INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND MATHEMATICS
Vol. 6Issue 8, December 2017, (Special Issue)
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Approach and Impute the Missing Data using Rough Set Rule Induction Method

B. S. Panda*
Ashok Misra**
S. S. Gantayat***

Abstract

In real world missing data are a common in every field and can have a significant effect on the results that can be drawn from the data. Many logics and techniques for handling missing data [1]-[4] have been proposed in the previous literature. Most of these techniques are much more complex. This paper focus on imputation technique based on basic rough set [5]-[8] techniques. The impact of missing data on quantitative research can be difficult and serious, leading to biased estimates of parameters, loss of information, decreased statistical power and increased different standard errors. All researchers have faced the problem of missing quantitative data at some point in their work. Research informants may refuse or forget to answer a survey question, files are lost, or data are not recorded properly. In this literature, characteristic relations are introduced to analyze incompletely medical data result. It is observe that the fundamental rough set technique of lower and upper approximations for incompletely specified decision tables may be defined in a variety of other ways. The final results obtained using real data sets are given and they provide a meaningful and promising insight to the difficult of missing data.

In this paper the medical data set contains patients with the attributes Nausea, Headache, and Temperature with the decision result flu. The complete use of the original basic rough set model obtained to finding many more findings of reducing the input data. To clear this problem, a new fundamental approach basic rough set theory, rule generation and statistical logics and techniques are developed and implemented.

Keywords:

Rough set,
Rule induction,
Missing data,
Missing Attribute values,
Imputation.

Author correspondence:

B S Panda,
Assoc Professor,
Dept. of CSE, MITS, Rayagada, Odisha, India.

1. Introduction

In real data set some attribute values are frequently missing and incomplete. There are mainly two reasons for attribute values to be missing: either they are 'lost' (e.g., were removed) or they are 'do not care' conditions.

* B S Panda, Assoc Professor, Dept. of CSE, MITS, Rayagada, Odisha, India.
** Ashok Misra, Professor, Dept of Mathematics, CUTM, Parlakhemundi, Odisha, India.
*** S S Gantayat, Professor, Dept. Of CSE, GMRIT, Rajam, AP, India.

(i.e., the actual values were not considered at all since they were different, and the decision to which concept a case belongs was taken without that information).

Different interpretations of missing values than 'lost' and 'do not care' conditions were clear represented. Decision tables with all incomplete attribute values that are 'lost' were discussed, within basic rough set theory, in Pawlak 1982 [9], where two techniques for basic rule induction from such data were presented. On the this method, the first attempt to study 'do not care' conditions using basic rough set theory , where a method for basic rule induction was developed in which missing attribute values were imputed by all possible values from the domain of the attribute. 'Do not care' conditions were also discussed later, where the indiscernibility-relation was again generalized, this time to analyze incomplete decision tables with 'do not care' conditions. The main technique, attribute value pair blocks. These attribute blocks are used to build characteristic sets, characteristic-relations, and rough set lower and upper approximations for decision tables with missing attribute values. We are estimating that the same decision table may contain both types of missing attribute values: 'lost' and 'do not care' conditions. An attribute-characteristic relation and set is a simple and generalization of the indiscernibility relation [10],[11].

2. Database Evaluation

This literature assume that input data for data mining are presented in a form of a decision table (or relation set) in which cases (or rows) are described by attributes (particular variables) and a decision (dependent variable). A sample medical data example of such a table is presented in Table-1, with the attributes Headache, Temperature, and Nausea and with the decision result Flu. Actually, many real-life data sets are incomplete and missing, i.e., some attribute values are not available. In Table-1 in complete attribute values are denoted by "?"s.

Case	***Attributes***			***Decision***
	Temperature	***Headache***	***Nausea***	***flu***
1	High	?	No	Yes
2	Very_high	yes	Yes	Yes
3	?	no	No	No
4	High	Yes	Yes	Yes
5	High	?	Yes	No
6	Normal	Yes	No	No
7	Normal	No	Yes	No
8	?	yes	?	yes

Table-1. Data with values Missing Attributes

3. Sequential Methods

In this method to manage missing attribute values original in-complete data sets, with in complete attribute values, are converted into complete data sets and then the main technique, e.g., attribute rule induction is conducted.

