© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

A Comparative study of Data Classification Techniques for Coronary Artery Disease

Vinod Babu P[1], B S V Prasad[2], Ch Seshadri Rao[3]

[1]Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India

[2]Department of CSE, Usha Rama College of Engineering and Technology, Telaprolu, Andhra Pradesh, India

[3]Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh, India

ABSTRACT

Now a day, the heart attack is one of the deadliest diseases patients face. This disease attacks a person so instantly that it hardly gets any time to be treated with. If you're 35-70 you should go for a heart health check-up to assess your risk of having a heart attack in the next 10 years. Identifying and managing a condition such as high blood pressure or high cholesterol or hypertension could help lower your chances of having a risk in the heart attack in the future. Data mining techniques can help in predicting the risk of heart attack of the person for the next ten years. Data classification algorithms such as Support Vector Machine (SVM), Logistic Regression, and Decision tree can help in classifying the given patients data can help in the prediction of the heart attack in next ten years. Results show the accuracy percentage of the prediction whether a person can have the heart attack in next ten year or not based on their medical data.

Keywords: Data Mining, Heart Attack, Classification, Support Vector Machine (SVM)

I. INTRODUCTION

Heart disease has been the leading cause of death worldwide since 1921. Coronary heart disease (CHD) normally happens when cholesterol accumulates on the artery walls, creating plaques. The arteries narrowing reduce blood flow to the heart. Sometimes, a clot can obstruct the flow of blood to the heart muscle. It is necessary to develop a CHD prediction model using data mining.

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease.

The organization of this document is as follows. In Section 2 the various methods are discussed like Logistic regression, Support Vector Machine (SVM), Extreme Gradient Boosting, The Naïve Bayesian and K nearest neighbour classifier. In Section 3 presented a simple pre-processing technique. The missing data, i.e., NA's are replaced with Mean of the attribute and the median of the attribute. The accuracy of the

algorithms are compared and found the logistic regression algorithm accuracy is 86%.

II. METHODS AND MATERIAL

The main aim of the developers is to develop a prediction model that can predict heart disease cases based on measurements taken from transthoracic echocardiography examination, and they have used the Knowledge Discovery in Database (KDD) methodology. Since the knowledge gained from the different experts are a high-level description of the problem from the medical point of view, a literature review was carried out and relevant works related to data mining and heart disease have been reviewed to have more knowledge about the domain. Furthermore, a real-time observation of the system was performed to understand the business process of the hospital. A key subgoal in this step is the determination of data mining goals and their success criteria. The goals are obtained by translating medical goals into data mining goals. To predict heart disease using classification techniques, the researchers have used three different supervised machine learning algorithms i.e., Decision Tree Classification, Bayesian Classifier, and Neural Network.

The researchers have introduced pincer search algorithm to discover the maximum frequent item set. It also reduces the number of times the database is scanned. The researchers also expressed about frequent itemset mining without the generation of conditional frequent pattern tress. The desired association rules are also discovered from the frequent item set. The other researchers also developed a predicting system to predict the heart disease. K-Means clustering technique is used to distinguish the risky and non-risky factors to categorize. The researchers developed MFFP-Tree Fuzzy Mining Algorithm to find out the linguistic frequent itemsets.

Methodology

The main goal of the prediction methodology is to design a model that can infer characteristic of predicted class fro.....m the combination of other data. The task of data mining in this experiment is to build models for prediction of the class based on selected attributes of the selected data set. In this experiment the following algorithms are implemented: Liner regression, Logistic Regression, Support Vector Machine, Naive Bayesian, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbour algorithm to classify and develop a model to diagnose heart attacks in the patient data set.

Linear Regression

Linear regression is a method used when the relationship between the variables and used for predicting the value of a dependent variable from an independent variable that can be described with a linear model. The best-fitting straight line through the points is determined by linear regression. The best-fitting line is called a regression line. The error of prediction represents the vertical lines from the points to the regression line. Mostly, the criteria used to determine the best-fitting line is the line that minimizes the sum of the squared errors of prediction. For non-linear relationships model, linear regression is often inappropriate. The outcome or dependent variable is continuous in linear regression. It can have any one of an infinite number of possible values.

