Indian Journal of Science and Technology, *Vol 8(12), DOI: 10.17485/ijst/2015/v8i12/58385, June 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Survey of Data Mining Techniques on Risk Prediction: Heart Disease

G. Purusothaman* and P. Krishnakumari

Department of Computer Applications (MCA), RVSCAS, Coimbatore, 641402, India; purusothamangd@gmail.com

Abstract

Comparison of classification techniques in Data mining to find the best technique for creating risk prediction model of heart disease at minimum effort. In Data mining, different methods used to find risk prediction of heart disease. There are two types of model used in analysis of data. First one is applying single model to various heart data and another one is applying combined model to the data. The combined model also known as hybrid model. This paper provides a quick and easy understanding of various prediction models in data mining and helps to find best model for further work. This is unique approach because various techniques listed and expressed in bar chart to understand accuracy level of each. These techniques are chosen based on their efficiency in the literature. In previous studies of different researcher expressed their effort on finding best approach for risk prediction model and here we found best model by comparing those researcher's findings as survey. This survey helps to understand the recent techniques involved in risk prediction of heart disease at classification in data mining. Survey of relevant data mining techniques which are involved in risk prediction of heart disease provides best prediction model as hybrid approach comparing with single model approach.

Keywords: Classification, Data Mining Algorithms, Heart Disease Risk Prediction, Hybrid Techniques, Survey of Data Mining Techniques, Prediction Models

1. Introduction

Heart Disease Prediction Model can support medical professionals and practitioners in predicting heart disease status based on the clinical data of patients. In biomedical field data mining and its techniques plays an essential role for prediction of various diseases. The physicians may not able to diagnose it correctly when the patients suffer from more than one type of disease of the same category. Because of missing concentration or unhealthy practices when prediction of disease category. The healthcare industry gives huge amounts of healthcare data and that need to be mined to ascertain hidden information for valuable decision making. Discover of hidden patterns and relationships often go unused[1]. The patient's record is classified and predicted if they have the symptoms of heart disease and using risk factors of disease. It is indispensable to find the best fit algorithm that has greater accuracy, less cost, speedy and memory utilization on classification in the case of heart disease prediction category.

1.1 Classification

Classification is a more useful data mining function that handle items in a collection to target categories or classes. The objective of classification is to accurately predict the target class for all case in the data. A classifier is able to learn based on given sample. The dataset used for training consists of information x and y for all data-point. Then x denotes what is generally a vector of observed characteristics for the data-item. And y denotes a group-label. The label y can take only a fixed number of values. Machine learning with classification can efficiently be applied for medical applications for complex measurements. Modern classification approaches provides more intelligent techniques for effective prediction of Heart Disease.

**Author for correspondence*

1.2 Heart disease

The heart is significant organ or part of our body. Human Life is itself dependent on efficient working of heart. If function of heart is not good then it will influence the other parts of human body such as brain, kidney etc. In case, the circulation of blood in body is inefficient the organs like brain and heart suffer. Generally blood arrest in brain is called as stroke, and in heart is called as attack. Life is completely dependent on efficient working of the heart and brain. The function of both is dependent of each one. There are number of factors which increase the risk of Heart disease[2]. Such as Heredity, Smoking, Cholesterol, Poor diet, High blood pressure, High blood cholesterol, Obesity, Physical inactivity and hyper tension.All heart diseases belong to the category of cardiovascular diseases. Types of cardiovascular diseases are Coronary heart disease, Angina pectoris, Congestive heart failure, Cardiomyopathy, Congenital heart disease, Arrhythmias and Myocarditis.

Many studies have been done that have focus on prediction of heart disease. Researchers have applied different data mining techniques for predicting heart disease and achieved different probabilities for different prediction models.

The work of Heon Gyu Lee et al.[3] was a novel approach to implement the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability). They found the linear and the non-linear properties of HRV for three recumbent positions like to be precise the supine, left lateral and right lateral position. They were conducted numerous experiment on linear and nonlinear characteristics of HRV indices to assess several classifiers.

Association Rule for classification of Heart-attack patients was proposed N. Deepika et al[12]. First stage, the data warehouse pre processed to make the mining work more efficient. The heart disease data warehouse is having the screening clinical data of heart patients. The extraction of significant patterns from the heart disease data warehouse was presented. The Association Rule used to pre process in order to handle missing values, and applied equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data.

Sellappan Palaniappan et al.[4] proposed a model called as Intelligent Heart Disease Prediction System (IHDPS) built with the aid of data mining techniques like Decision Trees, Naive Bayes and Neural Network. The results demonstrated the peculiar power of each of the methodologies in comprehending the objectives of the specified mining objectives.

