IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS)
e-ISSN: 2319-2380, p-ISSN: 2319-2372.Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64
www.iosrjournals.org

Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques

Shamsher Bahadur Patel [1], Pramod Kumar Yadav[2], Dr. D. P.Shukla[3]
[1]*Research Scholar, Department of Computer Science & Mathematics, Govt. P. G. Science College Rewa (M.P.) ,India,*
[2]*Research Scholar, Department of BCA & Physics, Govt. P. G. Science College Rewa (M.P.), India*
[3]*Dr. D. P. Shukla, Professor & Head, Department of Computer Science & Mathematics, Govt. P. G. Science College Rewa (M.P.), India*

Abstract: *The data mining can be referred as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays an important role in various applications such as business organizations, e-commerce, health care industry, scientific and engineering. In the health care industry, the data mining is mainly used for Disease Prediction.the objective our works to predict the diagnosis of heart disease with reduced number of attributes. Here fourteen attributes involved in predicting heart disease. But fourteen attributes are reduced to six attributes by using Genetic algorithm. Subsequently three classifiers like Naive Bayes, Classification by Clustering and Decision Tree are used to predict the diagnosis of heart disease after the reduction of number of attributes.*
Keywords: *Data Mining, Naive Bayes, Decision Tree, Classification by Clustering.*

I. Introduction

Heart disease is a general name for a variety of diseases, conditions and disorders that affect the heart and the blood vessels. Symptoms of heart disease vary depending on the specific type of heart disease. Congenital heart disease refers to a problem with the heart's structure and function due to abnormal heart development before birth. Congestive heart failure is when the heart does not pump adequate blood to the other organs in the body. Coronary heart disease or in its medical term Ischemic heart disease is the most frequent type of heart problem. Coronary heart disease is a term that refers to damage to the heart that happens because its blood supply is decreased, it leads to the fatty deposits build up on the linings of the blood vessels that provide the heart muscles with blood, resulting in them narrowing The paper identifies the risk factors for the different types of heart diseases.

Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data [1].These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" This is the main motivation for this research.

World Health Organization in the year 2003 reported that 29.2% of total global deaths are due to Cardio Vascular Disease (CVD). By the end of this year, CVD is expected to be the leading cause for deaths in developing countries due to change in life style, work culture and food habits. Hence, more careful and efficient methods of cardiac diseases and periodic examination are of high importance.

II. Related Work

Numerous works [3, 2] related to heart disease diagnosis using data mining techniques have motivated this study. The dataset, algorithms, methods used by the authors and the observed results along with the future work are studied for each paper. Large number of work is carried out in finding out efficient methods of medical diagnosis for various diseases. Our work is an attempt to predict efficiently diagnosis with reduced number of factors (i.e. attributes) that contribute more towards the cardiac disease using classification. Sellapan et al (2008), Asha et al(2010) developed an Intelligent Heart Disease Prediction System to predict the heart disease using three classifiers Decision Tree, Naïve Bayes and Neural Networks. Naïve Bayes performed with good prediction probability of 96.6%. Also, 13 attributes were used for prediction. Our works differs by reducing the number of attributes to 6 and were able to achieve the same performance. Carlos (2006) implemented efficient search for diagnosis of heart disease comparing association rules with decision trees. Our approach would be an another search for efficient diagnosis. The rest of the sections are organized in the following manner. In section 3 explains the material and methods, in section 4 explain the classifier evaluation measures, in section 5 explain the experiments and results and section 6 include the conclusion.

III. Material And Methods

A. Problem Statement:

Heart disease prediction using data mining is one of the most interesting and challenging tasks. The shortage of specialists and high wrongly diagnosed cases has necessitated the need to develop a fast and efficient detection system. The main objective of this work is to identify the key patterns or features from the medical data using the classifier model. The attributes that are more relevant to heart disease diagnosis can be observed. This will help the medical practitioners to understand the root causes of disease in depth.

B. Data Set

We have taken 14 attributes from medical Data [4]. These fourteen attributes are listed in figure 1. For simplicity, categorical attributes were used for all models. The number of attributes is reduced to six using Genetic Search are listed in figure 2. The reduced data set is fed to the three classification models.

S.N.	Attribute Name	Description
1	Age	Age in years
2	Sex	Male=1,Female=0
3	Cp	Chest pain type
4	Rbp	Resting Blood pressure upon hospital admission
5	Cholesterol	Serum Cholesterol in mg/dl
6	Fasting blood sugar	Fasting blood sugar>120 mg/dl true=1 and false=0
7	Resting ECG	Resting electrocardiographic Results
8	Thalach	Maximum Heart Rate
9	Induced Angina	Does the patient experience angina as a result of exercise (value 1: yes, value 0: no)
10	Old peak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Thal	Value 3:Normal ,value 6:fixed defect, value 7: reversible defect
13	CA	Number of major vessels colored by fluoroscopy(value 0-3)
14	Concept class	Angiographic disease status

Fig. 1 attribute of Heart Disease Date Sets

S.N.	Attribute Name	Description
1	Rbp	Resting Blood pressure
2	Oldpk	Old peak
3	Type	Chest pain type
4	Vsl	Number of major vessels colored
5	Eia	Exercise induced angina
6	Thal	Maximum heart rate achieved

Fig. 2 Reduced attributes list

C. Naïve Bayesion Classifier:

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naïve Bayes performs consistently after reduction of number of attributes.

