A Study on Data Mining Techniques and Tools for Big Data

Dr. R. Beena and C. Bhuvaneshwari

***Abstract*---**Big Data refers to large volume, growing data sets with heterogeneous, autonomous source such as Engineering, Genomics, Biology, Meteorology, Environmental research and many more. New technologies, systems and infrastructure must be developed in order to handle these data volumes. Deriving useful information from Big Data requires the development of increasingly sophisticated methods of mathematical and statistical analysis and the design of efficient algorithms.

The big data is constantly varying factor and newer algorithms and tools are continuously being developed to handle this big data. Big Data is all about exploring large volumes of unstructured, invaluable, imperfect, complex data and extract useful information or knowledge for future use.

The platforms such as GPU, Multicore CPUs etc. can be used to speed up the data processing. There tools like Hadoop, Spark, Dynamo, Pentaho, SAMOA etc., can be used to handle big data. Apart from the above mentioned big data platforms, there are many platforms available with different characteristics and choosing the right platform requires an in-depth knowledge about the capabilities of all these platforms.

This paper provides an in depth study on the various data mining algorithms and tools available for performing big data analytics.

***Keywords*---**Big Data Mining, Clustering, Classification, Big Data Tools, Hadoop, Spark, Pentaho, ASTERIX.

I. Introduction

DATA volumes and streaming rates are intensifying because most of the data are born digital as well as exchanged on internet today. Big data has rapidly developed into a hot topic that attracts extensive attention from academia, industry, and governments around the world. Although the advances of computer systems and internet technologies witnessed the development of computing hardware for several decades, the problems of handling the large-scale data still exist when entering the era of big data. The datasets that research now handles are not only large, but complex, containing unstructured, heterogeneous data, human language, image and video, and completely new approaches are required to handle them. Proper schemes are needed to manage and store data, and the proper management of metadata, including data on sample preparation, experimental parameters, and the data's provenance, is essential to enable Big Data to deliver trustworthy results.

II. Big Data

Big data is complex and heterogeneous data that spans across5 V'S: Volume, Variety, Velocity, Value and Veracity (In Fig.1).It is very difficult to manage and manipulate big data with the existing tools and techniques.

1. *Volume* - The size of data is very large and in terabytes and petabytes.
2. *Velocity*- It should be used when streaming in to the enterprise in order to maximize its value to the business. The role of time is very critical here.
3. *Variety* - It extends beyond the structured data, including unstructured data of all varieties: text, audio, video, posts, log files etc.
4. *Value*- The value dimension defines the value that the data can add to the organization.
5. *E.Veracity* - Big Data Veracity refers to the biases, noise and abnormality in data.

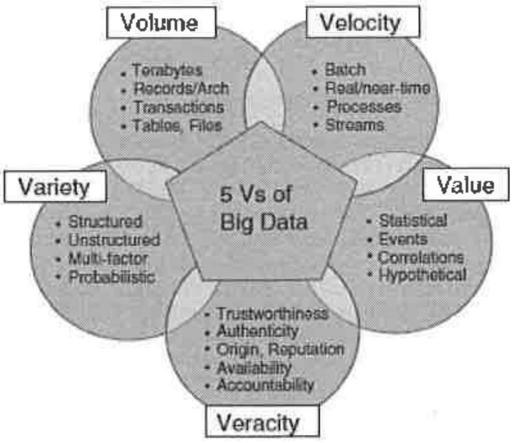

Fig 1-Five V's of BIG DATA

III. Big Data Characteristics

HACE Theorem: Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data.

Dr. R. Beena is with the Department of Computer Science, Kongunadu Arts & Science College, Coimbatore. E-Mail: beenamridula@yahoo.co.in

C. Bhuvaneshwari is with the Department of Computer Science, Kongunadu Arts & Science College, Coimbatore. E-Mail: bhuvaneshwari.c@grd.edu.in

A. Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors use their own schemata for data recording, and the nature of different applications also results in diverse representations of the data.

B. Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data sources is able to generate and collect information without involving (or relying on) any centralized control.

C. Complex and Evolving Relationships

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections which is one of the most important factors of the human society.

