ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (73-77), Month: October 2015 – March 2016, Available at: www.paperpublications.org

Intelligent Heart Attack Prediction System Using Big Data

[1]Prajakta Ghadge, [2]Vrushali Girme, [3]Kajal Kokane, [4]Prajakta Deshmukh

Savatribai Pune University

Abstract: **In today's world heart attack is the very common. So diagnosing patients correctly on timely basis is the most challenging task. Treatment of heart disease is quite high and not affordable by most of the patients. The main objective is to develop a prototype Intelligent Heart Attack Prediction System using Big data and data mining modelling techniques. This system can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot.**

By providing effective treatments, it also helps to reduce treatment costs. The healthcare industry collects huge amounts of big data which, unfortunately, are not "mined". Remedies can be provided by using advanced data mining techniques. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous.

Therefore, a medical diagnosis system like heart attack prediction system would probably be exceedingly beneficial.

Keywords: **Big data; Data Mining; Hadoop; Healthcare; Knowledge-Discovery; Risk Prediction.**

I. INTRODUCTION

The heart is vital part of our body. Life is completely dependent on efficient working of heart. If functioning of heart is not proper, it will affect the other parts of body human such as brain, kidney etc. Now a days heart attack has become a major cause of uncertain death world wide. Earlier only middle aged and old people were prone to heart attack, but due to today's unhealthy lifestyle it is affecting youngsters too. Heart attack prediction is a complex task for medical practitioners. So heart attack prediction system would prove to be beneficial by using data mining techniques.

Data mining has already established as a novel field for exploring knowledge from hidden patterns in the big datasets. "Data Mining" is a non-trivial extraction of implicit, previously unknown and potentially useful information from given data. In short, it is a process to analyze the data from different perspective and gather the knowledge from it. This discovered knowledge can be used in different application domains for example in healthcare industry (Heart attack prediction). So we will use data mining techniques along with big data and machine learning to develop software to assist doctors and other people in making decision of heart attack in early stages.

II. LITERATURE SURVEY

Investigation is carried out that focuses on diagnosis of heart disease. Different data mining techniques have been applied for diagnosis & achieved different probabilities for several methods.

•An Intelligent Heart Attack Prediction System is developed using data mining techniques. Neural Network, Naïve Bayes and Decision Tree proposed by Sellappan Palaniappan.

Each method has its own strength to get appropriate outputs. To setup this system hidden patterns and relationship between them is used. It is expandable, web based and user friendly system.

•Multi parametric feature with linear and nonlinear characteristics of Heart Rate Variability (HRV) a novel technique was developed by HeonGyu Lee. To achieve this, he has used several classifiers i.e. Bayesian Classifiers, Classification based on Multiple Association Rules (CMAR), Decision Tree and Support Vector Machine (SVM).

•The identification problem of constrained association rules for heart attack prediction was studied by Carlos Ordonez. The final dataset contains records of patients having heart disease. Three constraints to decrease the number of patterns are as follows:

1. The attributes should appear only on one side of the rule.

2. Separation of attributes into groups i.e. uninteresting groups.

3. In a rule, the number of attributes should be limited. The result will be separated into two groups of rules, the presence or absence of heart disease.

•In heart attack prediction, Blood Pressure and Sugar with the help of neural networks was proposed by Niti Guru. The dataset has 13 attributes in each record. The supervised network i.e. Neural Network along with back propagation algorithm is used for training and testing of data.

•Akhil Jabbarr proposed an efficient associative classification algorithm by using genetic approach for heart attack prediction. The main motive of using genetic algorithm in the discovery of high level prediction rules is that the discovered rules have high predictive accuracy and are highly comprehensible with great interestingness values.

•Kiyong Noh, for the extraction of multi parametric features by assessing Heart Rate Variability (HRV) from ECG, Data preprocessing and heart disease pattern used classification method. The dataset consisting of 670 people was distributed into two groups viz, normal people and patients with heart disease.

III. TECHNOLOGIES

Big Data:

Big data is a broad term for datasets, it is so large or complex that traditional data processing applications are insufficient. Challenges of big data include analysis, capture, data collection, search, sharing, storage, transfer, visualization, and information privacy. The term simply refers to the use of predictive analysis or other certain advanced methods to extract valuable information from data, and often refer to a particular size of dataset. Accurate analysis in big data may result in more confident decision making. And better decisions can provide us with greater operational efficiency, reduced risk and cost reduction.

Data mining:

Data mining is the analysis step of Knowledge Discovery in Databases or KDD. It is an interdisciplinary subfield of computer science. This is the process of discovering patterns in large datasets ("big data") involving methods at the intersection of artificial intelligence, machine learning, database systems and statistics. The final goal of the data mining process is to extract relevant information from a dataset and transform it into an understandable format for further use.

