ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Comparative Study of Detecting the Congenital Heart Disease Using Tree Algorithms.

M Karthi[1*], L Mary Gladence[2], and J Joshua[3].

[1]Department of Information Technology, St.Joseph's Institute of Technology, Chennai, Tamil Nadu, India.
[2]Faculty of Computing, Department of Information Technology, Sathyabama University, Chennai, Tamil Nadu, India.
[3]ETL-Developer Build, Preludesys india Ltd, Chennai-600063, Tamil Nadu, India.

ABSTRACT

Heart disease is one which mentioned for a large number of medical conditions related to heart. Latest statistical research summaries that the number of heart patients are becoming more as twice after 20 years especially in developing country like India. The main aim of our proposed system is to separate the lower risk patients with higher risk one. The J48 Decision Tree algorithm is used to develop the automatic classifier using high level attributes related to heart disease[1-18]. Also the performance metrics are compared with Bagging and CART algorithms. Our proposed classifier has given the highest performance in terms of performance measures such as True Positive, FalsePositive and some other accuracy terms like Precision and Recall

Keywords: Data mining, Bagging Algorithm, J48Decision Tree, Attributes

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Congestive Heart Failure is a critical heart disease condition. It is generally called as CHF. Recent researches have shown that CHF rate increase because of malfunction of cardiac. Due to this heart can't pump required blood to the body. Once heart couldn't function, it properly automatically person health leads to illness. CHF is measured with the help of New York Heart Association using NYHA rate scale which is used to prove the risk factor of the patient's death. The Heart rate variability is measured by the dissimilarity between heartbeats and these measures are predicted using Electrocardiographic signal.It is normally predicted by Electro Cardio Graphic source (ECG) within certain limits. HRV rate is studied in patients who are all affected by CHF. Many research studies represented that HRV term identify the risk factor which leads to death. Also, a new proposed classifier is working with long term HRV rate recorded from CHF patient's [14]. Data mining is used to find hidden patterns from clinical studies. Based on the report of the World Health Survey, seventeen million peoples death is due to heart failure. So we proposed an enhanced method which automates the diagnosis system for CHF (Congestive Heart Failure) and form a decision tree, to separate lower risk patients from higher risk patients. Congestive Heart Failure (CHF) is a one of the critical condition due to this heart can't function properly. Once this condition occurs for human, heart cannot pump the blood to the entire body. To avoid these circumstances, this work has proved which classifier is suitable for person those who are suffering from heart malfunctioning. This system can assist disease diagnosis and effective treatment.

The procedure discussed here is used to identify which one is suitable for heart related issues and which extent it can be avoided. The proposed classifier checks the risk, whether it is low or high using the procedure which we used in this work. Patients are given with their situation stating that whether they are in starting stage,Middle stage or at last stage.

RELATED WORK

To have a clear idea about the proposed work referred some papers related to the proposed work. Those paper's description is described below.

L. Mary Gladence, T. Ravi [4] has developed a New Sequential Pattern Mining Algorithm to mine the Customer behavior. This algorithm used the concept of Genetic Algorithm Procedures. Proposed work finds the customer behavior in terms of new algorithms and discovered patterns are classified based on the support count and its similarity. The performance results show that work done here, reduces the time complexity as well as it will help the authorized person to understand how customer buying trends changes day by day. Using this concept, they can easily optimize their work.

Zhaohui, Xiaoyun, Rohini Srihari [7] has concentrated on how entire work can be finished within the framework which they have developed. Moreover, they explored the usefulness of their work and through this, how is it is useful for the patients. This work is tested with the Multinomial Bayesian Belief Network as well as Regression Classifiers. Even though, positive and negative attributes are given according to the input it has given the best results. It is tested with imbalanced data to check whether the proposed work perform well or not.

Shereen Fouad and Peter Tino [9] Learning Using privileged Information (LUPI) is referred.It improves the supervised learning in the presence of privileged information.This information is available during the training phase, but not in the test phase. This Novel learning methodology is designed to assimilate privileged information in ordinal classification tasks, where there should be a natural order. Here global metric in the input space is changed by priviledged information which are revealed based on distance information. Experiments demonstrated that integrating privileged information via the proposed ordinal-based metric learning can improve the ordinal classification performance.

