Paper from Proceeding of the National Conference on Innovative Developments in Science, Technology & Management (NCIDSTM-2015) Organized by Ganga Technical Campus, Soldha, Bahadurgarh, Haryana (India) March 1st 2015 Published by International Journal of Engineering Sciences Paradigms and Researches (IJESPR) with ISSN (Online): 2319-6564, Impact Factor: 2.20 and Website: www.ijesonline.com

Diagnosis of Cardiovascular Abnormalities: A Data Mining Based Approach

Harish Ruhil
Lecturer, Department of Computer Science
Ganga Technical Campus, Bahadurgarh-India
harish.ruhil@gmail.com

Abstract

Now a day data mining is used in many areas directly or indirectly associated with our lives. Health care is also one of the domain which gets a lot of benefits and researches with the advent and progress in data mining. It plays a vital role in extracting useful knowledge and making scientific decision for diagnosis and treatment of disease. Data mining in medicine can resolve this problem and can provide promising results. Treatment records of millions of patients have been stored and various tools and techniques are applied to analyze the data. A classification approach in health care could be a method of diagnosing to determine if a patient has certain disease or not. In this paper, we make use of a large heart disease data set obtained from UCI machine repository applied J48 and Random Forest classifier after prepossessing the data sets and the result shows that J48 outperforms and gives promising results as compared to Random Forest.

Keywords*: Data Mining, Health care, Heart Disease*

1. Introduction

Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset. A number of algorithms have been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support [1].

The growing espousal of information technologies in medical field and the availability of vast data concerned with patients including investigation reports, clinical examination, treatment follow-ups vital parameters, and drug decisions etc. provide new opportunities for using analytics to control health outcomes. The Data mining is one of the considerable strategies that can be can be extremely useful for Medical practitioners for extracting hidden medical knowledge to predict the Disease possibility from the health record with approved feature of an individual. There are different types of diseases predicted using data mining techniques such as Heart disease, Hepatitis, Lung Cancer, Thyroid disease, Diabetes, Breast cancer, Liver disorder[2].

Data mining applications in healthcare include analysis of health care centers for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims [3].

Heart disease is the leading cause of death in the world over the past 10 years. The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths. The Economical and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular diseases, cancers, diabetes and chronic respiratory diseases. The Australian Bureau of Statistics reported that heart and circulatory system diseases are the first leading cause of death in Australia, causing 33.7% of all deaths. Statistics of South Africa reported that heart and circulatory system diseases are the third leading cause of death in Africa [4].

2. Data Mining Techniques

Data mining is a process of identifying and extracting hidden patterns and information from databases and data warehouses. There are various well known data mining techniques are available.

2.1. Classification

Classification is a supervised learning technique and it is most widely which employs a set of pre-classified examples to develop a model that can classify the population of records at large. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to classify the instances without class. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination [6].

2.2. Clustering

Clustering is unsupervised Unsupervised learning technique in which class labels are known. It is a descriptive technique which consists of identifying classes or groups in sets of unclassified data. It identifies groups of related records that can be used as a starting point for exploring further relationships [7].

2.3. Prediction

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict [8].

2.4. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little value [6].

3. Experimental Work

The experiment is based on using tool Weka and heart disease data set having 303 instances and 14 attributes available at UCI Machine Repository. The attributes are age, gender, cp, trestbps, fbs, restect, thalach, exang,oldpeak, slope, ca, thal, num.

Measures for performance evaluation

To measure the performance of a medical test, the concepts sensitivity and specificity are often used; these concepts are readily usable for the evaluation of any binary classifier. Say we test some people for the presence of a disease. Some of these people have the disease, and our test says they are positive. They are called true positives (TP). Some have the disease, but the test claims they don't. They are called false negatives (FN). Some don't have the disease, and the test says they don't - true negatives (TN). Finally, we might have healthy people who have a positive test result false positives (FP). Thus, the number of true positives, false negatives, true negatives, and false positives add up to 100% of the set [10].

