May-June 2020
ISSN: 0193-4120 Page No. 3685-3688

Analyzing the Data in Health Care Industry Using Big Data Tools

SP. Chokkalingam[1], G. Divya[2]
[1]Professor, [2]Assistant Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai
[1]cho_mas@yahoo.com, [2]mailtodivya16@gmail.com

Article Info
Volume 83
Page Number: 3685-3688
Publication Issue:
May - June 2020

Article History
Article Received: *19August 2019*
Revised: *27 November 2019*
Accepted: *29 January 2020*
Publication: *12 May 2020*

Abstract
Health Care Industry is one of the important sectors in current scenario. Terabytes of data are processed and managed for future prediction. Applying Traditional Approaches on large volume of data for interesting patterns, prediction is tedious process. In this paper, various big data tools is discussed and analyzed with healthcare data set. Organizing and managing the data in hadoop framework in an efficient manner. Map reduce is technique in the hadoop framework outperforms efficiently in measure of speed and accuracy. The experiment is carried out on various kinds of benchmark datasets and the performance of the big data tools is evaluated.

Keywords: *Surveillance camera, Object Detection, Deep Learning*

1. Introduction

In today's world, data is measured in quintillion byte of data per day. These data are comes from everywhere like sensor information from satellites, Social media like twitter, facebook, linkedln where people posting their views ,pictures and videos etc, In health care Industry data from x- rays EMR records clinical test reports which ranges in trillions per day. Generally Big data refers to 3 V's namely volume, variety and velocity. Volume -the size of the data for storage. Variety- These data are comprised of both structured data and unstructured data. Structured data includes the data available in the form of tuples, documents and are relevant for analysis. Unstructured data determines images, videos, and blogs. Data from multiples sources are mostly of unstructured data which is difficult to incorporate those data in to relevant shape. Velocity- How the data are processed in stipulated time.

1.1 Health Informatics

Big data in healthcare industry is most research topic area with data stream from various resources at every fraction of time. These data may be images, clinical reports, physician reports provide insights which are used for getting informative and interesting patterns. This pattern discovery is used for futuristic prediction analytics for prevention of diseases to increase the lifetime of patients. Data from multiple sources as shown in the figure 1 are huge in volume so it is difficult to maintain and processing the interesting for predictive analytics is very challenging task. Data analytical techniques such as statistical modeling, artificial intelligence, predictive analytics, machine learning techniques are induced to get efficient patterns.

Sujatha et al has used a method to analyse the report using statistical method to make a report on death rate on chronic disease . Various parameters were used in this research work. The next section introduces about related work carried out on data analysis model.

2. Related Work

This section represents about review of challenges, framework, techniques used in the big data with related to healthcare. Rituchauhan et al designed a robust model for healthcare system. This model is computed for effective patient taken care diagnosis.

Architecture of big data in healthcare systems is shown in the figure 2. The architecture will collect the information from various hospitals. Those information is in the form of sensors data, mobile data clinical reports and those information will be processed by big data platform like hadoop, hive, and produce interesting patterns. Redundancy, fault tolerance is an issue occurs in this framework. He concluded data privacy as a future scope of hiswork

Secured big data architecture is proposed by Archenaa et al to maintain the heterogeneous data from EMR(Electronic Medical Records). "Big data in health informatics" provides a general background on BigData in Health Informatics by Matthewherland et al. He

May-June 2020
ISSN: 0193-4120 Page No. 3685-3688

generalized that big data tools and its techniques for the health care analytics data collected at multiple levels. Multiple level of questions are described on human-scale biology, clinical –scale and epidemic- scale. Using Patient data, Physicians were able to predict the decisions using hadoop framework.. Bennett et al says there is vast gap between actual clinical care and research in clinical used in practice. Predictions are made mostly on general information based on the previous reports given by the experts who produce accurate, reliable and efficient in healthcare systems. Time consuming is very less when compared to machine learning algorithms.

