Heart Disease Prediction System Evaluation Using C4.5 Rules and Partial Tree

Purushottam Sharma, Kanak Saxena and Richa Sharma

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 and accurately and the correct automation would be very desirable. Every human being cannot be equally skilful 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. Then we evaluate and compare this system using C45 rules and partial tree. The performance of the system is evaluated in terms of different parameter like rules generated, classification accuracy, classification error, global classification error and the experimental results shows that the system has great potential in predicting the heart disease risk level more efficiently.

Keywords C4.5 · Heart disease prediction system · CVD · CAD · PART

P. Sharma (✉)
R.G.T.U., Bhopal, M.P, India
e-mail: puru.mit2002@gmail.com

K. Saxena
Department of Computer Application, S.A.T.I., Vidisha, M.P, India
e-mail: kanak.saxena@gmail.com

P. Sharma · R. Sharma
Amity University, Noida, Uttar Pradesh, India
e-mail: s.richa.sharma@gmail.com

H.S. Behera and D.P. Mohapatra (eds.), *Computational Intelligence in Data Mining—Volume 2*, Advances in Intelligent Systems and Computing 411, DOI 10.1007/978-81-322-2731-1_26

1 Introduction

In today's time at many places clinical test results are often produced based on doctors' intuition, skills and expertize rather than on the rich information available in many large databases. Many a times this process leads to error, unintentional biases and a huge medical cost. Sometimes it can affects the quality of service provided to patients drastically.

Today many hospitals installed some kind of patient's information collection systems to manage their healthcare or to collect patient data. These information systems usually 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.

Here we focus on Heart Disease Prediction using data Mining techniques. The motivation for this study is the estimation given by WHO. As per the WHO estimation by year 2030, almost 23.6 million people will die due to Heart disease. So to minimize the risk, prediction of heart disease should be done. The most difficult and complex task in healthcare sector is diagnosis of correct disease. Heart disease prediction using different parameters of a patient diagnostic tests is a multi-layered issue which may lead to false presumptions and unpredictable effects. Now a day's Healthcare sector generating a huge amounts of raw data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. This huge amount huge of raw data is the main resource that can be efficiently pre-processed and analysed for key information extraction that directly or indirectly motivates the medical society for cost-effectiveness and support decision making. Proper diagnosis of heart disease cannot be possible by using only human intelligence. There are lots of parameters that can affects the accurate diagnosis like less accurate results, less experience, time dependent performance, knowledge up gradation and so on. Lots of development and research happened in this field using multi-parametric attributes with nonlinear and linear features of Heart Rate Variability (HRV). A novel technique was proposed by Lee et al. [1]. To achieve this, many researchers have used many classifiers e.g. CMAR (Classification based on Multiple Association Rules), SVM (Support Vector Machine), Bayesian Classifiers and C4.5). Some of the latest techniques in this field described in [2]. In Healthcare, there is a very large scope and potential of Data mining applications usefulness but effectiveness of these application mostly reliable on accuracy of data and cleanliness. In this regard, it is very much desirable that the healthcare industry use such policies and methods so that data can be better prepared, stored, captured

and mined. Some probable methods and methodology we suggested includes the clinical data standardization, analysis and the data sharing across the related industries to enhance the accuracy and effectiveness of data mining applications in healthcare [3]. It is also advisable to explore the use of text mining and image mining for expansion the nature and scope of data mining applications in healthcare sector. Data mining application can also be explored on digital diagnostic images for application effectiveness. Some progress has been made in these areas [4, 5].

The question can be arises out of this available data:

"How can we use this data to generate useful information that can be used by healthcare practitioners to make effective clinical decisions?" This is the main objective of this research.

2 Background

In recent time, many organizations in healthcare sector uses data mining applications intensively and extensively on large scale. Another reason is that the healthcare transactions generated by this sector are too voluminous and complex to be analysed and processed by traditional methods. Decision-making can be improved majorly by using data mining applications in discovering trends and patterns in large volumes of typical data [6]. In recent trends analysis on these large dataset has become necessary due to financial pressures on healthcare industries. This extracted information can be used for decisions making based on the regress analysis of medical and financial data. Knowledge extraction can influence industry operating efficiency, revenue and cost using knowledge discovery from database by maintaining a top level of care [7]. Research shows that if we uses data mining applications in healthcare organizations then these organizations would be in better position to meet their short term goals and long-term needs, Benko and Wilson argue [8]. We can get very useful results from healthcare raw data by transforming raw data into useful information. A great reason that enables researchers in this field is that this is very useful for all stake holder involved in the healthcare sector. Like, if we consider Insurance provider, they can detect abuse and fraud, practitioner in healthcare can gain assistance in decisions making, like in customer relationship management. Healthcare providers (hospitals, physician, test laboratories and patient etc.) can also use data mining applications in their respective expert zone for expert decision making for example, by finding best practices and correct and effective treatments.

