Applications Of Big Data Analytics In Healthcare: A Research Perspective

M.S.Arun Kumar, R.S.Soundariya, M.Nivaashini, P.S.Dinesh, S.Iniya Shree

Abstract: The massive growth of smart devices have led to tremendous growth of data which paved the way for a new concept called as the Big Data. Big Data is actually a huge size of data which cannot be assembled and handled using traditional frameworks within a given timespan. Nowadays, the big data analytics play a crucial role in all fields such as education, healthcare, media, weather prediction, transportation etc. Among them health care is a notable sector in which huge data is generated by means of wearable devices, sensors, ECG etc. The extracted data can be utilized for further processing and results can be shared to the doctors as the evidence based records to diagnose the patients. The main challenge lies in processing such enormous amount of data using the existing frameworks in an efficient manner. Hence the proposed paper summarizes the advanced processing methodologies of various applications of big data analytics in healthcare sectors.

Index Terms: Big Data Analytics, Healthcare, Disease Diagnosis, Machine Learning, Hadoop.

1 INTRODUCTION

Big Data Analytics has been increasingly known as an embryonic technology and discipline in various aspects of our daily lives like medical and healthcare, financial industry government prospective, educational standards, business opportunities and social sciences. Big Data and its applications in health or medical science becomes even more noticeable because of new social arenas/media and networks (such as Facebook and Twitter), sensory/digital technology, and mobile devices with smartphone apps and personal sensor health data with real time digital data accumulations. [1, 2]. Diagnosing a disease and providing a proper and efficient treatment is very important for health care. During an outburst of new disease, it is difficult for the junior Doctors to identify it exactly. Medical knowledge from best Health care centres could be shared with every health centers using advanced technologies to boost the accuracy in diagnosis.

2 BIG DATA ANALYTICS IN HEALTH CARE

2.1 Big Data analytics for disease diagnosis and treatment recommendation

Nivedita et al conducted a detailed survey on medical applications of big data. The survey focused on prediction of various diseases like brain disease, heart diseases, kidney disease, tuberculosis and HIV/AIDS . The paper brought forth the drawbacks of using traditional relational database for disease prediction.

- *M.S.Arun Kumar is currently working as Assistant Professor in the department of CSE in Bannari Amman Institute of Technology, Sathyamangalam. E-mail: arunkumarms@bitsathy.ac.in*
- *R.S.Soundariya is currently pursuing research in the department of CSE in Bannari Amman Institute of Technology, Sathyamangalam. E-mail: soundariya@bitsathy.ac.in*
- *M.Nivaashini is currently pursuing research in the department of CSE in Bannari Amman Institute of Technology, Sathyamangalam E-mail: nivaashini@bitsathy.ac.in*
- *P.S.Dinesh is currently working as Assistant Professor in the department of CSE in Bannari Amman Institute of Technology, Sathyamangalam. E-mail: dineshps@bitsathy.ac.in*
- *S.Iniya Shree is currently working as Assistant Professor in the department of CSE in Sri Guru Institute of Technology, Sathyamangalam. E-mail: iniyashree@bitsathy.ac.in*

Hence by applying statistical techniques, accuracy and efficiency can be improved and thereby predictive models, intelligent systems can be built using advanced techniques like cluster analysis and hadoop framework [3]. Jianguo et al. introduced an algorithm called Density-Peaked Clustering Analysis (DPCA) by analysing large scale data sets of past examinations especially for diseases with multiple pathogenesis and proposed a Disease Diagnosis and Treatment Recommendation System (DDTRS). Association analysis is also performed on Disease-Diagnosis (D-D) rule and Disease-Treatment (D-T) rule using the Apriori algorithm, that helps to employ diagnosis and treatment knowledge among Health centres. They also enforce Doctors to update the inspection results in DDTRS during different stages to grab recommendations. Hadoop Distributed File system is used to store bulk medical data. They achieved high performance and low latency response by parallelizing DDTRS using Apache Spark cloud computing platform [4]. Clustering analysis of text formatted inspection data is carried out by building medical domain ontology library, here they have used PubMed biomedical database. Medical ontology of inspection dataset is drawn and similarities among them are found and PDCA algorithm is applied to it.Association analysis process is explained for D-T rules using Apriori algorithm. Multiple visits by the patient to the health centre is required in the treatment of certain diseases, which gives us multiple treatment records and these records are considered as association record. Frequent itemset are generated by setting minimum support and minimum confidence, then strong association rules are extracted from it. Along with the disease diagnosis, an interface between doctors and patients was also designed to share their recommendations.The proposed DDTRS lacks timely recommendation and also lacks in the efficiency of the disease symptom clustering, to overcome this Apache Spark cloud computing platform parallelized DDTRS. Finally the accuracy of the disease symptom clustering, robustness of the DPCA algorithm and quality of treatment recommendations are evaluated and found that it outperforms in all evaluation. Chitra et al. used Big data analytics with Evidence based methodology to analyze the health conditions of the patient and perfect diagnosis is done. For treatment, patients past medical evidence and current reports are considered. An analysis report is produced to monitor the recovery and success rate of the given diagnosis is monitored with the help of feedback collected from the patients on the drug prescribed.

