National Conference on Engineering Innovations and Solutions (NCEIS – 2018)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 6 | ISSN : 2456-3307

Heart Disease Classification: A Case Study using Machine Learning and Data Mining

Sourabh Kulkarni, Chaitra .D Bhat, Deepa Patil, Jovita Dara
Department of Computer Science and Engineering, K.L.E. Institute of Technology, Hubli, Karnataka, India

ABSTRACT

The diagnostic of heart disease remains more or less the most difficult and tedious task in the medical field and it various factors and symptoms of prediction which is involved in several layered issue that could engender the negative presumptions and unpredictable effects. Wu et al proposed that the integration of clinical decision support with relation to the computer- based system of the patient record could reduce the rate of errors in medical predictions, low the unwanted practice variation, enhance safety for patients, and the improvement of patient outcome. This knowledge provides a useful environment which can help to significantly improve the quality of clinical decisions. Many of hospital information in recent days are designed to implement patient billing, patient data storing, inventory management and generation of simple statistics computation. Most of the hospitals use decision support systems but they are still in most cases bounded. The majority of doctors are predicting heart disease symptoms based on their learning and working experience. In this case, prediction system should be implemented so that to reduce the risk of Heart Disease.

Keywords: Algorithms, Diseases, Heart-attack, Random forest, Decision trees , Data mining, Naive Bayes, Support Vector Machine.

I. INTRODUCTION

TheHeartdiseasehasbeenthemostsignificantcauseofdeathin the world during the past 10 years [1]. The use of heart monitoring systems, such as for example [2], and heart disease classification methodologies for decision support systems has been increasing accordingly. Unfortunately, many different factors can influence and complicate the detection of possible heart anomalies and can result in an inaccurate diagnosis or in a delay in a correct diagnosis.According to [3], due to the many and uncertain risk factors, sometimes heart disease diagnosis is difficult even for experts, who frequently require accurate tools that consider all these risk factors and give a clear result in a specific timeperiod.

Motivated by the need to acquire such an indispensable instrument for diagnosis and by the importance of avoiding as far as possible any unwanted biases, errors and excessive medical costs that might affect the quality of treatment provided to patients [4], many researchers have tried to find the most accurate machine learning techniques to discover the relationships between different heart disease and patient attributes in order toassist physicians [5], [6].

This paper provides a comparison of different machine learning classification techniques, such as

Decision Tree (DT), K Nearest Neighbour (K-NN), Random forest, and of their use in combination, through bagging, boosting and stacking on a heart disease data set. The dataset used is the Cleveland Heart Disease data set taken from the University of California, Irvine (UCI) learning data set repository, donated by Detrano.

Cardiovascular diseases (CVD) are the leading cause of death globally [1]. Diagnosis of CVD is a complicated and important taskthat needs to be executed accurately and efficiently. In order to improve the quality of health care and relieve the pressure of medical service, many data mining techniques are applied for clinical decision making supported by clinical decision support systems (CDSS). Classification is one of the most important algorithms in CDSS. The performance of classification is greatly affected by feature selection. It is a challenging problem to get good features. It is the motivation of thispaper.

Heart disease can be also known as (CVD) cardiovascular disease, contains a number of conditions that affect the heart including the heart attacks. In addition, Heart diseases possess some functional problems of the heart such as infections of the heart muscles like myocarditis (inflammatory heart diseases), heart-valve abnormalities or irregular heart rhythms etc. are the reasons that can be led to heart failure. The components that expansion the odds of heart attacks are smoking, absence of physical activities, hypertension, elevated cholesterol, unfortunate Eating routine, unfavourableutilisation of liquor, and high sugar levels. Cardio Vascular Disease (CVD) constitutes coronary heart, cerebrovascular or Stroke Hypertensive heart disease, inborn heart, fringe course, rheumatic heart disease and incendiary heart disease. Data mining is a learning revelation system to analyze data and typify it into valuable data.

