ISSN:0975-9646

# Retrieving the Missing Information from Information Systems Using Rough Set, Covering Based Rough Set and Soft Set

B.S.Panda[1], S.S.Gantayat[2], Ashok Misra[3]

[1]*Research Scholar, CUTM, India*
[2]*GMRIT, Rajam, AP, India*
[3]*CUTM, Paralakhemundi, Odisha, India*

**Abstract:In this paper, we study the various aspects of rough set theory, covering-based rough set and soft set theory to handle the missing information systems. The rough set theory, based on the indiscernibility relation, it helps to develop automated computational systems using mathematical model that can help to understand and to handle imperfect knowledge. Covering based rough set is an extension of the basic rough set. Soft set theory can be applied to problems that contain uncertainties in decision making problems. These concepts are applied to a real life malaria disease dataset to replace missing information in the malaria disease information table and compared the results.**

***Key words: Rough Set, Covering Based Rough Set, Soft Set, Missing information.***

## 1. INTRODUCTION

According to Little and Rubin [7], real time processing applications are highly dependent upon the data which causes the problem of missing input variables. Various heuristic methods of missing data imputation such as mean substitution and hot deck imputation also depend on the knowledge of the propagation of missing data. There are several reasons why the data may be missing, and as a result, missing data may follow an observable pattern.

Now-a-days 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. 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. 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.

In this paper we have tried to bridge the gap between generalized rough set based on coverings, and the possible-world approach to missing information. Each set in a covering is then the upper inverse image of an attribute value through a multiple-valued mapping, which describes incomplete knowledge on attribute values. Sometimes the set of objects are possibly indistinguishable. Due to the presence of uncertainty, in rough set the upper and lower approximations are ill-known, each being represented by two nested set. This interpretive setting leads us to choose, among possible covering-based rough set and soft set, some of them that look more appropriate than others.

## 2. BASIC CONCEPTS

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 in literature 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[8,9,10].

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. In the field of databases, various efforts have been made for the improvement and enhance of database or information table query process to retrieve the data. The methodology followed by different approaches like, Rough set [1], Covering based Rough set and [12], Soft set [13] and Statistically Similarity [14] etc.

## 3. ROUGH SET

Rough set, introduced by Pawlak [1, 15], is a notion of describing the objects based on the precise observations of attributes. The limited expressivity of the language defined by these attributes prevents set of objects. Only upper and lower approximations of set can be characterized in terms of unions of equivalence classes of the equivalence relations defined by the attributes. The set of objects defined by a property expressed in the language of the attributes may be ill-known for the attribute values of objects are imprecise or uncertain. In this case, the attributes become set-valued mappings. However, Dubois and Prade [2], shown that these set represent imprecision mutually exclusive values, one of which is the actual attribute value of the object under consideration. In this case, a set of objects is only approached in the form of an upper and a lower approximation which means the set of objects that possibly and necessarily satisfy the property.

Slowinski and Vanderpooten[3],have generalized the notions of lower and upper approximations of the Rough set by considering the reflexive binary relations instead of equivalence relations. They have extended many of the concepts related to basic rough set to this general setting. Also, adding the symmetry and transitivity of the relations separately and jointly a comparative study of these models have been made with the existing structures of rough set.

### 3.1 Definition and Notations

As per the definition of Rough set by Pawlak in [15], let R⊆UxU denote an equivalence relation on U, that is, R is a reflexive, symmetric and transitive relation. The equivalence class of an element x∈U with respect to R is the set of elements y∈U such that xRy. If two elements x, y in U belong to the same equivalence class, then we can say that x and y are indistinguishable with respect to relation R.

