Prediction of Coronary Heart Disease Using Apriori algorithm with Data Mining Classification

Rishabh Wadhawan
School Of Computer Science and Engineering, VIT University, Vellore.
Email: rishabhwadhawan1996@gmail.com

Abstract -Heart disease prediction is thought to be as a standout amongst the most confounded assignment in the field of Healthcare. Therefore there is a pressing need to build up a (DSS) Decision support system for identifying coronary illness of a patient. Around 70% of the total populace succumbs to the coronary illness. Albeit huge advance has been made in the analysis and treatment of coronary illness, still more progress is as yet required. Information mining, as an answer to extricate concealed patterns from the clinical dataset are applied to a database in this proposed method. The database comprises of different instances and attributes. The framework was implemented in visual studio c#. The fundamental goal of this paper to develop a system prototype which can help determine and extract unknown knowledge related with heart disease from a past heart disease database record. It can unravel convoluted questions for identifying coronary illness and consequently help medicinal experts to settle on savvy clinical choices which conventional traditional decision support systems were not ready to do. It can likewise help lessen the cost of medicines by giving precise and proficient expectations.

Keywords –Heart disease prediction, coronary heart disease, data mining techniques, heart disease symptoms, classification techniques
I. INTRODUCTION

Heart disease is the one of the largest cause of death among the world nowadays especially for people above the age of 55. Cholesterol, blood pressure, pulse rate are the major reason for the heart disease. Some non-modifiable factors are also there. Such as smoking, drinking also reason for heart disease.

Heart is the most important organ of our human body. The heart pumps new blood in our body and without the functioning of heart for even as mere as 5 seconds we may not be able to survive, the body might get paralyzed due to insufficient blood or may have other issues. High risk factors of heart disease include Family history of heart disease, Age, High blood pressure, Cholesterol, Excessive Drinking, Poor diet, Smoking. When blood vessels are overstretched, the risk level of the blood vessels is increased. This leads to the blood pressure. Blood pressure is typically measured in terms of systolic and diastolic.

Systolic rate shows the pressure in the veins when the heart muscle contracts and diastolic demonstrates the weight in the conduits when the heart muscle is in resting state. The level of lipids or fats expanded in the blood are causes the coronary illness. The lipids are in the supply routes thus the conduits end up thin and blood stream is likewise turned out to be moderate. Age is the non-modifiable hazard factor which likewise is an explanation behind coronary illness. Smoking is the explanation behind 40% of the passing of heart illnesses. Since it confines the oxygen level in the blood then it harm and fix the veins.

There are number of factors which increase the chances of Heart disease:

- Hypertension
- Physical inactivity
- Obesity
- Poor Diet
- High Blood Pressure
- High Blood Cholesterol
- Family History of Disease
- Smoking
Different information mining strategies, for example, Naïve Bayes, KNN calculation, Decision tree, Neural Network are utilized to foresee the danger of coronary illness. The Apriori calculation utilizes the user given values to discover the estimations of the elements of coronary illness. Decision tree calculation is utilized to give the grouped answer to the coronary illness. This strategy is utilized to anticipate the coronary illness through likelihood. The Neural Network gives the limited mistake of the expectation of coronary illness. In this previously mentioned systems the patient records are arranged and anticipated constantly. The patient movement is observed persistently, if there is any change in movement, at that point the hazard level of ailment is educated to the patient and specialist. The specialists can anticipate heart infections at a prior stage on account of machine learning calculations and with the assistance of PC innovation. This paper gives a knowledge about apriori information mining system used to foresee heart sicknesses.

II. REVIEW OF LITERATURE

In this section the work done in the field of heart disease prediction or existing systems for heart disease prediction are analysed.

