International Journal of Research in Engineering, Science and Management
Volume-2, Issue-2, February-2019
www.ijresm.com | ISSN (Online): 2581-5792

Prediction of Heart Disease based on Certain Attributes using ACR Data Mining Techniques

S. G. Divya[1], A. Amreen[2], R. Induja[3], M. Balamurugan[4]
[1,2,3]*UG Scholar, Dept. of Computer Science & Engineering, The Kavery Engineering College, Mecheri, India*
[4]*Professor & HoD, Dept. of Computer Science & Engineering, The Kavery Engineering College, Mecheri, India*

***Abstract*: Medical data mining has a great potential for exploring the hidden patterns in the data sets of medical domain. These data need to be collected in a standardized form. From the medical profiles certain attributes are extracted such as age, sex, blood pressure, blood sugar, trestbps: resting blood pressure, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num etc. can predict the likelihood of patient getting heart disease. The Association rule mining algorithms must perform efficiently. Also it is suggested to add upon the approach for association rule mining based on sequence number and clustering the transactional data base for heart attack prediction. These attributes are fed in to Fuzzy C- Means clustering Algorithm, applying the association, clustering and regression data mining technique to heart disease treatment; it can provide as reliable performance as that attained in diagnosing heart disease. By this medical industries could offer better diagnosis and treatment of the patient to attain a good quality of services. The main advantages of this paper are: early detection of heart disease and its diagnosis correctly on time**

***Keywords*: Heart disease; Data mining; heart disease prediction; Fuzzy c-means clustering; attributes.**

1. Introduction

In the modern life style health diseases are increasing tremendously. Our life style had a great impact on our health causing heart diseases and other health problems. Taking a survey of present population it is seen that about sixty percentages are suffering from heart diseases. Early detection of heart diseases can prevent the death rate, people are not aware about the detection of heart disease earlier due to lack of knowledge. Health care industries are aiming to diagnose the disease at early stages. In most cases it is noticed at the final stages of disease or after death. The cost of treatment for heart disease is very expensive. The treatment cost is not affordable for everyone. Therefore people are reluctant to do proper treatment at early stages of disease. The aim of our project is to diagnose the disease at early stage at affordable cost. By using data mining technique we can detect disease at early stage and we can completely cure the disease by proper diagnosis. Health care industry collect huge amount of data, which are not mined to discover hidden information. Remedy of this problem is data mining technique. Data mining is the process of analyzing large set of data and summarizing into useful information. Data mining techniques are:

1. Association
2. Clustering
3. Regression

Association rule mining is the data mining process of finding frequent patterns, correlations or casual sets of items or objects in transaction data base, relational database. Fuzzy c-means is a way of clustering which allows one piece of data to belong to two or more clusters. Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent and independent variable (s).

2. Related works

In this segment, we appraise the existing literature and confer about different aspects of data mining applications in prediction of heart disease:

In the year 2017 C.sowmiya, Dr. P. Sumithra [1] "Analytical study of heart disease diagnosis using classification techniques" in this paper the author analyze HD by classification technique with the proposed algorithm of apriori algorithm and svm in heart disease prediction.

In the year 2015 Ankita dewan, Meghna Sharma[02] "Prediction of heart disease using a hybrid technique in data mining classification" in this paper the author develop a prototype and extract unknown knowledge related with HD from a past HD database recordThe algorithm used here is neural networks , decision tree , navie bayes.

In the year 2018 D.Karthick , B.Priyadharshini [3]" predicting the changes of occurrence of cardio vascular disease (CVD) in people using classification technique with in fifty years of age" in this paper using some number of attributes of people data set which predicts the changes occurs in cvd. Classification algorithm like navie bayes algorithms used to develop a model. The output is predicted by binary classification which means when 1 occurs there will be a changes if 0 occurs there is no change.

In the year 2014 Deepali Chandna [04]" Diagnosis of heart disease using data mining algorithm" in this paper the principle of this study is , hence to extract hidden patterns by applying data mining technique .The study found that the accuracy for the proposed approach is 98.2% compared with other method.The Algorithms used is k-nearest neighbour's algorithm.T his work projected a system that uses method

called information gain and adaptive neuro-fuzzy inference, neural networks, system for heart disease diagnosis

In the year 2017 P. Sudeshna, S. Bhanumathi, M. R. Anish Hamlin[05]" Identifying symptoms and treatments for heart disease for biomedical literature using text data mining" In this paper they modify an automatic machine technique which is used for disease identification and correct medicine analysis based on evidence is achieved. The algorithm used is svm algorithm, decision tree, and sequential minimal optimization.

In the year 2015 Jyoti Soni et al. [6]"A survey of current techniques of data extraction from databases using data mining techniques" that are used in Heart Disease Prediction. The techniques used here are Naive Bayes, Decision List and KNN. Here the Classification based on clustering is not performing well.

