Machine Learning in Big Data Analytics

Saikat Das[1]
B.Tech (Information Technology)
Narula Institute of Technology
Kolkata, India

Sampita Mallick[2]
B.Tech (Information Technology)
Narula Institute of Technology
Kolkata, India

Shyamapriya Chatterjee[3]
B.Tech (Information Technology)
Narula Institute of Technology
Kolkata, India

Sujata Kundu[4]
B.Tech (Information Technology)
Narula Institute of Technology
Kolkata, India

***Abstract*—The term "Big Data" refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The concept of big data gain momentum in the early 2000's due the enormous use of data. Earlier 3Vs are use in big data but now the concept of 5Vs are used which are 'VOLUME', 'VELOCITY', 'VARIETY', 'VERACITY', 'VALUE'. While the potential of these massive data is undoubtedly significant. Machine learning is the artificial intelligence method of discovering knowledge for making intelligent decisions. This paper introduces methods in machine learning, main technologies in big data (case studies) and some application of machine learning in big data.**

Keywords—Machine Learning, Big Data, 5Vs

I. INTRODUCTION:

The term "Machine Learning" was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence. A representative book of the machine learning research during the 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973. In 1981 a report was given on using teaching strategies so that a neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal. Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. The term "Big Data" has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

BIG DATA PHILOSOPHY ENCOMPASSES UNSTRUCTURED, SEMI-STRUCTURED AND STRUCTURED DATA, HOWEVER THE MAIN FOCUS IS ON UNSTRUCTURED DATA.

A. Brief Analysis of Big Data

The term Big Data refers to a huge volume of data that cannot be stored processed by any traditional data storage or processing units. Data is generated at a very large scale and it is being used by many multinational companies to process and analyze in order to uncover insights and improve the business of many organizations.

I. Big Data is generally categorized into three different varieties. They are as shown below:

a) Structured Data: Structured data is data that adheres to a pre-defined data model and is therefore straightforward to analyze. Structured data conforms to a tabular format with relationship between the different rows and columns. Common examples of structured data are Excel files or SQL databases. Each of these have structured rows and columns that can be sorted. Structured data depends on the existence of a data model – a model of how data can be stored, processed and accessed. Because of a data model, each field is discrete and can be accesses separately or jointly along with data from other fields. This makes structured data extremely powerful: it is possible to quickly aggregate data from various locations in the database.2

b) Semi- Structured Data: Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contain tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure. Examples of semi-structured data include JSON and XML are forms of semi-structured data. The reason that this third category exists (between structured and unstructured data) is because semi-structured data is considerably easier to analyze than unstructured data. Many Big Data solutions and tools have the ability to 'read' and process either JSON or XML. This reduces the complexity to analyze structured data, compared to unstructured data. Example- Comma Separated Values(CSV) File.

c) Unstructured Data: Unstructured data is information that either does not have a predefined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structured databases. Common examples of unstructured data include audio, video files or No-SQL databases. Example- Audio Files, Images etc

B. 5Vs of Big Data:

1. Volume:

• The name 'Big Data' itself is related to a size which is enormous.

• Volume is a huge amount of data.

• To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'. This means whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.

• Hence while dealing with Big Data it is necessary to consider a characteristic 'Volume'.

• Example: In the year 2016, the estimated global mobile traffic was 6.2 Exabytes (6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 Exabytes of data.

2. Velocity:

•Velocity refers to the high speed of accumulation of data. • In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.

• There is a massive and continuous flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.

• Sampling data can help in dealing with the issue like 'velocity'.

• Example: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22% (Approx.) year by year.

3. Variety:

• It refers to nature of data that is structured, semi-structured and unstructured data.

• It also refers to heterogeneous sources.

• Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured, semi-structured and unstructured.

• Structured data: This data is basically an organized data. It generally refers to data that has defined the length and format of data.

