ROLE OF ROUGH MATRIX IN DECISION MAKING USING WEKA

Nirmala Rebecca Paul[(1)]
Department of Mathematics
Lady Doak College,Madurai
Tamil Nadu,India 625002

S.Pichumani Angayarkanni[(2)]
Department of Computer Sciences
Lady Doak College,Madurai
Tamil Nadu,India 625002

***Abstract* – The paper introduces rough matrix in terms of rough membership function. An algorithm is developed to find the core of an information system. This algorithm is applied to analyse the real-life problems. The algorithm is used to find the deciding factors for the disease Flu. Also the chemical parameters that decides the usage of water are also derived. The proposed algorithm is tested using Weka and the CORE attributes are determined using the attribute selection process using Rough set approach. The feasibility of the proposed Rough set matrix is tested for Chronic kidney disease dataset.**

2010 Mathematics Subject Classification:94D05,68T37

Index Terms – Rough sets, Rough matrix and core.

I INTRODUCTION

The Rough set theory represented by Pawlak[4] is used as an useful tool for representation and decision making. Rough set theory deals with approximations of sets in terms of the equivalence relations defined on the universe. Rough set theory is used for many practical problems to select the set of attributes necessary for classification of objects in the universe. Any information system consists of several attributes and it is necessary to pick the minimal attributes for the classification of objects. Rough topology [2] is defined in terms of approximations and boundary conditions. Pawlak has defined the rough membership function. This paper defines rough matrix in terms of rough membership function. An useful algorithm is developed to find the minimum number of attributes for the classification of objects in the universe called core. This algorithm is defined in terms of the rough matrix and it is applied to analyse the real life problems. The deciding factors for flu and the chemical parameters that decides the usage of water for drinking are found using the algorithm. The proposed algorithm is tested using the software ROSE2 for the Urology dataset . The CORE sets refers to the features which plays a vital role in classification are selected using the Rough Set approach.

II. PRELIMIMARIES

Definition 2.1[4]

Let U be a non-empty finite set of objects called the universe and R be an equivalence relation on U named as the indiscernibility relation. Then U is divided into disjoint equivalence classes. The pair (U,R) is said to be the approximation space. Let X be a subset of U

(i) The lower approximation of X with respect to R is defined as $L_R(X)=\{x \in U / R(x) \subseteq X\}$

(ii) The upper approximation of X with respect to R is defined as $U_R(X)=\{x \in U / R(x) \cap X \neq \phi\}$

(iii) The boundary of X with respect to R is defined as $B_R(X)= U_R(X)- L_R(X)$

where R(x) denotes the equivalence class determined by x.

Definition 2.2[4]

The rough membership function $\mu_X{}^R : U \to [0,1]$ is defined as $\mu_x{}^R(x)=\frac{|X \cap R(x)|}{|R(x)|}$ where $|X|$ denotes the cardinality of X.

III Rough Matrix

Rough Matrix is defined in terms of equivalece classes and the rough membership function.

Definition 3.1[4]

Let (U,A) be an information system. Here U is an universe, A is the set of attributes which is divided into a set of C of condition attributes and a set of D of decision attributes. The rows of the rough matrix represent the members of the universe and the columns represent the conditional attributes. The entries of the matrix are computed by the rough membership function.

Definition 3.2 [4] In an information system, not all condition attributes depict the decision attribute. The decision depends not on the whole set of condition attributes but on a subset of it is called the CORE.

Example 3.3

The following table gives the information of six patients suffering from "Flu".

