Int J Pharm Bio Sci 2015 Oct; 6(4): (B) 199 - 210

Review Article **Mathematics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

FUZZY PETRI NET GENERATED BY DATA MINING RULES FOR DIABETES DATA

S.JAYASUDHA*[1], K.RAMANATHAN[2] AND A.KUMARAVEL[3]

[1]Department of Mathematics, Bharath University, Chennai-73
[2]Department of Mathematics, KCG College of Engineering and Technology, Chennai
[3]Department of Information Technology, Bharath University, Chennai-73

ABSTRACT

Fuzzy Petri nets are capable of concurrent, reliable specification of business rule engines of a core of an expert system. An expert system based on Fuzzy rule based systems are common and specification of those systems by tools like Petri nets encourage more research work nowadays. The focus of this paper is to establish an iterative scheme using data mining techniques for extractingantimal set of rules with best accuracies of such models is devised and obtained result for generating theoptimal rule base for predicting the diabetes diagnosis results.

KEYWORDS: Fuzzy logic, WEKA, Fuzzy rule base, Fuzzy Petri net, Fuzzy Inference System and Receiver Operating Characteristics (ROC), Classification, Data Mining, Selected Attributes.

S.JAYASUDHA
Department of Mathematics, Bharath University, Chennai-73

1. INTRODUCTION

Diabetes care is typically complex and time consuming, drawing on many areas of health care management. There are two main types of cholesterol, Low Density Lipoprotein (LDL) known as the "bad cholesterol" capable of clogging the arteries, High Density Lipoprotein (HDL) known as the "good cholesterol" which transports some cholesterol back to the liver to be broken down. It is recommended to have lower LDL and higher HDL levels for a normal human being. Triglycerides are another type of blood fat that acts as temporary storage units of fat. High triglyceride levels can also contribute to plaque formation in the arteries. [14] The aim of this paper is to analyze how to evaluate the progression of disease by using Fuzzy Petri nets. Initially it is introduced 10 attributes of patients. They have stored as 442 instances download from original data base [1]. The data ranges of the attribute have been partitioned into several intervals based on certain intermediate values of the available data value [1]. Here we develop a framework and modeling approach for the classifying the progression of disease for diabetes mellitus by using Fuzzy Petri net. In Section II, the description of this application by constructing FPN model and mapping fuzzy rules to Fuzzy Petri net [2] is introduced. In Section III, we discuss about the WEKA tool. In Section IV, discusses the classification rule classifier and the various algorithms used for classification. In section V, we describe the ranges of attribute and also the statistical description for each attribute. In section VI, we present the comparison of different classification techniques using WEKA from the experimental results[9,13]. In section VII, we construct the Fuzzy Petri net and aconclusion is given in section VIII.

2. Fuzzy petrinet: [2]

In this section, the combination of Petri nets and fuzzy logic in order to extract both advantages namely, concurrency and conceptualizing the impreciseness is considered.

2.1 Fuzzy Petri Nets for Fuzzy Rules

The following Fuzzy Petri net (FPN) structure to model the fuzzy rules is introduced:
(P, P_s, P_c, T,TF,TRTF,A,I,O,TT,TTF,AEF,PR,PPM,TV) Where
P is a finite set of fuzzy places, in which
$P_s \subset$ P is a finite set of input places for primitive events or conditions.
$P_c \subset$ P is a finite set of output places for the actions or conclusions.
T is a finite set of transitions.
A$\subset$ (P X T $\cup T X P$) is a finite set of arcs for connections between places and transitions.
TT is a finite set of token types.
TTF: P $\rightarrow$ TT is token type function or a mapping which associates for each fuzzy place ϵ P to token type ϵ TT.
PR is a finite set of propositions, corresponding to either events or conditions or action conclusions.
PPM: P $\rightarrow$ PR, is a place to proposition mapping, where $|PR| = |P|$.
TV: P $\rightarrow$ [0,1] is a mapping for assigning truth values to places.
Hence FPN is a directed graph with edges indicating the linking the places to transitions and transitions to places. Similar to the conventional Petri nets except the nodes either places or transitions are associated with fuzzy values.