N. Deepika et al. [5] in their paper developed a method consisting of Association Rule for classification of Heart-attack patients. They presented the extraction of important patterns from heart disease data warehouse. The heart disease data warehouse contains the data that is obtained from various heart disease patients. Initially, data from the warehouse is pre-processed to make the mining process more efficient. The first stage of Association Rule is to pre-process the data in order to handle missing values. After pre-processing a method called equal interval binning with approximate values based on medical expert advice was applied on Pima Indian heart attack data. The significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a particular threshold level were chosen and they were used in the design and development of the heart attack prediction system for better accuracy. The, Pima Indian Heart attack dataset used was obtained from the UCI machine learning repository.
K. Srinivas et al.[10] presented Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks. They presented how data mining techniques such as Artificial Neural Network, Rule based, Naïve Bayes and Decision tree can be used for classification of massive Volume of healthcare data. The data mining tool used for exploratory data analysis, machine learning and statistical learning algorithms was Tangara. The training data set consists of 3000 instances with 14 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of heart disease. The performance of the classifiers is evaluated and their results are analyzed. The results of comparison are based on 10 tenfold cross-validations. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. The comparison made among these classification algorithms out of which the naïve apriori algorithm considered as the best performance algorithm.

M A. Jabbar et al.[9] proposed a method for Association Rule mining based on the sequence number and clustering for heart attack prediction. The entire database is divided into partitions of equal size. Dataset consisting of 14 attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation. Those frequent patterns which had a value greater than a particular threshold were chosen for accurate and precise prediction of heart attack. Three end result mining goals were defined based on data exploration and all the models proposed could answer complex queries in predicting heart attack.

Mai Shouman, et al.[7] proposed a method which consisted of k-means clustering along with the decision tree method to predict the heart disease. The enhanced algorithm proposed by them consisted of several centroid selection methods for k-means clustering to increase efficiency. The database consisting of 13 input attributes was collected from Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity, and accuracy in the proposed are calculated with different initial centroids selection methods and different numbers of clusters. For the random attribute and random row methods, ten runs were executed and the average and best for each method were calculated. It was also observed that decision trees with k means clustering gave better results than just using decision trees for prediction of heart disease. The proposed method which consisted of integrating k-means clustering and decision
tree achieved higher accuracy than the paging algorithm in the diagnosis of heart disease patients. At last the end result was that the accuracy achieved was 83.9% by the proposed method with two clusters.

S. Indhumathi, et al.[4] proposed a Web based health care monitoring system using naive Bayes algorithm. Poor clinical decisions nowadays can lead to disastrous consequences which can lead to a patient’s death and also reducing hospitals reputation and are therefore unacceptable. Hospitals have to minimize the cost of clinical tests so as to provide its operations to anyone in need. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most of the hospitals today use some sort of hospital information systems to manage their healthcare database or patient data. Some hospitals nowadays use decision support systems (DSS), but they are largely limited. The main objective of the research proposed in the paper was to develop an Intelligent Heart Disease Prediction System using Naïve Bayes algorithm which is easy to implement and also would result in cost efficient solution. The proposed technique is implemented as web based questionnaire application. After filling the questionnaire by the user, it can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database based on how the user fills the questionnaire. It can answer complex queries for diagnosing heart disease and therefore assist healthcare practitioners to make accurate and effective clinical decisions which traditional decision support systems cannot. It also helps in reducing cost by providing efficient and effective treatment details.

