Predicting Heart Disease Symptoms using Fuzzy C-Means Clustering

Lovepreet Kaur

Assistant Professor in Punjab Institute of Technology, Nandgarh

Abstract — The diagnosis of heart disease from various symptoms is a major issue which is not free from false presumptions. The healthcare industry gathers large amount of heart disease data that is not mined to determine the useful information for effective decision making by healthcare practitioners. The effort to utilize knowledge and experience of specialists and data of patients collected in database is a valuable option .Data Mining using a variety of techniques for decision making knowledge in the database and extracting these in a way that they can use in areas such as decision support, predictions, estimation. This research will provide an intelligent heart disease prediction system (IHDPS) able to help a physician as well as a health care system. In this research, the efficiency of heart disease system will enhance using if then else rules classification, fuzzy c means clustering and genetic algorithm. Various parameters like accuracy, time, specificity and sensitivity are calculated. The proposed algorithm provides better accuracy as compared to traditional algorithms.

Keywords: Data Mining, Disease Diagnosis, Heart Disease, classification using if then else rules, Fuzzy c means clustering, genetic algorithm.

I. Introduction

Predicting the outcome of disease is one of the most interesting and challenging task in data mining. The knowledge discovery database (KDD) process includes data mining techniques has become a popular research tool for medical researchers and it is able to predict the outcome of a disease using historical data records of patients. Firstly the heart disease was thought to be the problem of developed countries but now it is problem for developing countries too. Recent studies have documented that high cardiovascular diseases are reported in coal mining regions .The risk for cardiovascular disease is influenced by environmental, behavioural, health services variables. The automation of the system would be extremely beneficial for us. Regrettably all the doctors do not posses experience in every sub speciality and more ever there is shortage of resource persons at certain places [6]. Appropriate computer based information and decision system can aid in achieving test at the reduced cost. Today most countries face high and increasing rate of heart disease. It has become a leading cause of death. Due to increase in world's population, the health care industries are facing many challenges and issues based on patient's severity is to be reduced and detect it earlier in more efficient way. To save the life of patients and reduce the health care cost the medical error should be prevented [14].

Table 4 Description of attributes

S.NO.	Attribute	Description
1	Age	Age of patient in years
2	Sex	Male, female
3	Cp	Chest Pain Type 1.Angina=Typical Angina 2.Abnang=atypical Angina 3.Notang=Non Anginal Pain 4.Asympt-asymtomatic
4	Trestbps	Resting Blood pressure(in mm Hg on admission to the hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar(fbs>120=true, fbs<120=false)
7	Restecg	Resting electrocardiographic results are Hyper,Normal Abnm(abnormal)
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina(yes=true, No=false)
10	slope	The slope of the peak exercise ST segment,Up, Flat, down
11	Family history	True, false
12	ca	Number of major vessels(0-3) colored by fluoroscopy
13	No. cigrates per day	No. of cigrates per day
14	No. of years a smoker	Number of years a smoker

ISSN: 2278 – 1323

Heart Disease: The term heart disease encompasses the diverse disease that affects the heart. Before defining heart disease it is better to define what a heart is and functions performed by heart. Heart is one of the most important organs in our body. Essentially a pump, the heart is a muscle made up of four chambers separated by valves and divided into four chambers called an atrium and one called a ventricle. The arteries collect blood and the ventricles contract to push blood out of the heart. The right half of the heart pumps oxygen poor blood to the lungs where blood cells can obtain more oxygen. Then newly oxygenated blood travels from the lungs into the left atrium and left ventricle pumps the newly oxygen rich blood to the organs and tissues of the body. This oxygen provides body with energy and is essential to keep our body healthy.

Heart disease is a general name applies to number of illness, disorders, conditions that effect the circulatory system which consist of heart and blood vessels[4].Symptoms of heart disease depend on the specific type of heart disease. Classic symptom of heart disease is chest pain. Chest pain arises when the blood received by heart muscle is inadequate [4]. Sometimes there may be no symptoms in some people until life threatening complications occur.

Early signs of heart disease are dizzy spell, discomfort following meals especially if long continued, shortness of breath, fatigue, pain or tightness in chest, palpitation etc[4]. There is common set of risk factors that influence whether someone will ultimately be at risk for heart disease or not. These risk factors include age, gender, cholesterol, obesity, sedentary life style etc.