Patients	Headache (H)	Temperature(T)	Muscle Pain(M.P)	Flu
1	No	Yes	Yes	Yes
2	Yes	Yes	No	Yes
3	Yes	Yes	Yes	Yes
4	No	No	Yes	No
5	Yes	Yes	No	No
6	No	Yes	Yes	Yes

Here H,T and M.P. are conditional attributes and Flu is the decision attribute.U={1,2,3,,4,5,6} is the set of patients and the equivalence classes corresponding to all conditional attributes are given as{{1,6},{2,5},{3},{4}}. The equivalence classes corresponding to the conditional attribute H are {{1,4,6},{2,3,5}}. Similarly, the equivalence classes corresponding to the condition attributes T and M.P. are
{{ 1,2,35,6},{4}} and {{1,3,4,6},{2,5}} respectively. The equivalence classes corresponding to the decision attribute are {{1,2,3,6},{4,5}}.
Let A be the rough matrix whose first column entries are obtained by the membership function and the equivalence classes corresponding to all conditional attributes and X={2,3,5}. The second and third column entries are obtained as before with respect to the sets X={1,2,3,5,6},X={1,3,4,6}.

$$A=\begin{pmatrix}0&1&1\\1&1&0\\1&1&1\\0&0&1\\1&1&0\\0&1&1\end{pmatrix}$$ Similarly another rough matrix B can be obtained as before but the equivalence classes are {{1,2,3,6},{4,5}} and the sets are defined as in the matrix

$$A.\ B=\begin{pmatrix}\frac{1}{2}&1&\frac{3}{4}\\\frac{1}{2}&1&\frac{3}{4}\\\frac{1}{2}&1&\frac{3}{4}\\\frac{1}{2}&\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&1&\frac{3}{4}\end{pmatrix}$$

Definition 3.4 If A and B are two rough matrices of the same order the rough matrix C=A ∧ B is obtained by considering the minimum elements. $C=(c_{ij})=\min(a_{ij},b_{ij})$ for all i,j

Example 3.4

If $A=\begin{pmatrix}0&\frac{1}{2}\\1&\frac{1}{3}\end{pmatrix}$ $B=\begin{pmatrix}1&0\\\frac{1}{4}&\frac{2}{3}\end{pmatrix}$ then

$$C=A\wedge B=\begin{pmatrix}0&0\\\frac{1}{4}&\frac{1}{3}\end{pmatrix}$$

IV ALGORITHM

An algorithm is developed to find the the deciding factors or core to pick the minimum number of attributes necessary for classification of attributes.
Step 1: Find the equivalence classes corresponding to all conditional attributes
Step 2: Select X corresponding to the first conditional attribute.
Step 3: Find the first column entries of the rough matrix with the rough membership function and the equivalence classes corresponding to all conditional attributes.
Step 4: Find the other column entries corresponding to other conditional attributes and name it as A.
Step 5: Find the rough matrix B as in step 2, step 3 and step 4 corresponding to the equivalence classes of the decision attributes.
Step 6: Find the matrix A ∧ B.
Step 7: Find the maximum elements in each column.
Step 8: The resulting rough matrix will be a row matrix
Step 9: Find the subset of the conditional attributes with maximum values which represent the core of the information system.

Example 4.1

In the Example 3.2 $C=A\wedge B=\begin{pmatrix}0&1&\frac{3}{4}\\\frac{1}{2}&1&0\\\frac{1}{2}&1&\frac{3}{4}\\0&0&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}&0\\0&1&\frac{3}{4}\end{pmatrix}$.

The row matrix from C is $(\frac{1}{2}\ 1\ \frac{3}{4})$. Here 1 corresponds to the conditional attribute temperature. Hence core is the condition attribute temperature. Hence temperature is the key attribute that has close connection with the disease Flu.

Example 4.2

The following table gives eight samples of water taken to test the portability of water. If the sample contains the chemical parameters not within the permissible limit will have an impact which is given by the following table.

Parameters	Impacts
TDS	Bitter taste, formation of kidney stone
Hardness	Scale in utensils, formation of kidney stone
Nitrate	Blue baby disease in infant
pH	Leads to ulcer in the stomach
Flouride	Bone damage, Fluorosis

Samples	F	N	TDS	pH	Hard	usability
1	Yes	No	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes	No	Yes
3	Yes	No	No	Yes	Yes	No

