JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.7, Special Issue.3, Feb. 2019 E-ISSN: 2347-2693

Data Mining Algorithm and New HRDSD Theory for Big Data

Kamlesh Kumar Pandey[1*], Diwakar Shukla[2], RamMilan[3]

[1,2,3]Dept. of Computer Science & Applications, Dr. Harisingh Gour Vishwavidyalaya, Sagar, M.P., India

[*]*Corresponding Author: kameshamk@gmail.com*

Available online at: www.ijcseonline.org

Abstract— In a present time data is king for any organization. IT industry or any organization is based on data for making any type of decision like how to grow on the organization, market analysis, and consumer relationship analyses so on. In Present time data characters are changes in form of data to big data. Data is very helpful for making the decision for any organization through data mining. Big Data mining is the process of extract interesting knowledge from huge streams based databases, which hold characteristics of Big data volume, variability, velocity, variability, value, veracity, and visualization. When applying to data mining algorithm in the big dataset it gave some useful information and some not because our data mining algorithm is can't handle all characteristics of big data at a time. This paper presented to how to data converted tradition to big data, basic characteristics of big data, suitable big data mining algorithm. We can also propose HRDSD theory for Big data mining which is useful on developing new big data mining algorithm and framework.

Keyword:- *Big data, Data mining, Mining Algorithm, HRDSD, 3V..*

I. INTRODUCTION

Information technology is growing on day by day and its usage of also increasing in public domain. Uses of data are growing on very fast in form of machine and user data. Data is the collection of specific values and variables related to special sense. Data are collected and analyzed to create information suitable for making decisions. Information is only useful when data are studied a properly. Proper studied data provide a rich resource for knowledge discovery and decision support.

A. Big data

Big data term introduces by *John Mashey et al* (1998) published a book which title name is "Big Data and the Next Wave of Infra Stress" after that *Weiss and Indrukya et al* (1998) are discussed to Big Data mining on own book. In 2000, *Diebold* is the first author who's discussed to own academic contribution with the words Big Data. Hear above author define a large amount of data which is greater than Gigabytes is known as Big Data. *Usama Fayyad et al* (2012) are given some information about data generation in KDD BigMine. According to *Usama Fayyad et al* (2012) Google search engine handles more than 1 billion queries, Twitter handles more than 250 million tweets from all twitter user, Facebook handles more than 800 million all type of updates as comment, like, post, friend request so on and YouTube handles more than 4 billion video queries on per day[1]. Uses of information technology are defined some new large source of data like mobile devices, electronic device and biggest IT and social companies like Google, Apple, Facebook, Yahoo, Twitter etc[2]. According to International Data Corporation are predicts between 2005 to 2020 the global data volume will be grown by 300 factor [3]. Big data is *too big*, *fast* and *hard* for processing on existing tools for analysis and mining. A verbal concept *too big* defined the scale of petabyte collections of data, *Too fast* defined processed to large amount of data is very less time, *Too hard* defined existing processing tool is taken difficulty on data analysis because IT is changing readily [4]

*Doug Laney et al (*2001*)* was the first person those given basic characteristics of big data which are known as 3 V's characteristics of Big Data Management. These are a Volume, Variety, and Velocity. If any data comes under above any two or more characteristics they are known as big data [1][5]. *Volume* defines the amount of data in particular data storage which size is up to Gigabytes. Big data volume is always increasing in continues way like Facebook or Wal-Mart which above 1 million data are generated to one hour [2]. *Variety* defines different types of data and sources. In general, data can be structured, unstructured and semi-structured form. [6]. *Velocity* refers to the rate of data generated at particular source and the speed which are required on analyzed the static and dynamic data like airplane set up box are data is generated in high speed in continues way [7].

Structured data is organized in a proper format like RDBMS, OLAP, Data warehousing, spreadsheets and it is easier to analyze. Sources of structured data are web or any organization based application like business Applications such as retail, finance, bioinformatics etc. unstructured data have not as any format for storage and it can be difficult to analyze and categorization. Audio, video, pdf, email, image etc. comes from unstructured data through weblogs or machines. Semi-structured data is between structured and unstructured data. A data whose are suitable for conversion on structured data to unstructured data and unstructured data to structured data. Extensible Markup Language (XML) is a typical example of semi-structured data which are a textual language for exchanging data on the Web [5]

