applications exist and it is very difficult, if not infeasible, to provide a clustering techniques with the implementation of various data mining. Therefore I will focus on the second category and 'analytical Big Data processing'. This will include general functions of analytical applications, e.g. typical data processing steps, and infrastructure software that is used within the application like databases and frameworks as mentioned above.

II. DIGGING INTO BIG DATA

The term Big Data has recently appearing to define the traditional Data into a large amount of data which required an advance Data management system for storage in large Data Warehouse, Data processing, Data Analytics, and Visualization.

2.1. Evolution of Big Data:

In the late 80's and 90's the Information System (IS) started growing its enterprises and organization across many Industrial region which leads to rapid growth of IT industries. The Information Technology (IT) industries exponentially producing more and more amount of data which further creates a lot of issue in searching, analyzing . The Data Analysis process is executed inorder to find out different patterns of data format, trends, dependency, hidden patterns and etc. In the past decade the data remained untouched with the evolution of IT system which reached to a maturity level and strongly affects the data processing and Data Analysis method. The availability of existing memory (i.e. direct access, random access memory), computation power have been rapidly increasing. The data accessibility has modified into a complex data format, mostly the problem exist with the infrastructure set by IT Industries for Collection, searching, storing, Analyzing, managing these complex data is a big problem. The above observation lead to introduce the concept of Data warehouse system should be used for storing, analyzing, and visualization of massive data. A Date warehouse is a repository of Information collected from multiple sources, sorted under a unified schema, and usually residing a single database. The main role of data warehouse as for executing the data cleaning, data integration, data transformation, data loading, and continuing with periodic data refreshing.[ref data mining book] However it was observed that the data has changed the format, syntax, semantics and it leads to big confusing stage for the user in data accessibility, numerous problems can into existence that what will happen it these complex structure data are represented in a data warehouse, how the upcoming business analytics designer will be plan for company auditing that the recent available space for space will be worth full for future data or to handle the storage of massive data sources. In recent survey has stated that Big Data and its analysis are at the center of modern science and business. These data are emerged from online transactions, emails, videos, audios, images, click streams, logs, posts, search queries, health records, social networking interactions, science data, sensors and mobile phones and their applications [1, 2]. They are stored in databases grow massively and become difficult to capture, form, store, manage, share, analyze and visualize via typical database software tools. 5 Exabyte (1018 bytes) of data were created by human until 2003. Today this amount of information is created in two days. In 2012, digital world of data was expanded to 2.72 Zettabytes (1021 bytes). It is predicted to double every two years, reaching about 8 Zettabytes of data by 2015 [1]. IBM indicates that every day 2.5 Exabyte of data created also 90% of the data produced in last two years. A personal computer holds about 500 gigabytes (109 bytes), so it would require about 20 billion PCs to store all of the world's data. In the past, human genome decryption process takes approximately 10 years, now not more than a week. Multimedia data have big weight on internet backbone traffic and is expected to increase 70% by 2013. Only Google has got more than one million servers around the worlds. There have been 6 billion mobile subscriptions in the world and every day 10 billion text messages are sent. By the year 2020, 50 billion devices will be connected to networks and the internet [3]. In 2012, The Human Face of Big Data accomplished as a global project, which is centering in real time collect, visualize and analyze large amounts of data. According to this media project many statistics are derived. Facebook has 955 million monthly active accounts using 70 languages, 140 billion photos uploaded, 125 billion friend connections, every day 30 billion pieces of content and 2.7 billion likes and comments have been posted. Every minute, 48 hours of video are uploaded and every day, 4 billion views performed on YouTube. Google Support many services as both monitories 7.2 billion pages per day and processes 20 petabytes (1015 bytes) of data daily also translates into 66 languages. 1 billion Tweets every 72 hours from more than 140 million active users on Twitter. 571 new websites are created every minute of the day [23]. Within the next decade, number of information will increase by 50 times however number of information technology specialists who keep up with all that data will increase by 1.5 times. Now days Big Data Analytics has include the performance management tools, applications,

