ISSN:0975-9646

An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm

Rucha Shinde[(1)], Sandhya Arjun[(2)], Priyanka Patil [(3)], Prof. Jaishree Waghmare[(4)]

Trinity College of Engineering & Research, Pune

Abstract—Nowadays people work on computers for hours and hours they don't have time to take care of themselves. Due to hectic schedules and consumption of junk food it affects the health of people and mainly heart. So to we are implementing an heart disease prediction system using data mining technique Naïve Bayes and k-means clustering algorithm. It is the combination of both the algorithms. This paper gives an overview for the same. It helps in predicting the heart disease using various attributes and it predicts the output as in the prediction form. For grouping of various attributes it uses k-means algorithm and for predicting it uses naïve bayes algorithm.

Index Terms —Data mining, Comma separated files, naïve bayes, k-means algorithm, heart disease.

I. INTRODUCTION

The practice of examining large preexisting data bases in order to generate new information. It coverts raw data into useful information. It analyze the data for relationships that have not previously been discovered. [1]

The steps of data mining are: Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation.

Medical data mining is a domain of lot of imprecision and uncertainty. The clinical decisions are usually based on the doctors intuition. Therefore this may lead to disastrous consequences. Due to this there are many errors in the clinical decisions and it results in excessive medical costs. [1]

Serialization is also used in this system. It converts the data objects into streams of bytes and stores it into database.

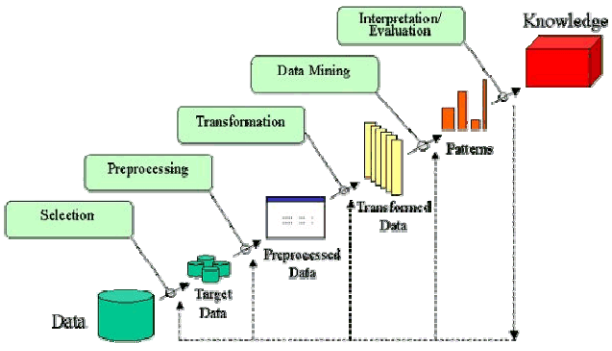

Figure1:-Data Mining Process

II. LITERATURE REVIEW

[1] Intelligent heart disease prediction system using data mining techniques:

In this paper heart disease prediction is done using data mining techniques such as decision trees, neural network and naïve bayes. This system answers "what if " query. It is implement on .net platform. It is used for heart disease prediction.

[2] An empirical study on applying data mining techniques for analysis and prediction of heart disease:

It is found that health environment is poor in extracting knowledge so in this paper data mining techniques are applied . this paper deals with application of data mining.

[3] Prediction system for heart disease using naïve bayes mining:

It is web-based classification. It retrives hidden data from database. It compare the value with trained dataset. In this paper it is mentioned that because of this system the treatment cost are reduced.

[4] Decision support in heart disease prediction system using naïve mining:

This research developed using data mining techniques mainly naïve bayes. It takes input as the patients attributes. It helps trained nurses and medical students to treat patients.

[5] Intelligent and effective heart attack prediction system using data mining:

In this paper k-means clustering is used. This system capable of predicting heart disease.

III. BLOCK DIAGRAM

The following block diagram represents the step by step implementation of the heart disease prediction system. The block diagram consist of two sets first one is the training set and the other one is prediction. In training set firstly the input is taken i.e the patients attributes then a dataset is being formed. After that dataset is given labels according to the name of the attributes. Then on the dataset transformation is done means the attributes are separated through comma separated vector i.e C.S.V files. After that on these dataset K-Means clustering algorithm is applied, here the grouping of the attributes is done and the attributes are added according to their groups. After this model is ready to apply prediction algorithm on it.

In prediction system, for prediction we used naïve bayes algorithm. Naïve bayes basically applies probability

concept. It integrates on each and every attribute and gives the result. The output which we would get will be prediction the person is having heart disease or he is likely to have heart disease. This system will help him to take the preventive measures from not getting the disease.

