# Binary Classifier For Prediction of Heart of Heart Disease Using Decision Tree

[1]Shwetali R. Rajguru , [2]Dr. prof. M. A. Pradhan, [3]Nikita R. Kurkure

[1]Student, [2]Senior Professor, [3]Student
[1]Computer Engineering,
[1] All India Shri Shivaji Memorial Society
College of Engineering, Pune-01, India

**Abstract :** This research paper deals with efficient data mining procedure for prediction of heart disease from medical records of patients. Heart disease is very common disease nowadays in all populations and in all age groups. In this approach binary classifier is used for detection of heart disease using decision tree. The dataset used is Statlog dataset from UCI Repository. The J48 algorithm is used.

***IndexTerms*** **- —Heart Disease, Decision tree, C4.5(J48) algorithm, Binary Classifier,Classification**

## I. INTRODUCTION

There are several database management systems for manipulating the data,but extraction of information or knowledge from data is complex than mere data manipulation.This technique includes a number of phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. Data mining has proven to be very beneficial in the field of medical analysis as it increases diagnostic accuracy, to reduce costs of patient treatment and to save human resources [1] . Data mining is a field which is a combination of machine learning, statistics, database technology**.** The J48 algorithm is used for classification.

.2. PROPOSED CLASSIFIER

Mathematically, a classifier can be represented as a function which takes features in p dimensional search space and assigns a label vector Lvc to it.[8]

C: $S_p$->$L_{vc}$

Where,

C is the Classifier which maps search space to label vectors

$S_p$ is the p dimensional search space

$L_{vc}$ is set of label vectors

The objective is to create ' C' using Decision Tree. During the training phase of the classifier, samples of the form {X1,X2,X3,X4,X5,X6,X7,X8} ε $S^8$ and associated label vector Lvc = {absence(1), presence(2)} are used to create classifier. Classification of heart data set is binary classification problem. i.e. there are only two label vector that can be assigned to each of the sample.

## 3. PREPROCESSING

The prediction accuracy of any classification algorithm is heavily dependent on the quantity of the data. The clinical data is generally collected as a result of the patient care activity to benefit the individual patient. So the clinical dataset may contain data that is redundant ,incomplete, Imprecise and inconsistent .Not all features used in describing data are predictive .The irrelevant or redundant features ,noisy data etc. affect the predictive accuracy .Hence, the clinical data requires rigorous preprocessing. Pre-processing of data is often a neglected but important stage in the data mining process.

**a) Normalization** : It require when the attribute values are not in specified range .It scales all the values for a given attribute such that they fall within a specified small range. It helps to avoid giving the undue significance to attributes having large range.

**b) Cross validation**: The standard way of predicting the error rate of learning technique given a single, fixed sample of data is to use stratified tenfold cross-validation. The data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in the full dataset.

**c) Sampling**: It used for data selection. Processing the entire dataset is time consuming and expensive .So a sample of dataset is taken which represent same properties as the original data. The different types of sampling are simple random sampling are simple random sampling, sampling without replacement, sampling with replacement and stratified sampling.

**d) Feature creation**: It creates the new attribute from existing attribute .The purpose is to capture the important information in a dataset more efficiently than the original attributes. Three main methods used for feature creation are:

- Feature extraction
- Mapping data to new space
- Feature construction

**e) Feature selection:** In data mining, feature selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons:

Simplification of models to make them easier to interpret by researchers/users,

- shorter training times,
- To avoid the curse of dimensionality,
- enhanced generalization by reducing over fitting (formally, reduction of variance)

Feature selection techniques should be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). Archetypal cases for the application of feature selection include the analysis of written texts and DNA microarray data, where there are many thousands of features, and a few tens to hundreds of samples. The central premise when using a feature selection technique is that the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information. Redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated. We have used correlation-based feature selection algorithm for selecting most predictive attributes.

## 4. CORRELATION-BASED FEATURE SELECTION ALGORITHM (CFS)

The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other"[6][7]. The following equation gives the merit of a feature subset S consisting of k features:

$$\mathbf{Merit}_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}.$$

Where,

$S_k$ is the heuristic evaluation function " merit" of a feature subset S containing k features,

$r_{cf}$ is the mean feature-class correlation(F€ S), and

$r_{ff}$ is the average feature-feature inter correlation.