W. Fan et al (2012) describe two more characteristics are Big Data. These known as Variability and Value [1]. *Variability* refers to the variation of Variability defines of data whose meaning and structures constantly change like two users are given the same key in Google and Google may be return different answer in same key [4]. *Value* defining an attribute of big data. It can be very useful for mining on knowledge in the structured, unstructured and semi-structured data. *Amir Gandomi et al (2015)* are defining one more characteristic of big data in own academic contribution these are known as Veracity. IBM coined *Veracity* for mining Purpose which represents the data quality and accuracy in some sources of data. [5]. *Sivarajah Uthayasankar et al (2017)* discussed next characteristics of big data which is known as known as visualization. *Visualization* refers to presenting the data in a proper manner which is easier to accessible and understandable form for the end user. These characteristics are very helpful for presenting on large volume and variety of data in the smaller way. In present time total, seven V’s are founded on Big Data. We will summarize these characteristics in **fig 1.**

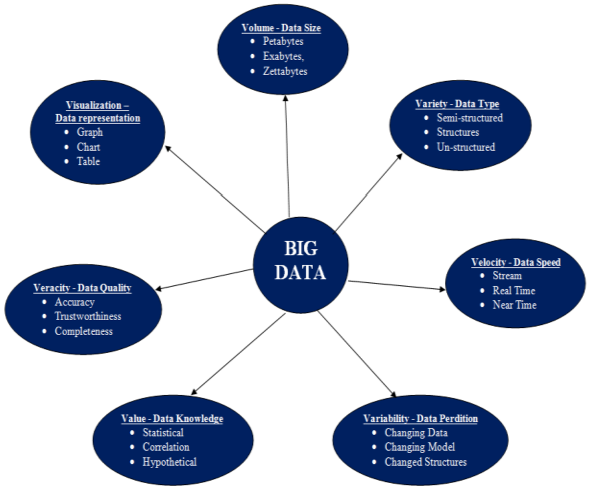

Figure 1- 7 V’s characteristics for Big Data

According to characteristics of the big data Relational database management system (RDBMS) are not suitable to process big data because the traditional database can be processed only structured data and it is not scalable as the rate of generation of big data is very high [3][8]. We also are given the difference between traditional data and big data on table 1 according to this V’s.

Table 1- Different between Tradition data and Big data

Characteristics	**Tradition data**	**Big data**
Volume	Volume of data is Megabyte/Gigabyte size.	the volume of data is Terabyte/ Petabyte size or greater them.
Variety	Analysis of data captured from limited and known sources and data is always in structured form.	Analysis of data captured from unlimited and unknown sources and data is always in structured data, semi-structured data and unstructured data
Velocity	Per hour	Per microsecond
Variability	Database structures are static	Database structures are constantly changes on per mintues
Value	Database attributes are pre-defined	Database attributes are pre-defined or may not.
Veracity	Data quality and Accuracy is handled by techniques.	Data quality and Accuracy is biggest issue on big data handling.
Visualization	Presenting and identifying on data in easier techniques.	Presenting and identifying on data in complex techniques.
Analysis	SQL approach use to data analysis with statistical method.	Massively parallel processing and No SQL approach use to data analysis with statistical method based on programming tools.
Framework	Relational database framework which can be handle historical, static data.	Hadoop framework, MapReduce framework or Spark framework used for Stream processing of real-time or live data.
Database Structures	Central database.	Distributed database.

B. Data Mining

Data mining is the process of discovering interesting knowledge such as associations, patterns, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses or other information repositories [9]. Alternative names are data mining is Knowledge discovery in databases (KDD), knowledge mining, knowledge extraction, data/pattern analysis etc. data mining process consists of seven steps sequence. First, four steps Cleaning, Integration, Transformation, and Selection of data come under data preprocessing and after three step Data mining, Pattern evaluation and Knowledge presentation are comes undermining process [10]. *Brachman and Anand et al (1996)* gave a KDD process. The first step is data mining is Selection where select a relevant data for analysis task in the database. The second step is data cleaning and preprocessing, in this step removing noise and inconsistent data. The third step is combining multiple data sources into one place or together at one database, these are called data integration. The fourth step is Transformation where data are transforming into appropriate forms for data mining processing. The fifth step is data mining process where we used various type of data mining algorithm for finding on interested knowledge. Six steps are Interpretation/Evaluation interpreting the patterns into knowledge by removing redundant or irrelevant patterns and translating the useful patterns into terms that human understandable form [8].

