IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning : A Science To Evaluate Big Data

[1] Prakhar Bhatnagar, [2] Dr. Devesh Katiyar, [3] Dr. Dinesh Kumar Singh, [4] Mr Gourav Goel

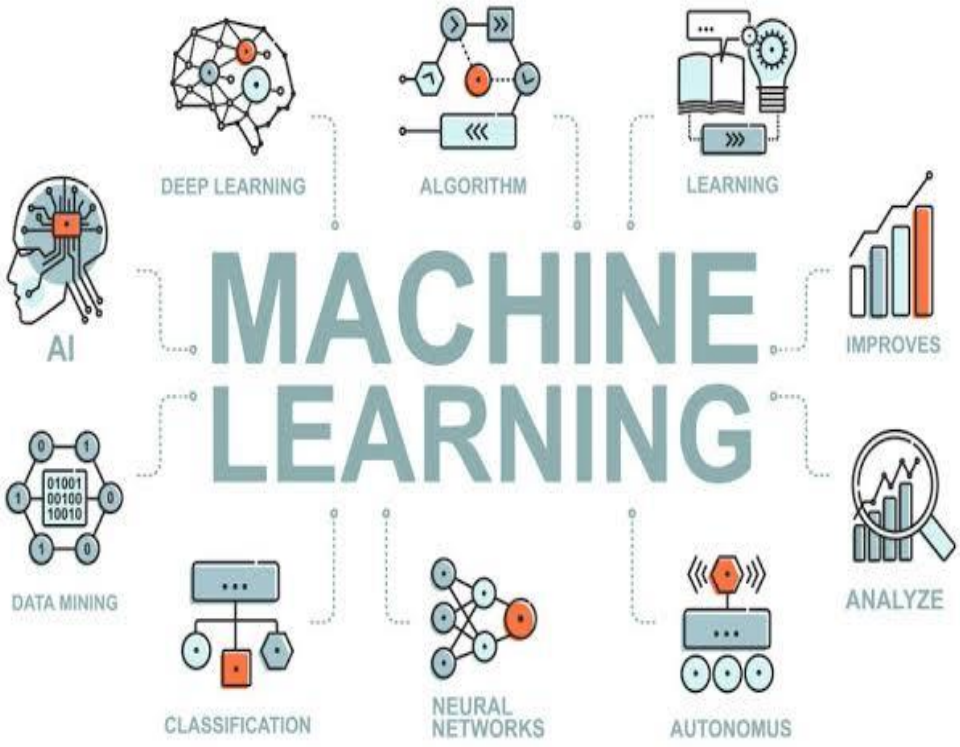
[1] Student, [2] Assistant Professor, [3] Assistant Professor, [4]Assistant Professor,
[1] Department of Computer Science,
[1] Dr Shakuntala Misra National Rehabilitation University, Lucknow, India

Abstract: Technology is emerging day by day. This emerging technology is creating zeta-bytes of data. The data created is not a waste, it's a fuel for the technology to help in evolving and bringing more advanced technologies. The 90% of today's data is created in the last couple of years and this trend will go on in the coming future. Sustainable computing studies in the information technology sector tell about the various methods by which computer engineers and scientists design various computers and their associated subsystems efficiently and effectively so that it has minimal bad effect over the environment. Current intelligent Machine learning systems today are only focusing over accurate prediction and classification of the data based over the training data set. Most of the machine learning algorithms are costly to operate in finding the accurate classification of the data. With the learning in the large data set the number of nodes within the network increases significantly, which results in exponential rise in computational complexities. This paper reviews the theoretical and practical data modelling techniques in large scale data intensive fields relating to model efficiency in large data sets and structure new algorithmic approaches that takes least memory in use and also processes to minimize computational cost while maintaining classification and prediction accuracy.

Keywords **-** Efficient Machine learning, big data, data mining, cost efficient computing, computational technique.

