Indian Journal of Science and Technology, *Vol 9(31), DOI: 10.17485/ijst/2016/v9i31/96740, August 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Preliminary Cardiac Disease Risk Prediction Based on Medical and Behavioural Data Set Using Supervised Machine Learning Techniques

Thendral Puyalnithi* and V. Madhu Viswanatham

School of Computing Science and Engineering, VIT University,Vellore - 632014,Tamil Nadu,India;
thendral_psg@yahoo.com,vmadhuviswanatham@vit.ac.in

Abstract

Objectives: The objective of the work is to detect the probable signs and symptoms which might further lead to detection of cardiac diseases using the learned system which is trained using data collected from previous patient.**Methods/Statistical Analysis**: The data set has been taken from reputed machine leaning data set repositories. The data has been cleaned, imputed and then the Outliers are removed before using them for training purpose. The Classification methods which are nothing but Supervised Learning of Machine Learning Technology is used to train the system. In this work Classification Tree, Naive Bayes, Random Forest and Support Vector Machine algorithms are used for training the Prediction System. The experiment has been conducted using the python based Data Mining tool called *Orange* and the scores have been evaluated.**Findings:**The comparison of precision of various supervised learning algorithms are analysed and has been found out that Classification Trees are efficient in Prediction. **Application/Improvements:**As future work the data of Hospitals and Health Research Institutes can be uploaded in Cloud and the data analysis can be done extensively to get an accurate Prediction which can be used across many hospitals and research institutes throughout the world

Keywords: Cardiac Disease Prediction, Classification Algorithms, Data Mining, Machine Learning, Supervised Learning

1. Introduction

Cardiac arrest is one of the major illnesses in the world which brings fatality. It is caused by a sudden occurrence of coronary thrombosis, typically which results in the death of a particular heart muscle and sometimes can be fatal. If the flow of oxygen abundant blood to a part of heart muscle is blocked due to fat deposition or fat detachment, then the heart cannot get enough oxygen and that leads to condition called heart attack. The plaque deposition inside the arteries is said to be atherosclerosis. The plaque build-up will happen through years and the rate of plaque build-up depends on various parameters and it varies from person to person. The Plaque area might break up. This leads to formation of blood clot on the surface of the plaque. The flow of the blood in the artery will be fully blocked if the clot gets bigger. If that block in the artery is not treated, he part of the heart muscle which gets blood from that artery will be affected and it leads to the death of the heart muscle and that will tissue will eventually turn into scar tissue. This heart condition may not be obvious and so this might not be detected at all until the unexpected sudden attack happens. Thus we can say the heart attack is proportional to the coronary heart disease. The coronary heart disease is nothing but the building up of plaque inside the coronary arteries. Early prediction is needed to diagnose the coronary heart condition. The cause of heart attack is severe spasm or we can say tightening of the coronary artery. The spasm can block the blood flow to the heart fully rendering the heart inactive and dead and leads to severe pain and resulting in death of the patient. We cannot say for sure that the Coronary Heart condition will always lead to heart attack and further it differs from person to person .A Person's genetic makeup also contributes to the extent of plaque deposition.

**Author for correspondence*

Since we have enormous quantity of clinical data related to coronary heart disease, Data mining tools are used to find the interesting relationships present among the data. Prediction or Regression, Classification and Clustering which are some of the Data mining techniques used for finding or analysing the pattern present the vast chunk of data. Regression and Classification are the supervised learning based to techniques of data mining. Supervised learning is applied where we have the data with label or we can say if we have labelled data we can go for supervised learning and we can train a system for prediction or classification. Regression is used to predict a continuous value and Classification is used to find the appropriate class, which is a discrete data. Clustering technique is unsupervised learning based technique. Unsupervised learning is applied when we have unlabelled data. So if we have unlabelled data we cannot train the system and so we directly go for grouping the similar data items in the data set. This process of grouping similar data is said to be Clustering. Once the similar data are grouped then we can have some number of clusters. So when new unlabelled data is encountered, we can place the unknown data in one of the groups by finding the closeness of new data with every group. Whichever group's data is similar to the new unlabelled data, then we place the new data in that group. Finally if you label some group, then the label applies to all the members of the group. There are many algorithms for performing Regression, Classification and Clustering. The focus of this paper is about creating a system trained with labelled data, since we have labelled data. So in this paper we are going to focus of regression and classification techniques not on clustering techniques. The algorithms we are going to use for performing classification are Naïve Bayes, Support Vector Machine (SVM), Random Forest (RF) and Classification Tree (ID3). The aim of this paper is to compare the performance of various classification algorithms which are used over coronary heart disease data set. The dataset is having health and habit of person as a record. This data set is not about clinical records, which are usually the values of blood, urine and other type of clinical tests Some of the attributes of the data set are whether a person has Stress, Blood Pressure, Congenital Disease, Diabetes, Drug abuse and so on. The classification models are built based on the algorithms mentioned above and the one which gives highest accuracy in detection will be used as a standard Model. If the new data is inducted into the data set then again the models are generated and the best one to do the prediction is found out.

