Medical Data Mining for General Hospitals

Dhiraj D. Jagtap[1], Komal S. Baviskar[2], Trupti V. Patil[3], Sohail Shaikh[4]

[1,2,3,4]*Student, B.E Computer Engineering, SSBT's COET, Maharashtra, India*

---***---

Abstract - *Healthcare industry have large amount of data but not all the data unfortunately are classified and mined for prediction of disease and provide solution using various techniques for classification and advance Big data techniques like MapReduce used for analysis and clustering and develops new methodology using Hadoop, it discovers the knowledge in database particularly in disease prediction. The proposed work aims at the prediction of patient's disease and suggestion can be given to the patient so that it will help to manage patient's risk by calculating Risk index of the patient disease. In the emergency situation the patients of respective disease can be alert by computing patient's Risk and patient's information.*

Key Words: **Data mining, MapReduce, Hadoop, Risk index.**

1. INTRODUCTION

The healthcare industry has history to generate large amounts of data, It consist of record keeping, regulatory requirements, compliance and patient care [1]. In healthcare industry most of the data is stored in traditional system like files and registers that is in hard copy form, now the recent trend is increasing towards the rapid digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare service mean while reducing the costs, these large quantities of data hold the promise of supporting a wide range of health care and medical functions, including clinical decision support and population health management.

By definition, large data in health care system refers to electronic health data sets so large and complex and these large data sets are very difficult to manage with traditional software and hardware. They cannot be easily managed with traditional or common data management tools and methods. Data in healthcare system is very great in amount because of its volume and also because of the diversity of data types and the speed at which it must be managed.

The data totality related to patient healthcare and well-being make up large data in the medical or healthcare industry. It includes clinical data and clinical decision supporting systems (physicians written notes and prescriptions, laboratory reports, pharmacy reports, insurance, and other administrative data); patient data in electronic patient records (EPRs); machine generated/sensor data, such as social media posts, including Twitter feeds, blogs, status updates on Facebook and other platforms, and web pages; and less about patient-specific information, including emergency care data, news feeds and articles in medical journals and books.

2. LITERATURE REVIEW

Numerous works in literature related with heart disease diagnosis using data mining techniques have motivated our work. Some of the works are discussed below:

Shantakumar B. Patil et.al [6], have proposed a proficient methodology for the extraction of significant patterns from the heart disease warehouses for heart attack prediction has been presented. Initially, the data warehouse is pre-processed in order to make it suitable for the mining process. Once the preprocessing gets over, the heart disease warehouse is clustered with the aid of the K-means clustering algorithm, which will extract the data appropriate to heart attack from the warehouse. But there are some Disadvantages like increase in processing time to process big volume of data of patient to predict patient's survival.

Byung Kwan Lee et.al [1], have proposed a Patient-customized Healthcare System based on the Hadoop with Text Mining (PHSHT) for an efficient Disease Management and Prediction. The PHSHT consists of four modules .Firstly, MDCM stores the big data like healthcare and medical information in the Hbase. Secondly, TMHM converts the unstructured data collected by MDCM to the structured data and stores the result in the Hbase. Thirdly, DRCM generates disease rules by using the Conditional Probability Set Theory (CPST) and stores them in the Hbase. The DMPM not only provides an efficient patient-customized medical service and information by comparing the disease rules generated by the DRCM with a patients information, a patients current status, and family history but also predicts a patients disease according to his health status.

Manaswini Pradhan et.al [2], have described the role played by information and communication technology has certainly a greater contribution for its effective delivery mechanism. The application of data mining is specifically relevant and it has been successfully applied in medical needs for its reliable precision accuracy and expeditious beneficial results. The various available application techniques have been discussed and analyzed for the purpose of the paper.

Cheryl Ann Alexander et.al [3], described the identification of the usage of Big Data analytics in heart attack prediction and prevention, the use of technologies applicable to big data, privacy concerns for the patient, and challenges and future trends as well as suggestions for further use of these technologies.

3. RELATED WORK

3.1 Hadoop

Hadoop mainly consists of Hadoop Distributed File System (HDFS), Hbase, and Hadoop MapReduce which can analyze large data. It is a open source framework that writes and implements an application program for processing large amount of data.

