FUZZY DATA MINING FOR HEART DISEASE DIAGNOSIS

S.Jayasudha

Department of Mathematics
Prince Shri Venkateswara Padmavathy Engineering College, Chennai
.

ABSTRACT:

We address the problem of having rigid values for attributes in the critical data set like heart disease data set. Here we propose the process of fuzzifying the crisp values by changing the appropriate intervals into linguistic variables. In this paper, we demonstrate the effect of such a process the mining tool WEKA. We obtain the error rate and accuracy by building the familiar learning model.

Keywords: Fuzzy Logic, WEKA, Classifiers, Absolute Error and Mean Absolute Error.

INTRODUCTION:

Heart disease describes a range of conditions that affect the heart. Cardio Vascular disease includes all the diseases of the heart and circulation including coronary heart disease, angina, heart attack, congenital heart disease and Stroke. Cardiovascular disease is a class of diseases that involve the heart or blood vessels.

A heart attack occurs when the blood flow to a part of the heart is blocked by a blood clot. Coronary heart disease (angina and heart attack) and stroke may be caused by the same problem. A risk factor is something that increases of getting a disease [1]. There are several risk factors for CVD, including smoking, high blood pressure, high blood cholesterol, being physically inactive, being overweight or obese, diabetes and family history of heart disease.

The aim of this paper is to study the pattern of analysis of data mining techniques applied in WEKA. The results of the classification techniques applied through this tool is summarized and gives the best one or the optimal one is suggested for our real world data analytics. Also the number of attributes needed for analysis can be reduced. Here we discuss about the importance of fuzzy logic in data mining in section I. WEKA tool analyzes nearly 31 instances collected from various clinical laboratories with 8 attributes in section II. In section III, We describe the ranges of attribute and statistical description for each attribute. In section IV, We present the comparison of different classification techniques using WEKA from the experimental results. In section V shows how the result analysis to evaluate the performance of different classification to produce the best accuracy and least error rate to diagnosis the data set. We conclude that in section VI.

1. FUZZY LOGIC IN DATA MINING:

Data mining is otherwise called as knowledge discovery in database which is a research area that considers the analysis of large databases in order to identifying valid, useful, meaningful unknown and unexpected relationships. Various techniques can be applied in data mining [4]. Fuzzy plays an important role in data mining analysis [2, 3]. Fuzzy set in data mining is a novel method because it gives the membership of data in a set by three measures by similarity, preference and Uncertainty.

Fuzzy logic is rule based one and results in IF- THEN rules which is our main motive in this paper[9,11].

II. DESCRIPTION OF WEKA TOOL:

WEKA represents the Waikato Environment for Knowledge Analysis. WEKA tool is to apply a dataset and analyze its output to extraction information about the data. The main aim of WEKA is a classifier and filter algorithms. We can perform preprocessing and classification in WEKA using different types of classifiers such as RIDOR, JRIP, ZERO R, PART and ONE R [8,9]. WEKA having facility to convert the numerical data into CSV format from using various parameters using different data sets. 10 fold cross validation is used for evaluation [10,11]. WEKA describes a list of instances sharing a set of attributes and statistical description is given clearly in Section III. WEKA is a very good tool used for solving various purposes of data mining. There are four WEKA application interfaces: explorer, experimenter, knowledge flow and simple command line.

III. STATISTICAL DESCRIPTION:

Here we introduce 31 instances and the number of attributes is 8. Input attributes are chest pain, blood pressure, serum cholesterol, number of years as a smoker, fasting blood sugar, heart beat rate and resting blood rate. The output attribute is the angiographic disease status of heart of patients [1] . The attributes are described in the following Table 1.