4	Yes	Yes	Yes	Yes	Yes	Yes
5	No	Yes	Yes	Yes	Yes	No
6	Yes	No	No	Yes	No	No
7	No	No	No	No	Yes	No
8	Yes	Yes	Yes	Yes	No	Yes

Here "No" represents that the parameter present in the sample is not within the permissible limit and "Yes" represents that the parameter present in the sample is within the permissible limit. The equivalence classes corresponding to all conditional attributes are {{1},{2,8},{3},{4}{5},{6},{7},{8}}. The clumn entries are obtained by taking X={1,2,3,4,6,8}, X={2,4,5,8},X={1,2,4,5,8}, X={1,2,3,4,5,6,8} and X={1,2,4,5,7}

The rough matrix A= $\begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & \frac{1}{2} \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}$

The equivalence classes of the decision attribute are{1,2,4,8},{2,5,6,7}. The matrix B is obtained by taking the this equivalence classes and the values of X as in A.

The rough matrix B= $\begin{pmatrix} 1 & \frac{3}{4} & 1 & 1 & \frac{1}{4} \\ 1 & \frac{3}{4} & 1 & 1 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} \\ 1 & \frac{3}{4} & 1 & 1 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} \\ 1 & \frac{3}{4} & 1 & 1 & \frac{1}{4} \end{pmatrix}$

The rough matrix C = $\begin{pmatrix} 1 & 0 & 1 & 1 & \frac{1}{4} \\ 1 & \frac{3}{4} & 1 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{3}{4} & \frac{1}{2} \\ 1 & \frac{3}{4} & 1 & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & 0 & 0 & \frac{3}{4} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{3}{4} \\ 1 & \frac{3}{4} & 1 & 1 & 0 \end{pmatrix}$ The row matrix is ($1 \quad \frac{3}{4} \quad 1 \quad 1 \quad \frac{3}{4}$). Thus the core is the set {F,TDS,Ph}. They are the key attribute which decide the usability of the samples.

V PROPOSED ROUGH SET ANALYSIS USING DATA MINING CONCEPT

Any dataset is represented in two forms

A. Information table
B. Decision table

The variables/attributes TDS, Hardness, Nitrate, pH and Flouride represent the columns.

The instances/cases represents the rows.

In the proposed approach we use decision table in which one of the variable is called a decision and the others are called attributes.

The Decision table for the Table(1) is represented using two elementary sets

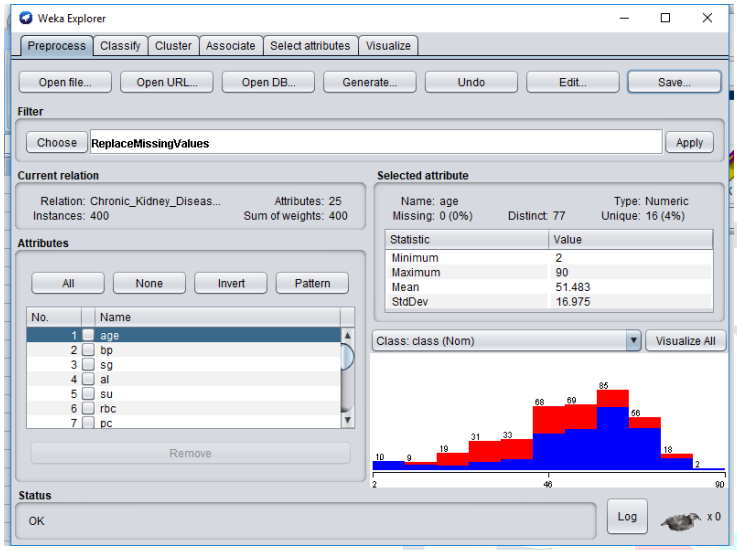
{usability} :{2,4,8}->yes
{1,3,5,6,7}->No

They are also called as concepts.