IV. Data Mining Techniques for Big Data

There are different types of techniques like classification, association rules, neural network, machine learning, clustering etc., The information is obtained from heterogeneous, multiple, autonomous sources with their complex relationship. Data is growing at a high speed.

It is very difficult for big data applications to retrieve manage and process data from large volume of data.

Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter. Data mining is an automated process used to extract valuable information from large and complex data sets. The several techniques in data mining classification and clustering are the main considerable point which is used to retrieve the essential knowledge from the very huge collection of data.

A. Decision Tree Induction Classification Algorithms

At the early stage Decision Tree was used to analyse big data. Tree structure has been widely used to represent classification models in the decision tree induction algorithms. Most of these algorithms follow a greedy top down recursive partition strategy for the growth of the tree. Decision tree classifiers break a complex decision into collection of simpler decision. Hall.et al. [16] proposed learning rules which generated a single decision system for a large set of training data. An efficient decision tree algorithm based on rainforest framework was developed for classifying large data set [17].

B. Evolutionary based Classification Algorithms

Evolutionary algorithms use domain independent technique to explore large spaces finding consistently good optimization solutions. There are different types of evolutionary algorithms such as genetic algorithms, genetic programming, evolution strategies, evolutionary programming and so on. Among these, genetic algorithms weremostly used for mining classification rules in large data sets [18]. Patil et al. [19] proposed a hybrid technique combining both genetic algorithm and decision tree to generate an optimized decision tree thus improving the efficiency and performance of computation. An effective feature and instance selection for supervised classification based on genetic algorithm was developed for high dimensional data [20].

C. K-means and Variant Partitioning Techniques

TQing He *et al.* [5] proposed a new technique for sole unseen feed forward neural network using ELM NMF (non-negative matrix factorization) and Extreme Learning Machine (ELM). The both methods are stable for very large datasets and are well-organized and support large applications. In [6], the authors discussed three new procedures that are used in clustering large data sets. The technique used was incremental k-means where clusters were using arithmetic data. In improved k-modes algorithm where the clusters were using definite data; while in the case of mixed arithmetic and definite figures, k- prototype is definite.

D. Other Partitioning Techniques

IshakSaida*et al.* [7] projected a novel meta heuristic procedure that can be used for data clustering which is founded on cuckoo search optimization (CSO) to eludetroublesomeness offered by k-means.

Xue-Feng Jiang [8] in his work has deliberated a global optimization technique that can be used for large scale computational difficulties. This technique uses the process of grouping but requires the skills to get paralleled.

Khadija Musayeva*et al.* [9] anticipated a novel clustering technique PFClust that will help in the discovery of ideal number of clusters spontaneously without any previous acquaintance of the amount of clusters.

E. Hierarchical-based

In hierarchical based algorithms large data are organized in a hierarchical manner based on the medium of proximity. The initial or root cluster gradually divides into several clusters. It follows a top down or bottom up strategy to represent the clusters.

Some of the well-known techniques of hierarchical clustering are BIRCH, CURE, ROCK and Chameleon. Hong Yu *et al.* [10] have offered ACADTRS (automatic clustering algorithm based on DTRS) a techniques based on hierarchical clustering, which routinely stops clustering once finding the required amount of clusters. They also suggested FACA-DTRS which is a quicker type of ACA-DTRS in terms of complication. Both techniques are proficient in terms of time and cost.

Buza*et al.* [11] presented a methodology to decrease the room that is wasted in the tick data which is swelling quickly. Using a novel clustering algorithm Storage Optimizing

Hierarchical Agglomerative Clustering (SOHAC) the preliminary tick data is separated by using clustering features into smaller data matrix. They also presented Quick SOHAC for hurtling up runtime.

Wang Shuliang*et al.* [12] presented in their paper Hierarchical Grid Clustering a new clustering technique Using Data Field (HGCUDF). Naim*et al.* [13] also suggested in their paper a clustering technique based on model, for high-dimensional large-sized datasets called SWIFT (Scalable Weighted Iterative Flow-clustering Technique).