Table 1 shows the Ranges of Attributes:

S.No.	INPUT	RANGE	FUZZY VALUE
1.	Chest Pain		Typical Angina
			Atypical Angina
			Nontypical Angina
			Asymptomatic Angina
2.	Blood Pressure	From 110 - 145	Mininmum
		From 145 Above	Maximum
3.	Serum Cholesterol	From 168 -240	Minimum Level
		From 240 Above	Maximum Level
4.	Smoking Habit (Years)	From 0 -30	Low Possibility
		From 30 Above	High Possibility
5.	Fasting Blood Sugar	From 60 - 120	No Sugar Level
		From 120 Above	Yes Sugar Level
6.	Maximum Heart Rate	From 50- 70	Minimum Heart Rate
		From 70 Above	Maximum Heart Rate
7.	Resting Blood Rate	From 90 – 140	Low Blood Rate
		From 140 Above	High Blood Rate
	OUTPUT		
8.	Angiographic disease status (diameter of coronary arteries)	Less than 50%	Mild
		More than 50%	Massive

IV Classification Accuracy:

It is the ability to predict categorical class labels. This is the simplest scoring measure. It calculates the proportion of correctly classified instances.

Accuracy = (Instances Correctly Classified/Total Number of Instances) *100

If the instance is positive and it is classified as positive. False Negative (FP): If the instance is positive but it is classified as negative. True Negative (TN) : If the instance is negative and it is classified as negative[5,6]. False Positive (FP): If the instance is negative but it is classified as positive.ROC(Receiver Operating Characteristics)

Area is a traditional to plot the same information in a normalized form with 1- false negative rate plotted against the

Classifier	Phase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Zero R	Cross validation	0.742	0.742	0.55	0.742	0.632	0.351
ONE R	Cross validation	0.55	0.817	0.516	0.55	0.532	0.367
PART	Cross validation	0.871	0.126	0.887	0.871	0.875	0.848
JRIP	Cross validation	0.806	0.149	0.853	0.806	0.817	0.832
RIDOR	Cross validation	0.742	0.497	0.723	0.742	0.729	0.622

false positive rate. Precision is the proportion of relevant documents in the results returned.
Mean absolute error, MAE, is the average of the difference between predicted and actual value in all test cases; it is the average prediction error. RMSE is frequently used measure of differences between values predicted by a model or estimator and the values actually observed from the thing being modeled or estimated [7,8].

Table 2 shows the Detailed accuracy by the classifiers chosen:

