Heart disease prediction using machine learning techniques: a survey

V.V. Ramalingam*, Ayantan Dandapani, M Karthik Raja

1Department of Computer Science & Engineering, SRM Institute of Science and Technology
*Corresponding Author E-mail: ramalingam.v@ktr.srmuniv.ac.in

Abstract

Heart related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need of reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyze their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers.

Keywords: Cardiovascular Diseases; Support Vector Machines; K-Nearest Neighbour; Naïve Bayes; Decision Tree; Random Forest; Ensemble Models.

1. Introduction

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases.

Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organisation, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality [1]. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or Cardiovascular diseases [2]. Thus, feasible and accurate prediction of heart related diseases is very important.

Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately.

2. Dimensionality Reduction

Dimensionality Reduction involves selecting a mathematical representation such that one can relate the majority of, but not all, the variance within the given data, thereby including only most significant information. The data considered for a task or a problem, may consists of a lot of attributes or dimensions, but not all of these attributes may equally influence the output. A large number of attributes, or features, may affect the computational complexity and may even lead to overfitting which leads to poor results. Thus, Dimensionality Reduction is a very important step considered while building any model. Dimensionality Reduction is generally achieved by two methods - Feature Extraction and Feature Selection.

A. Feature Extraction

In this, a new set of features is derived from the original feature set. Feature extraction involves a transformation of the features. This transformation is often not reversible as few, or maybe many, useful information is lost in the process. In [3] and [4], Principal Component Analysis (PCA) is used for feature extraction. Principal Component Analysis is a popularly used linear transformation algorithm. In the feature space, it finds the directions that maximize variance and finds directions that are mutually orthogonal. It is a global algorithm that gives the best reconstruction.

B. Feature Selection

In this, a subset of original feature set is selected. In [5], key features are selected by CFS (Correlation based Feature Selection) Subset Evaluation combined with Best First Search method to reduce dimensionality. In [6], chi-square statistics test is used to select the most significant features.
3. Algorithms and Techniques Used

A. Naive Bayes

Naive Bayes is a simple but an effective classification technique which is based on the Bayes Theorem. It assumes independence among predictors, i.e., the attributes or features should be not correlated to one another or should not, in anyway, be related to each other. Even if there is dependency, still all these features or attributes independently contribute to the probability and that is why it is called Naive.

\[
P(c | x) = \frac{P(x | c)P(c)}{P(x)}
\]

In [7], Naive Bayes has achieved an accuracy of 84.1584% with the 10 most significant features which are selected using SVM-RFE (Recursive Feature Elimination) and gain ratio algorithms whereas in [8], Naive Bayes has achieved an accuracy of 83.49% when all 13 attributes of the Cleveland dataset[25] are used.

B. Support Vector Machine

Support Vector Machine is an extremely popular supervised machine learning technique (having a pre-defined target variable) which can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall.

\[
P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \cdots \times P(x_n | c) \times P(c)
\]

Shan Xu et al. have used SVM to achieve an accuracy of 98.9% in People’s Hospital dataset [5]. In [9], SVM performs the best with 85.7655% of correctly classified instance and in [10] SVM is used with boosting technique to give an accuracy of 84.81%. HoudaMezrigui et al. have used SVM to attain a f-measure value of 93.5617 [11]. In [12] SVM classifies the pixel variation with an accuracy of 92.1% helping to identify the affected region accurately.

C. K – Nearest Neighbour

In 1951, Hodges et al. introduced a nonparametric technique for pattern classification which is popularly known the K-Nearest Neighbour rule [13]. K-Nearest Neighbour technique is one of the most elementary but very effective classification techniques. It makes no assumptions about the data and is generally be used for classification tasks when there is very less or no prior knowledge about the data distribution. This algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it.

In [10] KNN gives an accuracy of 83.16% when the value of k is equal to 9 while using 10-cross validation technique. In [14] KNN with Ant Colony Optimization performs better than other techniques with an accuracy of 70.26% and the error rates is 0.526. Ridhi Saini et al. have obtained an efficiency of 87.5% [15], which is very good.

D. Decision Tree

Decision tree is a of supervised learning algorithm. This technique is mostly used in classification problems. It performs effortlessly with continuous and categorical attributes. This algorithm divides the population into two or more similar sets based on the most significant predictors. Decision Tree algorithm, first calculates the entropy of each and every attribute. Then the dataset is split with the help of these variables or predictors with maximum information gain or minimum entropy. These two steps are performed recursively with the remaining attributes.

\[
\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i
\]

\[
\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)
\]

Fig. 2: Decision Tree

In [10] decision tree has the worst performance with an accuracy of 77.55% but when decision tree is used with boosting technique it performs better with an accuracy of 82.17%. In [9] decision tree performs very poorly with a correctly classified instance percentage of 42.8954% whereas in [16] also uses the same dataset but used the J48 algorithm for implementing Decision Trees and the accuracy thus obtained is 67.7% which is less but still an improvement on the former. Renu Chauhan et al. have obtained an accuracy of 71.43% [17]. M.A. Jabbar et al. have used alternating decision trees with principle component analysis to obtain an accuracy 92.2% [18]. KamranFarooq et al. have achieved the best results on using decision tree-based classifier combined with forward selection which achieves a weighted accuracy of 78.4684% [19].
E. Random Forest

Random Forest is also a popularly supervised machine learning algorithm. This technique can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, Random Forest technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. This technique is based on the belief that more number of trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality.

![Image of Random Forest](image_url)

Fig. 3: Random Forest

In [5], random forest performs exceptionally well. In Cleveland dataset, random forest has a significantly higher accuracy of 91.6% than all the other methods. In People's Hospital dataset, it achieves an accuracy of 97%. In [20] random forest has achieved an F-measure of 0.86. In [21], random forest is used to predict coronary heart disease and it obtains an accuracy of 97.7%.

F. Ensemble Model

In ensemble modeling two or more related but different analytical models are used and produce their results are combined into a single score. TahiraMahboob et al. [22] have used an ensemble of SVM, KNN and ANN to achieve an accuracy of 94.12%. The Majority vote-based model as demonstrated by Saba Bashir et al. [23] which comprises of Naive Bayes, Decision Tree and Support Vector Machine classifiers, gave an accuracy of 82%, sensitivity of 74% and specificity of 93% for UCI heart disease dataset. In [24] an ensemble model, consisting of Gini Index, SVM and Naive Bayes classifiers, has been proposed which gave an accuracy of 98% in predicting Syncope disease.

4. Conclusion

Based on the above review, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. Alternating decision trees when used with PCA, have performed extremely well but decision trees have performed very poorly in some other cases which could be due to overfitting. Random Forest and Ensemble models have performed very well because they solve the problem of overfitting by employing multiple algorithms (multiple Decision Trees in case of Random Forest). Models based on Naive Bayes classifier were computationally very fast and have also performed well. SVM performed extremely well for most of the cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and overfitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data.

5. Acknowledgment

We sincerely thank the staff of SRM Institute of Science and Technology, that have provided their immense support and guidance throughout the project.

Reference

[17] Renu Chauhan, Pinki Bajaj, Kavita Choudhary and Yogita Gigras et al. “Framework to Predict Health Diseases Using Attribute


