Intelligent Data Mining System for Heart Disease Prediction

Dr. D. Hevin Rajesh
Associate Professor, Department of Information Technology
St. Xavier's Catholic College of Engineering

***Abstract*—** The heart disease prediction system is an end user support and online application work. A web-based application that allows the user to get instant guidance on their heart disease through an online intelligent system. The application allows the user to share their heart related issues. It then processes user specific details to check for various illness based on some intelligent data mining techniques. Based on the result, the application automatically shows the specific guidance to the user.

***Keywords*—**Intelligent Systems, Prediction, Mining

I. INTRODUCTION

Recently World Health Organization (WHO) conducted a survey which shows approximately 17.3 million deaths globally are due to Cardio Vascular Diseases (CVD), heart attacks and strokes (1). The deaths due to heart disease in countries are due to exertion, work overload, mental stress and so on. Treatment and Diagnosis is complicated and is an important task that needs to be executed accurately and efficiently. The diagnosis is often based on doctor's experience and knowledge. This leads in some cases as unwanted outcomes and excessive medical costs of treatments for patients. Therefore, a medical diagnosis system is designed that takes advantage of collected database and decision from the previous records. Some hospitals have decision support systems, but they are limited.

In health industries, data mining plays a significant task for predicting diseases. Numeral number of tests must be requisite from the patient for detecting a disease. However, using data mining technique can reduce the number of tests that is required. Cardiovascular disease is the principal source of deaths widespread and the prediction of Heart Disease is significant at an untimely phase. In order to reduce number of deaths due to heart diseases there has to be a quick and efficient detection technique. Doctors as well as health care expert have their own experience in the bases of which they predict about particular heart disease of the patient. The healthcare industry produces huge amount of data but it is not effective and efficient decision making.

II. SUBJECTS AND METHODS

A. Data Collection

The Data set used is obtained from Data mining repository of California University, Irvine (UCI). Data set from Cleveland, Hungary, Switzerland, long beach set are collected. Cleveland, Hungary, Switzerland and long beach data set contains 76 attributes totally. But 14 attributes which are basically proven to be important is indicated in table1. Among all those Cleveland data set is the most commonly used data set; it has fewer missing attributes than others which helps in better result. Some sample of data set collected from the UCI repository.

B. Data Mining

Data mining techniques such as Classification, Clustering (2) and many more are used in extracting knowledge from database. Medical data is mined by using the techniques mentioned above and the diagnosis is carried out which is indicated in table2. Practical use of Data mining techniques in medical data (3) is explained below:

Table1. Data set Attributes

No	Name	Descriptions
1	Age	Age in Years
2	Sex	1=Male, 0= Female
3	CP	Chest pain type (1 = typical angina, 2=atypical angina, 3 = non- angina pain, 4 = asymptomatic).
4	Trestbps	Resting blood sugar (in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (1=true, 0=false)
7	Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T Wave abnormality, 2 = left ventricular hypertrophy)
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1=up sloping, 2=flat, 3= down sloping)
12	Ca	Number of major vessels colored by Fluoroscopy
13	Thal	3= normal, 6=fixed defect, 7= reversible defect
14	Num	Class (0=healthy, 1=have heart disease)

Table2. Mining Medical Data

Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
58	0	1	150	283	1	2	162	0	1	1	0	3	0
58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
58	1	3	132	224	0	2	173	0	3.2	1	2	7	3
60	1	4	130	206	0	2	132	1	2.4	2	2	7	4
50	0	3	120	219	0	0	158	0	1.6	2	0	3	0

C. Classification

Classification is done based on supervised machine learning Algorithm. K-means, Decision List Algorithm, Naïve Bayes, performance is based on accuracy and the time taken to build the model. Naïve bayes algorithm (4) commonly used and better from all since it takes only some to calculate the accuracy than other algorithm used and also it lead to lower error rates. Naïve Bayes algorithm gives 52.23% of accurate result (5). Table3 below shows the performance study of the algorithm.

Naïve Bayes Classification

A conditional probability is of some conclusion (6, 7), C, given some observation, E, where there is a dependence relationship between C and E. This probability is denoted as P(C |E) where:

$$P(C|E) = P(E|C)P(C) / P(E)$$

Naive Bayes or Bayes‟ Rule acts as the basis for many machine- learning and data mining methods. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.

