Analysis of Big Data Processing Using Machine Learning Algorithms

Ruqaiya[1], Dr. Ihtiram Raza Khan[2]
[1](Department of Computer Science, Jamia Hamdard, New Delhi, India
Ruqaiya786.279@gmail.com)
[2] (Assistant Professor, Department of Computer Science, Jamia Hamdard, New Delhi, India
Ehtiram2007@gmail.com)

ABSTRACT
Big data refers to size of data produced not only in terabytes but in Exabyte or even beyond this. The data being produced is structured, unstructured or even semi-structured. It has become difficult to find meaningful patterns hidden in these various data sets. The data is being produced at lightning speed. In this review paper, various machine learning algorithm have been reviewed for Big Bata processing.

Keywords– *Machine learning, Big Data.*

I. INTRODUCTION

As the name says ”Big Data” means the data is uncountable and is so massive that it can go beyond the size of hundreds of terabytes. Big data accounts structured as well as unstructured data. Big data has its roots in all the fields like science, engineering and technology. The size of big data is increasing rapidly at each passing second. Traditional database system has no capacity to handle such huge data. The world is creating 2.8 quintillion of data per day from unstructured data sources like social media platform like Facebook, Instagram, gmail, hike, photos, files, etc. Big data forms the three level structures of data produced that is structured data, unstructured data and semi-structured data.

In unstructured data, the elements within the data have no defined structure. The Big data Commission at the Tech American Foundation offers the following definition: “Big data is a term that describes large volume of high-velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.” Big Data is basically collection of heterogeneous data from various pool of data resources. These data carry hidden pattern which can give fruitful results if recognized. Various Machine learning techniques have given optimized results. Pattern recognition, K-means algorithm, clustering, line regression, Bayesian algorithm, Decision tree, neural network are the machine learning algorithm. The vital key features which define characteristics of Big data are as follows:-

DATA VOLUME: It is the amount of data that is available in terabytes or more. The data comes from social sites, media, post etc. With the ever increasing amount of data volume, it needs to be stored and analyzed and has to come up with optimal solution to solve its storage problem.
DATA VELOCITY: It is the amount of data that is produced at great pace. Data is increasing at every second with lightning speed and this makes time an important factor in several organization.
DATA VALUE: Big data comprises of heterogeneous collection of data and each data has value associated with it which varies economically. Finding hidden patterns and meaningful data from the bag of wide-ranging data and transforming it for analysis is a head to toe task.
DATA VARIETY: It is the heterogeneous trait of data that makes it complex, variable data coming from internet, e-mails, videos etc.

Machine Learning is a unique approach that yields answers to the problem of Big Data. Machine Learning is a field of computer science that is associated with Artificial intelligence, science, engineering and technology, bio-medical etc. Machine learning finds its root in data prediction and pattern evaluation. Various Machine Learning techniques stand in solving various problems of big data. Traditional machine learning systems were designed such that all the data would be loaded at one time and it would be processed after that. With the increasing amount of data at great speed, it has become difficult to use traditional data base systems.

SUPERVISED LEARNING: As the name suggest, it is supervised under a guider means trained data sets are readily available which consists of set of data inputs and desired outputs. Given rules are applied to the data sets to get outputs.
UNSUPERVISED LEARNING: Stands opposite to supervised learning, no trained data sets are available here. The outputs are totally based on prediction analysis. It helps to explain hidden structure from untrained data sets.
REINFORCEMENT LEARNING: It is based on the feedback provided by the environment. It is a decision making approach. The state of a game cannot be decided it is completed and the feedback is obtained in the form of won or lost.

Advance learning method are required that would handle unprecedented amount of data. It is classified into three main categories:

II. LITERATURE REVIEW

Steve Oberlin, et.al (2012) proposed various Machine Learning strategies for the Big Data processing. He applied Machine Learning and various techniques from Artificial Intelligence to the complex and powerful data sets. Recommendation engines used by Netflix to see the rating and preferences of audience are one of the applications of Machine Learning. Informatics and Data Mining in which IBM's "Watson" uses different Machine Learning approach to process and depict human language and answer the queries[1]. Linear regression, massaging the data, Perception, k- means are the few strategies used by him for uncovering the relationships and finding patterns in data. The choice of Machine Learning algorithm basically depends on the nature of prediction. The prediction can be estimate type or classification. He also discussed how increasing features can make the algorithm complex and increasing computational requirements.

Jainender singh, et.al (2014) proposed machine learning technique that would be providing promising results to security issues faced in applications, its technologies and theories. He emphasized on mining from sparse, incomplete and uncertain data that would give optimized results when hidden patterns are discovered from the data sets using machine learning algorithms like Support Vector Machine (SVM), Naïve Bays classifiers, clustering techniques which are used to create supervised learning[4]. It would give insight knowledge in health, education, trade and many more fields.

Yasir Safeer, et.al (2010) presented Machine learning Algorithm i.e. k-means clustering for finding a document from a vast collection of unstructured text documents. He proposed a technique to portray documents that would be improving clustering result[3]. He discussed about the stream of document clustering, implemented k-means and devised an algorithm for better representation of documents and proposed how systematic domain dictionary would be used to get better similarity results of documents.

T. Nelson Gnanaraj et.al (2014) presented a study of k-means clustering applied on structured as well as unstructured data. He overcomes the problem of applying k-means algorithm on both the structured as well unstructured data set together. Based on computational value he described and implemented k-means[2]. He also proposed CURE and BIRCH as clustering algorithm for extraction of knowledge from non-spherical shapes and various sizes of data. In CURE, the data samples are partitioned and then are partially clustered. The partitioned data sets are again clustered to get the desired output. BIRCH uses metric algorithm to reduce I/O cost and providing parallelism and dynamic performance based on knowledge about data sets.

