IMPUTATION OF MISSING DATA BASED ON ROUGH SET

Pallab Kumar Dey
Deptment of Computer Science, Kalna College, Kalna
pallabkumardey@gmail.com

Abstract
For Classification or other data mining task, Data imputations have great importance. Rough set is more robust method to deal with imprecision and uncertainty. Available techniques have been compared and extended model of Rough set based (ERSBA) algorithm has been proposed for missing value imputation. So using (ERSBA) algorithm complete data set may be generated which has a great importance for data mining. Efficiency and effectiveness of the proposed algorithm has been shown.
Keywords: *Missing-values, Data-mining, Tolerance-relation, Extended-valued-tolerance-relation, Rough-Set.*

1. Introduction

In this E-technological era large amount of data can be collected in every moment. These huge amounts of ideal data are required to enhance the quality of discovering knowledge. Though, ideal data is merely available. The data which are collected may be called noisy ideal data. So removal of noisy data is required to get better prediction[1-13]. Maximum data mining effort is involved with the preprocessing of data i.e. to remove noise from noisy ideal data. Missing values are also present due to different reasons. Data analysis may be erroneous due to missing values. Missing values handling is an important issue for data mining. Incompleteness of data may occur due to several reasons like data unavailability or not possible to collect data due to time constraints or cost efficiency. As maximum existing data mining algorithms are based on complete data so imputation of missing data is the best solution to use existing data mining algorithms effectively [11-13]. In this paper Rough set approach has been used to handle missing values for incomplete information as pre-processing tool. To handle uncertainty and impreciseness Rough set is the most important tool as no additional or prior information of data is required.

Many techniques are available for handling problems of incompleteness. But after looking into the matter deeply, it is clear that basic approaches are two types. First one is like ROUSTIDA[9] and RSDIDA [2] where missing values have to figure out by the suitable methods. Here classifier algorithm or data mining techniques can be applied after replacing missing values. So here first filling out the incomplete values then it is possible to apply any classifier (Fig.1). The second one is like LEM1 and LEM2 [8] where modified classifier algorithm can be applied directly for incomplete information system. But here it is not possible to use already available data mining algorithms which are based on perfect data. First one i.e. filling approach is better as existing data mining algorithm can be used.

Except these two approaches it is possible to classify another method called decomposition approach as in Fig.2. Decomposition approach is based on the decomposition of the incomplete information system (IIS) into some subset and after that applying the template evaluation function (**TEV**) and classifier; the rule is directly obtained [7]

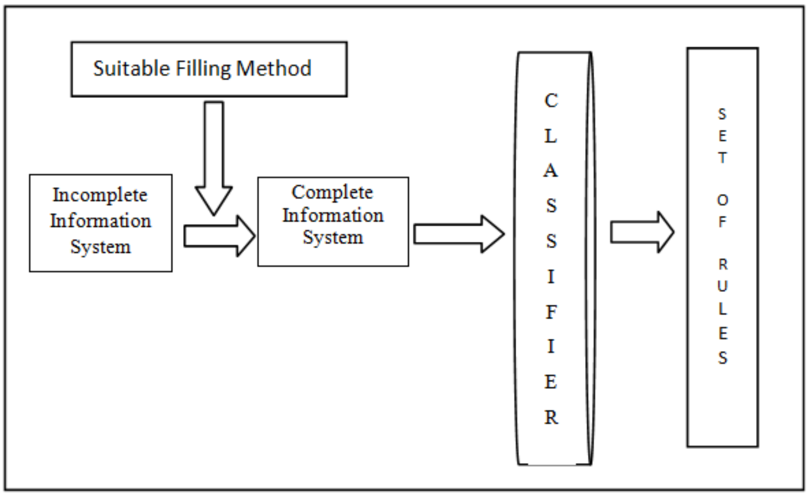

Fig. I: Rule Obtained from Incomplete Information System by Filling Method Approach.