The CFS criterion is defined as follows:

$$\mathrm{CFS} = \max_{S_k}\left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \cdots + r_{f_if_j} + \cdots + r_{f_kf_1})}}\right].$$

The $r_{cfi}$ and $r_{fifj}$ variables are referred to as correlations, but are not necessarily Pearson's correlation coefficient or Spearman's ρ. Dr. Mark Hall's dissertation uses neither of these, but uses three different measures of relatedness, minimum description length (MDL), symmetrical uncertainty, and relief. The CFS algorithm takes the training dataset as input and generates as input and generates a feature class and feature-feature correlation matrix which is then used to search the feature subset space. The criterion used for search is the best first search .The search starts with an empty set of features single feature expansion are created and the highest evaluated subset is expanded by adding single attribute .If expanding a subset result into improvement ,the search drop backs to next unexpanded subset and continues from there. CFS ranks feature subsets according to correlation based heuristic evaluation function. The subset that contain features that are highly correlated with the class and uncorrelated with each other are ranked as per evaluation function.

## 5. J48 DECISION TREE

We have implemented the classification algorithm J48. A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. Each node of the tree implements a decision rule that splits the examples into two or more partitions. New nodes are created to handle each of the partitions and a node is considered terminal or leaf node based on a stopping criteria. This approach to decision tree construction corresponds to a top-down greedy-algorithm that makes locally optimal decisions at each node. The sequence of the decisions made from the root node to the eventual labeling of a test input is easy to follow and also the approach can be extended to non-numeric domains where the attributes are categorical rather than numerical. [2, 3]

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found [4]. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable [5].

### 5.1 Basic Steps in the Algorithm: [4]

(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.

(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

**5.2 Features of the Algorithm**

(i) Both the discrete and continuous attributes can be handled by this algorithm. For handling continuous attributes threshold value is decided by C4.5. This value divides the data list into those who have their attribute value below the threshold and those having more than or equal to it.
(ii) This algorithm also handles the missing values and noisy data in the training data.
(iii) After the construction of tree, this algorithm performs the pruning of the tree.

**6. EXPERIMENTAL RESULTS**

The statlog heart disease datasets used for the traing and testing of the binary classifers.The performances obtained using the Statlog (Heart) dataset from the UCI machine learning database are compared in this context.

**6.1 Heart Disease Dataset**

| | | |
|---|---|---|
| 1 | Chest Pain: | 1.typical angina<br>2.atypical angina<br>3.non-anginal pain<br>4.asymptomatic |
| 2 | Resting electrocardiographic results: | 0:normal 1:having ST-T wave abnormality,2:showing probable or definte left venticular hypertrophy by Estes's criteria |
| 3 | Maximum Heart Rate achieve | 71-103,104-136,137-169,170-202 |
| 4 | Exercise induced angina | Yes,no |
| 5 | Old peak = ST depression induced by exercise relative to rest | 0-1.5,1.6-3.1,3.2-4.7,4.8-6.2 |
| 6 | Number of major vessels colored by fluoroscopy | 0,1,2,3 |
| 7 | Thal | Normal, fixed defect, reversible defect |

*Table 1 . Feature of Dataset*

**6.2 Statistical analysis of the attributes**

| Attribute | Mean | Std.Deviation | Maximum | Minimum |
|---|---|---|---|---|
| Age | 54.433 | 9.109 | 77 | 29 |
| Sex | 0.678 | 0.468 | 1 | 0 |
| Chest | 3.174 | 0.95 | 4 | 1 |
| Resting bps | 131.344 | 17.862 | 200 | 94 |
| Serum cholesterol | 249.659 | 51.686 | 564 | 126 |
| Fasting blood sugar | 0.148 | 0.356 | 1 | 0 |
| Resting ECG | 1.022 | 0.998 | 2 | 0 |
| Heart Rate | 149.678 | 23.166 | 202 | 71 |
| Exang-exercise | 0.33 | 0.471 | 1 | 0 |
| Oldpeak | 1.05 | 1.145 | 6.2 | 0 |
| Slope | 1.585 | 0.614 | 3 | 1 |
| No. of vessels | 0.67 | 0.944 | 3 | 0 |
| Thal | 4.696 | 1.941 | 7 | 3 |

*Table 2. Analysis of attributes*

**7. RESULT & DISCUSSION**

**7.1 Confusion Matrix**

Other than the classifier's overall predictive accuracy on unseen instances, it is often helpful to see a breakdown of the classifier's performance i. e. how frequently instances of class X were correctly classified as class X or misclassified as some other class. This information is given in confusion matrix. The confusion matrix is in the form of table. The body of the table has one row and one column for each possible classification. The rows correspond to the predicted classification. The value in the $i^{th}$ row and $j^{th}$ column gives the number of instances for which the correct classification is in $i^{th}$ class which are classified as belonging to the $j^{th}$ class. When the dataset has only two classes, one class is regarded as a class of principle interest and it is referred as positive and the others as negative. For proposed binary prediction models we have considered as presence of heart disease.

