CHAPTER 6

Approaches to handle Missing Information System

6.1 Introduction

(Rough sets Pawlak 1982)\textsuperscript{103} are often induced by descriptions of objects based on the precise observations of attributes. The limited expressivity of the language defined by these attributes prevents sets of objects defined in extension from being precisely described in this language. Only upper and lower approximations of sets can be characterised in terms of unions of equivalence classes of the equivalence relation defined by the attributes. On the other hand, sets of objects defined by a property expressed in the language of the attributes may be ill-known when the attribute values of objects are imprecise. In this case, the attributes become set-valued mappings. However, these sets represent imprecision, i.e. they contain mutually exclusive values, one of which is the actual attribute value of the object under concern. Again in this case, a set of objects is only approached via an upper and a lower approximation that is the sets of objects that possibly and necessarily satisfy the property, respectively (Dubois and Prade 1987)\textsuperscript{65}.

(Slowinski and Vanderpooten 1997)\textsuperscript{142} have generalized the notions of lower and upper rough approximations of a set by taking reflexive binary relations instead of equivalence relations. They have extended many of the concepts related to basic rough sets 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 sets.

A cover is a generalization of the notion of partition. Using covers instead of partitions, covering based rough sets have been introduced by (Zakowski 1983)\textsuperscript{162}. The covering based rough sets are models with promising potential for applications to data mining. There is only one lower approximation for such rough sets. However, there are as many as four different versions for the definition of upper approximation of such sets. Also, this has led to comparison of the different types of rough sets thus generated. Several properties of the different types of rough sets have been derived by different researchers the notion of kinds of rough sets is a topological property.

The theory of soft set (Molodtsov 1999)\textsuperscript{90} proposed, by a new method for handling uncertain data. Soft sets are called (binary, basic, elementary) neighborhood systems (Yao, Y.Y. 1998)\textsuperscript{158}. The soft set is a mapping from parameter to the crisp subset of universe. From such case, see the structure of a soft set can classify the objects into two classes (yes/1 or no/0). This means that the “standard” soft set deals with a Boolean-valued information system. The theory of soft set has been applied to data analysis and decision support systems. A fundamental notion supporting such applications is the concept of reducts. The restriction of those techniques is that they are applicable only for Boolean-valued information systems.

According to (Little and Rubin 2000)\textsuperscript{81} Real time processing applications that are highly dependent on the data often suffer from the problem of missing input variables. Various heuristics of missing data imputation such as mean substitution and hot deck imputation also depend on the knowledge of how data points become missing. 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 datasets is the major problem to get complete information of a particular object or to develop an Expert system to retrieve accurate information from the existing one. Most information systems usually have some missing values due to unavailability of data or after processing the data. Missing values minimizes the quality and quantity of classification rules generated by a data mining system. These values could influence the coverage percentage and number of rules generated and lead to the difficulty of extracting useful information from the data sets.

In this chapter try to bridge the gap between generalized rough sets 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 describing incomplete knowledge on attribute values. It is the set of objects that are possibly indistinguishable. Due to the presence of two sources of uncertainty, the rough set upper and lower approximations of a set are ill-known, each being bracketed by two nested sets. This interpretive setting leads us to choose, among possible covering-based rough sets, soft sets definitions, some of them that look more appropriate than others.

6.2 Literature Review

Even a small percentage 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 no results, and the results are not of the same type and absolutely better than the others (Little R. J. and Rubin D. B., 2002)\(^2\).

To handle such problems, researchers are trying to solve it in different approaches and then proposed to handle the information system in their way. The attribute values are important for information processing in 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 sets (Grzymala-Busse, J. 1988)\(^39\), Covering based sets (Tripathy, B. K. et.al 2009)\(^149\) and Soft sets (P.K. Maji, et.al 2002, 2003)\(^99\) and 100 and Statistically Similarity etc.

6.3 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 (Myers William R., 2000)\(^91\) and (Shalabi Luai Al 2006)\(^139\).

Statisticians have identified three classes of missing data. The easiest situation is when data are missing completely at random (MCAR). This means that the fact that a value is missing is unrelated to its value or to the value of other variables. When data are MCAR, the probability that a variable is missing is the same for every record. If the probability that a value is missing depends only on the value of other variables, we say that it is missing at random (MAR). In term of probabilities, if we have a variable \(Y\) with missing values, and another variable \(X\), the data are MAR. If the missingness depends on the missing value, data are not missing at random (NMAR), and this is a problem for many statistical Missing Data Techniques. This happens, for instance, when we collect data with a sensor which is not able to detect values over a particular threshold.
If we do not want to lose data and perhaps information, we may try to guess missing items. This process is generally called imputation. In fact, usually missing values depend on other values, and find a correlation between two variables, we may use it to impute missing items.

The table-6.1, show the attributes Temperature, Headache, Nausea and Vomiting and with the decision Malaria. However, many real-life data sets are incomplete. The missing attribute value is denoted by “?”.

<table>
<thead>
<tr>
<th>Case#</th>
<th>Temperature</th>
<th>Headache</th>
<th>Nausea</th>
<th>Vomiting</th>
<th>Malaria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mild</td>
<td>Mild</td>
<td>Moderate</td>
<td>Mild</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Moderate</td>
<td>Mild</td>
<td>Moderate</td>
<td>?</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Severe</td>
<td>Moderate</td>
<td>Mild</td>
<td>Mild</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Severe</td>
<td>Mild</td>
<td>?</td>
<td>Severe</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Moderate</td>
<td>Mild</td>
<td>Mild</td>
<td>Moderate</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Mild</td>
<td>Moderate</td>
<td>?</td>
<td>Mild</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Moderate</td>
<td>?</td>
<td>Severe</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>Mild</td>
<td>Mild</td>
<td>Mild</td>
<td>Moderate</td>
<td>No</td>
</tr>
</tbody>
</table>

Table -6.1. Data set with missing attributes

6.4 Results and Discussion

6.4.1 Example 1

To implement the concept of Rough set rule based technique to simplifying the diagnosis of malaria and impute the missing attributes have considered 8 data sets with four attributes out of 20 real data sets with ten attributes collected from different doctors. 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.

