A Machine Learning Based Massive Health Care Data Analysis

Raol Priyanka Ajaysinh[1], Saket Swarndeep[2]

[1]*M.E Student, Dept. of Computer Engineering, LJ Institute of Engineering & Technology, Ahmedabad, Gujarat, India*
[2]*Assitant Professor, Dept. of Computer Engineering, LJ Institute of Engineering & Technology, Ahmedabad, Gujarat, India*

ABSTRACT

***Abstract*—**Health Care is the maintenance or improvement of health. There are multiple processes going on within health sector. This process may help medical practitioners and data analysis helps in predicting the health related problems or disease. As the amount of healthcare related data is increasing every day, it is believed that extracting knowledge by data analysis process is essential. Big data analysis is not just an opportunity but a necessity. While many industries are successfully performing big data analysis, healthcare providers and investors are actively investing in data analytical capabilities. This move will help them to have a better understanding of the complexity of changing healthcare environment. This paper summarizes the role of big data analysis and prediction in healthcare and few machine learning techniques related to healthcare.

***Keywords*-** *Health Care, Machine Learning, Big data, Data analysis, Prediction.*

---

1. INTRODUCTION

The healthcare industry historically has generated large amounts of data. While most data is stored in hard copy form, the current trend is toward rapid digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as 'big data') hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management.

By definition, big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed. The totality of data related to patient healthcare and well-being make up "big data" in the healthcare industry. For the big data scientist, there is, amongst this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives and lower costs. Thus, big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions.

The potential for big data analytics in healthcare to lead to better outcomes exists across many scenarios, for example: by analyzing patient characteristics and the cost and outcomes of care to identify the most clinically and cost effective treatments and offer analysis and tools, thereby influencing provider behavior; applying advanced analytics to patient profiles (e.g., segmentation and predictive modeling) to proactively identify individuals who would benefit from preventative care. This article provides an overview of big data analytics in healthcare [6].

Data is considered to be the new warehouse for this era. Like the oil when it is unprocessed it is barely of use. But using the different analytical methods we can extract the precious information/knowledge hidden in it. Thus Big Data has a potential to make a big impact in healthcare industry

Massive data in healthcare domain provide us the huge ability to do the predictive analysis of these data and come up with solution of new problem which might be derived from the existing solution of the known problem. Through big data analytics we can parallelly process large number of information and find out the association among them which can help us to resolve the problem and provide us the effective solution of hidden problem. Healthcare domain can effectively use this analytical approach to cut down their cost or be better prepare with the needed equipment or resources, which might be needed as per the environment. For example, if there is an outburst of any disease in an area which has earlier occurred in different parts of the world, then using the analytical engine to predict the solution and spread of disease which might happen and thus be better ready with their resources and medicine which might be needed to solve the epidemic problem. Also hospitals and clinics can use the data to analyze and predict the patient's preferences, which opens a path to possible business. Predictive Analysis where we use various statistical method, machine learning techniques and, data mining approaches to process, analyze data and predict the outcome for the unknown bag of data. Healthcare domain is still in early stage to take up the new possibilities, which can be offered by big data solution and use it to do effective decision-making [7].

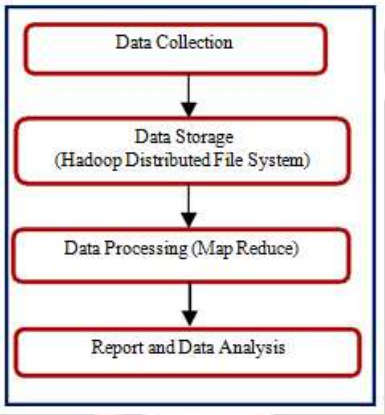

Fig-1: Big Data architecture [7]

Big data challenges in Health care

- Extracting knowledge from complex or unstructured data set.
- Understanding unstructured clinical notes in the right context.
- Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.
- Analyzing genomic data is a computationally intensive task and combining with standard clinical data adds additional layers of complexity.
- Big data analytics platform in healthcare must support the key functions necessary for processing the data.
- Real-time big data analytics is a key requirement in healthcare.
- The lag between data collection and processing has to be addressed.

