ISSN: 0975-766X
CODEN: IJPTFI
Available Online through ***Research Article***
www.ijptonline.com

ENDOWED HEART ATTACK PREDICTION SYSTEM USING BIG DATA

G.Kavitha[1] and Evangelin Hema Mariya.R[2]
[1]Research Scholar, School of Information Technology, VIT University, Vellore.
[2]Research Scholar, School of computer science and Engineering, VIT University, Vellore.
Email: hema122588@gmail.com

Received on: 15-02-2017 *Accepted on: 27-03-2017*

Abstract

Presently a day's heart assault has turned into a noteworthy reason for indeterminate demise around the world. Prior just center and old matured individuals were inclined to heart assault, yet today's way of life it is influencing adolescents as well. Heart assault is the main source of death and it is important to foresee it prior stages to spare the life of individuals. The Prescient investigation includes the separating of data from existing datasets in order to anticipate the results and patterns with a worthy level of precision and dependability. This paper breaks down a couple of parameters and anticipates heart assault. Thereby recommends a heart assault forecast framework in light of Naive Bayes arrangement and Hadoop and Mahout Framework structure. In machine learning, Naive Bayes classifiers are a group of basic probabilistic classifiers in view of applying Bayes hypothesis.

Keywords: Heart disease, Prediction, Hadoop, Mahout, Machine learning, Naive Bayes.

1. Introduction

Predictive analytics can be used effectively to evaluate enormous data generated by healthcare industry to extract useful information and establish relationships amongst the variables. The health care suppliers have just began to catch of predictive analytics but are rapidly becoming aware that they have to make change as the healthcare industry demands are changing. Unlike traditional statistical methods to reveal surprising associations which doctors would never even suspect. Pharmaceutical companies, hospitals and insurance providers will see changes from past treatment outcomes, latest medical research and database like fewer complications, short hospital stay, fewer readmissions.

Currently healthcare departments are turning to big data technology to improve and manage medical systems. The rapidly expanding field of big data has started to play an important role in the evolution of healthcare practices and

research. Healthcare system is the preservation of mental and physical health by preventing or treating illness through services offered by the provision.

The physicians are well trained and experienced however they simply cannot retain the record of all the medical problems that they come across with their solutions. Possibly if they may have access to the huge amount of medical repository, they would still require considerable amount of time to analyze the data and tailor it in line with the patient's profile. More and more doctors are the hassle predictive analysis for healthcare due to these reasons. Predictive analysis uses the database to forecast outcomes for the patient. The database could include data from previous medical history and other publications and review papers. This uses techniques such as statistical methods, data exploration and machine learning how to create a profile from earlier cases. This model is utilized to realize a new individual patient with an instant conjecture and a precise diagnosis.

2. Literature Survey

Numerous works in literature related to the prediction of disease has motivated this work. A brief literature survey is presented here.

Analysis of Classification Algorithm for Heart Disease Prediction and its accuracies was described by Jothikumar et al [1]. The accuracies of different classifiers are analyzed with the use of UCI Cleveland healthcare dataset is experimented with the support of Rapid Minor software. The test and training datasets were passed as input to the Random tree, Naive Bayes, Decision Tree and Random Forest. It is found that Naive Bayes better accuracy of 79.25%, Decision Tree with 78.24%, Random tree with 75.14% and Random Forest with 74.16%. Accuracies of these algorithms can be improved by preprocessing the datasets as the dataset may subject to noisy, inconsistent, missing and outdated values. It can be used by healthcare professionals to predict the heart disease earlier.

Dangre et al., [4] proposed Improved Study of Heart Disease Prediction System Using Data Mining classification Techniques. Huge amount of data were used to extract hidden information for making intelligent medical diagnosis. The system was used 15 attributes and a Multi-Layer perceptron Neural Networks (MLPNN) that maps a set of input data onto a set of appropriate. The total records are divided in to 2 datasets first one is training dataset it contains 303 records and the second one is testing dataset it contains 270 records. The system was used Neural Network as a classification technique.

Sonam Nihar et al., [5] presented Prediction of heart disease using machine learning algorithms. In this paper, the author collected record set with 76 medical attributes from Cleveland Heart disease database. But they worked on reduced attribute set with 19 attributes. These attributes are called useful attributes for prediction of heart disease. The records are split into two types of datasets one is training dataset, the second one is testing dataset. The datasets were used for preprocessing to make desired form. Different classification techniques Naive Bayes and Decision tree were performed over the preprocessing data to predict the accuracy of heart disease. From this experiment Naive Bayes and Decision tree with information gain calculation provides better results in the prediction of heart of disease and better accuracy. The accuracy of the classification algorithm can be calculated using confusion matrix. In order to improve the consistency and efficiency of the prediction system, genetic algorithm will be implemented in MATLAB.