Paolo Melillo, Leandro Pecchia Nicola, De Luca Marcello Bracale [14] have developed a heart failure diagnosis scoring model and these has been tested with patients records and their results are shown clearly. Comparing to previous studies, this work has given the labeling of whether particular person comes under which category by almost forty percentages more. These are described clearly while we look into results. Finally accuracy has reached up to Eighty Eight percentages where previous work accuracy is only seventy five percentages. Work described here is attainable for hear failure's early stage. F.

Syed Umar Amin, Kavita Agarwal,[15] Earlier researches done their heart disease prediction with six attributes, but this work is with only four attributes. Diagnosed result is based on fuzzy rules and are classified as Normal, Low Risk, Medium Risk, High Risk. Final results, keep in the database for calculating the efficiency and for the record maintenance of the patients.

PROBLEM STATEMENT

According to the previous researches many people who die because of heart diseases is very difficult to tell when a person might die because of heart disease. So our work developed the classifier for patients who are all suffering from heart related problem and this issue is tested using 11 attributes while the previous work concentrated on eleven features.They are Chest pain type (Cp), Slope, CA , Thalach , Fasting blood sugar , cholesterol, Exang, Restecg, Resting blood pressure Old peak , Thal. The Chest pain type ,Resting blood sugar, Height, Weight are general causes for CHF and we choose the other attributes like Exang, Oldpeak, Slope, CA, Thal are extracted from the long term HRV that is EGC record rate.

After taking all the eleven attributes, the required patient's data are entered in the above mentioned attribute values and then the result is displayed. Results shows in tree format and check whether the person is having congestive heart failure (CHF) or not. CHF occur wherever heart function pump is inadequate to deliver the efficient oxygen rich blood to the body. If a person is having CHF that will weaken the heart muscles, make severer tiredness, some breathing problem, etc. To determine whether the heart beat is correct or nor, heart rate variation is used . Therefore, the "goal" of work is to identify patients who are suffering from heart disease. To perform this work, data set which I have used is extracted from uci repository. The objective of the proposed work is by applying the predictive data mining techniques to the heart disease patients for diagnosing &also for the treatment of heart disease.. There are many trees based modeling is used. To solve this problem we chose a three data mining techniques such as J48 Decision Tree, CART, Bagging algorithm.

CLASSIFICATION TECHNIQUES

CART

CART is a not an assumption based decision tree learning technique. CART supports categorical as well as numerical value to produce the results in the form of trees. CART algorithm has Tree growing and Tree pruning stages. Tree growing is used in the generation of purer child nodes by splitting a tree. Tree pruning helps to omit unwanted nodes. Many functions are proposed to measure of impurity of each node. To perform Tree growing as well as Tree Pruning Cross-Validation procedure is used. From those functions here, Gini Index is used for binary classification. Here, the output is displayed in the form of tree like structure where initial nodes represent testing part and leaf nodes ,which are displayed in the last level represent the results.

BAGGING

To improve the performance of classification, machine learning algorithms are used. Bagging algorithm is one of the prominent machines learning classification algorithm which is used to improve the stability & accuracy of algorithms by subsamples. The decision tree is plotted from these output samples using voting. So the final decision tree is one with major priority. This is the only method which can be combined with any method. Generally Bagging model produces different structures based on the input given. It reduces the complex procedures as well as it improves the accuracy. Comforts behind this method are

- Level of accuracy which can be provided by this method cannot be obtained by a large single tree model.
- Creating a single decision tree from a collection of tree is not difficult.
- It helps in avoiding the problem of over fitting since random samples are used.