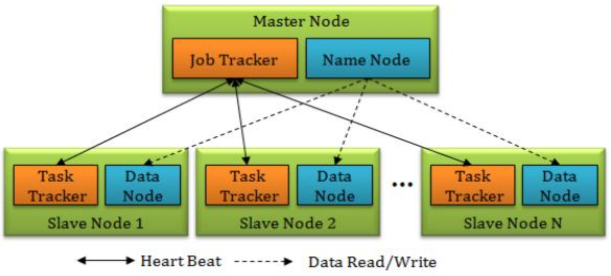

Fig.1: The Components of HDFS

HDFS is made up of a Master Node and several Slave nodes. Fig.1 shows Components of HDFS[1]. The Master Node, it consists of a Name Node that controls an access to a client file and a Job Tracker which accomplishes the scheduling about the given jobs. The Master Node manages the name space of HDFS in Hadoop. The storage on each node is managed by slave node and a Task Tracker accomplishes the jobs assigned by a Job Tracker.

3.2 MapReduce

The MapReduce, is a Distributed and Parallel processing model of data and it is based on a Key/Value pair. MapReduce provides scalability to data growth caused by Distributed and Parallel processing and it minimizes network traffic caused by data movement among nodes. The MapReduce in the following figure generates a intermediate result with the key/value by accomplishing MapReduce based on input data. The intermediate result grouped by key value is transferred to a Reduce Task. The Reduce Task integrates all the intermediate keys and transfers the final result to the Hbase.

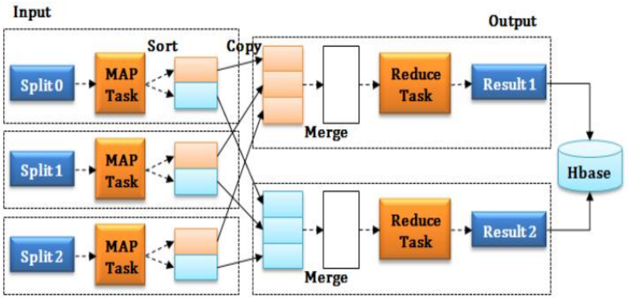

Fig.2: The Flow of MapReduce

3.2.1 The Map process of the Hadoop

The information from the patient's data, medical prescription, and clinical notes are stored in the Hbase. A Map task generates the intermediate data sorted with the data which is stored in the Hbase shown in the Fig.3 based on the Key and transfers it to a Reduce process step[1].

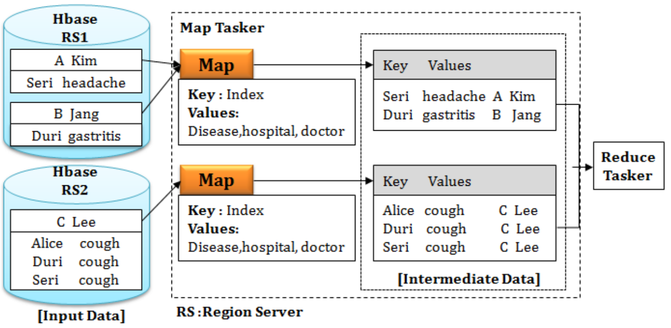

Fig.3: The Map process of Hadoop

3.2.2 The Reduce process of the Hadoop

A Reduce task sorts the transferred intermediate result values from a Map task on the basis of key values. The final result merged with the same key is stored in the Hbase. Thus, the stored final result is used to predict a patient's disease.

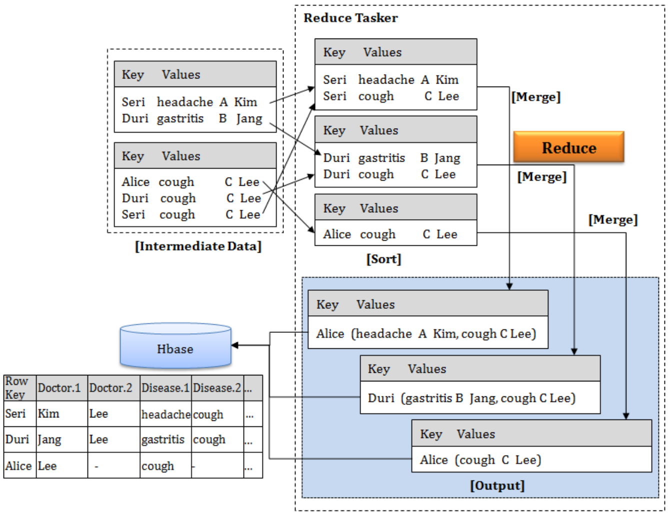

Fig.4: The Reduce process of Hadoop

3.3 Hbase

The Hbase is a distributed column-based database implemented in the HDFS and provides random access to a big data set in real time. It also supports the retrieval in new version. The data in the Hbase is stored in the row and column table. One row in the table consists of a row and column key and the column is grouped with column-family. The row is sorted with a main row key in the table and in a byte order. And the Hbase's access to the table is done through the row key of the table.

