A Study on Machine Learning Algorithms for Big Data Analytics

Mrs. T.Kavipriya[1] and Mr.N.Kumar[2]
[1]Assistant Professor, Dept. of Computer Technology, Hindusthan College of Arts and Science,Coimbatore
[2]Assistant Professor, Dept. of Computer Science, Dr.N.G.P. Arts and Science College, Coimbatore

Abstract: *- A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. Machine learning is the essence of artificial intelligence. Machine Learning learns from past experiences to improve the performances of intelligent programs. Machine learning system builds the learning model that effectively "learns" how to estimate from training data of given example. In this new era, Machine learning is mostly in use to demonstrate the promise of producing consistently accurate estimates. The basic objective of this paper is to explore the potential impact of big data challenges, and also machine learning algorithms for Big data analysis. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically.*

Keywords: *-Ada Boost, Big Data, Classification, Hadoop, Machine Learning,*

I. Introduction

A collection of large and complex data sets which are difficult to process using common database management toolsor traditional data processing applications. Big data is not just about size. Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data. These are available in structured, semi-structured, and unstructured format in petabytes and beyond. Formally, it is defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi structured etc. The fourth V refers to veracity that includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques [1]. Some of these extraction methods for obtaining helpful information was discussed by Gandomi and Haider [2]. The following Figure 1 refers to the definition of big data. However exact definition for big data is not defined and there is a believe that it is problem specific. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective.

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

The processes involved in machine learning are similar to that of data mining and predictive modeling. Both require searching through data to look for patterns and adjusting program actions accordingly. Nowadays people are familiar with machine learning from shopping on the internet and being served ads related to their purchase. This happens because recommendation engines use machine learning to personalize online ad delivery in almost real time. Beyond personalized marketing, other common machine learning use cases include fraud detection, spam filtering, network security threat detection, predictive maintenance and building news feeds.

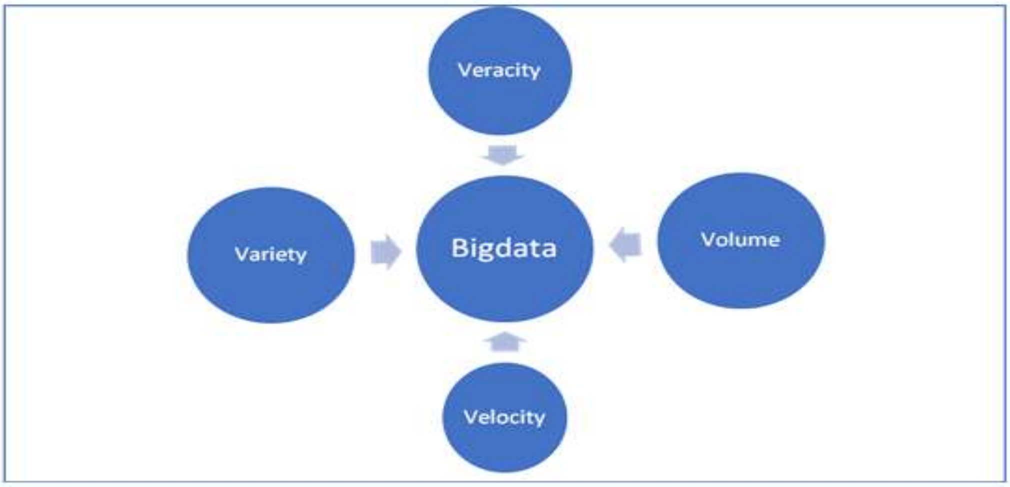

Figure.1. Big Data Characteristics

As you know Big Data involves working with huge chunks of both structured and unstructured data. The volume of data that data scientists have to work on sometimes exceeds over millions of rows and it becomes too tedious to prepare for the work, albeit doing it. That is when these technologies become interdisciplinary. With machine learning and artificial intelligence, a data scientist can make his or her work of process Big Data easily. Considering the volume of data sets, software models and conventional databases turn out to be less effective. This is exactly when machine learning can be applied to Big Data.

