Handling Incomplete Information System a Rough Set Theory based Approach

[1] Neha Saini, [2] Dr. Sachin Patel

[1] M Tech 4th, [2] Asso. Prof., C.S.E. Department
[1]C.S.E. Department S.I.R.T Indore M.P. India,
[2] C.S.E. Department S.I.R.T Indore M.P. India,

Abstract : Rough set theory is a new method that deals with uncertainty in decision making. This method has been developed to manage uncertainties from information that presents incompleteness and noises. Lower and upper approximations can be used by rough set for the representation of the concerned set. The approximation is the main concept acquired from data is the main objective of the rough set analysis. Rough sets have been used for a very wide variety of applications. Main advantages of the rough set approach includes, It does not need any additional information about data, It provides efficient methods, algorithms and tools for finding hidden patterns in data, It allows to reduce original data to find minimal sets of data. It allows evaluating the significance of data. In this paper we used rough set theory and its properties to handle incomplete information. We divide the data set into complete and incomplete information and used Degree of dependency, Indiscernibility and POS to construct incomplete information. The proposed approach is simple and easy to understand. By the experimental analysis it is observed that proposed approach construct more accurate incomplete information as compared to other statistical approach

***IndexTerms* – Rough Set, Incomplete, Lower, Upper, Approximations, Indiscernibility.**

I. Introduction

Rough Set Theory is defined as an extension of the conventional set theory that supports approximations in decision making. Rough set theory is the approximation of a clear concept. A pair of fixed concepts that classify the domain of interest into disjoint categories, called lower and upper approximations. The description of the domain objects which are known with certainty to belong to the subset of interest is called the lower approximation, whereas the description of the objects which possibly belong to the subset is called the upper approximation.

The RST has been applied in several fields including image processing, data mining, pattern recognition, medical informatics, knowledge discovery and expert systems. There are several research works have been proposed the rough set theory with other domain like artificial intelligence, neural networks, fuzzy logic, Data mining, Genetic Algorithms additionally to other methods resulting in some good results. The use of rough set theory to solve a specific complex problem has attracted world-wide attention of further research and development. Rough set theory has been extending to the original theory and increasingly widening fields of application. Rough set as a computationally efficient technique it presents a basic significance to many theoretical developments. Rough set addresses many problems such as data significance evaluation, hidden pattern discovery from data, decision rule generation, data reduction etc. Rough sets are applied in mal life applications such as, medicine, finance, telecommunication, vibration analysis, conflict resolution, intelligent agents, image analysis, pattern recognition, control theory, process industry, marketing, banking risk assessment etc.

II. Presentation of Information with RST

Data set**:-**A data set is represented as a table, where each row represents a case, an event, a patient or simply an object. Every column represents an attribute. It is a pair S = (U,A), where U is a non-empty finite set of objects called the universe and A is a non-empty finite set of attributes such that $a:U \rightarrow V_a$ for every a ∈ A. The set V_a is called the value set of a. Based on the information given, if the decision column is not given then the table is called the information table. Data set has followed properties

- A set of object can be characterized in terms of attribute values.
- It is possible to find total or partial dependencies between objects
- Data reduction
- The more significant attributes can be discovered
- Generation of decision rules

Indiscernibility: Let S=(U,A) be an information system, and $P \subseteq A$. A binary relation $INDs(P)$ defined in the following way $\mathbf{IND_s(P)=\{(x,y)\in U^2| \ \forall \ aP \ ,a(x)=a(y)\}}$ is called a P-indiscernibility relation. If $(\boldsymbol{x},)\in \boldsymbol{IND_s(P)}$, then x and y are indiscernible (or indistinguishable) by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted by $[\boldsymbol{x}]$.

III. Approximation Spaces and Set Approximation

Let us consider given a set of objects U called the Universe and an indiscernibility relation $R\subseteq U\times U$ representing our lack of knowledge about elements of U.Let us assume that R is an equivalence relation.