Hadoop and Mahout:

Apache Hadoop is an open-source software framework written in Java for distributed processing and distributed storage of huge datasets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a significant assumption that hardware failures of individual machines, or racks of machines are common place and should be handled automatically in s/w by the framework. The core of Apache Hadoop consists of a storage part Hadoop Distributed File System (HDFS) and a processing part Map Reduce. Hadoop files are split into large number of blocks and are distributed among the nodes in the cluster. For data processing, Hadoop Map Reduce transfers packaged code for fnodes for parallel processing, based on the given data each node needs to be processed. This approach takes advantage of data locality nodes, manipulating the data that they have on hand to allow the data to be processed more efficiently and

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (73-77), Month: October 2015 – March 2016, Available at: www.paperpublications.org

faster than it would be in a more conventional supercomputer architecture that depends on a parallel file system where data and computation are connected through high-speed networking.

Apache Software Foundation produced Apache Mahout project for free implementations of distributed or scalable machine learning algorithms which primarily focus on the areas of collaborative filtering, classification and clustering. Apache Hadoop platform is used by many applications. It also provides Java libraries for common math operations which focus on linear algebra and statistics and primitive Java collections.

Naive Bayes:

Naive Bayes classification is based on Bayes theorem. This classification algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent on values of other attributes.

The Bayes theorem is as follows:

Let,

A={a1, a2, …….an} be a set of n attributes.

In Bayesian theorem, A is considered as evidence and H be some hypothesis mean value, the data of A is a subset of class C.

We have to evaluate P (H|A), the probability that the hypothesis H holds given evidence i.e. data sample A.

According to Bayesian theorem the P (H|A) is given as,

P (H|A) = P (A| H) P (H) / P (A)

IV. ARCHITECTURE

System architecture is described below:

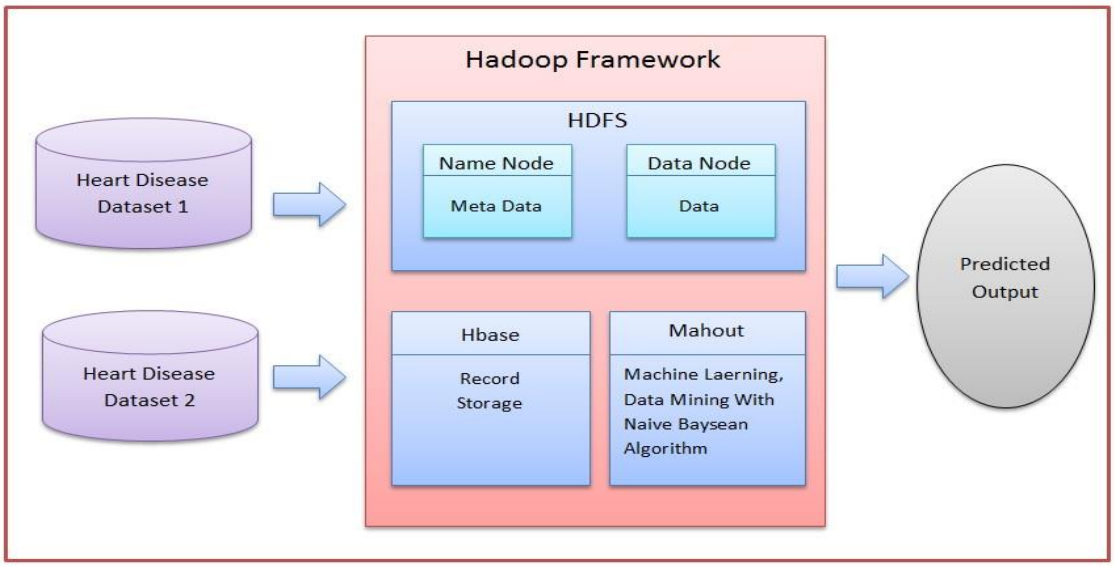

Fig 1. Architecture

- This system consists of two datasets. First is original big dataset and the other one is updated dataset.
- HDFS: It is a Java-based file system which provides a user with reliable and scalable data storage.

Name node: It is the centre piece of an HDFS. It keeps the record of all files in the file system, and tracks location of the file in a cluster. The name node cannot store the file data by itself.

Data node: Data Node stores data in the Hadoop File System. A functional file system consists of more than one Data Node, with data replication.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (73-77), Month: October 2015 – March 2016, Available at: www.paperpublications.org

- Hbase: It is used when one needs random, real time read or write access to Big Data. The goal of hbase is to host very large tables i.e. billions of rows X millions of columns.