In order to clear the problem, it is observed that to logically, technically and combinationally analyze and observe patterns in sequences by checking the impertinence of local patterns that consistently result in decision flu. For such an analysis, rule generation in rough set provides a data mining techniques based on the notions of attribute values reduction and reduced decision techniques. The main advantages of rough set data mining is that it can develop reduced and consistent decision rules and techniques by logically checking all type combinations of condition and decision attributes in a real-data system. The basic rough set theory can be used to develop essential attributes towards value attribute reduction of logical combinations. The pattern of sequential mining algorithms have not been well studied in the context of basic rough set theory. Extending this technique to sequential pattern mining entails a logical and technical analysis of local patterns in minimal computing; this is different from the frequency analysis of sequential patterns.

3.1 Case Deleting with Missing Attribute Values

This technique is based on removing cases with missing attribute values. It is also called case wise deletion (complete-case analysis or list wise-deletion) in statistics. All cases with missing attribute values are deleted from the data set.

Case	***Attributes***			***Decision***
	Temperature	***Headache***	***Nausea***	***flu***
1	Very_high	yes	Yes	Yes
2	High	Yes	Yes	Yes
3	Normal	Yes	No	No
4	Normal	No	Yes	No

Table 4. Cases Deleting with Missing Attributes

3.2 MCVA (The Most Common Value of an Attribute)

In this technique, one of the easiest methods to manage missing attribute values, such values are replaced by the most predicted value of the attribute. The different types, a missing attribute value are imputed by the most known probable attribute value, where such probabilities are taken by relative frequencies of corresponding value attributes.

Case	***Attributes***			***Decision***
	Headache	***Temperature***	***Nausea***	***flu***
1	Yes	High	No	Yes
2	Yes	Very_high	Yes	Yes
3	No	High	No	No
4	Yes	High	Yes	Yes
5	Yes	High	Yes	No
6	Yes	Normal	No	No
7	No	Normal	Yes	No
8	Yes	High	Yes	yes

Table-3. Most Common Values

3.3. The Most Common Value of an Attribute Restricted to a Concept

Case	*Attributes*			*Decision*
	Temperature	*Headache*	*Nausea*	*flu*
1	High	Yes	No	Yes
2	Very_high	Yes	Yes	Yes
3	Normal	No	No	No
4	High	Yes	Yes	Yes
5	High	No	Yes	No
6	Normal	Yes	No	No
7	Normal	No	Yes	No
8	High	Yes	Yes	Yes

Table 4 Most Common Values

For example, in Table-1, case 1 belongs to the concept (1, 2, 4, 8) all known values of Headache, restricted to (1, 2, 4, 8), are yes, so the missing attribute value is replaced by yes. On the other hand, in Table-1, case 3 related to the concept (3, 5, 6, and 7) and the value of Temperature is missing. The known values of Temperature, restricted to (3, 5, 6, and 7) are high (one time) and normal (two times), so the missing attribute value is replaced by normal.

3.4 Assigning All Possible Attribute Values to a Missing Attribute Value

This method the different case with missing attribute values is exchanged by the set of cases in which every missing attribute value is exchanged by all known possible values. This method is very simple easy to impute the small data sets.

Case	*Attributes*			*Decision*
	Headache	*Temperature*	*Nausea*	*flu*
1	yes	High	No	Yes
2	no	high	no	yes
3	yes	Very_high	Yes	Yes
4	no	high	No	No
5	no	very_high	No	No
6	no	normal	No	No
7	Yes	High	Yes	Yes
8	yes	High	yes	no
9	no	High	Yes	No
10	Yes	Normal	No	No
11	No	Normal	Yes	No
12	Yes	High	Yes	Yes
13	yes	high	no	yes
14	yes	very_high	yes	yes
15	yes	very_high	no	yes
16	yes	normal	yes	yes
17	yes	normal	no	yes

Table 5. all possible values

3.5 Assigning all Possible Attribute Values Restricted To a Concept

In this, every case with missing attribute values is replaced by the set of cases in which every attribute 'a' with the missing attribute value has its every possible known value restricted to the concept to which the case belongs.