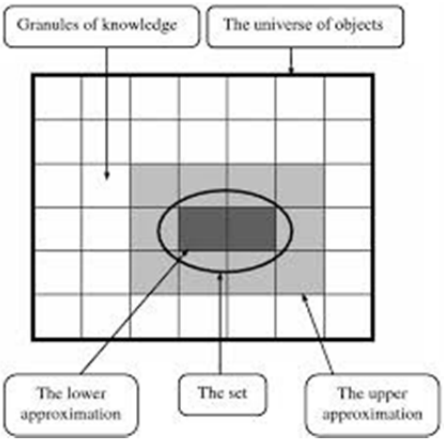


Fig-1. Rough Set Approach

Given an arbitrary set A⊆U, it may not be possible to describe 'A' precisely in the approximation space apr(R)=(U, R). But we may only characterize 'A' by a pair of lower and upper approximations. This leads to the concept of rough set. We define,

$$\underline{R}A=\bigcup\{Y\in U/R : Y\subseteq A\};$$

and $$\overline{R}A=\bigcup\{Y\in U/R : Y\bigcap A\neq\varphi\}.$$

$\underline{R}A$ and $\overline{R}A$ are respectively called the R-lower and R-upper approximation of A with respect to R.
It can be noted that

$$\underline{R}A=\{x\in U:[x_R]\subseteq A\}$$

and

$$\overline{R}A=\{x\in U:[x]_R\bigcap X\neq\varphi\}.$$

The set $BN_R(A)=\overline{R}A-\underline{R}A$ is called the *R-boundary* of A. The set $\underline{R}A$ contains the elements of U with certainty be classified as elements of A, employing the knowledge R. The set$\overline{R}A$ consists of all those elements of U which can possibly be classified as elements of A, employing the knowledge R. Set $BN_R(A)$is the set of elements which cannot be classified as either belonging to A or belonging to –A having the knowledge R. We say that a set A is *R-definable* if and only if $\underline{R}A=\overline{R}A$ . Otherwise A is said to be *R-rough*.
The borderline region is the un decidable area of the universe. We say that *X* is *rough* with respect to *R* if and only if$\underline{R}X\neq\overline{R}X$ , equivalently$BN_R(X)\neq\phi$. *X* is said to be R-*definable* if and only if$\underline{R}X=\overline{R}X$ or $BN_R(X)=\phi$
.

### 3.2 Application of Rough Set

We briefly highlight few applications of Rough set theory [16, 17, and 18]. The following applications show the use of Rough set.

- Identification and evaluation of data dependencies.
- Approximate pattern classification.
- Reasoning with uncertainty.
- Information-preserving data reduction.
- Knowledge analysis.
- Analysis of conflicts.

## 4. Covering Based Rough Set

The generalized notion of partition is a cover. Covering based rough set has been introduced by Zakowski [4]. The covering based rough set models have the promising potential for applications to data analysis. We know that there is only one lower approximation in rough set. However, four different types of upper approximations are in Rough set are introduced. This led to do comparison of the different types of rough sets generated using the four upper approximations.
There is a strong similarity between rough set and covering based set. However, the origin of uncertainty is completely different. In the case of rough set, attribute values are known but there are not enough attributes (or attribute domains are too coarse) to identify a single object by their descriptions using these attributes. In the case of covering based set there may be enough attribute values, but the lack of knowledge about objects prohibits a precise numeration of the contents of set defined by means of properties. These sources of uncertainty being unrelated and can be simultaneously present in different attributes. These problems are carried out in covering-based extensions of rough set proposed by Zhu [19, 20, and 21].

### 4.1 Definition and Notations

The following definitions of covering based Rough sets given by Zhu [19, 20, and 21]

**Definition 4.1.1:**Let U be a universe of discourse and C be a family of subset of U. C is called a cover of U if no subset in C is empty and ∪C = U. We call (U,C) the covering approximation space and the covering C is called the family of approximation set.

**Definition 4.1.2:** Let (U,C) be an approximation space and x be any element of U. Then the following family is called the minimal description of the object x.

$$Md\ (x) = \{K\in C: x\in K\},\ S\in C$$

**Definition 4.1.3**: For any set X□$\subseteq$ U, the family of set C*(X)= {K $\in$ C: K $\subseteq$ X} is called the family of set bottom approximating the set X.

