International Institution for Technological Research and Development
Volume 2, Issue 1, 2017

Efficient Mining of Cloud Based Electronic Health Records (EHR) for Clinical Decision Support System

Konica Dhingra
Department of Information Technology,
LJIET
konica.dhingra@gmail.com

Krunal Panchal
Asst. Professor, PG Department
LJIET
krunaljpanchal@gmail.com

Abstract **- In feats to support the growing trend of maintaining huge amount of clinical data in a common storage (Electronic Health Records) and predicting diseases based on the results of mining these data. Community clouds are in general trend where patient's medical history is stored and based on this training datasets, the diseases of new incoming patient is predicted using Clinical Decision Support System (CDSS). Data mining algorithms provide an efficient way to guess the disease of new incoming patient. The aim of this research focuses on developing parallelized classification algorithms which effectively uses computational benefits of cloud storage to accurately predict the disease. Our algorithm achieves parallelism by dividing the datasets into sub portions and feeding each sub portion into an individual processor (virtual machine) on cloud and then applying algorithms like Decision Tree induction, k-n-n classifier and Naïve Bayesian classifier. Laplacian correction technique is used to remove attributes whose values may be zero .At last, we have used ensemble methods like Random Forests and Bagging to find out the class with majority votes and assign that class to the incoming tuple (Patient's symptoms) which acts as the output of the CDSS.**

Keywords **- Electronic Health Records (EHR), Clinical Decision Support System (CDSS), classification, Decision tree induction, Naïve Bayesian classification, k-Nearest-Neighbor classifiers, Laplacian correction, Bagging, Random Forests, Parallel Laplacian Classification.**

1.Introduction

Data mining has great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify deficiencies and best practices that improve care and reduce costs. This could be a win/win overall. But due to the

complexity of healthcare data and a slower rate of technology adoption, our industry lags these others in implementing effective data mining and analytic strategies. An Electronic Health Record (EHR) is the digital version of a patients' medical chart, designed to contain and share information electronically with other health care providers and agencies involved in patient care. The volume of EHRs is growing because they are mandated in the United States by the 2009 American Recovery and Reinvestment Act (ARRA), and in particular, a section of ARRA called the HITECH Act (Health Information Technology for Economic and Clinical Health) [1] . To reduce the medical errors by improving the accuracy and clarity of medical records. To make the health information available, reducing duplication of tests, reducing delays in treatment, and patients well informed to take better decisions [2]. For Patients and families; improve coordination and population and public health. EHR data will improve clinical processes including patient controlled data, clinical decision support, health information exchange (HIE), and quality measurement and research [3]. Use of EHR data to improve health outcomes, quality of data, patient's safety, efficient way to analyze the diseases and precautions, and population health at the national level. Improve health outcomes, reduce medical errors, predict health trends, and demonstrate the comparative value of drugs and other treatments. To provide quality of care and Patient's safety, financial savings, technology advancements, aggregated and integrated the data and coordination of care.

We are using classification techniques to predict disease of any random patient using the model constructed. The model is constructed basically using the concept of parallel classification. The input data set (training data set) is obtained from community cloud storage where different hospitals store their medical data. The reason to store patient's medical records on cloud is to gain the computational benefits of cloud and faster access. Apart from that cloud also let a huge amount of medical data to be stored in a single place. This can be thought of as Electronic Health Records (EHR) where a single patient's medical data can be accessed by more than one authorities (Doctors, Pharmacies, and Claim Agents etc). Now, after obtaining the training data set from the Cloud based EHR, attribute selection measures are applied and the data set is divided into N sub portions. Where, N is the number of processors (virtual machines) on cloud that will be used to achieve parallelism. Each sub portion is feed to one processor. Each processor individually and synchronously process the data sets and applies any of the classification algorithm (Decision Tree Induction, Naïve Bayesian Classification, k-nearest-neighbor classification) on them. Also, Laplacian Correction technique is used to eliminate the possibility of zero attribute values. When the computation at each processor is finished, the models constructed with the help of different classification algorithms is used and then the majority vote class is assigned to the input data provided. This is known as the ensemble method, where M_k models are used to predict class for the same data set and the prediction is done based on majority

votes where each model M_1, M_2, M_3…, M_k has equal vote. A clinical decision support system (CDSS) is an application that analyzes data to help healthcare providers make clinical decisions [4]. A CDSS is an adaptation of the decision support system commonly used to support business management.