Business Intelligence tools (BI), various types of data warehouse management platform are used in diversified areas such as:
Finance: For Budgeting, Planning, Strategy.
Supply chain Management: logistics, Inventor
CRM-Customer Relation Management: sales, customer services, Marketing Analysis, Web Mining, and product and price optimization. Services operations: Banking, Education, Government, Healthcare, etc. Reporting and Analysis: Tools such as OLAP,OLTP, advance technologies that are used for data categorization, data discrimination, frequent pattern analysis, classification rules, cluster analysis and visualization of outliers(noise elimination). The Below figure depicts the overview of Big Data according to its dimensions (volume, velocity, variety):

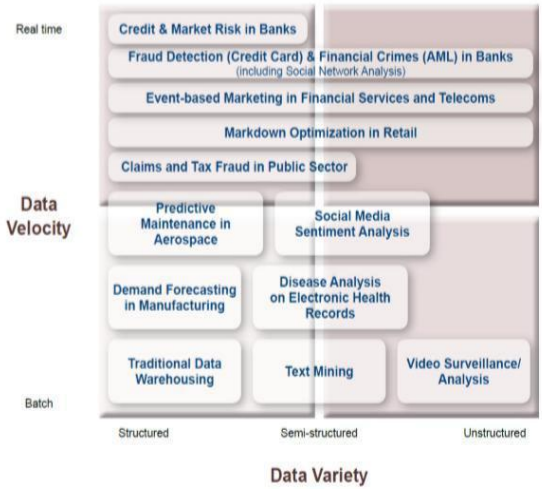

Figure 1: Use case diagram for Big Data with different dimensions

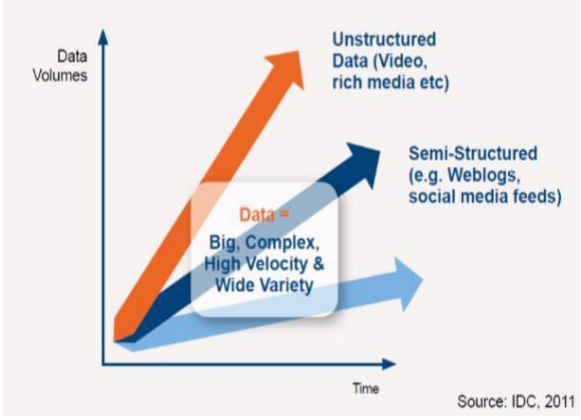

Figure 2: Big Data by IDC

III. LITERATURE REVIEW

In this paper the author had given a review of Big Data and its categorization. The term Big Data defined as massive data sets having a complex structure of representation and creates difficulties in storing, sorting, and analyzing and Data analysis. Such process of studying and evaluating Big Data into a secret correlation pattern is called as "Big Data Analytics". The existence of Big Data starts up with the change in the information system, latest technologies, recent trends in data accessing mode by users, social networking sites etc. Big Data is an informal term that encompasses 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 Facebook, 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 via typical database software tools. Till 2012 it was recorded that for every two days 2.72zettabytes of data sets are used and produce in various networks for user which is further predicted that it will be double in 2020. There have been 6 billion mobile user subscriptions in the world and every day 10 billion text messages are sent in every day. By the Year 2020, 50 billion devices will be connected to networks and the internet. A study by IDC found that, by 2020, the world will generate 50 times the amount of data that is now used and 75 times the number of information will be needed to store that data. Most of this information will be carried over mobile-broadband networks, and Ericsson forecasts that by 2017, Smartphone data traffic will increase to 1.1GB a month from its current 350MB.

IV. DESIGN AND IMPLEMENTATION

4.1 Proposed framework in Big Data using the concept of data mining and pattern evolution: In compression to past decades the IT industry has reckon and the data volume accessibility exceeds the capacity of current online storage and other processing system. The users are now allowed to access massive amount of data and the usage has extended from a range of Exabyte per year to Zettabytes per year. It has been observed that storage and data transport are such technology issue which is need to be focused inorder to optimize the existing problem. Today data scientist has announced that we all are living in the "Era of Digitization" and every fraction of seconds we are sinking to a deep sea of Big Data. Such exponential growth in Big Data cause various problem in storing, sorting, searching, Data Analysis and different hidden correlation data patterns makes the job data analytics as quite difficult. The persistence growth of computational data analysis has produced a tremendous flow of data streams as compared to past records. This tremendous increased in data accessibility and usage introduce many problem into field of Big Data. Such as the data are now represented in complex, multi-structured, unstructured, heterogeneous format and these data are mostly generated by different data sources. The major issue focuses as data processing, analyzing correct information from massive datasets, handling huge customer data in Big Data warehouse, etc. Thus inorder to overcome with Big Data issues and challenges that are observing in recent data analysis work can be resolved by the following points: Design an appropriate system to handle the data efficiently.

