NOVATEUR PUBLICATIONS
International Journal of Research Publications in Engineering and Technology [IJRPET]
ISSN: 2454-7875
VOLUME 3, ISSUE 6, Jun.-2017

HEART DISEASE PREDICTION USING NAIVE BAYES AND K-MEANS TECHNIQUES

MS.MEHDI KHUNDMIR ILIYAS
Department of Computer Science, M. E. Student, K.J College of Engineering and Management Research Pune, India
mehdiaims@gmail.com

PROF. VIKAS MARAL
Department of Computer Science K.J College of Engineering and Management Research Pune, India
vikasmaral@gmail.com

ABSTRACT:

Data mining techniques are started gaining its popularity nearly three decades ago. Till last few years data mining approach was not in been used in health care organization. Researchers have started paying attention towards this field, it's been found by the researcher health care sector is possessing very large volume of data but all this are highly unorganized. If this organized in a proper way using data mining technique it can be easily use for the prediction of various diseases. In this paper author has restricted to heart diseases only. In this paper author has developed a hybrid approach by using two technique Naïve Bayes and K – means algorithm and Hadoop technology. Different 14 parameters are considered by the author for prediction of the heart disease.

KEYWORDS: Data mining, Heart Disease, Hybrid approach, K-means, Naive Bayes, Hadoop etc.

I. INTRODUCTION:

DATA MINING:

Data mining technique widely used for computational and discovering patterns in large data sets. Data mining approach was found by researchers in the middle of 90's, and its been observed that it is very important technique for fetching unknowns patterns and vital information from large data set. In the literature it's been observed that if proper tool is made we can fetch a correct data or the pattern from the large database and can be used for decision making [1].

Many researcher assumes that data mining and knowledge discovery in database are same terms and they use it interchangeably, but many practitioners assumes that this are different terms and data mining is one stage for knowledge discovery in database(KDD). Fig.1. shows the process chart for the complete knowledge discovery in database.

APPLICATION OF DATA MINING INHEALTH SCIENCE:

1. Effective management of Hospital resource
2. Hospital Ranking
3. Better patient relation
4. Hospital Infection Control
5. Smarter Treatment Techniques
6. Improved Patient care
7. Decrease Insurance Fraud
8. Recognize High-Risk Patients
9. Health Policy Planning
10. Disease Prediction for future.

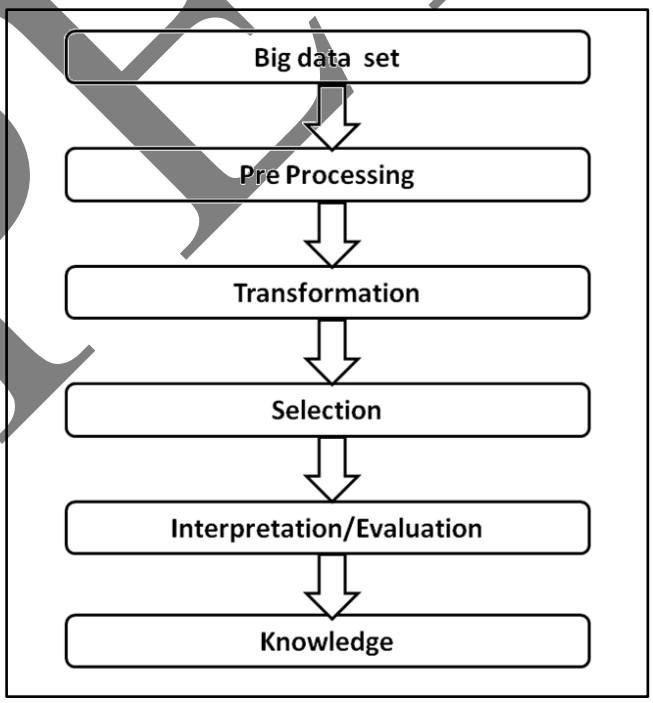

Fig.1. Process chart for the knowledge discovery in database

K MEANS ALGORITHM:

K means technique, traditionally used for the 'vector quantization', extensive use of this technique is found in signal processing. But after 20th century K-Means algorithm is extensively used in data mining technique for making partition of 'n' observation into various clusters. e.g.

Input

Given set of observations = $X_1, X_2, X3 \ldots\ldots X_n$

Where

Each observation is d-dimensional real vector.

Output
$S=S_1, S2, S3....S_K$
Where
Partitions with N observations are divided into k sets (k, n) for minimizing within cluster sum of squares.

K-MEAN ALGORITHM HAS FOLLOWING STEPS:

Let $X = \{x_1, x2, x3,.....,x_n\}$ be the set of data points and $V = \{v_1, v2,.....,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster
Centers.
3) Assign the data point to the cluster center whose distance
from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{j=1}^{c_i} x_i$$

Where, 'ci' represents the number of data points in ith cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3.