III. DATASET DESCRIPTION

The standard dataset, compiled in this research contains 209 records, which is gathered from a clinical facility containing records of patients with coronary illness. Information is accumulated from a solitary asset, so it blocks any joining tasks in this manner no mix procedure is required. Eight characteristics are used, from them, 7 are considered as sources of info which anticipate the condition of the quality "Determination". Every one of the attributes, alongside their values and data types are discussed about in Table underneath. On the off chance that Diagnosis comes to be 1 then the patient is at a high danger of coronary illness else okay of coronary illness otherwise he is at low risk.
<table>
<thead>
<tr>
<th>Attributes</th>
<th>Descriptions</th>
<th>Values</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age in years</td>
<td>28-66</td>
<td>Numeric</td>
</tr>
<tr>
<td>Nominal Rest Blood Pressure</td>
<td>Patient’s resting blood pressure in mm Hg at the time of admission to the hospital</td>
<td>94-200</td>
<td>Numeric</td>
</tr>
<tr>
<td>Blood Sugar</td>
<td>Below 120 mm Hg-Normal Above 120 mm Hg-High</td>
<td>High = 1 Normal = 0</td>
<td>Nominal Binary</td>
</tr>
<tr>
<td>Rest Electrocardiographic</td>
<td>Normal, Left Ventricular Hypertrophy (LVH) ST_T wave abnormality</td>
<td>Normal=1 Left Vent Hyper = 2 ST_T wave abnormality = 3</td>
<td>Nominal</td>
</tr>
<tr>
<td>Maximum Heart Rate</td>
<td>maximum heart rate attained sport test</td>
<td>82-188</td>
<td>Numeric</td>
</tr>
<tr>
<td>Exercise Angina</td>
<td>It includes two conditions of positive and negative</td>
<td>Positive = 1 Negative = 0</td>
<td>Nominal Binary</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>It includes two conditions of positive and negative</td>
<td>Positive = 1 Negative = 0</td>
<td>Nominal Binary</td>
</tr>
</tbody>
</table>

Table1: Shows the parameters and attributes being used.
IV. TECHNIQUES AND METHODOLOGY

PROPOSED TECHNIQUE

The given project has been implemented in c# as it is easy to implement such algorithms in an object oriented programming language. A Prediction method using Apriori algorithm proposed in this study. Apriori algorithm is a technique which is used to mine frequent item sets in a database with association rules. For this project there are two steps. The Initial step consists of providing with the Database which consists of all the relevant values and also pre-process it to remove redundant data and blank values. All the attributes need to be in 0 or 1. After this there is a Classifier module in which data is trained through algorithm and classified as either low risk or high risk. This classified data is then compared with the patient’s data to predict whether he/she is at low risk or high risk of a heart disease.

In the classifier module the association rules are calculated based on the minimum confidence and support values. The stronger the association rules the greater the prediction capability of the software.

Brief working of the Apriori Algorithm in the proposed system:

In the proposed strategy after the classification of the data by k means clustering the data is then analysed using apriori algorithm which in turn makes the candidate list from dataset and after reducing the number of candidates at each iteration based on a certain value of confidence it gives the final candidate list.

After obtaining the candidate list the apriori algorithm the candidate and support values to make association rules. The stronger the rules the better the result. Support, confidence and association rules are defined below.

1. Support:

The support of an item set, is the proportion of transaction in the database in which the item X (In our case the characteristic) appears. It signifies the popularity of that item (characteristic) set in the database. In our case it will be the values of the characteristic that appear more frequently. Furthermore the
calculated support helps us in identifying strong association rules and also helps in predicting whether those attributes result in low risk or high risk for heart disease.

2. Confidence:

Confidence of a rule is defined as follows:

It signifies the likelihood of item Y being possible when item X is possible. In our case let us assume that if blood sugar is high and exercise is low then there is a possibility of a heart disease

3. Association Rules

Using association rules, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. As one of the example where it is used, the association technique is used in heart disease prediction as it is used to tell us the relationship of different attributes used for analysis and sort out the patient with all the risk factor which are required for prediction of disease as is done in this.

Fig 1: shows the working of Apriori in the project.
V. IMPLEMENTATION AND ARCHITECTURE:

A. ARCHITECTURE

![Architecture Diagram]

Fig 2: Shows the architecture of the proposed system

The proposed system consists of two given modules

1. Patient: The patient can Login if he/she is an existing user. If not then the patient can register himself by filling up the registration form. After the registration the patient can login and has to upload his medical details and after the results are ready then the patient can request for his report.

2. Doctor: The doctor first has to upload the medical dataset containing the previous records of the heart disease patients which is then pre-processed to remove redundant and missing values. After pre-
processing the dataset is classified using k means algorithm into low risk and high risk. After this the data uploaded by the patient is compared against this and the report is generated which is then sent to the patient by the doctor upon patient’s request.