2.2 Fuzzy Production Rules [2]

Let R be a set of fuzzy production rules:
R ={ R_1,R_2,…..R_m}, and a fuzzy production rule R_i is as shown as follows
R_i : If d_j then d_k, (CF = μi)

IF all propositions in the antecedent dj have value true THEN the propositions in the consequent d_k are true.
Where di = {d_{j1}, d_{j2},….d_{jn}}, represents the antecedent part which comprises of one or more propositions connected by either " AND" or "OR" in the rule;
Dk = {d_{k1}, d_{k2},…. d_{kn}} represents the consequent part which comprises of one or more propositions connected by "AND" operator;
μi denotes the certainty factor (CFi) of the rule Ri. Generally, FPRs are classified into four types as follows:
Type 1: IF d_j , THEN d_k , (CF =μ),
Type 2: IF d_{j1} and d_{j2} and …..and d_{jn} THEN d_k (CF = μ),
Type 3: IF d_{j1} or d_{j2} or …..or d_{jn} THEN d_k (CF = μ),
Type 4: IF d_j THEN $d_{k1}$and d_{k2} and ……and d_{kn} (CF = μ),
FPN models are classified into 4 types of composite fuzzy production rules.

2.3 Fuzzification of Petri nets[2]

We described that a Petri net is generally constructed by using 4 types of objects, transitions, places, tokens and arcs. All of these objects may be fuzzified [4]. A fuzzy token is a generalization of the token in the standard Petri net which either does or does not belong to a place. Thus, a token may be seen as an individual having either the truth value 0 or 1. By allowing this truth value to be a value in the unit interval a fuzzy token is created. A more powerful idea is that the token has linguistic value, such as low, medium and high defined as a membership for a linguistic variable. A fuzzy place has a predicate or property associated with it. A fuzzy transition may for instance correspond to an if – then fuzzy production rule and is realized by truth values such as fuzzy inference algorithms. A fuzzy arc may specify a required value of a corresponding input token. A rule –base consists of a set of linguistic statements, called rules. These rules are of the form IF premise is composed of fuzzy input variables connected by the logical functions and the consequent is a fuzzy output variable.

2.4 Defuzzification

Executing the rule in the rule base generates multiple shapes representing the modified membership function. Defuzzification is the transformation of this set of percentages into a single crisp value based on how they perform this transformation, defuzzifier are divided with number of categories. The most commonly used defuzzifiers are the centre of gravity, maximum method etc.,

3. WEKA TOOL [7]

WEKA is a collection of machine learning algorithms for data mining tasks. WEKA contain tools for: - data preprocessing and classification. [4,] Classification is a data mining (machine learning) technique used to predict group membership for data instances. It is the problem of finding the model for class attribute as a function of the values of other attributes and predicting accurate class assignment for cross validation test. [4] All the classifiers like lazy , tree, rules and naïve comes under the categories only[9]. 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 [4].

3.1 Association Rule Mining [3]

Association rules are capable of revealing all interesting relationships in a potentially data base. [3] Many other classifications systems have been built based on association rules. In the research paper, there is an implementation of an association ruled – based classifier system in the WEKA framework. [7] The researcher has selected the data set given in the Table II which depicts the information about the different attributes in diagnosis of diabetes[6]. Classification rule is the subset of association rules.

4. Methodology Applied

Different rule based classifiers are used in this work to evaluate the effectiveness of those classifiers in a classification problem [4,] Figure 1 shows clearly the steps considered for our proposed method. The Classifiers applied are:

4.1 Zero R Classifier

In the Zero R method, the result is the class that is in the majority when the attributes are categorical and when they are numerical. Zero R is always considered as the base case for data mining [4] . Applications that work on the principles of data mining should not provide results worse than Zero R.

In this classifier, the test option is cross validation with 10 folds. Classification accuracy is 32.80 which is least value other than the classifiers and also the highest error is 0.2088 and number of rules is one, time taken by zero second.