In Carlos Ordonez[5] study, the problem of identifying constrained association rules for heart disease prediction. The assessed data set covered medical records of people having heart disease with attributes for risk factors. The heart perfusion measurements and artery narrowing was identified.

The work of Franck Le Duff et al.[6] might be executed for each medical procedure or medical problem. And it would be practical to build a decision tree rapidly with the data of a service or a physician. The main drawback of the study was knowledge acquisition. And the need to collect adequate data to create an appropriate model was necessary.

K. Srinivas et al.[13] proposed Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks. The prospective use of classification based data mining techniques such as Rule based, Decision tree, Naive Bayes and Artificial Neural Network to the massive Volume of healthcare data. Tanagra data mining tool was used for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 3000 instances with fourteen different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease.

In Latha Parthiban et al.[7] presented a model on basis of Coactive Neuro-Fuzzy Inference System (CANFIS) for prediction of heart disease. The CANFIS model diagnosed the presence of disease by merging different techniques that includes the neural network adaptive capabilities, the fuzzy logic qualitative approach and further integrating with genetic algorithm.

In Lei Yu and Huan Liu[8] proposed a novel method. It was named as predominant correlation, and proposed a fast filter method which can identify relevant features as well as redundancy among relevant features without pair wise correlation analysis.

J. Shreve, H. Schneider, O. Soysal[9] was proposed a methodology for comparing classification methods through the assessment of model stability and validity in variable selection. This study provides a systematic design for comparing the performance of six classification methods using Monte Carlo simulations and illustrates that the variable selection process is integral in comparing methodologies to ensure minimal bias, enhanced stability, and optimize performance.

Sudha et al.[14] to propose the classification algorithm like Naive Bayes, Decision tree and Neural Network for predicting the stroke diseases. The classification algorithm like decision trees, Bayesian classifier and back propagation neural network were adopted in this study. The records with irrelevant data were removed from data warehouse before mining process occurs. Data mining classification technology consists of classification model and evaluation model.

Feature selection for SVM via optimization of kernel polarization with Gaussian ARD kernels proposed by Tinghua Wanga, Houkuan Huang, Shengfeng Tian, Jianfeng Xu[10]. This work focused on effective feature selection method for Support Vector Machine (SVM). Unlike the traditional combinatorial searching method, feature selection is translated into the model selection of SVM.

M. A. Jabbar et al. proposed Association Rule mining based on the sequence number and clustering for heart attack prediction[15]. The entire database is divided into partitions of equal size. The dataset with fourteen attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation.

S. B. Patil and Y. S. Kumaraswamy[11] proposed a MAFIA (MAximal Frequent Itemset Algorithm) to extract the data relevant to heart attack from the warehouse. Then the significant weightage of the frequent patterns are calculated. Further, the patterns significant to heart attack prediction are chosen based on the calculated significant weightage. These significant patterns can be used in the development of heart attack prediction system.

Mai Shouman, et al[16] proposed k-means clustering with the decision tree method to predict the heart disease. In their work they suggested several centroid selection methods for k-means clustering to increase efficiency. The thirteen input attributes were collected from Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity, and accuracy are calculated with different initial centroids selection methods and different numbers of clusters.

Syed Umar Amin et al.[17] developed genetic neural network hybrid system. It uses the global optimization advantage of genetic algorithm for initialization of neural network weights. A back propagation algorithm trains the networks with optimize initialization of synaptic weights.

Ranjana Raut et al.[19] developed and proved the dimensionally reduced MLP Neural Network method is fastest network. They observed that MLP NN is simple in design and synthesis, lowest average MSE, highest accuracy and ROC analysis is perfect. They did Experiments with the Switzerland heart disease database on attempting to distinguish presence and absence.

Venkatesan P. et al.[20] studied and compared the effectiveness of the three popular classification algorithms namely C4.5, ID3 and CART to classify Tuberculosis dataset. It is more useful in medical research to construct algorithms for disease classification and prediction. The observation shown that C4.5 and CART performs better in terms of accuracy.

The performance of different algorithms shown on bar chart (Figure 1).

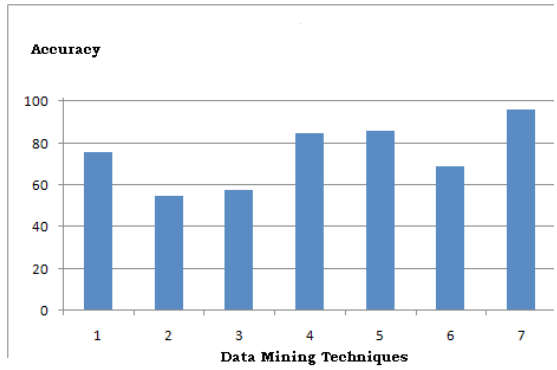

Figure 1. Performance level of data mining algorithms.

