Role of Machine Learning In Harnessing Big Data

K S Spoorthi[1]
[1]*PG Scholar, dept of MCA*
Dayananda Sagar College Of Engineering(VTU)
Bangalore, Karnataka, India
spoorthiks606@gmail.com

Dr. Chandrika M[2]
[2]*Assistant Professor, dept.of MCA*
Dayananda Sagar College Of Engineering(VTU)
Bangalore, Karnataka, India
chandrika-mcavtu@dayanandasagar.edu

Abstract **- The important component of most of the companies or organizations is the data which has been produced in the vast amounts due to progress of digitalization. This data is referred to as big data. The production of large amounts of data gives opportunities and the obstacles for gaining or acquiring meaningful information to make appropriate choices. This big data can provide opportunities to develop various branches like machine learning, IOT etc. Machine learning can efficiently work using big data which would be beneficial for the organizations.**

This research paper is revolves around mainly the facts and keys components of machine learning , about big data and its characteristics as well the description of how the big data is being used by machine learning. Later on it focuses on the process of extracting the meaningful data from the huge set of data, its steps to refine the data and make it applicable. And then the list of applications of it such as recommendation systems that everyone uses daily.

Next the paper focuses on the benefits that the organizations will have after implementing the data into the applicable form. Then the challenges that will be encountered by the organizations while processing the data for implementation. Next the very crucial aspect is it focuseson future trends that can be implemented using machine learning.

Keywords – Machine learning, IOT, Big data, Recommendation System, Future trends.

I.INTRODUCTION

Machine learning along with the big data helps in the automation of all human activities.

A.CONCEPT OF BIG DATA

Data that is in the form of large amounts coming from various sources and is of different kinds is termed as big data [4]. It is the data which is from the organizations working daily, and is difficult to manage using the traditional or oldest processing systems. The data that is produced by them(organizations) have become the huge challenge to manage. The growing data is one of the best resource that the organization can possess but it is difficult to manage and get the value or the required data from it because most of the times, the data will be in it's raw format or in a unstructured or a semi-structured form. This makes organizations difficult to decide whether they can keep the data or not [5].

1) CLASSIFICATION OF BIG DATA

There are mainly three forms of big data that can be found those are:

a) Structured data
b) Unstructured data
c) Semi-structured data

a) *Structured data:* The structured data refers to the data in its fixed state, that is in the form which can be stored, accessed and processed. The advancement in technology of computer science has successfully used this form of data to derive the required value out of it besides the fact that format is known in advance.
b) *Unstructured data:* The unstructured data refers to the form of data whose format is unknown and voluminous. This kind of data poses the difficulties to the organizations while deriving the value out of it. This form of data is the mixture of text, images as well as videos known to be as heterogeneous data. Even though there is a rich data, it has become a huge task for organizations to derive the value from it as it is raw or unstructured.
c) *Semi-structured data:* The semi-structured data refers to as the data which is the combination or the mixture of both structured and unstructured data. The data in semi-structured format may not be in a table format or not defined but it might be stored like xml file.[5]

2) CHARACTERISTICS OF BIG DATA

The big data which is produced by organization can have various characteristics which is visible. Characteristics are:

a) Volume
b) Variety
c) Velocity
d) Variability

a) *Volume*: The volume of data in big data plays a very crucial role as it determines whether the value can be derived or not. It size of the data also determines whether it can be called big data or not. Hence, the data which is voluminous should be considered one of the characteristic of big data.
b) *Variety*: The variety in big data refers to the diverse data. The data may be structured or

unstructured , and the data is diverse which means it comes in various forms like text, images, audio, video etc. these are indeed a unstructured which pose trouble to store, analyze and access.

c) *Velocity*: The velocity in big data refers to as the data with the speed, the real potential of the data can be seen through the speed of data generation along with the data processing speed. The data flows with the speed from various sources like business processes, networks, social media sites, sensors etc .which is massive and continuous.

d) *Variability*: The variability in big data refers to the inconsistency, which usually inhibits the data to manage and use effectively [5].

B. CONCEPT OF MACHINE LEARNING

Machine learning is some branch under Artificial Intelligence where the machine learns from the data it is fed with and enhance itself with the data it has been exposed to. Machine copies or imitates like human brain while learning[1]. Conversion of data into some useful and important information is the main agenda behind Machine Learning. There are various techniques that machine learning uses to convert data which is random, complex and huge into useful as well as data which can be utilized for solving complex problems[2].

1) FUNDAMENTAL COMPONENTS OF MACHINE LEARNING

There are various number of algorithms used in machine learning and numerous number of algorithms developed and added frequently.