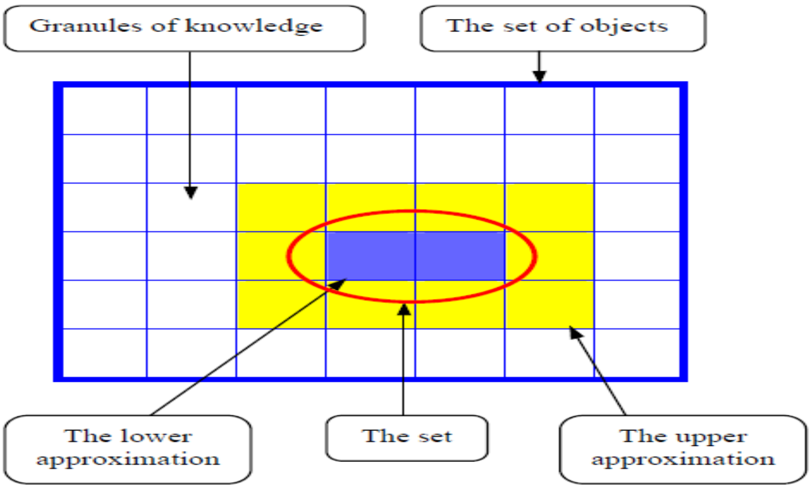


Figure 2 Graphical representations of Spaces and Set Approximation

A pair $(U,)$ is an approximation space where U is the Universe and R is an equivalence relation on U. Let X be a subset of U, i.e. $X\subseteq U$ The main goal is to characterize the set X with respect to R. By *R(x)*, denote the equivalence class of R determined by elements x .The indiscernibility relation represents they lack of knowledge about the universe U. Equivalence classes of the relation R, called granules represents an elementary portion of knowledge they able to perceive due to R. Using this indiscernibility relation they are not able to observe individual objects from U but only the accessible granules of knowledge described by this relation.

Lower approximation:

The set of all objects which can be with certainty classified as members of X with respect to R is called the *R-lower approximation* of a set X with respect to R, and denoted by $\boldsymbol{R*(x)}$.

$\mathbf{R*(X)=\{x: R(x)\subseteq X\}}$.

Upper approximation:

The set of all objects which can be only classified as possible members of X with respect to R is called the *R-upper approximation* of a set X with respect to R, and denoted by $\boldsymbol{R*(x)}$.

$\mathbf{R*(X)=\{x: R(x)\cap X\neq\emptyset\}}$.

IV. Literature Survey

In 2010 Yiyu Yao proposed "The rough set theory approximates a concept by three regions, namely, the positive, boundary and negative regions". A positive rule makes a decision of acceptance, a negative rule makes a decision of rejection, and a boundary rule makes a decision of abstaining. They provides an analysis of three-way decision rules in the classical rough set model and the decision-theoretic rough set model. The results enrich the rough set theory by ideas from Bayesian decision theory and hypothesis testing in statistics. The connections established between the levels of tolerance for errors and costs of incorrect decisions make the rough set theory practical in applications. They represent the results of a three-way decision of acceptance, rejection, or abstaining [1].

In 2011 Yiyu Yao and Xiaofei Den proposed "Sequential Three-way Decisions with Probabilistic Rough Sets". Rules obtained from the three regions are recently interpreted as making three-way decisions, consisting of acceptance, deferment, and rejection. This framework is further extended into sequential three-way decision making, in which the cost of obtaining required evidence or information is also considered. This enables us to consider both the cost of various misclassifications and the cost of obtaining the necessary evidence for making a classification decision. Although the latter is very important in decision-making, it has not received sufficient attention. They reported some preliminary results on the topic. The exploration of multiple representations of objects for decision-making is a useful direction in granular computing. Typically, decisions made at a higher level of granularity or abstraction may be less accurate or reliable but with a lower cost of resources [2].

In 2012 Feng Hu and Guoyin Wang proposed "Knowledge Reduction Based on Divide and Conquer Method in Rough Set Theory". They proposed the knowledge reduction approaches based on divide and conquer method, under equivalence relation and under tolerance relation, are presented, respectively. Systematic approach, named as the abstract process for knowledge reduction based on divide and conquer method in rough set theory, is proposed. Based on the presented approach, two algorithms for knowledge reduction, including an algorithm for attribute reduction and an algorithm for attribute value reduction, are presented. Some experimental evaluations are done to test the methods on uci data sets and KDDCUP99 data sets[3].

In 2013 Thabet Slimani proposed "Application of Rough Set Theory in Data Mining". They introduces the fundamental concepts of rough set theory and other aspects of data mining, a discussion of data representation with rough set theory including pairs of attribute-value blocks, information tables reducts, indiscernibility relation and decision tables. Additionally, the rough set approach to lower and upper approximations and certain possible rule sets concepts are introduced. Finally, some description about applications of the data mining system with rough set theory is included The applications of data mining based on the original approach of rough set theory, have been attempted valuable methods to generate decision rules in recent years (about 20 years now). The obtained results need more research, particularly, when quantitative attributes are involved[4].

