The Predictive Analytics of Machine Learning and Big Data for Value-Based Health Care

Baljinder Kaur
Member, Chandigarh University

***Abstract*—Normally healthcare is said to be information rich and to extract hidden data from such information-rich industry is difficult. It becomes necessary for healthcare informatics to deal with the advancement in technology and big data. This can be done by changing three core areas namely, to manage record electronically, integrate big data and computer-aided diagnosis. To resolve the mentioned challenges machine learning provides a range of techniques, algorithm, and different frameworks. This chapter focuses on handling big data of health care to predict disease using machine leaning approaches.**

***Index Terms*—Predictive Analytics, Machine learning, Big data, Health care.**

I. INTRODUCTION

Machine learning provides a wide range of algorithm to perform analysis of data to explore hidden patterns from the huge amount of data. Numerous machine learning algorithms are available which is used to learn from given data and helps to explore hidden patterns .Data mining is one of the most important aspects of machine learning. The main aim of machine learning is to generate learning algorithms which are efficient as well as fast to make predictions on given data set. Machine learning models are trained from available data in case of data analytics and then trained model is used to make predictions for new dataset.

Real life application of Machine learning includes Healthcare, marketing, financial services, prediction while travelling, social media services, extraction, image and speech recognition, suspicious activity detection from CCTV.

II. BIG DATA AND MACHINE LEARNING

It's been said more than 3 quintillion bytes of data is generated everyday .This will increase day by day ,so there is need to deal with such voluminous data. Some of the facts of big data are:

- In every 60 seconds more than 16 millions text messages are sent
- In every 24 hours nearly billion photos are shared on google photos.

Here challenge is not only to store such voluminous data but also to manage and to perform data analysis is another big confront. In other words for human experts big data will be surplus to tackle and finding correlation from such big data is another challenge .Big data can be explained easily with 5 V's

1. Volume: This represent amount of data being produced every second.
2. Velocity: This represent rate at which data is produced.
3. Variety: It focus on diversity of data produced.
4. Veracity: It deals with quality of data.
5. Value: This is meaningful information received from data.

Big data Analytics provide us ability to obtain knowledge by analyzing big data to make a better decision .It uses data mining to derive the relationship between different unknown attributes by applying machine learning algorithm , statistics. Big data Analytics[15] consist of an approach known as KDD(Knowledge Discovery in database) which is further used to extract pattern from huge data. Data integration, data storage, preprocessing, feature selection and data processing are stages of KDD process. Data integration is said to be the one of the most challenging stage due to heterogeneous nature of data. As data can be in structured form or unstructured form as well [16]. To deal with voluminous data there is need of a efficient approach.

Now consider one organization, gathering huge amount of information and finding clues from the this voluminous data that will help to make better, faster and smart decisions. Here comes the role of machine learning algorithms in finding clues from the big data

.As it will be difficult for traditional database system to deal with such voluminous data.

Machine learning model is said to be better if we are having huge amount of data. More data we are having in learning stage, more it can perform efficiently in testing stage. After learning from big data, model is used to find clues to make smarter decisions. Model trained with machine learning algorithms doesn't work properly if amount of data available for learning is not adequate[21].

A. BIG DATA TOOLS

Numerous Big Data tools are available which helps to retrieve the meaningful facts from the voluminous data for decision making by minimizing the time of analysis. Big Data tools like Hadoop, MongoDB, CartoDB, Tableau are compared on basis of Big Data Analytic process [7].Comparison of different Big Data tools is discussed below in table 1:

