*ISSN. 0972 - 8406*
*The NEHU Journal Vol.* ***XVII,*** *No. 1 (January - June) and No. 2 (July - December) 2019, pp. 133-138*

# A preliminary note on big data and machine learning technologies

*[*]Sandeep Sharma and N.C. Barwar*
Faculty of Engineering and Architecture
Jai Narain Vyas University, Jodhpur, India
[*]Corresponding author : sandeepsharma8892@gmail.com

## Abstract

*Big data is high-volume, high-velocity and variable information assets which demand cost-effective, innovative forms of processing for enhanced insight and decision making. An essential quality of the Big Data is the large volume which is heterogeneous and of different dimensions. Data mining and machine learning systems are utilized to separate the important and concealed examples from the huge volume of data. Many machine learning strategies are coordinated with big data analytics tools.*

**Keywords:** Machine learning, Big data, Data mining.

## Introduction

In recent times there has been an exponential production of data from various sources of the web, smart phones or smart sensors, which has lead to generation of big data. The term big data can be referred to as enormous, fast, arising, various classes and with parts of undesirable noises that are hard to store, process, analyze, translate, expend and settle for better decision in the field of medicinal services, funds, and business or industries. Gigantic data have originated from people through the usage of PC, advanced mobile phones, gadgets which are utilized to share message and recordings with companions in internet based life such as Facebook, Instagram, Whatsapp, etc, for sharing short clips, share their perspectives and purchase where data gathering has developed enormously and is already past the capacity of commonly utilized software tools to capture, manage, and process inside a "tolerable elapsed time" (Wu *et al.,* 2014) (Blazquez and Domenech, 2018)

Indeed, the activity of people and their exercises are recorded by smart sensors which are set in part of urban communities and in diverse public places. The most fundamental challenge for Big Data applications is to investigate the enormous volumes of data and focus on helpful data or information for future activities. In many cases, the learning/extraction procedure must be productive and near continuous on the grounds that putting away all watched data(Wu *et al.,* 2014).

Big Data begins with huge volume of heterogeneous, self-ruling sources with distributed and decentralized control, and tries to investigate complex and advancing connections among data which is known as HACE Hypothesis. These attributes make

it an extraordinary test for helpful learning from Big Data (Wu *et al.,* 2014, Chen and Zhang, 2014, Ke and Shi, 2014.)

**Big Data:** Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, and updating and information privacy (Saidulu *et al.,* 2017). Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. (Khine and Shun, 2017)

One of the essential qualities of the Big Data is the large volume of data which is heterogeneous and of different dimensions. Distinctive data gatherers lean toward their very own schemata or protocols for data recording, and the idea of various applications brings about varying data representations (Wu *et al.,* 2014, Ke and Shi, 2014).

However, there are traditional techniques for managing huge data and learning from these enormous data, open doors for Machine learning (ML). It is a sort of artificial strategy which is utilized for finding information from enormous data for settling on better astute decisions (Saidulu *et al.,* 2017). Machine learning algorithms arrange the learning task in three classes, *viz.* supervised, unsupervised and reinforcement learning. (Lakshmi and Sheshasaayee, 2015, Sheshasaayee and Lakshmi, 2017.)

**Machine Learning:** Machine Learning is a plan to gain from models and experience, without being expressly customized. Rather than composing code, one can feed data to the conventional algorithm, and it constructs relation dependent on the data given (https://towardsdatascience.com/introduction-tomachine-learning-db7c668822c4).