In 2014 K Anitha proposed "Rough Set Theory Approach to Generating Classification Rules". They emphasize the role of Reducts, Core and their approximations. Data from UCI repository have been taken to exhibit rules for soybean data set by using

ROSETTA. They proposed rule generation using Rough Set which is implemented for Soybean (Large) data set using ROSETTA software. From these set of rules Reducts, Upper and Lower Approximations have been evaluated. Reducts are minimal representation of the original dataset, and this minimal representation is useful in classification analysis for the entire dataset[5].

In 2015 Qing-Zhao Konga,b, and Zeng-Xin Weic proposed “Covering-based fuzzy rough sets”. They investigates the properties of covering-based fuzzy rough sets. In addition, operations of intersection, union and complement on covering-based fuzzy rough sets are investigated. Finally, the corresponding algebraic properties are discussed in detail. They proposed the covering-based fuzzy rough set model and discussed its corresponding properties. Although many researchers have studied many properties of rough sets, the operations of intersection, union and complement on rough sets have yet to be investigated. They proposed the concept of monotone covering and researched the operations of intersection, union and complement on covering-based fuzzy rough sets. Thus, the construction of the covering-based fuzzy rough set model is a meaningful generalization of rough set theory[6].

In 2016 Qinghua Zhang ,Qin Xie a proposed “A survey on rough set theory and its applications”. They proposed the basic concepts, operations and characteristics on the rough set theory are introduced firstly, and then the extensions of rough set model, the situation of their applications, some application software and the key problems in applied research for the rough set theory are presented. The rough set theory has been researched for more than thirty years. And it has made many achievements in many fields, such as machine learning, knowledge acquisition, decision analysis, knowledge discovery in database, expert system, decision support system, inductive inference, conflict resolution, pattern recognition, fuzzy control, medical diagnostics applications and so on. And for research on granular computing, it has become one of the main models and tools [7].

In 2017 Abbas Mardani and Mehrbakhsh Nilashi proposed “Recent Fuzzy Generalizations of Rough Sets Theory: A Systematic Review and Methodological”. Rough set theory has been used extensively in fields of complexity, cognitive sciences, and artificial intelligence, especially in numerous fields such as expert systems, knowledge discovery, information system, inductive reasoning, intelligent systems, data mining, pattern recognition, decision-making, and machine learning. Rough sets models, which have been recently proposed, are developed applying the different fuzzy generalisations. Currently, there is not a systematic literature review and classification of these new generalisations about rough set models. Therefore, in this review study, the attempt is made to provide a comprehensive systematic review of methodologies and applications of recent generalizations discussed in the area of fuzzy-rough set theory[8].

In 2018 Zhenquan Shi and Shiping Chen proposed “A New Knowledge Characteristics Weighting Method Based on Rough Set and Knowledge Granulation”. The current rough set weighting methods could not obtain a balanced redundant characteristic set. Too much redundancy might cause inaccuracy, and less redundancy might cause inefectiveness. They proposed new method based on rough set and knowledge granulation theories is proposed to ascertain the characteristics weight. Experimental results on several UCI data sets demonstrate that the weighting method can effectively avoid subjective arbitrariness and avoid taking the non-redundant characteristics as redundant characteristics. They proposed approach based on rough set theory and knowledge granularity theory, the weights of knowledge characteristics is determined. Experimental results show that the proposed method can effectively avoid taking non-redundant characteristics as redundant characteristics and can effectively determine the weights of knowledge characteristics[9].

In 2019 Shoubin Sun , Lingqiang Li proposed “A New Approach to Rough Set Based on Remote Neighborhood Systems”. Te notion of remote neighborhood systems is initial in the theory of topological molecular lattice, and it is abstracted from the geometric notion of “remote”. Terefore, the notion of remote neighborhood systems can be considered as the dual notion of neighborhood systems. In this paper, we develop a theory of rough set based on remote neighborhood systems. They construct a pair of lower and upper approximation operators and discuss their basic properties. In Furthermore, they use a set of axioms to describe the lower and upper approximation operators constructed from remote neighborhood systems[10].