I. INTRODUCTION

Since stone age, humans are inventing tools and machines to accomplish their tasks and to increase their efficiency. As the human brain is evolving so are the machines, in older times machines and tools were simple to use and were easily designed. But now in this modern era machines are designed in more complicated manner with the help of chips and smart processors. Machines are further divided in software and the hardware. As the designs are getting complicated machines are becoming more human friendly as they are easy to use and carry. Some machines require only mechanical parts because they are used to carry out heavy tasks but some machines require both hardware and software to work. Now days or according to the future requirements there is a need that machines need to be developed in efficient intelligent models to cope with advancements that will happen in future as well as the energy saving needs as energy efficient devices will take over the market in future. Such energy efficient oriented data modelling will affect a large data modelling algorithms and the industries associated with such data modelling techniques. Coming back to Machine learning, this technology is highly dependable over large data sets, which can be from various fields such as biology, geography, accounts, physics and many more. Every day new algorithms are being introduced in the field of Machine learning, this is why the old research papers on this topic seems some what old or little bit outdated. The large data sets that are evaluated for the machine learning are related to the real world phenomenon. These real world phenomenon create two types of data sets basically known as training data set and the test data set. The big data evaluation can be made easy by the machine learning technology, that will provide lots of latest algorithms to work upon. Machine learning algorithms are not easy to understand but their predictions and results can be evaluated by any non technical person. A professional machine learning programmer can write and understand the codes as well as the data sets available for evaluation. The results generated by the machine learning algorithms can be favorable for the real world problems that may further help to make easy predictions.