Coronary heart disease: Most frequent type of heart disease is coronary heart disease. It is the main cause of the heart attack. Coronary heart disease occurs when oxygen and blood supply to the heart is decreased. A sudden blockage of a coronary artery generally due to a blood clot results in a heart attack[4].There are few factors that are responsible for coronary artery disease high cholesterol that can increase fat concentration in our blood create build up of fatty deposits.

Fuzzy c means clustering: In hard clustering data is divided into distinct clusters where each data element can belong to one clusters like k means clustering. In fuzzy clustering data is divided into distinct clusters, one data element can belong to one or more clusters. Fuzzy c means clustering is also known as soft clustering. In fuzzy c means clustering each element is associated with set of membership level.

- Unlike K-means clustering where each data point belongs to exactly one cluster but in fuzzy c means data element can belong to more than one cluster.

- Gives best result for overlapped dataset and comparatively better than K-means algorithm.

II. Research Methodology

Diagnostic test are used to determine the presence or absence of disease. Confusion matrix is the primary source of measurement in classification problems. Classification instance classifies each instance into two classes either positive or negative. This gives four classifications for each instance: a true positive, a true negative, a false positive, a false negative. Given m classes ,a confusion matrix is a table of at least size m by m .When peoples are tested for a disease the test outcome can be positive(sick) or negative (healthy) while the actual status of the patient may be different then following four conditions may occur.

Accuracy: The overall accuracy of a classifier is estimated by dividing the total number of correctly classified instances by the total number of instances [26]

Accuracy = (TP+TN) / (TP+TN+FN+FP)

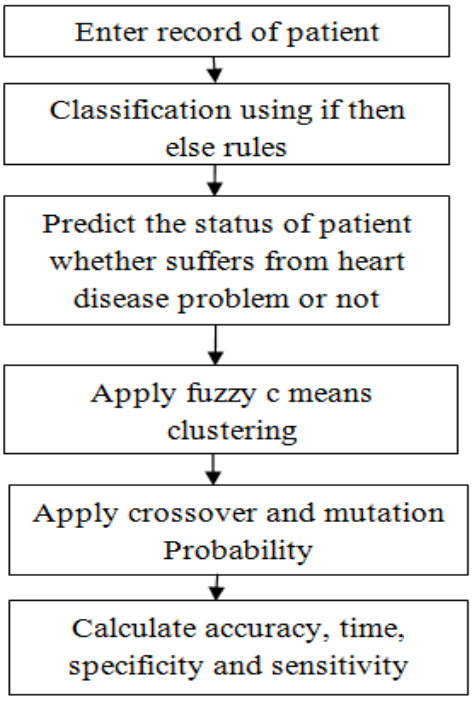

Figure 1.1: Flow of proposed work

In this research Slope, Exercise induced angina and no. Of major vessels coloured by fluoroscopy are three main factors after testing these factors other factors are considered.

[1] Healthy people correctly identified as healthy called "True negative (TN)".
[2] Healthy people wrongly identified as sick called "false positive (FP)".
[3] Sick people correctly identified as sick called true "positive (TP)".
[4] Sick people wrongly diagnosed as healthy called "false negative (FN)".

Table 1 Confusion Matrix

TEST RESULT	ACTUAL CONDITION		
	PRESENT		ABSENT
	POSITIVE	CONDITION PRESENT+POSITIVE RESULT=TRUE POSITIVE	CONDITION ABSENT+POSITIVE RESULT=FALSE POSITIVE
	NEGATI-VE	CONDITION PRESENT+NEGATIVE RESULT=FALSE NEGATIVE	CONDITION ABSENT+NEGATIVE RESULT=TRUE NEGATIVE

ISSN: 2278 – 1323

The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. Accuracy, specificity and sensitivity are the basic performance measurements.

Sensitivity: Sensitivity refers to the proportion of people with disease who have a positive test result. Specificity is ratio of number of true Positives to the number of true positives plus number of false negatives.

$$\text{Sensitivity} = TP / (TP + FN)$$

Specificity: Specificity refers to the proportion of people without the disease who have a negative test result. Specificity is the ratio of number of true negatives to the number of number of true negatives plus number of false positives [26].

$$\text{Specificity} = TN / (TN + FP)$$

III. Result

Experimental results deals with the output. The output result of proposed algorithm i.e. applied on Cleveland heart disease dataset is shown. Cleveland database has 303 records but in this dissertation only 58 records are used. Missing value records are discarded. In order to analyze the performance comparison will be made with existing algorithms.

