An Effective Prediction Factors for Coronary Heart Disease using Data Mining based Classification Technique

A. K. Shrivas
Dept. of IT, Dr. C. V. Raman University,
Bilaspur, India
e-mail:akhilesh.mca29@gmail.com

Rajat Kumar Yadu
Dept. of IT , Dr. C. V. Raman University,
Bilaspur, India
e-mail:rajyadu28@gmail.com

***Abstract*—** Identification of diseases are very challenging task in field of medical science. Heart disease is very critical issues facing by the people. In our proposed work we have used data mining based classification techniques for analysis and classification of different level of heart disease namely Cleveland, Switzerland, Hungarian and Long Beach. We have used WEKA and Rapid miner data mining tools for analysis of heart disease data set and compared the performance of different classification techniques with four heart disease data set using WEKA and Rapid Miner data mining tool. The proposed SVM gives better accuracy as 66.67% with Hungarian data set in case of WEKA data mining tool while Decision Stump gives better accuracy as 63.94% with same Hungarian data set in case of Rapid miner data mining tool. The Hungarian data set gives better performance with our proposed data mining tools and classification techniques which can help the people to predict effective factors about Coronary Heart Disease.

Keywords- *Coronary Heart Disease, Classification, Data Mining.*

*****

I. INTRODUCTION

Now Days, Coronary artery disease (CAD) is the most common type of heart disease. It is the leading cause of death now days approximately 90% of individuals with coronary heart disease (CHD) have at least one antecedent, traditional factor such as Cholesterol, fasting blood sugar, chest pain. The term heart disease [1] is related to all the diverse diseases affecting the heart. The healthcare industry generates huge amounts of data that are too difficult to be analyzed by traditional methods. Data Mining Software application [2] includes various methodologies that have been developed by both medical and heart disease research centre. Heart disease is a major cause of morbidity and mortality in the modern society. Medical heart disease diagnosis is extremely important but complicated task that should be performed accurately and efficiently. These techniques have been used for Healthcare coronary heart disease (CHD) and heart attack. These risk factors also increase the chance that existing CHD. So we have to need to develop prediction factor system .There are different authors done in the field of heard prediction, which are as follows –

A. Dhanasekar et al. [1] have developed model for heart disease prediction using stream associative classification and Association rules and compared to predictive rules mined with decision trees. M. A. Jabbar et al. [2] proposed an efficient associative classification algorithm using genetic approach for heart disease prediction and develop a decision support System for predicting heart disease of a patient. A. Jarad et al. [11] have used Naïve Bayes , Decision List and K-NN for classification of heart disease and compare the accuracy of models. The proposed Naïve Bayes gives 52.33% of accuracy as best classifier. J. Patel et al. [12] have compared the performance of different decision tree classifier like C4.5, LMT and Random Forest. The proposed J48 gives better accuracy as 56.76% for classifying heart disease. N. Kishore et al.[13] have suggested Genetic algorithm and artificial neural network(ANN) for cardiac analysis and classification of ECG Signal. They have compared the performance in terms of accuracy, FAR and FRR. The proposed algorithm outperforms others. S. Raghavendra, et al. [14] have suggested Logistic Regression (LR) and Artificial Neural Network (ANN) with forward and backward feature selection technique for classification and prediction of medical datasets . M. Singla et al. [15] have used various clustering techniques like k-mean, EM and the farthest first algorithm for the prediction of heart disease. The proposed farthest first algorithm is better among others.

II. PROPOSED METHODOLOGY

The proposed research would be deals with many data mining based classification techniques to predict factor of coronary heart disease. Classification plays important role that is used to classify the data based on its features. Figure 1 shows that proposed work of our research work. In this research work we have explored the four categories of heart disease data sets, Data mining tools, various classification techniques and performance measure.