Training a system requires lot of data. If we have got more number of labelled data the trained system will be accurate in classifying the unknown data. Classification Tree (ID3) uses Information theory concept to create the classification tree. Once the ID3 tree is created, when the unlabelled data arrives, it will be fed through the root of the classification tree, when the choice selection reaches the leaf the tree, we have to stop going further and that indicates class for that unknown data. Support Vector Machine based classifiers uses vector concepts to create the trained system and it's usually effective for small data sets. Random Forest classifiers uses set of Classifier tree (ID3) Models to classify a particular unknown data. For creating random forest, we usually split the data sets into many sub-data sets and each sub-data set is used for creating a classification tree(ID3) and the unknown data will be classified by each tree and based on voting system one class will be taken as correct classification. Naive Bayes is a classifier which works based on Probabilistic model.

2. Previous Works

Nowadays technology is assisting the physicians to detect a disease. Diseases can be detected though many ways. Some of the ways where technology assist physicians to detect are diseases are though Body Fluid Analysis(Blood, Urine), Imaging (CT,MR) and Electric Pulse Measures (EEG,ECG)

Image analysis is used to detect some diseases. For example, cancer detection is usually done using the CT scan inputs. Lund caner Prediction using Back propagation way of Neural network has been experimented by Ada et al. in a research paper[1].

The patient data records can also be clustered using Clustering analysis. Many clustering algorithms are available. If the data set is not labeled, then clustering techniques of machine learning which are said to be Unsupervised learning techniques are quite handy in analyzing. In the paper *A Survey on Clustering Techniques for Big Data Mining*[2]a variety of clustering algorithms are compared. As a future work, the clustering algorithms can also be used to assist the knowledge discovery in health care domain.

The work in this paper is towards using only Supervised Learning Algorithms, so we are restricting to only classification algorithms. The performance of Various ID3

versions of Classification Treesare analyzed by Teli et al. in *A Survey on Decision Tree Based Approaches in Data Mining*[3]

Comparison of Classification Algorithms Performance for measuring Students' performance has been performed by*An Analysis of students' Performance using Classification algorithms*[4]. The same kind of work is done for Health care analysis in this paper.

In [5] about implementing many family of classification algorithms in analyzing medical records for diagnosis. They have worked on data sets obtained from UCI Machine learning repository for analysis of Breast Cancer, Liver Disorder and Hepatitis Data sets.

In[6] analyzed the performance of various classification algorithms for to improve the system in e-learning. This kind of Classification algorithms' performance evaluation substantiates that the supervised learning algorithms usage is justified for Decision making systems.

In[7]has compared classification results of Random Forest and J48 based classification tree over 20 sets. They have compared the precision of classifiers with the size of the data sets that is the number of instances. It holds the result that if number of instances are more the prediction precision increases.

In [8] has used Naïve Bayes, Max entropy classifier, Random forest and Boosted Tree classifiers for Document classifiers. The use of classification algorithms in many fields proved its effectiveness for real life applications.

In the Paper, Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques[9], classification algorithms are used to classify lung cancer data. In the paper Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability[10], the author Edeki C has used classification techniques to predict Cancer Survivability for patients.

Thus the literature survey has made us to come to the conclusion that the classification algorithms are effective in predicting health care related data for diagnosing diseases.

