International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

Data Science: A Revolution of Data

Sunil Kumar[1], Dr. K.L. Bansal[2]

[1] Research Scholar, Department of Computer Science, Himachal Pradesh University, Shimla, Himachal Pradesh, India
[2] Professor, Department of Computer Science, Himachal Pradesh University, Shimla, Himachal Pradesh, India

---

Abstract *- This paper aims to analyze some of the different areas where humongous (Big-Data) amount of data which is being generated on daily basis from various sources across the globe can be processed and exploited in order to effectively and efficiently use the information hidden in that data. This big-data when combined with artificial intelligence, machine learning, deep learning, and data science concepts and algorithms, helps decision makers in efficient decision making. We have enlisted various application areas where data is being getting used for designing efficient system which can give more accurate results, thus more accurate predictions can be made.*

Keywords: Big-data, artificial intelligence, machine learning, deep learning, data science, algorithms, decision making.

1. INTRODUCTION

Every part of a business generates some data which may not seem useful to those who generates it but for decision makers that data plays a very crucial part in decision making especially decisions related to various operations of the business. Surge in data creation is mainly due to advancement in technology, growth of www and smartphone over the past decade. Data is being collected from various sources viz. primary and secondary. Data now includes text, audio and video information that is data can be structured, unstructured, and semi-structured.

Data thus collected needs to undergo ETL (Extraction, Transformation and Loading) process so that only such data can be extracted from the bunch of data which is useful to the organization, data thus extracted needs to be transformed to a common structure and then loaded to a staging area or data warehouse where various analytical tools can be applied on the data thus extracted, transformed and loaded.

Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Data Science (DS) all these fields combines the power of statistics, mathematical modeling, automation and programming for creating an efficient system which can solve real world problems within fraction of seconds, the system thus created possesses the capability of learning on itself.

2. DATA

The way a human sees various things, observe various features of that thing, extract important features [9] and considering the importance of selected features take necessary action. Similarly, machines or computer programs are required to extract important features from the data fed to the program and then take or suggest required decision or action. The data fed to the program for observation and processing can be either of the following types:

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. The data which doesn't follows traditional row and column structure falls in the category of unstructured data. Texts, pictures, videos etc. are the examples of unstructured data.

3. BIG-DATA

Data nowadays is increasing at a rapid scale. It has been projected that by 2025, the global data creation will be more than 180 zettabytes. One can imagine how valuable that huge data (Big-Data) can be. Big-Data has following characteristics:

Volume:

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.

Velocity:

The velocity aspect of Big Data demands analytic algorithms that can operate data in motion, ie, algorithms that do not assume that all the data is available all the time for decision making, and decisions need to be made "on the go," probably with summaries of past data.

Variety:

Data coming from various data sources is expected to be of different types like structured, semi-structured and unstructured data. Such data may come from inside or outside of the organization.

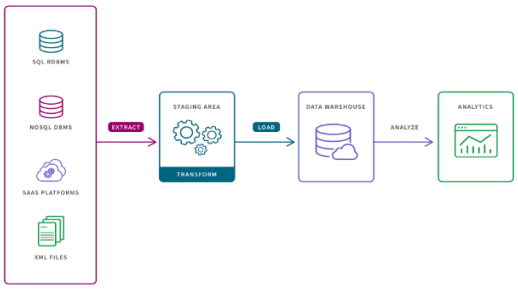
Veracity:

Veracity refers to the truthfulness or quality or accuracy of data and trustworthiness of data source from which data is being collected.

Value:

Data no matter how big, is of no use to organization until it is processed and useful information is collected from it. Such useful information helps decision makers to take necessary decision which helps in the growth of organization.

In order to use data for decision making, useful data must be first extracted from various data sources (primary and secondary). The data present in these data sources can be either of the forms.



After extracting data from data sources, data needs to be transformed into standard form. In transformation process various operations are performed like feature selection, cleaning of null values or removal of missing value data for consideration, joining of multiple attributes into one or vice-versa, etc. After transforming data, data is being loaded into the data warehouse from where it is fed to On-Line Analytical Processing (OLAP) system which helps the organization in retrieving useful and valuable insights from the data.

4. ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) is a branch of computer science in which we create intelligent machines or programs. The machine or program thus created behaves like a human, thinks like humans, and is able to make decisions on its own. The term "Artificial Intelligence" is a combination of two words "Artificial" and "Intelligence". The word "Artificial" refers to the something which is created by human to mimic something natural and the word "Intelligence" refers to the ability of thinking on its own, so it can be viewed as "thinking power created by human" [8]. The process of creating intelligent machines and programs goes through several stages of planning, reasoning, analyzing data, prediction of outcomes and acting accordingly. AI also involves the use of statistics and probability and various other mathematical tools (neural networks and machine learning is mostly based on these).

5. MACHINE LEARNING

Unlike Artificial Intelligence which solves a problem mimicking human intelligence, Machine Learning solves a problem by learning from data and making predictions. Machine learning is subset of Artificial Intelligence. So one can say that all machine learning is artificial intelligence but not all artificial intelligence is machine learning.

Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications.

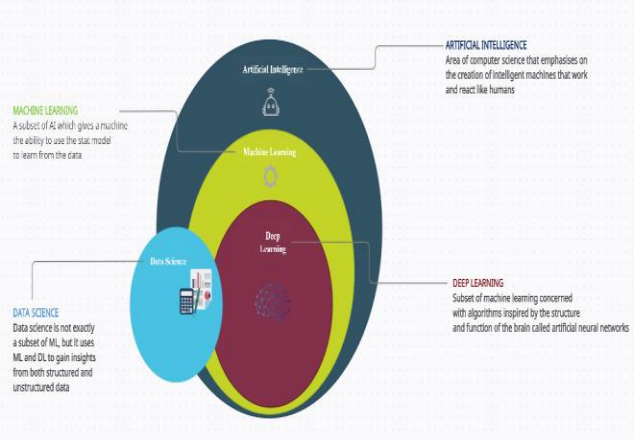
6. DEEP LEARNING

Deep learning is a new learning paradigm which is a subset of machine learning (ML) which is a subset of artificial intelligence (AI). Advancement in image analysis and image processing using deep learning resulted into generating massive interest in this field due to its learning capabilities from the given data [1]. A deep learning algorithm and support vector machine (SVM) is used to categorize the image as healthy or COVID-19 infected by

the feature extraction process with chest X-ray images. Various deep learning models like Inception-v3, AlexNet, VGG16, Inception-ResNet-v2, VGG19, ResNet-18, ResNet-50, GoogLeNet, ResNet-101, DenseNet201, and XceptionNet were used and achieved a 95.38% accuracy with ResNet50 and SVM. Deep learning has been successfully applied to several application areas which include image recognition [3,4], speech recognition [5], natural language understanding [6] and computational biology [7].

7. DATA SCIENCE

Data science is the process of finding meaningful information or insights from the data in a particular domain where deep learning can play a vital role in data analytics and decision making. Data science is an emerging discipline which came into limelight in the last few years. Data science is a collective discipline which combines various concepts of different fields. It merges the concepts of computer science (algorithms, programming, artificial intelligence, machine learning, and deep learning), mathematics (statistics and optimization), and domain knowledge (business, applications, and visualization) to extract insights from data and transform it into actions that have an impact in the particular domain of application. Implementing these concepts of computer science and mathematically tweaking the data with domain knowledge, results into accurate and effective insights from data. These previously hidden insights from data can be extremely important for the organization, as such valuable information may help the decision makers in taking more accurate and efficient decisions for the benefit of organization especially in the current era of extremely competitive market.



8. APPLICATIONS OF DATA SCIENCE

8.1 Fraud and Risk Detection

One of the major application area of data science is finance. Banks and companies dealing with finance were in huge loss due to various factors which includes bad debts, banking fraud etc. Even though these organizations had a lot of data about their customers including transactional data but were lacking resources to draw insights form this data. As data science started evolving, these financial organizations started using this data for drawing meaningful insights. They started listing their customers as per their expenditures and profiles and on various other factors which might help in deciding whether the customer can default on his loan or not. This also helped them to do target marketing based on customer's purchasing power.

8.2 Healthcare

The healthcare sector, especially, receives great benefits from data science applications. Using data science in healthcare sector may help health experts in diagnosing the disease at an earlier stage and in taking important decision combining their own expertise at right time so that better treatment can be given to patients. Following are some of the healthcare areas where data science can be used more efficiently:

8.2.1 Medical Image Analysis

Procedures such as detecting tumors, artery stenosis, organ delineation employ various methods to find optimal parameters for tasks like lung texture classification. It applies machine learning methods for bio-medical image analysis which helps in texture classification and detection.