Investigate and identify the issues that are associated in Big Data storage, management and processing.

Figure 4.1 Big Data Framework

Hence in this thesis work we are working on the grid density based clustering algorithm in Big Data. The application of clustering techniques in Big Data helps in analyzing and pattern evaluation by study the cluster formed using different massive datasets. Different Researchers has developed different types of clustering Algorithms that varies in there clustering properties, cluster model and further each cluster models was implemented with different types of Algorithms. The visualization of this cluster models varies from one another significantly with their respective properties. Defining the clustering 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".

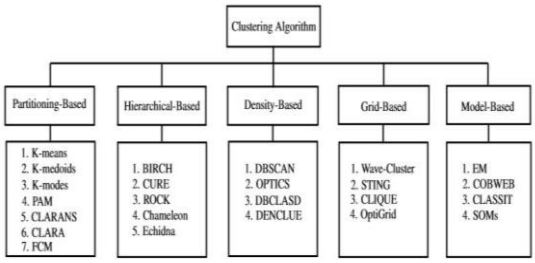

Figure 4.2: Classification of clustering Algorithm

The Cluster analysis is executed by applying of various types of clustering algorithm and the most common clustering methods are as follows:

4.2 Implementation and Environmental Setup:

MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation. Using the MATLAB product, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and FORTRAN. You can use MATLAB in a wide range of applications, including signal and image processing, communications, control design, test and measurement, financial modeling and analysis, and computational biology. Add-on toolboxes (collections of special- purpose MATLAB functions, available separately) extend the MATLAB environment to solve particular classes of problems in these application areas.

4.1.1 Introduction to MATLAB:

MATLAB (matrix laboratory) is a numerical computing environment and fourth generation programming language. Developed by Math Works, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, and FORTRAN. Although MATLAB is intended primarily for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities. An additional package, Simulink, adds graphical multi-domain simulation and Model-Based Design for dynamic and embedded systems. In 2004, MATLAB had around one million users across industry and academia. MATLAB users come from various backgrounds of engineering, science, and economics. MATLAB is widely used in academic and research institutions as well as industrial enterprises.

4.2.2 MatLab Requirement:

The below figure give the details of MatLab for windows operating system:

Windows

Operating Systems	Processors	Disk Space	RAM
32-Bit and 64-Bit MATLAB and Simulink Product Families			
Windows 7 or Service Pack 1 Windows Vista Service Pack 2 Windows XP Service Pack 3 Windows XP x64 Edition Service Pack 2 Windows Server 2008 Service Pack 2 or R2 Windows Server 2003 R2 Service Pack 2	Any Intel or AMD x86 processor supporting SSE2 instruction set*	1 GB for MATLAB only; 3–4 GB for a typical installation	1024 MB (At least 2048 MB recommended)

Figure 4.3 MatLab Specification for Windows operating system

Linux

Operating Systems	Processors	Disk Space	RAM
32-Bit and 64-Bit MATLAB and Simulink Product Families			
Qualified distributions*: Ubuntu 10.04 LTS, 10.10, and 11.04 Red Hat Enterprise Linux 5.x and 6.x SUSE Linux Enterprise Desktop 11.x Debian 5.x	Any Intel or AMD x86 processor supporting SSE2 instruction set**	1 GB for MATLAB only; 3–4 GB for a typical installation	1024 MB (At least 2048 MB recommended)

Figure 4.4 MatLab Specification for Linux operating system.