4.2 One R classifier

The One R algorithm creates a single rule for each attribute for training data and then picks up the rule with the least error rate. [4]. If two or more rules have same error rate then the rule is selected at random. To generate a rule for an attribute the most recurrent class for each attribute value must be established [6]. Table 2 shows the detailed accuracy for this classifier. Table 3 shows the classification accuracy, Mean Absolute Error and also Root Mean squared Error for this classifier. For this classifier, the test options are cross validations with 10 folds.

4.3 JRip Classifier

JRip is an inference and rules-based learner (RIPPER) that tries to come up with propositional rules which can be used to classify elements [4]. In this classifier, the test option is cross validations with 10 folds. JRip produces the best accuracy and also least error. Number of rules is 8, time taken by 0.05 seconds.

4.4 NNge Classifier

Non-Nested Generalized Exemplars (NNGE) is an algorithm introduced by Brent,1995. It performs generalization by merging exemplars, forming hyper rectangles in attribute space that represents conjunctive rules with internal disjunction [4,12]. The algorithm forms a generalization each time a new example is added to the database, by joining it to its nearest neighbor of the same class.

4.5 Decision Table Classifier

Two variants of decision table classifiers are available. The first Classifier, called DTMaj (Decision Table Majority) returns the majority of the training set if the decision table cell matching the new instances is empty. i.e., it does not contain any training instances [4]. The second classifier, called DTLoc (Decision Table Local) is a new variant that searches for a decision table entry with fewer matching attributes (larger cells) if the matching cell is empty. This variant therefore returns an answer from the local neighborhood.

4.6 Ridor Classifier

Ripple- down Rule learner first generates the default rule. The exceptions are generated for the default rule with the lowest (weighted) ERROR RATE [4]. Then it generates the “best” exceptions for each exception. Thus it carries out a tree- like expansion of exceptions and its leaf has a default rule without exceptions.

4.7 PART Classifier

This is a class for generating a PART decision list. It uses separate-and –conquer approach and builds a partial C4.5 decision tree in each iteration and makes the “best” leaf into a rule[4].

4.8 Conjunctive Rule Classifier

It is a decision –making rule in which the intending buyer assigns least values for a number of factors and discards any result which does not meet the have minimum value on all of the factors [4].

4.9. System diagram for the Proposed Method

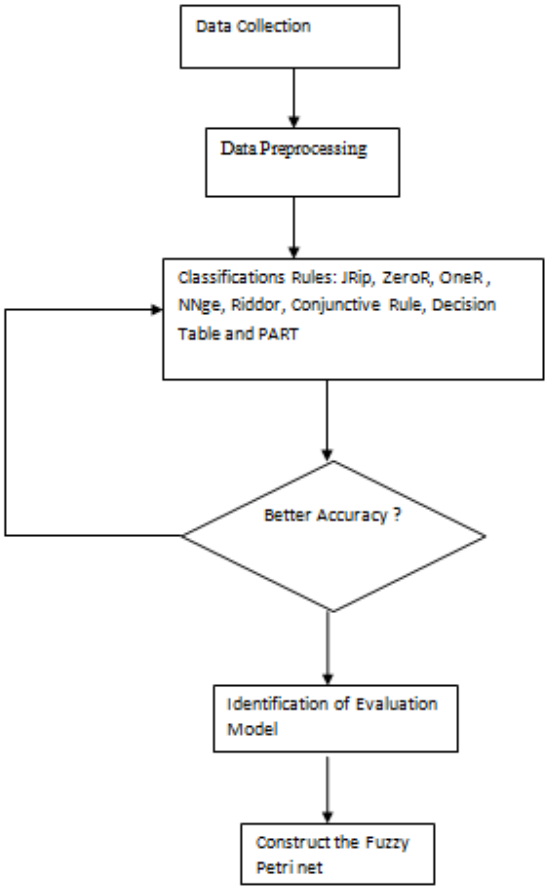

Figure 1
Iterative model for extracting the rules for diabetes data

5. DESCRIPTIONS

In this section, we describe the attributes (input fields) of diabetes data set and the process of fuzzification to be carried out[14].