II. LITERATURE REVIEW

The researchers [8] used Machine leaning and data mining methods in predicting models in the domain of cardiovascular diagnoses. The experiments were carried out using classification algorithms Random Forest, Decision Tree, K-NN and results proves that the decisiontrees provides better results than othercounterparts.

In the last few years, several studies have been dedicated to an evaluation of the classification accuracies of different classification algorithms applied to the Cleveland heart disease database [7] freely available at an online data mining repository of the UCI. Since its creation, this database has been used by many researchers investigating different classification problems with various classification algorithms. Detrano in [8] used a logistic regression algorithm and obtained a 77.0% classification accuracy.

In this paperwork [7] there are three different data mining techniques such as Random Forest, K-NN, Decision tree were addressed to analyse the dataset. In this paperwork, the experiment has been performed by the use of 3000 instances training dataset with 14 different attributes. The data set is classified into two categories in which we have 70% of the data were used for training while 30% were used for testing. Considering these experimental results, it is shown that the classification accuracy of decision tree algorithm is better compared to other algorithms.

Gudadhe et al. [5] realized an architecture base with both the MLP network and the SVM approach. This architecture achieved an accuracy of 80.41% in terms of the classification between two classes (the presence or absence of heartdisease, respectively). On the other hand, HumarKahramanli and NovruzAllahverdi [10] obtained an accuracy of 87.4% by using a hybrid neural network that combines a fuzzy neural network (FNN) with an artificial neural network (ANN).

Another study on heart disease prediction has been proposed and implemented by SY Huang, AH Chen, CH Cheng, PS Hong and EJ Lin. The classification and prediction was trained via learning Vector Quantization Algorithm which is one of Artificial Neural Network learning technique. There were three steps in their methodology. The first one was to select of 13 clinical features which are important compared to others, i.e., age, cholesterol, chest pain type, exercise induced angina, max heart rate, fasting blood sugar, number of vessels colored, old peak, resting ecg, sex, slope, thal and trestbps. Second one was using Artificial Neural Network algorithm for classification. Lastly, the heart disease prediction system was developed. The accuracy of prediction rate which was obtained from the study is near 80%. [4] Soni et al [9] provided a survey of current techniques in Data mining for heart disease prediction. Experiments has been conducted with various sorts of techniques using the same dataset out of which Decision tree shown high accuracy than that of the Bayesian classification, KNN, neural networks. The accuracy has been further improved by applying genetic algorithm with Decision trees. The work can be extended by using real dataset from health care organizations for the automation of Heart Diseaseprediction.

Rafiah et al [10] using Decision Trees, Naive Bayes, and Neural Network techniques developed a system for heart disease prediction using the Cleveland Heart disease database and shown that Naïve Bayes performs well followed by Neural Network and Decision Trees. The relationship between attributes produced by Neural Network is more difficult to understand than that of the other models used to predict heart disease. Continuous data can be used instead of categorical data and text mining methods can be incorporated to mine vast amount of unstructured data available in healthcare databases.

III. RESEARCH METHODOLOGY

Important concepts such as the data set, data portioning models and data mining techniques are described following.

1. The Cleveland DataSet

The data set used in the current research contains 303 instances with a total number of 76 attributes. However, the majority of the studies use a maximum of 14 attributes [11] as these are closely linked to heart disease [12]. The features included are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, oldpeak, slope, number of vessels colored and thalassemia, respectively. The main class has two values, “False” and “True”, corresponding to the absence or presence, respectively, of any heartdisease.

2. Machine LearningTechniques

In the current study, seven classifiers, namely DT, NB, MLP, RFB, SCRL, K-NN and SVM, and combinations of these classifiers, using ensemble learning methods such as bagging, boosting and stacking, are discussed. In each scenario, the performance is calculated using the standard metrics, namely accuracy, precision, recall and F-measure. In addition, the Receiver Operation Characteristic (ROC) curve area has been employed to compare the performance of each classifier

1) Decision Tree (DT): A Decision Tree is a flow chart- like structure that includes a root node, branches, and leaf nodes. The dataset attributes are defined through the internal nodes. The branches are the outcome of each test against each node. It is a popular classifier because it is simple, fast, and easy to interpret, explain and implement. It requires no domain knowledge or parametersetting.