F. Density-based

In density based algorithms clusters are formed based on the data objects regions of density, connectivity and boundary. A cluster grows in any direction based on the density growth. Techniques such as DBSCAN, OPTICS, DBCLASD and DENCLUE use a method to sifter out clatter (outliers) and determine clusters of arbitrary shape.

Kim *et al.* [14] suggested DBCURE algorithm to cluster very large data sets. This uses density based clustering technique along with parallelization. The authors also proposed a framework map reduce called DBCURE-MR. Both techniques that were proposed are proficient and will help in finding out clusters precisely for different data sets.

G. Grid-based

In grid base algorithms space of data objects are divided into number of grids for fast processing. OptiGrid algorithm is one such algorithm based on optimal grid partitioning [15].

Few of the typical examples are the Wave-Cluster and STING.

H. Model-based

In model based clustering algorithms clustering is mainly performed by probability distribution The most widely used model based technique is MCLUST, but recently several new additional good techniques have come up, such as EM (which uses density model mixture), conceptual clustering (such as COBWEB), and neural network approaches (such as self-organizing feature maps). Likelihood methods are used in neural networks in defining the clusters. To signify each resulting notion probabilistic images are typically used. A set of associated input/output units is used in neural network method, where single joining has a weight related with it. The widespread use of neural networks in clustering is because of their numerous properties. Firstly, parallel and distributed processing architectures are used in neural networks. Secondly, the learning method of the neural networks is done by regulating their interconnection masses which is used to fit the data. This permits the neural network to standardize or prototype. Arrangements act as features (or attributes) extractors for the numerous bunches.

Each group is made as an exemplar and this is the idea behind the grouping in neural network. An exemplar does not essentially have to resemble an entity, it just acts as a model of the group. Thirdly, measurable features are used in neural networks to identify the numerical trajectories. Many grouping tasks switch only arithmetical data or can alter their data into measurable structures if required.

V. BIG DATA TOOLS

A. Hadoop

Hadoop is the most significant open source distributed data processing platform for big data analytics. It belongs to the class NoSQL technologies (others include CouchDB and MongoDB) that have evolved to aggregate data in unique ways. Hadoop can serve as a data organizer or as an analytics tool. Hadoop offers a great deal of potential in enabling enterprises to harness the data that was, until now, difficult to manage and analyze. Specifically, Hadoop makes it possible to process extremely large volumes of structured as well as unstructured data. There are two important modules in Hadoop suchasHadoop Distributed File System (HDFS) and Map Reduce.

In HDFS files are split into one or more blocks and these blocks are stored in a set of Data Nodes. This facilitates the underlying storage for the Hadoop cluster.

MapReduce. MapReduce provides the interface for the distribution of the subtasks and then the gathering of the outputs. Similarly, when tasks are executed, MapReduce tracks the processing of each server/node. If it detects any anomalies such as reduced speed, going into a hiatus, or reaching a dead end, the task is transferred to another server/node that holds the duplicate data.

Some of the more notable Hadoop-related application development-oriented initiatives include Apache Avro (for data serialization), Cassandra and HBase (databases), Chukka (a monitoring system specifically designed with large distributed systems in view), Hive (provides ad hoc Structured Query Language (SQL)-like queries for data aggregation and summarization), Mahout (a machine learning library), Pig (a high-level Hadoop programming language that provides a data flow language and execution framework for parallel computation), Zookeeper (provides coordination services for distributed application).

B. HBase

HBase is an open source, non-relational, distributed database modelled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing BigTable-like capabilities for Hadoop.

C. Apache Hive

Apache Hive is a data warehouse infrastructure built on of Hadoop. It provides data summarization, query, and analysis of big data.

D. Dynamo

Dynamo, is another highly available data storage technology and has scalable distributed data store. Dynamo is used to manage the state of services that have very high reliability requirements and need tight control over the tradeoffs between availability, consistency, cost-effectiveness and performance. Dynamo has a simple key/value interface, is highly available with a clearly defined consistency window, is efficient in its resource usage, and has a simple scale out scheme to address growth in data set size or request rates. Each service that uses Dynamo runs its own Dynamo instances.