Case	*Attributes*			*Decision*
	Headache	*Temperature*	*Nausea*	*flu*
1	yes	High	No	Yes
2	yes	Very_high	Yes	Yes
3	no	normal	No	No
4	no	high	No	No
5	Yes	High	Yes	Yes
6	yes	High	yes	no
7	no	High	Yes	No
8	Yes	Normal	No	No
9	No	Normal	Yes	No
10	Yes	High	Yes	Yes
11	yes	high	no	yes
12	yes	very_high	yes	yes
13	yes	very_high	no	yes

Table 6. All Possible Values

3.6 Replacing Missing Attribute Values by the Attribute Mean

Case	*Attributes*			*Decision*
	Temperature	*Headache*	*Nausea*	*flu*
1	100.2	?	No	Yes
2	102.6	yes	Yes	Yes
3	?	no	No	No
4	99.6	Yes	Yes	Yes
5	99.8	?	Yes	No
6	96.4	Yes	No	No
7	96.6	No	Yes	No
8	?	yes	?	yes

Table 7. Replacing Missing Attribute by Mean

In this technique, every missing attribute value for a numerical attribute is exchanged by the arithmetic mean of different known attribute values. Data set in which missing attribute values are replaced by the attribute mean and the most common value.

Case	*Attributes*			*Decision*
	Temperature	*Headache*	*Nausea*	*flu*
1	100.2	yes	No	Yes
2	102.6	yes	Yes	Yes
3	99.2	no	No	No
4	99.6	Yes	Yes	Yes
5	99.8	yes	Yes	No
6	96.4	Yes	No	No
7	96.6	No	Yes	No
8	99.2	yes	yes	yes

Table-8. Replacing Missing attribute by Mean

4. PM (Parallel Methods)

In this type of method we will observe on manage missing attribute values in pm (parallel method) with basic rule induction. Here two types of missing attribute values: 'lost' and 'do not care' conditions. Here we will introduce some useful techniques, such as bunch of attribute-value pairs, characteristic value sets, and basic characteristic value relations, rough set lower and upper approximations [11]-[13]. After we here with explain that how to introduce rules using the same characteristic blocks of attribute-value pairs that was used to compute lower and upper approximations. Input data sets are not prepared the same way as in sequential methods; technically, the rule learning algorithm is revised to learn rules and logics directly from the original and incomplete data sets.

4.1 Blocks of Attribute-Value Pairs

The decision table defines a logic that links the direct product of the set U (universe) of all medical cases and the set A of all attribute values into the set of all possible values. In this method it will observe that all un-known and missing attribute values are denoted either by '?' or by '*', 'lost' values will be denoted by '?', 'do not care' conditions will be denoted by '*'. Thus, it will assume that all possible missing attribute values from Table-4.9 are lost. On the other hand, all attribute values from Table-4.10 are do not care conditions.

Case	*Attributes*			*Decision*
	Temperature	*Headache*	*Nausea*	*flu*
1	High	?	No	Yes
2	Very_high	yes	Yes	Yes
3	?	no	No	No
4	High	Yes	Yes	Yes
5	High	?	Yes	No
6	Normal	Yes	No	No
7	Normal	No	Yes	No
8	?	yes	?	yes

Table-9. Missing Attributes

Case	*Attributes*			*Decision*
	Temperature	*Headache*	*Nausea*	*flu*
1	High	*	No	Yes
2	Very_high	yes	Yes	Yes
3	*	no	No	No
4	High	Yes	Yes	Yes
5	High	*	Yes	No
6	Normal	Yes	No	No
7	Normal	No	Yes	No
8	*	yes	*	yes

Table-10. Do-not care attributes

Any decision table defines a function P that maps the direct product of the set U of all cases & the set A of all attributes into the set of all values. Let (a,v) attribute-value pair complete decision table [a,v)] is the set of all X for which P (x,a)=v.
For missing decision table P(x,a)=?
For don't care P(x,a)= *

For Table – 9	**For Table-10**
[(Temp,high)]= {1,4,5}	[(Temp,high)]= {1,3,4,5,8}
[(Temp,very_high)]={2}	[(Temp,very_high)]={2,3,8}
[(Temp,Normal)]={6,7}	[(Temp,Normal)]={3,6,7,8}
[(headache,yes)]={2,4,6,8}	[(headache,yes)]={1,2,4,5,6,8}
[(headache,no)]={3,7}	[(headache,no)]={1,3,5,7}
[(Nausea,No)]={1,3,6}	[(Nausea,No)]={1,3,6,8}
[(Nausea,yes)]={2,4,5,7}	[(Nausea,yes)]={2,4,5,7,8}

4.2 Characteristic Sets
The characteristic set $K_B(x)$ is the intersection of blocks of attribute-value pairs (a,v) for all attributes a from B for which p(x, a)is known and p(x, a)=v. For Table-9 and B=A.