Every algorithm has a following three components:

a) *Representation*: The way to present the data using various algorithms like decision trees, support vector machine etc.

b) *Evaluation*: To assess the candidate hypotheses by using evaluation processes like accuracy prediction and recall, posterior priority etc.

c) *Optimization*: It is the search process which is used to generate the candidate programs. The various methods used are combinatorial , convex optimizations and many other.[1]

2) CLASSIFICATION OF MACHINE LEARNING

The data fed to the machine maybe labelled data or unlabelled data or both. There are 3 subcategories under machine learning.

a) *Supervised learning*: It is one of the most widely used machine learning model. In this form of learning , the machine is fed with the labelled data which is the input data as well as the required or expected results for the particular input data. Considering an example, there is a data about pets along with their characteristics and images, to classify them into the particular class of dogs, cats and fishes. Based on the training set of data available the machine will learn the characteristics and patterns through which it can categorize the input data into the particular class.

b) *Unsupervised learning*: In this form of learning, the machine is fed with the unlabelled data which is unlike Supervised learning and has the data unclassified into expected outputs. The task of the machine is to find the similarities or the characteristics which closely match the data.For example, consider the data which consists about sports likings of every person in a city. It can have data about number of people and a sport's name. The machine can cluster or categorize the data based on the frequency of people.

c) *Reinforcement learning*: In this type of learning, the machine has to think while making the decisions as it's action will be assessed by giving rewards or a punishment. Based on which it learns whether the decision it has taken is correct or not using trial and error method [3]. The machine here interacts with the surroundings or the environment to take proper decisions.

II. ANALYZING BIG DATA USING MACHINE LEARNING

There are various steps involved in machine learning and big data to derive the value or the required data out of enormous, unstructured data. The following are the main steps necessary for processing the data.

A. Data Pre-processing:

1) *Data cleaning*: This step involves getting rid of the replicas, filling values which are lost as well as modifying the faults in the dataset. It escalates the feature as well as the evenness of the data. This step is necessary because improper data may lead to uneven data and will create wrong estimates.

2) *Data alteration*: This step is to alter the data in the way which can be prepared for the algorithms that are going to be used in machine learning. It involves scaling the data for chosen algorithms, standardizing it and converting the variables. This step is necessary for refining the model performance.

3) *Feature engineering*: This is one of the most significant step as it contains choosing and mining the features from the raw data to be used in an algorithm. This allows purifying of redundant data which decreases over fitting and also abridges the procedure of learning.

4) *Feature Scaling:* In order to guarantee a persistent range of values through all input variables, feature scaling involves organising or creating the features. This phase can aid in enhancing the performance and stability of machine learning algorithms, particularly those that are delicate to the size of input features, such distance-based algorithms and gradient-based optimisation techniques.

5) *Handling extreme data:* The data may get imbalanced when the dissemination of classes in the dataset becomes unequal which may lead to

incorrect machine learning models. Methods like resampling and adjusting the learning algorithm.

B. Model Selection and Assessment:

This step involves picking the algorithm that matches the variables, and cross-validating the performance of model, and the description of evaluation metrics used on models, as well as hyperparameter tuning and model interpretability.

1) *Opting the precise algorithm*: Based on the data, it's structure and it's complexity, the algorithm can be chosen for implementation. It would be beneficial if various algorithms are used for experimenting with the data, and comparing the performance of them. It would help to identify the perfect algorithm for a particular problem.
2) *Cross-justification*: This method is used for evaluation of the performance of the model. The data is split into numerous subsections, the model is educated and verified using these combinations of subsections. It diminishes the threat of over fitting and it offers the more precise estimate of model's generalization capabilities.
3) *Estimation metrics*: It is a computable measure that is used to gauge the performance of machine learning model. There are numerous different metrics used for diverse problems, those are accuracy , recall and many more. The main agenda of this step is to opt the appropriate evaluation metric to understand the boons or banes of model.
4) *Hyperparameter Tuning*: Parameters of the algorithm which are fixed by the user but not well-read by the algorithm are termed as hyper parameters. This is a method to catch the best values for these parameters which will develop the model performance. The common methods for hyper parameter tuning are grid , random search and many more.
5) *Model Interpretability:* It refers to the model's skill to understand the estimates and to explain it. Methods to develop the model interpretability are feature importance analysis, partial dependence plots etc.

C. Machine learning pipelines for big data:

This step involves various other steps which takes care from storage to the deployment of the model and also monitoring it after deployment.