ISSN: 0975-8585

J48 DECISION TREE

To perform classification, it first create a decision tree using available training data's attribute. Whenever the training set is prepared, identify attribute and that should discriminates instances clearly. This feature will clearly classify the data instance with highest information gain. If there is any inexactness occurs, terminate that and assign target values which we have obtained. There are some standards and terms which work well with J48 Decision trees are:

- Result of prediction output and actual value is identical, then it is referred as True Positive Rate
- Result of prediction output and actual value is not identical, then it is referred as False Positive Rate
- Exactness or quality measure is called Precision
- Completeness or quantity measure is called Recall

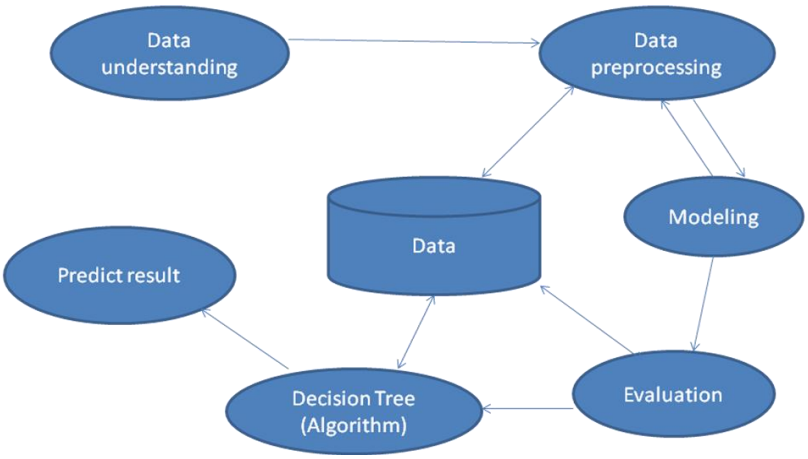

Figure 1: The Structure of our proposed method to predict Heart Disease

RESULTS &DISCUSSION

The overview of the CHA classification model is represented in" Fig 1".In overview, the various steps are used for predicting the heart disease such as Preprocessing, Modeling, Evaluation. Comparison of various data mining techniques with respect to 11 and 13 attributes is shown in "Table.3"that includes both low and high level features of a data set. The proposed method achieved higher performances compared to other methods. Initially to fill the missing values in the dataset, we use the Replace Missing Values Filter method of Weka Tool. Using this all the attribute values are filled. Here preprocessing method i.e Data cleaning is used. After cleaning process, Classification approaches such as J48 Decision tree, CART and Bagging algorithm were applied. The minimum number of decision trees, which improves the classification result performed by the bagging method, was found.

In our experiment the decision tree having 2 classes, they labeled as A and B .Using the 2x2 confusion matrixes the final results are described in the form of the class label. They are Class A and Class B .Where the class A denotes that the patient has heart disease. (ie) YES .Also the class B denotes that the a patient has no heart disease (ie) NO

These results and Binary Search Trees are clearly displayed in the "Fig.2".The leaf nodes in the binary tree are a pictorial representation is similar to If…then rule which are followed in Decision tree. To describe this, the leaf node 3 in the" Fig.2" is representing that: "if Rbs is lower than 151 ms2, the patient is treated as a low risk patient". The cross validation value is used to divide the training set into multiple part (N).(i.e) subsets of D/N size. From that one subset is selected as the training set. The remaining subsets have played the important role in test data sets. Mainly it applies to decision tree methods, but this algorithm may use with any method.