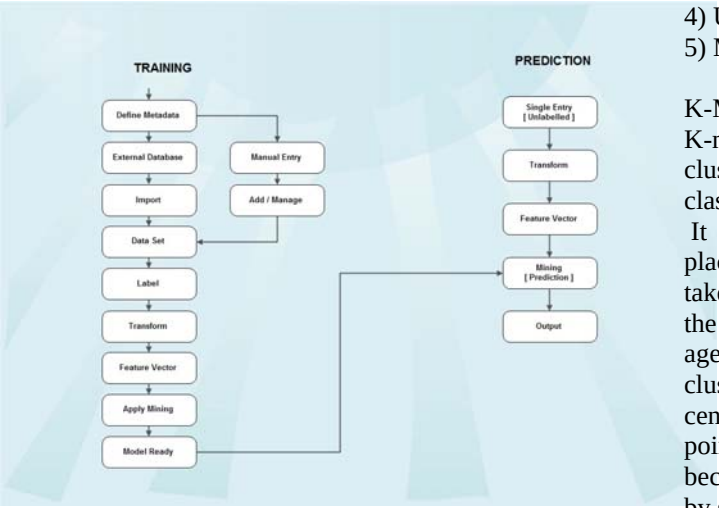

Fig. 2:Block diagram

IV. COMMA SEPARATED VALUES(CSV)

The full form of CSV is Comma Separated Values. In older days CSV is originally named as CSL i.e Comma Separated List.[2][3]

A CSV file stores plain text form into tabular data. Plain test contains character with no data as binary number. A CSV file contains records which are separated by some character or strings mainly by comma or tab. CSV refers to large family of formats.

It is supported by consumer and business applications. It is mainly used when user needs to transfer information from a database programs to a spreadsheet that uses a completely different format.[2][3]

In CSV files records are divided into fields separated by delimiters. They work both with UNICODE and ASCII. CSV files translates one character set to another.

CSV files cannot represent object oriented database because CSV records excepted to have same structures.

CSV files are also called as flat files.[2][3]

In this system we take various attributes such as age, obesity, gender, cholesterol, smoker ,blood pressure, chest pain ,blood sugar, ECG results etc. As we take this input one by one this inputs are separated using CSV files. This inputs are converted into a tabular format and are separated using comma.

Because of CSV files data appear in a sophisticated and in well- represented manner.

V. K-MEANS CLUSTERING ALGORITHM

Clustering is the process of grouping of data objects that are same to one other within the cluster. They even grouped dissimlar objects into another cluster. It is also called as data segmentation in some applications because it divides large data set into groups according to the similarities.[4]

Requirements of clustering in data mining:-

1) Deals with different types of attributes.
2) Deals with noise data
3) It requires minimum knowledge to determine input parameter.
4) Usability
5) More dimensionality

K-MEANS CLUSTERING

K-means is most simplest learning algorithm to solve the clustering problems. The process is simple and easy, it classifies given data set into certain number of clusters.

It defines k centriods for each cluster. They must be placed as much as possible far away from each other. Then take each point belonging to given data set and relate into the nearest centriod. If no point is pending then an group age is done. Then we re-calculate k new centroid for the cluster resulting from previous steps. When we get the k centroid a new binding is to be done between sane data points and nearest centroid. A loop is been generated because of this loop key centriod change the location step by step until no more changes are done.[4]

The advantages of k means clustering algorithm are simplicity and speed.

Algorithm:-

1) Select k center from the problem(random)
2) Divide data into k clusters by grouping points.
3) Calculate the mean of k cluster to find new centers.
4) Repeat steps 2 and 3 until centers do not change.

In this system we mainly used clustering for grouping the attributes. As we take almost 10 attributes such as age In this system we take various attributes such as age, obesity, gender, cholesterol, smoker ,blood pressure, chest pain ,blood sugar, ECG results etc. this attributes are grouped using K-Means clustering algorithm

Eg:- If we took an attribute such as age and we considered the age of the person between 0-100. After applying K-means algorithm on this dataset of age it will find the centriod and divide it into groups. It calculate the mean. Here, age will be divided into 3 groups such as from 0-30,31-60,61-100.