ISSN: 2321-8169
813 – 816

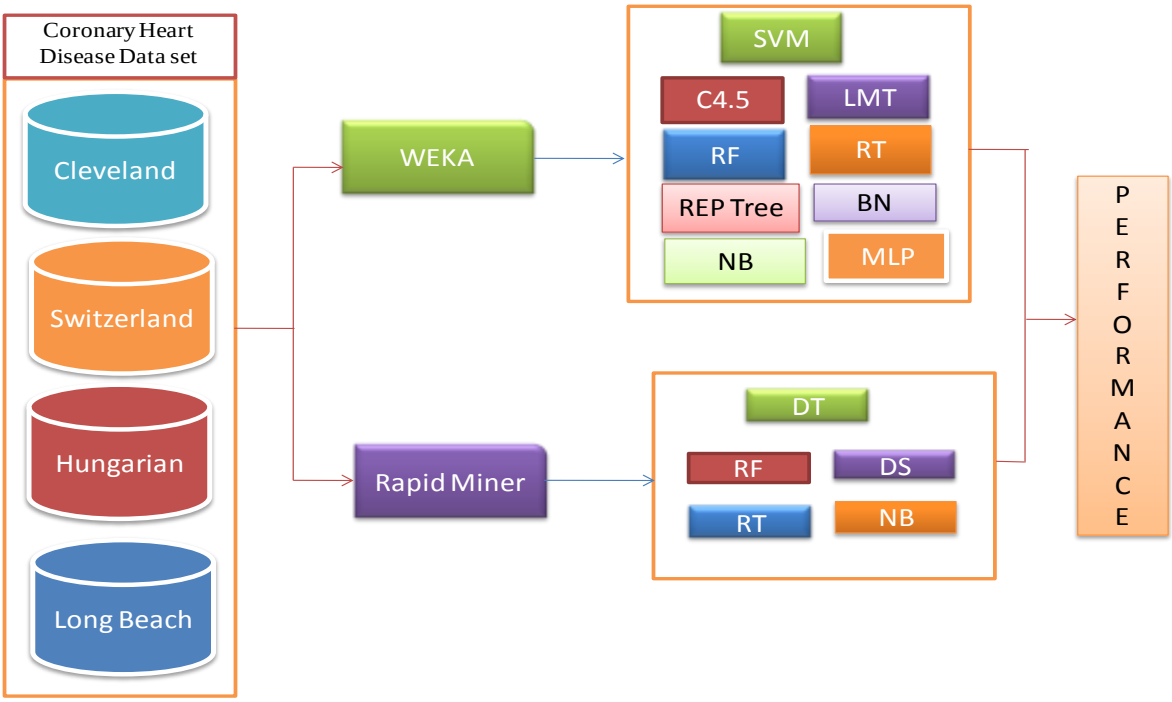

Figure 1. Architecture of proposed model

Heart Disease Dataset

We have used four different heart disease data sets collected from UCI repository [5]. The data sets are heart-disease namely Cleveland, Switzerland, Hungarian and Long Beach. The number of features, number of instances and number of class are described in Table I. Table II shows that class distribution of different heart disease data set.

Table I. Different heart diseases data set and their description

Name of Dataset	No. of Features	No. of Instances	No. of Classes
Cleveland	**13**	**303**	**5**
Switzerland	**13**	**123**	**5**
Hungarian	**13**	**294**	**5**
Long Beach	**13**	**200**	**5**

Table II. Class distribution of heart disease data set

Database	Class Label				
	0	**1**	**2**	**3**	**4**
Cleveland	164	55	36	35	13
Switzerland	8	48	32	30	5
Hungarian	188	37	26	28	15
Long Beach	51	56	41	42	10

A. Data Mining Tool

Tools and techniques are very important role to analysis of data and find out the results. In this research work we have used WEKA [3] and Rapid miner data mining [9] tools for classification of different level of heart disease data. Both the software is open source data mining tools which is free to use. It contains modules for data preprocessing, classification, clustering, association rule, extraction and visualization. WEKA and Rapid Miner tools are used for research, education, and applications.

B. Classification Technique

Classification is data mining based technique used to analysis and classification of data. There rae various classification techniques are used in this research work for classifying the data with different level of heart disease.

- *Decision Tree*

Decision tree (Tang, Z. et al., 2005) [7] is one of the important data mining and classification technique. The principle idea of a decision tree is to split our data recursively into subsets so that each subset contains more or less homogeneous states of our target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute. When this recursive process is completed, a decision tree is formed. In this research work, we have used C4.5, Random Forest [6], Random Tree, Decision Stump and REP Tree [10].

- *Naive Bayes*

Naïve Bayes (Han, J. et al., 2006) [4] is a statistical classifier. This classifier is a simple Bayesian classifier known as the Naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

- *Multilayer Perceptron (MLP)*

MLP (Pujari, A. K., 2001) [6] is a development from the simple perceptron in which extra hidden layers (layers additional to the input and output layers, not connected externally) are added. More than on hidden layer can be used. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes of a trial-and-error process, determined by the nature of the problem at hand. The transfer function generally a sigmoid function.

