ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 4, April 2016

A Study on Data mining Classification Algorithms in Heart Disease Prediction

Dr. T. Karthikeyan[1], Dr. B. Ragavan[2], V.A.Kanimozhi[3]

Abstract: **Data mining (sometimes called knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. By using data mining techniques, in medical field it takes less time for the prediction & detection of the disease with more accuracy. Main objective of this study is to give an analysis on various data mining classification algorithms used for the prediction of heart disease.**

Index Terms: **Data mining, Heart disease prediction, Knowledge Discovery Process, Medical Data mining, Classification algorithms.**

I. INTRODUCTION TO MEDICAL DATA MINING

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. But due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining and analytic strategies.

Huge and complex volumes of data are generated by healthcare activities; un-automated analysis has become impractical; DM can generate information that can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices. [10]

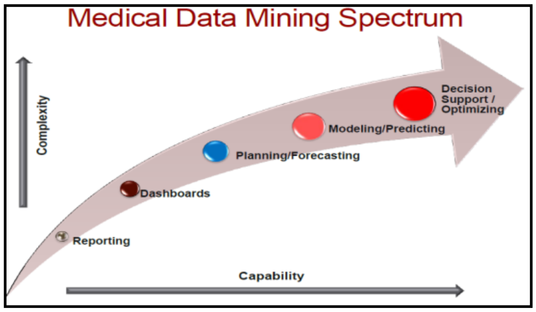
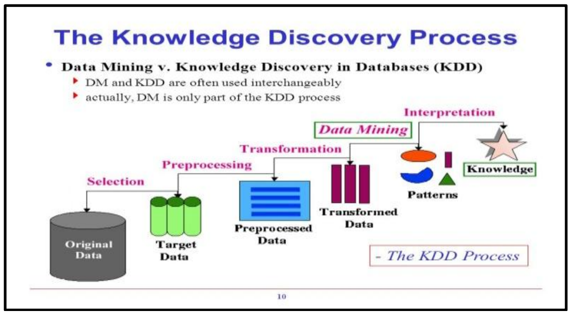


Fig. 1 The Medical Data Mining Spectrum

The following figure shows important steps in the Knowledge Discovery process:


The Knowledge Discovery Process

- Data Mining v. Knowledge Discovery in Databases (KDD)
  - DM and KDD are often used interchangeably
  - actually, DM is only part of the KDD process

Fig. 2 Steps in Data Mining Process

II.LITERATURE SURVEY

Table 1: Shows various data mining techniques used in heart disease prediction with accuracy

Author	**Purpose**	**Techniques Used**	**Accuracy**
Ms.shtake S.H & Prof.Sanap S.A. [1]	The aim of this paper is to Develop a prototype Intelligent heart disease prediction system.	Decision Tree	94.93%
		Naive Bayes	95%
		Neural Networks	93.54%
Chaitrali S.Dangare [2]	This paper has analysed prediction systems for Heart disease using more no.of input attributes.	Decision Tree	90%
		Naive Bayes	99.62%
		Neural Networks	100%
Jyoti Soni [3]	The focus is on using different algorithms and	Decision Tree	89%
		Naive Bayes	86.53%

	combinations of several target attributes for intelligent and effective heart attack prediction using data mining.	Neural Networks	85.53%
AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin [4]	The aim of this paper is to develop a heart disease prediction system that can assist medical professionals in predicting heart disease status based on the clinical database.	Neural Networks	80%
Vikas Chaurasia, [5]	The aim of this paper is to develop prediction models for heart disease survivability.	CART	83.49%
		ID3	72.93%
		Decision Table	82.50%
Andrea D'Souza [6]	The aim of this paper is to Heart Disease Prediction using data mining techniques.	Neural Networks	79.38%
		K-Means Clustering	63.29%
Milan Kumari [7]	The aim of this paper is to analyze various Data mining techniques on cardiovascular disease dataset.	Decision Tree	79.05%
		Neural Networks	80.06%
		SVM	84.12%
Abhishek Taneja [8]	The aim of this paper using various data mining techniques an attempt to assist in the diagnosis of the heart disease.	Naive Bayes	86.53%
		Decision tree	89%
		Neural Networks	85.53%
Sellappan Palaniappan Rafiah Awang [9]	Aim of this paper is to develop a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques	Naive Bayes	86.53%
		Neural Networks	85.53%
		Decision Tree	89%

III.DATA MINING ALGORITHMS:

Data mining algorithms provides a way to use various data mining tasks such as classification and clustering in order to predict solution sets for a problem.