3. Proposed Method

The data set is obtained from UCI Machine learning repository. First the data is pre-processed and then stored in proper format. The model creation and evaluation is based on two methods, k-fold Cross Validation and Leave One-out. In the Leave one out category 75% of the data set is taken for training and remaining 25% samples are used for testing the trained system. In k-fold Cross validation the total samples are split into k equal folds and (k-1) folds are used for testing leaving the remaining for testing. Then the evaluation is done for the trained system using the remaining k test samples. This process is repeated till all the (k-1) combinations are taken for training the system having the remaining test samples for testing, then average value of evaluation parameters are found out. Leave one out and Cross validation are followed for all the four algorithms SVM,Classification Tree(ID3),Naïve bayes, and Random Forest and the system is evaluated for based on the following evaluation parameters AUC,Accuracy,F1,Precision and Recall. Figure 1 shows the Proposed Model.

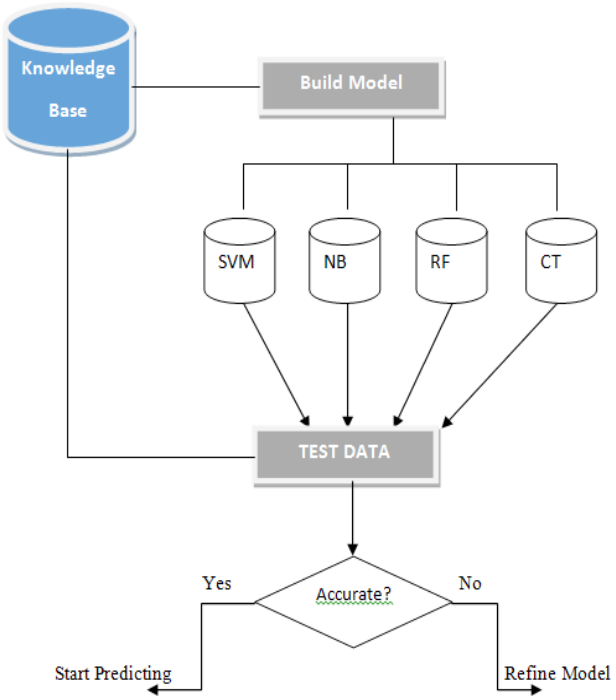

Figure 1. Proposed model.

In the first stage of the proposed method, the data has been pre-processed. Before the labelled data is getting used for training the outliers should be removed. So the data has gone through outlier detection techniques and after the removal, the classifier models are created. Four classifier are used for crating models based two methods leave one out and cross validation. So totally there are 8 models. The eight models are evaluated and results are shown in the tables. Orange tool has been used for generating the models. The classification Tree that has been generated is shown in the Figure 2.

The next step is to choose the best classifier among them. To decide the best classifier, we are using the following evaluation parameters AUC(area under the curve), Classification Accuracy (CA), Precision and Recall.

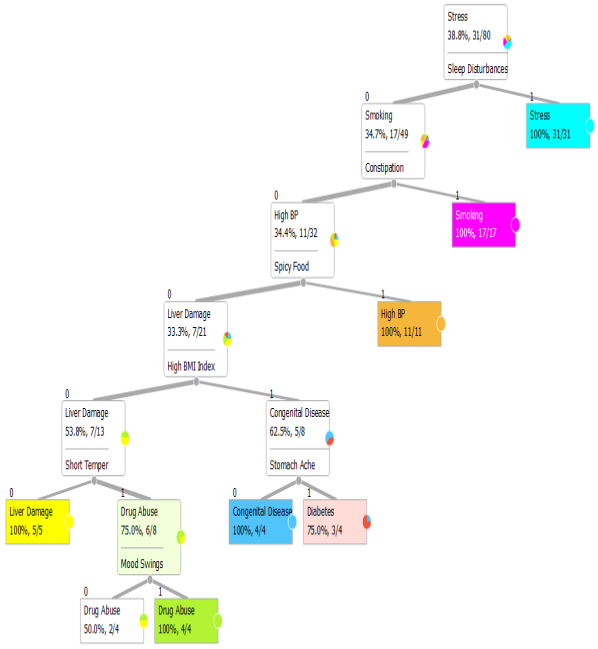

Figure 2. Decision tree for given data set.

Area Under an ROC Curve is one of the evaluation parameter where True positive rate and false positive rate are plotted and the curve is analysed to find the effectiveness.

Classification Accuracy = Correct Classifications/ Total Classifications

Precision = TP/(TP+FP) where TP is True Positive, and FP False Positive. Here Positive refers to, Heart Attack being happened and Negative refers to No Heart Attack.

Recall = TP/(TP + FN)

In Table 1 the results of Various Classification algorithms based on Leave One Out is shown.