The performance study of different data mining algorithms in risk prediction of Heart disease as given below Table 1. The study performed based on accuracy of algorithms at different Heart disease data set.

Table 1. Performance study of data mining algorithms

S.No.	The Algorithm Used	Accuracy (%)
1	Decision Tree	76
2	Association Rule	55
3	K-NN	58
4	Artificial Neural Network	85
5	SVM	86
6	Naive Bayes	69
7	Hybrid Approach	96

2. Conclusion

Classification techniques are accomplished of processing a large amount of data. It is one of the most widely used methods of Data Mining in Healthcare organization. The widespread classification techniques used in risk prediction of heart disease are Bayesian Networks, Artificial Neural Network, Nearest Neighbour method, Fuzzy logic, Fuzzy based Neural Networks, Decision trees, Genetic Algorithms and Support Vector Machines[18]. Also, applying hybrid data mining techniques has revealed promising results in the diagnosis of heart disease. There is being required to have reliable model for predicting the existence or absence of heart disease with known and unknown risk factors. Some time poor clinical decisions lead to mortality. And all clinicians are not equally good in predicting the heart disease. In the case of heart disease time is precious, proper risk identification at the right time saves life of many patients..

3 References

1. Dangare CS, Apte SS. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications. 2012; 47(10):44–8.
2. Yanwei X, Wang J, Zhao Z, Gao Y. Combination data mining models with new medical data to predict outcome of coronary heart disease. Proceedings International Conference on Convergence Information Technology; 2007. p. 868–72.
3. Lee HG, Noh KY, Ryu KH. Mining biosignal data: coronary artery disease diagnosis using linear and nonlinear features of HRV. LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining; 2007 May. p. 56–66.
4. Shreve J, Schneider H, Soysal O. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. Decision Support Systems. 2011; 52:247–57.
5. Ordonez C. Improving heart disease prediction using constrained association rules. Seminar Presentation at University of Tokyo; 2004.
6. Lemke F, Mueller J-A. Medical data analysis using self-organizing data mining technologies. Systems Analysis Modeling Simulation. 2003; 43(10):1399–408.
7. Parthiban L, Subramanian R. Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences. 2008; 3(3).
8. Li W, Han J, Pei J. CMAR: accurate and efficient classification based on multiple association rules. Proceedings of 2001 International Conference on Data Mining; 2001.
9. Shreve J, Schneider H, Soysal O. A methodology for comparing classification methods through the assessment of model stability and validity in variable selection. Decision Support Systems. 2011; 52:247–57.
10. Wanga T, Huang H, Tian S, Xu J. Feature selection for SVM via optimization of kernel polarization with Gaussian, ARD kernels.. Expert Systems with Applications. 2010; 37:6663–8.
11. Patil SB, Kumaraswamy YS. Extraction of significant patterns from heart disease warehouses for heart attack prediction. International Journal of Computer Science and Network Security (IJCSNS). 2009; 9(2):228–35.
12. Deepika N,. Chandrashekar K. Association rule for classification of Heart Attack Patients. International Journal of Advanced Engineering Science and Technologies. 2011; 11(2):253–57.
13. Srinivas K, Rani KB, Govrdhan A. Application of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering. 2011; 2(2):250–5.
14. Sudha A, Gayathiri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications. 2012; 43(14):26–31.
15. Jabbar MA, Chandra P, Deekshatulu BL. Cluster based association rule mining for heart attack prediction. Journal of Theoretical and Applied Information Technology. 2011; 32(2):197–201.
16. Shouman M, Turner T, Stocker R. Integrating decision tree and K-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. Proceedings of the International Conference on Data Mining; 2012.
17. Amin SU, Agarwal K, Beg R. Genetic neural network based data mining in prediction of heart disease using risk factor. Proceeding of IEEE Conference on Information and Communication Technologies (ICT); 2013 Apr. p. 1227–31.
18. Kumari M, Godara S. Comparative study of data mining classification methods in cardiovascular disease prediction. International Journal of Computer Science and Technology. 2011 Jun; 2(2):304–8.

19. Raut R, Dudul SV. Maximum heart rate resting blood pressure scatter plot for the prominent features abnormal normal design and performance analysis of MLP NN based binary classifier for heart diseases. Indian Journal of Science and Technology. 2009 Aug; 2(8):43–8.

20. Venkatesan P, Yamuna NR. Treatment response classification in randomized clinical trials: a decision tree approach. Indian Journal of Science and Technology. 2013 Jan; 6(1):3912–17.