In 2020 A. K. Sinha “Mathematical Modeling Of Lung Cancer Using Rough Sets”. They presents the dynamics of tumor cells' growth with anti-tumor treatment. It deals with the applications of the Rough set method in the patterns of life expectancy in lung tumors in uncertain situations through information knowledge and data intensive computer-based solutions. This realistic clinical data evaluation strategy shows that the system performance accuracy for the pattern of life expectancy in lung tumors is 98.00 % by the Rough set method, whereas the accuracy found in other ways (ANN, Boosted SVM) was less in the previous studies This article exhibits the dynamics of tumor cells' growth with anti-tumor treatment. It deals with the importance of the Rough set in the patterns of life expectancy in lung tumors in uncertain conditions through information science and data-intensive computer-based solutions. The Rough set is the appropriate scientific method to deal with imperfect knowledge and uncertainty and a reliable way for the patterns of life expectancy in lung tumors. This good clinical data evaluation strategy reveals that the system performance accuracy for the pattern of life expectancy in lung tumors better than the accuracy found in other ways in the previous studies[11].

V. Problem Statement

1. The uncertainty in the dataset is the major problem to get complete information of a particular attribute or to develop an Expert system to retrieve accurate information from the existing one.
2. Due to the digital transmission of information, information systems usually have some missing values. The causes are due to unavailability of data or after processing the data the information may lost or there will be an ambiguity.
3. Missing values give erroneous classification rules generated by a data mining system. It influences the percentage coverage and the number of rules generated and lead to the difficulty of extracting useful information from the data set.
4. Even a small amount of missing data can cause serious problems with the analysis leading to draw wrong conclusions and imperfect knowledge. There are many techniques developed to manipulate the knowledge with uncertainty and manage data with incomplete items, but are no such results came, and sometimes the results are not of the similar type and absolutely better than the others.

To handle such problems, researchers are trying to solve it in different approaches and then proposed to handle the information system in their way. It is observed from the experience that the attribute values are more important for information processing from a data set or information table.

VI. Proposed Approach

The steps that algorithm follows to predict the missing value are

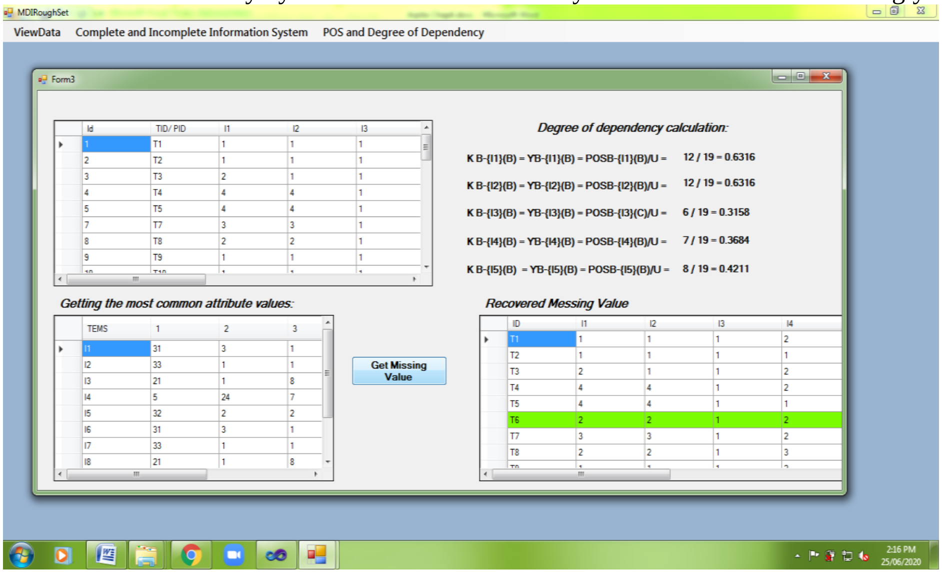
1. Separation the decision table to two tables (complete information system table and incomplete information system table)
2. Getting the most common value of each attribute.
3. Calculation of Degree of dependency:
   - The model calculates Indiscernibility relation for the complete attribute. The model calculates Indiscernibility relations for the complete attribute except for each attribute individual.
   - The model calculates the POS`s of the complete attribute except for each attribute individual.
   - The model calculates the degree of dependency by dividing each by the .The degree of dependency is between 0 and 1 POS U
   - The model eliminates the attribute b if is the biggest k.
4. Calculations of the distance function between every case in incomplete information system table and complete information system table.
5. Getting the smallest distance for every case in the incomplete information system table.
6. If the smallest distance unique, then the missing attribute value equal the value of the same attribute which its case has the smallest distance.
7. If the smallest distance is repeated and its records of complete information system table have the same value of the missing attribute, then the missing attribute value equal this repeated attribute value.
8. If the smallest distance is repeated in more records in the complete information system table and the records have different values of the missing attribute, then:
   a) The attribute which has small effects on the information system table will be done by using the degree of dependency.
   b) Calculation of the distance is repeated again and the above sequence will be repeated but with only the cases of complete information system table which have the smallest distance.
9. If there is no matching case, the algorithm supposes that the missing attribute value is the most common attribute value of this attribute.

