Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis of Data Mining Techniques for Diagnosing Heart Disease

[1]Jyoti Rohilla*, [2]Preeti Gulia
[1]Research Scolar, [2]Assistant Professor
[1, 2] Dept of Computer Science & Applications, Maharshi Dayanand University
Rohtak, Haryana, India

Abstract— Heart disease (HD) is a major cause of morbidity and mortality in the modern society. Almost 60% of the world population fall victim to the heart disease. Usage of data mining techniques in healthcare industry increases as health care industry gathers a huge volume of data and discovers useful trends or patterns that are used in diagnosis and decision making. In addition, it is most inspiring area of research. The quality of prediction can be improvised by using data mining techniques suitably. In this paper, we analyse various data mining techniques which are introduced in recent years for heart disease prediction. The proportion of deaths caused by heart disease is the greater then other disease. For the analysis we have used some of data mining algorithms like ID3 and J48 etc to develop the prediction models using Heart disease dataset. We have used 10-fold cross validation methods to measure the unbiased estimate of these prediction models. The observations reveal that Id3, a decision tree technique has outperformed over all other data mining techniques.

Keywords— Data mining, Heart disease, Prediction, Classification

I. INTRODUCTION

According to the World Health Organization heart disease is the first leading cause of death occurs equally in both men and women. By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs). The growth of medical databases is very high. This rapid growth motivates researchers to mine useful information from these medical databases. Data mining techniques find patterns and extract knowledge to provide better patient care moreover, effective diagnostic capabilities.

Predication should be done to reduce the risk of Heart disease. Diagnosis is usually based on signs, symptoms and physical examination of a patient. Almost all the doctors predict heart disease by these factors or symptoms. The diagnosis of disease is a difficult and tedious task in medical field. Predicting Heart disease from various factors or symptoms is a multi-layered issue which may lead to false presumptions and unpredictable effects. Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. The large amount of data is a key resource to be processed and analysed for knowledge extraction that enables support for cost-savings and decision making. Only human intelligence alone is not enough for proper diagnosis. A number of difficulties will arrive during diagnosis, such as less accurate results, less experience, time dependent performance, knowledge up gradation is difficult etc.

Medical Data mining in healthcare is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases. This research paper aims to analyse the several data mining techniques proposed in recent years for the diagnosis of heart disease. Many researchers used data mining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease in which several data mining techniques are used in the diagnosis of heart disease such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, Genetic Algorithm, Naive Bayes, Decision tree, WAC which are showing accuracy at different levels.

Automated Heart disease prediction can benefit healthcare sector. This automation will save not only cost but also time. This paper presents different data mining techniques which are deployed in these automated systems. Various data mining techniques can be helpful for medical analysts for accurate heart disease prediction.

A. The Risk Factor for Heart Disease

1) *Family History of Heart Disease:* - Heart disease can be hereditary, it can run in families. If anybody has a family history of heart disease, that person may be at greater risk of that kind of heart disease.
2) *Smoking:* - Smoking is major cause of heart attack and other Heart disease. Acc to a survey nearly 40% of all people die from smoking. There is rapid increase in heart and blood vessel disease due to smoking.
3) *Cholesterol:* - Abnormal levels of lipids (fats) in the blood are risk factor of heart diseases. Cholesterol is a soft, waxy substance found among the lipids in body's cells. High level of triglyceride combined with high levels of LDL (low density lipoprotein) cholesterol speed up atherosclerosis increasing the risk of heart diseases.
4) *High Blood Pressure:* - High blood pressure also known as HBP or hypertension. High blood pressure increase the risk of our blood vessels walls becoming overstretched and injured.

5) *Obesity:* - Obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Obesity is higher risk for health problem such as heart disease, stroke, high blood pressure, diabetes and more.