• Semi- Structured data: This data is basically a semi-organized data. It is generally a form of data that do not conform to the formal structure of data. Log files are the examples of this type of data.

• Unstructured data: This data basically refers to unorganized data. It generally refers to data that doesn't fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.

4. Veracity:

• It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.

• Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.

• Example: Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

5. Value:

• After having the 4 V's into account there comes one more V which stands for Value! The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.

• Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V's

II. BRIEF ANALYSIS OF MACHINE LEARNING (ML):

A. History:

• 1642 - Blaise Pascal invents a mechanical machine that can add, subtract, multiply and divide.

• 1679 - Gottfried Wilhelm Leibniz devises the system of binary code.

• 1834 - Charles Babbage conceives the idea for a general all-purpose device that could be programmed with punched cards.

• 1842 - Ada Lovelace describes a sequence of operations for solving mathematical problems using Charles Babbage's theoretical punch-card machine and becomes the first programmer.

• 1847 - George Boole creates Boolean logic, a form of algebra in which all values can be reduced to the binary values of true or false.

• 1936 - English logician and cryptanalyst Alan Turing proposes a universal machine that could decipher and execute a set of instructions. His published proof is considered the basis of computer science.

. • 1952 - Arthur Samuel creates a program to help an IBM computer get better at checkers the more it plays.

• 1959 - MADALINE becomes the first artificial neural network applied to a real-world problem: removing echoes from phone lines.

• 1985 - Terry Sejnowski and Charles Rosenberg's artificial neural network taught itself how to correctly pronounce 20,000 words in one week.

• 1997 - IBM's Deep Blue beat chess grandmaster Garry Kasparov.

• 1999 - A CAD prototype intelligent workstation reviewed 22,000 mammograms and detected cancer 52% more accurately than radiologists did.

• 2006 - Computer scientist Geoffrey Hinton invents the term deep learning to describe neural net research.

• 2012 - An unsupervised neural network created by Google learned to recognize cats in YouTube videos with 74.8% accuracy.

• 2014 - A chatbot passes the Turing Test by convincing 33% of human judges that it was a Ukrainian teen named Eugene Goostman.

• 2014 - Google's AlphaGo defeats the human champion in Go, the most difficult board game in the world.

- 2016 - LipNet, DeepMind's artificial-intelligence system, identifies lip-read words in video with an accuracy of 93.4%.
- 2019 - Amazon controls 70% of the market share for virtual assistants in the U.S

B. Types of Machine Learning:

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches:

- Supervised Learning
- Unsupervised Learning
- Semi- Supervised Learning
- Reinforcement Learning

C. What is it really?

Machine Learning is a subfield of Artificial Intelligence which evolved from Pattern Recognition and Computational Learning theory. Arthur Lee Samuel defines Machine Learning as: Field of study that gives computers the ability to learn without being explicitly programmed. So, basically, the field of Computer Science and Artificial intelligence that "learns" from data without human intervention. But this view has a flaw. As a result of this perception, whenever the word Machine Learning is thrown around, people usually think of "A.I." and "Neural Networks that can mimic Human brains (as of now, that is not possible)", Self 4 Driving Cars and what not. But Machine Learning is far beyond that. Below we uncover some expected and some generally not expected facets of Modern Computing where Machine Learning is in action.

D. The Expected Part From ML:

We'll start with some places where you might expect Machine Learning to play a part.

1. Speech Recognition (Natural Language Processing in more technical terms) – You talk to Cortana on Windows Devices. But how does it understand what you say? Along comes the field of Natural Language Processing, or N.L.P. It deals with the study of interactions between Machines and Humans, via Linguistics. Guess what is at the heart of NLP: Machine Learning Algorithms and Systems (Hidden Markov Models being one).

2. Computer Vision - Computer Vision is a subfield of AI which deals with a Machine's (probable) interpretation of the Real World. In other words, all Facial Recognition, Pattern Recognition, Character Recognition Techniques belong to Computer Vision. And Machine Learning once again, with it wide range of Algorithms, is at the heart of Computer Vision.