Table 1. Leave-one out analysis results

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.959	0.938	0.938	0.939	0.938
Naive Bayes	0.945	0.912	0.912	0.923	0.912
Random Forest	0.890	0.838	0.819	0.815	0.838
SVM	0.782	0.738	0.700	0.744	0.738

Cross validation analysis is the best method to solve the problem of over fitting.In Table 2 the results of Various Classification algorithms based on Cross Validation is shown.

Table 2. Cross validation analysis results

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.940	0.925	0.922	0.925	0.925
Naive Bayes	0.942	0.912	0.912	0.923	0.912
Random Forest	0.917	0.887	0.868	0.851	0.887
SVM	0.731	0.688	0.652	0.727	0.688

4. Results and Discussion

The evaluation parameter values in the above tables shows that the Classification Tree (ID3) based classifier works very well. The Classification Tree (ID3) has the best values for accuracy, precision and recall. Now from this we come to know that the main risk factor which might lead to heart attack. We can now say the percentage of chances for getting the Heart attack. The confusion matrix shows the performance of Classification Tree. Figure 3 shows the confusion matrix for classification tree.

Predicted

Actual	Congenital Diseases	Diabetes	Drug Abuse	High BP	Liver Damage	Smoking	Stress	∑
Congenital Diseases	3	0	1	0	0	0	1	5
Diabetes	1	2	0	0	0	0	0	3
Drug Abuse	0	0	5	0	1	0	0	6
High BP	0	0	0	11	0	0	0	11
Liver Damage	0	0	1	0	5	0	1	7
Smoking	0	0	0	0	0	17	0	17
Stress	0	0	0	0	0	0	31	31
∑	4	2	7	11	6	17	33	80

Figure 3. Confusion matrix for classification tree.

Correctly Classified Instances	745	99.866 %
Incorrectly Classified Instances	1	0.134 %
Kappa statistic	0.9984	
Mean absolute error	0.0097	
Root mean squared error	0.0263	
Relative absolute error	4.1035 %	
Root relative squared error	7.636 %	
Total Number of Instances	746	

Figure 4. Performance evaluation of classification tree.

5. Conclusions

Thus we can find the which major risk factor can lead to heart attack for a particular person. Thus by making a person going through simple screening process which records the BP,Stress,Behavioural habits and other data for a person and the tool can predict a major risk data for that trait and the person can take precautionary measures to alleviate the risk of getting Heart attack thereby reducing the fatality rate. This data set is a preliminary data set which coarsely predicts the risk parameter. If the person's risk parameter is high, then the system trained with Clinical data (Bloodwork, Scan, ECG) should be used as a next step for further prediction of risk of getting heart attack.

6. References

1. Kaur R.Using some data mining techniques to predict the survival year of lung cancer patient. International Journal of Computer Science and Mobile Computing.2013; 2(4):1–6.
2. Sajana T, Rani CMS, Narayana KV.A survey on clustering techniques for big data mining. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–12.DOI: 10.17485/ijst/2016/v9i3/75971.
3. Teli S, Kanikar P.A survey on decision tree based approaches in data mining. International Journal of Advanced Research in Computer Science and Software Engineering. 2015;5(4):1–5.
4. Mythili MS,Shanavas ARM. An analysis of students' performance using classification algorithms. IOSR Journal of Computer Engineering.2014; 16(1):63–9.
5. Godara S, Singh R. Evaluation of predictive machine learning techniques as expert systems in medical diagnosis. Indian Journal of Science and Technology. 2016 Mar; 9(10):1–14. DOI: 10.17485/ijst/2016/v9i10/87212.
6. Murugananthan V, Kumar BLS.An adaptive educational data mining technique for mining educational data models in e-learning systems. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–5.DOI: 10.17485/ijst/2016/v9i3/86392.
7. Ali J, Khan R, Ahmad N, Maqsood I.Random forests and decision trees. International Journal of Computer Science. 2012;9(5):1–61.
8. Gupte A, Joshi S, Gadgul P, Kadam A. Comparative study of classification algorithms used in sentiment analysis. International Journal of Computer Science and Information Technologies. 2014;5(5):1–4.
9. Krishnaiah V. Diagnosis of lung cancer prediction system using data mining classification techniques.International Journal of Computer Science and Information Technologies. 2013;4(1):39–45.
10. EdekiC. Comparative study of data mining and statistical learning techniques for prediction of cancer survivability. Mediterranean Journal of Social Sciences. 2012;3(14).