Big data and data mining both are different techniques and concept for data analysis. Big data and data mining are related to high volume of databases to handle the collection or reporting of data. Data mining algorithms uncover interesting patterns and hidden relationships in a large volume. When data mine technique fails for find out on the relationship between unstructured or structures data in high speed then comes to Big Data techniques. Data mining techniques are also used in big data analysis in upgraded form. Present time very needs for upgrading on data mining technique for works less than three V's of data mining. We try to give in **table 2** on the difference between data mining and big data [6] [10].

TABLE 2- DIFFERENCES BETWEEN BIG DATA AND DATA MINING CONCEPT

Big data	**Data mining**
Big Data is defined as a large set of data that is available structured, unstructured, semi-structured in non-defined data source.	Data mining is defined as a large set of data that is available in structures from in predefined source.
Big data is capable for given relations of among data set with finding interesting knowledge.	Data mining is not capable for given relations in every time. Data mining is capable for finding interesting patterns or knowledge from datasets.
Distributed database based mining techniques used.	Centralized database based mining techniques used.
Big data handles lot of mining algorithm for at a time.	Data mining handles on one data mining algorithm for at a time.
Big data mining is based on real time.	Traditional Data mining based on non-real time.

II. DATA MINING ALGORITHM FOR BIG DATA

There are predictive, descriptive and prescriptive types of analysis can be done in order to retrieve needed information from big data. Different type of analysis will hold different impact and result. Data mining technique depends on the type of business problem, types of data and sources, dynamic or static data etc that you are trying to solve. In this section, we introduce common data mining algorithm which is suitable for big data [11].

1 Classification Analysis: - It is a supervised based data mine algorithm. Classification is a systematic process of grouping the similar data into different classes or identifying to which of a set of categories different types of data on the basis of structures using discrete function and obtaining relevant information about data and metadata. Some classification technique is Decision tree, Bayesian classification, rule base Classification, neural network, K-NN algorithm etc [12].

A. Decision tree induction classification algorithms: - In decision tree induction algorithms is suitable for analyze and categories the big data. Decision tree classifiers are useful for break a more complex decision into a collection of the simpler decision. In decision tree structure all internal node represents a test on an attribute, all branch represents a result of the test, each leaf node represents a class label and topmost node in a tree is the root node. [13]

B. Evolutionary-based classification algorithms: - Evolutionary algorithms used for selecting proper data for analysis in good optimization solutions and solution of the multi-objective problem. There are different types of evolutionary algorithms such as genetic algorithms, evolution strategies, evolutionary programming and so on. Genetic algorithms were mostly used for mining classification rules in large datasets [14]. *Patil et al (2006)* proposed a hybrid technique using for genetic algorithm and decision tree that improving the efficiency and performance of computation.[15][16].

C. Neural Network (NN) based Approach: - neural networks are non-linear statistical data modeling approach. This approach is suitable for compute the complex relationship between input and output and finding patterns in a large amount of data [16].

D. Bayesian classification: - Bayesian classifiers are a type of statistical classifiers approach. This approach predicts data or class membership using Bayes' theorem probabilities, such as the probability that a given tuple belongs to a particular class [17].

***E. Support Vector Machine:* -** Support Vector Machine is an algorithm for the classification of both linear and nonlinear data. This approach transforms the original data in a higher dimension, from where it can find a hyperplane for separation of the data using essential training tuple called support vectors. [17]

***2. Clustering analysis:* -**It is an unsupervised based data mine algorithm. Clustering is a process of grouping similar objects into classes. The cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering analysis is the process of identifying data sets that are similar to each other to understand the differences as well as the similarities within the data [18]. These can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods and constraint-based methods [17].

***A. Partitioning based clustering algorithms:* -** In this approach, large data sets are divided into a number of partitions known as *k* partitions, where each partition represents a cluster they known as K-mean[13]. *J. C. Bezdek et al* proposed Fuzzy- C Mean's approach using K-means technique for distributed large dataset [16].

B. Hierarchical based clustering algorithms: - In this approach, large data are organized in a hierarchical manner based on the medium of proximity. First initial node is called root cluster which can derive several child clusters. It follows a top-down or bottoms up a strategy to represent the clusters [13]. *T. Zhang et al* proposed to Birch algorithm using hierarchical clustering which handles streaming data in real time and extracting semantic content was defined in Hierarchical clustering for concept mining [21].

***C. Density-based clustering algorithms:* -** in this approach, clusters are formed based on the data objects regions of density, connectivity, and boundary. Each cluster grows in any direction based on the density growth [11]. *A. Hinneburg et al* proposed DENCLUE algorithm using density based algorithms which can handles, separating on a different type of data and mining large amount of data [21].