$K_A(1)$ ={1,4,5}∩{1,3,6}={1}
$K_A(2)$ ={2}∩{2,4,6,8}∩{2,4,5,7}={2}
$K_A(3)$ ={3,7}∩{1,3,6}={3}
$K_A(4)$ ={1,4,5}∩{,2,4,6,8}∩{2,4,5,7}={4}
$K_A(5)$ ={1,4,5}∩{2,4,5,7}={4,5}
$K_A(6)$ ={6,7}∩{,2,4,6,8}∩{1,3,6}={6}
$K_A(7)$ ={6,7}∩{3,7}∩{2,4,5,7}={7}
$K_A(8)$ ={2,4,6,8}

and for Table 10 and B=A

$K_A(1)$ ={1,3,4,5,8}∩{1,3,6,8}={1,3,8}
$K_A(2)$ ={2,3,8}∩{1,2,4,5,6,8}∩{2,4,5,7,8}={2,8}
$K_A(3)$ ={1,3,5,7}∩{1,3,6,8}={1,3}
$K_A(4)$ ={1,3,4,5,8}∩{1,2,4,5,6,8}∩{2,4,5,7,8}={4,5,8}
$K_A(5)$ ={1,3,4,5,8}∩{2,4,5,7,8}={4,5,8}
$K_A(6)$ ={3,6,7,8}∩{3,7}∩{1,2,4,5,6,8}∩{1,3,6,8}={6,8}

$K_A(7) = \{3,6,7,8\} \cap \{1,3,5,7\} \cap \{2,4,5,7,8\} = \{7\}$
$K_A(8) = \{1,2,4,5,6,8\}$

Incomplete decision tables in which all missing attribute values are "do not care" conditions, from the view point of rough set theory, were observed for the first time in , where a method for rule induction was introduced in which each missing attribute value was imputed by all values from the domain of the attribute. Technically such values were replaced by all values from the entire domain of the attribute values, later, by attributes restricted to the same logic to which a case with a missing attributes belongs. Like such incomplete decision tables, with all possible missing attribute values being 'do not care conditions'.

4.3 Rough Set Lower and Upper Approximations
In all finite union of characteristic data sets of B is called a B-definable data set. The rough set lower approximation of the concept X is the largest possible data sets that are contained in X and the upper approximation of X is the minimal possible data set that contains X.

$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq x\}$

B-Upper approximation

$\overline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \theta\} = \cup\{K_B(x) \mid x \in X\}$

From table – 9: The rough set lower and upper approximations are.

$\underline{A}\{1,2,4,8\} = \{1,2,4\}$
$\underline{A}\{3,5,6,7\} = \{3,6,7\}$
$\bar{A}\{1,2,4,8\} = \{1,2,4,6,8\}$
$\bar{A}\{3,5,6,7\} = 3,4,5,6,7\}$
From decision table -10: The rough set lower and upper approximations are.

$\underline{A}\{1,2,4,8\} = \{2,8\}$
$\underline{A}\{3,5,6,7\} = \{7\}$
$\bar{A}\{1,2,4,8\} = \{1,2,3,4,5,6,8\}$
$\bar{A}\{3,5,6,7\} = \{1,2,3,4,5,6,7,8\}$

4.4.4 Rule Induction - MLEM2
Each incomplete attribute value was exchanged by all values from the domain of the attribute values. The MLEM2 rule induction method is a revised version of the algorithm LEM2. The techniques and rules induced [14], [15] from the rough set lower approximation of the concept definitely describe the concept, so they are called certain and concrete. On the other way, rules and logics induced from the rough set upper approximation of the technique describe the concept only possibly (or plausibly), so they become called possible or particular. MLEM2 may introduce both the particular and possible rules and logics from a decision table with some in complete attribute values being 'lost' and some missing attributes being 'do not care' conditions, while some different attributes may be numerical.

For this rule induction of decision tables with numerical attributes. MLEM2 manages in complete attribute [16] values by computing (in a way than different in LEM2) character blocks of attribute-value pairs, and then value characteristic sets and rough set lower and upper approximations. All these techniques are revised according to the both previous sub sections; the algorithm itself remains the same and no change.