Desirable feature of any efficient algorithm =>

- To reduce I/O operations and
- At the same time be efficient in computing

A. *Classification & Clustering & Prediction :*

By simple definition, in classification/clustering we analyze a set of data and generate a set of grouping rules which can be used to classify future data. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. [12]

Classification :

Classification is a form of supervised learning, where the class labels of some training samples are given, these samples are used as examples to supervise the learning of a classification model. 2 basic classification processes are Learning and Classification.

Learning : Training data are analyzed by a classification algorithm.

Classification: Test data are used to estimate the accuracy of the classification rule. If the accuracy is considered acceptable, the rules can be applied to the classification of new tuples.[12]

Clustering :

Clustering is a form of unsupervised learning, in which no class labels are provided. Instead data records need to be grouped based on how similar they are to other records.[12]

Prediction:

Prediction (or Regression) is an example of data mining task that fits in the predictive model of data mining. The use of regression tasks enable us to forecast future data values based on the present and past data values. [12]

IV.DATA MINING - CLASSIFICATION ALGORITHMS

A)DECISION TREE:

A decision tree is a classification scheme which generates a tree and set of rules, representing the model of different classes, from a given data set. The set of records available for developing classification methods is generally divided into 2 disjoint subsets – a training set and a test set. The former is used for deriving the classifier, while the latter is used to measure the accuracy of the classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified. [11]

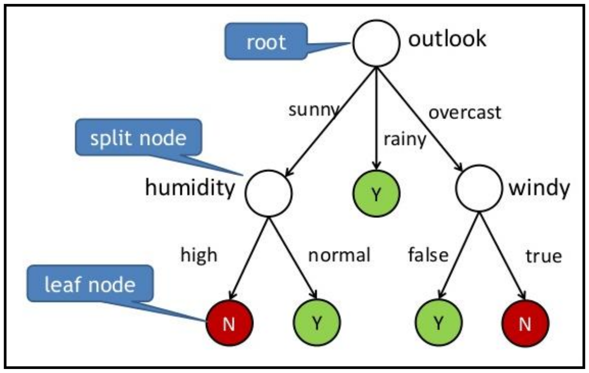


Fig 3: Sample Decision Tree

Algorithm Description :

The algorithm is called with 3 parameters :

D – Data distribution

Attribute list – List of attribute describing the tuples

Attribute selection method – It is a heuristic procedure for selecting the attribute that "best" discriminate the given tuples according to class.

- **Information Gain** – The expected information needed to classify a tuple in D is given by,

Info(D) = - $\sum_{i=0}^{m}$ $p_i \log_2(p_i)$

Where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|$ /|D|.

This Info(D) is also known as the entropy of D.

- **Gini index** – The Gini index is used in CART. This Gini index measures impurity of D, a data partition or set of training tuples, as Gini(D) = 1 - $\sum_{i=1}^{m}$ p_i^2

Where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|$ /|D|.[11]

B)ASSOCIATION RULES

The problem of deriving associations from the data was formulated by Agarwal. This is referred as "Market Basket Problem". The task is to find relationships between the presences of various items within these baskets. It can help the retailer develop marketing strategies by gaining insight into matters like "Which items are most frequently purchased by customers.

Algorithm Description :

For a given transaction database T, an association rule is an expression of the form X=>Y, where X and Y are subsets of A and X=>Y holds with confidence tow, if tow% oftransactions in D X also support Y. The rule X=>Y has support sigma in the transaction set T if sigma% of transactions in T support X U Y.

