A REVIEW ON VARIOUS DATA-MINING TECHNIQUES FOR BIG DATA

MANISHA R. THAKARE¹, PROF. S. W. MOHOD²

2. Assistant Professor, Department of CSE, B.D. College of Engineering, Wardha, India

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: There exist large amounts of heterogeneous digital data. This phenomenon is called Big data which will be examined. The examination of Big data has been launched as Big data analytics. Big data is large volume, heterogeneous, distributed data. Big data applications where data collection has grown continuously, it is expensive to manage, capture or extract and process data using existing software tools. Fast retrieval of the relevant information from databases has always been a significant issue. Clustering is a main task of exploratory data analysis and data mining applications. Clustering is one of the data mining techniques for dividing dataset into groups. Clustering is a kind of unsupervised data mining technique.

Keywords: Data Mining, Clustering, Classification, Clustering Algorithms, Big Data, Map-Reduce.
INTRODUCTION

Big data is a largest buzz phrases in domain of IT, new technologies of personal communication driving the big data new trend and internet population grew day by day. The need of big data generated from the large companies like facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured form or even in structured form. Google contains the large amount of information. The need of Big data analytics which is stored in relational database systems in terms of five parameters-variety, volume, value, veracity and velocity.

- **Volume**: Data is ever-growing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.

- **Variety**: Data sources are extremely heterogeneous. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more.

- **Velocity**: The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organizations data velocity is main challenge.

- **Value**: Value is main buzz for big data because it is important for business, IT infrastructure system to store large amount of values in database. It is a most important v in big data.

- **Veracity**: The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data.

  - **Challenges In Big Data Mining**

The challenges at (Tier I) focus on low-level data accessing and arithmetic computing procedures. Challenges on information sharing and privacy. Big data often stored on different location and it is continuously growing. An effective computing platforms to take distributed large scale data storage into consideration for computing. (Tier II) concentrate on high-level semantics, application domain knowledge for different applications of Big data and the user privacy issues. This information provides benefits to Big data access but also add a technical barriers to Big data access (Tier I) and mining algorithms (Tier II). The outmost tier is (Tier III) which challenges the actual mining algorithms. (Tier III) contains three stages. In first stage
sparse, heterogeneous, uncertain, incomplete and multisource data is preprocessed by data fusion technique. In second stage after preprocessing stage complex and dynamic data are mined. Third stage is for local learning and model fusion, where the global owledge is obtained by local learning and model fusion is tested and the relevant information is feedback to preprocessing stage.

○ Data Mining Techniques

Data mining having many type of techniques like clustering, classification, neural network etc but in this paper we are consider only two techniques.

- Clustering

Clustering is the most significant task of data mining. It is an unsupervised method of machine learning application. In clustering the classes are divided according to class variable. Two important topics are: (1) Different ways to group a set of objects into a set cluster. (2) Types of clusters. The result of the cluster analysis is a number of heterogeneous groups with homogeneous contents. The first document or object of a cluster is defined as the initiator of that cluster. The initiator is called the cluster seed. The procedure of the cluster analysis with four basic steps are as follows:

- Feature Selection or extraction. Feature extraction utilizes some transformations to generate useful and novel features from the original ones. Feature selection chooses distinguishing features from a set of candidates. An elegant selection of features can greatly decrease the workload, and simplify the subsequent design process. Ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret.
• Clustering algorithm design or selection. Patterns are grouped according to whether they resemble one another. The construction of a clustering criterion function makes the partition of clusters an optimization problem. Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems in specific fields. Therefore, it is important to carefully investigate the characteristics of the problem on hand, in order to select or design an appropriate clustering strategy.

• Cluster validation. Different approaches usually lead to different clusters and even for the same algorithm, parameter identification or the presentation order of the input patterns may affect the final result. Therefore, effective evaluation standards and criteria are important to provide the users with a degree of confidence, for the clustering results derived from the used algorithms. Generally, there are three categories of testing criteria: external indices, internal indices, and relative indices. These are defined on three types of clustering structures, known as partitional clustering, hierarchical clustering, and individual clusters.

• Result interpretation. The ultimate goal of clustering is to provide users with meaningful insights into original data, so that they can effectively solve the problems encountered. Experts in the relevant fields interpret the data partition. It may be required to guarantee the reliability of the extracted knowledge.

• Classification

Classification is a simple process to finding a model that describes and distinguishes data classes of test. It is both types supervised learning and unsupervised. It consists of two steps:

1. Model construction: It consists of set of predefined classes. The set of tuple used for model construction is known as training set. These model can be represented as classification rules, decision trees.

