Efficient Heart Disease Prediction System using Decision Tree

Purushottam
Research Scholar (PhD.),
R.G.T.U. Bhopal
(M.P), India.
puru.mit2002@gmail.com

Prof. (Dr.) Kanak Saxena,
Prof & Head, Department of
Computer Application, S.A.T.I.
Vidisha (M.P), India.
kanak.saxena@gmail.com

Richa Sharma
Assistant professor
Amity University Uttar Pradesh
Noida,India
s.richa.sharma@gmail.com

***Abstract*— Cardiovascular disease (CVD) is a big reason of morbidity and mortality in the current living style. Identification of Cardiovascular disease is an important but a complex task that needs to be performed very minutely, efficiently and the correct automation would be very desirable. Every human being can not be equally skillful and so as doctors. All doctors cannot be equally skilled in every sub specialty and at many places we don't have skilled and specialist doctors available easily. An automated system in medical diagnosis would enhance medical care and it can also reduce costs. In this study, we have designed a system that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health. The rules can be prioritized based on the user's requirement.**

The performance of the system is evaluated in terms of classification accuracy and the results shows that the system has great potential in predicting the heart disease risk level more accurately.

Keywords*— *C4.5, Heart disease prediction System, , Decision tree, CVD, CAD.

I. INTRODUCTION

In today's time at many places clinical test results are often made based on doctors' intuition and experience rather than on the rich information available in many large databases. Many a times this process leads to unintentional biases, errors and a huge medical cost which affects the quality of service provided to patients.

Today many hospitals installed some sort of patient's information systems to manage their healthcare or patient data. These information systems typically generate large amounts of data which can be in different format like numbers, text, charts and images but unfortunately, this database that contains rich information is rarely used for clinical decision making. There is a lot of information stored in repositories that can be used effectively to support decision making in healthcare. This raises an important question:

"How can we turn data into useful information that can enable healthcare practitioners to make effective clinical decisions?" This is the main objective of this research.

II. RISK LEVEL PREDICTION FROM HEART DISEASE DATABASE

The extraction of risk level from the heart disease data base is presented in this section. The heart disease database contains the screening clinical data of heart patients. Initially, the database preprocessed to make the mining process more efficient.

III. HEART DISEASE DATASET DESCRIPTION

Source Information:

(a) Creators of the used dataset: V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. (b) Donor: David W. Aha (aha@ics.uci.edu) (714) 856-8779

The "num" attributes notify to the presence of heart disease in the patient. The range of this attribute is from 0 (no presence) to 4.

Most of the experiments associated with Cleveland database are focused on absence (Num" value 0) and presence ("Num" values from 1 to 4)

Due to personal security patient's personal identification information replaced with dummy values.

Number of Instances: Cleveland: 303

The directory contains a dataset related with heart disease diagnosis. The data was collected from the following locations:

Cleveland Clinic Foundation (cleveland.data).

The Cleveland database contains total 76 raw attributes, but in experiments only 14 of them is actually used. The dataset used in this experiment contains different important parameters like ECR, cholesterol, chest pain, fasting sugar, MHR (maximum heart rate) and many more.

The detailed information about these attributes and their domain range are as follows:

@relation Cleveland
@attribute age real [29.0, 77.0]
@attribute sex real [0.0, 1.0]
@attribute cp real [1.0, 4.0]
@attribute trestbps real [94.0, 200.0]
@attribute chol real [126.0, 564.0]
@attribute fbs real [0.0, 1.0]
@attribute restecg real [0.0, 2.0]
@attribute thalach real [71.0, 202.0]
@attribute exang real [0.0, 1.0]

ISBN:978-1-4799-8890-7/15/$31.00 ©2015 IEEE

@attribute oldpeak real [0.0, 6.2]
@attribute slope real [1.0, 3.0]
@attribute ca real [0.0, 3.0]
@attribute thal real [3.0, 7.0]
@attribute num {0, 1, 2, 3, 4}
@inputs age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
@outputs num

We have used the Classification model by covering rules (based on decision trees) as C4.5Rules [2],[3] on the above modified dataset and find out the generated rule sets with different priority.