E. Apache Spark

Spark is an open source, scalable, massively parallel, in-memory execution environment for running analytics applications. Spark distribute data across a cluster, and process that data in parallel. The difference between MapReduce and Spark is that, unlike MapReduce—which shuffles files around on disk—Spark works in memory, making it much faster at processing data than MapReduce. Spark also includes prebuilt machine-learning algorithms and graph analysis algorithms that are especially written to execute in parallel and in memory. It also supports interactive SQL processing of queries and real-time streaming analytics.

F. Berkeley Data Analytics Stack

Today's data analytics tools are slow in answering even simple queries, as they typically require sifting through huge amounts of data stored on disk, and are even less suitable for complex computations, such as machine learning algorithms. These challenges are addressed by Berkeley Data Analytics Stack (BDAS), an open source data analytics stack. At the core of BDAS is Spark, an in-memory parallel execution engine, which enables us to provide unified support for batch, streaming, and interactive computations, as well as support sophisticated graph based and machine learning algorithms. Today, Spark and other BDAS components are used in production by tens of companies and institutions. In this talk, I'll present the architecture and the main design decisions we made in Spark, as well our future plans.

G. ASTERIX

ASTERIX is an Open Source System for big data management and analysis. With the help of ASTERIX Semi structured data can be easily ingested, stored, managed, indexed, retrieved and analyzed. Many of the drawbacks of Hadoop and similar platforms such as single system performance, difficulties of future maintenance, inefficiency in extracting data and awareness of record boundaries etc., are easily overcome by ASTERIX

H. SAMOA

SAMOA (Scalable Advanced Massive Online Analysis) is a upcoming platform for online mining in a cluster/cloud environment. It features a pluggable architecture that allows it to run on several distributed stream processing engines such as S4 and Storm. SAMOA includes algorithms for the most common machine learning tasks such as classification and clustering.

I. Pentaho

Pentaho provides big data tools to extract, prepare and blend the data, plus the visualizations and analytics within a single platform. Pentaho empowers business users and analysts to easily visualize, analyze, and report on data across multiple dimensions without depending on IT or developers.

VI. CONCLUSION

Big Data is regarded as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. This paper gives an overview of several data mining algorithms which explores knowledge from terabyte- and petabyte-scale datasets. Implementing the correct algorithm is very important to mine the required data within short span of time. This paper also gives an idea of the tools that support Big Data mining,

REFERENCES

[1] UzmaShafaque,ParagD. Thakare,Mangesh M. Ghonge,Milindkumar V. Sarode, "Algorithm and Approaches to Handle Big Data", International Journal of Computer Applications (0975 –8887) National Level Technical Conference "X-PLORE 14".

[2] Antonio Fernando Cruz Santos, I'talo Pereira Teles,Ota´vioManoel Pereira Siqueira, and Adicine´ iaAparecida de Oliveira , "Big Data: A Systematic Review", # Springer International Publishing AG 2018 S. Latifi (ed.), Information Technology – New Generations, Advances in Intelligent Systems and Computing 558, DOI 10.1007/978-3-319-54978-1_64

[3] Manisha R. Thakare,S. W. Mohod,A. N. Thakare, "Various Data-Mining Techniques for Big Data", International Journal of Computer Applications (0975 – 8887) International Conference on Quality Upgradation in Engineering, Science and Technology (ICQUEST2015) 9

[4] SaurabhArora, InderveerChana , "A Survey of Clustering Techniques for Big Data Analysis", 2014 5th International Conference- Confluence The Next Generation Information Technology Summit (Confluence) , 978-1-4799-4236-7/14/$31.00c©2014 IEEE

[5] Qing He, Xin Jin, Changying Du, FuzhenZhuang, and Zhongzhi Shi , "Clustering in extreme learning machine feature space" Neurocomputing, 12S:SS {9S, 2014} .

[6] R M adhuri, M Ramakrishna Murty, JVR Murthy, PVGD Prasad Reddy, and Suresh C Satapathy, "Cluster analysis on different data sets using k modes and k-prototype a lgorithmsN In ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society ofIndia-VolII, pages 137 {144. Springer, 2014.

[7] Ishak B oushakiSaida" KamelNadj et, and Bendjeghaba Omar, "A new algorithm for dataclustering based on cuckoo search optimization" in Genetic and Evolutionary Computing, pages SS {64. Springer, 2014.