1) *Data consumption and storage*: It is the process of gathering, importing it and handling data from diverse sources for use in big data initiatives involving machine learning. The organization and management of data is referred to as data storage, and it can be accomplished through relational or NoSQL databases, HDFS ,or cloud storage options.
2) *Data handling and exploration:* It includes all the pre-processing steps along with the fact-finding data analysis which is used to check the data features and its connections.
3) *Model training and assessment:* It is the important step in which the model is trained by feeding the preprocessed data , to absorb the patterns and have some understanding. Next the estimation metrics and cross validation methods to gauge the performance of the machine.
4) *Machine Deployment and Monitoring*: Deployment is the process of integrating the trained model into production environment for making predictions and for decision-making processes. After deploying, the model's performance should be tracked to find the model's potential to work. It allows to keep track of accuracy of the model and also keep it updated [7].

III. APPLICATIONS

A. Recommendations in the ecommerce app: The big data is produced daily by the various ecommerce applications, if used with the machine learning which can predict the customers interests and recommend it to them. Lot of data that is produced is filtered to know the customers interests and is used to recommend the products that the customer is interested. For Example: the search history of customers is tracked and also the products viewed by the customers.

B. Healthcare: The big data with machine learning can be used in the healthcare industry too. The machine learning algorithms can be used to diagnose the patient and know about the illness. They also help to predict the disease using all the data that is fed into it. The data about diseases and the patients records are fed into the machine. For example, the machine looks at the symptoms that the patient is having and try to predict the disease which is considered to be the most effective way, because there is a least chance of human errors in prediction and diseases. This can be even used to give the proper treatment to the patients.

C. Entertainment industry: The huge amount of data is being produced by the entertainment industry on a daily basis. This can be used along with machine learning in order to be effective. The OTT platforms that are used by customers recommend the similar type of movies based on their interests. It actually work like ecommerce applications recommendation system. For example, the search history of the users, the movies, web series that are searched by users of particular genre and language will be recommended to them.

D. Estimating trends: The big data along with the machine learning can be used to predict or estimate the trends and the future growth of a particular industry. The current data and the past data of any particular industry can be fed to find and predict the trend. For Example: considering the present cell phone industry and previous years data , the machine can predict the which kind of cell phone

will be in a trend, which type of model the users like to use and feel comfortable to use irrespective of the age of the user[6].

E. *Chatbots*: The chatbots that are generated for a particular system allows user to interact with people. The chatbots are machine learning products which are used to provide the better service to the people or users. Ex: it acts a voice assistant that converts text to text. The data is fed to the machine that what kind of services that particular system will offer to users, based on that it generates the response to users questions.

IV. BENEFITS

A. *Quick analysis of data*: Machine learning is quick at handling large amounts of data, evaluating it and produces the accurate analysis out of it. Due to this feature, organizations send the client communications that are related and appropriate to customer's activities and interactions. The machine learning algorithm which has built a model using various source of information finds the connections between variables. It reduces integration issues and allows for get accurate findings.

B. *Real-time data prediction*: Big data analysts who use machine learning think of it as one of the most effective tool to predict real world activities precisely. It takes in voluminous amount of data along with related data or activities, combine it and produce the data which is more concise and useful. This allows analysts to make better advancements in other fields of research or study[8].

C. *Personalization*: The personalization allows to understand the customer behavior and likings to recommend the products that suit the taste of customers. This is one of the major benefits that the organizations can have to increase the business by improving the interactivity with the customers and to provide greater service which would benefit both organizations as well as consumers.

D. *Trend and profit predictions*: The huge organizations have the turnovers in the billions which is hard to calculate as well as to predict the future business. So the machine learning algorithms will be useful to gain meaningful insights of the future business trends as well as the profits. The machine will produce nearly accurate predictions using the data which it possesses [9].

V. CHALLENGES

A. *Bad data*: This is one of the biggest problems faced by algorithm implementation. Such problems arise due to noise and dirty data in the data. This reduces data efficiency. Therefore, it is important to carefull y follow the steps to filter out redundant data and create clear, clean data that can be used in algorithm.

B. *Absence of training materials*: Sometimes there may be less previous material that can be used as trainin g material. This can lead to incorrect estimation of p roblems. The data must be large enough to be used i n the algorithm to increase the accuracy of the predi ction. For example, consider if there is an algorithm that can predict whether an animal is a cat or a dog. And the lack of information about what the two ani mals' ears and eyes look like can lead to an inaccura te guess.

C. *Less data*: There may be less data before or after cleaning, making the process more difficult to use or use as training data. Entering less data into the machine can affect the forecast and cause data to be inaccurate.