5.1 Age

This input field is divided into 4 fuzzy sets (Young, Mild, Old, and Very Old). These fuzzy sets with their ranges will be shown in Table 1:

Table 1
Age classification

Input field	Range	Fuzzy Sets
Age	< 35	Young
	35- 45	Mild
	45-60	Old
	>60	Very Old

5.2 Sex

This input field just has 2 values (1,2) and sets (Female, Male). Value 1 means that patient is Male and Value 2 means that the patient is Female. This attribute happens to be crisp set only.

5.3 Blood Pressure

This input variable has divided into 4 fuzzy sets. Fuzzy sets are low, Medium, High and Very high. These fuzzy sets will be shown in Table2:

Table 2
Blood Pressure Classification

Input	Range	Fuzzy Sets
Blood Pressure	<134	Low
	127 – 153	Medium
	142 – 172	High
	>154	Very High

5.4 Table 3
Shows the Range of Six Blood Serum Measurements

S.NO.	INPUT	RANGE	FUZZY SETS
1.	Total Cholestrol(TC)	Below 200	Desirable
		Betwn200 -220	Borderline
		Betwn 220 -240	High
		Above 240	Highly Risk
2.	LDL	Below 100	Optimal
		115 - 145	Near Optimal
		145- 175	Borderline
		180 - 190	High
		Above 190	Highly Risk
3.	HDL	40-60	Optimal
		60-70	Normal
		30 -40	Highly Risk
4.	Triglycerides(TGH)	Below 150	Normal
		150 - 350	Borderline
		Above 350	Highly Risk
5.	AIC	Below 4.7	Low
		4.7 – 5.7	Normal
		5.7 – 6.4	Borderline
		Above 6.5(percent)	High
6.	Glucose	Below 140	Low
		140 -199	Normal
		200 -399	High
		Above 400	Very High

5.5 Statistical Description

Total Number of instances is 442 and the number of attributes is 11. The attributes are described in the following table:

Table 4
Statistical Descriptions of Attributes

S.No.	Attribute	Mean	Standard deviation
1.	Age	48.51809955	13.0794
2.	Sex	1.468326	0.498432
3.	BMI	26.37579	4.408137
4.	Blood Pressure	94.64701	13.80003
5.	Total Cholestrol	189.1403	34.52984
6.	LDL	115.4391	30.34435
7.	HDL	49.78846	12.90497
8.	Triglycerides	4.070249	1.287534
9.	ITG	4.641411	0.52121
10.	Glucose	91.26018	11.47035

6. EXPERIMENTAL RESULTS

6.1 Accuracy Measure

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

True Positive (TP)

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. 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 false positive rate. Precision is the proportion of relevant documents in the results returned. [10]

Table 5
Shows the Detailed accuracy by the classifiers chosen

Classifier	Phase	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ZeroR	Cross validation	0.328	0.328	0.108	0.328	0.162	0.0486
JRIP	Cross validation	0.765	0.098	0.732	0.765	0.742	0.872
Decision Table	Cross validation	0.713	0.122	0.665	0.713	0.677	0.868
Conjunctive Rule	Cross validation	0.502	0.213	0.297	0.502	0.366	0.686
RIDOR	Cross validation	0.688	0.112	0.672	0.688	0.679	0.788
PART	Cross Validation	0.724	0.112	0.676	0.724	0.696	0.868
ONE-R	Cross validation	0.624	0.165	0.574	0.624	0.581	0.73
NNGE	Cross validation	0.652	0.118	0.64	0.652	0.645	0.767

6.2 Error Rate

6.2.1 Mean Absolute Error[4]

Mean absolute error, MAE, is the average of the difference between thepredicted and actual value in all test cases; it is the average prediction error. The formula for calculating MAE is given in equation shown below:

MAE = $(|a_1 - c_1| + |a_2 - c_2| + \ldots\ldots |a_n - c_n|)/n$

Assuming that the actual output is a expected output is c

6.2.2 Root Mean –Squared Error [4,]

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[5,8]. It is just the square root of the mean square error as shown in equation given below:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots . (a_n - b_n)^2}$$

The classification accuracy, mean absolute error and root mean squared error are calculated for each machine algorithm.