ISSN: 0975-8585

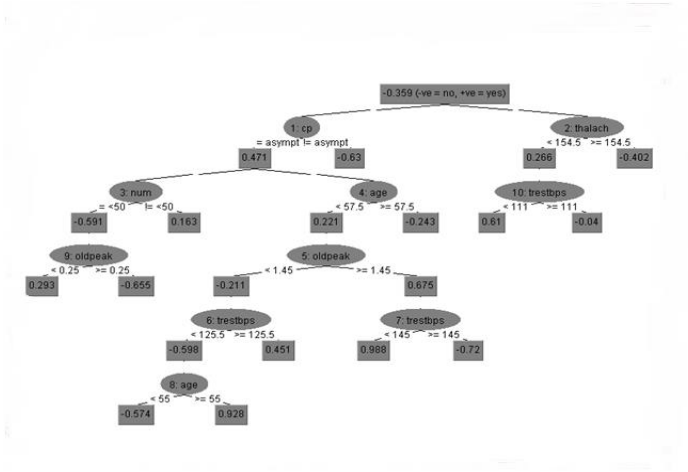

Figure 2: Final BST tree model of combination of attributes

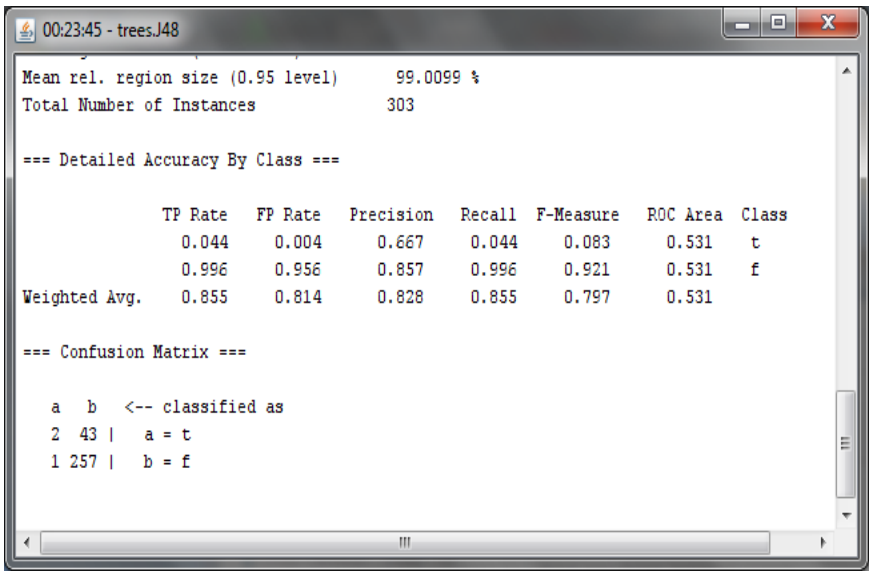

Figure 3. Final output of proposed model

Using this particular person is suffering from heart disease (CHF) or not has been predicted. In addition to this, the proposed work is compared with Classification and Regression Tree (CART) & Bagging algorithm with respect to the measures True Positive, False Positive, Precision (accuracy) and Recall and discovered values are described in Table 1. And the results are described in "Fig.3".Measures are calculated in the following manner.Suppose projected value and real value shows the same result ,it is referred as TruePositive.If the Projected value and Real values shows negative,it is referred ass TrueNegative.If Predicted value is ngative and Real value is Positive,it si referred as FalseNegative and Predicted value is Positive and Real value is Negative,it is referred as FalsePositive.“Fig.3” shows that how the confusion matrix is classified as based on J48 Decision tree algorithm. Here totally 303 instances are initially selected and our proposed method is applied. The final output is in the format of true-t, false-f. We represent those results as class labels A and B. The true (t) value means a class A, false (f) value means class B Also the new working model is having a less error rate compare to Bagging and CART methods.

ISSN: 0975-8585

Table 1: Classification Performance Measure

Classifier	TruePositive	FalsePositive	Precision	Recall	F-Measure
J48 DECISION TREE	0.855	0.814	0.828	0.855	0.792
CART	0.851	0.851	0.725	0.851	0.781
BAGGING ALGORITHM	0.848	0.852	0.725	0.848	0.782

Table 2: Results acquired with 11 attributes and 13 attributes

ALGORITHM USED	EXACTNESS	
	11 Attributes	13 Attributes
J48 DECISION TREE	85.47	85.14
CART	85.15	85.01
BAGGING ALGORITHM	84.81	84.4

CONCLUSION

In this study, we select the some tree based classification algorithm from various mining methods to construct a classification tree. The tree concludes that how to separate the patients based on their disease risk level . suggested work splits the least affected patients from most affected patients, by observing the heart rate variation for certain hours. At Initial part of the work the missing values are filled with Replace missing values filter using WEKA Tool. After this, tenfold cross validation matrix is used to reduce the misclassification error and tree like the decision is shown in the "Fig.2".After performing these,the proposed method results are compared with CART and Bagging Algorithm.This work shows the best performance in terms of measures like TruePositive, FalsePositive, and some accuracy terms like Precision and Recall.Through "Table.2"we can easily identify work which are done using eleven features has attained better result while compared it with thirteen features. This work extension can be done using time and frequency based feature selection which will consider the impact of environmental conditions in day to day life.