The existing RDBMS has several limitations on massive scalability and distributed processing, but the Hbase is linearly extensible by appending nodes only. Besides, it can Process data fast by paralleling big data sets with MapReduce.

3.4 C4.5 algorithm

C4.5[1]is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

In general, steps in C4.5 algorithm to build decision tree are:

-Choose attribute for root node

-Create branch for each value of that attribute

-Split case according to branches

-Repeat process for each branch until all cases in the branch have the same class choosing which attribute to be a root is based on highest gain of each attribute. To count the gain, following formula is used to count the gain.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{S_i}{S} * Entropy(S_i)$$

Here, {S1, ..., Si, ... Sn} is the partitions of S according to values of attribute A. n is the number of attributes A. |Si| is the number of cases in the partition Si .|S| is the total number of cases in S. The entropy is gotten by following formula.

$$Entropy(S) = \sum_{1}^{n} -\rho_i \times \log_2 \rho_i$$

Here, S is case set, and n is the number of cases in the partitions S. Pi is the Proportion of Si to S.

4. SYSTEM ARCHITECTURE

System architecture consists of design of a Patient-customized Healthcare System based on the Hadoop with MapReduce framework. The proposed system consists of 4 modules as shown in Figure.5

Firstly, the MDCM (Medical Data Collection Module) stores large data such as a patient's health and medical information in the Hbase which occurs in a hospital or a portable healthcare system.

Secondly, the TMHM (MapReduce based Hadoop Module) analyzes the collected data with MapReduce based Hadoop, and merges the data and stores it in Hbase again with a MapReduce Framework.

Thirdly, the DRCM (Disease Rule Creation Module) generates disease rules by using the disease information stored in the Hbase and C4.5 algorithm and stores them again in the Hbase.

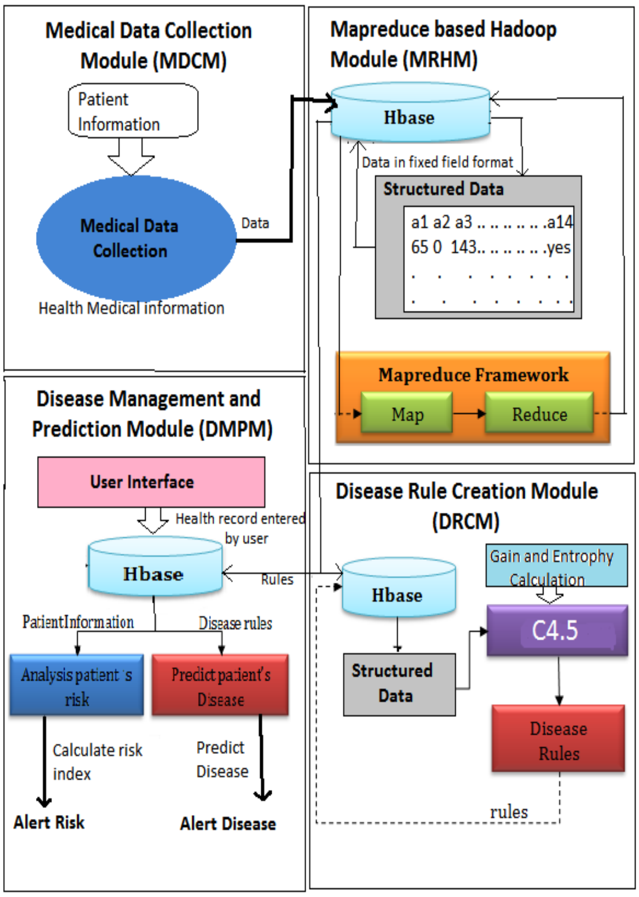
Medical Data Collection Module (MDCM)
Patient Information
Medical Data Collection
Data
Health Medical information
Mapreduce based Hadoop Module (MRHM)
Hbase
Data in fixed field format
Structured Data
a1 a2 a3 .. .. .a14
65 0 143.. .. .. .yes
Mapreduce Framework
Map
Reduce
Disease Management and Prediction Module (DMPM)
User Interface
Health record entered by user
Hbase
Rules
PatientInformation
Disease rules
Analysis patient's risk
Predict patient's Disease
Calculate risk index
Predict Disease
Alert Risk
Alert Disease
Disease Rule Creation Module (DRCM)
Hbase
Gain and Entrophy Calculation
C4.5
Structured Data
Disease Rules
rules