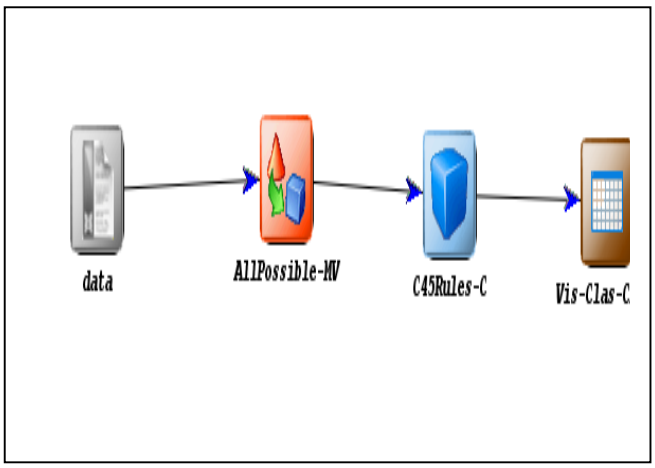
We have used WEKA tool [25] for dataset analysis and KEEL [24], [27] to find out the classification decision rules.

IV. EXPERIMENT DESIGN FORMAT WITH KEEL

We have used KEEL (Knowledge Extraction based on Evolutionary Learning) tool. KEEL is an open source (GPLv3) Java software tool to assess evolutionary algorithms for Data Mining problems.

We have designed an Experiment using the Cleveland dataset as given in the figure 1.In the preprocessing phase we have used an AllPossible-MV [1] algorithm to fill the missing values in the dataset.

Fig 1



A. *Generated Decision Tree*

```
if ( thal <= 3.000000 ) then
{
        if ( oldpeak <= 2.100000 ) then
        {
                if ( cp <= 3.000000 ) then
                {
                        num = "0"
                }
                elseif ( cp > 3.000000 ) then
                {
                        if ( ca <= 0.000000 ) then
                        {
                                num = "0"
                        }
                        elseif ( ca > 0.000000 ) then
                        {
                        if ( chol <= 282.000000 ) then
                                {
                                        num = "1"
                                }
                        elseif ( chol > 282.000000 ) then
                                {
                                        num = "0"
                                }
                        }
                }
        }
        elseif ( oldpeak > 2.100000 ) then
        {
                if ( exang <= 0.000000 ) then
                {
                        if ( slope <= 2.000000 ) then
                        {
                                num = "1"
                        }
                        elseif ( slope > 2.000000 ) then
                        {
                                num = "0"
                        }
                }
                elseif ( exang > 0.000000 ) then
                {
                        if ( ca <= 0.000000 ) then
                        {
                                num = "0"
                        }
                        elseif ( ca > 0.000000 ) then
                        {
                                num = "3"
                        }
                }
        }
}
elseif ( thal > 3.000000 ) then
{
        if ( sex <= 0.000000 ) then
        {
                if ( chol <= 295.000000 ) then
                {
                        num = "3"
                }
                elseif ( chol > 295.000000 ) then
                {
                        num = "1"
                }
        }
        elseif ( sex > 0.000000 ) then
        {
```

```
if ( oldpeak <= 2.400000 ) then
{
if ( fbs <= 0.000000 ) then
        {
        if ( thal <= 6.000000 ) then
                {
if ( trestbps <= 135.000000 ) then
                        {
                        num = "0"
                        }
elseif ( trestbps > 135.000000 ) then
                        {
                        num = "2"
                        }
                }
        elseif ( thal > 6.000000 ) then
                {
        if ( chol <= 205.000000 ) then
                        {
        if ( age <= 54.000000 ) then
                        {

num = "3"
                        }
                                elseif
( age > 54.000000 ) then
                        {

num = "2"
                        }
                }
elseif ( chol > 205.000000 ) then
                        {
if ( trestbps <= 114.000000 ) then
                        {

if ( age <= 53.000000 ) then

{

        num = "1"

}

elseif ( age > 53.000000 ) then

{

        num = "2"

}
                                                }
                                                elseif
( trestbps > 114.000000 ) then
                                                {

if ( chol <= 266.000000 ) then
{

        if ( ca <= 0.000000 ) then

        {

                num = "0"

        }

        elseif ( ca > 0.000000 ) then

        {

                if ( age <= 61.000000 ) then

                {

                        num = "1"

                }

                elseif ( age > 61.000000 ) then

                {

                        num = "0"

                }

        }

}

elseif ( chol > 266.000000 ) then

{

        if ( restecg <= 1.000000 ) then

        {

                num = "1"

        }

        elseif ( restecg > 1.000000 ) then

        {

                num = "3"

        }

}
                                                                        }
```

```
                                        }
                                }
                        }
                elseif ( fbs > 0.000000 ) then
                        {
                        if ( exang <= 0.000000 ) then
                                {
                if ( trestbps <= 156.000000 ) then
                                        {
                                        num = "0"
                                        }
                elseif ( trestbps > 156.000000 ) then
                                        {
                                        num = "3"
                                        }
                                }
                elseif ( exang > 0.000000 ) then
                                {
                                        num = "2"
                                }
                        }
                }
                elseif ( oldpeak > 2.400000 ) then
                {
                if ( thalach <= 124.000000 ) then
                        {
                                num = "3"
                        }
                elseif ( thalach > 124.000000 ) then
                        {
                        if ( cp <= 3.000000 ) then
                                {
                                        num = "4"
                                }
                        elseif ( cp > 3.000000 ) then
                                {
                        if ( exang <= 0.000000 ) then
                                        {
                                        num = "4"
                                        }
                elseif ( exang > 0.000000 ) then
                                        {
                                        num = "2"
                                        }
                                }
                        }
                }
        }
}
```

