A Study on Implementation of different Data Mining Techniques on Healthcare

Shabeena T.
Assistant Professor,
Dept of Computer Science, Calicut, Kerala, India

ABSTRACT
Data mining is one of the richest areas of research that is more popular in health organizations. It is the process of pattern discovery and extraction where huge amount of data is involved. The data generated by the health organizations are very vast and complex. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data. Disease prediction plays an important role in data mining. More data mining classification algorithms like decision trees, neural networks, Bayesian classifiers, Support vector machines, etc are used to diagnosis the heart diseases. The aim of this paper is to summarize some of the current research on predicting heart diseases using different data mining techniques, analyze the various combinations of mining algorithms used and conclude which technique(s) are effective and efficient.

Keywords
Heart disease, Data Mining, Decision Tree Techniques, Naive Bayes, Neural Networks.

1. INTRODUCTION
In the past decade, heart disease has been the leading cause of death in different countries in the world. In today's modern world, cardiovascular diseases are the highest flying diseases and in every year 12 million deaths occur over the world due to heart problem.In cardiovascular disease ,the heart and blood vessels are affected and as a result of which the blood is not pumped and circulated properly throughout the body parts. Cardiovascular disease includes coronary artery diseases such as angina and myocardial infarction. In coronary artery disease, the heart does not get sufficient blood that it requires because of cholesterol and fat that is deposited inside the wall of the arteries that supply the blood to heart. In myocardial infarctions which is also known as a heart attacks in which the path in the coronary artery is blocked due to the clotting of blood on the wall of the artery that supply the blood to the heart. In angina, chest pain occurs due to inadequate supply of blood to the heart as a result of which it does not function properly[1]. There are various factors which increases the risk of heart disease. Some of them are high blood pressure, cholesterol, family history of heart disease, obesity, hypertension, smoking, etc.

Data mining combines statistical analysis, machine learning algorithms and database technology to extract hidden patterns and relationships from large databases [2]. Nowadays, many hospitals keep their present data in an electronic form through some hospital database management system. This data may be used to extract meaningful information which may be used for decision making. Data mining concentrates on finding meaningful patterns from large datasets . Researchers have been applying different data mining techniques to help health care professionals with improved accuracy in the diagnosis of heart disease. Neural network, Naïve Bayes, Decision Tree and classification via clustering are some techniques used. The aim of all these techniques is to achieve better accuracy and to make the system more efficient so that it can predict the chances of heart attack .

Prediction is done by inserting a few attributes that be a symptom of heart disease and assign the weight to each attribute[3]. More weight is assigned to an attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions. The healthcare industry collects large amounts of health-care data and that need to be mined to discover hidden information for effective decision making.

2. LITERATURE REVIEW
Literature review was conducted inorder to obtain knowledge of previous researches.Many papers have been implemented using different Data mining techniques.for diagnosis of heart disease such as Naives Bayesian classification,Support Vector Machines(SVM),Decision Trees,Artificial Neural Network (ANN) etc showing different levels of accuracies[4]. One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success.

A system for classification of myocardial heart disease from ultrasonic images by optimizing the fuzzy membership functions by using genetic algorithm based method is proposed by Tsai and Watanabe. In this method by using the texture features obtained from ultrasonic images, the gaussian distributed membership function is constructed and the genetic algorithm based fuzzy classifier is used in classification. In this technique 96% of classification accuracy is achieved.

Genetic algorithm is also used in another approach by Anbarasi et al. where number of tests that are to be conducted by patient is reduced by determining the attributes that involved in the prediction of heart disease. In this approach three classifiers were used and these classifiers were fed with reduced attributes, but the system takes more time for model construction.

Haung and Wang have proposed a system using Support Vector Machine where feature selection and optimization of Support Vector Machine parameter is done using genetic

algorithm. This system uses less number of input parameters for support vector machine and it is evaluated on 11 real world datasets with improved accuracy of 89.6% .

In year 2013, S. Vijiyaraniet. al. performed a work, an efficient Classification Tree Technique for Heart Disease Prediction. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

B. Venkatalakshmi and M.V Shivsankar in year 2014 performed an analysis on heart disease diagnosis using data mining techniques Naïve Bayes and Decision Tree techniques. Different sessions of experiments were conducted with the same datasets in WEKA 3.6.0 tool. Data set of 294 records with 13 attributes was used and the results revealed that the Naïve Bayes performed better than the Decision tree techniques.