Rupali and Patil [3] explain Heart Disease Prediction System Using Naive Bayes and Jelinek-mercer smoothing. This research work was developed a decision support in Heart disease Prediction system using data mining modeling technique Naive Bayes and smoothing technique. The system extracts hidden knowledge from historical heart disease database. The smoothing technique could answer complex queries. The system can expandable in the sense that more number of records or attributes can be incorporated and new significant rules can be generated using underlying data mining technique.

3. Technology Used In Proposed System

3.1 HADOOP: Apache Hadoop is an open-source software context written in Java for distributed processing and distributed storage of huge datasets on computer clusters constructed from commodity hardware. Hadoop are designed with a significant assumption that hardware disappointments of distinct machines or racks of machines are common place and should be handled automatically in s/w by the framework in all modules. Apache Hadoop core consists of Hadoop Distributed File System (HDFS) storage part and a processing part Map Reduce. Hadoop files are split into large number of blocks and are distributed between the nodes in the cluster. Hadoop Map Reduce transfers packaged code for nodes for parallel processing, based on the given data each node needs to be processed. This approach takes benefit of data locality nodes, working the data that they have on hand to allow the data to be processed more efficiently and faster than it would be in a more traditional supercomputer architecture that depends on a parallel file system where data and computation are connected through high-speed networking.

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

3.3 Mahout:

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

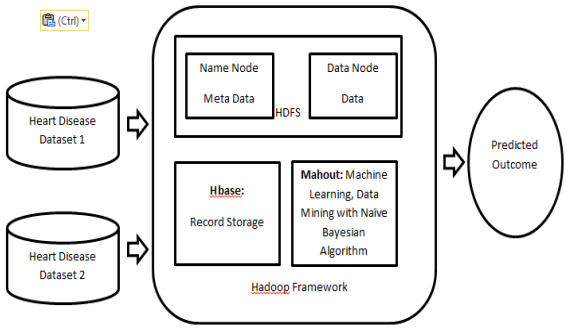

Figure 1. Architecture of Heart Disease Prediction.

Figure 1 describes the overall architecture of heart disease prediction system. The system uses heart disease dataset from UCI Machine learning repository.

Table 1. Attributes of heart disease and description.

Attributes Name	Description
Age	Age in years
Sex	Sex (Value1 : Male, Value 0 : Female)
CPT	Chest Pain Type Value 1 : typical type 1 angina Value 2 : typical type 2 angina Value 3 : non-angina pain Value 4 : asymptomatic
Chol	Serum Cholesterol in mg/dl
Fbs	Fasting blood sugar value > 120 mg/dl
Restecg	resting electrocardiographic results Value 0 : normal Value 1 : having ST-T wave abnormality Value 2 : showing probable or

	definite left ventricular bypertrophy
Hrtach	maximum heart rate achieved
Exang	exercise induced by angina Value 1 : yes Value 0 : No
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of peak exercise ST segment Value 1: up sloping Value 2: flat Value 3: down sloping
Ca	number of major vessels (0-3) colored by fluoroscopy
Thal	3= normal, 6=fixed defect, 7=reversible defect
Num	diagnosis of heart disease Value 0 : No risk Value 1 : Low risk Value 2 : Risk Value 3 : High risk Value 4 : Higher risk

The datasets are classified as two types first one is training dataset and second one is testing dataset. This experiment uses 13 attributes are age, sex, cp, trestbps, chol, ca, thal, fbs, restecg, hrtach, exang, oldpeak, slope and num for heart disease prediction. Table 1 explains the description of the attributes.

4. Proposed Methodology

***4.1 Naive Bayes*:** The Bayesian Classification represents a supervised learning method and statistical method designed for classification. Assumes a fundamental probabilistic model and it permits us to capture uncertainty about the model in a p way by defining probabilities of the results. It can resolve diagnostic and analytical problems. Naive Bayes algorithm is based on Bayesian Theorem. Figure 2 describes the working procedure of Naive Bayes.

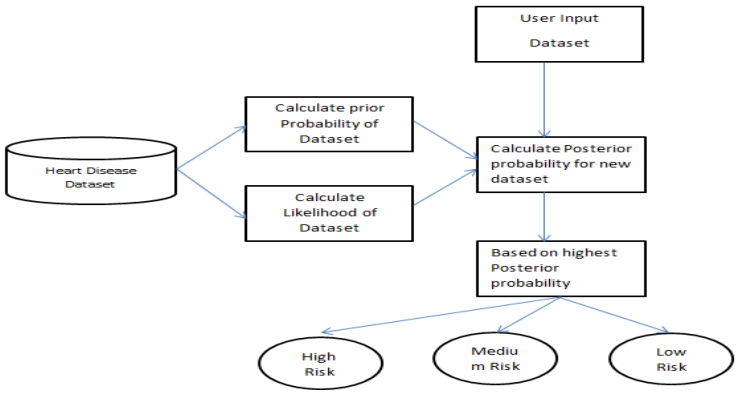

Figure 2. Naive Bayes working Procedure.