Junfei Qiu, et.al (2017) proposed some of the latest advances of Machine Learning for processing Big Data. Representation Learning, a new advanced learning method in which data representation is useful and meaningful by extracting helpful information while constructing classifiers and predictors. It aims to capture vast input which would give computation as well as statistical efficiency. Feature selection, Feature extraction and Metric learning are the subtopic of Representation learning. Active learning is another advanced Machine learning method applied for big data processing like biological DNA identification, image classification. It is a case of semi-supervised Machine learning in which it query the users to get desired output from subset of critical labeled instances availablethus minimizing the cost and giving higher accuracy and optimized results. He also discussed about the challenges and issues of Machine learning for Big Data processing. Heterogeneous nature of data, data produced at lightning speed, uncertainty and incomplete data, its vastness are some of the major concerns about Big Data. He also gave remedies for the same. Alternating direction

methods of multipliers(ADMM) is a promising method for parallel and distributed large scale data processing. It splits the multiple variables in an efficient way thus helping to find solution to a large scale of data. For handling high speed of data, Extreme Learning Method (ELM) has been introduced to provide faster learning speed, great performance and with less human interference.

Alexandra L'Heureux, et.al (2017)presentednew ways of processing Big Data through Machine Learning Algorithms. Due to Big Data characteristics, traditional tools are now not capable of handling its storage, transport or its efficiency. Machine Learning is regarded as a fundamental component of Data Analytics as it has power to learn from data and provides data driven insights, prediction and decision. The tremendous increase in size, space and time complexity of Support Vector Machine(SVM) would affect both the complexities thus making computational efficiency infeasible. Curse of Modularity in which increase in size of data leads to collapse of the given boundary of algorithm is solved by Map Reduce [6]. It is a programmable and scalable paradigm used for processing large data sets on various nodes byfollowing parallelism. It follows iterative approach. K-means can also be used to overcome shortcoming of Curse of modularity. Online Learning, one of the Machine Learning paradigms that would bridge the efficiency gaps produced by Big Data. It helps in processing large amount of data solution. Due to its adaptive nature, it is able to handle dirty and noisy data.
Roheet Bhatnagar, et.al (2018) presented about role of Machine Learning and Big Data Processing and Analytics (BDA). The development of Machine Learning and Big Data Analytics is complementary to each other. He discussed various future tends of Machine learning for Big data. Data Meaning implies how Machine Learning can be made more intelligent to acquire text or data awareness [5]. Technique Integration, another trend used to integrate data and process it. Classification, regression, cluster analysis are some of the techniques of machine Learning which are used to perform analytics and predict future from existing patterns find correlation among the given data sets.

III. CONCLUSION

Big data has found its roots in all streams including science and engineering. This review paper tries to accumulate all the new Machine Learning Algorithms which are used to process Big Data. Traditional algorithms are not capable of processing the data as data has become vast, heterogeneous in nature, value based and the speed at which it is produced is beyond the reach of these traditional algorithms. Various new paradigms have been discussed in this review paper which would find hidden patterns and predict the knowledge given by the data sets available. K-means clustering, line regression, Active learning and Support Vector Machine(SVM) are few of the Machine Learning tools which have given promising results during the processing of Big Data. As the features of Big Data increases the complexity to find hidden and meaningful patterns also increases. The algorithm designed for the above purpose has to be flexible and should be least human dependent as dependencies increases the chances of failure. Choosing the best suited algorithm for data prediction would give fruitful results.

REFERENCES

[1]Steve Oberlin, "Machine Learning, Cognition, and Big Data," 2012.

[2] T. Nelson Gnanaraj, Dr. K. Ramesh Kumar and N. Monica, "Survey on miming clusters using K-means algorithm from structured and unstructured data," *International Journal of Advances in Computer* Science and Technology, 2014.

[3] Yasir Safeer, Atika Mustafa and Anis Noor Ali, "Clustering Unstructured Data (Flat Files)," *International Journal of Computer Science and Information Security*, 2010.

[4] Jainendra Singh, "Big Data Analytics and Mining with Machine Learning Algorithm," *International Journal of Information and Computation technology*, 2014.

[5]Roheet Bhatnagar, "Machine Learning and Big Data Processing :A Technological Perspective and Review", 2018.

[6]Alexandra L'Heureux, Katarina Grolinger, Hany F. Elyamany, Miriam A.M. Capretz "Machine Learning With Big Data: Challenges And Approaches" *IEEE Access, 2017.*

[7]Alexandra L'Heureux, Katarina Grolinger, Hany F. Elyamany, Miriam A.M. Capretz "A survey of Machine

Learning for Big Data Processing" *EURASIP Journal on Advances in Signal Processing, Springer, 2016.*

[8]Long Xu Yihua Yan "Machine Learning for Astronomical Big Data Processing: Challenges And Approaches" Visual Communication And Image Processing, *IEEE, 2017.*

[9] https://en.wikipedia.org/wiki/Machine_learning.

[10] https://en.wikipedia.org/wiki/Big_data.

[11] Lidong Wang, Cheryl Ann Alexander "Machine Learning In Big Data*", International Journal of Mathematical, Engineering and Management Sciences, 2016.*

[12]Zaharaddeen Karami Lawal, Rufai Yusuf Zakari, Mansur Zakariyya Shuaibu, Alhassan Bala "A review: Issues and Challenges in Big Data from Analytic and Storage perspectives", *International Journal Of Engineering And Computer Science, 2016.*

[13] "Unstructured data: A big deal in big data," Game changing Technology to meet agency missions by sponsored reports FCW.

[14] K.V.Kanimozhi1, Dr.M.Venkatesan, "Unstructured Data Analysis-A Survey," *International Journal of Advanced Research in Computer and Communication Engineering, 2016*