Table 6

Shows the Classification Accuracy and Simulation Error

Classification model	Phase	classification Accuracy	Mean Absolute Error	Root Mean - Squared Error	Number of Rules	Time (seconds)
Conjunctive Rule	Cross validation	50.226	0.1713	0.2926	1	0.02
Decision Table	Cross validation	71.27	0.146	0.2542	30	0.09
JRip	Cross validation	76.4706	0.1059	0.2355	8	0.05
NNge	Cross validation	65.61	0.0995	0.3155	129	0.08
OneR	Cross validation	62.44	0.1073	0.3276	6	0.02
Part	Cross validation	72.3982	0.1063	0.2449	18	0.05
Ridor	Cross Validation	68.78	0.0892	0.2987	25	0.03
Zero R	Cross validation	32.80	0.2088	0.3226	1	0

From the above table, it is observed that JRIP algorithm attains least error rate. Therefore JRIP classification algorithms performs well because it contains least error rate and also highest accuracy when compared to other algorithms. [11,13]

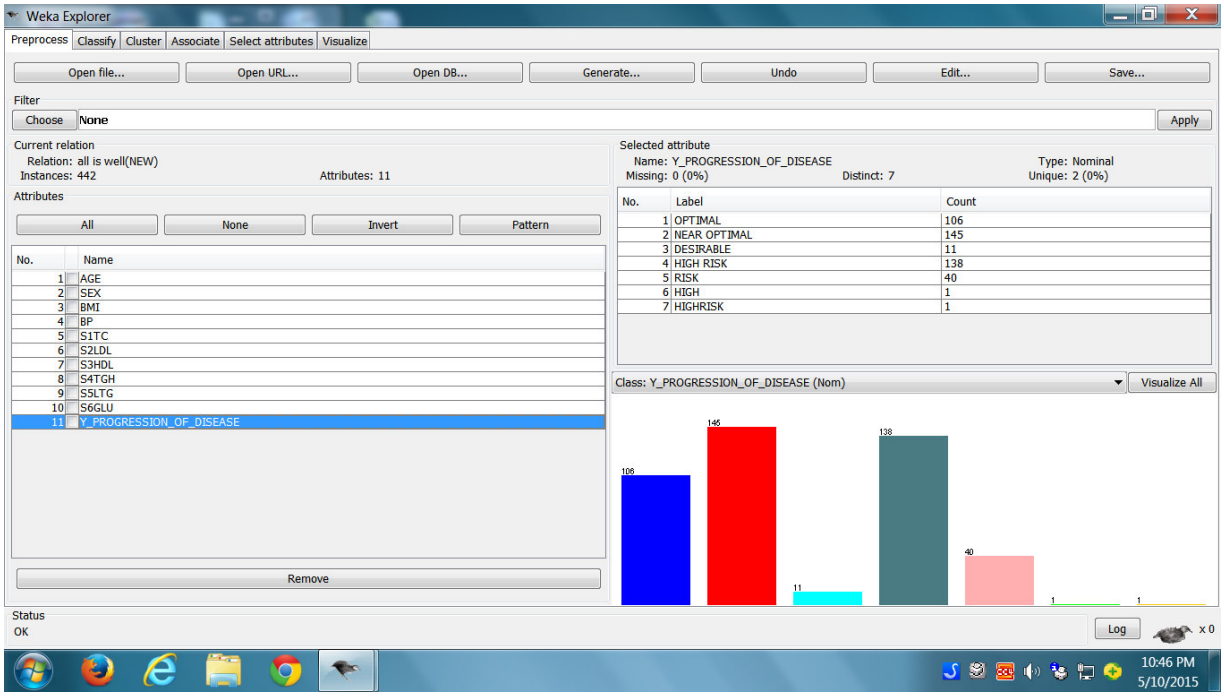

Table 7
Print Screen of WEKA 3.6 Environment

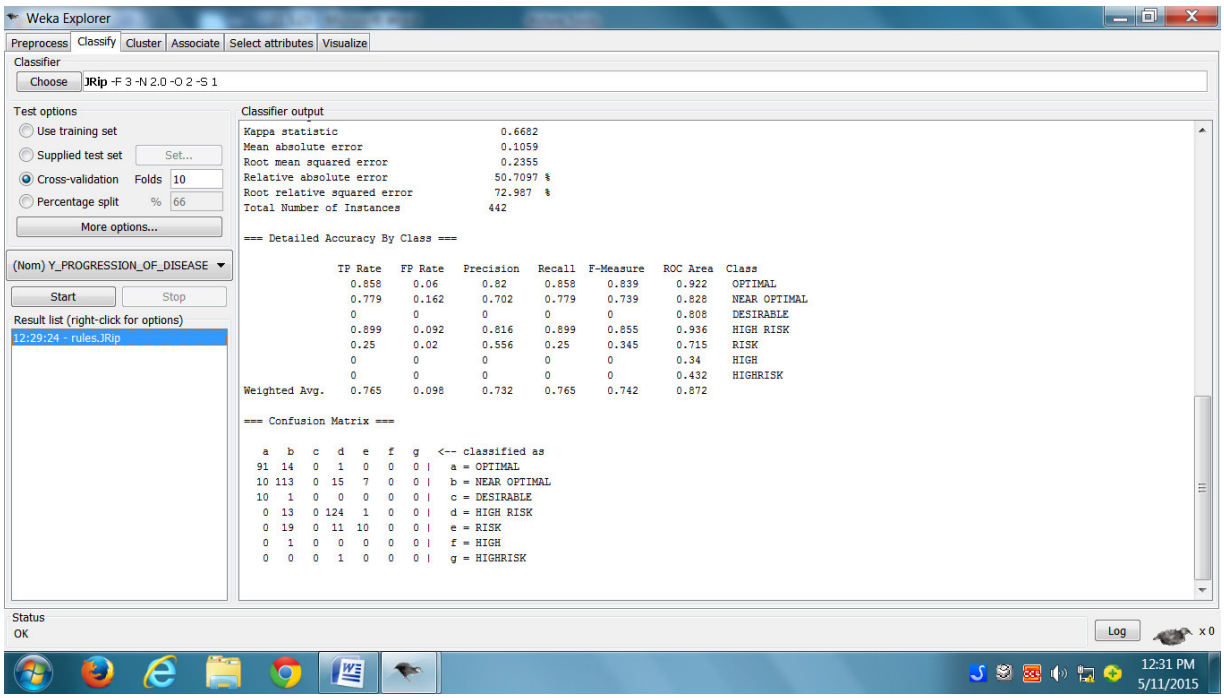

Table 8
Classifier Output of the JRIP model