2. Related Work

In [1] author has proposed generic architecture for big data healthcare analytic by using open sources, including Hadoop, Apache Storm, Kafka and NoSQL Cassandra. Big data computing is a new trend for future computing with a large-scale data set and can be divided into two paradigms: batch-oriented computing and real-time oriented computing (or stream computing). Batch computing is in general efficient in processing high volume data. The data are collected, stored, and processed in batches to produce the results. Apache Hadoop (Hadoop) [1] is an example of batch-oriented computing. In contrast, stream computing involves continual input and outcome of data. It emphasizes on the velocity of data. Stream computing paradigm has been used in many applications such as Tweeter Storm, Yahoo S4 and Microsoft Time Stream. Big data stream computing (BDSC) provides Real-time computing, to gain useful knowledge from big data. In the healthcare system, data are collected from many sources such as clinical data, genomic data, and personal behavior data that emphasizes the variety characteristic of healthcare data. Those data spread among multiple clinical centers, hospitals, health insurers, research labs, government entities and

continuously grow at an overwhelming speed, which indicates the velocity of the big data in healthcare. Furthermore, each of these data repositories are growing in complexity, not only by the volume, velocity, variety, but also by veracity due to the inconsistency of data. Nowadays, BDSC plays an important role in big data analytics to obtain the value of big data in health care. Author has proposed a generic architecture for BDSC in health care, both in real-time and batch-oriented computing. Multiple streams of messages that are generated from Apache Kafka's producers are processed at Storm, then stored in a distributed storage NoSQL Cassandra system.

In [2] author has shown few machine learning algorithms in Health Care environment

1. Multi-Layer Perception (MLP)

MLP is a type of feed forward Neural Network. It consists of input layer, one or more hidden layer and an output layer. Each Layer has number of Neurons that connect a layer to another. Neuron in the input layer is the number of input features (i.e. disease or Symptoms) in the data set. Each data instance is fed into the input layer then the weighted outputs of these input neurons pass to the hidden layers. Once the weighted outputs pass through all the hidden layers, they are fed into the output layer where there are as many output neurons as the number of output features.

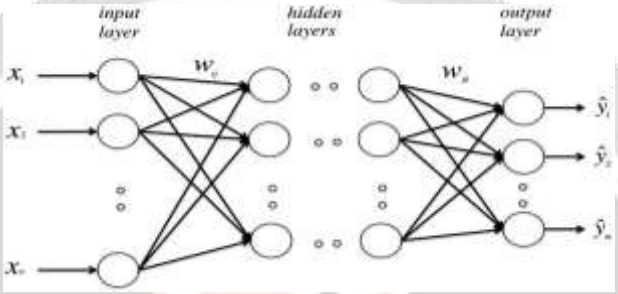

Fig-2: Basic structure of MLP [2]

2. Support Vector Regression Machine (SVR)

SVR is an extension of SVM that can be applied to regression problem. It attempts to minimize the error between the target and prediction, by finding optimal hyper plane such that the prediction error for each training data does not exceed a threshold.

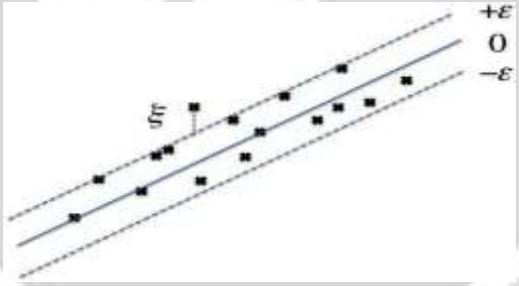

Fig-3: SVR's ε precision and slack variable [2]

3. Generalized Regression Neural Network (GRNN).

GRNN is an instance based learning algorithm introduced by Specht (1991). GRNN predicts by first calculating the distance between the new input data instance and training data instances.

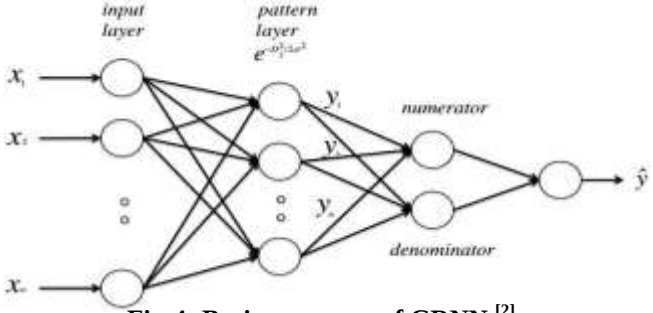

Fig-4: Basic structure of GRNN [2]

4. K Nearest Neighbour Regression (KNN).

kNN predicts the function value $\hat{f}(x)$ given a new input data using the *k* closest neighboring data in the training dataset *T.* KNN helps in finding the related nearest disease which will help in prediction of disease. *k* is the closest neighboring data.