VII. Experimental Analysis

We evaluate the performance of Rough Set. We implement proposed approach by using VB dot net 2013 as front end for user interface. In the implementation of proposed approach there are various option are given to the user, user can select option as per the requirement working. In the implementation we divide the data set into two parts complete data set with no missing values and incomplete data set with missing values. We apply properties of Rough Set and finally we construct the missing values. We used to existing statistical methods to compare the performance of the proposed approach. We used statistical method like replacing missing value by most common value, replacing missing value by mean etc.

We have taken 1000 transaction and 10 products for experimental analysis. We used SQL server 2010 to store data set. We used different number of transaction with different number of missing values to check the accuracy of the proposed approach and statistical method.

We have taken 500 records with 19 missing values 1000 records with 26 missing values and 1500 records 35 missing values. We apply all three methods proposed approach, replacing with most common value and replacing by mean. We compare how many missing value are constructed correctly by all methods and how many values are constructed wrongly.


MDIRoughSet
ViewData
Complete and Incomplete Information System
POS and Degree of Dependency
Form3
Get Missing Value
2:16 PM
25/06/2020

Id	TID/ PID	I1	I2	I3
1	T1	1	1	1
2	T2	1	1	1
3	T3	2	1	1
4	T4	4	4	1
5	T5	4	4	1
7	T7	3	3	1
8	T8	2	2	1
9	T9	1	1	1

Degree of dependency calculation:

K B-{I1}(B) = YB-{I1}(B) = POSB-{I1}(B)/U = 12 / 19 = 0.6316

K B-{I2}(B) = YB-{I2}(B) = POSB-{I2}(B)/U = 12 / 19 = 0.6316

K B-{I3}(B) = YB-{I3}(B) = POSB-{I3}(C)/U = 6 / 19 = 0.3158

K B-{I4}(B) = YB-{I4}(B) = POSB-{I4}(B)/U = 7 / 19 = 0.3684

K B-{I5}(B) = YB-{I5}(B) = POSB-{I5}(B)/U = 8 / 19 = 0.4211

Getting the most common attribute values:

TEMS	1	2	3
I1	31	3	1
I2	33	1	1
I3	21	1	8
I4	5	24	7
I5	32	2	2
I6	31	3	1
I7	33	1	1
I8	21	1	8

Recovered Messing Value

ID	I1	I2	I3	I4
T1	1	1	1	2
T2	1	1	1	1
T3	2	1	1	2
T4	4	4	1	2
T5	4	4	1	1
T6	2	2	1	2
T7	3	3	1	2
T8	2	2	1	3

Figure 2 Constructing missing values rough set approach

Table 1 Number of records and missing values constructed correctly

Number of missing value for	Rough set based approach	Replacing missing values with most common values
500 (19)	16	10
1000 (26)	21	13
1500 (35)	29	19