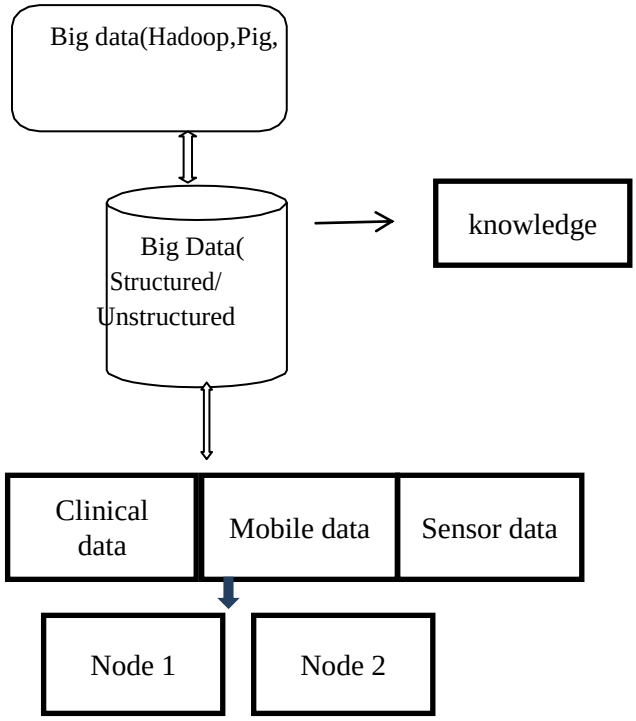


Figure 2: Architecture of Big Data in Healthcare Systems

Haferlach et al. says about gene expression formula used as profiling classifier to place patients into 18 different subclasses of either myeloid or lymphoid leukemia. He formulated the study based on 3,334 patients where 2,143 patients are used as training data set and the rest is used for testing. 54,630 gene probe samples. He uses pair-wise classification design using Difference of Quantile Normalized Values(DQN approach to find out the mean difference between perfect match and mismatch. Using this approach 99.8% for the classification with 14 sub classes in terms of lymphoid and myeloid. Of 6 and 8.

Zhang et al. discussed about real time predictions using data streams with two categories prognosis and diagnosis by a system called "Clinical Support System". This system uses an algorithm called Very Fast Decision Tree(VFDT) for streaming the data in sequential manner can make prediction both diagnostically and prognostic. Zhang et al.'s. method is not tested on real world data and would need to be before its usefulness can be determined and can be legitimately compared to IBM's method.

Ashish et al. created a message board forum in social media like twitter, facebooketc and that platform called Smart Health Informatics Program (SHIP). In this platform, patients share their medical experiences in the message forum. Around 50,000 discussions including 400,000 post from different websites like inspire.com, medhelp.com. Pipeline structure is used taking all the discussions and posts from the message forum and store them in a database for retrieval of information. Lucene is an java based search engine library by optimizing for indexing, dynamic score, boosting and local caching. Table 1 shows test cases with 5 expression of interesting categories includes personal experience, information, support, advice and outcome in terms of precision and accuracy. Saravana Kumar et al. had done analysis for diabetic patients using predictive algorithm in Hadoop environment. This approach provides types of diabetes, complications and types of treatment to cure the patients in advance with minimal cost. Figure 3 shows the architecture of predictive analytics using healthcare system explains about the data from various fields are collected extract relevant information and then it's processed in to Hadoop platform. Using machine learning algorithm like clustering , classification, Association rule mining for prediction and then it generated as reports and shared to server(S1,S2) and distributed to each node(N1, N2,N3,N4). This design yields best result when compared to previous approaches

May-June 2020
ISSN: 0193-4120 Page No. 3685-3688

Table 1: Results of SHIP

Category	Precisions	Accuracy
Personal experience	0.87	0.82
Advice	0.91	0.62
Information	0.93	0.91
Support	0.89	0.90
Outcome	0.80	0.58

Keith C.C. Chan et al. Explained that Big data measures in the range of volume. The data is related to pharmaceutical like gene data, molecular data, ECG data, and drug data. Even the data are collected from online health forum like social media which helps to perform sentimental analysis and prediction of earlier disease. This helps to create awareness among people who are fear of disease symptoms.

To optimize the classification a portion of the images from the entire dataset is taken for training process with a predefined learning rate for CNN. The abnormality of the lung images is obtained by wrong values(odd) in the feature vector. Generally, feedback or backpropagation are used for fine tuning the layers in the network.

The classification accuracy is improved by finetuning the data. . In order to verify the performance (quality of the image) of the noise removal function, the performance metrics are calculated such as MSE, SNR, PSNR, and SSIM between the input and pre-processed images

Data acquisition method is a function which retrieves, collect and prepare the image data for a specific process to provide an expected outcome. Most of the medical images are in the form of analog images and it will be represented in discrete values.

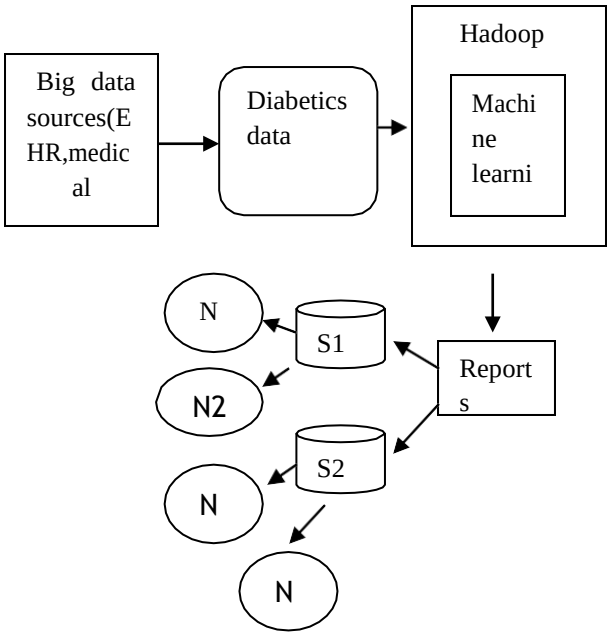
Ho Ting Wong et al says that the big data plays important role in health care industry. He has framed novel approach for cost effective which help to predicts the patient report as earlier. Big data, by contrast, provides flexibility in analyzing a data set from different perspectives. Moreover, because governmental organizations already maintain a significant amount of big data, research using big data can be conducted at a very lowcost.