**Definition 4.1.4:** The set X* = $\cup$C*(X) is called the lower approximation of the set X.

## 5. SOFT SET

The theory of soft set proposed, by Molodtsov in 1999, is a new method for handling uncertain data[4, 5].Soft set are called either binary or basic or elementary neighborhood systems [6]. The soft set is a mapping from parameter to the crisp subset of universe. The structure of a soft set can classify the objects into two classes either yes/1 or no/0. This means that the “standard” soft set deals with a Boolean-valued information system. It can be applied to data analysis and decision support systems. Reduct is a fundamental notion that supports the applications. The drawback of this notion is that it is applicable only for Boolean-valued information systems. Soft set can be used as a generic mathematical tool for dealing with uncertainty. Soft set is a parameterized general mathematical tool which deals with a collection of approximate descriptions of objects. Each approximate description has two parts one is a predicate and another one is an approximate value set. So, Molodtsov introduced the notion of approximate solution. The initial description of the object has an approximate in nature, and we do not need to introduce the notion of exact solution to a problem. The absence of any restrictions on the approximate description in soft set theory is very convenient and easily applicable in practical real life problems. For any parameterization, we can prefer any words and sentences, real numbers, functions, mappings and so on.

Maji et al. [13] proposed the idea of reduct and decision making using soft set theory. The application of soft set theory to a decision making problem was considered with the help of Pawlak’s rough set concept, which uses the maximal weighted value among objects related to the parameters.

Zou proposed a new technique for decision making of soft set theory under incomplete information systems [22]. The notion is based on the calculation of weighted-average of all possible choice values of object and the weight of each possible choice value is decided by the distribution of other objects in the information table.

### 5.1 Definition and Notations

The following definition are considered from Molodtsov [4,5] and Maji et al. [13].

**Definition 5.1.1:** A pair (F,A) is called a soft set over an universal set U where F is a mapping given by

$$F: A \rightarrow P(U)$$

Here we consider that a soft set over U is a parameterized family of subset of the universe U. For X$\in$A, F(X) may be considered as the set of X-elements of the soft set (F,A) or as the set of X-approximate elements of the soft set.

**Definition 5.1.2:** Let S = (U, A, V, f) be an information system and let B be any subset of A. Two elements x, y$\in$U are said to be B-indiscernible (indiscernible by the set of attribute B $\in$ A in S) if and only if f (x, a) = f (y, a), for every a $\in$B.

**Definition 5.1.3:** Let S = (U, A, V, f) be an information system and let B be any subset of A. A rough approximation of a subset X$\in$U with respect to B is defined as a pair of lower and upper approximations of X, i.e. $\{\underline{B}(X), \overline{B}(X)\}$

**Definition 5.1.4:** Let S = (U, A,V, f ) be an information system and let B be any subset of A in information system S. Attribute b$\in$B is called dispensable if

$$U/(B -\{b\}) = U / B$$

### 5.2. Applications of Soft Set

Soft set theory has potential applications in many different fields which include the smoothness of functions, game theory, operations research, Riemann integration, Peron integration, probability theory, and measurement theory and so on [4, 5]. Also, application of soft set theory to the problems of medical diagnosis in medical expert system was discussed. Maji et al.[13] gave first practical application of soft set in decision making problems. It is based on the notion of knowledge of reduction in rough set theory. N. Cagman and S. Enginoglu [24] defined soft matrices and their operations to construct a soft max-min decision making method which can be successfully applied to the problems that contain uncertainties. T. Herawanet al., [25], gave an alternative approach for attribute reduction in multi-valued information system under soft set theory. In their work they had shown that the reducts obtained using soft set are equivalent with Pawlak’s rough reduction.