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

Temperature  = \{\{1,6,8\}, \{2,5,7\}, \{3,4\}\}
Headache     = \{\{1,2,4,5,8\},\{3,6,7\}\}
Nausea       = \{\{1,2\},\{3,5,8\},\{4,6\},\{7\}\}
Vomiting     = \{\{1,3,6\},\{2\},\{4\},\{5,7,8\}\}
Malaria      = \{\{1,5,6,8\},\{2,3,4,7\}\}

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-6.1.

After replacing proper values to the missing data, get the different classes from the above table as follows.

Temperature  = \{\{1,6,8\}, \{2,5,7\}, \{3,4\}\}
Headache     = \{\{1,2,4,5,8\},\{3,6,7\}\}
Nausea       = \{\{1,2\},\{3,5,6,8\},\{4,7\}\}
Vomiting $= \{\{1,3,6\},\{4\},\{2,5,7,8\}\}$
Malaria $= \{\{1,5,6,8\},\{2,3,4,7\}\}$

6.4.2 Example 2

(Covering Lower and upper approximations)

The covering lower approximation is defined as
$$X^* = U \{K_B(x)/x \in C \text{ and } K_B(x) \subseteq x\} \quad \text{and} \quad \text{The covering upper approximation is defined as}$$
$$X^\prime = U \{\text{neighbor}(x)/x \in X\}$$

From table-6.1 consider $U = \{1,2,3,4,5,6,7,8\}$ and $X = \{2,4,6,7\}$ then $X = \text{neighbor}(2) = \{\text{moderate, mild, moderate}\}$,
$\text{Neighbor}(4) = \{\text{severe, mild, severe}\}$,
$\text{Neighbor}(6) = \{\text{mild, moderate, mild}\}$ and
$\text{Neighbor}(7) = \{\text{moderate, severe, moderate}\}$ then $X^* = \{\text{moderate, severe, mild, moderate}\}$

6.4.3 Example 3

The proposed approach is based on Soft Set to impute the missing attributes from the table 6.1 analyze the relations between the attributes and define the notion of association degree to measure the relations. The priority gives relations between the attributes due to its higher reliability. When the mapping set of a attribute includes incomplete data, firstly look for another attribute which has the stronger association with the parameter.

$$(F,c_2) = \{\{\text{Temperature} = 2\}, \{\text{Headache}=1\}, \{\text{Nausea}=2\}, \{\text{Vomiting}=0\}\}$$
$$(F,c_4) = \{\{\text{Temperature} = 3\}, \{\text{Headache}=1\}, \{\text{Nausea}=0\}, \{\text{Vomiting}=3\}\}$$
$$(F,c_6) = \{\{\text{Temperature} = 1\}, \{\text{Headache}=2\}, \{\text{Nausea}=0\}, \{\text{Vomiting}=1\}\}$$
$$(F,c_7) = \{\{\text{Temperature} = 2\}, \{\text{Headache}=0\}, \{\text{Nausea}=3\}, \{\text{Vomiting}=2\}\}$$

Where mild $= 1$ moderate $= 2$, Severe$=3$.

Then computed elementary sets

$$(F,c_2) = \{\{\text{Temperature} = 2\} \cap \{\text{Headache}=1\} \cap \{\text{Nausea}=2\} \cap \{\text{Vomiting}=0\}\} = \{2\}$$
$$(F,c_4) = \{\{\text{Temperature} = 3\} \cap \{\text{Headache}=1\} \cap \{\text{Nausea}=0\} \cap \{\text{Vomiting}=3\}\} = \{3\}$$
$$(F,c_6) = \{\{\text{Temperature} = 1\} \cap \{\text{Headache}=2\} \cap \{\text{Nausea}=0\} \cap \{\text{Vomiting}=1\}\} = \{1\}$$
$$(F,c_7) = \{\{\text{Temperature} = 2\} \cap \{\text{Headache}=0\} \cap \{\text{Nausea}=3\} \cap \{\text{Vomiting}=2\}\} = \{2\}$$

After computation the result set become $\{2\}, \{3\}, \{1\}$ and $\{2\}$

<table>
<thead>
<tr>
<th>Case#</th>
<th>Common Attribute</th>
<th>Actual Attribute</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>C02</td>
<td>Moderate</td>
<td>Mild</td>
<td>Yes</td>
</tr>
<tr>
<td>C04</td>
<td>Severe</td>
<td>Moderate</td>
<td>Yes</td>
</tr>
<tr>
<td>C06</td>
<td>Mild</td>
<td>Mild</td>
<td>No</td>
</tr>
<tr>
<td>C07</td>
<td>Moderate</td>
<td>Mild</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 6.2. The Missing Data Filled with Observed Data
6.5 Conclusion

This chapter summarized the basic concepts of rough sets, covering based sets and soft sets the manner in which rough sets are related to covering based sets and soft sets, and then presented a detailed theoretical study of soft sets, which led to the definition of missing data handling. The missing data properties of rough sets, covering based rough sets and soft sent in a group structure. If the mapping set of an attribute includes incomplete data, the data filled 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 sets and soft sets. The methods can be used to handle various applications involved incomplete data sets.