VI. FUTURE TRENDS

A. *Data fabric*: It supports on-premises and cloud platforms. It can be used by organizations to incorporate the data storage on the cloud, which allows access to and allows sharing data on a scattered environment. Architecture of data fabric allows organizations to accumulate and fetch information, organizations can easily use the data fabric as it has supervisory environment guaranteeing that the data provided for analyzing is secure.

The artificial data which is produced algorithmically can be used as an alternative for manufacturing data which is in turn used to endorse the mathematical models but not to prepare algorithms.

Now, the focus is mainly on preparation of algorithms using artificial data, which yields huge set of training data. To make the data look a lot like the original data, various techniques like conflicting networks and simulators are used.

B. *Data as a Service*: In the old systems, the data used to be put into storage inside the data stores which were intended for particular applications to access, DaaS used cloud technology to deliver its users as well as applications an on-demand access of data irrespective of their location. DaaS were not usually built for supervising large data jobs but were expected to host the applications and store data only.

As the cloud is affordable, the DaaS can use the cloud-based platforms for supervising and processing large amounts of data with high speed and efficiency.

C. *Active Metadata*: The main concern is the enhancement of active metadata by human interaction, machine learning and process output. There exist several classifications of data , and the metadata is the data about data which informs users about the data. To guarantee that big data is properly conveyed and understood, the metadata supervision strategy is necessary.

The steps like gathering, processing and removing unnecessary data, should be followed to show there is a decent data management.. By using active metadata, there would be improved management and data will be available in a various forms.

D. *Edge Computing:* It refers to the process which shifts a process that is currently running on a local system such as system of a user or an IOT device, or a server, to another system. Edge computing is a significant trend in big data analytics, which enables data to be controlled at the edge of a network, lessening the amount of large distance connections between the server and the consumer. The edge computing reduces the latency period and increases data streaming. Edge computing is considered efficient as it takes lesser bandwidth and also reduces company's expenditures. It allows easy execution of distant software.

E. *NLP*: The process of analyzing the given text or speech data is termed as NLP which is the natural way of processing the data(language). It uses machine learning algorithms to achieve the language exchange between the computers and its users. It is a tool for effective communication between them. The algorithms used will read the text or listen to the speech and decode it to human language and finds the comprehension. The algorithm used will use the grammar rules to fetch the required data from sentence. The semantic analysis takes care of interpretation and syntactic analysis will take care of the format[10].

VII. CONCLUSION

There are many machine learning applications which have been used, in those many are used daily by the users knowingly or unknowingly. One such application is the recommendation system in any social media site. The customers find satisfaction when they find the related products when searched. The machine learning algorithm keeps the data of the customers taste to provide the best service possible. . These kind of innovations certainly bring the advancement as well as ease to the users who are using them. They reduce time consumption and are quick at processing the data.

In this modern era, everything around us is turned into automation which is reducing the human intervention. The artificial intelligence branch is one such branch which is expected to grow rapidly in future generations which has sub branches like machine learning and more. The machine learning using the big data is one of the most widely used as well as most dependent and trusted.

The machine learning has played the important or the most crucial role in connecting the big data. The big data which is incoming from various sources can only be effectively used by new technologies like cloud computing, IOT, Cyber security and machine learning. The machine learning considers data as a treasure or an asset which is the most necessary source of its working. The data that is incoming in large volumes is produced by the social media sites, networks, and various other businesses. To properly use the data machine learning finds the suitable algorithms based on the form of data. The data which is turned into a resource as it can be utilized for implementation.

REFERENCES

[1] Machine learning, explained bySara Brown Apr 21, 2021

[2] ANALYSIS OF MACHJNE LEARNING AND ITS ALGORITHMS Gomathi R 1, Ponnieshwari S2, Aparna R3 1,2student 3 Assistant Professor 7 ' · 60001 S.S.S. Shasun Jain College for Women, T. Nagar, Chennai,07 Feb,2020

[3] Machine Learning Basics: Components, Application, Resources and More Sep 26, 2022 By Chainika

[4] What is Big Data? Introduction, Types, Characteristics, Examples By David Taylor Updated May 13, 2023

[5] What is Big Data? A Quick Introduction for Analytics and Data Engineering Beginners by Siddharth Sonkar — Published On November 25, 2020 and Last Modified On May 30, 2023.

[6] Top 10 Real-World Machine Learning Applications By Simran Kaur Arora 07 Dec, 2022

[7] Machine Learning for Big Data: A Beginner's Guide – Netnut

[8] Benefits of Machine Learning for Big Data Analytics By Sangeeta Mittal APRIL 16, 2018

[9] Uses and Benefits of Machine Learning for Your Enterprise By Matillion - 05.09.2022

[10] 10 latest future trends in Big data analytics By InData Labs 4 October 2022