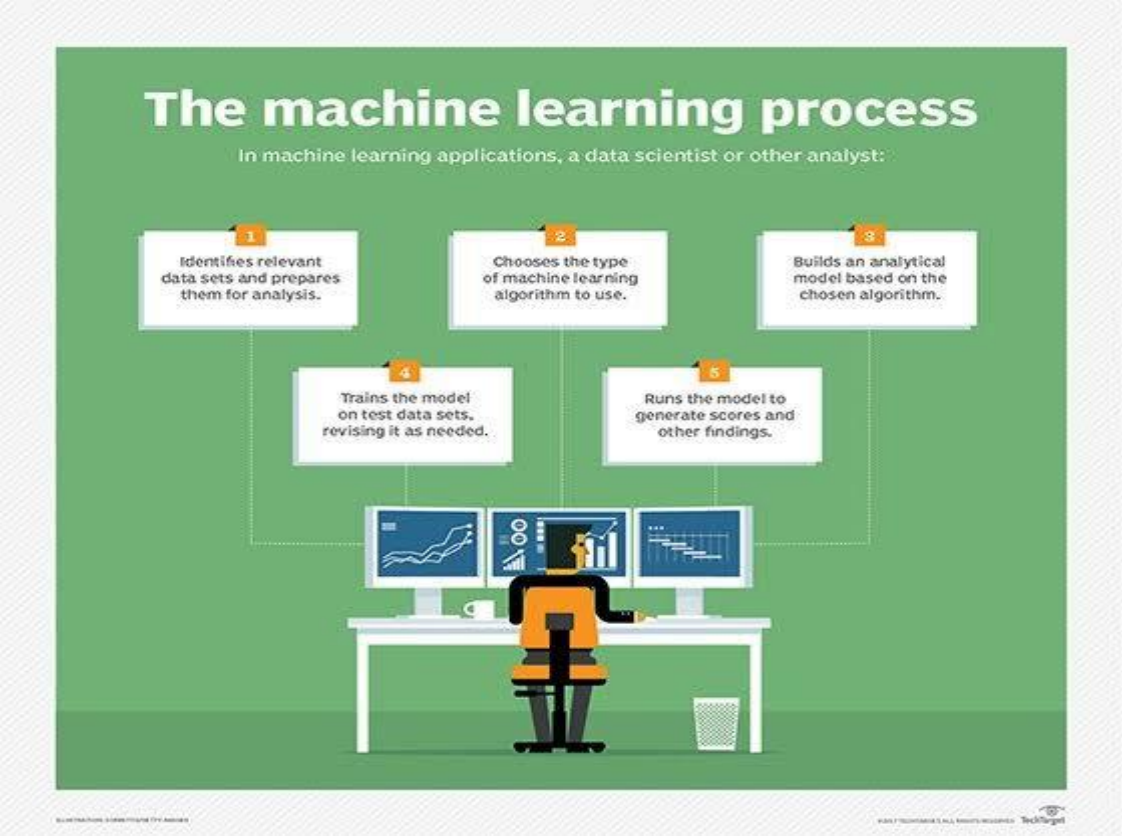


II. MACHINE LEARNING TO MAKE MACHINES LEARN

Machine learning research in today's era is conducted in a way that it impacts whole world. The results generated by the algorithms is impacting day to day life of a common person on this planet. Machine learning comprises of both theory and practical knowledge over data. For developing algorithms so that machines can learn both theory and practical work should be done simultaneously, both are important aspects for a good algorithm. The researchers claiming theory versus practical in terms of Machine learning can't be relied upon according to the present scenario of machine learning. These types of research does not affect Machine learning at a large point but if they are not eliminated they can have a bad effect over the results as well as the data. The use of old dataset for the algorithm makes machines to learn easily as well as reduces the machines load. There are many algorithms that can work on old datasets of other algorithms which reduces machines load and the machines are also aware of the data so they work fast and hence reduce energy. The very first step is to collect the data required for your algorithms, the data should be errorless, with no empty cells, with no duplicates. If the data collected contains any one of the above mentioned faults, then the very first task is to clean the data, fill the empty cells, remove the duplicate entries so that algorithms work on exact data. Now for the machines to learn itself an algorithm is required that will feed the dataset into the machines. Accuracy of algorithms depends on the dataset. Algorithms may malfunction if the dataset is not up to date or has some sort of error. Algorithms once implemented in the machines will make machines smarter and by the time machines will go on learning new data everyday and will itself keep increasing its knowledge. Every time that installed algorithm is used in the machine it will gain more data that will help it in increasing its knowledge. More the dealing with the data more the smarter is the machine. To make machines learn a correct combination of dataset with a good algorithm is a must. Even a slight error in algorithm or in the dataset will not make machines smart, enfact it can alter the functioning of the machine

III. FOCUS ON BENCHMARK DATASETS

Machine learning totally depends upon a good algorithm and a good dataset. Algorithm plays around 40% role in machine learning and dataset performs 60% role. An error in algorithm can be corrected at any stage but same is not the case for the dataset. Algorithm is the way or the path through which data will be sent in the machine to make machines learn and make them smart. ***The data is the oil on which machine learning algorithms work.*** An error or duplicacy or empty data cells can cause error in the functioning of the algorithm. Dataset with a good and relevant information is very much needed in machine learning algorithms. A dataset can contain thousands of pictures or videos or MS excel sheets with large sets of information. Dataset should contain many photos of a single object taken from every side and angle. Similar is the case for MS excel sheets used here, consider a case where a machine learning algorithm is to be used to predict the price of a property. So the MS excel sheet will contain some parameters on which the price of the property will be predicted and further those several entries will have thousands of entries in each column. These thousands of entries play a vital role in making machines smarter. These thousands of entries are the various categories through which a single parameter can be judged through different scenarios or situations.



IV. FOCUS ON EXTRACTION OF DATA

The large datasets provided to the machine via algorithms are of no use until the algorithm extracts the meaningful trends from the datasets. The meaningful trends are the useful information taken out from the data. The machine learning algorithms need good quality of complete dataset to work correctly and predict the future of some scenarios. The extraction of data can be done by deeply studying the data and understanding the trends hidden so that the results can be predicted correctly. Its very easy to download the dataset or create a dataset but even some little errors will bring out other results, that is it's a very sensitive content for the machine learning algorithm. Suppose there is a dataset for the closed eyes and there are by mistake some open eyes pictures are added then there creates a scenario that machine may get confused between these two and algorithm may predict or bring out faulty results and the root cause is the wrong extraction of information from the dataset as the information in the dataset is also wrong. As told above dataset is the oil, it is the most precious element for the algorithms to work smoothly and correctly. Extraction of information from the raw data is a very old traditional technology in the computer world, similarly the machine learning algorithms extract information as per the need and work according to the information gathered from the data set.