REFERENCES

[1] O'Donovan, Claire E., et al. *Cardiology in the Young* 2016; 26.01 : 100-109.

[2] Bhatnagar, Aakanksha, Shweta P. Jadye, and Madan Mohan Nagar. International Journal of Engineering Research &Technology (IJERT), ESRSA Publications 2012;1:9,pp1-3

[3] Pecchia, Leandro, Paolo Melillo, and Marcello Bracale, *IEEE Transactions* 2011; 58.3:pp 800-804.

[4] Gladence Mary L, and T. Ravi M.Karthi, International Journal of Applied Engineering Research ISSN 0973-4562-2014, 9: 21 pp. 8593-8602.

[5] G. Eason, B. Noble, &I. N. Sneddon capacity in clinical &research applications: An advisory from the committee onexercise, rehabilitation, &prevention, council on clinical cardiology, American heart association;Circulation, 2000; 102.13, pp. 1591–1597.

[6] MaryGladence L,Ravi T, Karthi M :Heart Disease Prediction using Naïve Bayes Classifier-Sequential Pattern Mining, the International Journal of Applied Engineering Research ISSN 0973-4562; 2014, 9:pp. 8593-8602

[7] zhaohui,xiaouyun,Rohini :ACM SIGMOD international conference on Management of data, 2007;6.1.

[8] MaryGladence L,Karthi M,MariaAnu V : A Statistical Comparison of Logistic Regression and different Bayes Classification Methods for Machine Learning ; ARPN Journal of Engineering and Applied SciencesISSN 1819-6608 ,2015: 10, pp 5947-5953

[9] Shereen Fouad S and peter Tiňo ; Ordinal-based metric learning for learning using privileged information; IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2013, pp. 1-8.

[10] MaryGladence L,Ravi T :Heart Disease Prediction and Treatment Suggestion, Research Journal of Pharmaceutical, Biological and Chemical Sciences ISSN: 0975-8585 , 7:2,pp 1274-1279

[11] RajendraAcharya U, Paul Joseph K, Kannathal, Lim C. M., & Suri J. S, Med. Biol. Eng. Comput.,2006 ; 44:12,pp. 1031–1051.

[12] Clerk Maxwell J, A Treatise on Electricity &Magnetism, A Treatise on Electricity &Magnetism- Oxford: Clarendon, 3-2, 1892, pp.68–73.

[13] MaryGladence L,Karthi M,Ravi T :A Novel Technique for Multi-Class Ordinal Regression-APDC; Indian Journal of Science&Technology, 2016;9:10,pp 1-6 .
[14] Laguna, Pablo, George B. Moody, and Roger G. Mark , IEEE TRANSACTION ;1998;45.6 : 698-715.
[15] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan ,Beg, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)2013; 2: 1 pp 218-223
[16] MaryGladence L,Ravi T : Mining the Change of Customer behavior with the aid of Similarity Computation Index(SCI) and Genetic Algorithm (GA); the International Review on Computers and Software(IRECOS), 2013 8.11:2552-2561
[17] MaryGladence L,Ravi T, Karthi M :An Enhanced Method For Detecting CongestiveHeart Failure-Automatic Classifier, IEEE International Conference on AdvancedCommunication Control and Computing Technologies (ICACCCT), 2014: pp 586-590.
[18] Deepika M, Mary Gladence L, MadhuKeerthana R , Research Journal of Pharmaceutical, Biological and Chemical Sciences ISSN: 0975-8585, 2016 7:1, pp808-814.