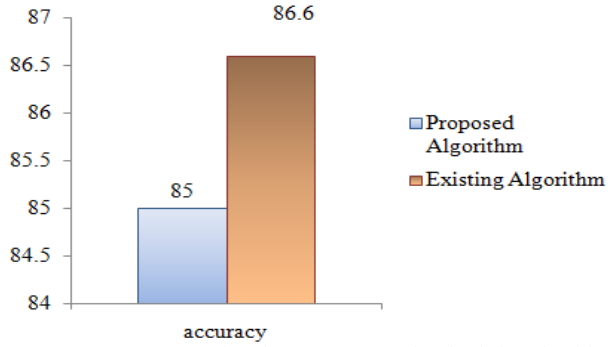

Fig 1.2 Accuracy comparison between proposed and existing algorithm

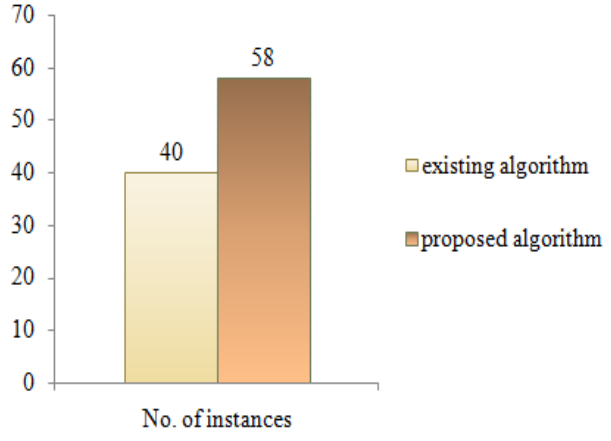

Fig 1.3 No. of instances used in existing algorithm and proposed algorithm

IV. Conclusions

In this research work performance of heart disease prediction system is evaluated. Data is collected from Cleveland heart disease dataset. This dataset contains total 303 instances and 76 attributes. In this research work 58 records and 14 attributes are used. Data mining Classification using if then else rules, fuzzy c means clustering, genetic algorithms are used in this research work to improve the accuracy of the system. The performance of the heart disease prediction system is evaluated using performance measures accuracy, Time, specificity, Sensitivity etc. The prediction system shows 86.6% accuracy, 32 milliseconds time, 0.44 specificity,0.45 sensitivity.

V. Future Work

In future to increase the performance of heart disease prediction system more datasets can be used. Sensitivity and specificity can be further improved. In future intelligent heart disease prediction system can be build that can work on specific type of heart disease.

VI. References

[1]My Chau Tu,Dongil Shin,Dongkyoo Shin "*Effective Diagnosis of Heart Disease Through Bagging Approach*" 2nd International conference on Biomedical Engineering and Informatics,2009,pages(s):1-4,IEEE

[2]Farhad Soleimanian Gharehchopogh,Zeynab Abbasi Khalifelu "*Neural Network Application in Diagnosis of Patient: A Case Study*" International conference on computer networks and information Technology(ICCNIT),2011,Page(s):245-249,IEEE

[3]Sellappan Palaniappan ,Rafiah Awang "*Intelligent Heart Disease Prediction System Using Data Mining Techniques*" International Conference on computer systems and applications,AICCSA 2008,Page(s):108-115,IEEE

[4]V.V.Jaya Rama krishniah,D.V.Chandra Sekar.K.Ramchand H Rao "*Predicting the Heart Attack Symptoms Using Biomedical Data Mining Techniques*"The International Journal of Computer Science and Applications(TIJCSA)Volume 1,No.3,May 2012 ISSN-2278-1080

[5]K.Srinivas,Dr.G.Raghavendra Rao,Dr.A.Govardhan "*Analysis of Coronary Heart Disease and Prediction of Heart Attack in coal Mining Regions using Data Mining Techniques*" The 5th International Conference on computer science and education ,China.August 24-27,2010,IEEE

[6]Jyoti Soni,Uzma Ansari,Dipesh Sharma,Sunita Soni "*Intelligent and Effective Heart Disaese Prediction System Using Weighted Associative Classifiers"*International Journal on computer science and Engineering(IJCSE),Volume no. 3,6 June 2011,ISSN:0975-3397

ISSN: 2278 – 1323

[7]Shantakumar B.Patil,Dr.Y.S.Kumaraswamy "*Extraction of Significant patterns from heart Disaese Warehouse For Heart Attack Prediction*"IJCSNS International Journal Of Computer science and network security,Vol.9 No.2,February 2009