Linear regression gives an equation, which is of the form, is as shown below

$Y = MX + C$, means equation with degree 1

Where m is the intercept

Logistic Regression

Logistic regression is a method for analyzing a dataset in which the output is determined by one or more independent variables. The outcome variable a

dichotomous variable i.e., the variable in which there are only two possible outcomes. Logistic regression is used to find the best fitting model in order to describe the relationship between the dependent variable i.e. , the outcome variable and a set of independent variables to arrive at the solution. Logistic regression uses maximum likelihood method. In this technique, the outcome or dependent variable is a categorical, i.e., discrete. In Logistic regression, the coefficients of a formula are generated in order to predict logic transformations of the probability of the presence of the characteristic of interest:

$$\text{Logit}(p) = b0+b1X1+b2X2+.....+bkXk.$$

Where p is the probability of the presence of the characteristic of interest

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm. SVM is used to analyze the data for classification and regression analysis. SVM is mostly used for classification problems. In SVM algorithm, each data item is plotted as a point in n-dimensional space, where n is the number of features you have and the value of each feature being the value of a particular coordinate. The classification of the dataset is done by finding the hyperplane that differentiates the two classes.

- **Linear Support Vector Machine**

Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time, which scales linearly with the size of the training dataset.

- **Non-Linear Support Vector Machine**

SVM has a technique called the kernel trick. These are functions which take low dimensional input space and transform it to a higher dimensional space i.e. it converts not separable problem to separable problem, these functions are called kernels. It is most useful in non-linear separation problem.

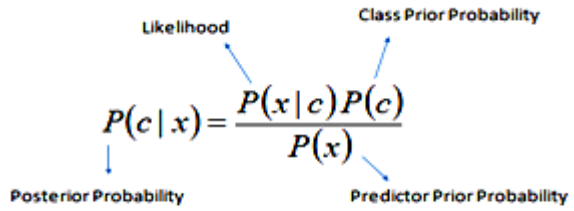
Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) algorithm is used in the implementation of gradient boosted decision trees to increase speed and performance. XGBoost is a supervised learning algorithm, where the training data having multiple features are used to predict a target variable or outcome variable. The implementation of the algorithm is to increase the efficiency of computing time and memory resources. The design goal of XGBoost was to make the best use of available resources to train the model. Gradient boosting uses an approach to predict the residuals or errors of prior models based on the new models that are created and then added together to make the final prediction. This algorithm uses a gradient descent algorithm to minimize the loss when adding new models; hence it is called gradient boosting. Extreme Gradient Boosting algorithm is used for classification and regression predictive modeling problems.

Naïve Bayesian

The Naïve Bayesian Classification represents a supervised learning method for classification. The Naive Bayesian classifier is based on Bays theorem. Naive Bayesian is applied with the assumption of independence between the inputs. Given a set of variables, X = {x1, x2, x3..., x_n}, we want to construct the posterior probability for the event C_j among a set of possible outcomes C = {c1, c2, c3..., c_n}. Technically we can say that X is the predictors and C is the set of categorical levels present in the dependent variable.

- P(*c*/*x*) is the posterior probability of *class* given *predictor* (*attribute*).
- *P*(*c*) is the prior probability of *class*.
- *P*(*x*/*c*) is the likelihood which is the probability of *predictor* given *class*.
- *P*(*x*) is the prior probability of *predictor*.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

KNN (K-Nearest Neighbour)

K nearest neighbour is a simple algorithm used for classification and regression problems. This algorithm stores all available cases and classifies new cases based on a distance function. A case is classified by a majority vote of its neighbours. The most popular distance functions used in KNN is Euclidean's distance function. The other distance functions are Manhattan distance, Hamming distance, Minkowski distance. Based on the properties of the data we can choose the best distance metric.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Patients Dataset

The work is carried out on the Framingham heart study data set. Initially, the dataset contains 40 attributes but for this experiment, we needed 16 attributes. For this project, dataset consists of 16 attributes and 4240 tuples. Initially, the dataset is pre-processed in order to make it suitable for the mining process. Fig 2.1 describes the attributes of the dataset.

Attributes of the dataset

S no	Attribute	Description
1	Male	Gender(men=1 , women=0)
2	Education	Some high school (1), high school/GED (2), some college / vocational school (3), college (4)
3	Age	Age of the person
4	Current smoker	Yes=1 , no=0
5	Cig per day	
6	BP Meds	On blood pressure medication at time of first examination (Yes=1, no=0)
7	Prevalent stroke	Previously had a stroke (Yes=1, no=0)
8	Prevalent Hypertensive	Currently hypertensive (Yes=1, no=0)
9	Diabetes	Currently has diabetes(Yes=1, no=0)
10	Total cholesterol	Total cholesterol (mg/dL)
11	Systolic BP	Systolic blood pressure
12	Diastolic BP	Diastolic blood pressure
13	BMI	Body Mass Index, weight (kg)/height (m)2
14	Heart rate	Heart rate (beats/minute)
15	Glucose	Blood glucose level (mg/dL)
16	Ten years CHD	ten year Coronary Heart Disease (CHD) (Yes=1, no=0)

III. RESULTS AND DISCUSSION

In this experiment, during the pre-processing of data first, we replaced the NA's with Mean of the attribute and next time we replaced the NA's with the median of the attribute. The confusion matrix obtained for different classification techniques when NA's are replaced with Mean of their attribute are as follow:

Linear Regression

	FALSE	*TRUE*
0	*3591*	*620*
1	*5*	*24*

Logistic Regression

1. Logistic Regression with the random splitting of the dataset into train and test set

	FALSE	*TRUE*
0	*1254*	*5*
1	*210*	*15*

2. Logistic Regression with the linear splitting of the dataset into train and test sets

	FALSE	*TRUE*
0	*1071*	*11*
1	*172*	*19*

Support Vector Machine (SVM)

1. SVM - linear

	FALSE	*TRUE*
0	*1259*	*225*
1	*0*	*0*

2. SVM - radial

	FALSE	*TRUE*
0	*1257*	*224*
1	*2*	*1*

Extreme Gradient Boosting (XGBOOST)

1. XGBoost with the random splitting of the dataset 1into train and test sets

	FALSE	*TRUE*
0	*1082*	*7*
1	*171*	*12*

2. XGBoost with the index splitting of the dataset into train and test sets

	FALSE	*TRUE*
0	*1059*	*173*
1	*26*	*14*

Naïve Bayesian

	FALSE	*TRUE*
0	*1024*	*149*
1	*54*	*45*

K-Nearest Neighbor

	FALSE	*TRUE*
0	*1018*	*155*
1	*56*	*44*

The confusion matrix obtained for different classification techniques when NA's are replaced with Median of their attribute values are as follow:

Linear Regression

	FALSE	*TRUE*
0	*3590*	*620*
1	*6*	*24*

Logistic Regression

1. Logistic Regression with the random splitting of the dataset into train and test sets

	FALSE	*TRUE*
0	*1252*	*7*
1	*211*	*14*

2. Logistic Regression with the linear splitting of the dataset into train and test sets

Support Vector Machine (SVM)

	FALSE	*TRUE*
0	*1221*	*196*
1	*38*	*29*

1. SVM - linear

	FALSE	*TRUE*
0	*1259*	*225*
1	*0*	*0*

2. **SVM - radial**

	FALSE	*TRUE*
0	*1257*	*224*
1	*2*	*1*

Extreme Gradiant Boosting (XGBOOST)

1. XGBoost with the random splitting of the dataset into train and test sets

	FALSE	*TRUE*
0	*1220*	*197*
1	*49*	*28*

2. XGBoost with the index splitting of the dataset into train and test sets

	FALSE	*TRUE*
0	*1045*	*190*
1	*23*	*14*

Naïve Bayesian

	FALSE	*TRUE*
0	*1005*	*158*
1	*71*	*38*

K-Nearest Neighbor

	FALSE	*TRUE*
0	*1014*	*196*
1	*46*	*16*

IV.CONCLUSION

In this, we have implemented different data mining classification algorithms to predict whether a person can have the heart attack in next ten years or not using Framingham heart attack dataset. The algorithms are Support Vector Machine, XGBoost, Linear Regression, Logistic Regression, KNN, Naïve Bayesian and we compare the accuracies of the algorithms .We observed that when NA's are replaced by Mean in data cleaning process ,the accuracy of different algorithms when is mostly greater or sometimes equal to the accuracy of different algorithms when NA's are replaced by Median in data cleaning process. For our dataset the Logistic regression with indexing split when NA's are replaced by Mean in data cleaning process has the more accuracy compared to all others algorithms that we implemented, the accuracy of the logistic algorithm is 86%.

V. REFERENCES

1. Usha Rani G, Vijaya Prakash R, Govardhan A. Mining Multilevel Association Rule Using Pincer Search Algorithm. International Journal of Scientific Research 2013;
2. Meera Narvekar, Shafaque Fatma Syed. An Optimized Algorithm for Association Rule Mining using FP Tree. International Conference on Advanced Computing Technologies and Applications 2015; 45:101-110.

3. Alagugowri S, Christopher T. Enhanced Heart Disease Analysis and Prediction System [EHDAPS] Using Data Mining. International Journal of Emerging Trends in Science and Technology 2014; 1:1555-1560.
4. Tzung-Pei Hong, Chun-Wei Lin, Tsung-Ching Lin. The MFFP-Tree Fuzzy Mining Algorithm to Discover Complete Linguistic Frequent Itemsets. International Journal of Computational Intelligence 2014; 30:145-166.
5. Data classification using Support Vector Machine by Durgesh K. Srivastava and Lekha Bhambhu.
6. XGBoost : A scalable Tree Boosting System by Tianqi Chen and Carlos Guestrin.
7. Support Vector Machine by Vikram Aditya Jakkula.
8. Support Vector Machines for Data Mining by Robert Burbidge and Bernard Buxton.