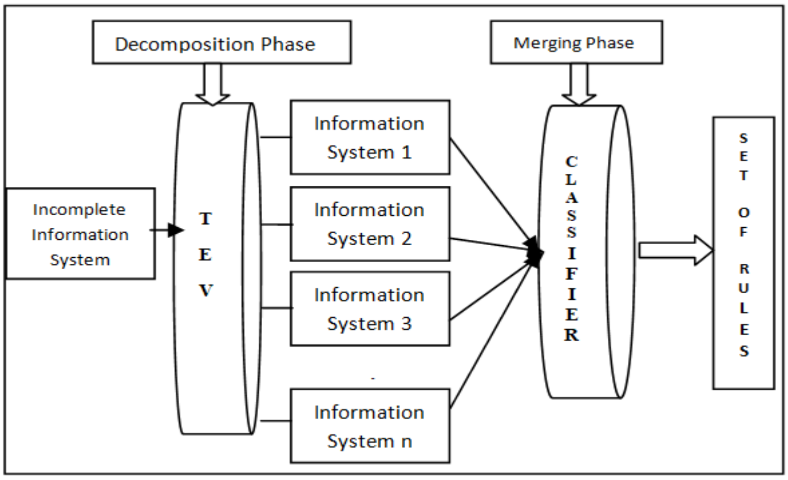

Fig. 2: Rule Obtained from Incomplete Information System by Decomposition Approach.

This method actually comprises of two main component the TEV and corresponding classifier. This may be considered as the special case of modified classifier. This paper will compare such methods for evaluation system. Presently we are available with few filling methods which may be classified into followings.

II. Filling Methods:

***A.** Reduction Approach*

By reduction approach the objects with missing values are deleted and the data mining is carried out with the remaining complete data. It is very clear that deliberate deletion causes loss of information and hence best knowledge discovering hope is diminishes. In this paper it has been shown that this method can be applied to few cases without losing the efficiency of data mining out come. Though this method becomes dangerous for small amount of data, and then it causes serious affect on the discovered rules.

B. Extension Approach

By extension approach the missing values are replaced by all possible values [4] and then any data mining technique can be used. Though the database gets larger in volume and hence the computation cost becomes high. Also the discoverable knowledge is incorporated with *'out of context information'*. Another drawback of this method is that it can't be applied into numerical missing values.

$$VTR(x, y) = \begin{cases} 0, & \text{if } T \neq (x,y) \\ \prod_{a\epsilon\,(D\cup C)} P(x,y), & \text{if } T = (x,y) \end{cases}$$

C. *Statistical Approach*

In *statistical approach* missing values are substitute with the help of some statistical methods. These methods are applied on the basis of the trend of present observed values [10-13]. For example, mean of a particular attribute values can be used to replace the missing values. Main disadvantage of these methods is the statistical hypothesis.

D. *Concept Similarity Approach using Rough Set*

It's a very new one. After introduction of rough set the idea of indiscernibility relation in information system has evolved. The same concepts were applied for incomplete information system, named as similarity relation. Degree of similarity is represented as valued tolerance relation [3]. In [2,6] new idea of extended valued tolerance relation has been proposed. This is much more efficient method.

Table: 1 Comparison Of Four Filling Methods

	Reductio	Extensio	Statistica probabili distributi	Concept Similarity Approach using Rough Set
KDD Efficienc	May Change	May Change	Constant	May Change
Roughne	*	*	Constant	Increase
Database	Reduced	Expand	No	No
Class	Constant	Expand	Expand	Constant
Computi Complex	Easy	Tuff	Easy	Easy

*Cannot applicable

Comparison of these methods has been shown in table 1. It is clear from table 1 that concept similarity approach is superior among all other methods. Now it has been discussed about the different methods available for the similarity approach and compares them.

III. Definitions

A. Incomplete Information System

Incomplete information system I = (U, C, D, V, *,f)
Where,
U denotes the Universe of discourse,
C = Set of all conditional attribute,
D = Set of all decision attribute,
V = Set of all values,
* = Missing value,
f is mapping as,
U X (C,D) ← V
UxC← *

B. Tolerance Relation [8]

Tolerance relation can be defined as follows,

T = {(x, y) € U X U | ∀ a € (D U C)(a(x) =a(y) or a(x) = * or a(y) = *) }

For incomplete information, tolerance relations describe similarity between two objects. Tolerance relation does not provide similarity comparison i.e. which object is more similar to a object.