S. No	Stage	Tool	Usage	Data Extraction Way
1	Data Storage	CloudEra	Cloudera is said to a secure and modern tool for providing flexible services using cloud	It uses Apache Hadoop to develop big data applications along with recent open source tools.
2		Hadoop	This tool is open source used for storing distributed data of voluminous dataset. Easy to deal with structured as well as unstructured data.	Map Reduce, YARN,HDFS
3		MongoDB	The main objective of MongoDB is for data storage which is voluminous itself. This tool works well for Real Time Mining. Also it has an ability to deal with data which is frequently changing.	This is used to give a single view for different systems. It helps in managing content and delivering applications
4		Talend Open Studio	Talend is open source tool used for Big Data.Also it provide an oppurtunity to connect with other Big Data distributions like Cloudera, MapReduce.	Talend is used to make business decisions by providing high scale and fast data processing.
5	Data Processing	Statwing	Statewing tool is used to perform data analysing at new level.It has an ability to explore voluminous data quickly and generate charts as an result of analysis.	Statwing is used to translate the result in English so that if someone not having knowledge of statistical analysis canstill get fruitful benefits of the result.
6		Qubole	Qubole is tool used for both structured as well as unstructured data.It is said to be Hadoop platform based on cloud.	This tool provide an interface to user for analyzing data even in absence of Hadoop.It provides the oppurtunity to user to access any tool any engine, any cloud.
7		BigML	This tool make things easy for machine learning and provides an interactive interface to user.	Basically this tool is mainly used for predictive analytics
8	Data Visualization	Tableau	This tool is mainly used for buisness purposes , so that buisness users can find valuable facts quickly in their dataset to make better decisions .	Tableau removes the he dependency on advanced query languages as it provides the graphical analysis interface that will make it simple for stakeholders to deal with big data
9		CartoDB	Basically it is having cloud computing framework which is used for making maps.	This tool is used to represent location without programming. This tool is also known as location intelligent tool.

III. LITERATURE REVIEW

If you are using Word, use either the Microsoft Equation Editor or the MathType add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation or MathType Equation). ―Float over text‖ should not be selected.

1. J.L Berral-Garcia in their paper represented machine learning algorithms which are frequently used for big data analytics .Algorithms used in this paper are Decision Tree ,K-means, KNN, Bayesian, SVM. Along with this several framework like Map-Reduce, Microsoft's Azure-Ml are discussed. Machine learning tools like WEKA , Kibana are used to implement above mentioned learning algorithms.[1]
2. M.U. Bokhari et al. represented architecture for analyzing big data which is having three layers .In this architecture of three layer ,one is for gathering of data , second is for data storage and third is for data analysis and generation of report.[3].Different high speed nodes are required at data gathering layer for collecting huge data .While in second layer named as data

storage layer HDFS is used for efficient storage of big data .While in third layer named as analysis layer machine learning algorithms are used for getting clues or patterns from the big data stored by second layer .

3. M.R. Bendre et al. in their paper[4] focused on big data in precision agriculture. In this author discussed that usage of big data helps to uncover wide range of farming problems by analyzing large amount of data .To process big data in this paper author used map reduce framework .Along with this data prediction of farming problem is done by using linear regression. For this research work data is collected from KVR (Krishi Vidyapeeth Rahuri)for training stage .Trained model is used for forecasting which further helps for making a better decisions in field of agriculture.
4. M.Chen et. al in their paper predicted disease by using machine learning over big data. In this paper author experiments the prediction model on data which is real and collected from central china(2013-2015).A model is used to deal with incomplete data by reconstructing missing data .In this research CNN(Convolution Neural Network) is used to deal with unstructured data. Machine learning algorithms like Naïve Bayesian, K-Nearest Neighbor ,Decision tree are used for prediction of disease.[5]
5. Qiu et al.[2] focused on signal preprocessing using machine learning algorithms for Big Data. In this paper author discussed five critical issues with respect to big data : Variety of data ,speed of data ,incomplete data ,high volume of data ,uncertain data ,data having low value density).In this research author also discussed different learning techniques from big data.
6. Abd et al. [11] applied machine learning approach to E-medication system to diagnose Sickle Cell Disease with help of Smartphone .An intelligent system is proposed by an author by providing self care and monitoring system with the application embedded in their smart phone.
7. Now a days Big Data analytics is used in different sectors to perform analysis of Big Data .In this paper author[13] proposed data pipeline based on Cloudera - Hadoop to perform analysis of data.In this US stocks are used as an input for analysis to predict gains on daily basis . In this paper Apache Hadoop framework is used to handle Big Data and prediction is done using Machine learning module of Spark.
8. Mehdi Assefi et. al focused on platform named as Apache Spark MLlib to perform big data analytics as quintillion bytes of data is generated on daily basis[14].Also Big Data analytics is becoming trending day by day with increase in data. Apache Spark MLlib offers numerous admirable functionalities for machine learning tasks. Different experiments were performed in this paper to analyze the qualitative as well as quantitative features of the platform.
9. Surabhi Dwivedi and Kumari Roshni focused on a recommended model for big data in the field of education[17].In this paper students were given guidance to select an appropriate elective course as per their skills and past performance. To generate recommendations author used mahout machine learning library over Hadoop . As per this research the machine learning approaches and big data approaches are quite helpful to perform analysis of voluminous data and finding meaningful results.
10. Yin zhang Qiu and Meikang Qiu in their paper discussed about cyber physical system [19] to provide health care to the patient using Big Data and cloud.Along with this author introduced a new logical method to provide services related to health care . This proposed model contains three layers named as data collection layer , data management layer and data oriented service layer. This paper highlights the benefits of using Big Data techniques in the field of healthcare.
11. Kobashi et. al.[18] presented an approach to predict post operative implanted knee function .TKA(Total knee arthroplasty) is the knee surgery which is very common now a days. But difficult task is to select appropriate TKA implant for the patient as there are different types of TKA implant available.

