International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue
National Conference "NCPCI-2016", 19 March 2016
Available online at www.ijrat.org

Heart Disease Prediction Using Data Mining Techniques

Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir and Prof Y.K.Sharma
Vishwakarma Institute of Information Technology
Email ID:achhabbi@gmail.com,lakhanahuja.4@gmail.com

Abstract—The health industry accumulates huge amount of health related data which can be processed and used to discover hidden patterns that can be further used for taking effective decisions. Data mining techniques can be implemented to develop a rationally intelligent system that can discover such hidden patterns from the previously collected data and can answer complex query questions related to prediction of heart disease. The user is supposed to answer a set of predefined questions. The system matches the answers to the hidden patterns formed by the training dataset and then produces the result. The developed system can assist health care practitioners of level fresher to experienced and can also reduce the treatment cost.

Keywords—Data mining, Naive Bayes,Kmeans,Classification,Clustering,Heart Disease,Prediction.

I. INTRODUCTION

Motivated by the increase in the rate of mortality of heart disease patients each year, researchers have been using data mining techniques to assist health care patients. Data mining techniques convert huge amount of unprocessed data into useful information. The data is analyzed for relationships within its content that have not been discovered previously. The techniques used for prediction of heart disease in this paper are nave Bayes and improved k-means algorithm. Nave Bayes falls under a classification type of data mining technique and improved k-means comes under the clustering type of data mining technique. The logic behind combining these two clustering and classification algorithms is to make the developed system efficient in handling labeled as well as unlabeled data. By clubbing the two data mining algorithms the developed system will have better efficiency and also better accuracy in handling any type of data.

II. ALGORITHMS USED IN DEVELOPING THE HEART DISEASE PREDICTION SYSTEM

Data mining is an iterative process. The iterative process consists of following steps like Data cleaning, Data Integration, Data Selection, Data transformation, Data Mining, Pattern Evaluation. Data mining algorithms are divided into two types supervised algorithm and unsupervised algorithm. Supervised algorithm requires training and testing dataset and is capable of handling labeled data set only. Unsupervised algorithm does not require any training or testing data set and are capable of handling unlabeled data. Clustering using improved k-means: Clustering is a process of grouping the data into clusters such that similar data is located inside a cluster and dissimilar data are located outside a cluster. In this paper, we use an improved version of the traditionally used k means algorithm. In improved k-means we remove the dependency of the value k. The value k denotes the number of clusters to be formed. By removing the dependency of value k, the cluster accuracy increases and it also removes the need of domain knowledge by the user. In this paper, the clustering algorithm is used as an input to the classification algorithm i.e. nave bayes. Nave bayes Nave bayes is a classification algorithm. It requires training and testing data set. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. This algorithm is based on the bayes theorem of probability. 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.

III.DATA SETS AND INPUT ATTRIBUTES

A Dataset with medical attributes was obtained by the UCI repository. This dataset is been used to extract patterns that are significant to predict heart disease prediction. This dataset was split equally as testing dataset and training dataset.

A. Input attributes

- Age in Year
- Sex (value 1: Male; value 0: Female)
- Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic)
- Fasting Blood Sugar (value 1: ¿120 mg/dl; value 0: ¡120 mg/dl)
- Restecg resting electrographic results (value 0:normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)

- Exang - exercise induced angina (value 1: yes; value 0: no)
- Slope the slope of the peak exercise ST segment (value 1: unsloping ; value 2: flat; value 3: downsloping)
- CA number of major vessels colored by fluoroscopy
  (value 0-3)
- Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
- Trest Blood Pressure (mm Hg on admission to the hospital)
- Serum Cholesterol (mg/dl)
- Thalach maximum heart rate achieved
- Old peak ST depression induced by exercise
- Smoking (value 1: past; value 2: current; value 3: never)
- Obesity (value 1: yes; value 0: no)

IV. PROPOSED ALGORITHM

The modified K-means algorithm which does not require number of clusters (K) as input is proposed below. In this algorithm two clusters are created initially by choosing two initial centroids which are farthest apart in the data set. This is done so that in the initial step itself we can create two clusters with the data members, which are the most dissimilar ones.

A. Input

D: The set of n tuples with attributes A1, A2, . . . , Am where m = no. of attributes. All attributes are numeric

B. Output

Suitable number of clusters with n tuples distributed properly.

C. Method

Method:Compute sum of the attribute values of each tuple (to find the points in the data set which are farthest apart) Take tuples with minimum and maximum values of the sum as initial centroids. Create initial partitions (clusters) using Euclidean distance between every tuple and the initial centroids. Find distance of every tuple from the centroid in both the initial partitions. Take d=minimum of all distances. (other than zero) Compute new means (centroids) for the partitions created in step 3. Compute Euclidean distance of every tuple from the new means (cluster centers) and fmd the outliers depending on the following objective function: If Distance of the tuple from the cluster mean¡d then not an Outlier. Compute new Centroids of the clusters. Calculate Euclidean distance of every outlier from the new cluster centroids and find the outliers not satisfying the objective function in step 6. Let B=Y1,Y2,. ....Y p be the set of outliers obtained in step 8 (value of k depends on number of outliers).

Create a new cluster for the set B, by taking mean value of its members as centroid.

- Find the outliers of this cluster, depending on the objective function in step 6.
- If no. of outliers = p then

  ◦Create a new cluster with one of the outliers as its member and test every other outlier for the objective function as in step 6. ◦Find the outliers if any.
- Calculate the distance of every outlier from the Centroid of the existing clusters and adjust the outliers in the existing which satisfy the objective function in step 6.
- B = Z1, Z2 .... Zq be the new set of outliers.(value of q depends on number of outliers)

V. CONCLUSION

In this paper we are proposing heart disease prediction system using nave bayes and Improved k-means clustering. We are using Improved 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.

ACKNOWLEDGMENT

We would like to take this opportunity to thank our guide Prof. Y.K.Sharma for giving us all the help and guidance we needed. We are really grateful to him for his kind support. His valuable suggestions were very helpful. We are also grateful to Dr.S.R.Sakhare, Head of Computer Engineering Department, VIIT for his indispensable support, suggestions. In the end our special thanks to VIIT for providing various resources such as laboratory with all needed software platform and continuous Internet connection.

REFERENCES

[1] Mrs G.Subbalakshmi, Decision Support in Heart Disease Prediction System using Naive Bayes, Indian Journal of Computer Science and Engineering.

[2] Shadab Adam Pattekari and AsmaParveen, Prediction System for Heart Disease using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences.

[3] Mai Shouman, Tim Turner and Rob Stocker ,Integrating Naive Bayes And K-means Clustering With Different Initial Centroid Selection Methods in The Diagnosis of Heart Disease Patients

[4] SellappanPalaniappan, RafiahAwang, Intelligent Heart Disease Prediction System Using Data Mining Techniques

[5] K.Rajalakshmi, Dr.S.S.Dhenakaran, N.Roobini, Comparative

a. Analysis of KMeansAlgorithm in DiseasePrediction,
b. International Journal of Science, Engineering and Technology Research (IJSETR)

[6] R. Chitra and V. Seenivasagam, Review of heart disease prediction system using data mining and hybrid intelligent techniques, ICTACT journal on soft computing.

[7] Sivagowry, Dr.Durairaj and Persia, An Empirical Study on applying Data MiningTechniques for the Analysis and Prediction of Heart Disease,International conference on information communication and embedded systems.

[8] Anupama Chadha, An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K,International Conference on Reliability, Optimization and Information Tec.