Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection

Indu Yekkala, Vardhaman College of Engineering, Hyderabad, India
Sunanda Dixit, Dayananda Sagar College of Engineering, Banglore, India

ABSTRACT

Data is generated by the medical industry. Often this data is of very complex nature—electronic records, handwritten scripts, etc.—since it is generated from multiple sources. Due to the Complexity and sheer volume of this data necessitates techniques that can extract insight from this data in a quick and efficient way. These insights not only diagnose the diseases but also predict and can prevent disease. One such use of these techniques is cardiovascular diseases. Heart disease or coronary artery disease (CAD) is one of the major causes of death all over the world. Comprehensive research using single data mining techniques have not resulted in an acceptable accuracy. Further research is being carried out on the effectiveness of hybridizing more than one technique for increasing accuracy in the diagnosis of heart disease. In this article, the authors worked on heart stalog dataset collected from the UCI repository, used the Random Forest algorithm and Feature Selection using rough sets to accurately predict the occurrence of heart disease.

KEYWORDS
Classification, Feature Selection, Heart Disease, Random Forest, Rough Set

1. INTRODUCTION

Data mining plays a significant role in extracting the hidden patterns from datasets of medical domain and these patterns helpful for early detection of heart disease, which is the cause of approximately 12 million deaths occur across worldwide. Coronary heart disease deaths are observed more in the United States than in any other developed and developing countries (Soni, 2011). Coronary heart disease is also known as coronary artery disease (CAD) due to the fact that plaque or a wax like substance gets accumulated and contracts the coronary arteries. These arteries supply blood and oxygen to heart. This, in turn, affects the functioning of the heart and other organs. Atherosclerosis is one of the ailments which occur when a substance like a plaque or fatty material builds up on the artery walls. Though women in their 40s have a lower risk of CAD but when compared to men, their risk factor increases as they grow in age. Some of the diseases/attributes that increase the risk factor in women are but not limited to a) high LDL b) high BP c) diabetes d) smoking e) cholesterol f) obesity (Jabbar, 2015) and along with the major risk factors are Electrocardiographic pattern of left ventricular hypertrophy, Elevated serum cholesterol, and hyper-tension. These factors signify that there is a chance of coronary heart disease (CHD) (Kannel, 2011). CHD is a major cause to increased mortality rate across worldwide (Chaurasia, 2013).

DOI: 10.4018/IJBDAH.2018010101

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Data mining plays a vital role in health sector for prediction of heart disease (Vijiyarani, 2013). Rough Set Theory (RST) was introduced in 1982 as a methodology for data analysis. RST works on the principle of in-discernibility relation i.e. the ability to find the difference between objects, based on their attribute values. Here, we construct upper approximations (UP) and lower approximations (LU). Lower approximation (LU) represents the objects that surely belong to a given response class, whereas upper approximation is those objects that possibly belong to the decision class. An important factor of RST is that it doesn’t require any additional parameter for analysis of data. RST has been used for feature selection, instance selection, classification and regression (Riza, 2014). Feature Selection (FS) is a pre-processing step in machine learning; it helps to categorize the different categories such as relevant, irrelevant and redundant which indirectly increases learning accuracy (Jabbar, 2013). Once the data is categorized, redundant and irrelevant features are removed. Feature selection (FS) is one of the crucial steps in the medical industry for prediction of medical diseases (Jabbar, 2013).

Random Forests is the most popular ensemble technique used for prediction and probability estimation. In Random Forests N-number of decision trees are generated and each decision tree is selected based on a vote, then the decision of the class is considered. It is an excellent method for handling huge amounts of data and missing values (Jabbar, 2016). The paper is organized as follows – Section 2 discusses the earlier work done by researchers and our work in this perspective. Section 3 describes different Machine learning concepts, ensemble technique – Random Forests, Feature selection and Rough Set. Section 4, represents the proposed method for early diagnosis of heart disease using Rough Set and Random Forests. Experimental results are discussed in section 5 and concluding remarks in section 6.

2. RELATED WORK

Various research papers related to the application of data mining techniques on healthcare data for prediction of heart disease have been presented earlier.

(Liu, 2017) uses a hybrid classification method based on RFRS (Relief F and Rough Set) to aid early diagnosis of heart diseases. Proposed system is a hybrid classification system which is a combination of Feature Selection using RFRS and the results of this is applied to ensemble classifier based on C4.5 decision tree technique is proposed. This hybrid classification system helps in early diagnosis of heart diseases. (Jabbar et al., 2012) tested on Andhra Pradesh heart disease data set to generate lazy association rules and Principal Component Analysis (PCA) for feature subset selection and observed the results on both medical and non-medical data. (Manogaran, 2017) proposed wearable health monitoring system which collects the data of patients such as BP, sugar level and heart rate. Here author used apache flume to store the data and from there Amazon SW storage is used for analysis. Electronic Health Records (EHR) is used to build a prediction model by using stochastic gradient and logistic regression technique. This model is applied on sensor data of the patient and predicts the heart disease status of a patient and achieved an accuracy of 81.75%.

(Kavitha, 2016) proposed a new model using an optimization technique PCA (Principal Component Analysis) for feature extraction and Information gain is used for ranking the feature attributes and feature subset selection. (Chitra, 2013) proposed a hybrid intelligent algorithm using data mining classification techniques and compared all the techniques on the same data set for prediction of heart disease. They noticed that Decision tree outperforms than Bayesian Classification (BC) and also concluded that the accuracy of Bayesian Classification (BC) and Decision Tree (DT) can be improved by applying genetic algorithm, this technique helps to reduce the size of the actual data for selecting optimal subset features from original data size. (Xu, 2017) worked on Cleveland heart disease dataset for prediction of heart disease. For feature subset evaluation, he combined CFS and best-first method to improve the accuracy. To build the model author worked on two datasets Cleveland Heart Disease dataset and people’s hospital dataset, the experimental results shows that
Prediction of Heart Diseases Using Data Mining Techniques: Application on Framingham Heart Study
[www.igi-global.com/article/prediction-of-heart-diseases-using-data-mining-techniques/223163?camid=4v1a](www.igi-global.com/article/prediction-of-heart-diseases-using-data-mining-techniques/223163?camid=4v1a)

A Survey on Prediction Using Big Data Analytics
[www.igi-global.com/article/a-survey-on-prediction-using-big-data-analytics/197438?camid=4v1a](www.igi-global.com/article/a-survey-on-prediction-using-big-data-analytics/197438?camid=4v1a)