V. RESULTS

Data set:

Clinical records have accumulated large quantity of data about patients and their medical diagnosis reports. The term cardiovascular disease involves the diverse diseases that affect the heart. Record set with medical attributes was obtained from the Cleveland Heart Disease database which is available on web. With the assistance of the dataset, the patterns necessary for heart attack prediction are extracted.

Input attributes:

1. Age in Year.

2. Sex- (value 1: Male; value 0: Female).

3. Thal - (value 3: normal; value 6: fixed defect; value 7: reversible defect).

4. CA – number of major vessels colored by fluoroscopy (value 0-3).

5. Old peak – ST depression induced by exercise.

6. Exang - exercise induced angina (value 1: yes; value 0: no).

7. Chest Pain Type -(value 1:typical type1 angina, value 2: typical type 2 angina, value 3:non-angina pain; value 4: asymptomatic).

8. Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy).

9. Serum Cholestrol (mg/dl).

10 .Fasting Blood Sugar- (value 0: <120 mg/dl:value 1: >120 mg/dl;).

11. Trest Blood Pressure (mm Hg on admission to the hospital).

12. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping).

13. Thalach – maximum heart rate achieved.

14. Heart Disease Present - 0:No 1: Yes.

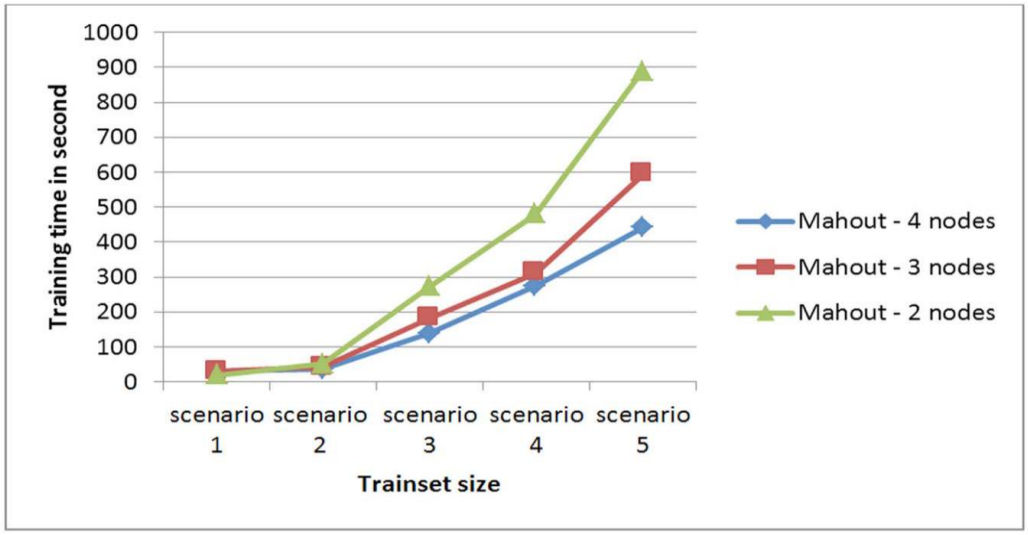

Fig 2. Training time for Mahout with different number of nodes

VI. CONCLUSION

Here we have studied heart attack prediction system using big data solution. Our proposed solution gives big data infrastructure for both predictive modeling and information extraction. We study the effectiveness of our proposed system with an experiment set, consisting of scalability and quality. Future scope of our system aims at giving big data infrastructure for our designed risk calculation tools, for designing more sophisticated prediction models and feature extraction techniques and extending our proposed system to predict other clinical risks.

REFERENCES

[1] *Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients* by Kiyana Zolfaghar, Naren Meadem, Ankur Teredesai, Senjuti Basu Roy, Si-Chi Chin Institute of Technology, CWDS, UW Tacoma. IEEE 2013.

[2] *Prediction of Heart Disease using Classification Algorithms by* Hlaudi Daniel Masethe, Mosima Anna Masethe, USA

[3] *Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction* by Jyoti Soni, Ujma Ansari, Dipesh Sharma International Journal of Computer Applications

[4] *Heart Disease Prediction System using Associative Classification and Genetic Algorithm* by M.Akhil jabbar , Dr.Priti Chandrab, Dr.B.L Deekshatuluc, Research Scholar, JNTU Hyderabad, A.P INDIA

[5] *Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing* by Ms.Rupali R.Patil Asst. Professor, Jawaharlal Nehru College of Engineering.

[6] *Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques* Chaitrali S. Dangare Sulabha S. Apte, PhD.