The intuitive meaning of such a rule is that a transaction of the database which contains X tends to contain Y. Given a set of transactions, T, the problem of mining association rules is to discover all rules that have support and confidence greater than or equal to the user specified minimum support and minimum confidence, respectively.

Association rule is a process to search relationships among data items in a given data set.[11]

X,Y = itemsets

X=>Y - Association rules.

C) Naïve Bayes :

Naïve Bayes is one of the simplest classifiers and works well for many applications especially those involving text classification. Bayesian classifiers are statistical classifiers. Bayesian classification is based on Bayes' theorem.

Algorithm Description :

Let D be a training set of tuples and there are m classes, $C_1,C_2,\ldots.C_m$. Given a tuple, X the classifier will predict that X belongs to the class having highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$P(C_i|X) > P(C_j | X)$ for $1\leq j\leq m$, j=i. [11]

Thus we maximize $P(C_i | X)$. The class C_i for which $P(C_i$)for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem, $\mathbf{P(C_i|X)}$ is

$$\frac{\boldsymbol{P(X|\mathrm{Ci})P(\mathrm{Ci})}}{\boldsymbol{P(X)}}$$

D) Support Vector Machines:

Support Vector Machines are learning machines that can perform binary classification and regression estimation tasks. It is an algorithm that work as follows: It uses a non linear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (that is, a "decision boundary" separating the tuples of one class from another).

With an appropriate non linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors("essential" training tuples) and margins (defined by the support vectors).

Algorithm Description :

The training data is converted into n-dimensional data using non-linear transformation method. Then the algorithm searches for the best hyper-plane to separate transformed data into two different classes. SVM performs classification tasks by maximizing the margin of the hyper-plane separating both classes while minimizing the classification errors.[11]

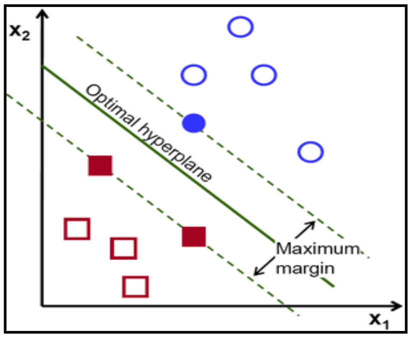


Fig 4: Support Vector Machine

E) Neural Network :

A neural network is a set of connected input/output units where each connection has a weight associated with it. During the learning process, the network learns by adjusting the weights to predict the correct class label of the input samples.

Artificial neural networks mimic the pattern-finding capacity of the human brain and hence some researchers have suggested applying Neural Network algorithms to pattern-mapping. Neural networks have been applied successfully in a few applications that involve classification.

Typical NN structure for classification:

- One output node per class
- Output value is class membership function value

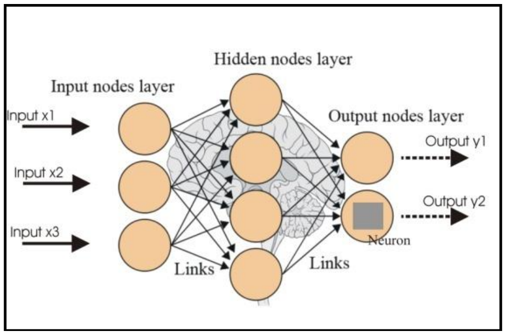


Fig 5: Neural Network

Neural network Issues

• Number of source nodes , Number of hidden layers, Training data, Number of sinks, Interconnections, Weights, Activation Functions, Learning Technique, When to stop learning. [12]

Table 2: Shows various Criteria for comparing Classification algorithms.

Criteria	Classification
Accuracy	The accuracy of classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data.
Speed	This refers to the computational cost in generating and using the classifier.
Robustness	It refers to the ability of classifier to make correct predictions from given noisy data.
Scalability	Scalability refers to the ability to construct the classifier efficiently; given large amount of data.
Interpretability	It refers to what extent the classifier understands.