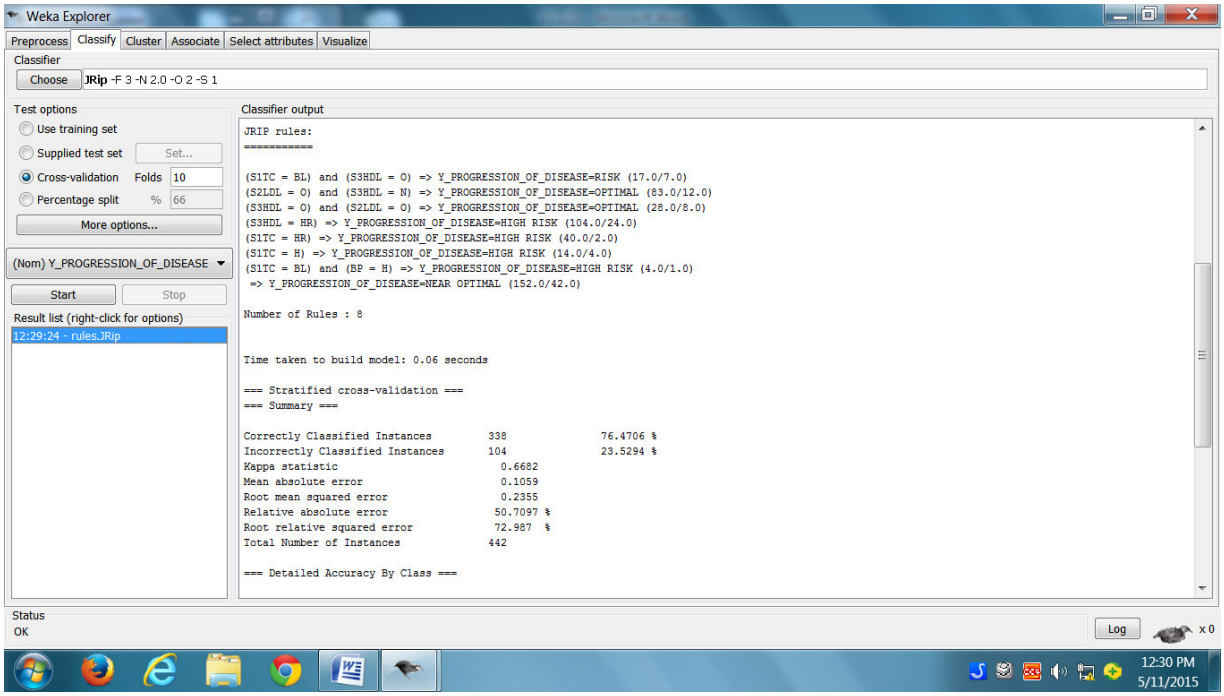

Table 9
Classifier output of the JRIP Rules

7. Construction of Fuzzy Petri Net

Based on the overall results, JRip Classifier produces the better accuracy and also gives the minimum error. Using WEKA tool, Number of rules found in JRip Classifier are eight. From these rules, the researcher constructs the Fuzzy Petri net. Introduce Input Four Places S_1,S_2,S_3 and S_4 and 8 transitions and the outcome three places Desirable, Risk and Highly Risk.