B. *Valued Tolerance Relation [6]*

Valued tolerance relation(VTR) is the measures degree of equivalence between two objects in an incomplete information system. Valued tolerance relation give similarity degree by which we can predict which object is more similar to other object. It can be defined as above,
Where, T =(x, y) denotes that there is a tolerance relation between x and y, [x,y € U].
P(x.y) may be defined as,

$$P(x, y) = \begin{cases} 1, & \text{if} \forall\, a, a(x) \neq * \ and\ a(y) \neq * \\ 1/|V|, & \text{if} \forall\, a, a(x) \neq * \ and\ a(y) = * \end{cases}$$

The point to be noted over here is, consider P(x, y) only when there is a tolerance relation between x and y i.e for all a, a(x) and a(y) are either same or anyone of them is missing(*) or both are missing.

D. *Extended Valued Tolerance Relation[2]*

Extended Valued tolerance relation(EVTR) is the measures degree of equivalence between two objects in an incomplete informationsystem. Here for filling missing values similarity of object is considered with filling ability.
Missing attribute set MAS is the collection of all missing attribute for an object. MAS of any object whose attribute values are missing, can be defined as,

MAS(x) = {k | Ck, € C(x), k = 1,2,3 ... |(C U D)|}

Extended Valued tolerance relation can be defined as,

$$EVTR(x, y) = \begin{cases} 0 & \text{if } MAS(x) \subseteq MAS(y) \\ \prod_{a\in (D\cup C)} P(x,y), & \text{else} \end{cases}$$

Where, [x,y € U].
P(x,y) may be defined as,

$$P(x,y) = \begin{cases} 1 & a(x) \neq * \wedge a(y) \neq * \wedge a(x) = a(y) \\ \frac{1}{|V|} & (a(x) = * \wedge a(y) \neq *) \vee (a(x) \neq * \wedge a(y) = *) \\ \frac{1}{|V|^2} & a(x) = * \wedge a(y) = * \\ 0 & a(x) \neq * \wedge a(y) \neq * \wedge a(x) \neq a(y) \end{cases}$$

The point to be noted here is, P(x,y) has been consider only when no. of missing value attribute of x is less than that of y i.e. x object is much more known than y object.

IV. Comparison Between Best Two Similarity Methods

It is seen that concept similarity approach using rough set is the most prominent method for that time being. Basically two better algorithms ROUSTIDA and RSDIDA are better for fulfil our need.

A. *ROUSTIDA[9]*

In this algorithm, a very simple idea of tolerance relation is used without going deep into the problem of similarity degree. Objects, having tolerance relation with eachother, can replace one another missing attribute values. Conflict arises when we can find two objects are in tolerance relation with a third object. E.g. In Table: 2. tolerance relations T = (xl,x2), T = (xl,x3), T=(x2,x3). It shows, ultimately all xl, x2, x3 will leads to a same object and it just enhance the support of the object at later stage of data mining. Is the degree of tolerance in each pair same? Apart from the complexity, this unsolved question is also a limitation of ROUSTIDA.

Table:2 Incomplete Information System

U	a1	a2	a3	a4	d
xl	*	2	3	1	y
x2	1	*	*	*	y
x3	1	2	*	*	y

B. *RSDIDA[2]*

This method can solve the limitations of the previous one. From this system (Table: 2) RSDIDA will compute Extended Valued Tolerance matrix as,

EVTR matrix	x1	x2	x3
x1	0	1/81	1/27
x2	1/81	0	1/243
x3	1/27	1/243	0

The unsolved query of the previous section can be solved with the help of Extended valued tolerance relation. Now V = {1,2,3}