True Positive Rate

It is the proportion of people that tested positive of all the positive people tested; that is (true positives) / (true positives + false negatives). It can be seen as the probability that the test is positive given that the patient is sick. The higher the sensitivity, the fewer real cases of diseases go undetected. It can be defined as:

$$sensitivity = \frac{number\_of\_true\_positives(TP)}{number\_of\_true\_positives(TP) + number\_of\_false\_negatives(FN)}$$

True Negative Rate

It is the proportion of people that tested negative of all the negative people tested; that is (true negatives) / (true negatives + false positives). As with sensitivity, it can be looked at as the probability that the test is negative given that the patient is not sick. The higher the specificity, the fewer healthy people are labeled as sick. It can be defined as:

$$specificity = \frac{number\_of\_true\_negative(TN)}{number\_of\_true\_negative(TN) + number\_of\_false\_positive(FP)}$$

Accuracy

It is simply a ratio of ((no. of correctly classified examples) / (total no. of examples)) *100). Technically it can be defined as:

$$accuracy = \frac{TP + TN}{(TP + FN) + (FP + TN)}$$

Confusion Matrix

		actual value		
		p	n	total
prediction outcome	p'	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

Decision Tree

Decision tree as a predictive model can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification [9].

Confusion Matrix for J48

	NO	Yes
NO	190	29
Yes	30	54

Confusion Matrix: Random Forest

	No	Yes
No	198	21
Yes	40	44

Decision Tree J48

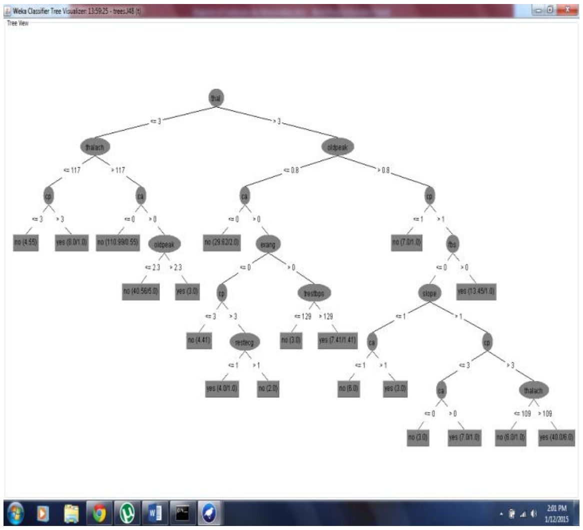
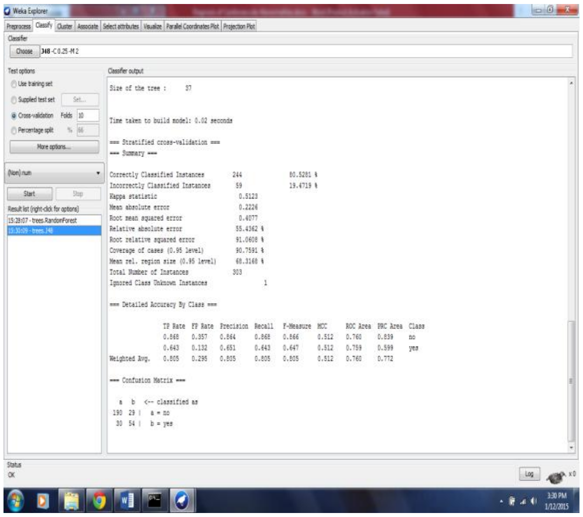