***D. Grid-based clustering algorithms:* -** in this algorithm, clusters are divided into a number of grids for fast high processing. *A. Hinneburg et al* proposed OptiGrid algorithm which handles terabytes volume data and its according to this approach cluster defined as a finite number of a cell that forms a grid structure [21].

***E. Model-based clustering algorithms:* -** *A. P. Dempster et al* proposed clustering algorithms for incomplete data. According to this algorithm, clustering is performed by a probability distribution.

3.Association rule learning:- Association rule learning enables the discovery of interesting relations between different attribute and data in large databases. This approach uncovers hidden patterns in the data that can be used to identify variables and attributes within [16].

4. Regression: - Regression algorithm is based on a statistical prediction model. Classification technique predicts categorical classes and prediction technique predicts the class using continuous valued functions. Regression is a linear, nonlinear and generalized linear technique for prediction. At most all nonlinear problems are converted to linear problems by performing transformations using predictor variables. Like decision tree, Regression tree and model tree are used for prediction. In regression trees, each leaf stores a continuous values prediction and model tree, each leaf holds a regression model [17].

III. HRDSD THEORY

In a present time, a lot researcher proposed various data mining technique for the traditional database management system. The concept of big data is coming in the year 1998 but uses of big data is *increases* to after 2014, according to [2][5] in present time not available proper data mining technique for big data mining because of a till date not coming to a proper definition of big data. Like in 2001 Big data definition defined three characteristics volume, verity, velocity after that in 2012 Big data definition defined two more characteristics Variability and Value after that in 2015 and 2017 Big data definition defined two more characteristics Veracity and Visualization. In general biggest IT companies like Facebook, Google, Yahoo, Amazon, Intel IBM etc. say if any data hold to volume, verity, and velocity with or without any other characteristics then this data comes under to Big Data. Various researchers have given to a lot of algorithm for individual characteristics of big data but they not given to any specific algorithm for big data mining. In this cause, we proposed to HRDSD theory for design on proper big data mining framework or algorithm. This theory is very helpful for design big data mining algorithm, framework or proper definition because this theory covers basic concept, characteristics, and nature of big data.

HRDSD theory has five characters which defined natures of big data. First **H** defined a high volume, speed, and dimensions, **R** defined complex relationships between among them, **D** defined distributed source and deferent type of data, **S** defined streaming or continues the way of data generation in real-time and **D** defined data. If any Data were taken to **H**igh storage space

When nature and behavior of traditional data follow to HRDSD characters so these come under Big data. According to this theory in present time "data generated in very **H**igh Speed then data volume is growing to very **H**igh in per minutes. Growing on the data in **S**treaming or continues from in **D**istributed sources where every source stored **D**ifferent type of data like structures, unstructured and semi structures with **H**igh dimensions with complex **R**elationships".

This theory defined if we merge to various or suitable data management technique then we obtained goods Big data mining algorithm like we merge suitable algorithm for the stream, distributed, association, unstructured so on data management technique then the new Big data mining algorithm can do to handle any type of data which are generated are stream way in distributed source and we also find out the relationship among them. We gave to some notable point which can help to selection on algorithm under the data management under to this theory.

1. Choose those classifications, clustering algorithm or another algorithm which are able to handle a large amount of data and separated to a different type of data from each.

2. Choose those classifications or clustering algorithm or another algorithm which are very helpful for finding patterns or knowledge in streaming data in less time.

3. Choose that algorithm which cans preprocessed and communicates on distributed database system or network support system.

4. Choose to best association rules or statistic tools which are able to extract complex relationship and dimensions between any type of data, attributes in less time from a large volume.

IV. CONCLUSION

Big data is growing to very fast in last few years. In present time a lot of problems is involved in big data mining like if we apply stream base data mining technique in dataset then other feathers of big data mining do not work to properly like data mining technique fails to handle unstructured data management, distributed data management. This paper describes what is acutely meaning of big data how to apply data mining algorithm in big data, which data mining algorithm is suitable for big data mining, what is basic feathers of big data so on in section first and second. In third section of this paper we proposed to HRDSD theory for designing to big data mining algorithm and framework. This theory holds to all basic characteristics of big data and given to the concept and proper way of design of big data mining algorithm with their framework.

REFERENCES

[1]. Fan Wei and Bifet Albert (2012): "Mining Big Data: Current Status, and Forecast to the Future", ACM SIGKDD Explorations Newsletter, V-14, I-2, pp 1-5.

[2]. Pal Kaushika and Saini R. (2013): "A Study of Current State of Work and Challenges in Mining Big Data", International Journal of Advanced Networking Applications (ISSN No: 0975-0290), V-12, I-3, pp 73-76.