- *Support Vector Machine (SVM)*

Support vector machines (SVMs) (Olson, D. L. et al., 2008) [8] are supervised learning methods that generate input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. SVMs belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. They are also said to belong to "kernel methods". In this research work we have used polynomial kernel function.

C. *Performance Measure*

Performance of classifiers can be evaluated using well-known statistical measures like accuracy. These measures are defined by true positive (TP), true negative (TN, false positive (FP) and false negative (FN). Classification accuracy measures the proportion of correct predictions considering the positive and negative inputs. It is calculated as follows:

Classification accuracy = (TP+TN)/N.
Where N is total number of samples.

III. RESULTS AND DISCUSSION

This experiment is carried out using two data mining tools with different data mining based classification techniques for analysis and classification of different level of heart disease. We have used WEKA and Rapid miner open source data mining tools for analysis of four different categories of coronary heart disease data set like Cleveland, Switzerland, Hungarian and Long Beach. We have used 10-fold cross validation to train and test the models where data is divided into 10 fold in which one by tenth partition is used to test the model and rest of the partition is used to train the model. This process continuous until all the partition used as testing data. .This chapter is divided into two sections: firstly analysis of classification techniques with WEKA data mining tools and secondly analysis of classification techniques with Rapid miner data mining tool.

A. Analysis of Models with WEKA Data Mining Tool

In this section we have used WEKA data mining tool for analysis and classification of different level of heart disease. We have used different data mining based classification techniques for classification of safe and different level of heart disease. Table III shows that tuning parameter of different techniques like C4.5, Random Forest, Random Tree, REP Tree, Naive Bayes, MLP and SVM. The tuning parameters are used to tune the techniques to improve the performance of model. Table IV shows that accuracy of models with four coronary heart disease data set with 10-fold cross validation. SVM gives better accuracy are 58.74%, 39.83% and 66.67% in case of Cleveland, Switzerland and Hungarian data set respectively while Naive Bayes gives best accuracy as 37.50% in case of Long Beach heart disease data set. Finally SVM gives best accuracy as 66.67% in case of Hungarian heart disease data set.

Table III. Tuning parameter different techniques with WEKA data mining tool

Techniques	**Tuning Parameter**
C4.5	Batch size=200, Confidence factor=0.50, MinNumObj=2, Seed=1
Random Forest(RF)	Bag Size percentage=100, Batch size=100,mim depth=0(unlimited), seed=1
Random Tree (RT)	Batch size=100, max depth=0(Unlimited),seed=1
REP Tree	Batch size=100,maxdepth=-1,minNum=2.0,seed=1
Naïve Bayes	Batch size=100,Num decimal places=2
MLP	Batch size=100, hidden layer=1, learning rate=0.3, momentum=0.2
SVM	Batch size=100,Calibrator=Logistic, Kernel=Poly kernel, randomseed=1

Table IV. Accuracy of models with different heart disease data set with 10-fold cross validation

Techniques	Cleveland	Switzerland	Hungarian	Long Beach
C4.5	49.83	34.95	65.98	28.50
Random Forest (RF)	57.42	36.58	66.32	35.50
Random Tree(RT)	48.84	33.34	58.84	27.50
REP Tree	57.42	36.58	63.26	28.00
Naive Bayes	56.43	31.70	62.58	**37.50**
MLP	57.09	34.95	63.60	30.00
SVM	**58.74**	**39.83**	**66.67**	32.00

B. Analysis of Models with Rapid Miner Data Mining Tool

Similarly previous section, we have used Rapid Miner data mining tool for analyzing and classification of different level of heart disease. We have also used different classification techniques for classification of safe and different level of heart disease. Table V shows that tuning parameter of different techniques like Decision Tree (DT), Random Forest, Decision Stump, Random Tree and Naïve Bayes. Table VI shows that accuracy of models with four

different coronary heart disease data with 10-fold cross validation. In case of Cleveland and Long Beach data set , Naïve Bayes gives better accuracy as 55.43% and 28.50% respectively. Decision Tree, Random Forest and Random Tree gives 39.10% as best accuracy in case of Switzerland heart disease data set while Decision Stump gives as best 63.94% of accuracy in case of Hungarian heart disease data set. Finally, Decision Stump is best classifier and achieved better accuracy as 63.94% in case of Hungarian data set.