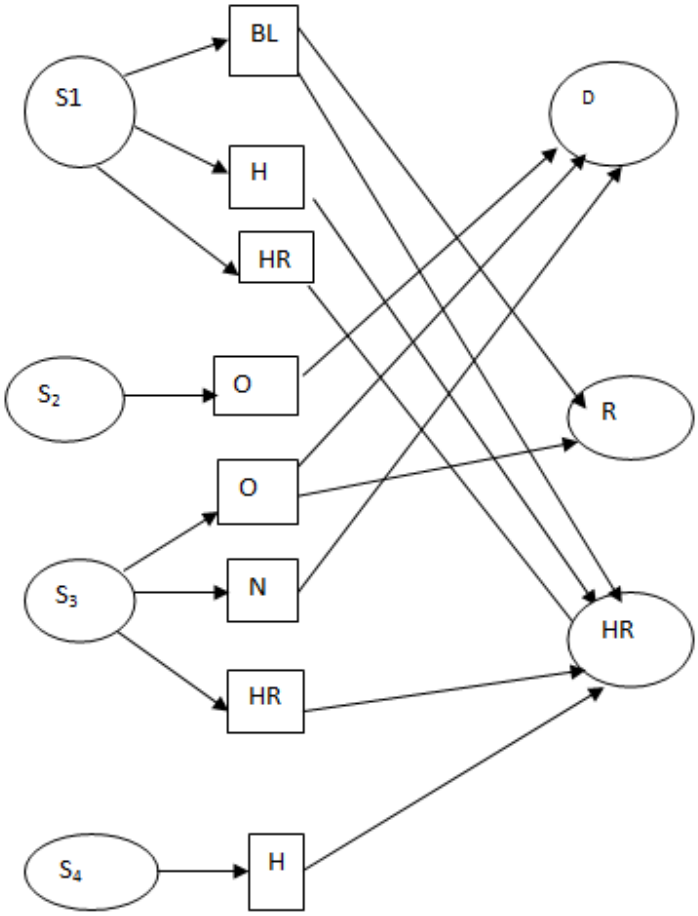

Figure 2
Fuzzy Petri net

8. CONCLUSION

This work is performed using Machine learning tool to predict the effectiveness of all the rule based classifiers. The performances of the various algorithms measured in classification Accuracy. Comparisons among classifiers based on the accuracy, Mean Absolute Error and Root Mean squared values also considered. Comparison among classifier based on the correctly classified instances are shown in Table 6. Based on the results, JRIP classifier produces the better accuracy and the lowest error in MAE and RMSE. In JRIP classifier, a number of rules are 8 is given above. From the rules, finally fuzzy petri net is constructed. For the purposes of comparing the classification accuracy obtained with the same number of rules, some parameters were tuned for better results.

ACKNOWLEDGEMENT

The author's thanks for the Referee's comments

REFERENCES

1. American Diabetes Association. Economic costs of diabetes in the US in 2002. Diabetes Care; 26: 917-932, 2003.

2. S.Jayasudha, A.Ammu Qudsiya, K.Ramanathan, "Fuzzy Petri net Model for Dynamic Alert Management System", International Journal of Computer

Applications Volume 95 – Number 16,(0975-8887) (2014).

3. Shilpa Dhanjibhai Serasiya, Neeraj Chaudhary, “Simulation of various classifications results using WEKA” International Journal of Recent Technology and Engineering (IJRTE), , Volume -1, Issue -3, ISSN:2277-3878,(2012).
4. 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, , special issue, ISSN 2250 – 1061 (2011).
5. Payal Dhakate, K.Rajeswari, Deepa Abin, “ Analysis of Different Classifiers for Medical Dataset using various Measure” International Journal of computer Applications Volume 111- No.5, (0975 - 8887) (2015).
6. 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) , 681-682., 2013,
7. P.Yasodha, M.Kannan, “Analysis of a Population of Diabetic Patients Databases in Weka Tool”, International Journal of Scientific & Engineering Research, Volume 2, Issue 5, (2001).
8. Khairul.A. Rasmani,Jonathan.M, Garibaldi, Qiang Shen and Ian O.Ellis, “ Linguistic Rulesets Extracted from a Quantifier – Based Fuzzy Classification System, FUZZ – IEEE, 20-24, (2009).
9. 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.
10. S.Vijarani, M.Muthulakshmi, “Evaluating the Efficiency of Rules Techniques for File Classification”,IJRET,International Journal of Research in Engineering and Technology,volume:02, Issue :10 eISSN: 2319 – 1163/PISSN:2321-7308.(2013)
11. A.Kumaravel and Pradeepa.R, “Efficient Molecule Reduction For Drug Design by Intelligent Search methods” International Journal of Pharma and Bio Sciences, ISSN 9075 – 6299, 4(2): (B) 1023 -1029. (2013)
12. M.Sudha, and A.Kumaravel, “ Performance comparison based on Attribute Selection Tools for Data Mining” , International Journal of Science and Technology, ISSN: 0974 -6846 Vol 7 (S7), 61-65, (2014).:
13. R.Karthikeyan, A.Kumaravel and V.Khanna, “ Significance of Information Gan Ratio for improving classification of Heart Diseases”, International Journal of Pharma and Bio Sciences , ISSN 0975 – 6299 6(2):B(182 – 190),. (2015)
14. A.Kumaravel and D.Udhayakumarapandian, “A Novel subset selection For Classification of Diabetes Data set By Iterative Methods “, Int J Pharm Bio SCi; 5(3): (B) 1-8.(2014).