[8]Jyoti Soni,Ujma Ansari, Dipesh Sharma "*Predictive Data Mining For Medical Diagnosis: An Overview of Heart Disease Prediction*"International Journal of Computer applications(0975-8887) Volume 17- No 8,March 2011

[9]Mrs.G.Subbalakshmi,Mr.K.Ramesh ,Mr.M.Chinna Rao "*Decision support in Heart Disease Prediction System Using Naïve Bayes*" Indian Journal of computer science and engineering(IJCSE),Vol.No.2 Apr-May 2011,ISSN:0976-5166

[10]Nidhi Bhatla,Kiran Jyoti "*An Analysis Of Heart Disease Prediction Using Different Data Mining Techniques*" International journal of engineering Research and Technology(IJERT),ISSN:2278-0181,Vol.1 Issue 8,October-2012

[11]Shashikant Ghumbre,Chetn Patil and Ashok Ghatol "*Heart Disease Diagnosis Using Support Vector Machine*" International Conference on computer science and information Technology(ICCSIT 2011)Pattaya Dec.2011

[12]R.Chitra and Dr.V.Seenivasagam "*Heart Disease Prediction System Using Supervised Learning Classifier*" Bonfring International Journal of Software Engineering and Soft Computing,Vol.3,No.1,March 2013

[13]G.Parthiban,S.K.Srivatsa "*Applying machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients*" International Journal of Applied Information Systems(IJAIS)-ISSN:2249-0868,Volume 3-No.7,August 2012

[14]A.Sudha,P.Gayathri,N.Jaisankar "*Utilization of data mining approaches for prediction of life threatening diseases survivability*" International Journal of computer applications(0975-8887)Volume 41-No.17,March 2012

[15]Roohallah Alizadehsani,Jafar Habibi,Behdad Bahadorian,Hoda Mashayekhi,Asma Ghandeharioun,Reihane Boghrati,Zahra Alizadeh Sani "*Diagnosis of Coronary Arteries Stenosis Using Data mining*"JMMSS/July 2012,Vol2,No.3

[16]Sangeeta Goele,Nisha Chanana "*Data Mining Trends in past,current and future*"International journal of computing and business Research,ISSN(online):2229-6166

[17]Usama M.Fayyad "*Data Mining and knowledge discovery in databases:Applications in Astronomy and Planetary science*" AAA-96 Proceedings.copyright© 1996,AAAI

[18] Usma Fayad, Gregory Piatetsky-Shapiro and Padhraic Smyth "*The KDD process useful for extracting useful Knowledge from large volumes of data".*

[19]Gonzalo Mariscal,Oscar Marban and Covadonga Fernandez "*A Survey of data mining and knowledge discovery process models and methodologie" the knowledge engineeringreview".*,vol.25:2,137-166& Cambridge University press,2010 doi:10.1017/SO269888910000032.

[20] Fadzilah Siraj and Mansour Ali Abdoulha "*Mining Enrolment Data Using Predictive and descriptive approaches*"

[21]Daniel T Larose "*Discovering knowledge in Data*", An Introduction to data mining,Wiley

[22] M.A. Chatti, A.L. Dyckhoff, U. Schroeder, and H. Thüs "*A Reference model for learning analysis" International Journal of Technology Enhanced Learning (IJTEL) – Special Issue on "State-of-the-Art in TEL*"

[23]Michael J.A.Berry,Gorden S.Linoff, "*Data Mining Techniques for Markiting,sales and customer relationship Management*" Second Edition,Wiley publishing,Inc.

[24] Dr.Yashpal Singh,Alok Singh Chauhan "*Neural Network in Data Mining*" Journal of Theoretical and Applied Information Technology,2005 - 2009 JATIT

[25] Mrs. Bharati R.Jipkate and Dr.Mrs.V.V.Gohokar "*A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms*" International Journal of computation engineering research,ISSN:2250-3005

[26] Devashish Sharma, U.B. Yadav, Pulak Sharma "*The concept of specificity and sensitivity in relation to two types of errors and its application in medical research*" Journal of Reliability and Statistical Studies (ISSN: 0974-8024) Vol. 2, Issue 2(2009): 53-58

[27] E.W.T.Ngai, Li Xiu, D.C.K.Chau "*Application of data mining techniques in customer relationship management: A literature review and classification*" Expert Systems with Applications 36 (2009) 2592–2602

ISSN: 2278 – 1323