ADVANTAGES:

1. Very fast, Robust and Reliable
2. Easy to understand
3. produces better results when data set are distinct.

NAIVE BAYES:

Naive Bayes technique is based on the 'Bayes Theorem'. Bayes theorem describes the probability of an event based on previous conditions which are relating to particular event. e.g. if Cardiovascular disease (CVD) are related to cholesterol level, then by using Bayes theorem, cholesterol level of person can be used more accurately for assessing the probability of CVD's. In machine learning, Naïve Bayes classifiers are treated as trouble free probabilistic classifiers and it is independent from assumptions between the features.

HADOOP:

Hadoop is reframing the computer science technology in modern generation. Basically it is a open source software, based on Java programming. Framework developed handling of extremely large data specifically in distributed computing environment. Hadoop is developed by the apache software foundation. By using the Hadoop we can run the application on thousands of commodity hardware nodes and it can easily handle thousands of terabytes. Hadoop uses a distributed file system, it facilitates the rapid transfer of data amongst nodes.

Now a days patients data is increasing day by day and we have to store that huge data and analyze it,for that purpose huge database systems is required . Hadoop is preferred for huge data.

II. IMPLEMENTED SYSTEM:

The data is taken from UCI repository (i.e The University of California, Irvine) where four databases are available I.e Cleveland, Hungary , Switzerland and the VA Long Beach.

These four databases is taken as input file for training the system. These four databases has following instances or records.

Table.1. Database instances used

Sr.No	Database	instances
1	Cleveland	303
2	Hungarian	294
3	Switzerland	123
4	Long Beach VA	200

It contains 76 attributes out of which 14 attributes are considered which are important as follows.

Table.2. Attributes used

Sr.No	Attribute	Description
1	Age	Age in years
2	Gender	1=male, 0=female
3	Cp	Chest pain type (1 = typical angina 1, 2= typical angina, 3 = non-angina Pain, 4 = asymptomatic).
4	Trestbps	Resting blood pressure (in mm Hg on admission to hospital).
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (1=true, 0=false).
7	Restecg	Resting electrocardiographic results (0= normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy).
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1=up sloping, 2=flat, 3= down sloping)
12	Ca	Number of major vessels colour by Fluoroscopy
13	Thal	Value 3:normal, 6:fixed defect, value7:reversible defect
14	Heart Disease	Value 0 to 3

SYSTEM DESCRIPTION:

Prediction of the heart disease is very complicated task, and in current world it mainly depends upon the individual medical practitioner. If all individual medical practitioners are combined on one data set it will be very useful for younger generation of the medical practitioner and ultimately it will helpful to the people. In this paper for heart attack prediction hybrid approach is been used, combination of the most popular clustering technique called 'K-Means' and as a Classifier 'Naive Bayes' algorithm are used. Because of hybrid approach this technique is most suitable for any complex problem and it produces results with very good accuracy; also it gives flexibility to work with Hadoop. Fig.2. is detailed block diagram of the proposed system.

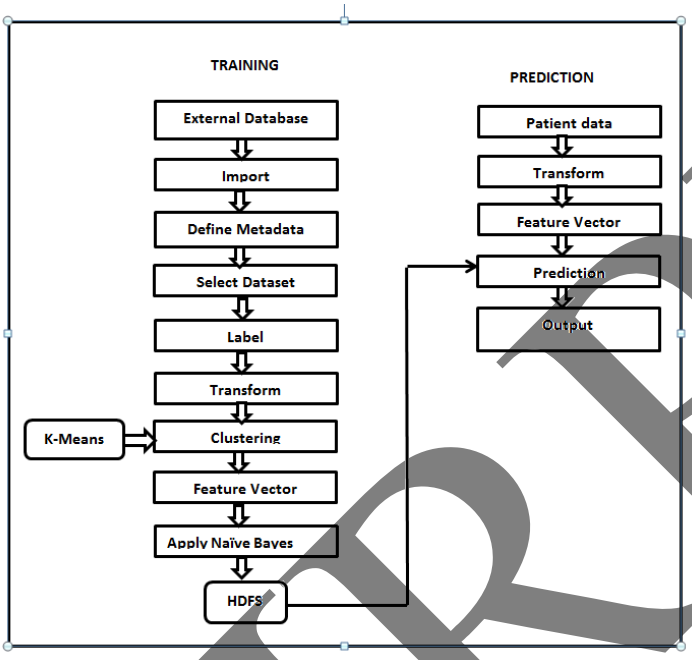

Fig. 2. Block diagram of the proposed system.

In implemented system data is taken from UCI repository. Where four different databases with 76 attributes are available. Here 14 attributes are taken which are important. e.g. age, gender,chest pain , cholesterol , Thalach etc. These database is uploaded to designed software.

