EasyChair Preprint

№ 5955

Performance Evaluation of Classification Methods for Predicting Heart Disease

S. Poonguzhali, P. Sujatha, P. Sripriya, V. Deepa and K. Mahalakshmi

June 29, 2021

PERFORMANCE EVALUATION OF CLASSIFICATION METHODS FOR PREDICTING HEART DISEASE

S.Poonguzhali[1],P.Sujatha [2],P.Sripriya[3], V.Deepa [4], K. Mahalakshmi[5]

[1]*Assistant Professor, School of Computing Sciences,Vels Institute of Science, Technology & Advanced Studies,Chennai, India.*
poonguzhali.research@gmail.com

[2]*Professor,School of Computing Sciences,Vels Institute of Science, Technology & Advanced Studies,Chennai, India.*
sujathap.research@gmail.com

[3]*Professor, School of Computing Sciences,Vels Institute of Science, Technology & Advanced Studies,Chennai, India.*
sripriya.phd@gmail.com

[4]*Research Scholar, School of Computing Sciences,Vels Institute of Science, Technology & Advanced Studies,Chennai, India.*
deepamanagunan@gmail.com

[5]*Research Scholar, School of Computing Sciences,Vels Institute of Science, Technology & Advanced Studies,Chennai, India.*
rkmahalakshmi2016@gmail.com

Abstract

In this modern era industries are generating large amount of data. The most important asset of any kind of organization is data.Digitalization reduces human effort and makes data easily accessible. With the increased access to the data, all industries are using data for making decisions. Health care industries are one among the top in generating data. To mine the complex data advance algorithms and techniques are needed. The data extraction techniques are used to convert these raw facts asmeaningful information. One of the popular data extraction techniques is data mining and machine learning. With the patient data Health care industries arenow focusing on optimizing the efficiency and quality of the treatment using various data analytical tools. Data mining and Machine learning has been used by many industries, howeverthey are the proven methodology in health care. Non communicable disease such a heart disease, diabetics and cancer are major reason for the death around the world. Heart disease is one among the top reason for death. In this research paper we have implemented popular data mining algorithms viz., Support vector machineand decision tree with the relevantheart disease data set using Python. The performance of the algorithms is evaluated using various evaluation metrics.

Keywords

Data Mining, Machine learning, Health care, Decision Tree, support vector machine, heart disease.

I. INTRODUCTION

In the modern world, data is used for predicting the occurrence of the future events which leads to better conclusion. Data mining and Machine learning techniques are used for predicting the results by analyzing the data. They are becoming the hotspot of research works especially in healthcare industries. Heart ailment is one of the top reasons for the death around the world. The major reason for heart disease is unhealthy life style. Heart disease is curable if it is predicted in the early stage. Data Mining is the proven technology in health care industry in treatment process. Data mining algorithms build predictive model which provides reliable predictions and improve treatment process. The data mining is used to discover the useful and understandable patterns or relationships by analyzing the large sets of data. These data patterns are used for prediction. Data mining algorithms are used by health care professionals to decide about patient health form raw facts. Major and popular techniques of data mining are association, classification, clustering, prediction, sequential patterns and regression etc. Popular data mining algorithms to analyze the data are as follows:

- Neural Networks
- Bayesian Networks
- Decision Trees
- Support Vector Machines
- K-means
- Apriori algorithm
- K Nearest Neighbor