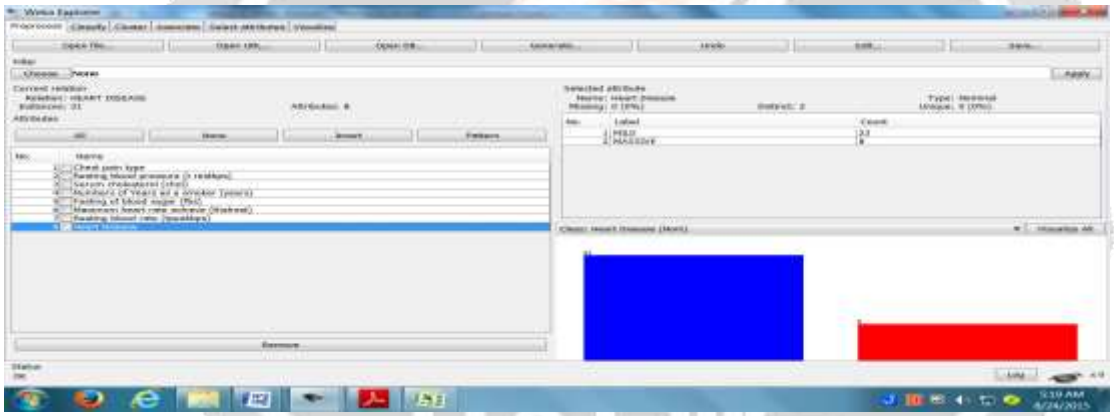

FIG1: Print Screen of WEKA 3.6 Environment

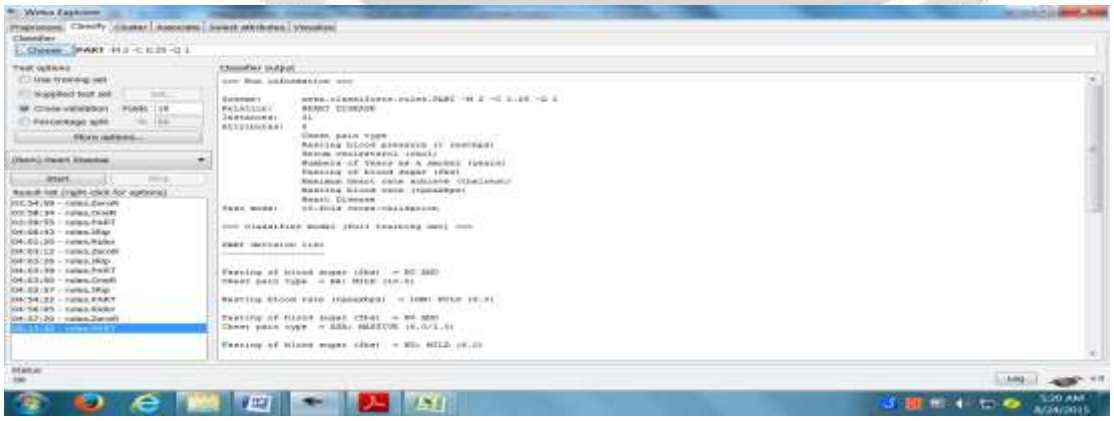

FIG 2: Classifier Output of PART Algorithm Rules

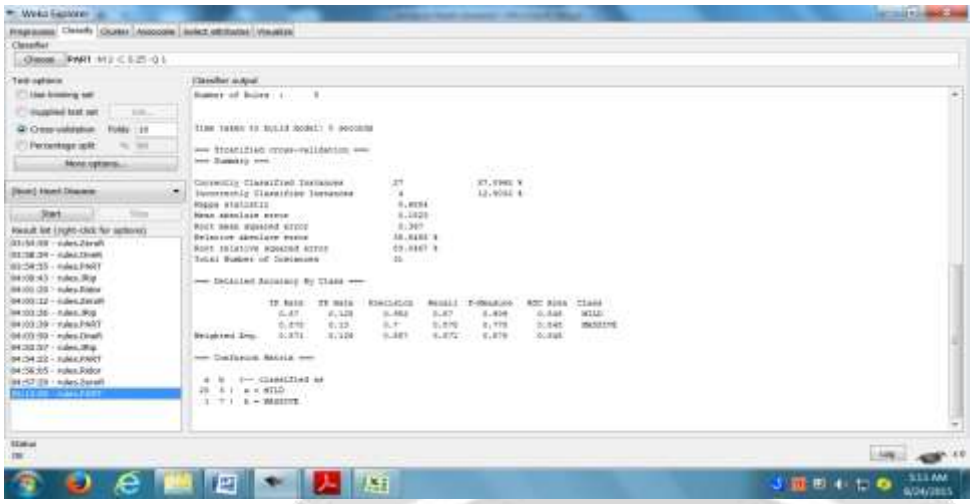

FIG 3: Classifier Output of PART Model

Table 3 shows the Classification Accuracy and Simulation Error:

Classification model	Phase	Classification Accuracy(%)	Mean Absolute Error	Root Mean -Squared Error	Number of Rules	Time (seconds)
Zero R	Cross validation	74.1935	0.3941	0.4419	01	0
ONE R	Cross validation	55	0.45	0.6708	0	0
PART	Cross validation	**87.0968**	**0.1523**	0.367	**5**	**0**
JRIP	Cross validation	80.6452	0.2165	0.393	2	0
RIDOR	Cross validation	74.1935	0.2581	0.508	3	0

From the above table, it is observed that PART algorithm attains least error rate and also gives the best accuracy (**87.0968%)** when compared to other algorithms. It generates the 5 rules in zero seconds. From the rule classifiers in WEKA tool, JRIP classifier gives the highest accuracy (80.6%) and next lowest error. It also generates the 2 rules in zero seconds.