II. LITRATURE REVIEW

Works done in heart disease diagnosis using data mining is discussed below:

Mohammad Taha Khan et al. [9] developed a prototype model for the breast cancer as well as heart disease prediction using data mining techniques, namely, C4.5 and the C5.0. He used Evidence Based Medicine (EBM), which is a new and important approach which can greatly improve decision making in health care. **Nidhi Bhatla and Kiran Jyoti [11]** analyze different data mining techniques and concluded that Neural Network has shown good accuracy out of all other Data mining technique. **Vikas Chaurasia and Saurabh Pal [15]** applied three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm on Heart disease Dataset. **Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte [8]** developed a Heart Disease Prediction system (HDPS) using data mining and artificial neural network (ANN) technique.

Mohammed Abdul Khaleel et al. [10] compare data mining techniques with conventional methods. **Divya Tomar and Sonali Agarwal [2]** explore the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. **Parvathi I, Siddharth Rautaray[12]** discuss how to enable the disease diagnosis and prognosis, and the discovery of hidden biomedical and healthcare patterns from related databases along with a discussion of the use of data mining to discover such relationships as those between health conditions and a disease, relationships among diseases.

Deepali Chandna [4] presents work that shows how information gain method, feature selection technique, can be used in collaboration with adaptive neuro fuzzy inference systems (ANFIS) in diagnosing new patient cases. **Abhishek Taneja[14]**design a predictive model for heart disease detection using data mining techniques from Transthoracic Echocardiography Report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography. **R. Chitra and V. Seenivasagam [13]** observe that Hybrid Intelligent Algorithm improves the accuracy of the heart disease prediction system.

III. HEART DISEASE DATA

The data used in this study is the Hungarian Institute of Cardiology. Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. While the databases have 76 raw attributes, only 11 of them are actually used. The data set contains 295 rows. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

A. Attribute Information:

1) Age (age in years)
2) Sex (1 = male; 0 = female)
3) Chest pain type (4 values)
   - Value 1: typical angina
   - Value 2: atypical angina
   - Value 3: non-anginal pain
   - Value 4: asymptomatic
4) Resting blood pressure
5) Serum cholesterol in mg/dl
6) Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7) Resting electrocardiography results (values 0, 1, and 2)
   - Value 0: normal
   - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8) Maximum heart rate achieved
9) Exercise induced angina (1 = yes; 0 = no)
10) The slope of the peak exercise ST segment
    - Value 1: up sloping
    - Value 2: flat
    - Value 3: down sloping
11) Num: diagnosis of heart disease (angiographic disease status)
    - Value 0: < 50% diameter narrowing
    - Value 1: > 50% diameter narrowing

B. Attribute Types:

Continuous, Discrete and Nominal

IV. DATA MINING MODEL

Experiments are conducted using Weka tool and 10-fold cross validation methods is used to minimize any bias in the process and improve the efficiency of the process. Weka tool is a workbench for machine learning algorithms written in Java. The classify panel enables the user to apply classification algorithms to the Heart disease dataset, to estimate the accuracy of predictive model, and to visualize erroneous predictions, or the model itself.

The classifiers like Naive Bayes, J48Decision Tree and Bagging algorithm etc were implemented in WEKA. These predictive models provide ways to predict whether a patient having heart disease or not. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

A. Advantages of using WEKA:

WEKA is a popular machine-learning workbench, which provides a unified interface, collects the classic machine learning algorithm and data pre-processing tools to variety of real-world problems.

- It is an Open Source toolkit, which can be modified by its users according to their requirements.
- It contains many algorithms
- Free (most other Data Mining tools are very expensive)
- Constantly under development (not only by the original designers)

V. METHODOLOGY AND EXPERIMENTAL RESULT

The goal of the prediction methodology is to design a model that can bring characteristic of predicted class from combination of other data and to evaluate the performance of the classification algorithms on diagnosis of heart disease problems using a variety of performance metrics.

The task of data mining in this research is to build models for prediction of the class based on selected attributes. The research applies the following algorithms:

A. ID3 (Iterative Dichotomiser 3):

In decision tree learning, is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The decision tree is used to classify future dataset. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node.