- Total Number of Nodes 27
- Number of Leafs 28
- Total Number of Nodes 27
- Number of Antecedents By Rule 5.928571428571429
- Number of Itemsets Instances 151
- Number of Correctly Classified Instances 132
- Percentage of Correctly Classified Instances 87.41722%
- Number of Incorrectly Classified Instances 19
- Percentage of Incorrectly Classified Instances 12.582782%

B. Rules inferred from decision Tree:

Ruleset 1:

```
if(thal>3.0 && sex<=0.0 && chol<=295.0) (4/4)
        output=3
else if(sex>0.0 && oldpeak>2.4 && thalach<=124.0) (5/5)
        output=3
else if(oldpeak<=2.4 && fbs>0.0 && exang>0.0) (2/3)
        output=2
else if(thal>3.0 && sex>0.0 && oldpeak>2.4 && exang<=0.0) (3/4)
        output=4
else if(thal<=3.0 && oldpeak<=2.1 && cp<=3.0) (51/53)
        output=0
else if(thal<=3.0 && ca<=0.0) (53/59)
        output=0
else if(fbs>0.0 && exang<=0.0 && trestbps<=156.0) (9/9)
        output=0
else if(thal>3.0 && chol>295.0) (4/12)
        output=1
else
        output=0
```

Ruleset 2:

```
        if(cp>3.0 && fbs>0.0 && oldpeak>1.2) (5/5)
        output=2
else if(fbs<=0.0 && ca>0.0 && thalach>131.0 && trestbps>118.0 && exang>0.0) (3/3)
        output=4
else if(fbs<=0.0 && thal<=6.0 && age>58.0 && age<=63.0) (6/8)
        output=1
else if(cp>3.0 && ca>0.0 && sex>0.0 && thalach<=131.0) (6/13)
        output=3
else if(cp<=3.0 && oldpeak<=0.5) (36/43)
        output=0
else if(fbs<=0.0 && sex<=0.0 && slope<=1.0) (24/25)
        output=0
else if(ca<=0.0 && exang<=0.0 && sex<=0.0) (25/26)
        output=0
else if(ca<=0.0 && thal<=6.0) (52/59)
        output=0
else
        output=1
```

V. PERFORMANCE EVALUATION

We have divided the dataset in two parts for performance evaluation

Part 1: Number of instances: 151
Part 2: Number of instances: 152
Total percentage of successes: 0.8675
Percentage of successes in each partition:

1 0.863
2 0.872
Confusion matrix (rows=real class; columns=obtained class):

163	1	0	0	0
13	38	2	2	0
3	4	25	3	1
2	5	1	27	0
1	0	1	2	9

We have used statistical method for analyzing the classifier performance and the Classifier performance results are given below.

