IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 6 (Jul. - Aug. 2013), PP 83-86
www.iosrjournals.org

A Heart Disease Prediction Model using Decision Tree

Atul Kumar Pandey[1],Prabhat Pandey[2],K.L. Jaiswal[3],Ashish Kumar Sen[4]
[1]*Assistant Professor of Computer Science, Department of Physics, Govt. PG Science College, Rewa(M.P.)-India,*
[2]*OSD, Additional Directorate, Higher Education, Division Rewa (M.P.)-India,*
[3]*Assistant Professor and In charge of BCA, DCA & PGDCA, Department of Physics, Govt. PG Science College, Rewa(M.P.)-India-486001,*
[4]*Assistant Professor, Department of Mathematics & Computer Science, Govt. PG Science College, Rewa(M.P.)-India,*

Abstract—*In this paper, we develop a heart disease prediction model that can assist medical professionals in predicting heart disease status based on the clinical data of patients. Firstly, we select 14 important clinical features, i.e., age, sex, chest pain type, trestbps, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise induced angina, old peak, slope, number of vessels colored, thal and diagnosis of heart disease. Secondly, we develop an prediction model using J48 decision tree for classifying heart disease based on these clinical features against unpruned, pruned and pruned with reduced error pruning approach.. Finally, the accuracy of Pruned J48 Decision Tree with Reduced Error Pruning Approach is more better then the simple Pruned and Unpruned approach. The result obtained that which shows that fasting blood sugar is the most important attribute which gives better classification against the other attributes but its gives not better accuracy.*

Keywords—*Data mining, Reduced Error Pruning, Gain Ratio and Decision Tree.*

I. Introduction

The diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data; this complexity leads to the excessive medical costs affecting the quality of the medical care [1]. According to the statistic data from WHO, one third population worldwide died from heart disease; heart disease is found to be the leading cause of death in developing countries by 2010. It shows one third American adult have one or more types of heart diseases based on American Heart Association report. Computational biology is often applied in the process of translating biological knowledge into clinical practice, as well as in the understanding of biological phenomena from the clinical data. The discovery of biomarkers in heart disease is one of the key contributions using computational biology. This process involves the development of a predictive model and the integration of different types of data and knowledge for diagnostic purposes.

Furthermore, this process requires the design and combination of different methodologies from statistical analysis and data mining [2,3].

In the past decades, data mining have played an important role in heart disease research. To find the hidden medical information from the different expression between the healthy and the heart disease individuals in the existed clinical data is a noticeable and powerful approach in the study of heart disease classification. Heart disease classification provides the critical basis for the therapy of patients. Statistics and machine learning are two main approaches which have been applied to predict the status of heart disease based on the expression of the clinical data [4,5].

Data mining (DM) is the core stage of knowledge discovery in databases (KDD), which is a "nontrivial extraction of implicit, novel, and potentially useful information from data" [6]. It applies machine learning and statistical methods in order to discover areas of previously unknown knowledge. As a rule, the KDD process involves the following steps: data selection, data pre-processing, transformation, DM (induction of useful patterns), and interpretation of results. Several data mining techniques are used in the diagnosis of heart disease such as naïve bayes, decision tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies.

One of the most common classification models is the decision tree, which is a tree-like structure where each internal node denotes a test on a predictive attribute and each branch denotes an attribute value. A leaf node represents predicted classes or class distributions [7]. An unlabeled object is classified by starting at the topmost (root) node of the tree, then traversing the tree, based on the values of the predictive attributes in this object. Decision-tree techniques assume that the data objects are described by a fixed set of attributes, where each predictive attribute takes a small number of disjoint possible values and the target (dependent) variable has discrete output values, each value representing a class label.

There are several known algorithms of decision tree induction: ID3 - which uses information gain with statistical pre-pruning, C4.5, an advanced version of ID3, and probably the most popular decision-tree algorithm [8], ART, which minimizes a cost-complexity function, See5 - which builds several models and uses unequal misclassification costs, and IFN – Info-Fuzzy Network which utilizes information theory to minimize the number of predictive attributes in a decision-tree model [9] [10]. In [9], the IFN algorithm is shown empirically to produce more compact models than C4.5, while preserving nearly the same level of classification accuracy.

II. Methods

A. Data sources

In this paper, we use the heart disease data from machine learning repository of UCI [11]. We have total 303 instances of which 164 instances belonged to the healthy and 139 instances belonged to the heart disease. 14 clinical features have been recorded for each instance.

B. Features description

Table 1 shows the 14 clinical features and their description.

. Table 1- Clinical features and their description

Clinical features	Description
Age	Instance age in years
Sex	Instance gender
Cp	Chest pain type
Trestbps (mmHg)	Resting blood pressure
Chol (mg/dl)	Serum cholesterol
Fbs	Fasting blood sugar
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Oldpeak	ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment
Ca	Number of major vessels (0-3)colored by flourosopy
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Num	Diagnosis of heart disease

In Table 1, there are 14 attributes used in this system, including 8 symbolic and 6 numeric: age (age in years), sex (male, female), Chest pain type (typical angina, atypical angina, non-angina pain, asymptomatic), Trestbps (resting blood pressure in mm Hg), cholesterol (serum cholesterol in mg/dl), fasting blood sugar < 120 mg/dl (true or false), resting electrocardiographic results (normal, having ST-T wave abnormality, showing probable or definite left ventricular hypertrophy by Estes' criteria), max heart rate, exercise induced angina (true or false), oldpeak (ST depression induced by exercise relative to rest), slope (up, flat, down), number of vessels colored by fluoroscopy (0-3), thal (normal, fixed defect, reversible defect), and class (healthy, with heart-disease).