The proposed roughest process is validated for Chronic kidney disease dataset using RoughSet approach in WEKA Software. The Fuzzy Roughset algorithm plays a vital role in attribute selection. Description about the Dataset:

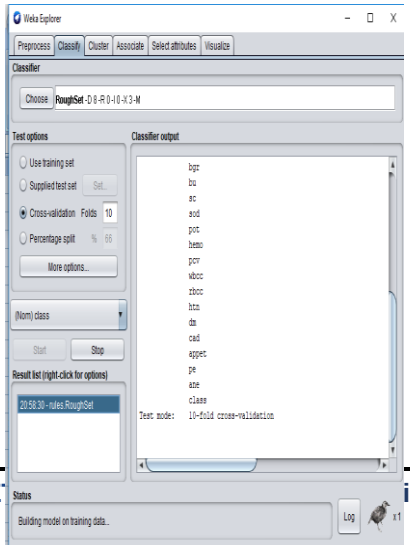
Data Set Characteristics:	Multivariate	**Number of Instances:**	400	**Area:**	N/A
Attribute Characteristics:	Real	**Number of Attributes:**	25	**Date Donated**	2015-07-03
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number**	108081

Step 1: Load the weka dataset



Step 1: Preprocessing technique:
Data is rarely clean and often you can have corrupt or missing values. It is important to identify, mark and handle missing data when developing machine learning models in order to get the very best performance.Therefore the RemoveMissingValue unsupervised filter is applied. A simple way to handle missing data is to remove those instances that have one or more missing values.

Step 2: Developing the java code for the Roughset matrix approach proposed above and converting to jar file. The java package is converted to jar file . The package provides 3 classifiers. Rule classifier based on rough sets uses the concepts of discernibility matrix, reducts and rules generated from reducts. It provides variety of algorithms generating reducts including faster for larger data sets local reducts and has modes to work with incomplete data and inconsistent data. K nearest neighbours classifier provides variety of distance measures that can work also for data with both numeric and nominal attributes and has built-in k optimization. It implements fast neighbours searching algorithm making the classifier work for very large data sets. The classifier has also the mode to work as RIONA algorithm. The third classifier is the extension of k nearest neighbours method that induces local metric for each classified object. It is dedicated rather to large data sets (2000+ training instances) and improves accuracy particularly in case of data containing nominal attributes.



The selected Attributes are :
=== Run information ===

Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search: weka.attributeSelection.RoughSet -T -1.7976931348623157E308 -N -1 -num-slots 1
Relation: Chronic_Kidney_Disease-weka.filters.unsupervised.attribute.RemoveUseless-M99.0-weka.filters.unsupervised.attribute.RemoveUseless-M99.0
Instances: 400
Attributes: 25
age
bp
sg
al
su
rbc
pc
pcc
ba
bgr
bu
sc
sod
pot
hemo
pcv
wbcc
rbcc
htn
dm
cad
appet

pe
ane
class

Evaluation mode: evaluate on all training data

= Attribute Selection on all input data ===

Search Method:
Greedy Stepwise (forwards).
Start set: no attributes
Merit of best subset found: 0.671

Attribute Subset Evaluator (supervised, Class (nominal): 25 class):
CFS Subset Evaluator
Including locally predictive attributes

Selected attributes:
2,3,4,6,10,12,13,14,15,16,17,19,20,22,23,24 : 16
bp
sg
al
rbc
bgr
sc
sod
pot
hemo
pcv
wbcc
htn
dm
appet
pe
an

VI CONCLUSION

The rough matrix is defined in terms of the rough membership function. An algorithm is developed to find the CORE of an information system. Two examples are discussed to find the CORE of different information system. The proposed algorithm is implemented using Weka and tested for Chronic kidney failure dataset. Out of 25 attributes the CORE attributes selected includes 17 and these attributes plays a vital role in decision making of the kidney failure.

VII REFERENCES

[1] Jansi Rani, P.G. and Bhaskaran R, Computation of reducts using topology and measure of significance of attributes, Journal of Computing 2(2010),50-55.

[2] Nirmala Rebecca Paul, Decision making in an information system via a new topology, Annals of Fuzzy Mathematics and Informatics, 9(2016),1-10.

[3] Salama A.S., Some topological properties of rough sets with tools for data mining, International Journal of Computer Science Issues8(2011),588-595.

[4] Pawlak z., Rough sets, International Journal of Information and Computer Sciences, 11(1982),341-356.

[5]Yuhua Qian, Chuangyin Dang, Jyeh Liang and Dawaei, Set Valued ordered Systems, Information. Sciences, 179(2009),2793-2809.