Due to the higher accuracy and learning rate the Artificial Neural Network k(ANN) algorithms can also be used in the prediction of heart disease . Kumaravel et al. have proposed automatic diagnosis system for heart diseases using neural network. In this system ECG data of the patients is used to extract features and 38 input parameters are used to classify 5 major types of heart diseases with accuracy of 63.6 - 82.9%.

This paper analyzes the classification tree techniques in data mining. The objective of this research is to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

3. METHODOLOGY

3.1 Data mining Tools:

Data mining tools provide ready to use implementation of the mining algorithms. Most of them are free open source software's so that researchers can easily use them. They have an easy to use interface. Some of the popular data mining tools are WEKA, RapidMiner, TANAGRA, MATLAB etc[5].

3.1.1 Weka :

This research uses Weka as data mining tool.WEKA stands for Waikato Environment for Knowledge Analysis.It is a data mining tool written in java developed at Waikato,New Zealand. Weka is a very good data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics. Explorer, Experimenter, Knowledge Flow and Simple CLI are the interfaces available in Weka that has been used by us.

A. Explorer: The explorer interface has several panels like preprocess, classify, cluster, associate, select attribute and visualize. But in this interface our main focus is on the Classification Panel.

B. Experimenter: This interface provides facility for systematic comparison of different algorithms on basis of given datasets. Each algorithm runs 10 times and then the accuracy is reported.

C. Knowledge Flow: It is an alternative to the explorer interface. The only difference between this and others is that here user selects Weka component from toolbar and connects them to make a layout for running the algorithms.

D. Simple CLI: Simple CLI means command line interface. User performs operations through a command line interface by giving instructions to the operating system. This interface is less popular as compared to other three.

For comparing various Decision Tree classification techniques, Cleveland dataset from UCI repository is used, which is available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset has 76 attributes and 250 records. However, only 13 attributes are used for this study & testing.The information of UCI repository is regularly introduced in a database or spreadsheet. In order to use this data for WEKA tool, the data sets need to be in the ARFF format (attribute-relation file format). WEKA tool is used to pre-process the dataset. With thorough comparison between different decision tree algorithms within WEKA tool and deriving the decisions out of it, would help the system to predict the likely presence of heart disease in the patient and will definitely help to diagnose heart disease well in advance and able to cure it in right time.

The following steps are performed in WEKA.

Start the WEKA Explorer.

Open CSV dataset file and save in ARFF format

Click on Classify tab. This is the area for running algorithms against a loaded dataset in Weka.

Click the "**Start**" button to run the algorithm.

After running the algorithm, you can note the results in the "Classifier output" section.

4. DATA MINING TECHNIQES USED FOR PREDICTION:

4.1 Artificial Neural Networks(ANN):

An artificial neural network is a computational model that attempts to account for the parallel nature of the human brain. It is a network of highly interconnecting processing elements (neurons) operating in parallel. These elements are inspired by biological nervous systems.

Artificial neural networks work as a leading tool that helps doctors to evaluate, model and get sensible results from complex data. Most applications of artificial neural networks in medicine are diagnostic systems, biomedical analysis, image analysis, drug development. Feed-forward neural networks are widely and successfully used models for classification, forecasting and problem solving. A typical feed-forward back propagation neural network is proposed for disease diagnosis.