This paper will focus on challenges in Big data analytics and few machine learning algorithms that support for Big Data processing. The following paper divide into three section, section II elaborate challenges in big data processing, Section III will explore few best machine learning algorithms and finally section IV concluded this paper.

II. Challenges In Big Data Processing.

Big data is new kind of precious item today. According to Bernard Marr in his article in forbes website said that , the amount of data created today is astonishing. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated. In this paper we identified few data generating sources.

Bigdata Generation.

Every day, fifty percent of web searches carried through Mobile devices.On an average google process 40000 searches per second and 70 percent of searches carries out using google. Modern human new love affair is a social media nearly 4,146,600 YouTube videos, closing 456,000 tweets are sent on Twitter. In Instagram users post 46,740 photos. According to Facebook, there are nearly 2 billion active users on Facebook along. Not only in social media, most of the communication are carried on internet. On an average nearly 16 million text message using internet In Various shops are mall nearly 9,90,000 swipes are done using debit/credit cards. In every min there are approximately 100 billons of emails were sent. Using skype, more than one million calls were made.At a nutshell Bigdata are created from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals, to name a few.

The above stats show that, every single individual generated data each day. Now our world becomes digitalized and data colossal.

Bigdata storage and processing.

A definition, given by McKinsey Global Institute (MGI) [1]: “Big Data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.It is clearly understandingthat Big Data has the potential to become a driving force for innovation and value creation.

Big Data analytics, is a might be the task for generating value from it. According to MGI, the “value” that can be derived by analyzing Big Data are as follows:

- Creating transparencies;
- Discovering needs, exposing variability, and improving performance;

- Segmenting customers; and
- Replacing are supporting human decision-making with automated algorithms. Innovating new business models, products, and services.

There is no single set formula for extracting value from Big Data; it will depend on the application. There are many applications where simply being able to comb through large volumes of complex data from multiple sources via Big Data Computing interactive queries can give organizations new insights about their products, customers, services, etc. Being able to combine these interactive data explorations with some analytics and visualization can produce new insights that would otherwise be hidden. They key limitations in Big Data are

- Shortage of talent necessary for organizations to take advantage of Big Data.
- Shortage of knowledge in statistics, machine learning, and data mining.

In Big Data processing has certain issues in the following areas:

- Access to data
- Data policies
- Industry structure
- Technology and techniques

From the technology perspective, Big Data challenges classified into three dimensions are as follows

1. Data
2. Processes
3. Management

We are well aware one the main characteristics of Big Data is its Volume of data, mostly are machine generated. Here the challenge is how to deal with its size. Next most challenges are its Variety, data are unstructured or complex structured that is hard to store in columns and rows. Data come from sensors, smart devices, and social collaboration technologies.Data are not only structured, but raw, semi-structured,unstructured data from web pages, web log files (click stream data),search indexes, e-mails, documents, sensor data, etc. The challenges here is its types, data sources and formats.

The Velocity of data generation is another key challenge, applications need to reach on time. Another key aspect is veracity it means data availability and its quality. There are several challenges which includes Data uncertainty, imprecision, missing values and missing truths.

Data Discovery: In this huge collection of data, how do we find high-quality of data.
Personal Identifiable Information: Today majority of information are about people, here how do we extract data without compromise personal privacy.

Process Challenges

The following are some of the key Big Data process challenges.

- Capturing Data
- Aligning data from different sources.
- Data Modeling, either or some form of simulation
- Understanding the output, visualizing and sharing the results, on various devices.
- Data transformation, Transforms data into a suitable form for analysis.

The main management challenges are: - Data privacy, Security, Governance and Ethical.

Big Data Platforms: State of Art.

Big Data handling technology are still is in immature state, at the same time clear shortage analytical experience for new data. In the Big Data era the old conventional method, moving data to the application program would not work anymore. Instead, the application logic should come closure to data. Transporting petabytes of data is highly impossible using current technologies.