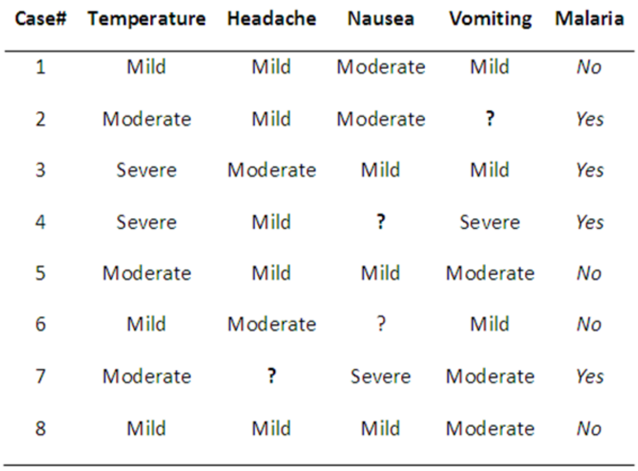
## 6. MISSING ATTRIBUTE VALUES

Missing attribute values commonly exist in real world data set. They may come from the data collecting process or redundant diagnose tests, unknown data and so on. Discarding all data containing the missing attribute values cannot fully preserve the characteristics of the original data. Various approaches on how to cope with the missing attribute values have been proposed in the past years [29, 30]. Various techniques to resolve missing information or data also carried out by different researchers [9, 10, 22, 27, 28].

The table-1, show the attributes *Temperature*, *Headache*, *Nausea and Vomiting* and with the decision *Malaria*. However, many real-life dataset are incomplete. The missing attribute value is denoted by "?".

## 7. RESULTS AND DISCUSSION

**Example 7.1:** To implement the concept of Rough set rule based technique to simplifying the diagnosis of malaria and impute the missing attributes out of 20 real data set with ten attributes collected from different doctors. we have considered 8 data set with four attributes. The technique used on knowledge of domain experts (five medical doctors), applied with rough set theory. Rough set classification is utilized to remove uncertainty, ambiguity and vagueness inherent in medical diagnosis.



| Case# | Temperature | Headache | Nausea | Vomiting | Malaria |
|---|---|---|---|---|---|
| 1 | Mild | Mild | Moderate | Mild | *No* |
| 2 | Moderate | Mild | Moderate | **?** | *Yes* |
| 3 | Severe | Moderate | Mild | Mild | *Yes* |
| 4 | Severe | Mild | **?** | Severe | *Yes* |
| 5 | Moderate | Mild | Mild | Moderate | *No* |
| 6 | Mild | Moderate | ? | Mild | *No* |
| 7 | Moderate | **?** | Severe | Moderate | *Yes* |
| 8 | Mild | Mild | Mild | Moderate | *No* |

Table -1. Data set with missing attributes

From the above the different classes can be generated using rough set concept as follows.

Temperature = {{C1,C6,C8}, {C2,C5,C7}, {C3,C4}}
Headache = {{C1,C2,C4,C5,C8},{C3,C6},**{C7}**}
Nausea = {{C1,C2},{C3,C5,C8} ,**{C4,C6},**{C7}}
Vomiting= {{C1,C3,C6},{C**2**},{C4},{C5,C7,C8}
Malaria = {{C1,C5,C6,C8},{C2,C3,C4,C7}}

In the above classification the bold classes are the missing data, which will be filled in the currently proposed technique.
The bold word in the above table shows the replacement of approximated information with the missing data in the Table-1.
After replacing proper values to the missing data, we get the different classes from the above table as follows.

Temperature = {{C1,C6,C8}, {C2,C5,C7}, {C3,C4}}
Headache = {{C1,C2,C4,C5,C8},{C3,C6,C**7**}}
Nausea = {{C1,C2},{C3,C5,C**6, C**8},{C**4**,C7}}
Vomiting = {{C1,C3,C6},{C4},{C**2**,C5,C7,C8}}
Malaria = {{C1,C5,C6,C8},{C2,C3,C4,C7}}

**Example 7.2:**
Using the Covering based Rough set, the following result is discussed.
The covering lower approximation is defined as
$X_{*}=\cup\{K_B(x)/ K_B(x)\in C \text{ and } K_B(x)\subseteq X\}$ and
The covering upper approximation is defined as
$X^{*} = \cup\{\text{neighbor}(x)/x\in X\}$
From table-1 we consider
U={C1,C2,C3,C4,C5,C6,C7,C8}and
X={C2,C4,C6,C7}then
*neighbor*(C2)={moderate, mild, moderate},
*neighbor*(C4) = {severe, mild, severe},
*neighbor*(C6) = {mild, moderate, mild} and
*neighbor*(C7) = {moderate, severe, moderate}
then *X**= {moderate, severe, mild, moderate}