Author used machine learning approach to make prediction of post operative knee function in this paper. Principal component analysis is used to extract features and perform mapping from pre operative to post operative space.

IV. ALGORITHMS

In field of healthcare there is vast amount of data. In data analytical process data handling and processing is the most difficult part .Big data tools and algorithms are required to deal with such voluminous data .Machine learning algorithms are used for data

analysis so that meaningful patterns can be extracted from the data . The machine learning approaches are discussed below:

1. Artificial Neural Network(ANN) : ANN is a approach based on working of biological neurons.This model is also named as multi layer perceptron model to analyze the working of neuron .This model mainly contains three layer named as input layer to receive input ,hidden layer which is also known as middle layer and output layer to generate an output .Hidden layer in center can be multiple also .In ANN feed forward neural network multiple hidden layers are used along with one target .. Activation function is used in hidden layer of model to compute the result .There is no direct relationship between input and the target as shown in given Figure 1. Convolution Neural Network is widely used for the cases where images are taken as an input. In ConvNet neurons are arranged in different pattern as compared to regular neural network. In this neurons are arranged in three dimensions including height, width and depth. In ConvNet a three dimensional image is converted to a single vector .Along with all the merits NN also have some demerits as for large neural network it requires high processing time.

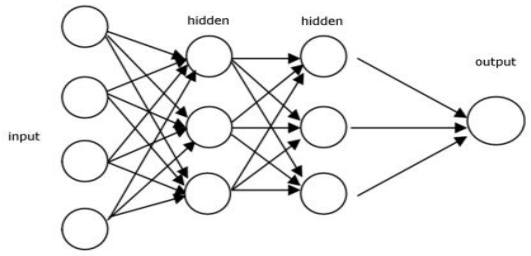

Fig 1: Feed forward Neural Network

2. Naive Bayes
   Naive bayes is one the most important approach of classification. This algorithms assumes all attributes are independent of each other in a data set .Changing value of one does not affect other attribute. Also naive bayes doesn't require much data to estimate outcome. The only basis of Naive Bayes is Bayes theorem

- Bayes theorem

Bayes theorem uses prior knowledge to computes conditional probability of event .Basically conditional probability is the one which reflects happening of one event on probability of other event. Terms related to bayes theoren are:

Prior probability: This is the original probability of an event before referring to any additional information obtained.