EVTR(xl,x2) = 1/34= 1/81
EVTR(x2,x3) = 1/35= 1/243
EVTR(xl,x3) = 1/33= *1/27*

It is now clear object x1 and object x3 are having greatest value of tolerance relation. Perhaps we can say object x1 and x3 are much more similar among three objects. But, this method applies divide and conquer ideology i.e. IISis first decomposed into some subset. The decomposition is made with respect to the decision attribute values. Then the Extended valued tolerance relation matrix is prepared. This matrix is used for filling up each of the decomposed IIS. As a result we lost the conflict set(object set with similar conditional attribute values but different in decision attribute values) which may be required for decision making purpose. According to RSDIDA decision table(Table: 3) will be divided into two IIS based on the decision attribute values. One IIS is contained with object x1 and x3, another with object x2 and x4. The missing value of attribute a4 for the object x2 i.e. a4(x2) will be replaced by a4(x4) and it is 2. The point to be noted over here is EVTR(x2,x4) = 1/27. Whereas, in case of no decomposition a4(x2) value will be replaced by a4(xl) and it is 1. Moreover, EVTR(x2,xl) = 1/9.

Table: 3 Incomplete Information System

U	a1	a2	a3	a4	d
x1	*	2	3	1	y
x2	1	2	3	*	n
x3	*	1	3	*	y
x4	1	*	*	2	n

The entire comparison can be presented in a tabular format in Table: 4. The point to be mentioned over here is that the conflict item set may be lost in ROUSTIDA and it must be lost in case of RSDIDA, but in case of our proposed algorithm it must be preserved.

Table: 4	ROUSTIDA	RSDIDA
Filling Ratio	Usual	Better
Complexity	Tuff	Easy
Reliability	Usual	Good
Conflict Set	May Lost	Lost

Comparison between two popular similarity approaches

Keeping in mind the above problems, here is a proposed algorithm(ERSBA)which may be used as computation algorithm.

V. Proposed Algorithm

Algorithm: ERSBA

Input: An Incomplete Information System, IIS0 = (U, C, D,V, *f);
Output: Complete Information System, IS0= (U, C, D, V, f);
Method: Main(IIS0)

```
n ← |C|
m ← |U|
T(m,m) ← 0
For i = 1 to m
   Do
         CEVTM(IIS0,i,T)
  End
IS0 = PFILL(IIS0,T)
```

The main algorithm ERSBA consist of two main subroutine CEVTM(IIS0,i,T) for computation of extended value toleration relation and another subroutine PFILL(IIS0,T) for filling object with suitable object value.

Algorithm: CEVTM

```
Input: An Incomplete Information System,
IIS0 = (U, C, D, V, *,f);

Output: Tolerance Value of object i,

Subroutine: CEVTM (IIS0, i, T)
For j = 1 to m
Do
  If((i==j) OR (MAS(i)⊆ MAS(j))
            T(i,i) =0;
  Else

    For k = 1 to m
      Do
        If(IIS0(i,k) == * AND IIS0(j,k) == *)
                T(i,j) = 1/|Vk|²
        elseIf(IIS0(i,k) == * OR IIS0(j,k) == *)
                T(i,j) = 1/|Vk|
        elseIf for all k (IIS0(i,k) == IIS0(j,k))
                T(i,j) = 1
      End
End
Algorithm: PFILL
```

```
Input:An Incomplete Information System,
IIS0 = (U, c, D, V, *,f);
Extended Valued Tolerance relation, T;
Output: Complete Information System,
          IS0 = (U, c, D, V, f);

Subroutine: PFILL(IIS0, T)
IS0 ← IIS0
For i = 1 to m
 Do
  For j = 1 to m
   Do
    If( Max(T(i,j)))
      For k = 1to n
        Do
          ISO(i,k) = ISO(j,k)
        End
   End
End
Return ISO
```

VI. Results And Discussion

Now to compare the efficiency of the proposed algorithm.Let us observe the result that is obtained from the same Incomplete Information table using two algorithms, RSDIDA and the proposed one.

A. Applying RSDIDA Method

Information in Table 3 will be processed as the following diagram shows. First, it will be divided into two subset, according to RSDIDA method. Then Extended Valued Tolerance matrix have been calculated. After that These incomplete subset have been change into complete subset by applying RSDIDA method. Then again these two subset have been merge to get the desire complete table.