2.Background Theory

Data mining is a component of a wider process called knowledge discovery from databases. It involves scientists from a wide range of disciplines, including mathematicians, computer scientists and statisticians, as well as those working in fields such as machine learning, artificial intelligence, information retrieval and pattern recognition [5]. Data mining or knowledge discovery in database, as it is also known, is the process of extraction of understandable patterns in data, previously unknown and potentially useful information from the data. Data mining techniques can be applied in the field of medical data mining to predict the disease of patient based on his/her medical history and other symptoms. For example, classification and clustering techniques can be used to know if a patient is suffering from Tuberculosis or not by seeing the medical history of similar patients. Cloud Storage is a model of data storage in which the digital data is stored in logical pools, the physical storage spans multiple servers (and often locations), and the physical environment is typically owned and managed by a hosting company [6]. Cloud storage is based on a virtualized infrastructure with accessible interfaces, which provides flexible and portable on demand access to stored resources. Types of cloud storage are: Public, Private and Hybrid Clouds. Cloud based EHR are used to store data and apply predictive analytics on it to predict a patient's disease and suggest the further treatments based on past data present in cloud. Implementation is much easier with cloud-based EHR systems. We just have to store the data in cloud storage rest while retrieving data back, we just need to use the computational abilities of cloud. Practices do realize tremendous savings from cloud-based EHR systems. As, data is stored on cloud, storage is fast, reliable and available. Resource requirements like hardware and software are significantly reduced. Web-based software provides superior accessibility and collaboration. Cloud based EHR are fast, safe, easy, reliable and provides an on-demand way of accessing medical records. Scalability is simplified with cloud-based systems. We can store a huge amount of data in cloud storage.

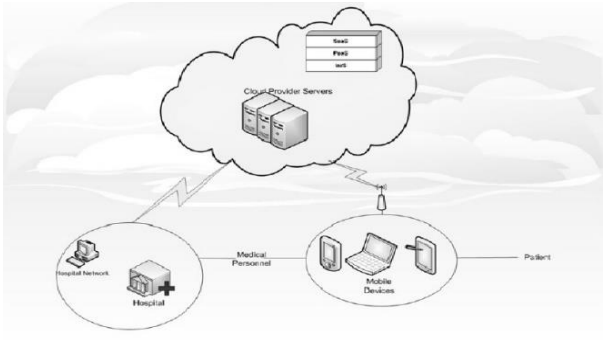

Figure 1: Cloud based solution for a large hospital [7]

Reduced expense for both software and hardware. Satisfaction levels are higher among mobile & tablet EHR users. It has been observed that patients who use

mobile/web based EHR and who rely on EHR for their day to day medical activities and medications and satisfied with their health conditions.

3.Literature Survey

Parallel classification concept is utilized. To improve the efficiency of classification algorithms by running them in parallel over multiple virtual machines (in cloud). Classification algorithms like Decision tree induction, Naïve Bayesian classifier and knn classification is used. Classification algorithms like Decision Tree Induction, k-NN classifier and Naive Bayesian classifier are run in parallel across multiple processors (Virtual Machines) in cloud. The 3 parallel classifier's outputs are given as an input to boosting algorithm which is an ensemble method to improve accuracy of parallel classifier. In this paper, when the patient consults the doctor, his/her medical history is stored in cloud based EHR. Then these data are processed and Principal Component Analysis (PCA) technique is applied on it for attribute selection. Now, the selected training samples are given as input to the classification algorithms being applied on different virtual machines. Classification accuracy is improved by applying ensemble method called Bagging. When a new input comes in, it is feed to the classification algorithm to predict the class of new incoming data. Thus, in this way the whole model is systematically planned so that any redundant data or outliers are not present in the data before classification. The time complexity of the algorithm considerably decreases because of the parallelization process. The test data is assigned to a class with majority of votes from each classifier. Classifier accuracy can be evaluated using specificity and sensitive matrix.