**Example 7.3:**
The proposed approach is based on Soft Set to impute the missing attributes from the table-1.We have analyzed the relations between the attributes and define the notion of association degree to measure the relations. In our method, we give priority to the relations between the attributes due to its higher reliability. When the mapping set of an attribute includes incomplete data, we firstly look for another attribute which has the stronger association with the parameter.

( F,C2)= {{Temperature =2}, {Headache=1},
{Nausea =2}, **{**Vomiting= ?}}
( F,C4)= {{Temperature =3}, {Headache=1},
{Nausea = ? }, **{**Vomiting= 3}}
( F,C6)= {{Temperature =1}, {Headache=2},
{Nausea = ? }, **{**Vomiting= 1}}
( F,C7)= {{Temperature =2}, {Headache= ?},
{Nausea =3}, **{**Vomiting=2}}

Here mild =1 moderate = 2, Severe=3.
We have computed elementary set

( F,C2)= {{Temperature =2} ∩ {Headache=1} ∩
{Nausea =2} ∩ **{**Vomiting= ?}} = {2}
( F,C4)= {{Temperature =3} ∩ {Headache=1}∩
{Nausea = ? } ∩ **{**Vomiting= 3}} = {3}
( F,C6)= {{Temperature =1} ∩{Headache=2}∩
{Nausea = ? } ∩ **{**Vomiting= 1}}={1}
( F,C7)= {{Temperature =2} ∩ {Headache= ?}∩
{Nausea =3}∩ **{**Vomiting=2}}={2}

After computation the result set become {C2},{C3},{C1} and {C2}

The final result of the imputation of missed data, from above analysis is given below.

| Case # | Common Attribute | Actual Attribute | Diagnosis |
|---|---|---|---|
| C2 | **Moderate** | Mild | Yes |
| C4 | **Severe** | Moderate | Yes |
| C6 | **Mild** | Mild | No |
| C7 | **Moderate** | Mild | Yes |

Table 2.The Missing Data Filled with Observed Data

## 8. Conclusion

This paper summarized the basic concepts of rough set, covering based set and soft set the manner in which rough set are related to covering based set and soft set. We then presented a detailed theoretical study of soft set, which led to the definition of missing data handling. If the mapping set of an attribute includes incomplete data, we filled the data according to the value in the corresponding attributes. This work focused on imputes the missing values through the above techniques. To extend this work, one could study the properties of rough set and soft set. The methods can be used to handle various applications involved incomplete information system.

## REFERENCES

[1] Pawlak, Z., Rough Set-Theoretical Aspects of Reasoning about Data, Kluwer Acad. Publ., 1991.

[2] D. Dubois, H. Prade, "TwoFold Fuzzy Set and Rough Set -Some Issues in Knowledge Representation", Fuzzy Set and Systems, 23, pp. 3-18, 1987.

[3] Slowinski, R. and Vanderpooten, D., "Similarity Relation as a Basis for Rough Approximations", in P.P. Wang, Editor, Advances in Machine Intelligence & Soft-Computing, Vol. IV, Duke University Press, Durham, NC, pp. 17-33,1997.

[4] Zakowski, W.,"Approximation in Space (U, Π)", Demonstratio Mathematica, 16, pp. 761-769, 1983.

[5] Molodtsov,D. "Soft Set Theory-First Results". Computers and Mathematics with Applications. 37, pp. 19–31, 1999.

[6] Yao, Y.Y. "Relational Interpretations of Neighborhood Operators and Rough Set Approximation Operators", Information Sciences, 111, pp. 239–259, 1998.

[7] Little, R.J. A., Rubin, D. B. Statistical Analysis with Missing Data. 2nd edition, New York: John Wiley, 2002.