[8] Xu e-Feng Jiang, "Application of parallel annealing particle clustering algorithm in data mining" TELKOMNIKA Indonesian Journal of Electrical Engineering, 12(3):211S{2126, 2014.

[9] Khadija Musayeva, Tristan Henderson, John BO Mitchell, and LazarosMavridis, "Pf clust: an optimised implementation of a parameter-free clustering algorithm" Source code for biology and medicine, 9(1):5, 2014. Copyright © SMART -2016 ISBN: 978-1-5090-3543-4, Different Clustering Algorithms for Big Data Analytics: A Review.

[10] Hong Yu, Zhanguo Liu, and Guoyin Wang, "An automatic method to determine the number of clusters using decision-theoretic rough set" International Journal of Approximate Reasoning,SS(I): IOI {IiS, 2014.

[11] KrisztianBuza, Gabor I Nagy, and AlexandrosNanopoulos,"Storage optimizing clustering algorithms for high-dimensional tick data" ExpertSystems with Applications, 2014.

[12] Shuliang WANG, Jinghu a FAN, Meng FANG, and Hanning YUAN, "Hg cudf: Hierarchical grid clustering using data field" Chinese Journal of Electronics, 23(I), 2014.

[13] Iftekhar Naim, Suprakash Dana, Jonathan Rebhahn, James S Cavenaugh, Tim R Mosmann, and Gaurav Sharma, " Swift scalable clustering for automated identification of rare clle populations in large, high-dimensional flow cytometry datasets, part I : Algorithmdesign" CytometryPartA, 2014.

[14] Younghoon Kim, et al., —DBCUREMR: An efficient density – Basedclustering

[15] A. Hinneburg, D. A. Keim, et al. "Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering", Proc. Very Large Data Bases (VLDB), pp. 506–517, 1999.

[16] Lawrence 0. Hall, NiteshChawla , Kevin W. Bowyer, "Decision Tree Learning on Very Large Data Sets", IEEE, Oct 1998.

[17] Thangaparvathi, B., Anandhavalli, D An improved algorithm of decision tree for classifying large data set based on rainforest framework, Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on Oct. 2010Page(s):800 – 805.

[18] D. L. A Araujo., H. S. Lopes, A. A. Freitas, "A parallel genetic algorithm for rule discovery in large databases" , Proc. IEEE Systems, Man and Cybernetics Conference, Volume 3, Tokyo, 940-945, 1999.

[19] Mr. D. V. Patil, Prof. Dr. R. S. Bichkar,"A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets", IEEE, 2006.

[20] Ros, F., Harba, R.; Pintore, M. Fast dual selection using genetic algorithms for large data sets, Intelligent Systems Design and Applications (ISDA), 12th International Conference on Date of Conference:27-29 Nov. 2012 Page(s):815 – 820, 2012.

[21] R. XU and D. Wunsch, "Survey of clustering algorithms," IEEE Trans. Neural Network, vol. 16, no. 3,p p. 64S_678, May 2005 .

[22] Giuseppe DeCandia, DenizHastorun, MadanJampani, GunavardhanKakulapati, AvinashLakshman, Alex Pilchin, SwaminathanSivasubramanian, Peter Vosshall and Werner Vogels"Dynamo: Amazon's Highly Available Key-value Store", Amazon.com.

[23] Rajkumar.D, Usha.S, "A Survey on Big Data Mining Platforms, Algorithms and Handling Techniques", International Journal for Research in Emerging Science and Technology, Volume-3, Special Issue-1, p.50-55, NCRTCT'16.

[24] Gianmarco De FrancisciMorales,"SAMOA: A Platform for Mining Big Data Streams".

Online References

[1] http://www.cubrid.org/blog/dev-platform/platforms-for-big-data/

[2] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-014-0008-6#Sec33

[3] http://www.ittoday.info/ITPerformanceImprovement/Articles/2014-07Raghupathi.html

[4] https://www.microsoft.com/en-us/research/video/big-data-platforms/

[5] http://www.ibmbigdatahub.com/blog/what-sparka