Heart Disease Prediction System Using Data Mining Techniques

Prof. Madhuri Gawali¹, Nikita Shirwalkar², Anupama Kalshetti³
¹,²,³Dept. of Information Technology
¹,²,³Jayawantrao Sawant College of Engineering
Hadapsar Pune, India

July 7, 2018

Abstract

In today's world large volume of material is generated by utilizing patients medical history. Not all systems use the available scientific material for inclination purpose. The proposed system work as a decision support system. Hence, we are using data mining method that shows the consequence depending on patients scientific history. Use of features such as blood pressure, gender, cholesterol, etc. is done to estimate the possibility of heart disease for a person. The possibility of heart disease is computed using data mining techniques Naive Bayes and Improved K-Means.

Keywords: Naive Bayes, improved k-means, Data mining, Heart Disease, Prediction, Classification.

1 Introduction

Heart disease is that the biggest reason behind death nowadays. Blood pressure, cholesterol, pulse rate are the major reason for the
heart disease. The heart is an operating system of our human body. If the function of heart is not done properly it affect other body part also. There are some risks factors of heart disease such as Family history, High blood pressure, Cholesterol, Age, Poor diet, Smoking. When blood vessels are overstretched, the risk levels of the blood vessels are increased. This leads to the blood pressure. Blood pressure is typically measured in terms of systolic and diastolic. Systolic shows the pressure in the arteries when the heart muscle contracts. Diastolic indicates the pressure in the arteries when the heart muscle is in resting state. The level of fats increased in the blood are causes the heart disease. The lipids are in the arteries hence the arteries become narrow and blood flow is also become slow. Age is the non-modifiable risk factor which also a reason for heart disease. Smoking is the reason 35% of the death of heart diseases. Because it limits the oxygen level in the blood then it damage and tighten the blood vessels. Different data mining techniques such as Nave Bayes, KNN algorithm, Decision tree, Neural Network are used to predict the risk of heart disease [1]. The Nave Bayes method is used to predict the heart disease through probability. Improved k means is used to provide the classified report for the heart disease. In many hospitals, more experience doctors are there. They know how analyze the patient within a given period. But in some hospitals new doctors dont have more experience. They cant take quickly decision for e.g. heart disease prediction so they use this system for easy and quicker decision. In medical college, students who study about how to diagnose heart i.e. PHD students. So it easy to learn using this heart disease system for those students.

2 Proposed System Description

User download the data set (UCI database) [6], while running the application this data set load into our system. Transform our data set into one form to another form and make clusters from that data set using improved k-means algorithm this is known as discretization. After discretization pass the new data for training is done through naive Bayes algorithm. Trained data becomes trained model.
In training predict initial probability whether person has heart disease or not. Training model we can predict in which level of heart disease person has by calculate generate ratio probability array i.e. normal, stage 1, stage 2, stage 3.

When person enter their details using 13 different parameter such as age, gender, Thai, blood pressure, cholesterol, etc. after person submit their details it goes to one format that is transform data and discretize the data same as done in training this is known feature table. We are giving feature table as a input to the trained model and then predict in which level of heart disease person has i.e. normal, stage 1, stage 2 or stage 3[3].

Proposed System

![Proposed System Diagram]

Fig.(1) Proposed System

3 Related Work

1. Monika Gandhi, S. Singh etc.al described as follow, Paper describe details about various techniques of data pro-
cessing for predict heart problem (ANN,NB, Decision tree algorithm)[1].

2. Theresa Princy, J. Thomas, etc.al described as follow. Paper gives the survey about data mining classification techniques depends on sex, age, cholesterol, pulse rate, bp by taking Naive Bayes, KNN, Algorithm of decision tree, and ANN[6].

3. Vishal Jadhav, Devendra Ratnaparakhi, Tusdhar Mahajan etc.al described as follow, this paper gives us details about Heart Disease Prediction System which is totally web application. It fetches the data from stored database and compare them with stored dataset [10].

4. S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita etc.al as follow, this paper for perfect analysis of heart disease, output of each algorithm is combined. After combined output it will be compared. Different algorithms are used such as Decision tree, ANN, Nave Bayes and SVM algorithms[11].

4 DATA MINING TECHNIQUES

A. Naive Bayes

Naive Bayes Classifier is based on Bayes theorem. It has high accuracy and efficiency when applied to large database. Naive Bayes Classifier assumes that the effect of each attribute on a given class is independent from the other attributes.

Bayes theorem,

\[ P(C_i \mid X) = \frac{P(x \mid C_i)P(C_i)}{P(X)} \quad (1) \]

The first term represent contribution probability for each class. The next term based on the fraction of each class in the training data. The following equation is used to estimate contribution for text variable:

\[ P(X \mid C_i) = \pi_{k=1}^nP(x_k \mid C_i) \quad (2) \]
If attribute contains a continuous variable, use probability density rather than a probability. The following equation is normal distribution probability density function:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

(3)

B. Improved k-means Algorithm

According the K-means algorithm, the centroid $m_i$ of cluster $i$ could be calculated through formula:

$$m_i = \frac{1}{n} \sum_{j=1}^{n} P_j$$

(4)

where $p_j$ is the position of object $j$ in cluster $C_i$. There are $n$ objects in cluster $C_i$.

$$P = (m_1, m_2..., m_k)$$

(5)

Where $P$ is represented as spacial vector and $k$ is the clustering number.

5 Results

Fig(2) Analysis of Mining algorithms
6 Conclusion

The main motivation of this paper is to provide an insight about detecting heart disease risk rate using data mining techniques. Various Data mining techniques and classifiers are discussed in many studies which are used for efficient and efficacious heart disease. As per the analysis mode, many authors use various technologies and different number of attributes for their study. Hence, different technologies give different precision depending on a number of attributes considered. Using Naive Bayes and Improved K-means algorithm the risk rate of heart disease was detected and accuracy level also provided for different number of attributes.

References


[8] Mrs.G.Subbalakshmi (M.Tech), Kakinada Institute of Engineering & Technology (Affiliated to JNTU-Kakinada), Yanam Road, Korangi-533461, E.G.Dist., A.P., India.