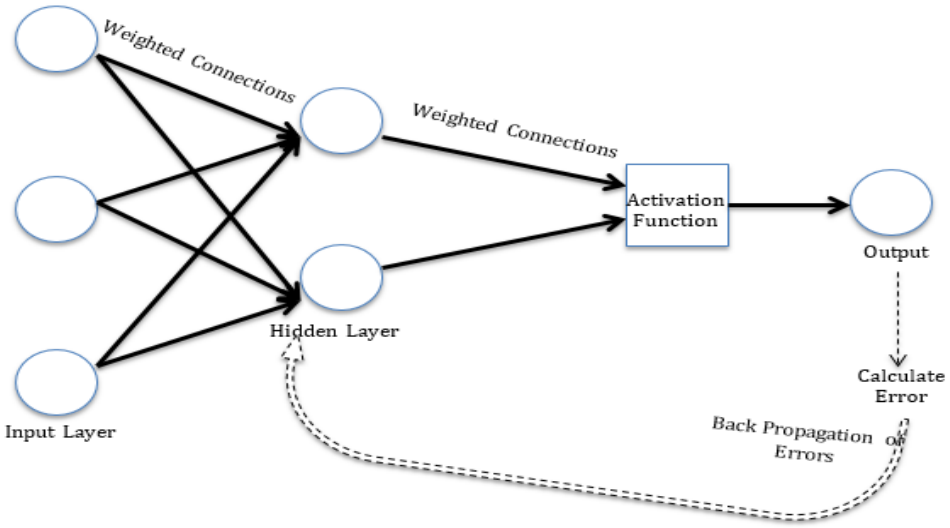

Fig.4.1.1.Principle of Artificial Neural Network

Feed- forward neural network, is an emulation of biological neural system. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of Yj = f (Σ wji xi) Where f is a simple threshold function such as sigmoid[6].

4.2 Naive Bayes:

Naïve bayes is based on machine learning and data mining methods. It is used to create the prediction model. It shows the predictable state for probability of each attribute. This model produces a more efficient output compare with other output. The main advantage of naïve bayes algorithm is that it requires a small amount of training data to estimate the parameter for classification. it is capable of calculating the most probable output depend on the input. it is easy to add new data at runtime for a better classifier[7].

Naive Bayes algorithm is preferred in the following cases.

- When the dimensionality of data is high
- When the attributes are independent of each other.
- When we expect more efficient output, as compared to other methods output.
- Exhibits high accuracy and speed when applied to large databases.

Steps:

- Convert the dataset into frequency table.
- Create likelihood table by finding the probability.
- Naïve bayes to calculate the posterior probability of each class
- The class with highest priority probability is the outcome of prediction

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$.

Naive Bayes classifier considers that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

$$P(c|x) = P(x1/c)\times P(x2/c)\times\ldots..\times P(xn/c)\times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute) of class.
- P(c) is called the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor of given class.
- P(x) is the prior probability of predictor of class.

4.3 Decision Trees:

Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. Decision Trees (DTs) are a non-parametric supervised learning method used for classification[8]. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The structure of decision tree is in the form of a tree. Tree size should be relatively small that can be controlled by using a technique called pruning. Decision trees classify instances by starting at the root of the tree and moving through it until a leaf node. This technique is commonly used in operations research, mainly in decision analysis . Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition. It can handle input data like Nominal, Numeric and Text. It is able to process erroneous datasets or missing values.

Decision tree is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label. The node at the top most labels in the tree is called root node. There are many popular decision tree algorithms ID3, C4.5, CART, and J48

4.3.1 ID3:

ID3 stands for Iterative Dichotomiser 3 is an algorithm introduced by Ross Quinlan utilized to make a decision tree. ID3 adopt a greedy (i.e. nonbacktracking) approach in which decision trees are constructed in a top-down recursive divide and conquer manner. The resulting tree is used to classify the future samples.

4.3.2 C4.5:

C4.5 is the latest adaption of ID3 induction algorithm. It builds decision tree from a set of training data in the same way as ID3 using the concept of information entropy.C4.5 is often called as Statistical Classifier. It can be used as a training tool to train nurses and medical students to diagnose patients with heart disease.

4.3.3 CART: -

Classification and regression trees (CART) are a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. It was invented by Breiman in 1984.

4.3.4 J48:

J48 decision tree is the implementation of ID3 algorithm, developed by WEKA project team. J48 is a simple C4.5 decision tree for classification. With this technique, a tree is constructed to model the classification process.