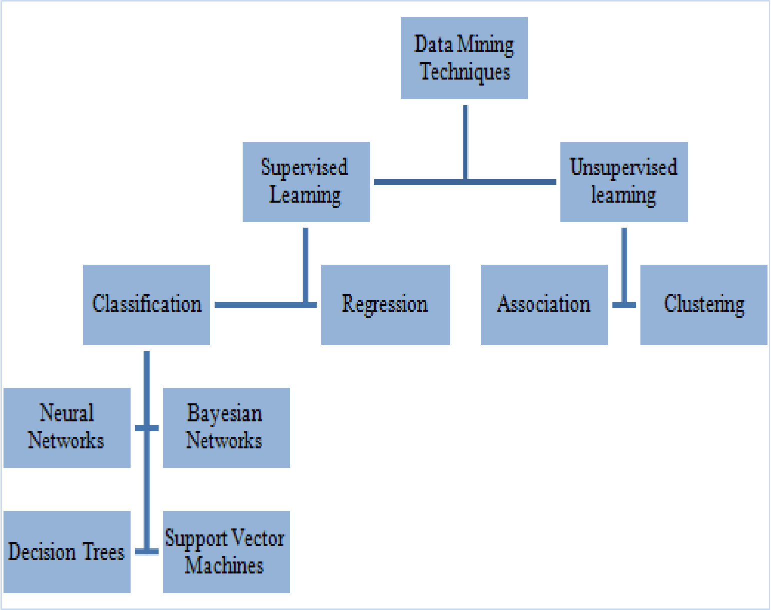

Figure 1- Classificationof Data Mining Algorithms

This research work implements SVMand decision tree withset of data relevant to heart disease using Python to predict the heart disease. The performance of those two algorithms is evaluated based on the metrics such as accuracy, precision, recall and F1Score.

II. REVIEW OF LITERATURE

Existing research works for predicting Heart Disease using Data mining algorithms are discussed in this Section.

The hybrid associative classification [1] is proposed to predict heart diseases using KNN algorithm on weka environment. The study carried out on 14 attributes of Cleveland heart disease database. Hybrid classification associate rules obtained a prediction accuracy of 99.19%.

Thiswork [2]includes two modules viz., classifier module and prediction module. Data is trained through KNN in classifier module. Using ID3 algorithm, the menace of heart disease found. This model achieved accuracy of 40.3% using the basic attributes. Prediction accuracy increased up to 80.6% by adding more number of attributes. The proposed system found that the accuracy is high with more number of attributes.

Random forest with 14 attributes of Cleveland database suggested for prediction of patients with heart disease using Apache Spark and Hadoop HDFS [3]. The proposed model achieved the accuracy of 98% with 600 dataset records. The experimental result shows that, the accuracy raised from 88% to 98%.

The data mining Classification techniques such as decision tree and neural network suggested to predict heart disease using MATLAB R20 [4]. The drawback of Back propagation algorithm is local minima. Efficient Optimizing Genetic algorithm is used in this model to solve local minima problem. The author concluded that,for non linear data, Neural Network classifier achieves more accuracy.

III. METHODLOGY

Popular classification algorithms such as Support vector machineand decision tree are implemented with heart disease dataset and their performance are analyzed based on important metrics. The algorithms are implemented using Python 3.7.In the following Figure 2, the entire flow is given.

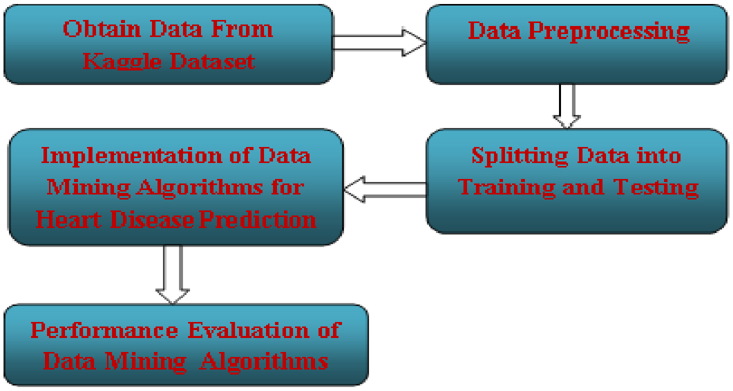
Obtain Data From Kaggle Dataset
Data Preprocessing
Splitting Data into Training and Testing
Implementation of Data Mining Algorithms for Heart Disease Prediction
Performance Evaluation of Data Mining Algorithms

Figure 2 – Schematic Diagram

A. Data Set Description

This research work retrieved the set of data from kaggle database[5]. This dataset consists of 14 attributes with 303 records. Table 1 describesa few of the attributes of the data set.