In order to analyze big data, the current technology is using parallel databases or NoSQL data store with Hadoop. Hadoop is used for processing the unstructured Big Data. Today is widely used platform for Big Data Processing.

Hadoop is a new open source platform to analyze and process BigData. It was inspired byGoogle's MapReduce and Google File System (GFS) papers.

Hadoop Stack

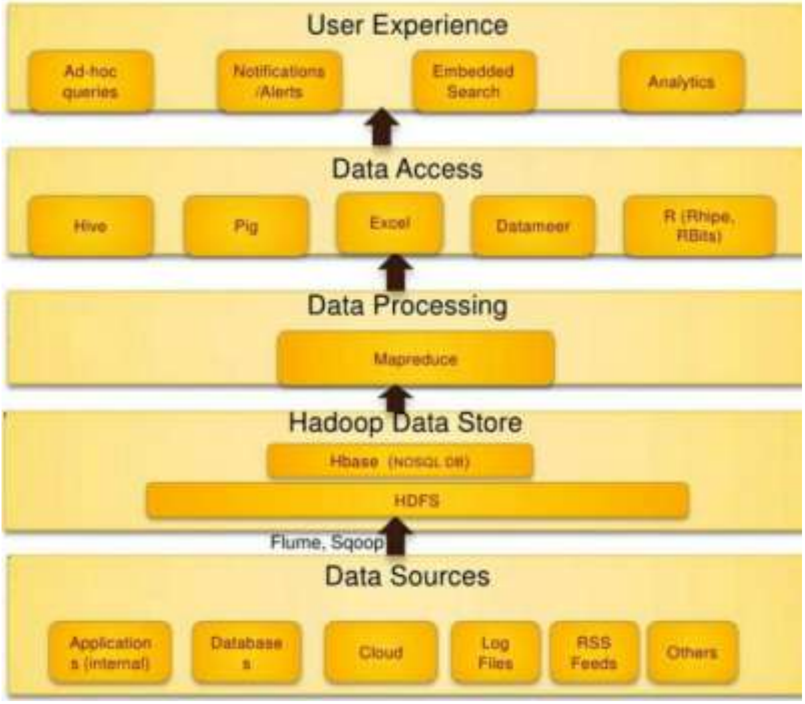

Figure 2. Hadoop Stack

The Hadoop stack includes more than a dozen components, or subprojects, that are complex to deploy and manage.Installation, configuration and production deployment at scale is challenging.

The main components include:

- Hadoop. Java software framework to support data-intensive distributed applications
- ZooKeeper. A highly reliable distributed coordination system
- MapReduce. A flexible parallel data processing framework for large data sets
- HDFS. Hadoop Distributed File System
- Oozie. A MapReduce job scheduler
- HBase. Key-value database
- Hive. A high-level language built on top of MapReduce for analyzing large data sets
- Pig. Enables the analysis of large data sets using Pig Latin. Pig Latinis a high-level language compiled into MapReduce for parallel data processing.

The Big Data technologies are largely contributed by some big web search companies includes:- Yahoo!, Facebook, GoogleBigTable, Amazon Dynamo

Hadoop Pros

- Open source.
- Ability to schedule very large jobs in smaller chunks.
- Support for replication and machine fail-over without operationintervention.
- The combination of scale, ability to process unstructured data alongwith the availability of machinelearning algorithms, and recommendationengines create the opportunity to build new game changingapplications.
- Does not require a schema first.
- Provides a great tool for exploratory analysis of the data,

Hadoop Cons

- Hadoop is difficult to use.
- Can give powerful analysis, but it is fundamentally a batch-orientedparadigm.
- Hadoop file system (HDS) has a centralized metadata store(NameNode), which represents a single point of failure withoutavailability. When the NameNode is recovered, it can take a longtime to get the Hadoop cluster running again.