Sensitivity : TP/ TP+FN

Specificity : TN/FP+TN

Where, TP : True positive , TN : True Negative , FP : False Positive , FN : False Negative.

In another paper, the main objective is to present a novel classifier for classification in the field of Medical Data Mining. Laplacian Correction is applied to Decision Tree Classification algorithm, as well as, Naïve Bayesian Classification algorithm. Hence, an improved version of the two classifiers, namely, Decision-Tree Classifier (modified) and Naïve-Bayesian Classifier (modified) is proposed. The Adaptive Classifier has been created by combining the techniques of Rule-Based Classifier, Decision-Tree Classifier (modified) and Naïve-Bayesian Classifier (modified). The data for this research work has been collected from Nalanda Medical College Hospital (NMCH), Patna, Bihar, India for the non-infectious disease Diabetes and the infectious disease Tuberculosis. The attributes of the patients included in the sample medical dataset are: Name, Age, Gender, Location, Occupation, Economic Status, Disease Symptoms and Disease Status. The raw data collected underwent a few pre-processing techniques so as to obtain better results, and to improve the performance of the proposed model [8].

4.Proposed System

I am going to propose a new adaptive enhanced way for analyzing the data by "Efficient Mining of Cloud based Electronic Health Records (EHR) for Clinical Decision Support System (CDSS)". Data will be stored on cloud with Apache Spark, which gives faster analyzing of the process rather than Hadoop. Also parallel classification is good process to analyze the data. Moreover the missing attributes can be handle by the Laplacian Correction which removes the zero consistency values with proper suggestive value. So based on the Laplacian Correction use for the modification of the classification algorithms. N number of data needs to be classified with splitting the data attributes in equal number of classifiers to work in parallel.

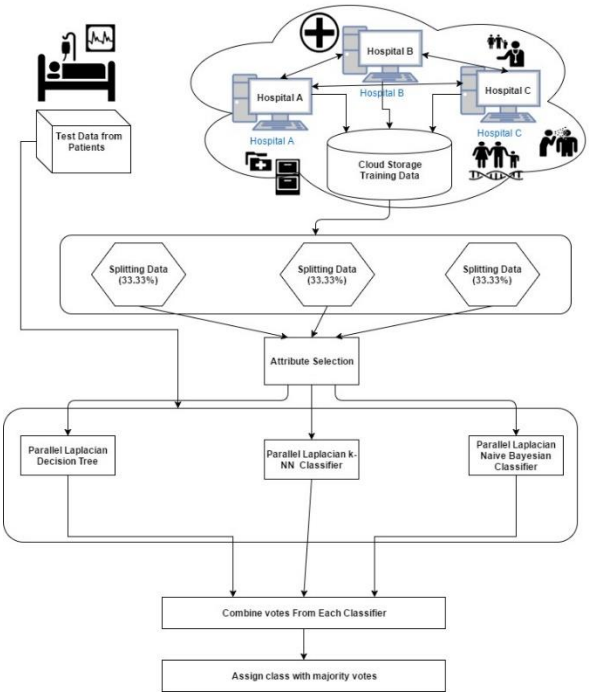

Figure 2: Block diagram of proposed system

The datasets are taken from a) https://vincentarelbundock.github.io/Rdatasets/datasets.html and b) https://archive.ics.uci.edu/ml/datasets.html : .Also, these data files are placed on cloud server , which acts like a community cloud for multiple hospitals.

For simulation, RapidMiner[9] tool is used. In order to integrate the proposed with Hadoop , Radoop[10] (RapidMiner + Hadoop) extension is used. For Hadoop environment setup , Oracle VirtualBox and CloudEra[11] extension for the same is used. By implementing the steps as described in proposed model, we are able to build and effective and efficient predictive analytics model. The graphs generated through simulation are as given below.