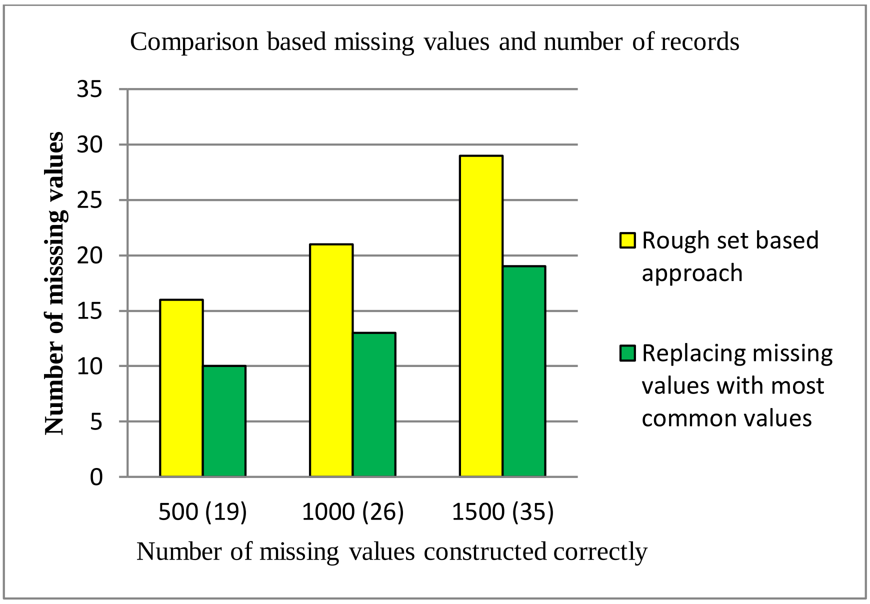


Figure 3 Number of records and missing values constructed correctly

VIII. Conclusion

The use of rough set theory to solve a specific complex problem and has attracted world-wide attention of further research and development, extending the original theory and increasingly widening fields of application. Rough set as a computationally efficient technique it presents a basic significance to many theoretical developments and practical applications of computing and automation, especially in the areas of machine learning and data mining, decision analysis and intelligent control. In the proposed work we use rough set theory to construct missing values for incomplete information system. We some of the properties of rough set theory and construct missing value we found that proposed approach construct missing values more correctly as compared some of the statistical method . With the experimental analysis we compared performance of proposed approach with statistical method

References

[1] In 2010 Yiyu Yao "Three-way decisions with probabilistic rough sets" Information Sciences, Vol. 180, No. 3, pp. 341-353, 2010.

[2] Yao, Y.Y., Sequential Three-way Decisions with Probabilistic Rough Sets, 10th IEEE International Conference on Cognitive Informatics and Cognitive Computing, pp. 120-125, 2011.

[3] Feng Hu and Guoyin Wang "Knowledge Reduction Based on Divide and Conquer Method in Rough Set Theory" Hindawi Publishing Corporation Mathematical Problems in EngineeringVolume 2012, Article ID 864652, 24 .

[4] In 2013 Thabet Slimani " Application of Rough Set Theory in Data Mining College of Computer Science and Information Technology, Taif University.

[5] K Anitha and P Venkatesan "Rough Set Theory Approach to Generating Classification Rules" International Journal of Computational Intelligence and Informatics, Vol. 4: No. 3, October - December 2014.

[6] Qing-Zhao Konga,b, and Zeng-Xin Weic "Covering-based fuzzy rough sets" Journal of Intelligent & Fuzzy Systems 29 (2015) 2405–2411 DOI:10.3233/IFS-151940 IOS Press.

[7] Qinghua Zhang , Qin Xie , Guoyin Wang A survey on rough set theory and its applications Available online at www.sciencedirect.com Science Direct CAAI Transactions on Intelligence Technology (2016) 323e333 http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology

[8] Abbas Mardani, Mehrbakhsh Nilashi, "Recent Fuzzy Generalisations of Rough Sets Theory: A Systematic Review and Methodological Critique of the Literature" Hindawi Complexity Volume 2017, Article ID 1608147, 33 pages https://doi.org/10.1155/2017/1608147.

[9] Zhenquan Shi1, and Shiping Chen " A New Knowledge Characteristics Weighting Method Based on Rough Set and Knowledge Granulation" Hindawi Computational Intelligence and Neuroscience Volume 2018, Article ID 1838639, 9 pages https://doi.org/10.1155/2018/1838639.

[10] Shoubin Sun , Lingqiang Li ,and Kai Hu2 "A New Approach to Rough Set Based on Remote Neighborhood Systems" Hindawi Mathematical Problems in Engineering Volume 2019, Article ID 8712010, 8 pages https://doi.org/10.1155/2019/8712010

[11] A. K. Sinha and Nishant Namdev "Mathematical Modeling Of Lung Cancer Using Rough Sets" International Journal of Advanced Research in Engineering and Technology (IJARET) Volume 11, Issue 3, March 2020, pp. 1-9, Article ID: IJARET Available online at Journal Impact Factor (2020): 10.9475 (Calculated by GISI) www.jifactor.com ISSN Print: 0976-6480 and ISSN Online: 0976-6499