It will give them values such as

0-30=0
31-60=1
61-100=2

For gender attribute it will divide into groups such as

Male=0
Female=1

K-means will be applied on each and every attribute mentioned above.

After that the attributes and their values will be added in a dataset accordingly. Then the model is being ready for prediction.

VI. NAÏVE BAYES ALGORITHM

Naïve Bayes classifier is based on Bayes theorem. It has strong independence assumption. It is also known as independent feature model.

It assumes the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature in the given class.

Naïve bayes classifier can be trained in supervised learning setting. It uses the method of maximum similarity. It has been worked in complex real world situation. It requires small amount of training data. It estimates parameters for classification. Only the variance of variable need to be determined for each class not the entire matrix.[5][6]

Naïve bayes is mainly used when the inputs are high. It gives ouput in more sophisticated form. The probability of each input attribute is shown from the predictable state.

Machine learning and data mining methods are based on naïve bayes classification.

Bayes theorem:-

$$P(H|X) = \frac{P(X|H)\ P(H)}{P(X)}$$

Where

P(H|X) is posterior probability of H conditioned on X

P(X|H) is posterior probability of X conditioned on H

P(H)is prior probability of H

P(X) is prior probability of X

Naïve bayes will basically predict the output whether the patient will have chances of getting the heart disease or not.

The model dataset which we get after applying K-Means algorithm will compared the values of dataset with a trained dataset. It will apply the bayes theorem and the probability will be obtained whether the patient will have heart disease or not.[5][6]

VII. Input Attributes

1) Age
2) Gender
3) Obesity
4) Smoking
5) Electrographic result
6) Heart rate
7) Chest pain
8) Cholesterol
9) Blood pressure
10) Blood sugar

VIII. Conclusion

In this paper we are proposing heart disease prediction system using naïve bayes and k-means clustering. We are using k-means clustering for increasing the efficiency of the output. This is the most effective model to predict patients with heart disease. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy

References

[1] Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques"Department of Information Technology Malaysia University of Science and Technology Block C, Kelana Square, Jalan SS7/26 Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia .

[2] "CSV File Reading and Writing" (http:/ / docs. python. org/ library/ csv. html). . Retrieved July 24, 2011. "is no "CSV standard""

[3] Y. Shafranovich. "Common Format and MIME Type for Comma-Separated Values (CSV) Files" (http:/ / tools. ietf. org/ html/ rfc4180) Retrieved September 12, 2011.

[4] home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html "A tutorial on clustering algorithms".

[5] Shadab Adam Pattekari and Asma Parveen "Prediction System For Heart Disease Using Naïve Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[6] Mrs.G.Subbalakshmi (M.Tech), Mr. K. Ramesh M.Tech, Asst. Professor Mr. M. Chinna Rao M.Tech,(Ph.D.) Asst. Professor, "Decision Support in Heart Disease Prediction System using Naive Bayes" G.Subbalakshmi et al. / Indian Journal of Computer Science and Engineering (IJCSE)2011.

[7] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen "Association rule mining to detect factors which contribute to heart disease in males and females" Expert Systems with Applications 40 (2013) 1086–1093.

[8] Oleg Yu. Atkov (MD, PhD), Svetlana G. Gorokhova (MD, PhD), Alexandr G. Sboev (PhD), Eduard V. Generozov (PhD), Elena V. Muraseyeva (MD, PhD), Svetlana Y. Moroshkina,Nadezhda N. Cherniy "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters" Journal of Cardiology (2012) 59, 190—194.

[9] Shantakumar B.Patil Y.S.Kumaraswamy **"**Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network" European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656**.**

[10] Sivagowry, Dr. Durairaj. M2 and Persia. "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease" 2013.