4.2 Bayseian Theorem:

Given training data X, posterior probability of a hypothesis theory H, P(H|X), monitors the Bayes theorem P(H|X)=P(X|H)P(H)/P(X)

4.3 Algorithm:

Bayesian theorem is based on Naive Bayes algorithm as given by steps

Steps in algorithm are as follows: Each data sample is considered by an n feature vector dimensional, X = (x1, x2…. xn), representing n attributes made on the sample from n measurements respectively A1, A2, An.

1. Assume that there are m classes, C1, C2……Cm. Given X, an indefinite data sample (i.e., having no class label), the classifier will calculate that X belongs to the class having the maximum latter probability, Conditioned: P(Ci/X)>P(Cj/X) for all 1< = j< = m and j! = i Thus we exploit P(Ci|X). The Ci class for which P(Ci|X) is maximized is called posteriori hypothesis. By Bayes theorem,
2. As P(X) is persistent for all classes, only P(X|Ci) P(Ci) need be exploited. If the class prior probabilities are not recognized, then it is generally assumed that the classes are similarly like, i.e. P(C1) = P(C2) = …...= P(Cm), and we would hence exploit P(X|Ci). Otherwise, we exploit P(X|Ci) P(Ci). Note that the class prior probabilities may be projected by P(Ci) = si/s, where Si is training samples of class Ci, and s is the complete number of training samples On X. That is, the naive probability assigns an indefinite sample X to the class Ci.

4.4 Performance Evaluation

The proposed solution gives big data infrastructure for both predictive modeling and information extraction. The effectiveness of the heart disease prediction system is observed in terms of scalability, accuracy and quality. Scalability is achieved by using hadoop frame work. Confusion matrix is a table that is often used to describe the performance of a classification model or classifier on a set of test data for which the true values are known. It is used to show the accuracy of the classification.

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

Accuracy =correctly classified records / total records.

10-fold Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

- Break data into 10 sets of size n/10.
- Train on 9 datasets and test on 1.
- Repeat 10 times and take a mean accuracy.

5. Conclusion

The objective of this work is to provide a technique that can be employed in automated heart disease prediction systems. Various techniques are defined in this paper as related work which is emerged in recent years for efficient and effective heart disease diagnosis. Each technique has its own pros and cons the selection of model should be done based on the objective of the model and the data. Motivated by the world wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using predictive analytics models in the diagnosis of heart disease. Applying predictive models helps healthcare professionals in the diagnosis of heart disease. The predictive analytics model is used to identify a suitable treatment for heart disease patients has received less attention.

References:

1. Jothikumar and Sivabalan, "Analysis of Classification Algorithm for Heart disease Prediction and its Accuracies" Middle-East Journal of scientific Research, PP (200-206), ISSN 1990-9233; IDOSI Publications, 2016.
2. Prajakta Ghadge, Vrushali Girme, Kajal Kokane and Prajakta deshmukh, "Intelligent Heart diseases Prediction System Using Big Data" International Journal of Recent Research in Mathematics Computer Science and Information Technology vol. 2, Issue 2, PP: (73-77), Month: October 2015 – March 2016.
3. Rupali Patil, "Heart disease prediction system using Naive Bayes and jelinek mercer smoothing" International journal of advanced research in computer and communication engineering VOl-3, issue 5, May 2014.
4. Chitra and Seenivasagam, "Review of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques" ICTACT Journal of Soft Computing, July 2013, Volume: 03, Issue: 04.
5. Sonam Nikhar and Karandikar, "Prediction of heart disease using machine learning algorithms" International journal of advanced engineering, Management and science (IJAEMS) vol-2, Issue-6, June – 2016.

6. Florence, Bhuvaneswari Amma, G.Annapoorani and K.Malathi, "Predicting the risk of heart diseases using neural network and decision tree" International journal of innovative research in computer and communication engineering, VOL 2, issue 11, November 2014.
7. Venkatalakshmi, shivsankar, "Heart disease diagnosis using predictive data mining" Internal Journal of innovative research in science, Engineering and technology Volume 3, Special Issue 3, March 2014.
8. Marco Viceconti, Peter Hunter and Rod house, "Big Data, Big Knowledge: Big Data for Personalized healthcare" IEEE Journal of biomedical and health informatics, vol. 19, No. 4, July 2015.
9. Ximeng Liu, Rongxing Lu, jianfeng Ma, Le chen and Baodong Qin, "Privacy – Preserving patient centric clinical decision support system on naive Bayesian classification" IEEE journal of biomedical and health informatics VOL 20. NO. 2 March 2016.