V. RESULT ANALYSIS:

In this paper, we evaluate the performance of different classification methods that could produce accuracy and absolute average error to diagnosis the data set.

From the above Table 2 shows the detailed accuracy of different classifiers and also Table 3 gives the Classification accuracy and the simulation error.

From the table 3, it is observed that PART Algorithm is the best accuracy is 87.0968% and also gives the least error rate. The next accuracy is 80.6% belongs to JRIP Classifier and also produce the lowest error. FIVE rules are

generated by PART algorithm with ZERO seconds. The lowest accuracy is 55% belongs to One R classifier and also gives the highest error.

VI . CONCLUSION:

The performances of the various algorithms are measured through Classification Accuracy and Error rate. Comparisons among classifiers based on the accuracy, Mean Absolute Error and Root Mean squared values also considered. Comparisons among classifier based on the correctly classified instances are shown in Table 3. Based on the results, PART classifier produces the better accuracy and the lowest error in MAE and RMSE. In PART classifier, a number of rules is 5 from the above Fig. 3. The number of rules obtained by our proposed fuzzy method is less than that of earlier method.

REFERANCES:

[1] Manisha Barman, J.Pal Choudhury, "A Fuzzy Rule Base System for the Diagnosis of Heart Disease", International Journal of Computer Applications, (0975 – 8887), Volume 57- Number 7, November 2012.

[2]] Shilpa Dhanjibhai Serasiya, Neeraj Chaudhary, "Simulation of various classifications results using WEKA" International Journal of Recent Technology and Engineering (IJRTE),
ISSN: 2277 -3878, Volume -1, Issue -3, August 2012.

[3] C.Lakshmi Devasena, T.Sumathi, V.V. Gomathi and M.Hemalatha, "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set, Bonfring International Journal of Man Machine Interface, Vol – 1, ISSN 2250 – 1061, special issue, December 2011.

[4] Payal Dhakate, K.Rajeswari, Deepa Abin, " Analysis of Different Classifiers for Medical Dataset using various Measure" International Journal of computer Applications (0975 -8887) Volume 111- No.5, February 2015.

[5]Murlidhar Mourya, Phani Prasad, "An Effective Execution of Diabetes Dataset using WEKA" International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol. 4(5) , 2013,681-682.

[6] Khairul.A. Rasmani,Jonathan.M, Garibaldi, Qiang Shen and Ian O.Ellis, " Linguistic Rulesets Extracted from a Quantifier – Based Fuzzy Classification System, FUZZ – IEEE, 2009, Korea, August 20-24, 2009.

[7] F.Ibrahim,N.A.Abu Osman, J.Usman and N.A.Kadri(Eds),"Comparison of Different Classification Techniques Using WEKA for Breast Cancer" Biomed 06, IFMBE, Proceedings 15, pp.520-523, 2007, www.springerlink.com C.Springer-Verlag Berlin Heiddberg , 2007.

[8] M.Sudha, and A.Kumaravel, " Performance comparison based on Attribute Selection Tools for Data Mining" , International Journal of Science and Technology, Vol 7 (S7), 61-65, November 2014, ISSN: 0974 -6846.

[9] S.Jayasudha, A.Ammu Qudsiya, K.Ramanathan, "Fuzzy Petri net Model for Dynamic Alert Management System", International Journal of Computer Applications (0975 – 8887) Volume 95 – Number 16, June 2014.

[10] S.Jayasudha, C.vennila,R, Rajalakshmi,T.Manimozhi"Fuzzy Classification Algorithm: Performance Comparison with WEKA and MATLAB", National Conference in K.C.G. College of Engineering & Technology, Chennai.

[11] S.Jayasudha,C.vennila,R.Rajalakshmi,TManimozhi, "An Implementaion Of FIS For Diabetes Data With Fuzzy Petri Net Representation", International Journal of Applied Engineering Research (IJAER), ISSN 0973-4562 Volume 9, Number 22 (2014) pp.17537 – 17553. Research India Publications.