Fig.5: The Modules of Proposed System

Fourthly, the DMPM (Disease Management and Prediction Module) informs a patient or his family doctor of a risk index or a result of disease prediction, after analyzing a patient's risk with a patient's collected information or predicting a patient's disease by comparing the disease rules generated by the DRCM with the patient's collected information.

5. RESULTS

Hadoop framework is used for faster processing of data using Map Reduce technique. C4.5 algorithm is used to generate the rules for decision making in Hbase. The decision making rules are store again into database. By applying decision rules on the patient's information the risk index can be calculated and message can be given to the patient.

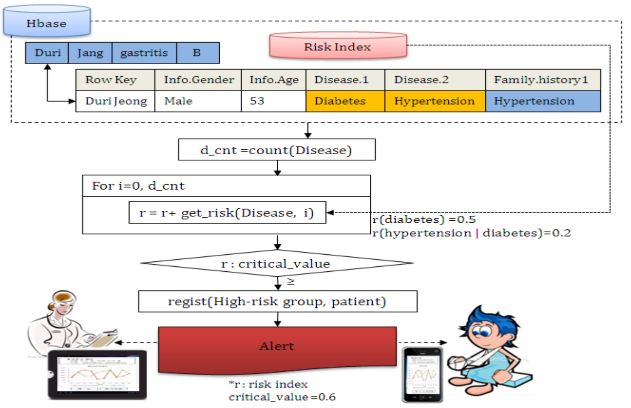

Fig.6: The Flowchart of disease prediction

i. If Risk index is greater than 0.9 then Emergency alert is given to the patient.

ii. If Risk index is between 0.5 and 0.9 then notice is given to the patient.

iii. If Risk index is less than 0.5 then ignore.

6. CONCLUSIONS AND FUTURE WORK

The proposed system a PHSHT design extracts the patients information from the Medical data about a patient by using Hadoop based on MapReduce and provides the disease management and the disease prediction with the extracted information. In particular, the Disease Rule Creation Module in the proposed system designs the C4.5 algorithm which generates disease decision rules by using the attributes with high Gain. The C4.5 algorithm reduces the number of the attributes used in disease rules and improves the accuracy of rules. The Disease Management and Prediction Module in the proposed system alert a patient's emergency situation by computing a patient's disease risk. In addition, it predicts diseases by comparing a patient's information to the disease rule the DRCM generates.

In future the dynamic model of patient health care system can be developed where the record of new patient will be added into the Hbase and accordingly intermediate and new rule will be generated to give more accurate results about patient's disease.

REFERENCES

[1] ByungKwan Lee, EunHee Jeong, "A Design of a Patient-customized Healthcare System based on the Hadoop with Text Mining (PHSHT) for an efficient Disease Management and Prediction" Vol.8, No.8 (2014).

[2] Wei Dai and Wei Ji, "A MapReduce Implementation of C4.5 Decision Algorithm",school of Economics and Management, Hubei University,Hungashi 435003, P.R.Chaina.

[3] Alexander and Wang, J Nurs Care,"Big Data Analytics in Heart Attack Prediction"(2017), 6:2 DOI: 10.4172/2167-1168.1000393

[4] Mimoh Ojha,Dr. Kirti Mathur, "Proposed Application of Big Data Analytics in Healthcare at Maharaja Yeshwantrao Hospital" 2016 3rd MEC International Conference on Big Data and Smart City.

[5] Ms Manaswini Pradhan,"Data Mining and Health Care: Techniques of Application",PG Dept of I and CT, F M University, Balasore, Odisha 756019 India,Volume 1 Issue 1;Page No. 18-26

[6] Shantakumar B.Patil Y.S.Kumaraswamy,"Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network",ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656

[7] Heart Disease,http://archive.ics.uci.edu/ml/machine/learningdatabases/statlog/heart/