- Hadoop assumes that the workload it runs will belong running, soit makes heavy use of checkpointing at intermediate stages. Thismeans parts of a job can fail, be restarted, and eventually completesuccessfully—there are no transactional guarantees.

III. Machine Learning Algorithms

From the above section II, we understand that, Big Data involves working with huge chunks of both structured and unstructured data. The volume of data, exceeds over millions of rows and it becomes too tedious. With machine learning algorithms we can easily process big Data. Considering the volume of data sets, software models and conventional databases turn out to be less effective. This is exactly when machine learning can be applied to Big Data.

Machine learning is an integral part of Artificial Intelligence. There are three types of algorithms in machine learning that can be used for Big Data classification – Supervised, semi-supervised and unsupervised.

Supervised Learning

Supervised learning is the task of inferring a function from labeled training data. By fitting to the labeled training set, we want to find the most optimal model parameters to predict unknown labels on other objects (test set). If the label is a real number, we call the task regression. If the label is from the limited number of values, where these values are unordered, then it's classification.

Classification and regression are two classifications of supervised learning. Classification is when the class attribute of a set is discrete and regression is when it is continuous. Without getting too technical, let us simply understand that some of the classification methods include

- Decision Tree Learning
- k-Nearest Neighbour
- Naïve Bayes Classifier

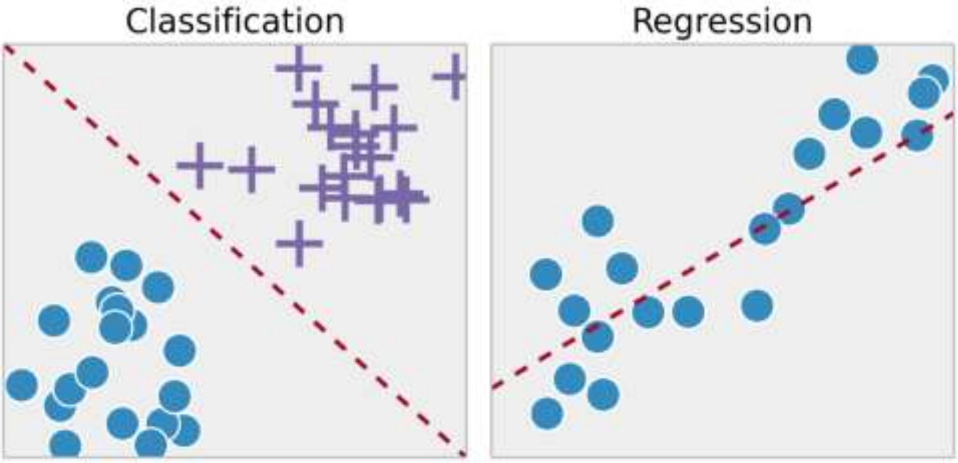

Figure 3: Classification and Regression

When it comes to regression techniques, they include linear and logistic regression techniques.
As far as supervised learning algorithms, some of the most commonly used ones include –

- Maximum Entropy Method (MaxENT)
- Support Vector Machines (SVM)
- Naïve Bayes
- Boosting Algorithm

Maximum Entropy Method (MaxENT)

Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information.That is to say, when characterizing some unknown events with a statistical model, we should always choose the one that has Maximum Entropy.

Maximum Entropy Modeling has been successfully applied to Computer Vision, Spatial Physics, Natural Language Processing and many other fields.