3 UCI 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)

The "num" attributes indicate the presence and level or absence of heart disease in the patient. The range of this attribute is from 0 (no presence) to 4 (severe).

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 our experiments only 14 of them is actually used because all published experiments till now using a subset of 14 only and the data is also given only for these 14 attributes. 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],@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 C4.5 Rules [9–11] and partial tree on the above modified dataset and find out the generated rule sets with different priority. We have also generated pruned and classified rules. Further we have used WEKA tool [12] for dataset analysis and KEEL [13, 14] to find out the classification decision rules and partial tree generation.

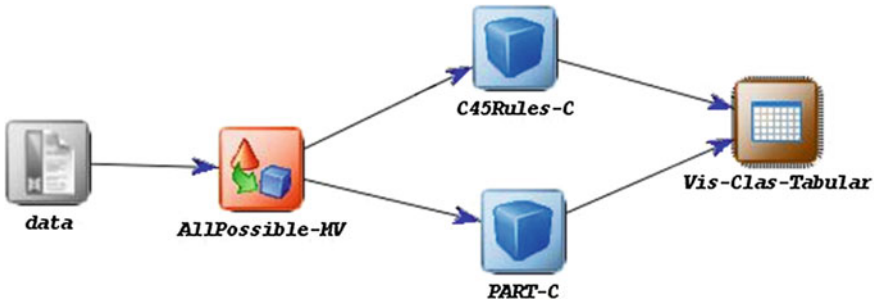

Fig. 1 Heart disease prediction model

4 Experiment Design with KEEL

We have used KEEL (Knowledge Extraction based on Evolutionary Learning) [14] 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 Fig. 1. In the preprocessing phase we have used an AllPossible-MV [15] algorithm to fill the missing values in the dataset.

5 Classification Rule Generation

5.1 Pruned rule set generated by the experiments:

if(ca<=0.0 && thal>6.0 && oldpeak>0.3 && thalach<=129.0 && thalach>122.0) output=4 else if(cp>3.0 && ca>0.0 && slope<=2.0 && ca<=1.0 && age>62.0) output=2 else if(cp>3.0 && thal<=6.0 && restecg<=1.0 && exang>0.0 && age<=56.0) output=2 else if(cp>3.0 && slope<=2.0 && sex>0.0 && ca>1.0 && restecg<=1.0) output=3 else if(ca>0.0 && slope>2.0) output=3 else if(cp>3.0 && thal>6.0 && oldpeak<=0.3) output=1 else if(cp<=3.0 && oldpeak<=1.9 && fbs<=0.0 &&thalach>126.0) output=0 else if(cp<=3.0 && fbs>0.0 && oldpeak<=1.2) output=0 else if(ca<=0.0 && thal<=6.0 && restecg>1.0) output=0 else if(cp<=3.0 && oldpeak<=1.9 && age<=57.0) output=0 else if(cp<=3.0 && ca<=1.0 &&

slope<=1.0) output=0 else if(cp<=3.0 && fbs<=0.0 && thalach>126.0 && sex<=0.0) output=0else if(ca<=0.0 && thal<=6.0 && restecg<=1.0 && exang<=0.0) output=0 else output=1

The rules generated based on partial tree:

(cp<=3.0 && sex<=0.0)-> 0 (cp<=3.0 && sex>0.0 && age<=63.0 && oldpeak<=2.0 && trestbps<=152.0 && fbs<=0.0 && exang<=0.0 && thalach<=142.0)-> 1
(cp<=3.0 && sex>0.0 && age<=63.0 && oldpeak<=2.0 && trestbps<=152.0 && fbs<=0.0 && exang<=0.0 && thalach>142.0 && oldpeak<=0.5)-> 0
(cp<=3.0 && sex>0.0 && age<=63.0 && oldpeak<=2.0 && trestbps<=152.0 && fbs<=0.0 && exang<=0.0 && thalach>142.0 && oldpeak>0.5 && age<=57.0)-> 0
(cp<=3.0 && sex>0.0 && age<=63.0 && oldpeak<=2.0 && trestbps<=152.0 && fbs<=0.0 && exang<=0.0 && thalach>142.0 && oldpeak>0.5 && age>57.0)-> 1
(cp<=3.0 && sex>0.0 && age<=63.0 && oldpeak<=2.0 && trestbps<=152.0 && fbs<=0.0 && exang>0.0)-> 0 (cp<=3.0 && sex>0.0 && age<=63.0 && oldpeak<=2.0 && trestbps<=152.0 && fbs>0.0)-> 0

Partial tree generated by the experiments

```
if ( cp <= 3.000000 ) then
{
      if ( sex <= 0.000000 ) then
      {
            num = "0"
      }
      elseif ( sex > 0.000000 ) then
      {
            if ( age <= 63.000000 ) then
            {
                  if ( oldpeak <= 2.000000 ) then
                  {
             if ( trestbps <= 152.000000 ) then
                        {
                   if ( fbs <= 0.000000 ) then
                              {
                  if ( exang <= 0.000000 ) then
                                    {
                   if ( thalach <= 142.000000 ) then
                                          {
                                         num = "1"
                                          }
       elseif ( thalach > 142.000000 )  then
 {
          if ( oldpeak <= 0.500000 )     then
{
    num = "0" }elseif ( oldpeak > 0.500000 ) then
{ if ( age <= 57.000000 )  then

{ num = "0"

 }
elseif ( age > 57.000000 )     then

{  num = "1"

  }  }      }     }
  elseif ( exang > 0.000000 ) then
                                    {
                                   num = "0"
                                    }
}
            elseif ( fbs > 0.000000 ) then
                                    {
                                      num = "0"
                }              }              }        }
}
```

6 Evaluation Results

We have used 5 folds for training and 5 folds for testing to evaluate the classification accuracy using different parameter.

6.1 Classification Results by Algorithm and by Fold

We have evaluated the classification accuracy of C4.5 Rules and Partial Tree classifier and the results using different classifier fold wise are as follows

Test Results using Partial Tree Classifier
Fold 0 CORRECT=0.540983606557377 N/C=0.0
Fold 1 CORRECT=0.5454545454545454 N/C=0.0
Fold 2 CORRECT=0.540983606557377 N/C=0.0
Fold 3 CORRECT=0.5238095238095238 N/C=0.0
Fold 4 CORRECT=0.5882352941176471 N/C=0.0

Global Classification Error + N/C: 0.45210668470070586
Stddev Global Classification Error + N/C: 0.02148925320086861
Correctly classified: 0.5478933152992942, Global N/C: 0.0

Train Results using Partial Tree Classifier
Fold 0 CORRECT=0.5503875968992248 N/C=0.0
Fold 1 CORRECT=0.5494071146245059 N/C=0.0
Fold 2 CORRECT=0.5503875968992248 N/C=0.0
Fold 3 CORRECT=0.5546875 N/C=0.0
Fold 4 CORRECT=0.5378486055776892 N/C=0.0

Global Classification Error + N/C: 0.451456317199871
Stddev Global Classification Error + N/C: 0.0056511365140908335
Correctly classified: 0.548543682800129, Global N/C: 0.0

Test Results using C4.5 Rules Classifier
Fold 0 CORRECT=0.5081967213114754 N/C=0.0
Fold 1 CORRECT=0.5 N/C=0.0
Fold 2 CORRECT=0.6065573770491803 N/C=0.0
Fold 3 CORRECT=0.47619047619047616 N/C=0.0
Fold 4 CORRECT=0.4558823529411765 N/C=0.0

Global Classification Error + N/C: 0.4906346145015383,
Stddev Global Classification Error + N/C: 0.05195453555220856
Correctly classified: 0.5093653854984617, Global N/C: 0.0

Train Results using C4.5 Rules Classifier
Fold 0 CORRECT=0.7093023255813953 N/C=0.0
Fold 1 CORRECT=0.6482213438735178 N/C=0.0
Fold 2 CORRECT=0.6550387596899225 N/C=0.0
Fold 3 CORRECT=0.62890625 N/C=0.0
Fold 4 CORRECT=0.6772908366533865 N/C=0.0

Global Classification Error + N/C: 0.33624809684035556,
stddev Global Classification Error + N/C: 0.02752991241516823
Correctly classified: 0.6637519031596444, Global N/C: 0.0

6.2 Global Average and Variance

The global average and variance measured using C4.5 Rules classifier and Partial Tree classifier are given in Table 1.