3. Google's Self Driving Car - Well. You can imagine what drives it actually. More Machine Learning goodness.

E. The Unexpected Part From ML:

Let's visit some places normal folks would not really associate easily with Machine Learning:

1. Amazon's Product Recommendations - Ever wondered how Amazon always has a recommendation that just tempts you to lighten your wallet. Well, that's a Machine Learning Algorithm(s) called "Recommender Systems" working in the backdrop. It learns every user's personal preference and makes recommendations according to that.

2. YouTube/Netflix - They work just as above!

3. Data Mining / Big Data - This might not be so much of a shock to many. But Data Mining and Big Data are just manifestations of studying and learning from data at a larger scale. And wherever there's the objective of extracting information from data, you'll find Machine Learning lurking nearby.

4. Stock Market/Housing Finance/Real Estate - All of these fields incorporate a lot of Machine Learning systems in order to better assess the market, namely "Regression Techniques", for things as mediocre as predicting the price of a House, to predicting and analyzing stock market trends.

III. BIG DATA MEETS MACHINE LEARNING:

Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y).

$$Y = f(X)$$

This is a general learning task where we would like to make predictions in the future (Y) given new examples of input variables (X). We don't know what the function (f) looks like or its form. If we did, we would use it directly and we would not need to learn it from data using machine learning algorithms. The most common type of machine learning is to learn the mapping $Y = f(X)$ to make predictions of Y for new X. This is called predictive modeling or predictive analytics and the main aim is to make the most accurate predictions possible. If we use big data for storing bulk amount of information and manipulation, is one thing but extracting useful information from these will possible through machine learning. With this machine learning we can extract efficient patterns. Machine-learning algorithms become more effective as the size of training datasets grows. So when combining big data with machine learning, we benefit twice: the algorithms help us keep up with the continuous influx of data, while the volume and variety of the same data feeds the algorithms and helps them grow. Let's look at how this integration process might work: By feeding big data to a machine-learning algorithm, we might expect to see defined and analyzed results, like hidden patterns and analytics that can assist in predictive modelling. For some companies, these algorithms might automate processes that were previously human - centered. But more often than not, company will review the algorithm's findings and search them for valuable insights that might guide business operations. Here's where people come back into the picture. While AI and data analytics run on computers that outperform humans by a vast margin, they lack certain decision-making abilities. Computers have yet to replicate many characteristics inherent to humans, such as critical thinking, intention and the ability to use holistic approaches. Without an expert to provide the right data, the value of algorithm-generated results diminishes, and without an expert to interpret its output, suggestions made by an algorithm may compromise company decisions.

IV. HOW TO APPLY MACHINE LEARNING IN BIG DATA?

Machine Learning provides efficient and automated tools for data gathering, analysis, and assimilation. In collaboration with cloud computing superiority, the machine learning ingests agility into processing and integrates large amounts of data regardless of its source.

Machine learning algorithms can be applied to every element of Big Data operation including:

- Data Segmentation
- Data Analytics
- Simulation

All these stages are integrated create the big picture out of Big Data with insights, patterns, which later get categorized and packaged into an understandable format. The fusion of Machine Learning and Big Data is a never-ending loop. The algorithms created for certain purposes are monitored and perfected over time as the information is coming into the system and out of the system.

V. CASE STUDY OF MACHINE LEARNING IN BIG DATA:

Big data at Walmart–

Walmart is the largest retailer in the world and the world’s largest company by revenue, with more than 2 million employees and 20000 stores in 28 countries. It started making use of big data analytics much before the word Big Data came into the picture. Walmart uses Data Mining to discover patterns that can be used to provide product recommendations to the user, based on which products were brought together. Walmart by applying effective Data Mining has increased its conversion rate of customers. It has been speeding along big data analysis to provide best-in-class e-commerce technologies with a motive to deliver superior customer experience. The main objective of holding big data at Walmart is to optimize the shopping experience of customers when they are in a Walmart store. Big data solutions at Walmart are developed with the intent of redesigning global websites and building innovative applications to customize the shopping experience for customers whilst increasing logistics efficiency. Hadoop and No-SQL technologies are used to provide internal customers with access to real-time data collected from different sources and centralized for effective use.