TABLE 1. LIST OF ATTRIBUTES OF SAMPLE DATA SET

S.No	Attributes	Description
1.	Age of a Person	Age as continuous value
2.	Gender	1 - Male 0 - Female
3.	Chest Pain	Type of chest pain like anginal or non-anginal pain
4.	RBP	Resting Blood Pressure
5.	SC	Serum Cholesterol
6.	FBS	Fasting Blood Sugar
7.	ECG	Electrocardiogram
8.	MHR	Maximum Heart Rate
9.	Exercise induced angina	Blood supply during excercise
10.	OP	Old peak
11.	SE	Slope
12.	Nca	Number of vessels
13.	Thl	Thallium Stress Test
14.	Tar	1 - Prediction of Heart Disease 0 – No Heart Disease

B. Data Preprocessing

The high quality data used to build better model for prediction. Data preprocessing is the process of increasing the quality of the data. It is the process of transforming raw data into meaningfulinformation. Data is often incomplete, inconsistent and also contain noise or wrong data. All these factors reduce the quality of the results. Heart disease data set from kaggle is also analyzed for inconsistency. As the data set is very small in size, inconsistencies are not found in the heart disease dataset.

C. Implementation of Data mining Algorithms

In this analysis, heart disease is predicted by employing popular algorithms such as SVM with polynomial kernel and decision tree classification algorithms. SVM is mainly used for classification, regression and outlier detection. It is used to solve linear and non-linear problems. The SVM algorithm creates hyperplane (line) that separates the data into classes. SVM takes data as input and outputs a hyperplane that separates the class. In real world it is very difficult and takes lot of time to find the perfect class for large amount of training data. Tuning parameters can be used to regularizing parameter. SVM with tuning parameters can be used to increase the accuracy of SVM. Major part of the SVM is kernel trick.

Decision tree is used forperforming classification and regression. DT is tree representation in which the internal node represents attributes (features) and leaf node represents class labels. The DT algorithm will give set of rules. These rules can be used to predict the test dataset. The main complexityof Decision Tree is recognizing the attributes of the root node at each level. Some of the popular measures for attribute selection are Information Gain and Gini Index[7].

SVM with polynomial kernel and decision tree are implemented in kaggle heart disease dataset using python. Dataset consists of 303 records and 14 attributes. All 14 attributes are considered for predicting heart disease. We split the dataset in the ratio of 80:20 to create training dataset and testing dataset. After dividing the data we can train the algorithms and the trained algorithm is used to predict the heart disease.

IV. RESULTS &DISCUSSION

Python is used for implementing the classification algorithms. The dataset consists of 303 records. The 303 records are divided as training and testing dataset in the ratio of 80:20 respectively. Out of 303 records 242 records for training the algorithm and remaining 61 records are used for testing the algorithm.SVM and decision tree algorithms were employed to predict the heart disease. The performance of the algorithms is evaluated for checking the results. Confusion matrix, accuracy, f1score, precision and recall are used in the research paper to analysis the efficiency of the algorithm

Confusion matrix describes the performance of the algorithm on a collection of test data. The following figure 3& figure 4 shows the confusion matrix obtained by support vector machine and decision tree classification techniques.

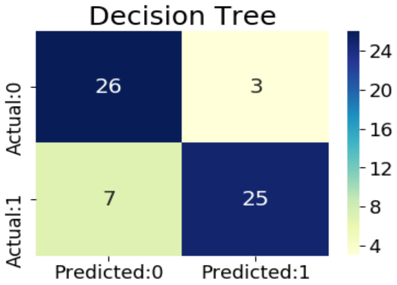

Figure 3 - Confusion matrix of decision tree

Figure 3 describes that, out of those 61 records, the decision tree classifier predicted “Yes“28 times and predicted “No” 33 times. Twenty Six people were correctly predicted as “No” and really they do not have the heart disease. Twenty Five people were correctly predicted as "Yes" and also they have heart disease in fact. Seven people were incorrectly predicted as "No" even though they have heart problem. Three people were incorrectly predicted as "Yes" when they do not have heart problem.