Posterior probability: This is probability computed on basis relevant information. It is written as :

$$P(m|n) = \frac{P(n|m) * P(m)}{P(n|m) * P(m) + P(n|\neg m) * P(\neg m)}$$

where

P(m) and P(n) are prior probabilities of m and n respectively

P(n/m) is posterior probability of n given m

P(m/n) is posterior probability of m given n

P(n/¬m) is probability of n given m is false

P(¬m) is probability of m being false

For feature vector x1,x2....xn and classes C1,C2,... ,Ck , Bayes theorem can be written as:

$$P(C_j|x_1,x_2,....x_n) = \frac{P(x_1,x_2,....x_n|C_j) * P(C_j)}{Px_1,x_2....x_n}$$

Using assumptions of independence

$$P(x_i|C_j,x_1,\ldots,x_n) = P(x_i|C_j),$$

for all i .Further this can be written as:

$$P(C_j|x_1,x_2,.....x_n) = \frac{P(C_j)\prod_{i=1}^{n} P(x_i|C_j)}{P(x_1,x_2,\ldots,x_n)}$$

Further P(x1,x2.....xn) is constant and equation is written as:

$$P(C_j|x_1,x_2,.....x_n) \propto P(C_j)\prod_{i=1}^{n} P(x_i|C_j)$$

Here we are calculating conditional probability of object with given feature vector x1,x2.....xn with respect to particular class.

3. Linear Regression Method : This is statistical approach for modeling .In this model a relationship between dependent variable and one or more independent variable is represented through linear regression .If independent variable is one then this is known as simple linear regression but if number of independent variables are multiple then this is known as multivariate linear regression. A straight line is represented in Figure 2 representing best fit line and this can be plotted using below equation

$d = a_0 + (i * a_1)$

Where d represents the dependent variable and i represents the independent variable. The main objective is to find the best value of a_0 and a_1

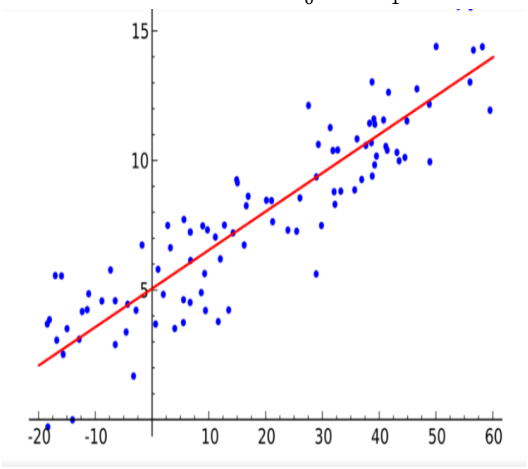

Fig 2: Linear Regression

This method is implemented using approach named as least square .In this we need to analyze the variable which is changing as a straight line function of another variable. Limitation of this approach is that this approach will not work for the cases where we are having high dimensional data[9].

4. SVM: Support Vector machine is said to be a supervised machine learning algorithm which can be used for regression as well as classification. The main idea of this algorithm is to classify the big data into different classes by generating an optimal hyperplane[10]. Line classifying two different classes is said to be optimal if distance between support vectors and line is maximum .Support vectors are the points from both the classes which are closest to the dividing line.
5. Logistic Regression Model: This model is almost similar to linear regression model .The main difference between logistics and linear regression model is that logistic is preferred for the case when dependent variable is binary in nature while in linear regression dependent variable is continuous in nature[12].

In logistics regression model dependent variable is represented in two possible cases (True/False or 0/1).It can be used to check the possibility of the event that can be fail or successful .This can be achieved by using sigmoid function as represented in equation :

$d = a_0 + (i * a_1)$

$$P = \frac{1}{1 + e^{-d}}$$

The graph of sigmoid function is represented below:

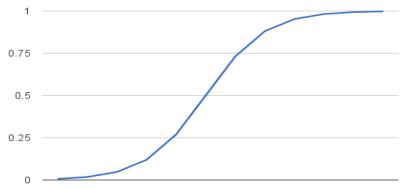

Figure 4: Logistics Regression

V. CONCLUSION

With advancement in big data, prediction and analytics is also growing day by day. But issue with big data is that it is present in inconsistent format i.e structured as well as unstructured format .To extract some useful information from the voluminous data is also a difficult task. So some special techniques and tools are required to deal with such huge and heterogeneous data .Also to deal with voluminous data efficient machine learning approaches are required to extract patterns and relationship among different parameters in field of healthcare .This chapter gives an overall idea of challenges of Big Data in field of healthcare and Big data tools to perform analytics and the advanced machine learning algorithms to provide optimized solutions.