[8] Wohlrab Lars, FürnkranzJohannes. "A Review and Comparison of Strategies For Handling Missing Values in Separate-and-Conquer Rule Learning", Journal of Intelligent Information Systems, 36(1), pp 73-98,February 2011.

[9] S. S. Gantayat, Ashok Misra, B. S. Panda, "A Study of Incomplete Data – A Review", In: FICTA-2013, LNCS-Springer, pp. 401-408, 2013.

[10] Panda, B. S.,Gantayat, S. S., Misra, Ashok, "Rough Set Rule Based Technique for the Retrieval of Missing Data in Malaria Diseases Diagnosis" In: Springer FMB Series, 2014.

[11] Maji,P.K., Roy,A.R., Biswas,R., "An Application of Soft Set in a Decision Making Problem", Comput. Math. Appl. 44, 2002.

[12] Tripathy, B. K. and Mitra, A., "Some Topological properties of Covering based Rough Set", International Conference on Emerging Trends in Computing {ICETiC), January, 2009.

[13] P.K. Maji, R. Biswas, A.R. Roy, "Soft Set Theory", Comput. Math. Appl., 45, pp. 555–562,2003.

[14] Grzymala-Busse, J., "Knowledge Acquisition under Uncertainty – A Rough Set Approach", J. Intelligent and Robotics Systems, 1, pp. 3-16, 1988.

[15] Pawlak, Z., "Rough Set", J. Inf. Comp. Sc. II, pp. 341- 356, 1982.

[16] Pawlak, Z., and Skowron, A., "Rough Set- Some Extensions", Information Sciences. 177(1), pp. 28-40, 2007.

[17] Panda, B. S., Rahul Abhishek and Gantayat, S. S., "Uncertainty Classification of Expert Systems - A Rough Set Approach". in ISCON proceedings with IJCA, ISBN: 973-93-80867-87-0, 2012.

[18] Panda B. S., Gantayat S. S., Misra Ashok, "Rough Set Approach to Development of a Knowledge-Based Expert System" In: International Journal of Advanced Research in Science and Technology (IJARST), 2(2), pp.74-78, 2013.

[19] Zhu, W., "Basic Concepts in Covering-based Rough Set", IEEE-Third International Conference on Natural Computation (ICNC), 2007.

[20] Zhu, W., "Topological Approaches to Covering Rough Set", Information Sciences, 177, pp. 1499-1508, 2007

[21] Zhu, W., "On Three Types of Covering-Based Rough Set", IEEE Transactions on Knowledge and Data Engineering, 19(8), pp. 1131-1143, August 2007,

[22] Zou, Y. and Xiao, Z. "Data Analysis Approaches of Soft set under Incomplete Information", Knowledge Based Systems, 21, pp. 941–945, 2008.

[23] H. Yang and Z. Guo, "Kernels and Closures of Soft set Relations, and Soft set Relation Mappings", Comp. Math. Appl. 61, pp. 651-662 2011.

[24] N.Cagman, F.Citak, and S.Enginoglu. "FP-Soft Set Theory and its Applications", Annals of Fuzzy Math. Inform. 2(2), pp. 219-226, 2011.

[25] Herawan,T., Ghazali,R. and Deris,M. M., Soft set Theoretic Approach for Dimensionality Reduction, Information Journal of Database Theory and Application 3(2), 2010.

[26] Chen, D., Tsang, E.C.C., Yeung, D.S., and Wang, X. "Some Notes on the Parameterization Reduction of Soft set". Proceeding of International Conference on Machine Learning and Cybernetics, IEEE Press,3, pp.1442–1445, 2003.

[27] Myers William R., "Handling Missing Data in Clinical Trials: An Overview", Drug Information Journal, 34, pp. 525–533, 2000.

[28] ShalabiLuai Al,Najjar Mohannad and Kayed Ahmad Al, "A Framework to Deal with Missing Data in Data Set",Journal of Computer Science, 2 (9), pp.740-745, 2006.