International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 8; Issue 03 (G); March 2019; Page No.18001-18003
DOI: http://dx.doi.org/10.24327/ijcar.2019.18003.3433

Research Article

HEART DISEASE PREDICTION SYSTEM USING K-MEANS CLUSTERING AND NAÏVE BAYES ALGORITHM

Anjaiah A, Kannareddy Rakshith Reddy, Ramidi Ashish Kumar Reddy and Harshavardhan

Department of Computer Science Engineering, St. Peter's Engineering College, Kompally, Hyderabad, Telangana

A R T I C L E I N F O

Article History:
Received 6th December,, 2018
Received in revised form 15th January, 2019
Accepted 12th February, 2019
Published online 28th March, 2019

Key words:
Heart Disease, K-means, Naïve bayes,

A B S T R A C T

Now-a-days people work on computers hours together, finding no time to take care of their health. Due to hectic schedules and consumption of junk food, their health is being affected which leads to heart diseases. So we are implementing a heart disease prediction system using data mining technique Naïve Bayes and k-means clustering algorithms. 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 clustering algorithm and for predicting it uses naïve bayes algorithm.

INTRODUCTION

Amajor challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. [1]

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like "Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease." Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. [1] [2]

Fig 4 Data Mining Process

The diagnosis of diseases is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Providing precious services at affordable costs is a major constraint encountered by the healthcare organizations (hospitals, medical centers). Valuable

Corresponding author:* **Anjaiah A
Department of Computer Science Engineering, St. Peter's Engineering College, Kompally, Hyderabad, Telangana

quality service denotes the accurate diagnosis of patients and providing efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained. [2] [3]Besides, it is essential that the hospitals decrease the cost of clinical test. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Naive Bayes is 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. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables with immense knowledge and accurate data in that field. Large corporations invest heavily in this kind of activity to help focus attention on possible events and risks that are involved. Such work brings together all available past and current data, as a basis on which to develop reasonable expectations about the future. [4]

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 dissimilar 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. [1]

Requirements of Clustering in Data Mining

a. Deals with different types of attributes.
b. Deals with noise data
c. It requires minimum knowledge to determine input parameter.
d. Usability
e. More dimensionality

K-Means Clustering

K-means is 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 centroids 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 centroid. 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 centroid change the location step by step until no more changes are done.[1] [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 E.g.:- 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 centroid and divide it into groups. It calculates 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.

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 output 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. [7][8]

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

Block Diagram

The following block diagram represents the step by step implementation of the heart disease prediction system. 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.[9] [10]

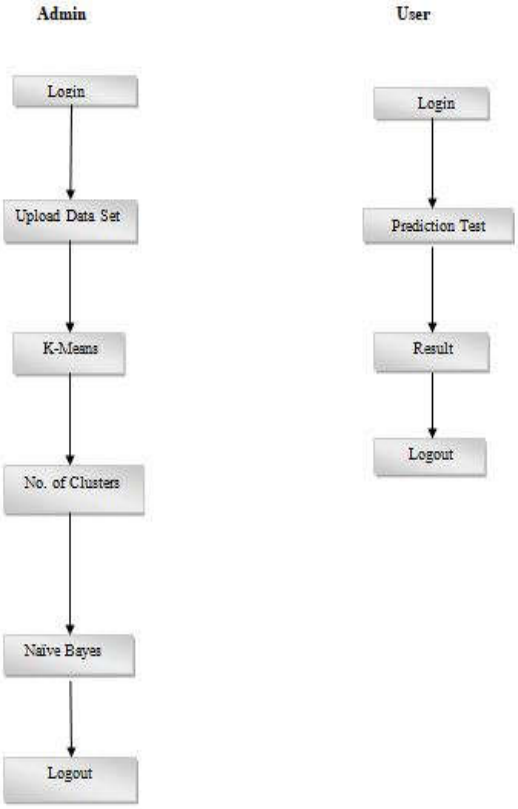

Fig 2 Block Diagram

CONCLUSION

In this paper, we proposed 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. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease.

Intelligent Heart Disease Prediction System can be further enhanced and expanded. For example, it can incorporate other medical attributes. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available in healthcare databases. Another challenge would be to integrate data mining and text mining

References

1. Shinde, R., Arjun, S., Patil, P., &Waghmare, J. (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. IJCSIT) *International Journal of Computer Science and Information Technologies,* 6(1), 637-639.
2. 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 .
3. "CSV File Reading and Writing" (http:/ / docs. python. org/ library/ csv. html). . Retrieved July 24, 2011. "is no "CSV standard"'
4. Y. Shafranovich. "Common Format and MIME Type for Comma Separated Values (CSV) Files" (http:/ / tools. ietf. org/ html/ rfc4180) Retrieved September 12, 2011.
5. Shadab Adam Pattekari and AsmaParveen "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 RaoM.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.

How to cite this article:

Anjaiah A *et al* (2019) 'Heart Disease Prediction System Using k-Means Clustering and Naïve Bayes Algorithm', *International Journal of Current Advanced Research,* 08(03), pp. 18001-18003.
DOI: http://dx.doi.org/10.24327/ijcar.2019.18003.3433