Table V. Tuning parameter of different techniques with Rapid miner data mining tool

Techniques	Tuning Parameter
Decision Tree (DT)	Criterion=gain ratio, minimal size for split=4,minimal leaf size=2,minimal gain=0.1,maxmimal depth=20,confidence=0.25
Random Forest (RF)	Number of tree=10, Criterion=gain ratio, minimal size for split=4,minimal leaf size=2,minimal gain=0.1,maxmimal depth=20,confidence=0.25
Decision Stump (DS)	Criterion=gain ratio, minimal leaf size=1
Random Tree (RT)	Criterion=gain ratio, minimal size for split=4,minimal leaf size=2,minimal gain=0.1,maxmimal depth=20,confidence=0.25
Naïve Bayes	Estimation mode=greedy,min badwith=0.1,number of kernel=10

Table VI. Accuracy of models with different heart disease data set with 10-fold cross validation

Models	Cleveland	Switzerland	Hungarian	Long Beach
Decision Tree (DT)	50.78	**39.10**	57.78	28.00
Random Forest	54.13	**39.10**	63.28	27.00
Decision Stump	54.45	38.33	**63.94**	28.00
Random Tree	54.45	**39.10**	62.59	28.00
Naïve Bayes	**55.43**	32.69	56.59	**28.50**

IV. CONCLUSION AND FUTURE WORK

In this research work we have used effective predictive techniques for prediction of coronary heart disease with different classifier available in WEKA and Rapid Miner data mining tool. We have achieved 66.67% of accuracy with SVM as best classifier in case of Hungarian heart disease data set with WEKA data mining tool. We have also achieved 63.94% of accuracy with Decision Stump as best classifier with Hungarian data set in case of Rapid miner data mining tool. Finally we have suggested our classifier achieved better classification accuracy with Hungarian data set.

Identification and categorization is very important key factor to diagnosis of heart disease. Our experiment focused on analysis of data with individual's techniques. In future, we will try to develop hybrid model which will improve the classification accuracy. We will also use different feature selection techniques to reduce the features and improve the performance of model. In future, we will also predict the risk factor of coronary heart disease in Chhattisgarh region.

REFERENCES

[1]. A. Dhanasekar and R. Mala, "Analysis of Association Rule for Heart Disease Prediction from Large Datasets", International Journal of Innovative Research in Science, Engineering and Technology" , Vol. 5, Issue 10, 2016.

[2]. M. A. Jabbar, P. Chandra ,and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection", 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 628-634,2012.

[3]. Web source: http:// www.cs.waikato.ac.nz/~ml/weka/ last accessed on Feb. 2017.

[4]. J. Han and M. Kamber," Data Mining Concepts and Techniques, published by Morgan Kauffman,2nd ed.2, 2006.

[5]. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html].(Browing date: Feb 2017).

[6]. A. K. Pujari, " Data Mining Techniques", Universities Press (India) Private Limited, 4th ed., 2001.

[7]. Z. Tang and J. Maclennan, "Data Mining with SQL Server 2005",Willey Publishing, Inc, USA, 2005.

[8]. D. L. Olson and D. Delen, "Advanced Data Mining Techniques", USA, Springer Publishing, 2008.

[9]. RapidMiner Tutorials (2009),"User Guide Operator Reference Developer Tutorials". (Browning date: Jan. 2017).

[10]. H. W. Ian and F. Eibe "Data Mining Practical Machine Learning Tools and Techniques", Morgan Kaufmann, San, 2nd ed., 2005.

[11]. A. Jarad, R. Katkar, A. R. Shaikh and A. Salve," Intelligent Heart Disease Prediction System with MONGODB", International Journal of Emerging Trends & Technology in Computer Science,Vol. 4, Issue 1,pp. 236-239, 2015.

[12]. J. Patel, T. Upadhyay and S. Patel , "Heart Disease Prediction Using Machine learning and Data Mining Technique", IJCSC, Vol. 7, No. 1, pp. 129-137, 2016.

[13]. N. Kishore and S. Singh, "Cardiac Analysis and Classification of ECG Signal using GA and NN", International Journal of Computer Applications, Vol. 27, No. 12, pp. 23-27, 2015.

[14]. S. Raghavendra, and M. Indiramma, "Classification and Prediction Model using Hybrid Technique for Medical Datasets". International Journal of Computer Applications, Vol. 127 , No.5 , pp. 20-25, 2015.

[15]. M. Singla and K. Singh, "Heart Disease Prediction System using Data Mining Clustering Techniques", International Journal of Computer Applications and International Conference on Advances in Emerging Technology, pp. 1-5.