B. IMPLEMENTATION DETAILS

The proposed technique combines rule mining using Apriori algorithm and mafia algorithm as well as classification using KNN algorithm to efficiently predict the heart diseases and also to increase the accuracy.

In the proposed system the medical database is uploaded first which is pre-processed to remove redundancy and missing values. After pre-processing the database is classified using the KNN classification algorithm using a k value of 2 after which we get 2 clusters basically consisting of parameters with low risk values and high risk values. It uses the attributes in database to classify each record as high risk or low risk. After the classification of the records, the association rules are identified using Apriori and mafia algorithm.

The stronger the rules higher will be the accuracy. For mining strong rules, confidence and support values are to be used and set accordingly. After all this the patient’s medical records consisting of various attributes is given as input and accordingly it is predicted whether he/she is at high risk or at low risk.

VI. EXPERIMENTAL RESULTS

First the heart disease database taken into account is pre-processed successfully by eliminating identical records and providing missing values. The refined heart disease data set, resulting from pre-processing, is then collected by K-means algorithm with the K value of 2. It was observed that the higher the confidence and support of a rule, the more it is accurate thus is more close to representing a regular pattern in the dataset. If these measures are relatively low, then any inconsistency would be less strong than it would be for rules with high confidence and high support. The end result would be a classified database against which the user entered when compared would tell the user whether he is at high risk of disease or at low risk.
The precision and recall can be further calculated as follow:

The metrics for above can be derived from the confusion matrix and can be easily converted to true-positive (TP) a false-positive (FP) metrics.

Precision = \( \frac{TP}{(TP+FP)} \)

Recall = \( \frac{TP}{(TP+FN)} \)

True Positive (TP): The total percentage of members classified as Class A that actually belong to Class A.

False Positive (FP): The total percentage of members classified as Class A but in reality do not actually belong to Class A.

False Negative (FN): The total percentage of members of Class A incorrectly classified as not belonging to Class A.

True Negative (TN): The total percentage of member which do not belong to Class A and are also classified as not belonging to Class A. It can also be given as (100%-FP).

<table>
<thead>
<tr>
<th>Technique</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>K means using Apriori algorithm</td>
<td>.78</td>
<td>.67</td>
<td>74%</td>
</tr>
</tbody>
</table>

Table 2: Shows the result obtained after applying the specified algorithm.

The below images show dataset after pre-processing and also the dataset when it is classified with K means value showing low risk in Heart attack low diagnosis and high risk in heart disease diagnosis and output the user gets after entering his medical records.
VII. CONCLUSION

Different Apriori strong rules were contrasted with aim to foresee coronary illness. A one of a kind model comprising of one filter and evaluation techniques are developed. Three strong rules and distinctive assessment techniques, are applied to find the superior software. Apriori rules are thought about with respect to their exact number of support, better exactness, and considering strong rules. The high performance software was presented. The trial can serve as an instrument for doctors to successfully anticipate dubious cases and prompt appropriately.

The target of our work is to give an investigation of Apriori calculation based information mining procedures that can be utilized in computerized coronary illness expectation frameworks. Different methods and information mining classifiers are characterized in this work which have risen as of late for productive and compelling coronary illness conclusion. The examination demonstrates that distinctive innovations are utilized in the papers taken into account also taking diverse number of attributes for
datasets taken. So, different technologies used shown the different accuracy to each other. In some papers it is shown that neural networks give a better result. But using Apriori algorithm with strong rules we can get results as good as a neural network.

The future work with regards to this project can be to plan, design and build up a proficient heart attack prediction framework with Patient Prescription Support utilizing web mining and data warehouse strategies which can be extremely useful for effective and precise prediction system and furthermore a shoddy one that is accessible effortlessly.

REFERENCES


Issue: 10, pp.3003-08, October 2014.