B. J48:

C4.5 algorithm is a classification algorithm producing decision tree based on information theory. C4.5 is from Ross Quinlan (known in Weka as J48 J for Java). J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data.

C. Naive Bayes:

Naive Bayes is a statistical classifier. It is based on Bayes's rule and "naively" assumes independence- it is only valid to multiply probabilities when the events are independent. Naïve Bayes is based on supervised learning.

D. Simple CART:

The term *Classification And Regression Tree (CART)* analysis is an umbrella term used to refer to both of the classification tree and regression tree, first introduced by Breiman et al. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split. Simple Cart (Classification and regression tree) is a classification technique that generates the binary decision tree.

E. Logistic Regression:

Logistic regression refers to methods for describing the relationship between a categorical response variable and a set of predictor variables (Larose, 2005). Logistic regression describes a function of mean (which is a probability) as a function of the exploratory variables. The function of mean it uses is the logit function. It assumes that the relationship between the response and the predictor is a non-linear. It produces linear segmentation of classes.

F. Bagging:

Bagging means Bootstrap aggregation which is an ensemble method to classify the data with good accuracy. Bagging can be applied on neural networks, Bayesian algorithms, Rule based algorithms, Support vector machines, Associative classification, and Distance based methods and Genetic Algorithms.

We analyse heart data set by applying different data mining algorithm. Table I shows the experimental result. We have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting heart patients.

Table I Performance of the Classifiers

Classification Technique	**Accuracy**
Naïve bayes	80%

Bagging	80%
Id3	88%
J48	84%
Simple cart	82%
Logistic regression	82%
REPTree	84%

As accuracy is very important in the field of medical domain, the performance measure accuracy of classification is considered in this study. So Id3 classifier has more accuracy than other classifiers.

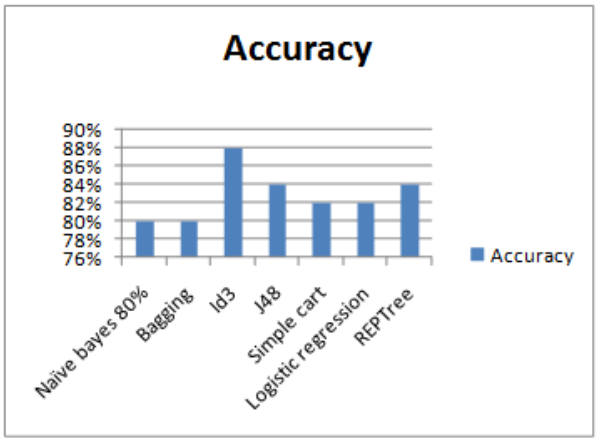

Figure 1: Graphical Representation of Classifiers.

Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results are shown in Tables II.

Table II TRAINING AND SIMULATION ERROR

Classifiers	**Evaluation Criteria**				
	KS	**MAE**	**RMSE**	**RAE**	**RRSE**
Simple cart	0.6349	0.2429	0.3853	48.9968 %	77.3668 %
Bagging	0.6008	0.2777	0.3683	56.0274%	73.9484%
Naïve Bayes	0.5976	0.3077	0.39	62.078%	78.3005%
Id3	0.7576	0.111	0.3438	22.3947%	65.012%
J48	0.6755	0.203	0.3675	40.9614%	73.7904%
Logistic regression	0.6393	0.2794	0.3946	56.3777%	79.2291%
REPTree	0.6765	0.244	0.3752	49.2352%	75.3389%

KS= Kappa statistic
MAE= Mean absolute error
RMSE= Root mean squared error
RAE= Relative absolute error
RRSE= Root relative squared error

Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity.

VI. CONCLUSIONS

In this paper, different classifiers are analysed to find the best classifier for predicting heart disease. The experiment helps health care professionals in the diagnosis of heart disease. A comparative study is conducted in this paper for Heart disease medical databases. Results have shown that most of decision tree based methods implemented have outperformed i.e. Id3 method. An added advantage of decision tree based methods is that it is easier to produce interpretability for the medical practitioners and may help in both the validation of the method and in developing further knowledge of the problem.