Certain rules from Table -9

(Number of attribute value) (Number of example) (Training cases)
2 1 1 (Temp,High) & (Nausea, No) →(Flu:Yes)
2 2 2 (Headache,Yes) &(Nausea,Yes) →(Flu:Yes)
1 2 2 (Temp, Normal) →(Flu:No)
1 2 2 (Headache, No) →(Flu:No)

Possible Rule Set Table-9

1 3 4 (Headache,Yes) →(Flu:Yes)
2 1 1 (Temp, High)&(Nausea,No) → (Flu:Yes)
2 1 2 (Temp, High) & (Nausea, Yes) → (Flu:No)
1 2 2 (Temp, Normal) → (Flu:No)
1 2 2 (Headache, No) → (Flu:No)

Certain rules from Table -10

2 2 2 (Temp, Very_High)& (Nausea, Yes) → (Flu:Yes)
3 1 1 (Temp, Normal) & (Headache, No) & (Nausea, Yes) → (Flu:No)

Possible Rule set from Table-10

1 4 6 (Headache, Yes) → (Flu:Yes)
2 1 1 (Temp, Very_High) →(Flu:Yes)
1 2 5 (Temp, High) →(Flu:No)
1 3 4 (Temp,Normal) → (Flu:No)

4.4 Conclusion

There exist many techniques to manage missing data. In particular, we are interested in preprocessing methods, which can be used before any further analysis. Some of them are very easy to apply, but have too many drawbacks, which limit their use to uninteresting situations. This is the case for list wise deletion, and for naive imputation techniques, such as deterministic mode or mean imputation.

The technique of an attribute-value pair block, the main logic for rough set rule induction used in this chapter, is both very simple and useful technique. It is particularly useful for in complete decision tables, from it is used to determine value characteristic sets, value characteristic relations, rough set lower and upper approximations, and, finally, it is used in rule induction method.

References:

[1] Allison, P.D. (2001) "Missing Data. Sage Publications", Thousand Oaks.

[2] B. S. Panda, S. S. Gantayat, Ashok Misra (2017) "A Comparative Study of Handling Missing Data in Student Data Analysis using Rough Set and Soft Set" In: International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 8, ISSN : 2229-3345.

[3] B. S. Panda, S. S. Gantayat, Ashok Misra (2016) "Retrieving the Missing Information from Information Systems Using Rough Set, Covering Based Rough Set and Soft Set" In: International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7 (3), 1403-1407-2016.

[4] B. S. Panda, S. S. Gantayat, Ashok Misra, (2014) "Rough Set Rule Based Technique for the Retrieval of Missing Data in Malaria Diseases Diagnosis" In: Springer FMB Series, ISBN: 978-981-287-260-9.

[5] S. S. Gantayat, Ashok Misra, B. S. Panda (2013) "A-Study-of-Incomplete Data – A Review" In: FICTA, LNCS Springer. Pp 401-408. ISBN: 978-3-319-02930-6.

[6] B. S. Panda, S. S. Gantayat, Ashok Misra (2013) "Rough-Set Approach to Development of a Knowledge-Based Expert System" In: International Journal of Advanced Research in Science and Technology (IJARST), Vol. 2, Issue 2, pp. 74-78. ISSN: 2319-1783.

[7] Greco, Matarazzo and Slowinski, (1999) "The Use of Rough Sets and Fuzzy Sets in MCDM" in Multicriteria Decision Making Volume 21 of the series International Series in Operations Research & Management Science pp 397-455.

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

[9] Pawlak, Z. (1982): Rough Sets. J. Inf. & Comp. Sc. II, 341- 356.

[10] Grzymala-Busse, J.W.: (1991) "On the unknown attribute values in learning from examples". Proc. Of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina, October 16–19, 1991. Lecture Notes in Artificial Intelligence, vol. 542, Springer-Verlag, Berlin, Heidelberg, New York, pp. 368–377.

[11] Kerdprasop N., K., Y. Saiveaw and P. Pumrungreong, (2003)A comparative study of techniques to handle missing values in the classification task of data mining, 29th Congress on Science and Technology of Thailand, Khon Kaen University, Thailand.

[12] L. A. Zadeh. (2005). Toward a generalized theory of uncertainty (GTU) – an outline, Information Sciences. 172, 1-40.

[13] Saar-Tsechansky Maytal and Foster Provost, (2007) Handling Missing Values when Applying Classification Models, Journal of Machine Learning Research Vol 8, pp. 1625-1657.

[14] Stefanowski, J., and Tsouki`as, A.: (1999) On the extension of rough sets under incomplete information. In Zhong, N., Skowron, A., Ohsuga, S., eds.: Proceedings of the RSFDGrC '99. LNCS 1711, Springer, pp 73–81.

[15] Y. Y. Yao, (1998) A comparative study of fuzzy sets and rough sets, Information Sciences, v. 109, Issue 1-4, pp. 227 – 242.

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