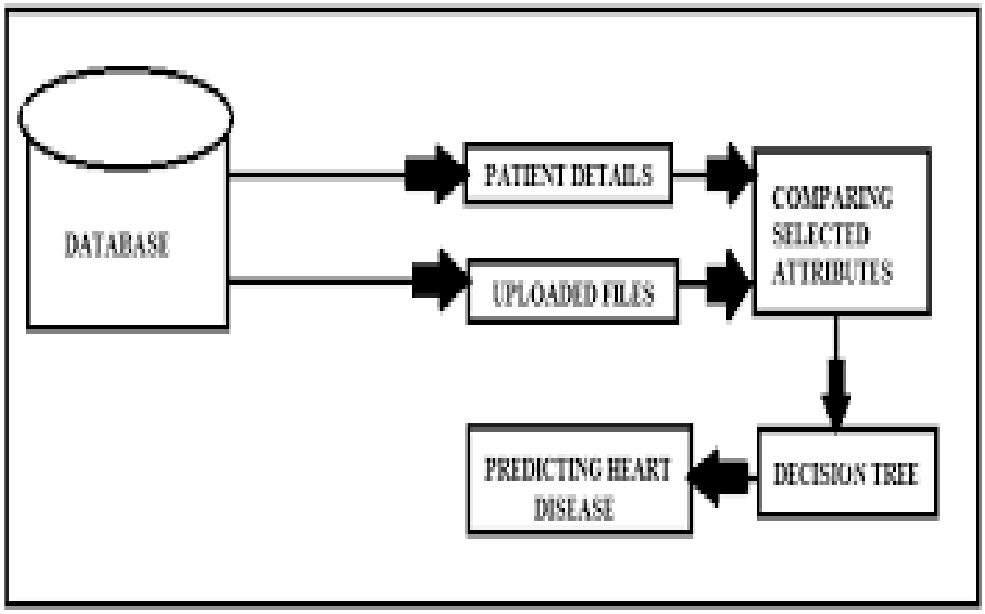

Fig.4.3.1.implementation of ID3 algorithm

5. DATA ANALYSIS

Different Classification Tree Algorithms Used are:

Decision Tree, Artificial Neural Network and Naïve Bayes

A total of 250 records were obtained from the Cleveland Heart Disease Database. The dataset consist of 3 types of attributes. Input, Key and Prediction attributes. Commonly used attributes such as Age, Gender, Blood pressure, Pulse rate ,Cholesterol etc are considered as input attributes of which Age and gender are non-modifiable attributes . Age is continuous and dynamic in nature where gender is static and constant. The other parameters have continuous and Random Values. In order to get more appropriate results additional attributes such as Smoking and history of heart diseases also are included in this study. Smoking and history of heart diseases are the Modifiable attributes. Constant values are given to the smoking and history of heart disease to predict the risk rate of heart disease. Patient id is considered as a key attribute which is unique for each and every user. Using this key attribute the patient and doctor can retrieve the record. Authenticity of the user is taken care by the application. The Prediction Attribute found the risk level of the disease. The risk level is classified into three levels namely low risk, high risk and normal risk which indicates lesser than 50%, greater than 50% and 0 respectively. The Dataset has 76 attributes . However, only 13 attributes are used for this study and the results are shown in the following Table.

Table 5.1: Classification accuracy of Naïve Bayes Decision Trees and Neural Network algorithms

Classification Techniques used	Accuracy(%)with 13 attributes
Neural Networks	99.14
Decision Trees	96.2
Naïve Bayes	94.13

6. CONCLUSION

By analysing the experimental results, it is concluded that Neural Networks performed better in predicting the heart disease with more accuracy followed by Decision Trees and Naïve Bayes respectively. As a future work, I will use the research described here as a foundation for the development of effective prediction system to enhance medical care .That being said, this is a large topic and there are numerous opportunities for additional research that would significantly extend the functionality of the current research. For example: Considering other kinds of diseases that are biologically related to heart diseases. In my future work, I would like to explore different rules such as

Association, Clustering, K-means etc for better efficiency and ease of simplicity.

7. REFERENCES

[1] S. Palaniappan and R. Awang, "Intelligent heart disease prediction systemusing data mining techniques," pp. 108–115, 2008.

[2] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling schemes for heart disease classification," Applied Soft Computing,vol. 14, pp. 47–52, 2014.

[3] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," pp. 173–177, 2012.;3

[4] K.Srinivas, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks". (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255, 2010.

[5] Prajakta Ghadge,Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh, 2016,"Intelligent Heart Attack Prediction System Using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology,Vol. 2, Issue 2, pp.73-77, October 2015–March.

[6] K.Sudhakar, and Dr. M. Manimekalai, January 2014, "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1,pp. 1157-1160.

[7] G.-G. Wang, M. Lu, Y.-Q. Dong, and X.-J. Zhao, "Self-adaptive extreme learning machine," Neural Computing and Applications, pp. 1–13, 2015.

[8] Combination data mining methods with new medical data to predictingoutcome of coronary heart disease," in Convergence InformationTechnology, 2007. International Conference on. IEEE, 2007, pp. 868–872.