In the year 2018 Mingliu, Younghoon Kim. [07]"Classification of heart disease based on ECG signals using long short term memory" in this paper the author proposed a method of classifying heart disease using ECG signals that achieves high accuracy in short period. The accuracy of classification is 98.4% and the response time is much less than the method without preprocessing.

In the year 2014 B. Venkatalakshmi, M.V. Shivasankar [8]"Heart disease diagnosis using predictive data mining " dataset of 294 records with 13 attributes is used and the outcome reveals the Navie Bayes outperforms and sometime decision tree. In future genetic algorithm is used.

3. Proposed system

The main objective of this project is to analyze the Medical Data to mine the patterns and relationships associated with heart disease from heart disease data base using association rule mining and also to provide efficient Heart attack prediction methodology. Association rule mining is one of the fundamental research topics in data mining and knowledge discovery that finds interesting association or correlation relationship among the set of large data modules and predicts the associative and correlative behaviors for new data. The Association rule mining algorithms must perform efficiently. Also it is suggested to add upon the approach for association rule mining based on sequence number and clustering the transactional data base for heart attack prediction. The Clustering is done by Fuzzy C- Means clustering Algorithm. Further, the Prediction level is still improvised in our methodology by utilizing Multinomial Logistic Regression.

Advantage:

- Logistic regression doesn't require linear relationship between relative and unrelative variables. It can handle various types of relationships because it applies a non-linear log transformation to the prediction ratio.
- It is more suitable for medical data where large sample size is used because maximum likelihood estimates are more powerful.
- Data is also clustered and then the clustered data are associated using association rules for better prediction.
- Provides efficient way of heart attack prediction.
- Utilizes maximum of 16 attributes for improving the prediction level.

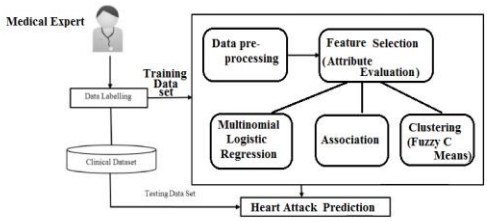

Fig. 1. System architecture

4. Modules

1. Data Collection & Data Preprocessing: The Heart attack dataset was obtained from the repository. The dataset consists of data from different heart attack cases containing 20 descriptive attributes and multiple classification variables. The original dataset is then divided into a training set and a test set. Models were constructed and tested using both the full variable data set as well as a limited dataset.

2. Data Repository Management: Heart Attack Prediction data Management System provides complete assistance in patient management. In the module of the patient management system, there is a facility to register patients and view their reports and history. Medical management system allows getting detailed information of patient's health care.

3. Feature Selection and Attribute Evaluation: This module enables the administrator or the medical expert to analyze the various features or parameters to predict the heart attack by analyzing the health status of the patients. In this system From the medical profiles certain 16 attributes are extracted such as age, sex, blood pressure, blood sugar, trestbps : resting blood pressure, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num are used for further processing of prediction.

4. Clustering and Association: Association rule mining is one of the fundamental research topics in data mining and knowledge discovery that finds interesting association or correlation relationship among a large set of data modules and predicts the associative and correlative behaviors for new data. The association is based on sequence number and clustering the transactional data base for heart attack prediction. The Clustering is done by Fuzzy C- Means clustering Algorithm.

5. Prediction using Regression: Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. Here Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the relative variable is binary in

nature. The parameters are chosen to maximize the chance of observing the sample values. If relative variable is multi class then it is known as Multinomial Logistic regression.

5. Proposed methodology algorithms

Algorithm 1: To detect the possibility of heart attack

Step 1: Load the details of patient from .dat file named dataset.
data = load('dataset.dat');
//where, data contains the files extracted from dataset.dat.

Step 2: Collect data and Convert the dataset values into an array.

Step 3: Apply fuzzy c-means clustering technique.
[center,U,obj_fcn] = fcm(data, n_clusters);
//where, center represents the center matrix of the clusters, U represents the membership function matrix.
//Obj_fcn represents the objective function.
//data is the matrix containing all the patient's
//information
//n_clusters represents the number of clusters to be
//formed which is 2.

Step 4: Compare the 2 clusters based on their membership values.
if(U(1,i)>U(2,i))
p(1,i) = U(1,i);
else
p(2,i)=U(2,i);
end
//where, p is the matrix containing compared values

Step 5: If (prone = true)
disp('you are prone to heart attack');
//send message
else
goto step 1 //read input of next patient

6. Fuzzy clustering algorithm

Fuzzy algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. In the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, addition of membership of each data point should be equal to one. After each iteration membership and cluster centers are upgrade according to the formula:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

Where,
'n' is the number of data points.
'vj' represents the jth cluster center.
'm' is the fuzziness index m € [1, ∞]
'c' represents the number of cluster center.
'µij' indicates the membership of ith data to jth cluster center.
'dij' represents the Euclidean distance between ith data and jth cluster center.

Main mottive of fuzzy c-means algorithm is to minimize:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \left\| x_i - v_j \right\|^2$$

Where,
'||xi – vj||' is the Euclidean distance between ith data and jth cluster center.