Fig. 1: J 48 Decision Tree

Results with parameters



J48 -C 0.25 -M 2

◉ Cross-validation Folds 10

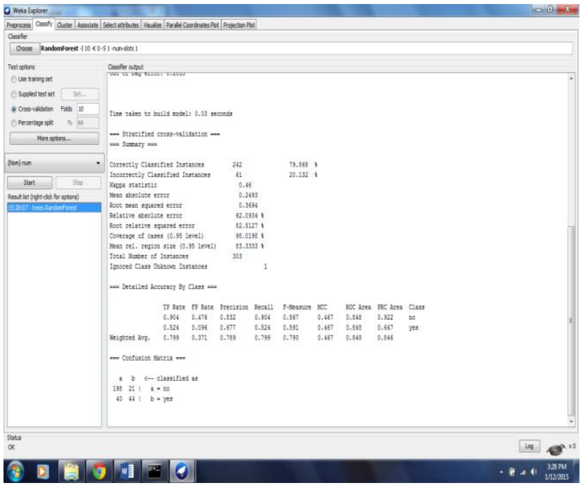
```
Size of the tree :      37


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         244               80.5281 %
Incorrectly Classified Instances        59               19.4719 %
Kappa statistic                          0.5123
Mean absolute error                      0.2226
Root mean squared error                  0.4077
Relative absolute error                 55.4362 %
Root relative squared error             91.0608 %
Coverage of cases (0.95 level)          90.7591 %
Mean rel. region size (0.95 level)      68.3168 %
Total Number of Instances              303
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.868    0.357    0.864      0.868    0.866      0.512    0.760     0.839     no
                 0.643    0.132    0.651      0.643    0.647      0.512    0.759     0.599     yes
Weighted Avg.    0.805    0.295    0.805      0.805    0.805      0.512    0.760     0.772

=== Confusion Matrix ===

   a   b   <-- classified as
 190  29 |   a = no
  30  54 |   b = yes
```

Fig. 2: J48 Result

J 48 classifier builds model in 0.02 seconds with 80.5% correctly classified instances and 19.4 % incorrectly classified instances.

1. Random Forest Classifier

Random Forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk) k=1, 2, ….} where the {Θk } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x. Random Forest generates an ensemble of decision trees. To achieve diversity among base decision trees, Breiman selected the randomization approach which works well with bagging or random subspace methods [5].



Classifier

RandomForest -I 10 -K 0 -S 1 -num-slots 1

Test options

○ Use training set

○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) num

Result list (right-click for options)

15:28:07 - trees.RandomForest

Classifier output

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         242               79.868 %
Incorrectly Classified Instances        61               20.132 %
Kappa statistic                          0.46
Mean absolute error                      0.2493
Root mean squared error                  0.3694
Relative absolute error                 62.0934 %
Root relative squared error             82.5127 %
Coverage of cases (0.95 level)          98.0198 %
Mean rel. region size (0.95 level)      83.3333 %
Total Number of Instances              303
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.904    0.476    0.832      0.904    0.867      0.467    0.848     0.922     no
                 0.524    0.096    0.677      0.524    0.591      0.467    0.848     0.647     yes
Weighted Avg.    0.799    0.371    0.789      0.799    0.790      0.467    0.848     0.846

=== Confusion Matrix ===

   a   b   <-- classified as
 198  21 |   a = no
  40  44 |   b = yes
```

Status

OK

Fig 3: Random forest result

Random classifier builds the model is 0.07 seconds with 79.868 % correctly classified instances.

4. Conclusion

In this paper we compared the Decision tree classifiers on Heart Disease data set. The result show that J 48 classifier give more accurate result that is 85.14 % as compared to Random Forest.

References

[1] Desouza, K.C. (2001) Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands: Institute for Healthcare Technology Management.

[2] Ramana Nagavelli et al, "Degree of Disease Possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining", IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, 2014, Jaipur, India

[3] Ruben, D.C.J., Data Mining in Healthcare: Current Applications and Issues. 2009.

[4] Mai Shouman et al, "Using data mining techniques in heart Disease diagnosis and treatment", IEEE Conference 2012 Japan-Egypt Conference on Electronics, Communications and Computers

[5] Vrushali Y Kulkarni et al, "Random Forest Classifiers :A Survey and Future Research Directions", International Journal of Advanced Computing, ISSN:2051-0845, Vol.36, Issue.1 April 2013

[6] Bharati M. Ramageri ,"DATA MINING TECHNIQUES AND APPLICATIONS" Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305 (ISSN : 0976-5166)

[7] A. Kusiak, K.H. Kernstine, J.A. Kern, K.A. McLaughlin and T.L. Tseng, Data Mining: Medical and Engineering Case Studies, Proceedings of the Industrial Engineering Research Conference (2000), Cleveland, Ohio, May 21-23, pp. 1-7

[8] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[9] Nikita Jain et al. "DATA MINING TECHNIQUES: A SURVEY PAPER ", International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11, Nov-2013.

[10] Varun Kumar et al, "Binary Classifiers for Health Care Databases: A Comparative Study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer", IJCST Vol. 1, Iss ue 2, December 2010.