2) 2) K-Nearest Neighbors (K-NN): K-Nearest Neighbors classifies an object by the majority vote of its closest neighbors. In other words,

based on some distance metrics, the class of a new instance will be predicted. The distance metric used in nearest neighbor methods for numerical attributes can be a simple Euclideandistance.

3) Support Vector Machine (SVM): Support Vector Ma-chine models are defined as finite-dimensional vector spaces in which each dimension represents a 'feature" of a particular object. It has been shown to be an effective approach in high-dimensional space problems. Due to its computational efficiency on large datasets this technique is usually used in document classification and sentimentanalysis.

DATA CLASSIFICATION USING RANDOM FOREST CLASSIFIER

1) **Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. It is one of the most accurate among the learning algorithmsavailable.**

For many data sets, it produces a highly accurate classifier. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler [12] the method combines Breiman's "bagging" idea and the random selection of features, in order to construct a collection of decision trees with controlledvariation.

Algorithm: Random forest

classifier Input:

1. Training Dataset N, Which is a set of training observations and their associated classvalues.

Output: Generates Decision trees

Each tree is constructed based on the followingsteps:

1. Let the number of training cases be N, and the number of variables in the classifier beM.

2. The number m of input variables to be used to determine the decision at a node of the tree; m should be much less thanM.

3. Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting theirclasses.

4. For each node of the tree, randomly choose m variables on which to base the decision at that node.Calculate the best split based on these m variables in the trainingset.

5. Each tree is fully grown and not pruned (as may be done in constructing a normal treeclassifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal nodeit ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction. done in constructing a normal tree classifier). For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forestprediction.

DATASETS

Table 1. Confusion Table

Prediction		Disease	
		+	-
Test	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

For the datasets, we met patients having past history of heart disease, gathered valuable information from

the doctor, compiled all the reports into the dataset consisting of more than 100 records. We also gathered necessary datasets form the internet for testing purpose. The following attributes with nominal values are considered: Patient Identification Number (replaced with dummy values), Chest Pain, Cholesterol, Fasting Blood Sugar, Rest ECG, Thalach (Maximum heartrate achieved), Exang (Exercise induced angina) and Slope (the slope of the peak exercise STsegment) Cleveland database was used for heart disease prediction system. Because Cleveland database is the most commonly used database by ML researchers. The dataset contains 303 instances and 76 attributes, but only 14 of them are referred by all published studies. The "goal" field which has varying values from 0(absence) to 4 denotes if heart disease present or not in the patient. Studies on the Cleveland database have focuses on distinguishing absence(value 0) from presence (values range from 1 to 4) [13].

The dataset has some missing values in it. Firstly missing values were filled with interpolation values. Then dataset was split into three parts: one for training (%70), second one for testing (%15) and third one for validation(%15). There are 213 instances and 13 attributes in training data. Test data and validation data contain 45 instances and 13 attributes 13 of the attributes listed below were used as input data for the networ. The remaining attribute, num which is predicting value, was used as output data for the network. The num can get values between 0 and 4. Only 0 means absence of disease.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Attribute information (only 14 used) is shown in Table

Table 2. Clinical features and their descriptions

Clinical Features	Description
Age	Age
Ca	Number of major vessels (0-3) colored by flourosopy
Chol(mg/dl)	Serum cholesterol
Cp	Chest pain type
Exang	Exercise induced angina
Fbs	Fasting blood sugar
Num	Diagnosis of heart disease
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segment
Thal	3=normal ; 6 = fixed defect; 7= reversible defect
Thalach	Maximum heart rate achieved
Trestbps(mmHg)	Resting Blood Pressure

IV. PERFORMANCE MEASURES

2) In this approach, the classification accuracy rates for the datasets were measured. For example, in the classification problem with two-classes, positive and negative, an single prediction has four possibility. The True Positive rate (TP)and True Negative rate (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative. A False Negative (FN)

3) occurs when the outcome is incorrectly predicted as negative when it is actually positive.