REFERENCES

[1] A.PRIYANGA and Dr.S.PRAKASAM The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness [Journal] // International Journal of Computer Science and Engineering Communications-IJCSEC.. - December 2013. - 1 : Vol. 1.

[2] Divya Tomar and Sonali Agarwal A survey on Data Mining approaches for Healthcare [Journal]. - [s.l.] : International Journal of Bio-Science and Bio-Technology, 2013. - 5 : Vol. 5. - pp. 241-266.

[3] Beant Kaur and Williamjeet Singh Review on Heart Disease Prediction System using Data Mining Techniques [Journal] // International Journal on Recent and Innovation Trends in Computing and Communication. - 10 : Vol. 2. - pp. 3003 – 3008. - ISSN: 2321-8169.

[4] Chandna Deepali Diagnosis of Heart Disease Using Data Mining Algorithm [Journal] // (IJCSIT) International Journal of Computer Science and Information Technologies. - 2014. - 2 : Vol. 5. - pp. 1678-1680. - ISSN:0975-9646.

[5] Chaurasia Vikas Early Prediction of Heart Diseases Using Data Mining Techniques [Journal]. - [s.l.] : Caribbean Journal of Science and Technology, 2013. - Vol. 1. - pp. 208-217.

[6] Hlaudi Daniel, Masethe Mosima and Anna Masethe Prediction of Heart Disease using Classification Algorithms [Journal] // World Congress on Engineering and Computer Science(WCECS). - October 22-24, 2014. - Vol. II.

[7] K.Sudhakar and Dr. M. Manimekalai Study of Heart Disease Prediction using Data Mining [Journal] // International Journal of Advanced Research in Computer Science and Software Engineering. - January 2014. - 1 : Vol. 4. - ISSN: 2277 128X.

[8] Miss. Chaitrali S. Dangare and Dr. Mrs. Sulabha S. Apte A DATA MINING APPROACH FOR PREDICTION OF HEART DISEASE USING NEURAL NETWORKS [Journal]. - [s.l.] : INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY(IJCET), October - December (2012). - 3 : Vol. 3. - pp. pp. 30-40. - ISSN.

[9] Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining [Journal]. - [s.l.] : International Journal of Applied Engineering Research(IJAER), 2012. - Vol. 7. - ISSN 0973-4562.

[10] Mohammed Abdul Khaleel, Sateesh Kumar and Pradham G.N. Dash A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases [Journal]. - [s.l.] : International Journal of Advanced Research in Computer Science and Software Engineering, August 2013. - 8 : Vol. 3. - ISSN: 2277 128X.

[11] Nidhi Bhatla and Kiran Jyoti An Analysis of Heart Disease Prediction using Different Data Mining Techniques [Journal] // International Journal of Engineering Research & Technology (IJERT). - October - 2012. - 8 : Vol. 1. - ISSN: 2278-0181.

[12] Parvathi I and Siddharth Rautaray Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain [Journal]. - [s.l.] : International Journal of Computer Science and Information Technologies(IJCSIT), 2014. - 1 : Vol. 5. - pp. 838-846. - ISSN:0975-9646.

[13] R. Chitra and V. Seenivasagam REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES [Journal]. - [s.l.] : ICTACT JOURNAL ON SOFT COMPUTING, July 2013. - 04 : Vol. 03. - ISSN: 2229-6956.

[14] Taneja Abhishek Heart Disease Prediction System Using Data Mining Techniques [Journal] // ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY. - December 2013. - 6 : Vol. 4. - pp. 457-466. - ISSN: 0974-6471.

[15] Vikas Chaurasia and Saurabh Pal Data Mining Approach to Detect Heart Dieses [Journal]. - [s.l.] : International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2013. - 4 : Vol. 2. - pp. 56-66. - ISSN: 2296-1739.