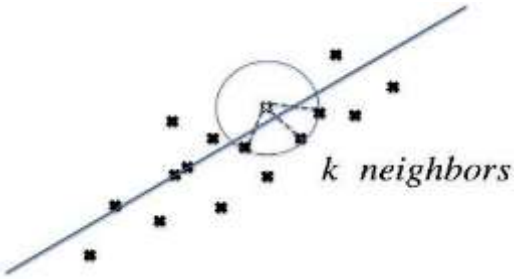

Fig-5: KNN scheme [2]

In [3] author has explained the Student Clinic Data with the help of Mapper and Reducer Function pseudo code and the work flow of map reduce technique. Author has select monthly-based data including student ID, basic clinic service fee, medicine fee and special medical treatment fee, and amount of allowed reimbursement. In the simple test, author has only calculated the total values. For example, in the test there are 24 clinic service fees in the sample data records, decompose the data into 4 data sets and construct 4 vector arrays. By using Mapper and Reducer pseudo code MapReduce mechanism with map() and reduce() functions can be calculated by MapReduce mechanism with map() and reduce() functions.

In [4] author has proposed a massive data management and analysis solution based on Hadoop to archive better performance, scalability and fault tolerance. Two different data analysis methods based on MapReduce and Hive are proposed. Hadoop contains a computational paradigm named MapReduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. On top of HDFS, Hadoop has an open-source, distributed, column-oriented database named HBase. HBase provides Bigtable-like capabilities for Hadoop. . Hive supports queries expressed in a SQL-like declarative language - HiveQL, which are compiled into map-reduce jobs that are executed using Hadoop. In addition, HiveQL enables users to plug in custom map-reduce scripts into queries. Hive actually lightens up the workload of data analysis by provide languages that can be easier to implement.

In [5] author has explained the naïve Bayes method to classify new cases or patients as they arrive. Model developed by author can predict the state of person's health on providing the symptoms as input and analyze the pattern of disease growth. Knowledge was gained from a bulk of unstructured data sets and implemented into a web system. Model is based on Naïve Bayes classification algorithm. Naïve Bayes algorithm recommends diseases very efficiently resulting in reduced treatment costs.

3. MACHINE LEARNING METHODS

1)Regression

Regression is a measure of the relation between the mean value of one variable and corresponding values of other variable. One wishes to find some simple pattern in the data – a functional relationship between the X and Y components of the data. For example, one wishes to find a linear function that best predicts a baby's birth weight on the basis of ultrasound measures of his head circumference, abdominal circumference, and femur length.

2) Neural Network

Networks of non-linear elements, interconnected through adjustable weights, play a prominent role in machine learning. They are called neural networks because the non-linear elements have as their inputs a weighted sum of the outputs of other elements— much like networks of biological neurons do. These networks commonly use the threshold element which we encountered in study of linearly separable Boolean functions.

3)Support Vector machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. When data are not labeled, supervised learning is not possible, and an

unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering.

4)K-Means Clustering
k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

5)Decision Tree
One of the most popular machine learning algorithms in use today, this is a supervised learning algorithm that is used for classifying problems. It works well classifying for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

6)Random Forest
A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

4. PROPOSED WORK

In this section, we have described proposed model for better healthcare prediction. Nowadays, accurate prediction is the major issue to overcome this issue; the proposed model will be useful for improvement of accuracy in the field of healthcare. In proposed model pre-processing is applied on unstructured data to convert it into structured format. NLP module is applied on structured data and symptoms are matched and mapped with the disease. If matching fails then it will go to the learning module for accurate prediction.