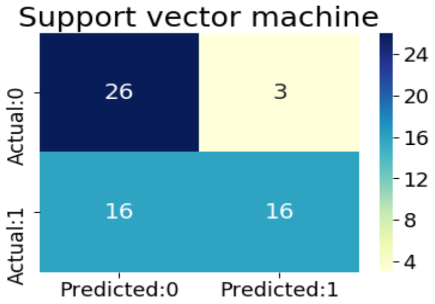

Figure 4 Confusion matrix of SVM

Figure 4 describes that, from 61 testing records, the decision tree classifier predicted 19 times "Yes"and predicted "No" 42 times. Twenty Six people were correctly predicted as "No" and they do not have the heart diseasein reality. Sixteen people were correctly predicted as "Yes" and also they have heart disease actually. Sixteen people were incorrectly predicted as "No" even if they have heart problem. Three people were incorrectly predicted as "Yes" though they do not have heart problem.

The following table 2& Figure 4 evident that decision tree predicts the disease with maximum accuracy of 83.6% compared to SVM that provides the accuracy of 68.8%.

Table 2: Performance Evaluation of Classification Algorithms

Classification Algorithms	**Accuracy**	**F1 Score**	**Precision**	**Recall**
Decision Tree	0.836066	0.833333	0.892857	0.78125
SVM	0.688525	0.627451	0.842105	0.50000

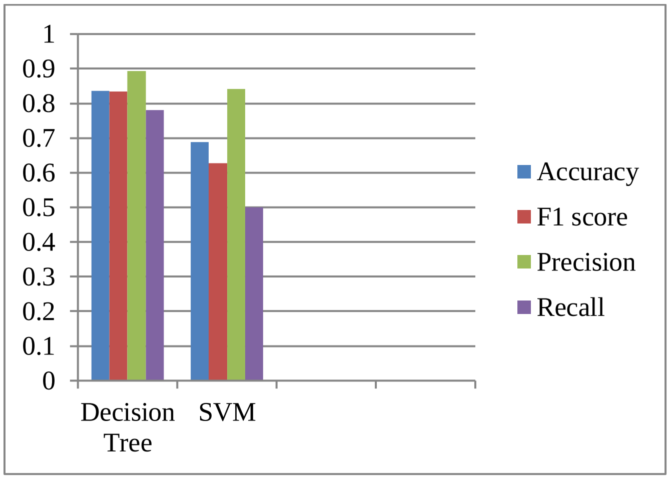

Figure 4. Performance Evaluation of Classification Algorithms

V. CONCLUSION

Heart disease covers any kind of disorder of the heart. It emerged as foremostsource of death. Heart disease prediction and prevention is necessary. Data mining algorithms helps in predicting the heart diseases, and predictions made by algorithms are quite accurate. Decision Tree and Support vector machine are employed on heart disease dataset from kaggle website. The classification algorithms are implemented using python. The decision tree algorithm achieved the maximum accuracy of 83.6%. Comparing with decision tree,SVM produces lowest accuracy of 68.85%.

From the experimental results we can conclude that decision tree algorithm performs better when compared to SVM.Our future of work is to combine classification algorithm to develop hybrid model to predict the heart ailment.

REFERENCES

[1] Jagdeep Singh, Amit Kamra and Harbhag Singh, “Prediction of Heart Diseases using Associative Classification,” IEEE 5th International conference on Wireless Networks & embedded systems (WECON) 2016.

[2] Theresa Princy.R and J.Thomas, “Human Heart Disease Prediction System using Data Mining Techniques,” International conference on Circuit, Power and Computing Technologies [ICCPCT], 2016.

[3] Rashmi G Saboji and Prem Kumar Ramesh, “ A Scalable Solution for heart disease prediction using classification mining technique,” International Conference on Energy, Communication, Data Analytics and Soft Computing(ICECDS), 2017.

[4] Ankita Dewan and Meghna Sharma, “Prediction of Heart Disease using a Hybrid Technique in Data Mining Classification,” 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015.

[5] https://www.kaggle.com/ronitf/heart-disease-uci?select=heart.csv

[6] https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/

[7] https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/