According to Bayesian theorem

$$P(A|B) = P(A)*P(B/A)/P(B), \text{Where } P(B|A)=P(A\cap B)/P(A)$$

Based on above formula, Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression it has the excellent interpretability same as decision tree and is able to use previous data to build analysis model for future prediction or classification

D. Decision Trees:

- Decision trees are powerful and popular tools for classification and prediction.
- Decision trees represent rules, which can be understood by humans and used in knowledge system such as database.

Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Our work uses J48 decision tree for classification. Observations show that Decision trees outperform the other two classifiers but take more time to build
the model.

E. Classification via clustering:

Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized
before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Observations show that Classification via clustering performs poor even after reduction of number of attributes when compared to the other two methods.

IV. Classifier Evaluation Measures

The four terms used in computing evaluation measures are used for evaluating the model and are described here. The True positives (T_Pos) refer to the positive tuples that are correctly labeled by the classifier, while True negatives (T_Neg) are the negative tuples that are correctly labeled by the classifier. False positives (F_Pos) are the negative tuples that are incorrectly labeled by the classifier, False negatives (F_Neg) are the positive tuples that were incorrectly labeled by the classifier. The confusion matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. Sensitivity is referred to as the true positive rate that is the proportion of positive tuples that are correctly identified.

$$\text{Sensitivity} = = \frac{\text{T\_pos}}{\text{pos}}$$

Specificity is the true negative rate that is the proportion of negative tuples that are correctly identified.

$$\text{Specificity} == \frac{\text{T\_neg}}{\text{neg}}$$

classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. It is a function of specificity and sensitivity.

$$\text{Accuracy} = = \frac{\text{T\_pos} + \text{T\_neg}}{\text{pos} + \text{neg}}$$

V. Experiments And Results

Experiments were conducted with Weka tool. All attributesare made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated., the genetic search resulted in 6 attributes which contributes more towards the diagnosis of the cardiac disease. The three classifiers such as Decision tree, Classification via clustering and Naïve Bayes were used for diagnosis of patients with heart disease. The classifiers were fed with reduced data set with 6 attributes. Results are shown in Table 1. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering performs poor compared to other two methods.

Data Mining Techniques	Accuracy	Model Construction Time	Mean Absolute Error
Decision Tree	99.2%	0.09 s	0.00016
Naïve Bayes	96.5%	0.02 s	0.044
Classification clustering	88.3%	0.06 s	0.117

Table 1. Comparative Table of the three classifiers

VI. Conclusion

In this paper three Classification function Techniques in Data mining are compared for predicting Heart Disease with reduced number of attributes .They are Naïve Bayes, Decision Tree and Classification by Clustering. In our work, Genetic algorithm is used to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. Fourteen attributes are reduced to 6 attributes using genetic search. Also, the observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection with relatively high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering performs poor compared to other two methods. Inconsistencies and missing values were resolved before model construction but in real time, that is not the case. Also, the intensity of the disease based on the results was unpredictable. We intend to extend our work applying fuzzy learning models to evaluate the intensity of cardiac disease.

References

[1] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[2] R. B. Rao, S. Krishan, and R. S. Niculescu(2006), "Data mining for improved cardiac care," ACM SIGKDD Explorations Newsletter., vol. 8, no. 1, pp. 3–10.

[3] Mai Shouman, Tim Turner, Rob Stocker,(2012),"Using Data Mining Techniques In Heart Disease Diagnosis And Treatment ",Proceedings in Japan-Egypt Conference on Electronics, Communications and Computers,IEEE,Vol.2 pp.174-177.

[4] UCI Machine learning Repository from http://archive.ics.uci.edu/ml/datasets/Heart+Di sease

[5] Boleslaw Szymanski, et al. (2006): Using Efficient Supanova Kernel For Heart Disease Diagnosis, proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16, pp. 305-310.

[6] Carlos Ordonez (2006): Comparing Association Rules and Decision Trees for Disease Prediction, ACM, HIKM'06, Arlington, Virginia, USA.

[7] Franck Le Duff, et al. (2004): Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method, Studies in health technology and informatics ,107 pp. 1256-1259.

[8] Harleen Kaur and Siri Krishan Wasan (2006): Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, ISSN pp.1549-3636.

[9] Krishnapuram B, et al. (2004): A Bayesian approach to joint featureselection and classifier design.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 6(9): pp. 1105-1111.

[10] Niti Guru, et al. (2007), Decision Support System for Heart Disease Diagnosis Using Neural Network, Delhi Business Review , Vol. 8,No. 1.

[11] Ordonez C, et al. (2001): Mining constrained association rules to predict heart disease. In IEEE ICDM Conference, pp. 433–440.

[12] Sellappan Palaniappan and Rafiah Awang (2008): Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244- 968- 5/08/ IEEE.

[13] Shantakumar B.Patil and Y.S.Kumaraswamy (2009): Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450- 216X Vol.31 No.4, pp. 642-656.

[14] Anchana Khemphila and Veera Boonjing (2011), "Heart Disease Classification Using Neural Network And Feature Selection", in Proc. 21st International Conference on Systems Engineering,IEEE,vol.3 pp. 406-409.

[15] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi,and Constantinos S. Pattichis(2010),"Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions On Information Technology In Biomedicine, Vol. 14, No. 3.pp.559-566.

[16] K.Srinivas , B.Kavita Rani, Dr. A.Govardhan (2010), Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks ,IJCSE Vol. 02, No. 02, pp 250-255.

[17] M. Karaolis, J. A. Moutiris, L. Papaconstantinou, and C. S. Pattichis(2009), "Association rule analysis for the assessment of the risk of coronary heart events," in Proc. 31st Annu. Int. IEEE Eng. Med. Biol. Soc. Conf., Minneapolis, MN, Sep. 2–6, pp. 6238–6241.