Table 1 Global average and variance

C4.5 rules				Partial tree		
	Average correctly classified	Variance correctly classified	Not classified	Average correctly classified	Variance correctly classified	Not classified
Test	0.509365	0.002699	0.00	0.547893	0.000461	0.00
Train	0.663751	0.000757	0.00	0.548543	0.000031	0.00

Table 2 Classification rate by algorithm and by fold

		C4.5 rules		Partial tree	
		Correctly classified	Not classified	Correctly classified	Not classified
Test	Fold 0	0.5081967213	0.00000	0.5409836066	0.000000
	Fold 1	0.5000000000	0.00000	0.5454545455	0.000000
	Fold 2	0.6065573770	0.00000	0.5409836066	0.000000
	Fold 3	0.4761904762	0.00000	0.5238095238	0.000000
	Fold 4	0.4558823529	0.00000	0.5882352941	0.000000
Train	Fold 0	0.7093023256	0.00000	0.5503875969	0.000000
	Fold 1	0.6482213439	0.00000	0.5494071146	0.000000
	Fold 2	0.6550387597	0.00000	0.5503875969	0.000000
	Fold 3	0.6289062500	0.00000	0.5546875000	0.000000
	Fold 4	0.6772908367	0.00000	0.5378486056	0.000000

6.3 Classification Rate by Algorithm and by Fold

To evaluate the performance of C4.5 Rules classifier and Partial Tree classifier fold wise on test data set and training data set are given in the Table 2.

7 Conclusions

Heart Disease Prediction System evaluation analysis shows the evaluation of the two classifier on different parameter with different statistics measures. Results shows that C4.5 classifier can correctly classified the heart Disease up to 70.93 %. It has been also observed that C4.5 classifier supersedes the partial classifier on the given dataset.

References

1. Lee, H.G., Noh, K.Y., Ryu, K.H.: Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV. LNAI 4819
2. Chhikara, S., Sharma, P.: Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases. I JRASET, pp. 396–402. (2014)
3. Cody, W.F., Kreulen, J.T., Krishna, V., Spangler, W.S.: The integration of business intelligence and knowledge management. IBM Syst. J. **41**(4), 697–713 (2002)
4. Ceusters, W.: Medical natural language understanding as a supporting technology for data mining in healthcare. In: Cios, K.J. (ed.) Medical Data Mining and Knowledge Discovery, pp. 41–69. PhysicaVerlag Heidelberg, New York (2001)
5. Megalooikonomou, V., Herskovits, E.H.: Mining structure function associations in a brain image database. In: Cios, K.J. (ed.) Medical Data Mining and Knowledge Discovery, pp. 153–180. Physica-Verlag Heidelberg, New York (2001)
6. Biafore, S.: Predictive solutions bring more power to decision makers. Health Manag. Technol. **20**(10), 12–14 (1999)
7. Silver, M., Sakata, T., Su, H.C., Herman, C., Dolins, S.B., O'Shea, M.J.: Case study: how to apply data mining techniques in a healthcare data warehouse. J. Healthc. Inf. Manag. **15**(2), 155–164 (2001)
8. Benko, A., Wilson, B.: Online decision support gives plans an edge. Managed Healthc. Executive **13**(5), 20 (2003)
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman Publishers, San Mateo-California (1993)
10. Quinlan, J.R.: MDL and categorical theories (continued). In: Machine Learning: Proceedings of the Twelfth International Conference, pp. 464–470. Lake Tahoe, California. Morgan Kaufmann (1995)
11. Tang, T.-I., Zheng, G., Huang, Y., Shu, G., Wang, P.: A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS **4**(1), 102–108 (2005)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 2009

13. Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F.: KEEL: a software tool to assess evolutionary algorithms to data mining problems. Soft Comput. 307–318 (2009)
14. Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Logic Soft Comput. **17**(2–3), 255–287 (2011)
15. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: 6th International Symposium on Methodologies for Intelligent Systems (ISMIS'91). Lecture Notes in Computer Science, vol. 542, pp. 368–377. Springer, Charlotte (USA) (1991)