[3]. Tiwarkhed A.S. and Kakde Vinit (2013): " A Review Paper on Big Data Analytics", International Journal of Science and Research (ISSN 2319-7064), V-4, I-4, pp 845-848.

[4]. Sagiroglu Seref and Sinanc Duygu (2013): "Big Data Review", Published in the Proc. Of International Conference on Collaboration Technologies and Systems (CTS), Published By IEEE, held in San Diego, CA, the USA at 20-24 May 2013, pp 42-47.

[5]. Gandomi Amir and Haider Murtaza (2015): "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, Published By Elsevier, V-35, pp 137-144.

[6]. Parmar Vinti and Gupta Itisha (2015): " Big data analytics vs Data Mining analytics", International Journal in IT and Engineering(ISSN: 2321 –1776), V-3, I-3, pp 258-263.

[7]. Laney D. (2001): "3-D Data Management: Controlling Data Volume, Velocity, and Variety", META Group Research Note Published By Springer, V- 6, pp 34-42.

[8]. Fayyad Usama and Piatetsky Shapiro Gregory (1996): "From Data Mining to Knowledge Discovery in Databases" Artificial Intelligence Magazine, V-17, I-3, pp 37-54.

[9]. K.U. Jaseena and David M. (2014): "Issue Challenges and Solution: Big Data Mining", Published in the Proc. Of SMTP-2014, Published By AIRCC Publishing Corporation, held in Chennai, India on 27-28 Dec 2014, pp 131-140.

[10] Pandey Kamlesh (2014): "An Analytical and Comparative Study of Various Data Preprocessing Method in Data Mining" International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459), V-4, I-10, pp 174-180.

[11]. Arun K.and Jabasheela L. (2014): "Big Data: Review, Classification and Analysis Survey", International Journal of Innovative Research in Information Security(ISSN: 2349-7009), V-1, I-3, pp 17-23.

[12]. Manisha R. Thakare, S. W. Mohod (2015): " Various Data-Mining Techniques for Big Data" Proceedings of the International Conference on Quality Up-gradation in Engineering, Science and Technology Published By IJCA Held on Wardha India in 9-13 October 2015.

[13]. A Sherin, Uma S and K Saranya K (2014): "Survey on BIG Data Mining Platforms, Algorithm, and Challenges", International Journal of Computer Science & Engineering Technology(ISSN:2229-3345), V-5, I-9, pp 854-862.

[14]. D. L. A Araujo., H. S. Lopes, A. A. Freitas (1999), "A parallel genetic algorithm for rule discovery in large databases", Published in the Proc. Of IEEE Systems Man and Cybernetics Conference, Published By IEEE at Tokyo, V-3, pp 940-945.

[15].Mr. D. V. Patil, Prof. Dr. R. S. Bichkar (2006),"A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large DataSets", Published in the Proc. Of IEEE International Conference on Industrial Technology, Published By in IEEE, Held in Mumbai, India on 15-17 Dec.

[16] Jiang Heling and An Yang (2016): "Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery", International Journal of Hybrid Information Technology(ISSN:1738-9968), V-9, I-3, pp 179-188.

[17]. Han Jiawei, Kamber Micheline and Pei Jian (2011):"Data Mining: Concepts and Techniques" Published By Morgan Kaufmann, ISBN 978-9380931913.

[18]. Jiang Heling and An Yang (2016):"Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery", International Journal of Hybrid Information Technology(ISSN:1738-9968), V-9, I-3, pp 179-188.

[19]. T. Zhang, R. Ramakrishnan, and M. Livny (1996), "An efficient data clustering method for very large databases" Published in the

Proc. Of ACM SIGMOD international conference on Management of data, held on Montreal, Quebec, Canada, V-25, pp 103–114.

[20]. A. Hinneburg and D. A. Keim (1998). "An efficient approach to clustering in large multimedia databases with noise", Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, held on New York on 27-31 August, pp 58–65.

[21]. A. Hinneburg, D. A. Keim (1999), "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering". Proceedings of the 25th International Conference on Very Large Data Bases held in the USA at 7-10 Sep pp 506–517.

[22]. Pandey Kamlesh (2018),: "Mining on Relationship in Big Data era Using Appiori Algorithm", Published in the Proc. Of National Conference on Data Analytics, Machine Learning and Security to be held on 15-16 February 2018 by Department of CSIT, GGV, Bilaspur, C.G, India, ISBN : 978-93-5291-457-9.