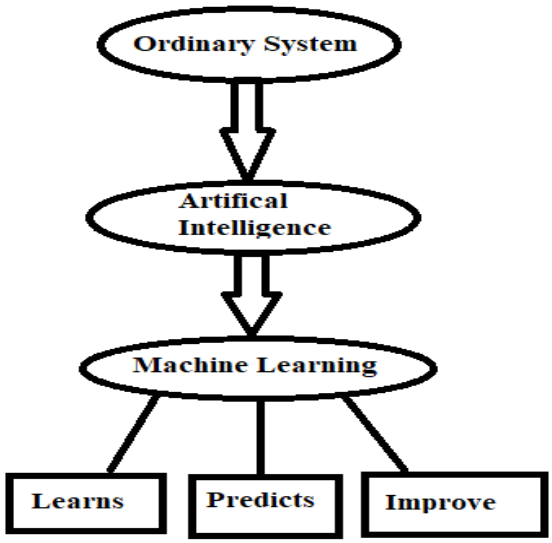
Machine learning is a data analytics procedure that instructs PCs to do what falls into place without any issues for people and creatures: gain from experience. Machine learning algorithms utilize computational techniques to "learn" information legitimately from data without depending on a predetermined condition as a model. The algorithms adaptively improve their exhibition as the quantity of tests accessible for learning increments. Deep learning is a specific type of machine learning (https://in.mathworks.com/discovery/machinelearning.html).

Machine Learning is a field which is raised out of Artificial Intelligence (AI). Applying Artificial Intelligence, we needed to manufacture better and keen machines. (Khine and Shun, 2017) In any case, besides a couple of unimportant undertakings, we were not able to program the unpredictable and always encounter difficulties. Therefore, the best way to have the option to accomplish such undertaking is to give machine a chance and gain from itself.

Machine learning, in present times is available in such a significant number of fragments of innovation that we don't understand while utilizing it (https://towardsdatascience.

com/introduction-tomachine-learning-db7c668822c4).

Example:




As an example, it can be assumed that a system is fed with input data that contains the photographs of the students of a college. At that point we do:

- Analysis of data. The system attempts to discover patterns, for example tallness, size, etc.
- Dependent on these patterns, the system attempts to predict various types of student have a place with specifics course, fittest student, etc and partition them.
- Finally, it keeps all tracks of the choices; to ensure that it is learning. At this point of time when requested, the machine can predict and isolate the various kinds of students. Thus, the machine does not repeat the whole procedure once more. That is how the machine learning works.

With the rise in big data, machine learning has become a key technique for solving problems in diverse areas, such as: **Automotive, aerospace, and manufacturing,** for predictive maintenance, **computational biology,** for tumor detection, drug discovery and DNA sequencing, **computational finance,** for credit scoring and algorithmic trading, **energy production,** for price and load forecasting, **image processing and computer vision,** for face recognition, motion detection, and object detection, **natural language processing,** for voice recognition applications. Generally, the field of machine learning is divided into three sub domains: Supervised learning, unsupervised learning, reinforcement learning (Mustafi, 2016)

**Supervised Learning:** The machine learns from the training data that is labeled. So one has to supervise the machine learning while training it to work on its own. Supervised

learning requires training with labeled data, which has inputs and desired outputs.

**Unsupervised Learning:** The machine learns from the training data but without label. Unsupervised learning finds hidden patterns or intrinsic structures in data. It draws inferences from data sets consisting of input data without labeled responses (https://in.mathworks.com/discovery/machinelearning.html.).

**Reinforcement Learning:** The machine learns on its own i.e. by its mistake and experiences. (Chen and Zhang, 2014)

Over the last few decades, heart disease has been the most common cause of casualties around the globe. An early detection of heart disease and continuous monitoring can reduce the mortality rate. The exponential growth of data from different sources such as wearable sensor devices used in health monitoring activities- thanks to Internet of Things. This resulted in generation of an enormous amount of data on continuous basis. The combination of streaming big data analysis and machine learning is a breakthrough technology that can have a significant impact in healthcare especially for an early detection of disease, so developing a distributed and real-time healthcare analytics system using traditional analytical tools is extremely complex, while exploiting open source big data technologies can do it in a simpler and more effective way. A real time heart disease prediction system based on apache Spark which stands as an effective large scale distributed computing platform which can be used successfully for streaming data event against machine learning through in-memory computations has been recently developed. (Sheshasaayee and Lakshmi, 2017, Lakshmi and Sheshasaayee, 2015).