REFERENCES

[1] J. L. Berral-Garcia, "A quick view on current techniques and machine learning algorithms for big data analytics", 18th International Conf. on Transparent Optical Networks, pp.1-4,2016. DOI: 10.1109/ICTON.2016.7550517J.

[2] Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A Survey of Machine Learning for Big Data Processing," EURASIP Journal on Advances in Signal Processing, vol. 67, pp. 1–16, 2016.

[3] M. U. Bokhari, M. Zeyauddin and M. A. Siddiqui, "An effective model for big data analytics", 3rd International Conference on Computing for Sustainable Global Development, pp. 3980-3982, 2016.

[4] M. R. Bendre, R. C. Thool and V. R. Thool, "Big data in precision agriculture: Weather forecasting

for future farming", 1st International Conf. on Next Generation Computing Technologies, pp. 744-750, 2015. DOI:10.1109/NGCT.2015. 7375 220.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, ''Disease Prediction by Machine Learning over Big Healthcare Data,'' IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017

[6] Retrieved from wikipedia: https://en.wikipedia. org/wiki/BRENDA

[7] Chen,M.,Mao,S.,&Liu,Y.(2014).Big Data: A Survey. Mobile Networks and Applications, 19(2). 171-209 doi:10.100/s11036-013-0489-0.

[8] Zhou, Lina & Pan, Shimei & Wang, Jianwu & Vasilakos, Athanasios. (2017). Machine Learning on Big Data: Opportunities and Challenges. Neurocomputing. 237.10.1016/j.neucom. 2017. 01.026.

[9] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, andM. D. Wang, "–Omic and Electronic Health Record Big Data Analytics for Precision Medicine", IEEE Transactions on Biomedical Engineering, vol.64, no.2, pp.263-273, 2017.

[10] https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[11] Dhafar Hamed Abd,Jwan K. Alwan, Mohamed Ibrahim, Mohammab B Naeem, "The Utilisation of Machine Learning Approaches for Medical Data Classification and Personal Care System Mangement for Sickle Cell Disease", Annual Conference on New Trends in Information & Communications Technology Applications (NTICT'2017)7 9 March 2017, IEEE 2017

[12] https://techdifferences.com/difference-between-linear-and-logistic-regression.html

[13] Peng, Z. (2019). Stocks Analysis and Prediction Using Big Data Analytics. 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS).doi:10.1109 /icitbs.2019.00081

[14] Mehdi Assefi, Ehsun Behravesh, Guangchi Liu, and Ahmad P. Tafti - "Big Data Machine Learning using Apache Spark MLlib ", 2017 IEEE International Conference on Big Data(BIGDATA), 11-14 Dec. 2017,Boston, MA, USA.

[15] M. D. Anto Praveena and B. Bharathi, "A survey paper on big data analytics," in 2017 International Conference on Information Communication and Embedded Systems (ICICES), 23-24 Feb. 2017.

[16] Leo Willyanto Santoso*, Yulia, "Data Warehouse with Big Data Technology for Higher Education," in 4th Information Systems International Conference2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia

[17] Dwivedi, S., & Roshni, V. K. (2017, August). Recommender system for big data in education. In E-Learning & E-Learning Technologies (ELELTECH), 2017 5th National Conference on(pp. 1-4). IEEE.

[18] Kobashi, S., Hossain, B., Nii, M., Kambara, S., Morooka, T., Okuno, M., & Yoshiya, S. (2016, July). Prediction of post-operative implanted knee function using machine learning in clinical big data. In Machine Learning and Cybernetics (ICMLC), 2016 International Conference on (pp. 195-200). IEEE.

[19] Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., & Alamri, A. (2017). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. IEEE Systems Journal, 11(1), 88-95.

[20] P. Samuel Kirubakaran et al, "An Automated Health Care Computing Model for Continuous Monitoring of Patients for Immediate Medical Care during Emergency," International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1307-1311

[21] E.R. Sparks, S. Venkataraman, T. Kaftan, M.J. Franklin, B. Recht, KeystoneML: optimizing pipelines for large-scale advanced analytics, in: 2017 IEEE 33rd In-ternational Conference on Data Engineering, ICDE, IEEE, 2017, pp.535–546.