2. Model usage: This model is used for defining future or unknown objects. It is used unsupervised learning rule.

• Literature survey

G.P and MARTY et.al in (1994) examines in the paper, how Clustering technique is useful to identify different information by considering various examples. By examining one or more attributes or classes, you can group individual pieces of data together to form a structure
opinion. Clustering is using one or more attributes for identifying a cluster of correlating results. [1].

Fayyad et.al in (1996) describes the various data mining techniques that allow extracting unknown relationships among the data items from large data collection that are useful for decision making. The wide-spread use of distributed information systems leads to the construction of large data collections in business, science and on the web. Businesses can learn from their transaction data more about the behavior of their customers and therefore can improve their business by exploiting this knowledge. Science can obtain from observational data (e.g. satellite data) new insights on research questions. Web usage information can be analyzed and exploited to optimize information access. Thus data mining generates novel, unsuspected interpretations of data. [2].

Areej Shhad, Gongde and Daniel Neagu in (2005) They discussed about the machine learning techniques for data mining problems. In this paper they applied different artificial intelligence techniques on prediction data sets. This paper specifies that multiple classification methods are valuable for any certain application. Particularly they mentioned that artificial neural network model has higher capability for solving data mining problems. [3].

Jyothi Bellary, Bhargavi Peyakunta, Sekhar Konetigari in (2010) have worked on Hybrid Machine Learning Approach In Data Mining. Machine learning methods as “symbolic and sub-symbolic”. Decision trees, case-based reasoning, instance-based learnings are under symbolic classification and neural network, fuzzy rules, and genetic algorithms are under sub-symbolic classification. In their study they proposed a hybrid approach, which is combination of both case-based reasoning (symbolic) and neural network (sub-symbolic) are applied them for data mining problems. [4].

Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C in (2010) they describe the application of k-mean clustering algorithm to provide the result of students academic performance. The main aim is to analysis performance by using k-mean implementation in clustering. In this paper they combined the k-mean model with the deterministic model to analyze the result. They simply compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance. They provide better result compare the earliest model of k-mean. [5].

Varun Kumar, Nisha Rathee in (2011) explained about knowledge discovery from database using an integration of clustering and classification. Authors experimented and equated the outcome of a simple classification method with the outcome of integrated clustering and classification method with the WEKA (Waikato Environment for Knowledge Analysis) data
mining tool. As a result they found, the integration of clustering and classification techniques gives more accurate and robust result. In their study, they have shown that clustering is an unsupervised learning method and it creates classes by partitioning number of clusters according to their instances. After clustering they have applied the classification method to assign the attributes to these clustered classes, because at the time of clustering decision rules are obtained, which are very useful in classification. [6].

McKinsey (2011) Big data is seen as “The next frontier for innovation, competition and productivity” and as such the related applications will contribute to economic growth. The positive impacts of big data provide a huge potential for organizations. In order to achieve these aspirations several issues should be analysed and discussed in the complex systems and using systems approaches such as holistic thinking and system dynamics. [7].

Chen et al. in (2012) have compared data base management systems and analytics as well as ETL with using MapReduce and Hadoop. Hadoop was originally a distributed file system approach applying the MapReduce framework. Recently, Hadoop has been developed and used as a complex ecosystem that includes a wider range of software systems. Therefore in the recent conceptual approaches Hadoop is primarily considered an ecosystem or an infrastructure or a framework and not just the file system alongside Map Reduce components. [8].

Patel et al. in (2012) have explored a practical solution to big data problem using the Hadoop data cluster, Hadoop Distributed File System alongside MapReduce framework, and a big data prototype application scenarios. The results obtained from various experiments indicate promising results to address big data problem. [9].

Wu Yuntian in (2012) provided a survey based on Machine Learning of Data Mining to Further Explore. This paper emphasizes on data gathered, use data information, development data information resources. SQL query is trying to retrieve the hidden and not apparent information. This query language is helpful when it is known which information need to retrieve, but data mining is useful when knowledge is not clear about the core part. Therefore both these applications are complementary for any data repository. [10].

Neelamadhab Padhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi in (2012) provided a survey on Data Mining Applications and feature scope. According to authors there are varieties of methods, approaches which are applied for data mining process in different fields. Each application have their own properties which are very important and useful for do research. In their research they focused on the scope and benefits of number of applications. The domain
specific methods are distinguished as special technique for data mining, because these techniques are intended for particular purpose and to produce the precise information. [11].