Algorithmic steps for Fuzzy c-means clustering

Let X = {x1, x2, x3 ..., xn} be the group of data points and V = {v1, v2, v3 ..., vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the fuzzy membership 'µij' using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3) Compute the fuzzy centers 'vj' using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

$||U(k+1) - U(k)|| < \beta$.

where,
'k' is the iteration step.
'β' is the termination criterion between [0, 1].
'U = (µij)n*c' is the fuzzy membership matrix.
'J' is the objective function.

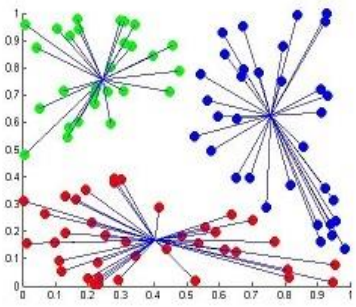

Fig. 2. Result of Fuzzy c-means clustering

Advantages of fuzzy Clustering

1) Gives best result for overlapped data set and comparatively better then k-means algorithm.
2) Disprate k-means where data point must exclusively related to one cluster center here data point is assigned membership to each cluster center as a result of which data point may relate to more than one cluster center.

7. Multinomial Logistic Regression

Logistic regression is used to find the probability of occurence=Success and occurence=Failure. We should use logistic regression when the relative variable is binary (0/ 1,

True/ False, Yes/ No) in nature. Here the value of Y extend from 0 to 1 and it can represented by following equation.

odds= p/ (1-p) = probability of occurrence / probability of not occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best purpose for this distribution. And, it is logit function. From the above equation, the attributes are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

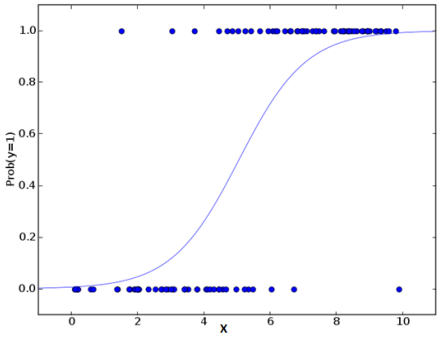

Fig. 3. Graph

8. Attribute list

The extracted attribute are as follows,

1. age: age in years
2. sex: sex (1 = men; 0 =women)
3. cp: chest pain type
   Value 1: typical angina
   Value 2: atypical angina
   Value 3: non-anginal pain
   Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (speeding blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
   Value 0: normal
   Value 1: having ST-T wave unusuall (T wave inversions and ST elevation or depression of > 0.05 mV)
   Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise also cover angina (1 = yes; 0 = no)
10. oldpeak = ST depression involve by exercise relative to rest
11. slope: the slope of the peak exertion ST segment
    Value 1: upsloping
    Value 2: flat
    Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. num: diagnosis of heart disease
    Value 0: < 50% diameter narrowing
    Value 1: > 50% diameter narrowing
15. Diagnosis : diagnosis classes
    Value 0: healthy
    Value 1:not healthy
16. painloc: chest pain location
    Value 0: substernal
    Vaue1:otherwise.

9. Conclusion

There is a need for combinational and more complex models to increase the accuracy for predicting the heart disease. It is concluded that, as identified through regression, it is a more efficient than other techniques as it is combinational and models to increase the accuracy of prediction set of cardiovascular diseases. This is a framework using combinations Association clustering and logistic regression which analyzes maximum of 16 attributes for more accuracy. Hence, we arrive at an accurate prediction of heart attack or heart disease data provides guidelines to train and test the system thus attain the most efficient model or rule based combinations. The Future work can be extended to include real time sensor based data collection of different patient health parameters and update to the medical database in the servers maintained through IOT connected devices.

References

[1] C.sowmiya , Dr.P.Sumithra "Analytical study of heart disease diagnosis using classification techniques" IEE international conference on intelligent techniques in control, optimization and signal processing, 2017.

[2] Ankita dewan , Meghna Sharma "Prediction of heart disease using a hybrid technique in data mining classification", 2015

[3] D. Karthick, and B. Priyadharshini " predicting the changes of occurrence of cardio vascular disease (CVD) in people using classification technique with in fifty years of age", 2018.

[4] Deepali Chandna" Diagnosis of heart disease using data mining algorithm", in International journal of computer science and information technologies, vol.5(2), 2014, pp. 1678-1680.

[5] P. Sudeshna, S.Bhanumathi, and M.R.Anish Hamlin" Identifying symptoms and treatments for heart disease for biomedical literature using text data mining" international conference on computing of power , energy , information and communication, 2017.

[6] Jyoti Soni et al., "A Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, Volume 17, No.8, March 2011.

[7] Mingliu, Younghoon Kim, "Classification of heart disease based on ECG signals using long short term memory", 2017.

[8] B. Venkatalakshmi, and M.V. Shivasankar, "Heart disease diagnosis using predictive data mining," ICIET'14, 2014.