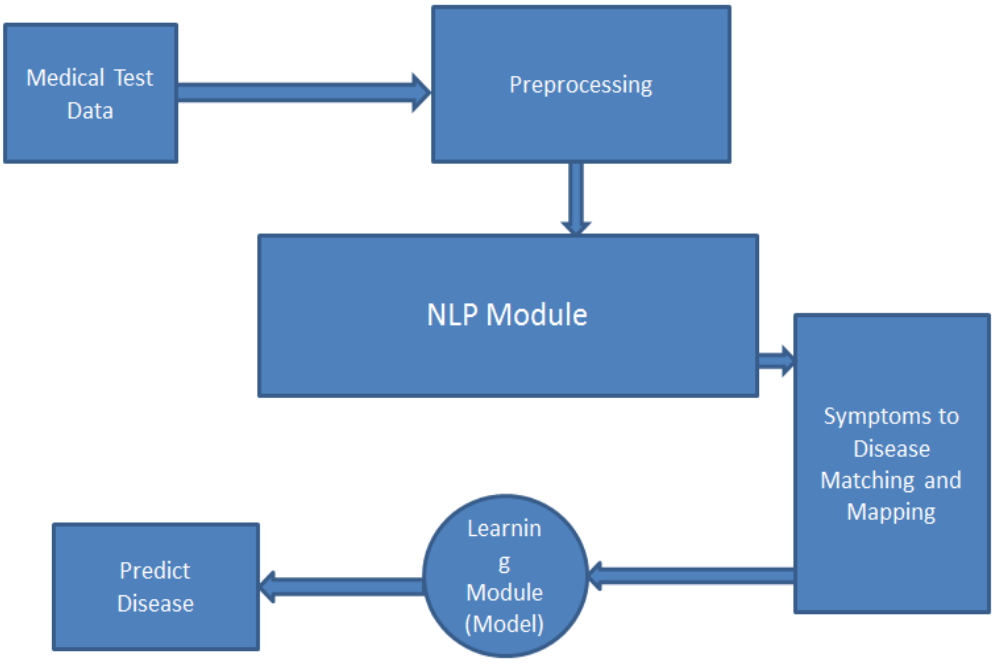

Fig-6: Proposed Model

5. RESULT ANALYSIS

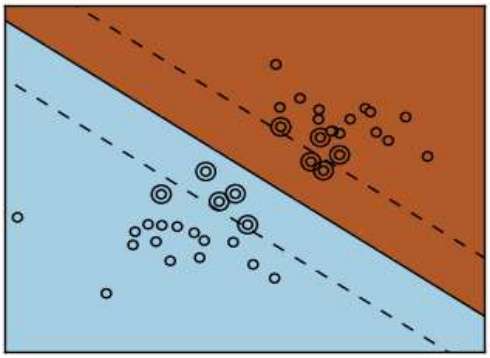

Fig-7: Untrimmed region

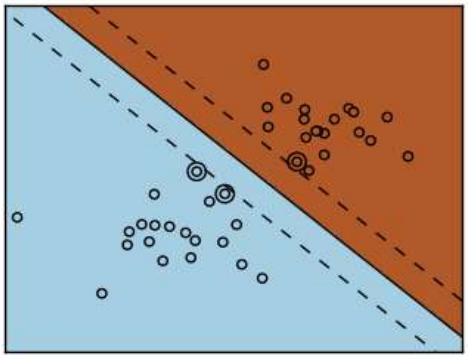

Fig-8: Trimmed region

Dotted line indicates hyper plane and dots indicate the predicted data. The generated graph contains hybrid svm-nb approach. Fig-7 indicates the untrimmed region of svm-nb hybrid approach algorithm and Fig-8 indicates the trimmed region of svm-nb hybrid approach algorithm.

6. CONCLUSION AND FUTURE WORK

There are urgent demands for healthcare professionals to work on analysis and prediction of healthcare disease. Health care system should be undertaken in the same spirit of continuous improvement. Analysing the symptoms and matching it with disease for prediction of better healthcare. Proposed model can predict the health related disease more accurately and faster, based on the already trained data. The hybrid machine learning algorithm will learn according to the environment and will help in predicting the data more accurately.

7. REFERENCES

[1] Van-Dai Ta, Chuan-Ming Liu, Goodwill Wandile Nkabinde**,"Big Data Stream Computing in Healthcare Real-Time Analytics",** IEEE, 2016, pp. 37-42, DOI:10.1109/ICCCBDA.2016.7529531

[2] Junheung Park, Kyoung-Yun Kim, **"COMPARISON OF MACHINE LEARNING ALGORITHMS TO PREDICT PSYCHOLOGICAL WELLNESS INDICES FOR UBIQUITOUS HEALTHCARE SYSTEM DESIGN",** IEEE, 2014, pp. 263-269, DOI: 10.1109/IDAM.2014.6912705

[3] Jun Ni, Ying Chen, Jie Sha, and Minghuan Zhang, **"Hadoop-based Distributed Computing Algorithms for Healthcare and Clinic Data Processing",** IEEE, 2015, pp. 188-193, DOI: 10.1109/ICICSE.2015.41

[4] Hongyong Yu, Deshuai Wang, "**Research and Implementation of Massive Health Care Data Management and Analysis Based on Hadoop**", IEEE, 2012, pp. 514-517, DOI: 10.1109/ICCIS.2012.225

[5] Weider D. Yu, Choudhury Pratiksha, Sawant Swati, Sreenath Akhil, Medarametla Sarath**,"A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment,"** IEEE, 2015, pp. 75-86, DOI: 10.1109/COMPSAC.2015.335

[6] Wullianallur Raghupathi, Viju Raghupathi, **"Big data analytics in healthcare:promise and potential"**, NCBI, 2014, DOI: 10.1186/2047-2501-2-3

[7] Dharavath Ramesh, Pranshu Suraj, Lokendra Saini, **"Big data Analytics in healthcare: A Survey Approach"**, IEEE, 2016, pp. 1 – 6, DOI: 10.1109/MicroCom.2016.7522520

[8] https://www.siam.org/meetings/sdm13/sun.pdf_accessed on 22/03/2017,05:30PM

[9] https://www.dezyre.com/article/top-10-machine-learning-algorithms/202_accessed on 25/03/2017,06:00PM