Aditya B. in (2012) defines big data problem using Hadoop and Map Reduce reports the experimental research on the Big data problems in various domains. It describes the optimal and efficient solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage data and Map Reduce framework for parallel processing to process massive data sets and records. [12].

Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta in (2012) in this paper presents the comparison of data mining algorithms for clustering. In this paper cover classification, clustering techniques. Data mining is used in every field for the analysis of large volumes of data. The k-means approach is use to predict the solution less sensitive to initialization and provides results at multiple resolutions, and k-means algorithm is also sensitive to the presence of outliers. KNN classification is an easy to understand and easy to implement classification technique. [13].

John A. Keane in (2013) proposed a framework in which big data applications can be developed. The framework consist of three stages (multiple data sources, data analysis and modeling, data organization and interpretation) and seven layers (visualization/presentation layer, service/query layer, modeling/statistical layer, processing layer, system layer, data layer/multi model) to provide big data application into blocks. The main motive of this paper is to manage and architect a massive amount of big data applications. The advantages of this paper is big data handles heterogeneous data and data sources in timely to get high performance and framework Bridge gap with business needs and technical realities. [14].

Xin Luna Dong in (2013) explained challenges of big data integration (schema mapping, record linkage and data fusion). These challenges are explained by using examples and techniques for data integration in addressing the new challenges raised by big data, includes volume and number of sources, velocity, variety and veracity. The advantage of this paper is identifying the data source problems to integrate existing data and system. The disadvantage of this paper is big data integration such as integration such as integrating data from markets, integrating crowd sourcing data, providing an exploration tool for data sources. [15].

Sagiroglu, S.; Sinanc, D. in (2013) provided a survey on “Big Data: A Review” describe the big data content, its scope, methods, samples, advantages and challenges of data. The critical issue about the Big data is the privacy and security. Big data samples describe the review about the atmosphere, biological science and research. By this paper, we can conclude that any
organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets. [16].

Keith C.C. Chan in (2013) authors describes large amount of structured and unstructured data collection, processing and analysis from hospitals, laboratories, pharmaceutical, companies or even social media and also discus about how to collect or analyse huge volume of data for drug discovery. The advantage of this paper is how big data analytics contributes to better drug safety efficacy for pharmaceutical regulators and companies. The disadvantage of this paper it needs the algorithms that are simple, scalable, efficient and effective for data discovery process. [17].

Madhuri V. Joseph, Lipsa Sadath and Vanaja Rajan in (2013) they represent comparision on various techniques in data mining. In data mining having various techniques which are used to exploring the important data from bulk amount of data and they can deal with different data type. In this paper they explain and compare some common techniques of data mining which are mainly used in our daily life and business environment. Every data mining techniques is an important role in the business environment according to its functionality so there is no anyone model which play all the roll in business environment. [18].

Aastha Joshi, Rajneet Kaur in (2013) they represent that clustering is a process in which we find the structure of unlabel data. In this paper they provide the reviews of six types of clustering techniques k-means clustering, Hierarchical Clustering, DBSCAN clustering, OPTICS, STING. The result of this comparison is that k-mean is better for large data set means it can increase the performance with increasing clusters. Hierarchical algorithm is useful in categorical data. Density based methods OPTICS, DBSCAN are designed to find clusters of arbitrary shape whereas partitioning and hierarchical methods are designed to find the spherical shaped clusters. [19].

Yaxiong Zhao in (2014) proposed data aware caching (Dache) framework that made minimum change to the original map reduce programming model to increment processing for big data applications using the map reduce model. It is a protocol, data aware cache description scheme and architecture. The advantage of this paper is, it improves the completion time of map reduce jobs. [20].

Wu et al. (2014) have introduce HACE theorem that described the key characteristics of the big data as huge with heterogeneous and diverse data sources, autonomous with distributed and decentralized control, and complex and evolving in data and knowledge associations. Big data a
collection of complex and large data sets that are difficult to process and mine for patterns and knowledge. [21].

- Future work

To tackle the Big data challenges and “seize the opportunities afforded by the new, data driven resolution”. This project seek to develop methods, algorithms, frameworks and research infrastructure that allow us to bring the massive amounts of data. Map reduce mechanisms suitable for large scale data mining by testing series of standards data mining tasks on cluster. Map reduce implementation mechanism evaluated the algorithm. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, the growth of data volumes and item relationships should help from legitimate patterns.

IV. CONCLUSION

Big data framework needs to consider complex relationships between samples, models and data sources along with their evolving changes with time and other possible factors. To support Big data mining high performance computing platforms are required. With Big data technologies we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at realtime.

V. REFERENCES