Summary of data, Classifiers:
A/cleveland/cleveland
Fold 0 : CORRECT=0.8622516556291391 N/C=0.0
Fold 1 : CORRECT=0.8710526315789473 N/C=0.0
Global Classification Error
+ N/C: 0.13334785639595677
Std.dev Global Classification Error
+ N/C:0.002400487974905742
Correctly classified: 0.8666521436040432
Global N/C: 0.0

REFERENCES

[1]. J.W. Grzymala-Busse, 1991, On the Unknown Attribute Values In Learning From Examples. 6th International Symposium on Methodologies For Intelligent Systems (ISMIS'91). Lecture Notes In Computer Science 542, Springer-Verlag 1991, Charlotte (USA) 368-377.

[2]. J.R. Quinlan. 1993, C4.5: Programs for Machine Learning. Morgan Kauffman Publishers, San Mateo-California.

[3]. J.R. Quinlan. 1995, MDL and Categorical Theories (Continued). In Machine Learning: Proceedings of the Twelfth International Conference. Lake Tahoe, California. Morgan Kaufmann, , 464-470.

[4]. Frawley and Piatetsky-Shapiro, 1996, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[5]. Eibe Frank and Ian H. Witten. 1998 Generating accurate rule sets without global optimization. In Proc 15th International Conference on Machine Learning, Madison, Wisconsin, pages 144-151. Morgan Kaufmann.

[6]. Chen, J., Greiner, R., 1999 "Comparing Bayesian Network Classifiers", In Proc. of UAI-99, pp.101–108.

[7]. Fu-ren Lin, Shien-chao Chou, Shung-mei Pan, Yao-mei Chen, 2000 "Mining time dependency patterns in clinical pathways", Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Vol. 1, pp. 8.

[8]. Wynne Hsu, Mong-Li Lee, Bing Liu, Tok Wang Ling, 2000, "Exploration mining in diabetic patients databases: findings and conclusions", KDD 2000: pp: 430-436.

[9]. Doug Burdick, Manuel Calimlim, Jason Flannick, Johannes Gehrke, Tomi Yiu, 2001"MAFIA: A Performance Study of Mining Maximal Frequent Itemsets", Proceedings of the 17th International Conference on Data Engineering, p.443-452, April 02-06

[10]. Li, W., Han, J., Pei, J., 2001 "CMAR: Accurate and Efficient Classification Based on Multiple Association Rules", In: Proc. of 2001 Interna'l Conference on Data Mining.

[11]. S Stilou, P D Bamidis, N Maglaveras, C Pappas, , 2001, "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare", Stud Health Technol Inform 84: Pt 2. 1399-1403.

[12]. "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", January 2003 Special report: 2001 – 2003, New Mexico.9.

[13]. Margaret R. Kraft, Kevin C. Desouza, Ida Androwich, 2003 "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population", Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03).

[14]. Ian H. Witten and Eibe Frank 2005 "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann,San Francisco.

[15]. Tzung-I Tang, Gang Zheng, Yalou Huang, Guangfu Shu, Pengtao Wang, June 2005, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", IEMS, Vol. 4, No. 1, pp. 102-108.

[16]. Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu , 2006 "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721-727.

[17]. Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, May 2007 "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66.

[18]. T Syeda-Mahmood, F Wang, D Beymer, A Amir, M Richmond, SN Hashmi, 2007 "AALIM: Multimodal Mining for Cardiac Decision Support", Computers in Cardiology, pp. 209-212

[19]. Galip Saracoglu, 2008 "Artificial Neural Network Approach for Prediction of Absorption Measurements of an Evanescent Field Fiber Sensor", Sensors, Vol. 8, pp. 1585-1594.

[20]. Latha Parthiban and R.Subramanian, 2008, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3.

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

[22]. Heart attack dataset from http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[23]. "Heart Disease" from http://chinese-school.netfirms.com/heart-disease-causes.html

[24]. J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, 2009 KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 307-318,

[25]. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten 2009; The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[26]. Shantakumar B. Patil, Y.S. Kumaraswamy, February 2009,"Extraction of Significant Patterns from Heart Disease

Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network. Security, Vol. 9 No. 2 pp. 228-235.

[27]. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. 2011, KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 255-287

[28]. Purushottam Sharma, Dr Kanak Saxena,"Temporal Weighted Association Rule Mining for Classification International Journal of Computer Theory and Engineering ISSN: 1793-8201 Vol. 4, No. 5, October 2012.