VI. MACHINE LEARNING APPLICATIONS FOR BIG DATA:

Today, machine learning is used in a wide range of applications. Perhaps one of the most well-known examples of machine learning in action is the recommendation engine that powers Facebook's News Feed. Facebook uses machine learning to personalize how each member's feed is delivered. If a member frequently stops to read a particular group's posts, the recommendation engine will start to show more of that group's activity earlier in the feed. Behind the scenes, the engine is attempting to reinforce known patterns in the member's online behavior. Should the member change patterns and fail to read posts from that group in the coming weeks, the News Feed will adjust accordingly. In addition to recommendation engines, other uses for machine learning include the following: Customer relationship management - CRM software can use machine learning models to analyze email and prompt sales team members to respond to the most important messages first. More advanced systems can even recommend potentially effective responses.

Business intelligence - BI and analytics vendor’s use machine learning in their software to identify potentially important data points, patterns of data points and anomalies.

Human resource information systems - HRIS systems can use machine learning models to filter through applications and identify the best candidates for an open position.

Self-driving cars - Machine learning algorithms can even make it possible for a semi-autonomous car to recognize a partially visible object and alert the driver.

Virtual assistants - Smart assistants typically combine supervised and unsupervised machine learning models to interpret natural speech and supply context.

VII. BIG DATA AND MACHINE LEARNING COMPARISON TABLE:

Basis For Comparison	Big Data	Machine Learning
Data Use:	Big data can be used for a variety of purposes, including financial research, collecting sales data etc.	Machine learning is the technology behind self-driving cars and advance recommendation engines.
Foundation For Learning:	Big data analytics pulls from existing information to look for emerging patterns that can help shape our decision-making processes.	On the other hand, Machine learning can learn from the existing data and provide the foundation required for a machine to teach itself.
Pattern Recognition:	Big data analytics can reveal some patterns through classifications and sequence analysis.	However, machine learning takes this concept a one step ahead by using the same algorithms that big data analytics uses to automatically learn from the collected data.
Data Volume:	Big data as the name suggest tends to be interested in large-scale datasets wherethe problem is dealing with the large volume of data.	ML tends to be more interested in small datasets where over-fitting is the problem
Purpose:	Purpose of big data is to store large volume of data and find out pattern indata	Purpose of machine learning is to learnfrom trained data and predicts or estimates future results.

IX. CONCLUSION:

ML (Machine Learning) is fundamental to address the difficulties postured by big data and reveal concealed patters, information, and bits of knowledge from enormous information keeping in mind the end goal to transform the capability of the last into genuine incentive for business basic leadership and logical investigation. Future scope of Machine learning analytics is how to make ML more declarative, so that it is easier for no experts to specify and interact with different type of data in different streams. In the future, we will enhance and assess the performance of machine learning techniques for different types of problems. One promising direction is to extend the machine learning approaches towards big data, which are efficient and highly scalable in the way they process high dimensional data.

REFERENCES

[1] https://www.edureka.co/blog/big-data-characteristics/
[2] https://www.geeksforgeeks.org/5-vs-of-big-data/
[3] https://www.educba.com/big-data-vs-machine-learning/
[4] https://searchenterpriseai.techtarget.com/definition/machine-learning-ML •https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/
[5] https://www.geeksforgeeks.org/demystifying-machine-learning/
[6] https://www.edureka.co/blog/machine-learning-and-big-data/
[7] https://towardsdatascience.com/machine-learning-and-big-data-real-world-applications-3ba3a3345cf5
[8] https://www.educba.com/big-data-vs-machine-learning/4