Evidence Based Medicine Analysis is achieved using Big data technique and it includes analysis of patients health(Height, Weight, BMI, Lipid Profile etc), clinical evidences (symptoms and biomedical investigations) and disease pattern analysis (inquiring the tropical region pattern, heredity pattern collected via opinion poll).The stacked up data are divided into little parts and processed in parallel using HDFS. Master process runs on a single node (name node) whereas the data node runs on all the nodes of the cluster. Creation, deletion and replication of data nodes are managed by Name nodes. Map reduction function is applied on it, with symptoms and diseases as key value pair and disease that has almost all the symptoms as mentioned by the patient is taken out and corresponding drug is prescribed. K-Means clustering method is applied to the set of d-dimensional vectors D={xi|i=1,2,…N} where xi belongs to d.Centroids is choosen either by sampling at random from data sets and setting them as output for small subset or by perturbing the global mean of data k times. Data assignment and relocation of 'means' is done until convergence. This system thus helps the health centres to identify the disease accurately from the symptoms, medical history with Big data analysis prior to the actual treatment [5].

2.2 Big Data to treat complications in medical field

Fetal growth curves are used in prenatal medicine to locate fetal growth problems, estimate prenatal outcome and treat complication within reach right away. Mario et al. aims at creation of customized curves using Big data techniques to give solution for poor precision that exists in currently chosen curves. Huge amount of input data is summarized by means of multidimensional views in addition with clustering and classification. Least square method, multidimensional analysis and clustering and classification techniques are commonly used techniques to assess the health of fetus-maternal. Even though Fetal growth are under risk of new variables every year because of new pollutants, pathologies, medicines, it is important to consider. Customized fetal growth curves are constructed by searching for patients who share almost identical growth patterns using multidimensional analysis techniques and also in searching for possible correlation of fetal growth alternative by fetal gender, maternal age, etc. Fetus with same gestational age, similar environmental conditions and constitute to similar genetics are subject to similar fetal growth curves and are referred as Homogeneous Patient Groups (HPG). When a growth parameter differs from those to which he/she belongs, we can check for whether the fetus is pathologic or not. Multidimensional view of the target is build, the HPG is initially unknown and can be determined by undergoing specific tests. Once HPGs to which the target belongs is identified, its health is tracked by comparing its current growth with the reference chart of HPG. HPG and reference growth curves are periodically updated and not static. Reference growth curves are associated with hundreds/thousands of HPGs instead of few ethnic groups. Experiment is done with 25000 records of 500 patients by considering 60 attributes grouped into 9 main categories and the results confirm the effectiveness of their proposed system [6]. Zhan et.al designed and developed an efficient text mining framework called Spark text on a Big data infrastructure which is composed of Apache Spark data streaming and machine learning methods combined with Cassandra No SQL database. They demonstrated the performance of their framework for classifying cancer types by employing Naive Bayes, Support Vector Machine (SVM) and Logical regression to build prediction models to the thousands of data from PubMed. The Sparktext mined the data set in 6 mins which is 93.81% accurate. Cancer may occur due to various causes including inherited genetic mutations, hormones, weak immune system, tobacco exposure, unhealthy diet etc [7]. K. Sharmila et.al developed a hybrid model for Big data using classification and clustering techniques in Hadoop. Incorrectly classified instances are stamped out using the K-Means clustering, the right cluster is picked up from it and fine tuned classification is performed using Support Vector Machine (SVM).The result from the hybrid model shows that patients with Diabetes and affected with the symptoms of Cardiovascular disease, Nephropathy and Retinopathy are predicted efficiently [8].

