Big Data and Medical Research

Elangeshwaran Kannabiran

MSc.Computational Neuroscience & Cognitive Robotics, University of Birmingham, Birmingham, UK

Deepa Narasappa

MSc. Medical Engineering, KTH Royal Institute, Stockholm, Sweden

Tejaswini Prakash

MSc. Data Analytics Engineering, George mason university, Fairfax, Virginia, USA

Email: elangeshwaran1@gmail.com,tejaswiniprakash1003@gmail.com, deepanayak1002@gmail.com

Phone: +91 9620000612, +91 8867903072, +91 9066628011

ABSTRACT- To include large data sets so that they can be used over common softwares, the term big data was introduced in the late 90s. It has been predicted that in 2025, the usage of Big Data in the field of Medicine will increase exceedingly. Big data in medicine may be used by commercial, academic, government, and public sectors such as electronic health data, biologic and biometric data. It has contributed to changes in a lot of research methodology. The influence of the changes in the clinical research on a large-scale biological data harvesting developed, analysed, and managed by normal computing technology with the support of flexibility between real-world data and academics, industry, regulators. The access via the internet has eased the rate of participation. The current world desires immediate solutions which can be obtained/ supplied by big data. It can reveal Health patterns and give out promising solutions from the data obtained from previous experiences. However, privacy, data ownership, public ignorance and international laws are the reason for Slowing down its progressive use. This paper reflects on how Big Data and different techniques, algorithms and methods of Data handling can be used in overcoming demands in data storage, processing and analysis in different fields of medical research like Image Analytics, Neuroscience and Bioinformatics.

***Keywords*- Big Data, enormous Data, Image Analytics, Neuroscience, Bioinformatics, Map reduce, Hadoop, Apache spark, NeuroPigPen, Hive, Machine Learning, Unsupervised Learning, Supervised Learning, Supervised Mining, CT, MRI, X-Ray, EEG, DNA, mRNA, histone.**

I. INTRODUCTION

Big Data

The enormous amount of data that is uncontrollable when using traditional softwares and internet- based platforms is termed generally as 'Big Data'. During the 1990's this was introduced so that the data sets that were too colossal to be used with common software could be included. Later on, for the transformation to be put into use, the asset of information was characterized with large volume, velocity and the diversity required for specific technology and analytics methods. These attributes along with a few more such as quality, value are of high use to make the Big data more effective. Since the 90's it is seen that the volume has been gradually increasing universally with the collection rate amplifying every 40 months. Post 2000s the increased amount of alphanumeric data, images, audio information through social media, internet based devices including smartphones, computers, IoTs, EHRs, insurance websites, geo-spatial data and videos. In modern times, every 30 minutes, more than 105TB of data is analyzed by Facebook.

Furthermore, the data being statistically complex and dynamic, it needs to be available real time, which is then allowed to be used immediately. Having enormous volume and diverse characteristics, it is given that Big Data requires appropriate management technologies along with software, framework and expertise. Variety of trends such as weather patterns, business census, shopping etc are shown by Big data. In 2014, the United Nations (UN) recognizing and perceiving the power of big data created a Global Working Group on big data under the UN Statistical Commission. For data sharing and economic benefit, the global statistical community was created by using big data technologies in the UN global platform and they made this their main vision. The uses of Big data in medical research is diverse.

Types of data visualizations

To administer big data tasks, one of the important infrastructures is Parallel Computing. On a big group of machines or supercomputers, this technique makes it efficient of algorithm tasks parallely. In modern times, Google has introduced a platform called Map

reduce which is a parallel computing model which is there to propose a new big data framework. For distributed data management, Apache released an open source by map reduce called Hadoop. Synchronous data access to cluster machines is supported by the Hadoop Distributed file system (HDFS). These services are also perceived as cloud computing platforms and it allows access to the data centrally or from an isolated location.

For sharing data and its assets over the network, cloud computing is said to be a novel model and it is also favorable as it can likewise serve as a platform or a software for providing integrated solutions. The system is said to have an increase in its speed and agility and its flexibility with the usage of cloud computing which makes the system run programs quickly. Majority of new big data technologies use this software.

In healthcare, big data is said to consist of billions of entries about drugs, prescriptions, patients, treatments, surgical procedures, research results, and many more. We have to think of a way to analyse the data and process it efficiently if we want to use all that on a regular basis.