1. Accuracy - It refers to the total number of records that are correctly classified by theclassifier.

2. Classification error - This refers to the misclassifieddatasets from the correctly classifiedrecords.

3. True Positive Rate (TP) - It corresponds to the number of positive examples that have been correctly predicted by the classificationmodel.

4. False Positive Rate (FP) - It corresponds to the number of negative examples that have been

wrongly predicted by the classificationmodel.

5. Kappa Statistics - A measure of the degree of nonrandom agreement between observers or measurements of the same categoricalvariable.

6. Precision - is the fraction of retrieved instances that are relevant.

$$Precision = \frac{TP}{TP + FP}$$

7. Recall - is the fraction of relevant instances that are retrieved.

8. Root-Mean-Squared-error - It is a statistical measure of the magnitude of a varying quantity. It can be calculated for a series of discrete values or for a continuously varying function.

Since the class label prediction is of multi-class, the result on the test set will be displayed as a two-dimensional confusion matrix with a row and column for each class. Each matrix element shows the number of test cases for which the actual class is the row and the predicted class is thecolumn.

V. CONCLUSION

Data mining plays an important role in the identification and prediction of various sort of metabolic syndromes and hence various sorts of diseases can be discovered. In the existing work, Decision tree classification algorithm has been used to assess the events related to CHD. The proposed work is mainly concerned with the development of a data mining model with the Random Forest classification algorithm. The developed model will have the functionalities suchaspredicting the occurrence of various events related to each patient record, prevention of risk factors with its associated cost metrics and an improvement in overall prediction accuracy. As a result, the causes and the symptoms related to each event will be made in accordance with the record related to each patient and thereby CHD can be reduced to a great extent.

VI. REFERENCES

1. "World Health Organization," accessed on 03-02-2017. [Online].Available: http://www.who.int/mediacentre/factsheets/fs310.pdf
2. G Sannino and G. De Pietro, "A smart context-aware mobile monitoring system for heart patients," in Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on. IEEE, 2011, pp.655–695.
3. A Adeli and M. Neshat, "A fuzzy expert system for heart disease diagnosis," in International Multi Conference of Engineers and ComputerScientists, vol. 1, 2010.
4. N Allahverdi, S. Torun, and I. Saritas, "Design of a fuzzy expert system for determination of coronary heart disease risk," in International conference on Computer systems and technologies, 2007, p. 36.
5. M Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in Computer and Communication Technology (ICCCT), 2010 International Conference on, 2010, pp. 741–745.
6. H Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multilayer perceptron-based medical decision support system for heart disease diagnosis," Expert Systems with Applications, vol. 30, no. 2, pp. 272–281,2006.
7. A Rajkumar and G. S. Reena, "Diagnosis of heart disease using data mining algorithm," Global journal of computer science and technology, vol. 10, pp. 38-43, 2010.
8. K Srinivas, G. Raghavendra, Rao, and A. Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques", 5th

IntConf on Computer Science &Education Hefei, China, pp. 1344-1349, 2010.

9. J Soni, U. Ansari, and D. Sharmaa "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" Int., Journal of ComputerApplications vol. 17, No. 8, 2011.
10. A.Rafiah and P.Sellappan "Intelligent Heart DiseasePrediction System Using Data Mining Techniques", 2008.
11. I.H. Karlberg, and S.L. Elo "Validity and utilization of epidemiological data: A study of Ischaemic heart diseaseand coronary risk factors in a local population", 2009.
12. L. Breiman and A. Cutler " www.stat.berkeley.edu".
13. A. Reena and S.G. Rajkumar, "Diagnosis of heartdisease using Data mining algorithm", Global Journal ofComputer science and Technology, Vol. 10, No. 10, 2010.