Once import is done to newly designed software tag is assigned to each attribute. Then we have to select any one column for clustering. The software applies K-Mean clustering on selected column. It asks user how many clusters have to form, then by using K-Means clusters are formed and centroids are identified. After this training is given to system .where Naïve Bayes is used to find the initial probability and data is stored inHadoop database.

In the next phase prediction of heart disease take place. Where individual patient data is taken as a input. Then the system predict weather the patient have heart disease or not. The sample data set is shown in table 2.

Table.3. Sample Data Set

Age	Gender	Chest Pain	Trestbps	Chol	Fbs	Resteg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Heart Disease
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
49	1	2	130	266	0	0	171	0	0.6	1	0	3	0

ALGORITHM:

1. Start.
2. Take data from external database.
3. Upload CSV file as input data.
4. Define Metadata.
5. Take dataset.
6. Label each dataset.
7. Select any one dataset for clustering.
8. Apply K-mean clustering.
9. Ask how many cluster to form.
10. Identify Centroids .
11. Form clusters.
12. Use Naïve Bayes for Initial Probability.
13. Train the system.
14. Store data in Hadoop database.
15. Start prediction phase.
16. Enter patient data.
17. Transform and find feature vector.
18. Mining take place and predict heart disease.
19. Result show patient have heart disease or not.
20. Stop

DEVELOPED SOFTWARE

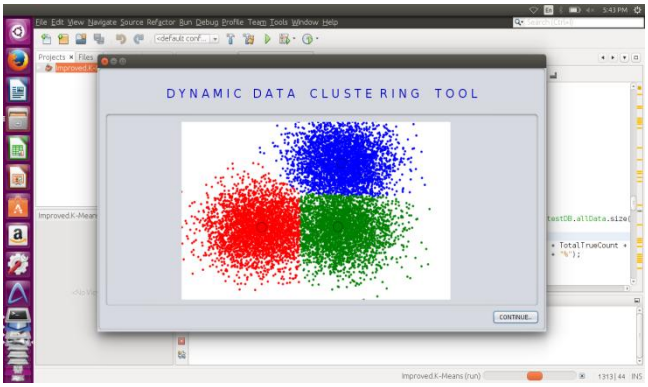

Fig.3. Welcome Screen of developed software

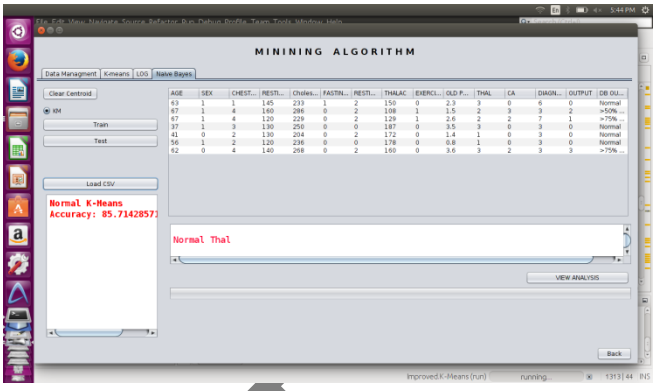

Fig. 7.Prediction of Heart Disease.