Big data can aid in supporting clinical treatment or monitoring efficiency of the healthcare companies, improving patient service, as a big-data healthcare repository, determining and implementing appropriate methods for patient treatment, determining and implementing appropriate methods for patient treatment,Electronic health records, Digitization of healthcare data and biomedical research and much more. We chose to put forth how Big Data is used in different streams of medical research in this paper.

METHODOLOGY

I. Big Data in Image Analytics

CT, X-ray, MRI, PET, EEG, and molecular imaging along with photo-acoustic imaging are the most common imaging techniques used in the world. The data collected in these high-resolution images are large. Radiologists, doctors and other health care professionals are doing wonders in evaluating the medical data which are in the format mentioned above and with the help of it target different diseases. Many systems like PACS (Picture Archiving and Communication Systems) have been developed for storing and easier access to medical images and report data in order to establish an efficient system for these professionals. However, to retrieve medical images, data exchange in PACS depends essentially on using structured data. This means there is a loss of unstructured data while retrieving these medical images. Moreover, you can easily miss a patient's data from these images. Hence a professional diagnosing a condition might easily miss this condition, especially when the condition is still in its developing stages. Hence, from biomedical images, biomarkers are being developed by imaging analytics in order to help avoid such errors. This approach uses pattern recognition and machine learning techniques to draw insights from large volumes of clinical image data to transform the diagnosis, treatment and monitoring patients.

To dig out the hidden information, image analysis is performed by a handful of software tools that have been developed based on the features such as generic, reconstruction, segmentation,registration, visualization,, simulation and diffusion. A freely available software which allows dynamic processing and analysis of 3D images from medical tests is Visualization Toolkit. SPM is a software which can process and analyse 5 different types of brain images (e.g. MRI, fMRI, PET, CT-Scan and EEG). GIMIAS, Elastix, and MITK are few other softwares which backs all types of images.

II. Big Data in Neuroscience

EEG data plays a relevant role in diagnosis and analysis of brain, neurological disorders and brain connectivity research. These signals are non-stationary and non-linear in identity making it difficult to perceive perfect properties. The use of Big data can be fruitful for neurological scientists for clinical diagnosis as SEEG data record activities of the functional brain in both temporal and spatial scales.

Big data is a complex form of large datasets, it is difficult to process data with the current data processing application but techniques in Big data like Hadoop, Map reduce algorithm, Apache spark, Pig and Hive can be useful in handling neurological data.

- Map reduce and Hadoop

Hadoop can be used to store data and Map reduce algorithms to perform tasks faster.

A simple word count program using map reduce, where during map stage a value and an associating value for the set of keys are generated. With the aid of words being the number of existence and the key of each word in every portion of the document is being calculated on a distributed file system. The Reduce stage, the general number of existences of the word in the aggregate of the dataset is gained. The

neurological computations are computed based on this approach.

Ensemble Empirical Mode Decomposition (EEMD) algorithm has made it possible to process neural signals as now it is compute-intensive and data-intensive.

- Apache spark

Machine learning and graph analysis for enlacing power in specific domains and to elucidate how handling pipelines applying it can be enhanced. Resilient Distributed Datasets (RDD) is the data stored in the Hadoop distributed file system in the Spark framework. The property of distributed storage and computation on distributed dataset and maintaining high-level interface along with writing of distributed code is made easier.

Graphics Processing Units (GPUs) plays here a key role in the area of neuroimaging and neuroscience. f-MRI in particular, both big data frameworks and GPU acceleration can make it useful where, increasing spatial and temporal resolutions and larger sample sizes lead to rapid increase in amount of data that has to be processed for a particular study.

- NeuroPigPen

The data partitioning method is very much essential for complete storing, processing, and analysing of enormous volumes of composite electrophysiological and for scalable data processing applications which can run fast, the parallel computing. To resolve the encounters from large scale electrophysiological signal, the combination of Apache, Hadoop and Pig developed a new toolkit knows as NeuroPigPen

To practice neurological signal data by previous data partitioning, data alteration and data handling performance, the NeuroPigPen toolkit was developed with several user defined functions in it with the help of Pig compiler.