2.3 Big Data analytics in medication during outburst of any disease

Chinmayee et al. divides the population dynamics by considering Dengue fever disease into three levels based on their severity as High, Mid, Low vulnerable. Depending upon the vulnerability level corresponding preventive measure is selected among Forced, Efficient and delayed preventive measure. Various symptoms of the disease is collected from various resources like hospitals, social media and from people affected by it. Collected data are pre-processed stored in HDFS (Hadoop Distributed File System), it breaks the larger file into smaller blocks and stores them as clusters. Further processing is done by Map-Reduce that works in two phases namely mapping and reducing. Map-Reduce divides the data set into smaller chunks and those chunks are mapped by mapper function in parallel manner. The output of the mappers are shuffled automatically and stored, according to the output they have determined the vulnerability and corresponding preventive measures [9]. P.SubhaPriya et al. proposes an enhanced data mining algorithm for health care applications using anomaly detection, clustering and classification. Random forest algorithm is opted for classification thereby increasing the accuracy of their system. They have used supervised learning for training set. Classifier ensembles with five base classifier has been used on five medical data sets, different types of decision tree algorithm are chosen and output is evaluated [10].

2.4 Big Data Dealing with Metabolic Syndrome

Gregory et al. aims to predict accurately the metabolic syndrome and its several factors on both at individual and population level by applying Big data analytic platform along with Reverse Engineering and Forward Simulation (REFS) to large data set of one nationwide customer. The REFS platform learns by Metroplois Monte Carlo sampling followed by Forward Simulation to study risk factor as well as the impact of interventions for individuals and population.The advantages of this platform includes generating insights faster, increases program impact and returns, learns model directly from data with consistent processing, allows extreme data heterogenity, safeguards against overfitting. It also overcomes the drawback in Archimedes model and general purpose statistical analysis platform such as SAS [11].

2.5 Big Data Handling -Omic and EHR Data

Pon-Yen et.al facilitates precision medicine by addressing the challenges in the -Omic and EHR data using Analytics of Big

Data. They also identified disease biomarkers from multi-omic data and incorporated -omnic information into EHR. Health care has various heterogeneous data in Bioinformatics, health informatics, imaging informatics and sensor informatics. To make use of these data, it should undergo big data analytics such as analysis, modeling, interpretation, quality control and validation. Higher Throughput -omic assays such as Next Generation Sequencing (NGS) and mass Spectrometry results in huge generation of -omic data. EHR data can be unstructured and may contain spelling errors, grammar errors, abbreviations and acronyms, so it is difficult to handle such heterogeneous data. Structured EHR can be either Administrative data (remains unchanged) or Ancillary data (discrete / continuous). Data collection frequency will vary from one data to other, for example genome needs one time data acquisitions where as few other data may need multi point data acquisitions. Similarly in EHR, Bedtime data can be more frequent when compared to lab test data. The quality of data may vary because of various reasons like contamination of sample, misinterpretation of original records, missing data due to variables that vary among clinics, low signal to noise ratio, etc. In addition to data collection frequency and variance of quality of data, there are other problems like high dimensionality and heterogeneous data [12]. Following are the approaches used for precision medicine using Big Data Analytics:

1) General Analytics for biomedical Big Data: High dimensional data takes more time to process and also results in low accuracy, hence it is necessary to reduce the dimensionality of data without affecting the original characteristics of data. It can be done with feature selection (filtering, wrapper or embedded methods) followed by feature extraction (artificial neural network, Principal Component Analysis).
2) Pre-processing of -omic data: Sequence mapping, alignment, baseline correction, peak detection are used for pre-processing.
3) Biomarker identification: Biomarker is a characteristic by which a particular disease can be identified. So it is necessary to identify those biomarkers even though the samples are collected at different stages of disease. -Omnic biomarkers includes differential protein-DNA binding, differential histone modification etc. This can be identified by quantifying and then fit abundance of each group to Poisson Distribution followed by statistical tests.
4) Systems Biology modeling: It is done either by static network analysis (identifying and decomposing the network scaffold and mathematically represent each decomposed scaffold for downstream simulation and analysis) or by dynamic temporal analysis.
5) EHR Data preprocessing: To solve missing data values issue it is necessary to follow missing data imputation methods like interpolation, multiple imputation, and maximum likelihood. Sensor fusion techniques are used to correct waveforms. Filtering techniques such as median filtering, model based filtering are used to handle noise.
6) EHR data mining: This can be done by static endpoint prediction with Regression analysis, classification and association Rule learning techniques. Temporal data mining can be done with hidden Markov model and conditional random field.