III. Decision Tree

The decision tree type used in this research is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain [15]. Information gain [12] is the difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top ranked features are chosen as the potential attributes used in the classifier. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula [14, 15]:

$$\mathbf{E} = \sum_{i=1}^{k} \mathbf{Pi}\ \mathbf{log_2 Pi}$$

Where k is the number of classes of the target attributes Pi is the number of occurrences of class i divided by the total number of instances (i.e. the probability of i occurring). To reduce the effect of bias resulting from the use of information gain, a variant known as gain ratio was introduced by the Australian academic Ross Quinlan [15]. The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values [13]. Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain Ratio = Information Gain / Split Information

Where the split information is a value based on the column sums of the frequency table [15].

IV. Results and Discussion

We demonstrate here the usefulness of the prediction model to the clinical data of heart disease where training instances 200 and testing instances 103 using split test mode.

Table 2- summarizes the results generated from different approches.

	J48 Unpruned tree	J48 Pruned tree	J48 Reduced Error Pruning
Percent Correct	72.82	73.79	75.73
Number Correct	75	76	78
IR Precision	0.68	0.72	0.74
No. of Rules	14	33	11
Tree Size	24	56	17
F Measure	0.78	0.77	0.79

After extracting the decision tree rules, reduced error pruning was used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small decision rules [16]. Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

```
thal = fixed_defect
|  ca <= 0: <50 (6.12/2.06)
|  ca > 0: >50_1 (6.0)
thal = normal
|  exang = no: <50 (93.56/15.0)
|  exang = yes
|  |  cp = typ_angina: <50 (2.0)
|  |  cp = asympt
|  |  |  ca <= 0: <50 (5.56/2.55)
|  |  |  ca > 0: >50_1 (7.0)
|  |  cp = non_anginal: <50 (2.0)
|  |  cp = atyp_angina: <50 (2.0)
thal = reversable_defect
|  cp = typ_angina: >50_1 (5.0/2.0)
|  cp = asympt: >50_1 (50.39/4.0)
|  cp = non_anginal
|  |  slope = up: <50 (4.39/1.0)
|  |  slope = flat: >50_1 (10.0/4.0)
|  |  slope = down: <50 (1.0)
|  cp = atyp_angina: <50 (7.0/3.0)
```

Fig 1- J48 pruned tree with reduced error pruning approach

V. Conclusion

This work will provide better diagnosis the heart patients than the previous. We develop a heart disease prediction model that can assist medical professionals in predicting heart disease status based on the clinical data of patients. Firstly, we select 14 important clinical features, i.e., age, sex, chest pain type, trestbps, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise induced angina, old peak, slope, number of vessels colored, thal and diagnosis of heart disease. Secondly, we develop a prediction model using J48 decision tree for classifying heart disease based on these clinical features against unpruned, pruned and pruned with reduced error pruning approach. Finally, the accuracy of Pruned J48 Decision Tree with Reduced Error Pruning Approach is more better then the simple Pruned and Unpruned approach. The result obtained that which shows that fasting blood sugar is the most important attribute which gives better classification against the other attributes but its gives not better accuracy.

References

[1.] Wu R, Peters W, Morgan MW. The next generation clinical decision support: linking evidence to best practice. J Healthc Inf Manag, 2002; 16:50-5.

[2.] Thuraisingham BM. A Primer for Understanding and applying data mining. IT Professional 2000; 1:28-31.

[3.] Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. Global Journal of Computer Science and Technology 2010; 10:38-43.

[4.] Anbarasi M, Anupriya E, Iyengar NCHSN. Enhanced prediction of heart Disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology 2010; 2:5370-76.

[5.] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. International Journal of Computer Science and Network Security 2008; 8:343-50.

[6.] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann, 2006.

[7.] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[8.] J.R. Quinlan: C4.5, Programs for MachineLearning, Morgan Kaufmann, 1993.

[9.] M. Last and O. Maimon, "A Compact and Accurate Model for Classification", IEEE Transactions on Knowledge and Data Engineering 2004; 16, 2: 203-215.

[10.] O. Maimon and M. Last, Knowledge Discovery and Data Mining – The InfoFuzzy Network (IFN) Methodology, Kluwer Academic Publishers, Massive Computing, Boston, December 2000.

[11.] UCI Machine Learning Repository [homepage on the Internet]. Arlington: The Association; 2006 [updated 1996 Dec 3; cited 2011 Feb 2]. Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[12.] Kwong-Sak Leung,kin hong Lee,Jin-Feng Wang,Eddie Y.T.Ng,Henry L.Y.Chan,Stephen K.W.Tsui,Tony S.K. Mok,Pete Chi-Hang Tse, Joseph Jao-yui Sung, Data Mining on DNA Sequences of Hepatitis B virus. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol 8,No 2,March/April 2011.

[13.] 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.

[14.] Han, j. and M. Kamber, Data Mining Concepts and Techniques. 2006: Morgan Kaufmann Publishers. Lee, I.-N., S.-C. Liao, and M. Embrechts, Data.

[15.] Bramer, M., Principles of data mining. 2007: Springer.

[16.] Esposito, F., D. Malerba and G. Semeraro, A Comparative Analysis of method for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 997. Vol. 19, No. 5.