One important developmental discovery made by software engineers is the choice of the language that is used in the implementation of these algorithms, which need to be adopted and optimized applications. Learning for large scale of data, learning for different types of data, learning for high speed of streaming data, learning for uncertain and incomplete data, learning for data with low value density and meaning diversity are some of the important issue for future.

**Challenges:** Data capture and storage, Data transmission, Data curation (data quality assurance), Data visualization, Data analysis, are the issue which the scientists need to face in near future (Ed-Daoudy and Maalmi, 2019.)

To process the huge sized unstructured, inconsistent, incomplete and vague data generated by computing machines is a challenging task. To perform operations in the data, present in higher dimensions may be more computationally complex procedure as well as the computational overhead in further training and testing phases of classification (Xindong *et al.,* 2014.) In recent past years, Rough set theory and Fuzzy logic evolved as an efficient machine learning methodology, which has grown as an important tool to perform big data analytics.

Presently we are in a period of big data, preparing and investigation of huge

estimated, unstructured, conflicting, deficient and uncertain data which is a big testing assignment. To perform activity and investigate deficiency in a data, present in higher measurements is simply muddled or complex.

With the rise in big data, machine learning has become a key technique for solving problems in areas (Ruiz *et al.,* 2017) such as: automotive, aerospace, and manufacturing, for predictive maintenance, computational biology, for tumor detection, drug discovery and DNA sequencing, computational finance, for credit scoring and algorithmic trading, energy production, for price and load forecasting, image processing and computer vision, for face recognition, motion detection, and object detection, natural language processing, for voice recognition applications, and many more (Chen and Zhang, 2014, Ke and Shi, 2014.)

**Acknowledgement**

The authors are thankful to Prof. S. R. Rao and his research scholar, Ms. Sohini Deb, Department of Biotechnology and Bioinformatics, North- Eastern Hill University, Shillong for revising the manuscript.

**References**

Ed-Daoudy, A. and Maalmi, K. 2019. Real-time machine learning for early detection of heart disease using big data approach. *Int. Conf. on Wireless Tech. Embedded and Intelligent Systems (WITS*), Fez, Morocco, pp. 1-5.

Sheshasaayee, A. and Lakshmi, J. V. N. 2017. An insight into tree based machine learning techniques for big data analytics using Apache Spark. *Int. Conf. on Intelligent Computing, Instrumentation and Control Technologies (ICICICT),* Kannur, pp. 1740-1743.

Lakshmi, J. V. N. and Sheshasaayee, A. 2015. Machine learning approaches on map reduce for Big Data analytics, *Int. Conf. on Green Computing and Internet of Things (ICGCIoT)*, Noida, pp. 480-484.

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, 2014. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26 (1): 97-107.

Blazquez, D. and Domenech, J. 2018. Big Data sources and methods for social and economic analyses. *In: Technological Forecasting & Social Change* 130, pp. 99–113.

Chen, C.P. and Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275: 314-347.

Ke, M. and Shi, Y. 2014. Big data, big change: In the financial management. *Open Journal of Accounting*, 3, 77-82.

Khine, P.P. and Shun, W.Z. 2017. Big Data for organizations: a review. *Journal of Computer and Communications*, *5*(3): 40-48.

Ruiz, Z., Salvador, J. and Garcia-Rodriguez, J. 2017, June. A Survey of Machine Learning Methods for Big Data. *International Work-Conference on the Interplay Between Natural and Artificial Computation.* pp. 259-267. Springer, Cham.

Saidulu, D. and Sasikala, R. 2017. Machine learning and statistical approaches for Big Data: issues, challenges and research directions. *International Journal of Applied Engineering Research*, *12*(21): 11691-11699.

Mustafi, J. 2016. Natural Language Processing and Machine Learning for Big Data. In: *Techniques and Environments for Big Data Analysis,* pp. 53-74. Springer, Cham.

https://towardsdatascience.com/introduction-tomachine-learning-db7c668822c4.

https://in.mathworks.com/discovery/machinelearning.html.

Internet Resources:

a https://towardsdatascience.com/introduction-tomachine-learning-db7c668822c4.

b https://in.mathworks.com/discovery/machinelearning.html.