2.6 Big Data to improve the accuracy of the classification algorithm

R. Saravana kumar et al. tried to deal with the complexity and accuracy issue that exists in traditional classification algorithms such as support vector matrix (SVM), neural networks, k-Nearest neighbors, decision trees, Differential evolution algorithm, Naive Bayes, machine learning algorithm for large amount of medical big data. Random Forest classifier algorithm is used here along with K-Means clustering algorithm. As K-Means clustering algorithm takes less time for partitioning the large data sets, it has high performance and hence it is used first to form clusters. They have done the mean estimation for each cluster and computed the difference between each cluster member and mean of cluster value, then RF classification is applied on this.While performing K-Mean clustering the value of K can be chosen using methods like cross validation, information criteria, information theoretic jump, G-Means algorithm or simply by comparing the results across different values of k the mean difference between data points and cluster centroid. Cluster centroid is selected and data assignment is done for it, then centroid is recomputed and steps are repeated again until no data points change clusters.. Random forest cherish the convenience of decision trees such as missing value, continuous and categorical predictions and majority vote is resolved from all individual trained trees after classification. They have done experiments with five datasets including USPS (United States Postal Service) dataset and found that training is faster in RF with high interpretation and few parameters when compared to LC-KNN and RC-KNN [13].

3 CONCLUSION

The developing field of Big Data science and related practices offered new prospects and is promising, however it accompanies numerous difficulties in all fields, particularly the biomedical and wellbeing science fields which makes improved understanding of human life. The synergistic system, supporting conditions and interdisciplinary approach with exceptionally prepared computational abilities and insolent utilization of available resources are the keys for improving the genuine estimation of real time big data for significant decision making and better outcomes in healthcare.The future developments of big data analytics in healthcare have the perspective of improving and accelerating collaborations among patients and doctors by minimising the costs and improving efficiency based on risk reduction, early prediction results and enhancing the personalized care.

4 REFERENCES

[1] Bollen J, Mao H, Zeng XJ ,"Twitter mood predicts the stock market", Journal of Computational Science,Volume .no.2,issue.1, Pages 1-8, March 2011.
[2] Bachrach Y, Kosinski M, Graepel T, Kohli P, Stillwell D "Personality and patterns of Facebook usage", WebSci '12: Proceedings of the 3rd Annual ACM Web Science Conference, pp: 24-32, 2012.
[3] Nivedita Das, Siddharth Rautaray, Manjusha Pandey, "Big Data Analytics for Medical Applications", International Journal of Modern Education and Computer Science, Volume.no.1,issue.2, February 2018.
[4] Jianguo Chen, Kenli Li, Huigui Rong, Kashif Bilal, Nan Yang, Keqin Li, "A Disease Diagnosis and Treatment Recommendation System based on Big Data Mining and

Cloud Computing", Information Sciences, Elsevier, 2018.
[5] Chitra Pasupathi, Vijaya Kalavakonda, "Evidence based health care system using big data for disease diagnosis", 2nd International Conference on Advances in Electrical, Electronics, Communication and Bioinformatics, Feb 2016.
[6] Mario A. Bochicchio, Antonella Longo, Lucia Vaira and Sergio Ramazzina, "Online Data Analysis of Fetal Growth curves", ICA3PP 2013, Part II, LNCS 8286, pp. 149-156, 2013.
[7] Ye Z, Tafti AP, He KY, Wang K, He MM " SparkText: Biomedical Text Mining on Big Data Framework" PLoS ONE 11(9): e0162721, 2016.
[8] K Sharmila, SA Vethamanickam, "MRK-SVM: An Effective Technique for Big Data In the HealthCare Sector", International Journal of Scientific & Engineering Research, 2016.
[9] Chinmayee Mohapatra, Leena Das, Siddharth Swarup Rautray, Manjusha Pandey, "Map-reduce based modeling and dynamics of infectious disease", International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Feb 2017.
[10] Subhapriya. P, Sujatha. R, Meghana. K, "Healthcare Prediction Analysis in Big Data using Random Forest", International Journal of Advance Research, Ideas and Innovations in Technology, pp. 494-496, 2017.
[11] Gregory B. Steinberg, Bruce W. Church, Carol J. McCall, Adam B. Scott, Brain P. Kalis, "Novel Predictive Models for Metabolic Syndrome Risk: A Big Data Analysis Approach", The American Journal of Managed Care, Volume 20, No. 6, June 2014.
[12] Wu P.-Y., Cheng C.-W., Kaddi C.D., et al. "Omic and electronic health record big data analytics for precision medicine" IEEE Trans. Biomed. Eng., 64:263–273, 2017.
[13] R. Saravana Kumar, P. Manikandan, "Medical Big Data classification using a combination of Random Forest classifier and k-means clustering", International Journal of Intelligent Systems and Applications, 11, 2018.
[14] Liang Y, Kelemen A ," Big Data Science and Its Applications in Health and Medical Research: Challenges and Opportunities" Journal of Biometrics and Biostatistics,Volume.no.7,issue.3, 2016.