V. BIG DATA EVALUATION BY MACHINE LEARNING ALGORITHM

The learnings in machines can be of various types such as representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, kernel based learning which are broadly known as the machine learning. The machine learning algorithm is basically divided into three sub-domains namely supervised learning, unsupervised learning and reinforcement learning. Among all the learning methods the deep learning is considered the best and most in the machine learning algorithms. Deep learning broadly uses two types of learnings to evaluate big data namely the supervised and unsupervised learning. The big data consists of more complicated hierarchically launched statistical patterns of inputs to achieve new information of greater discoveries predicting about the data. Although big data is often used in machine learning still it encounters lot of challenges while making machines learn. From the name it is very obvious that the big data consists of large volume of data which is a big issue for the machine learning algorithms to handle. According to a report Google handles around 24 petabytes of data on daily basis, but if we take other sources then the data becomes more large. There is no doubt that today there is voluminous data around us in the world created every minute and with the increase in data, it is becoming difficult to create such algorithms which can handle large amounts of data with ease and even demand of such machines which have good hardware strengths to handle large amount of data. In big organizations data to be handled by them is increasing from petabytes to exabytes. Due to this scenario the demand for the machines with good storage and a central processor to handle data easily is increasing day by day. This problem of handling big data can be solved by implementing distributed frameworks with parallel computing should be preferred. Nowadays alternating direction method of multipliers is serving as a promising computing framework to develop distributed, scalable online convex optimization algorithms is well suited to accomplish parallel and distributed large scale data processing. The key merits of ADMM is its ability to split or decouple multiple variables in optimization problems which helps us to find a solution to a large scale global optimization problem by coordinating solutions to smaller sub problems. The challenges related to large amount data handling can also be handled by implementing some practicable parallel programming methods which can be proposed and applied to learning algorithms which can easily deal with large volumes of data. Suppose there is a large amount of data for the machine learning algorithm to read so the very first task for the machine will be to read and access the whole given data at a very fast speed, as today's era is of fast moving technology. The data accessing speed of algorithm should be high so that the data processing can start in no time after reading. The velocity by which the data will be read should be good as all the further processes depend upon the data as read and understood by

the machine. To handle such problems of the algorithm various learning methods can be implemented that can increase the speed of data accessing such as online learning approach. It's a well established learning method whose strategy is to learn one instance at a time instead of forming a batch of data in one instance. It is tested that this method of learning has increased the speed of data accessing and reduced the time. Another issue is the trust over data. The website or the source of the data cannot be trusted every time, thus the data on which your algorithm will run can come in the category of faulty or corrupt files. The solution for such kind of problems is to take data from the trusted sources. Trusted sources means official websites, official data warehouses or if taking from any person then he should be the authorized person to hold the dataset. Lastly the data chosen for the algorithm should be related to the tasks to be done by the algorithm, dataset should not be away from the algorithm's work. Similarly the data fed to the algorithm for a single task should not be of different varieties. A single dataset should be provided at a time for testing and training, different datasets from different sources should not be taken altogether to train the machine. Although lots and lots of data is essential to make machines smart but not altogether, it should be given sequence wise that is one dataset at a time.

VI. SCOPE OF MY STUDY

Machine learning is no doubt the most trending and demanded topic of today's time. Its making today's machines intelligent on their own. The machines that will be made in future will also be enabled with machine learning algorithms. With the advancement in technology dependence of humans over machines will keep on increasing and humans will demand of more intelligent machines that will reduce human involvement while machines work. The smart machines will be needed everywhere in business organizations, houses, corporate sectors, banks, hospitals that will reduce human efforts and make tasks easy and fast. As the technology is moving towards artificial intelligence and machine learning more job opportunities will come in this sector. Technology such as machine learning is showing good results in the machines it is installed so human dependence on such machines has to increase automatically. Many surveys and reports have shown that there is a huge difference between the predictions of the simple machines and of those machines which have machine learning algorithms installed in it. According to my study more and more machine learning algorithms enabled machines should be made in every sector so that the sector works efficiently and less human interaction with machines is made making machines smart and intelligent.