V. SIMULATION AND EXPERIMENTATION OF PROPOSED ALGORITHM

5.1 Proposed Grid Based Clustering in Big Data:

Clustering is a common technique for the analysis of large data sets. The clustering algorithms are efficient in mining large multi-dimensional data sets called as GRID BASED CLUSTERING ALGORITHM IN BIG DATA (GCAB). In the following thesis work we have implemented to grid based clustering of very large data sets is presented using MatLab ver7.01.The GRIDCLUS algorithm uses a multidimensional grid data structure to organize the value space surrounding the pattern values, correlation rather than to organize the patterns themselves. The patterns are grouped into blocks and clustered with respect to the blocks by a topological neighbor search algorithm. The overall runtime behavior of the algorithm output performs all conventional hierarchical methods. And we focus on the comparisons between the different types of grid clustering methods. This algorithm divides the data space into a finite number of cells in grid structure and generates the cluster inform of grid structure. The clusters in the grid region are denser in data

points than their surrounding datasets. The benefits of GCAB algorithm are more significant in reducing time complexity specially when applied for very large data. The clustering points developed in the GCAB clustering algorithm easily calculate the neighboring data cluster points

5.2 Basic Outline of Grid structure

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.

Creation of Grid Bocks

A grid Block is constructed as d-dimensional hype-rectangle (rectangle shaped cube) containing maximum of B_s pattern called as block size.

Let $X = \{x_1, x_2 \ldots x_n\}$ be the set of n-patterns. $X_i = i^{th}$ pattern consisting of a tuple with describing features as $(a_{i1},a_{i2},a_{i3} \ldots\ldots\ldots a_{id})$ 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, $X_i \in B_j$ ..... eq(1)

$B_j \cap B_k = Ø$; if $j \neq K$ ..... eq(2)

$B_j \neq Ø$ and

$\cup B_j = X$

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 $[C_{lu1},C_{lu2},C_{lu3},\ldots\ldots\ldots C_{luwn}]$ 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

Cluj=Bj

Such that j =1, 2……b and $W_0 = b$.

Calculation of Density Indices:

In the grid density based algorithm (GCAB) 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.

$V_b = \sqcap e_{bi}$

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).

$\therefore Db = Pb/Vb$……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 .

VI. PRESENTATION AND ANALYSIS OF RESULTS

In this chapter we have given the details of different types of datasets used in GCAB algorithm and the result analysis figure using the MatLab 7.01 version.

6.1 Synthetic Datasets:

We generate synthetic data sets in matrix forms using the MatLab function which to generate a sequence of pseudorandom number in the available work space.

Syntax >>% rand(n) function used in MatLab to generate the uniformly distributed pseudorandom numbers such as r = rand(n) returns an n-by-n matrix containing pseudorandom values drawn from the standard uniform distribution on the open interval (0,1). rand(m,n) or rand([m,n]) returns an m-by-n matrix. rand(m,n,p,...) or rand([m,n,p,...]) returns an m-by-n-by-p-by-... array. rand returns a scalar. rand(size(A)) returns an array the same size as % randn used for normally distributed random number i.e. Generate a 5-by-5 matrix of normally distributed random numbers. % r = randn(5) gives the output as :

Using different type of function in mat file we generate the different types of numerical data and applied in the GCAB algorithm to study and verify the different cluster pattern generated in 2D and 3D grid structure.

The 2D, 3D GCAB algorithm output. 2D- GCAB Algorithm Elapsed time is 0.429920 seconds for generating 500 normal distributed random numbers in formation of 3 grid structure cluster pattern. The Elapsed time for 3D-GCAB Algorithm is 1.730009 seconds. The following table 7.1 gives the details of synthetic data used in GCAB algorithm.

6.2 Cancer Datasets:

This datasets is related to disease called as Adenoma cancer which is a most critical disease occur in many human being and nearly 1 millions of people suffer with the Adenoma cancer disease by 2000year. A per record studied 2% of total number of cancer patients increased every year worldwide. This datasets is used from the PRICETON University, New Jersey, USA (/genomics-pubs.princeton.edu/oncology/). The important Cancer Research Project is to identify the actual stage of Tumor causing the cancer in human begin and the diagnosis of such disease should be done in early stage. These cancer genes are growth is maximum in the maligent stage. It was observed that the gene expression intensity was found ad different types such as Adenoma Gene, Normal

Gene, sometime colon Adenoma carcinoma sample and paired normal sample were hybridized to Gene chips study by Affymetrix. The Gene expression level was studied by Gene chips 3.0 software analysis by Affymetrix. But approx. 2% of 4000 gene could not be analyzed by this software and many advance gene expression with high intensity between tumor and normal sample could not be studies well. The cancer dataset we have implemented in GCAB algorithm using the MatLab 7.01version and observed a well defined cluster and visualization cluster patterns. The total number instances in datasets is 7162 out of which we have used 5000 and different types of attributes which include the gene class as Adenoma and Normal Adenoma The elapsed time is recorded as 82.3469 sec. The cluster are represented using the different type of markers specified in MatLab 7.01version and the datasets represented as formation of grid based clustering output for clustering different Adenoma gene in class 1 , Normal Adenoma, and Adenoma gene class3.