The values in the cells of the confusion can be labeled as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). True positive (TP) is the no of positive instances that are classified as positive. True Negative (TN) is

the no of instances that are classifies as negative. False Negative (FN) is the no of positive instances that are classified as negative.

## Results Respective to Accuracy:

**Before Applying CFS algorithm:**

| Actual Class | Predicted Class | |
|---|---|---|
| | Absence | Presence |
| Absence of Heart Disease | 119 | 31 |
| Presence of Heart Disease | 32 | 88 |

*Table 3. Confusion matrix for absence class*

| *Actual Class* | *Predicted Class* | |
|---|---|---|
| | *Presence* | *Absence* |
| *Presence of Heart Disease* | *88* | *32* |
| *Absence of Heart Disease* | *31* | *119* |

*Table 4. Confusion matrix for presence class*

| False Negative rate | | False Positive rate | Sensitivity or Recall | Precision or positive Predicted value | Specificit-y | F measur-e | Error Rate | Classificatio-n Rate or Accuracy |
|---|---|---|---|---|---|---|---|---|
| Absence | 0.207 | 0.267 | 0.793 | 0.788 | 0.733 | 0.791 | 23.33% | 76.66% |
| Presence | 0.267 | 0.207 | 0.733 | 0.739 | 0.793 | 0.767 | 23.33% | 76.66% |

*Table 5. Performance Measure Summary*

**After Applying CFS algorithm:**

| Actual Class | Predicted Class | |
|---|---|---|
| | Absence | Presence |
| Absence of Heart Disease | 128 | 122 |
| Presence of Heart Disease | 29 | 91 |

*Table 6. Confusion matrix for absence class*

| Actual Class | Predicted Class | |
|---|---|---|
| | Presence | Absence |
| Presence of Heart Disease | 91 | 29 |
| Absence of Heart Disease | 22 | 128 |

*Table 7. Confusion matrix for presence class*

| False Negative rate | | False Positive rate | Sensitivity or Recall | Precision or positive Predicted value | Specificit-y | F measur-e | Error Rate | Classificatio-n Rate or Accuracy |
|---|---|---|---|---|---|---|---|---|
| Absence | 0.147 | 0.242 | 0.853 | 0.815 | 0.758 | 0.834 | 18.88% | 81.11% |
| Presence | 0.242 | 0.147 | 0.758 | 0.805 | 0.853 | 0.781 | 18.88% | 81.11% |

*Table 8. Performance Measure Summary*

## 8.Conclusion

In this paper, we presented an approach for the binary classification problem using J48 decision tree. From these works, it can be observed that Normalization, cross-validation, sampling and feature selection methods can improved the performance of single classifier algorithms in diagnosing heart disease.

## REFERENCES

**[1]** Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology And Secured Transactions, 2012 InternationalConferece For, pp. 471-472. IEEE, 2012.

**[2]** J. Han and M. kambar, " Data Mining:Concepts and Techniques" ,
Morgan Kaufman Publishers,(2004)

**[3]** M. Singh, P. K. Wadhwa and P. S. Sandhu ," Human Protein Function Prediction Using Decision Tree Induction" , International Journal of Computer Science and Network Security,vol.7, No.4, (2007), pp. 92-98. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

**[4]** Korting, Thales Sehn. "C4. 5 algorithm and Multivariate Decision Trees." Image Processing Division, National Institute for Space Research--INPE [5] Devasena, C. Lakshmi, et al. "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set." Bonfring International Journal of Man Machine Interface 1.Special Issue Inaugural Special Issue (2011): 05-09

**[5]** Nadali, A; Kakhky, E.N.; Nosratabadi, H.E., "Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system," Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.6, no., pp.161,165, 8- 10 April 2011

**[6]** M. Hall 1999, "Correlation-based Feature Selection for Machine Learning" .

**[7]** Senliol, Baris, et al. "Fast Correlation Based Filter (FCBF) with a different search strategy." Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on. IEEE, 2008.

**[8]** Madhavi Pradhan , G. R. Bamnote. "Efficient Binary Classifier for prediction of Dibetes using Data Preprocessing and Support Vector Machine".