V.ADVANTAGES AND DISADVANTAGES OF DATA MINING CLASSIFICATION ALGORITHMS

Data Mining Algorithms	Advantages	Disadvantages
Decision Tree	The Decision trees are able to generate understandable rules. They are able to handle both numerical and categorical attributes They provide a clear indication of which fields are most important for prediction or classification	Some decision trees can only deal with binary – valued target classes. Others are able to assign records to an arbitrary no.of classes, but are error-prone when the no.of training examples per class gets small. This can happen rather quickly in a tree with many levels and/or many branches per node. The The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field is examined before its best split can be found.
Association Rules	This algorithm was initially developed for retail store transaction mining It is perfect for categorical (non-numeric) data and it involves little more than simple counting.	This algorithm is for discovering frequent sets are not directly suitable, when the underlying database is incremented intermittently. Discovery of poorly understandable rules
Naïve Bayes	Bayesian classifier has the minimum error rate in comparison to all other classifiers. Easy to implement, less model complexity	The main disadvantage is that it can't learn interactions between features. In classification task we need a big data set in order to make reliable estimations of the probability of each class. We can use Naïve Bayes classification algorithm with a small data set but precision and recall will keep very low.
Support Vector Machines	Prediction accuracy is generally high Unlike the other other classification techniques SVM is minimize the expected error rather than minimizing the classification error. SVM is employ the duality theory of mathematical programming to get a dual problem that admits efficient computational methods. Works well with fewer training samples (number of support vectors do not matter much).	Problem need to be formulated as 2-class classification Difficult to understand the learned function (weights). Learning takes long time (QP Optimization).
Neural Network	Neural network are quite simple to implement. They are often used for generalization they're good at find solutions that often solve the problem but in some rare cases won't.	Neural Networks cannot be retrained. If you add data later, this is almost impossible to add to an existing network. Handling Handling of time series data in neural networks is a very complicated topic.

VI. CONCLUSION

In this paper a study among classification's algorithms have been carried out, these are (Decision tree and Bayesian network, neural network, association rules, support vector machine algorithms. In this all algorithms, decision Tree's algorithms have less error rate and it is the easier algorithm as compared to Association rule and Bayesian. Each and every algorithm has their own advantages and disadvantages. Based on our needs and criteria like size of data set, time duration, etc., we can choose any one of the above classification algorithms.

VII. REFERENCES

[1] Ms. Ishtake S.H , Prof. Sanap S.A. "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, April 2013.

[2] Chaitrali S. Dangare Sulabha, " Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, " Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[4] AH Chen, SY Huang, PS Hong, CH Cheng, EJ lin, "HDPS : Heart Disease Prediction System, Computing in cardiology", 2011 : 38:557-560.

[5] Vikas Chaurasia, Saurabh Pal, " Early Prediction of Heart Diseases using Data Mining Techniques", Caribbean Journal of Science & Technology, ISSN 0799-3757.

[6] Andrea D'Souza, "Heart Disease Prediction Using Data Mining Techniques", International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364, ISSN (Print): 2320-9356.

[7] Milan Kumari, Sunila Godara, " Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction" , International Journal of Computer Sci ence and Technology, IJCST Vol. 2, Iss ue 2, June 2011, ISSN : 2229 - 4333 (Print) | ISSN : 0976 - 8491 (Online).

[8] Abhishek Taneja, "Heart Disease Prediction System Using Data Mining Techniques", Oriental Journal Of Computer Science & Technology, ISSN: 0974-6471 December 2013, Vol. 6, No. (4).

[9] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August 2008.

[10] MEDICAL DATA MINING, Timothy Hays, PhD, Health IT Strategy Executive, Dynamics Research Corporation (DRC), December 13, 2012.

[11] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" Second Edition, University of Illinois at Urbana-Champaign.

[12] "Data mining" Typical data mining process for predictive modeling by BPB Publications.

VIII. AUTHORS PROFILE

First Author: Associate Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu, India.

Second Author: Associate Professor, Department of Biochemistry, PSG College of Arts & Science, Coimbatore, Tamilnadu, India.

Third Author: Research Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore, Tamilnadu, India.