- Hive

To handle the enormous data, hive was used to deliver a standard interface for database programmers that strictly match the SQL. Hive guides the Map Reduce work within Hadoop to recognize data with queries and a few commands. This simplifies the development of complex analysis that uses traditional database query steps.

III. Big Data in Bioinformatics

The new technologies of Big Data allows biologists to generate, manage and classify large amounts of data in the from of measurement in the images of physiological structures till genomic sequence. Measurements are performed throughout and beyond the components of central dogma in between sequence and structure. Such measurements include protein expression, metabolite concentrations, mRNA expression and transcription factor binding.

The amount of available data reduces the individual's performance ability. These computational algorithms from machine learning (ML) are efficient in identifying important patterns in large data anthology or compendia. ML falls in two broad classes namely Unsupervised Learning and Supervised Learning.

For both classes of methods the overall focus was an integrative approach that uses multiple measurements of various levels for example mRNA expression, miRNA expression, and transcription factor binding termed as deep data integration.

- Unsupervised Learning

The input is given as a set of unlabelled examples without predefined classes with the objective of discovering hidden signals within the data.

The challenge to discover molecular subtypes is best addressed with unsupervised methods. For example, use of k-means clustering to identify molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. They are also used to separate the human genome to different functional segments based on genome wide histone modification patterns from the "The Encyclopaedia of DNA Elements project".

- Supervised Learning

The input consists of a set of training samples with known labels. The algorithm analyses the training data and produces an inferred function which is used to classify new samples. It identifies the genes expressed in a single cell lineage from data measuring a complex mixture of cells in a solid tissue.

This method can precisely predict the expression level of genes in embryonic stem cells of mice (mESCs) from the knowledge of 12 transcription factors and 7 histone modifications. It can integrate levels of gene expression with the inputs given as TF binding and histone modification signals.

- Supervised Big Data Mining Without Programming

Large public data compendia contain substantial information about gene function. They are often used to answer questions about response of genes for a treatment or identify subtypes. They answer the question of interest and are uploaded to data repositories. Platform for Interactive Learning by Genomics Results Mining (PILGRM) is one such server in which the researcher defines a positive and negative standard, where positive standard represents genes involved in the process in which the researcher would like to discover additional players and the negative standard indicate genes that are not desired, which could be either highly specific or genes selected randomly.

IV. CONCLUSION

- PACS developed to store and access medical data, however one can easily miss a patient's data while being retrieved. While this is in its developing stages implementation of machine learning and pattern recognition approaches can help overcome these issues
- Hadoop can be used to store data and Map reduce algorithm to perform tasks faster
- Apache spark can make a parallel system and increase the processing speed 100 times faster than map reduce and can be useful in analysis with its concept of machine learning
- Pig and Hive can be useful in analysing neurological data
- Currently NeuroPigPen is the best toolkit for computing functional connectivity network and classification of seizure networks in epilepsy patients
- We can conclude that the selection of big data methods highly depends on its application Biological discoveries confront vast challenges in storage of data, analysis and processing which can be removed or reduced by endeavours performed by genetic scientists, experimental biologists and bioinformatics with the help of Big Data methods and algorithms.

V. REFERENCES

[1] Proceedings of the 12th INDIACom; INDIACom-2018; IEEE Conference ID: 42835 2018 5th International Conference on "Computing for Sustainable Global Development", 14th - 16th March, 2018 Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

[2] Big Data Bioinformatics: CASEY.S GREENE,1,2,3 JIE TAN,1 MATTHEW UNG,1 JASON H. MOORE,1,2,3,* and CHAO CHENG1,2,3,* PMCID: PMC5604462 NIHMSID: NIHMS868708 PMID: 24799088

[3] Big biological data: challenges and opportunities Yixue Li 1, Luonan Chen 2 Affiliations expand PMID: 25462151 PMCID: PMC4411415 DOI:10.1016/j.gpb.2014.10.001

[4] Big data in healthcare: management, analysis and future prospects: Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma & Sandeep Kaushik Journal of Big Data volume 6, Article number: 54 (2019) Cite this article|84k Accesses |32 Citations| 26 Altmetric| Metrics

[5] https://en.wikipedia.org/wiki/Apache_Hadoop

[6] https://en.wikipedia.org/wiki/Big_data