Support Vector Machines (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N—the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

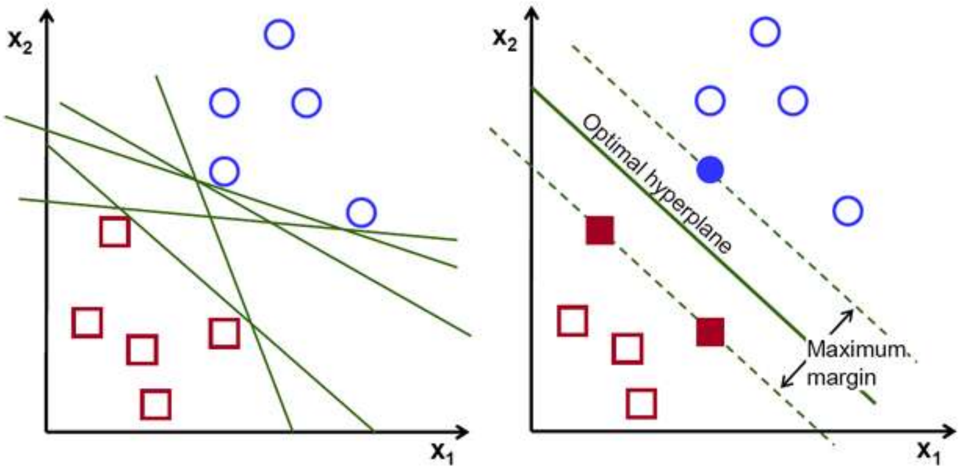

Figure 4: Hyper Plane and SVM

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

Naïve Bayes

Naive Bayes is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.

Boosting Algorithm: AdaBoost

AdaBoost, short for “Adaptive Boosting”, is the first practical boosting algorithm proposed by Freund and Schapire in 1996. It focuses on classification problems and aims to convert a set of weak classifiers into a strong one. The final equation for classification can be represented as

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)),$$

where f_m stands for the m_th weak classifier and theta_m is the corresponding weight. It is exactly the weighted combination of M weak classifiers.

Unsupervised Learning

In unsupervised learning, the algorithms take unlabeled data and classify it by drawing a comparison among data features.

- Clustering – which can be again classified into hierarchal, density-based, K-means
- Self-organizing Maps

Clustering can be further classified into three categories– supervised clustering, unsupervised clustering and semi-supervised clustering.

Supervised Clustering

Supervised clustering works on identifying clusters with high-probability densities with respect to individual classes. Supervised clustering works best when there are target variables and training sets that include the variables to cluster.

Unsupervised Clustering

Unsupervised clustering reduces the intercluster similarity and increases intracluster similarity. It works on a very specific object function and that is why hierarchal and k-Means are two of the most popular clustering techniques in unsupervised learning.

Semi-supervised Clustering

This class of clustering makes use of adjusting or guiding domain information in order to improve clustering. This guiding or adjusting domain information could be pairwise constraints prevalent between the target or observation variables for some observations.

IV. Conculsion

The Big Data, Big Data Analytics and Machine Learning techniques are emerging interleaved technologies for Big Data processing. This paper deeply expelling the Big Data Storage, Processing challenges and techniques. Some portion of this paper also provide insights into Hadoop platform and Hadoop stack. Section III provides some basic features of Machine learning algorithms. We strongly believe that in near future researcher will use these technologies to solve problems using Big Data and Machine learning.

References

[1]. M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in bigdata analytics, International Journal of Application or Innovation inEngineering & Management, 2(8) (2015), pp.228-232.

[2]. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods,and analytics, International Journal ofInformation Management,35(2) (2015), pp.137-144.

[3]. McKinsey Global Institute (MGI), Big Data: The next frontier for innovation, competition, and productivity, Report, June, 2012.

[4]. X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challengesof big data research, Big Data Research, 2(2) (2015), pp.59-64.

[5]. D. P. Acharjya and Kauser Ahmed P, International Journal of Advanced Computer Science and Applications, 7(2) (2016).

[6]. Osisanwo F.Y. et al, Supervised Machine Learning Algorithms: Classification and Comparison, International Journal of Computer Trends and Technology (IJCTT) – (48) (2017)

[7]. Pradeep, K. R. & Naveen, N. C. (2017). A Collective Study of Machine Learning (ML)Algorithms with Big Data Analytics (BDA) for Healthcare Analytics (HcA). International Journal of Computer Trends and Technology, 47(3),(2017).