8.2.2 Personalize treatments

Integrating data science concepts for understanding the impact of DNA on our health and finding biological connection between genes, disease, and drug response of an individual. This provides us meaningful insights of genetic issues in reactions to particular drugs and diseases. This helps in personalization of treatment for individuals.

8.2.3 Drug Development

Drug development often includes testing, huge financial and time expenditure. Usually this process takes over a

decade for successful drug development. Data science in combination with machine learning algorithms simplifies this drug development cycle from the very beginning to end that is from selection of drug compounds to the prediction of the success rate based on the biological factors. Simulating lab experiments using machine learning algorithms and modeling can forecast the reaction of drug compounds in the body with higher accuracy.

8.2.4 Virtual assistance for patients

Artificial intelligence powered virtual assistants can provide basic healthcare support in many cases in which it is not required for the patients to visit the doctor in person. Such virtual assistants are trained using various machine learning algorithms and helps the patient by asking him to describe the symptoms, past medical history and then receive key information about his medical condition.

These virtual assistants encouraging patients to make healthy decisions, saves their time waiting in line for an appointment, and allows doctors to focus on more critical cases.

8.3 Targeted Advertising

Data science helps the organizations in advertising their or their client's product to visitors of their website as per the likings or search of the visitor. Different visitors may find different advertisement in the same place at the same place. This is the reason why organizations are focusing on digitally advertising their products than placing hoardings near to the public places or on highways.

8.4 Advanced Image Recognition

Machine learning algorithms are being used for recognizing various objects in an image or video. Such algorithms work by detecting various objects in the image or video and then classifying the detected object according to their features. Machine learning algorithms are being used for training machine learning model with huge datasets which contains images or videos containing different objects. For instance, you upload an image containing faces of your friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm.

8.5 Speech Recognition

Machine learning is widely used in speech recognition systems. Amazon Alexa, Google assistant, Siri, Cortana etc. are some of the most widely used voice assistant systems which processes our voice, analyze it and then gives us the appropriate information or feedback as per our voice command. There are systems which convert our speech to text, so if you are not in a position to type a message, just speak it out and it will be converted to text.

9 CONCLUSION

In this research we have examined the ingenious topic of big-data, artificial intelligence, machine learning, deep learning and data science. All these topics gained lots of interest due to their perceived unprecedented opportunities and benefits. Variety of humongous data is being collected at a rapid speed, within this data lies the hidden knowledge which should be extracted and utilized. The literature was reviewed in order to provide an overview of application areas of Artificial intelligence, machine learning, deep learning and data science which are being researched.

Accordingly, this research enlisted various application areas viz. banking fraud detection and risk analysis, healthcare, targeted digital marketing, image recognition and speech recognition where with the help of AI, ML, DL and DS concepts, data is being used to design intelligent systems. Data acts as a base for designing intelligent decision making systems as without having data or enough data in hand we won't be able to make any learning model efficient and effective.

It is believed that data when used in a more efficient manner that is with appropriate technology and algorithm is of great significance in almost all areas which generates data.

REFRENCES

[1] Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers In Artificial Intelligence*, *3*. doi: 10.3389/frai.2020.00004

[2] Wan, L., Zeiler, M., Zhang, S., Cun, Y. L., and Fergus, R. (2013). "Regularization of neural networks using dropconnect," in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 1058–1066.

[3] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012a). ImageNet Classification with Deep Convolutional Neural Networks. Curran Associates, Inc.

[4] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521:436.

[5] Graves, A., Mohamed, A., and Hinton, G. E. (2013). "Speech recognition with deep recurrent neural networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[6] Sarikaya, R., Hinton, G. E., and Deoras, A. (2014). Application of deep belief networks for natural language understanding. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 778–784. doi: 10.1109/TASLP.2014.2303296

[7] Leung, M. K. K., Xiong, H. Y., Lee, L. J., and Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. Bioinformatics 30, 121–129. doi: 10.1093/bioinformatics/btu277

[8] Dhankar, M., & Walia, N. (2020). An Introduction to Artificial Intelligence. In Emerging Trends in Big Data, IoT and Cyber Security (pp. 105-108). EXCELLENT PUBLISHING HOUSE. Retrieved from https://msi-ggsip.org/wp-content/uploads/conference2020.pdf

[9] Guyon, I., Gunn, S., Nikravesh, M., & A. Zadeh, L. (2006). Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing, 207). Springer.