Output result:

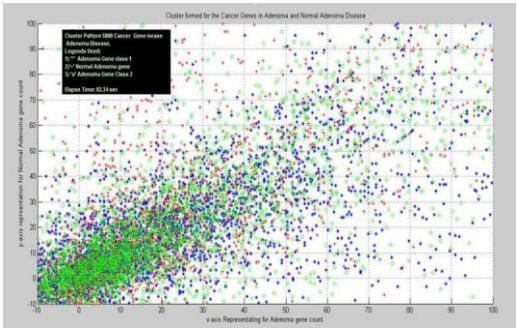

Figure 6.1 cluster pattern evaluation incase of Adenoma Cancer Diseases.

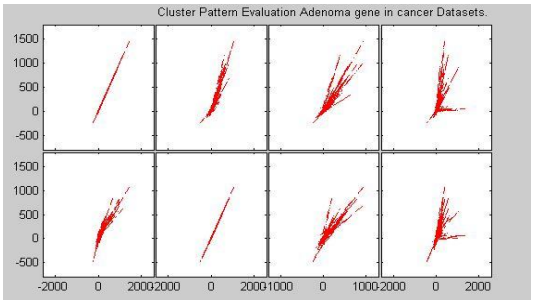

Figure 6.2: Representation of Adenoma Gene in Cancer Datasets.

6.3 Gene Expression in Tissue of Yeast:

Yeast micro array datasets used by UCI Machine Repository given with sample of different types of gene expression for a probabilistic clustering system for predicting the cellular localization sites of "Proteins". The resultant and pattern evaluation by using the yeast datasets was observed as 55% of yeast data are present in adhoc structure. This information about the Yeast Micro array tissue synthesis was given by the Institute of Molecular and Cellular Biology, Osaka, University in 1996 associated with the Ecoli Database. This datasets contains number of instance as 1484 and 8 predictive number of attributes along with the one attribute for naming of yeast tissue sample in Accession number (archive.ics.uci.edu/ml/datasets/Yeast). We have implemented the yeast dataset in MatLab 7.01version using the proposed GCAB algorithm and record the run time efficiency as 17.62 sec and the cluster pattern observed with protein count in Yeast Tissue.

Output:

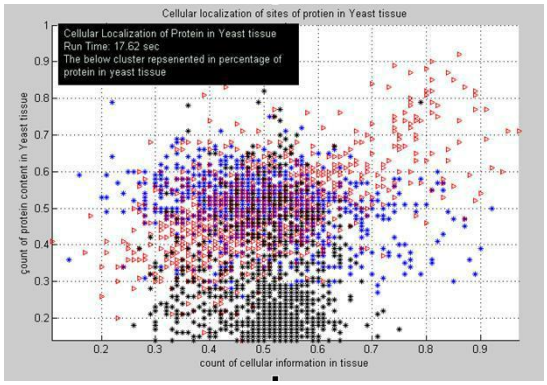

Figure 6.3 Gene Expression of Tissue in Yeast Dataset using GCAB algorithm

6.4 Real-time Datasets:

This is a sample real time dataset given by Spatialkey service provider such as the software application used for various full time potential of time and location based information collected by the red time observation images through satellite data in different point of observation (Latitude and Longitude) data points over a particular area in geographical distribution area. Such application of software tools is mostly associated with study of Big Data in the field of Market Analysis, Banking, and Business Intelligence etc. This datasets is related to insurance field containing total number of instance as 36634 in Florida for a Banking company in 2011-2012 report. The attribute of the datasets are associated with total insurance value (TIV) and geographical records longitude and latitude location, Customer-id, insurance policy-id.