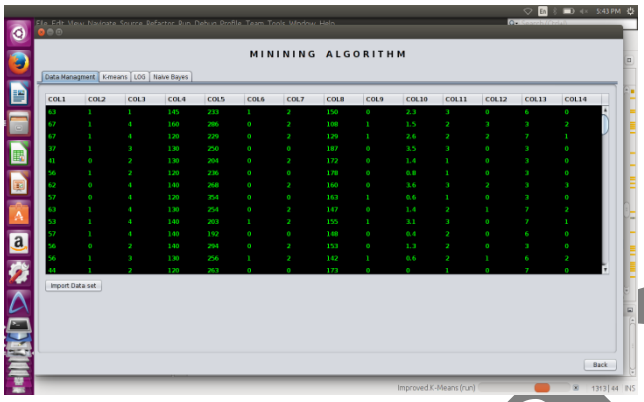

Fig. 4. Data Loading

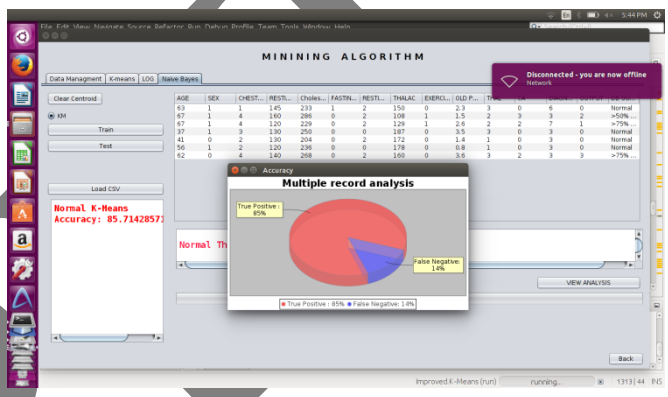

Fig. 8. Final accuracy

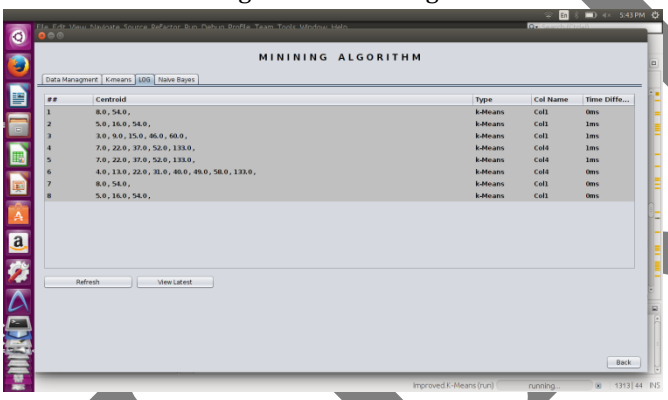

Fig.5 K-Mean Clustering

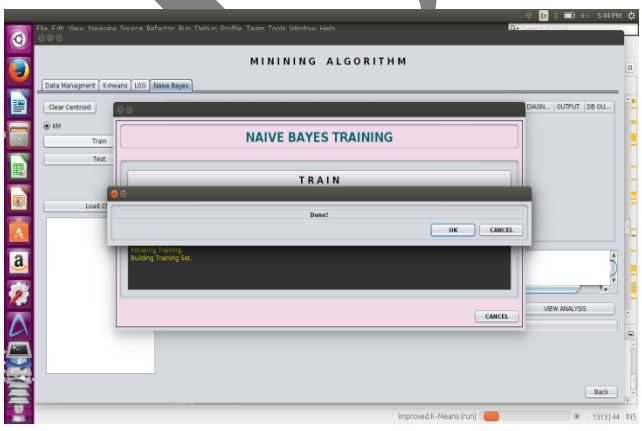

Fig. 6. Training Screen of developed software

III. RESULT DISCUSSIONS:

The result shows that whether the patient have heart disease or not. Here individual patient or list of patient data is given as input for prediction of heart disease. This system is hybrid approach which give prediction accuracy of 85% true positive and 14% false negative.

IV. CONCLUSION

Prediction of the heart disease system is developed by combining Naïve Bayes and K-Means algorithm in conjunction with Hadoop. The develop software haul out the knowledge from historical database made by medical practitioner or the clinical care units. In first step training of data set is needed then this data set is validated against previously available training dataset for true working of the developed software. As developed approach in hybrid it produces accuracy of order of 85%.

REFERENCES:

1) D. Hand, H. Mannila and P. Smyth, *"Principles of data,* MIT, (2001).
2) AkashJarad, Rohit Katkar, Abdul Rehaman Shaikh, Anup Salve, *International Journal of Emerging Trends & Technology in Computer Science* (IJETTCS) , January-February 2015.

3) Sellappan Palaniappan, Rafiah Awang , *Intelligent Heart Disease Prediction System Using Data Mining Techniques* , IEEE 2008.
4) Nidhi Bhatla KiranJyoti, *An Analysis of Heart Disease Prediction using Different Data Mining Techniques*, International Journal of Engineering Research and Technology (IJERT), 2012.
5) R. Thanigaivel, Dr. K. Ramesh Kumar*, Review on Heart Disease Prediction System using Data Mining Techniques*", Asian Journal of Computer Science and Technology (AJCST).
6) Shadab Adam Pattekari and Asma Parveen, *Prediction system for heart disease using Naive bayes*, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624.
7) Carlos Ordonez, Edward Omiecinski, *Mining Constrained Association Rules to Predict Heart Disease,* IEEE, Published in International Conference on Data Mining (ICDM), p. 433- 440, 2001.
8) Ms. Ishtake S.H, Prof. Sanap S. A., *Intelligent Heart Disease Prediction System Using Data Mining Techniques,* International J. of Healthcare & Biomedical Research,2013.
9) Rishi Dubey, Santosh chandrakar *Review on Hybrid Data Mining Techniques for The Diagnosis of Heart Diseases in Medical Ground Indian Journal Of Applied Research* , August2015.
10) G. Purusothaman , P. Krishnakumari , *A Survey of Data Mining Techniques on Risk Prediction: Heart Disease* , Indian Journal of Science and Technology , June 2015.
11) Mrs. G. Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao *Decision Support in Heart Disease Prediction System using Nave Bayes* . Indian Journal of Computer Science and Engineering (IJCSE)2011.
12) BalaSundar V, *Development of Data Clustering Algorithm for predicting Heart*, IJCA, Vol 48(7), June 2012.