VII. CONCLUSION

The extensive techniques of mining data to make machines smart is the main motive of machine learning. The algorithms are designed in such a way that they read the dataset and extract a good set of information hidden in it to make machines think and take decisions on their own. In order to make machine learning the best technology its algorithms must be designed in such a way that they can read data at a very fast speed and make the further tasks take less time than usually they take. Artificial intelligence is a broad category under which comes machine learning and data science. AI makes the machines human like intelligent but with help of knowledge that is created artificially and machine learning contributes to it with the help of algorithms those are giving data to machines and making them learn. Machine learning will keep on evolving in the coming times as this technology is seen with the best one for machines at least for today and in the near future. In today's era business organizations, households, hospitals, banks are normally using machine learning technologies to make their daily tasks easy and fast. With some changes machine learning technology will keep on growing and will keep on advancing so that the future of humans with machines can be interactive and interesting.

VIII. REFERENCES

[1] Dr. Devesh Katiyar -Performance evaluating system based on MapReduce in context of Educational Big Data- https://www.igi-global.com/gateway/article/197871#pnlRecommendationForm

[2] Dr. Devesh Katiyar – A Study of IoT and Big Data- https://www.ijsr.net/get_abstract.php?paper_id=SR20911122957

[3] W Breuer, BI Steininger- Recent Trends in Machine learning http://www.machinelearning.ru/wiki/images/0/07/Langley00crafting.pdf

[4] M Skorikov, S Momen : Machine learning approach to predicting decisions- https://ieeexplore.ieee.org/abstract/document/9172011/

[5] R Brisebois, A Abran, A Nadembega : An Assisted Literature Review using Machine Learning Models to Recommend a Relevant- https://books.google.co.in/books?hl=en&lr=&id=mm9aCwAAQBAJ&oi=fnd&pg=PR5&dq=machine+learning+papers&ots=1W5RD7YkDm&sig=LGfg8JuKNgo0bRfZ2okC4UEdrX4

[6] M Atzmueller, A Chin, F Janssen, I Schweizer : Social Environments, MUSE 2014, and First International Workshop on Machine Learning for Urban Sensor Data, Sense ML 2014, Revised Selected Papers- https://www.researchgate.net/profile/Apollinaire-Nadembega-2/publication/321194657_An_Assisted_Literature_Review_using_Machine_Learning_Models_to_Recommend_a_Relevant_Reference_Papers_List/links/5a145200a6fdccd697bbdc50/An-Assisted-Literature-Review-using-Machine-Learning-Models-to-Recommend-a-Relevant-Reference-Papers-List.pdf

[7] N Nissim, A Cohen, J Wu, A Lanzi, L Rokach : Sec-lib: Protecting scholarly digital libraries from infected papers using active machine learning framework-https://ieeexplore.ieee.org/abstract/document/8788686/

[8] TG Dietterich, DC Wilkins, NS Flann : Real estate trends with help of machine learning- https://web.engr.oregonstate.edu/~tgd/publications/mlj-ijcai85review.pdf

[9] J Quiñonero-Candela, I Dagan, B Magnini : Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment, First Pascal Machine Learning-

https://books.google.co.in/books?hl=en&lr=&id=TT73BwAAQBAJ&oi=fnd&pg=PA1&dq=machine+learning+papers&ots=cuX_GZXwe4&sig=ClCejiOkKMJrS4krMsZCNZ_5kec

[10] RS Geiger, D Cope, J Ip, M Lotosh, A Shah: Garbage in, garbage out revisited: What do machine learning application papers report about human-labeled training data- https://direct.mit.edu/qss/article-abstract/2/3/795/102771

[11] Dr. Devesh Katiyar – An Introduction on Machine Learning- Applications and Opportunities- https://ijsrem.com/volume-05-issue-03-march-2021

[12] Dr. Devesh Katiyar – Driver Drowsiness Detection using Machine Learning – https://www.google.com/url?sa=t&source=web&rct=j&url=https://ijcrt.org/papers/IJCRT2109018.pdf&ved=2ahUKEwjZgpqBvKT5AhWI-jgGHSgnCH8QFrnoECA0QAQ&usg=AOvVaw1mumd7y2_wozE_msute7UQ