Output:

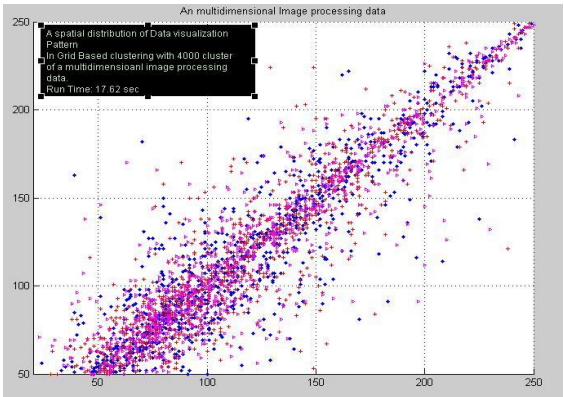

Figure 6.4: Data visualization of Image datasets using GCAB algorithm

VII. CONCLUSION

In this thesis work we studied Big Data, Big Data Analytics and its categorization i.e. Data Volume, Velocity, Variety and Veracity. These data format are represented in complex structure and created difficult in Big Data Analytics and define the challenges that big data are facing today for

storing, sorting, analyzing and visualization. The Clustering algorithms have emerged as an alternative powerful and meta-learning tool helps to analyze the massive volume of data (Big Data) generated by many applications. In general, we studied that Big Data creates a lot of confusion while categorization of Big Data. Therefore it is important to choose a relevant clustering algorithm to classify these massive data and help in data clustering formation and pattern evaluation process. such recent techniques used to handle Big Data and extract the hidden, relevant information for these complex massive data sets along with of execution of the massive data sets into a secrete correlation pattern called as "Big Data Mining". The details process of Big Data in Data mining given in the proposed framework. Further we have implemented the Grid Density Based Clustering in Big Data (GCAB) Algorithm in MatLab. Grid Density based clustering is concerned with the value space that surrounds the data points not with the data objects. This algorithm uses the grid data structure and use dense grids to form clusters. It first quantized the original data space into finite number of cells which form the grid structure and then perform all the operations on the quantized space. Grid based clustering maps the infinite amount of data records in data streams to finite numbers of grids. Its main uniqueness is the fastest processing time, since like data points will fall into similar cell and will be treated as a single point. It makes the algorithm self-governing of the number of data points in the original datasets. Grid density takes the advantage of the density and the grid algorithms. Grid density is suitable for handling noise. It can find the arbitrary shaped clusters used for high dimensional data. The grid density algorithm does not require the distance computation. K-mean knows the number of clusters in advance but the grid density does not. Grid density algorithm is better than the k-mean algorithm in clustering. The advantage of grid density method is lower processing time. Therefore, we implement the grid density clustering algorithm for analyze and increase the speed, and accuracy of the dataset.

Future Work

In this paper we have observed the main issue in Big Data and its categorization on the basis of data accessibility and define the challenges that big data are facing today for storing, sorting, and analyzing. we disclosure different types of clustering algorithms .We suggest and investigate different types of data clustering algorithms and its implementation to big data , optimize and calculate the efficiency of such algorithm suitable to handle massive Big Data and applicable for multi dimensional data sets . In Future we will be working on various Grid Based Clustering Algorithms used to handle the Multi Dimension datasets including the major concepts for verifying the Sensitivity areas in Grid Based Clustering Algorithm.

REFERENCES

[1] SAGIROGLU and Duygu SINANC, "Big Data: A Review", Gazi University Department of Computer Engineering, Faculty of Engineering Ankara .978-1-4673-6404-1/13/2013 IEEE.

[2] C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data" Mc Graw-Hill Companies, 978-0-07-179053-6, 2012

[3] http://en.wikipedia.org/wiki/Big_data , last access: 20/4/2015

[4] From Big Data to Meaningful Information - Insights from a webinar sponsored by KMWorld Magazine and SAS; Conclusion paper; 2013.

[5] Big Data Meets Big Data Analytics; SAS Institute - White paper; 2012.

[6] John Gantz and David Reinsel. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. URL www.emc.com/leadership/digital-universe/index.htm

[7] Doug Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. URL http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[8] http://www-01.ibm.com/software/in/data/bigdata/

[9] A. Fahad, N. Alshatri, Z. Tari, Member, IEEE , A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras. "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis" IEEE,2012

[10] Andrew McAfee and Erik Brynjolfsson. Big Data: The Management Revolution. Harvard Business Review, October 2012:60–68, October 2012