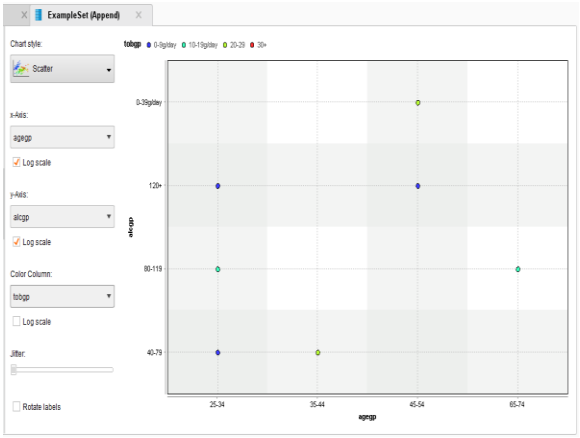

Figure 3: Scatter chart of proposed model with first dataset (esophCancer)

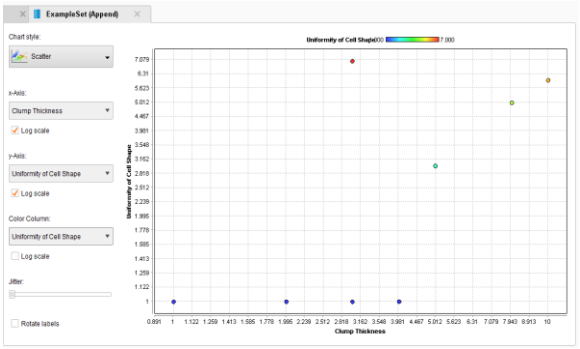

Figure 4: Scatter chart of proposed model with second dataset (wbcd)

5.Conclusion

The primary focus of this research is to modify existing predictive analytics systems to improve the quality of healthcare and treatment processes. The early detection of diseases can lead to a better health of people at national level. Thus, the main focus is put on efficiently mining the cloud based EHR so that it can correctly predict the disease and treatment for a specific patient whose data is completely new to the Clinical Decision Support System (CDSS) so that it can help physicians better.

6.References

[1] Haritha Chennamsetty, Suresh Chalasani, Derek Riley , "Predictive Analytics on Electronic Health Records (EHRs) using Hadoop and Hive", IEEE International ISSN: 978-1-4799-6085-9/15 , Pg No. : 1 , 2015

[2] http://medicaleconomics.modernmedicine.com/medicaleconomics/news/mining-ehr-data-quality-improvement?page=full, 29th Nov, 2016, 12:00 AM.

[3] http://ieet.org/index.php/IEET/more/hanraha n20131115, 29th Nov, 2016, 12:00 AM.

[4] https://www.healthit.gov/policy-researchers-implementers/clinical-decision-support-cds, 29th Nov, 2016, 12:00 AM.

[5] Arun K Pujari , "Data mining Techniques"; 1st Edition; Universities Press (India) Private Limited, Hyderabad, 2001, Pg. no. 2.

[6] http://searchcloudcomputing.techtarget.com/definition/cloud-computing, 29th Nov, 2016, 12:00 AM.

[7] Gonzalo Fernández-Cardeñosa, Isabel de la Torre-Díez & Miguel López-Coronado, Joel J. P. C. Rodrigues , "Analysis of Cloud Bases Solutions on EHRs Systems in Different Scenarios", Springer, April, 2012.

[8] Nazanin Shahrokhi, Roxana Dehzad and Soheila Sahami, "Targeting Customers with Data Mining Techniques: Classification", 2011 International Conference on User Science and Engineering (i-USEr), ISBN: 978-1-4577-1655-3/11/$26.00, Pg 212 - 215, IEEE, 2011.

[9] https://rapidminer.com/

[10] https://rapidminer.com/products/radoop/

[11] https://www.cloudera.com/