Table3. Performance Study

Algorithms used	Accuracy	Time taken
Naive Bayes	52.33%	609ms
Decision List	52%	719ms
KNN	45.67%	1000ms

D. K-means Clustering

Given a set of observations (x1, x2...xn), where each observation is a d-dimensional real vector, k-means clustering focuses to partition the n

observations into k-sets (k ≤ n) S = {S1, S2... Sk} so as to minimize the within-cluster sum of squares (WCSS): $J=|| X_i-C_j||^2$

The algorithm is composed of the following steps:

1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

2) Assign each object to the group that has the closest centroids.

3) When all objects have been assigned, recalculate the positions of the K centroids.

4) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

III. RESULTS AND DISCUSSION

The healthcare and medical fields are rich in information but is not properly used to its potential leads to weak or bad decision-making ability. The proposed work focuses on the data which is not mined. Here ten attributes are used to predict the chances of heart disease and thereby preventive measures can be taken to avoid serious effects (8). The proposed system is reliable web based and user-friendly application. The system provides platform for user to share

their heart related issues. So that it can be provided with effective medical guidance to the users. It reduces the time for medical treatment providing deduction of causes of diseases and identification of illness.

For each patient a unique authentication is done. After the patient is an authorized user there is an access to application GUI and enters the symptoms. The symptoms are stored in the database and can be loaded, selected via CSV format (9). Information can be imported in the database via Excel files. Database is saved and loaded. Mining techniques (10) are applied that is k-means and naïve bayes. Clustering is done with respect to above parameters considering age as the primary parameter. After the age is clustered, various groups are formed. Naïve bayes is applied which gives the conditional probability of the patient who will suffer heart disease in the future with respect to rest of the parameters as declared.

Intelligent System for heart disease is predicted by so many attributes. But 14 attribute which are basically proven to be important and give better results with a smaller number of tests. To predict with a smaller number of attributes and faster efficiency to predict the risk of having heart disease the Naïve Bayes algorithm gives 52.23% of accuracy in less time than others. It also having low error rate. So, it is one of the intelligent systems for prediction.

IV. CONCLUSION

The overall objective of our work is to predict accurately with a smaller number of tests and attributes the presence of heart disease. In this work fourteen attributes are considered which form the primary basis for tests and give accurate results more or less. Many more input attributes can take but our goal is to predict with less number of attributes and faster efficiency to predict the risk of having heart disease at a particular age span. Two data mining classification techniques were applied namely K- means and Naive Bayes. As shown above, it is clear that Naïve Bayes has better accuracy in less time than others. Other data mining technique can also be used for predication such as Neural Networks, Time series, Association rules.

REFERENCES

[1] AshaRajkumar and Sophia Reena, “Diagnosis of Heart Disease using Data Mining Algorithms”, Global Journal of Computer Science and Technology,2010;10: 38-43.
[2] BalaSundar V, “Development of Data Clustering Algorithm for predicting Heart”, IJCA, 2012; 48:8-13.
[3] Chapman, P., Clinton, J., Kerber, R. Khabeza, T.,Reinartz, T., Shearer, C., Wirth, R., “CRISP-DM 1.0:Step by step data mining guide”, SPSS, 2000;2: 1-78.
[4] Liangxiao. J, Harry.Z, Zhihua.C and Jiang.S, “OneDependency Augmented Naïve Bayes”, ADMA, 2005; 2: 186-194.
[5] Manjusha K. K, K. Sankaranarayanan, Seena P, “Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification”, International Journal of Advanced Research in Computer Science and Software Engineering, 2014; 4: 77-82.
[6] G.Subbalakshmi , K. Ramesh and M. ChinnaRao , “Decision Support in Heart Disease Prediction System using Naïve Bayes”, Indian Journal of Computer Science and Engineering, 2011.
[7] Shadab Adam Pattekari and AsmaParveen , “Prediction System for Heart Disease Using NaïveBayes”, International Journal of Advanced Computer and Mathematical Sciences, 2012; 3: 290-294.
[8] Chaltrali S. Dangare and Sulabha, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”,IJCA, 2012; 47: 44-48.
[9] CSV File Reading and Writing (http:/ / docs.python. org/ library/csv. Html) is no CSV standard, Retrieved July 24, 2011.
[10] K.R.Lakshmi, M.Veera Krishna and S.PremKumar, “Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability”, International Journal of Scientific and Research Publications, 2013; 3:6-9.