3. Big Data Analytics-Tools and Technology

A traditional method of data mining is not suitable processing of big data. NOSQL database were introduced to handle unstructured data. There are many tools and technologies were introduced for Real time analytics. . Real time analytics focused on specific time to analyze and process the data efficiently. Hadoop is the base storage developed by apache for distributed processing. Several tools like spark, Storm and kaftka used the data as a pattern analyzed on clusters. In this section, Each tool will be described indetail.

- **HDFS:**

It is used to store large volume of data and divides the data in to number of blocks. Each blocks three copies of data or file in the respective servers. It is broadly classifies into three nodes Name node, Data node and edge node. Name node acts as master node and it tracks all the information or records. It identifies the location where the file has been stored. Data node acts as slave node for doing particular task instructed by master node.



- **MapReduce**

Map reduce is an programming model which uses divide and conquer approach. It has two functions map() and reduce() where map() perform on master node where the tasks are divided into sub tasks. Each input acts key-value pair and merges the data using intermediate key is done in reduce function.

Map Reduce Components:

Name Node: manages HDFS metadata, doesn't deal with files directly.

Data Node: stores blocks of HDFS—default replication level for each block:

Job Tracker: schedules, allocates and monitors job execution on slaves—Task Trackers.

Task Tracker: runs Map Reduce operations

It works on top of the hadoop distributed file system and its infrastructure has three components driver program, cluster manager and worker nodes. User can run this application in programming language like R, python, or scala to analyze the complex data. Bilal et al. says spark is generally used for graph database and it process the waste data and analyze its efficiently

3.1 Storm

Storm is fault tolerant computing system for real time processing .It is similar to hadoop cluster but the only difference, hadoop uses topologies to process the message

May-June 2020
ISSN: 0193-4120 Page No. 3685-3688

between the worker nodes and master node rather than map reduce. It uses zookeeper as minion worker instead of hadoop cluster to process the data in the form of acyclic directed graphs. It is used to detect in crucial event through the processing of twitterfeeds.

3.2 Kaftka

It is an open source technology introduced by Apache. Message broker is important component able to transfers the number of message from thousand clients. It is a message queue system for transaction kept as log to make as best infrastructure to process multiplestreaming.

4. Conclusion

In recent years, data in health informatics is challenging task for predictive modeling. In this paper, we surveyed the various data from multiple sources using different techniques and the results were analyzed. Different big data tools and technologies also had been discussed for real time analytics in health care systems. The design of various predictive analytics system gives better view of Research on using these tools and techniques. Health Informatics is critical, because this domain requires a great deal of testing and confirmation before new techniques can be applied for making real world decisions across alllevels.

References

[1] Rituchauhan, Rajesh Jangade "A robust dataModel for Big healthcare data Analytics", IEEE Inc, 2016

[2] Archenna.J, Mary Anita.E.A "A survey of Big Data in Healthcare and Government", Elsevier Procedia conference2015

[3] Bennett C, Doub T "Data mining and electronic health records:", selecting optimal clinical treatments inpractice. CoRR abs/1112: 1668(2011)

[4] Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Béné MC, De Vos J, Hernández JM, Hofmann WK, Mills KI,Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Wm Liu, Williams PM, Fo R (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia:report from the international microarray innovations in leukemia study group. J ClinOncol 28(15):2529–2537.

[5] Zhang Y, Fong S, Fiaidhi J, Mohammed S "Real- time clinical decision support system with data stream mining",. J Biomed Biotechnol2012.

[6] Ashish N, Biswas A, Das S, Nag S, Pratap "The Abzooba smart health informatics platform (SHIP)™– from patient experiences to big data to insights", 2012.

[7] SaravanaKumar N M, Eswari.T, Sampath.P, Lavanya.S "Predictive Methodology for Diabetic Data Analysis in Big Data",ElsevierProcedia conference2015.

[8] Keith C.C. Chan. "Big data Analytics for drug discovery", 2013

[9] Ho Ting Wong, Qian yin, Ying Qi, " Big data as a new approach in emergency medicine Research", Elsevier journal of Acute Disease, 2015.

[10] Matthew herland, Taggi M Koshgoftaar, Randall Wald "A Review of data mining using Big dData In Health informatics", Journal of Springer.2014