[IIS0]

U	a1	a2	a3	a4	d
xl	*	2	3	1	y
x2	I	2	3	*	n
x3	*	1	3	*	y
x4	1	*	*	2	n

[IIS01]

u	a1	a2	a3	a4	d
x1	*	2	3	1	y
x3	*	1	3	*	y

[IIS02]

u	a1	a2	a3	a4	d
x2	1	2	3	*	n
x4	1	*	*	2	n

[TEV1]

u	x1	x2
x1	0	1/27
x3	1/27	0

[TEV2]

u	x1	x2
x2	0	1/27
x4	1/27	0

[S01]

u	a1	a2	a3	a4	d
x1	*	2	3	1	y
x3	*	1	3	1	y

[S02]

u	a1	a2	a3	a4	d
x2	1	2	3	2	n
x4	1	2	3	2	n

[S0]

U	a1	a2	a3	a4	d
xl	*	2	3	1	y
x2	I	2	3	2	n
x3	*	1	3	1	y
x4	1	2	3	2	n

B. Applying Proposed Method

[IIS0]

U	a1	a2	a3	a4	d
xl	*	2	3	1	y
x2	I	2	3	*	n
x3	*	1	3	*	y
x4	1	*	*	2	n

[TEV]

U	a1	a2	a3	a4
xl	0	1/9	0	0
x2	1/9	0	0	1/27
x3	0	0	0	1/81
x4	0	1/27	1/81	0

[S0]

U	a1	a2	a3	a4	d

xl	1	2	3	1	y
x2	1	2	3	1	n
x3	1	1	3	2	y
x4	1	2	3	2	n

These tables' data shows that proposed ERSBA algorithms imputation accuracy is better than other methods. Its reliability over other methods has been shown above. Error rate of proposed algorithm's imputation is lower than others methods. So it can be concluded that ERSBA algorithm perform better than other methods. ERSBA algorithms prediction is almost perfect considering all evaluation parameter. So for practical cases it may be used. So it can be adopted as a better method for missing value imputation.

VII. Conclusion

Rough set concept has been used for incomplete data set. Computations of tolerance relation, valued tolerance relation and extended valued tolerance relation have been shown. Extended valued tolerance has been used for imputation of missing data. For imputation in pre-processing approach it is always better to fill the missing values by available best object values. This concept has been used for missing data imputation with similar object, fetching from extended valued tolerance relation. So after application of ERSBA algorithm there is no chance to generate misleading information. Proper utilization of extended valued tolerance relations enhance the efficiency of filling missing data by considering most suitable object. ERSBA algorithm can be use as preprocessing tool for missing data imputation. This algorithm may be enhanced for applications of imputation with feature reduction methods to achieve more suitable data for data mining.

References

1. Zhang Qizhong. "An Approach to Rough Set Decomposition ofIncomplete Information Systems", *2007 Second IEEE Conferenceon Industrial Electronics and Applications.* pp 2455 – 2460.
2. Zaimei Zhang; Renfa Li; Zhongsheng Li: Haiyan Zhang; GuangxueYue, "An Incomplete Data Analysis Approach Based on the Rough Set Theory and Divide-and-Conquer Idea",IEEE conf on Fuzzy Systems and Knowledge Discovery,2007.FSKD2007.FourthInternational Conference on ,pp 119-123.
3. Kryszkiewicz M. "Rough set approach to incomplete information systems". *Information Sciences*, 1998,1 12,pp.39-49.
4. Jerzy W, Grzyrnala-Busse, Ming Hu. "A comparison of several approaches to missing attribute values in data mining". In: *Proc of the 2nd lnt' Conf on Rough Sets and Current Trends in Computing.* Berlin: Springer-Verlag, 2000, pp. 378-385.
5. G. Wang, "Extension of rough set under incomplete information systems," Fuzzy Systems, 2: pp. 1098-1103,2002.
6. Stefanowski J, Tsoukias A. "On the Extension of Rough Set,Under Incomplete Information". S Zhong, A Skowron, S Ohsuga(Eds.). *In:Proc. of the 7th Int'l Workshop on New DirectionsinRoughSets, Data Mining, and Granular Soft Computing.*Berlin:Springer-Verlag, 1999, pp.73-81.
7. *R. Laekowski, "On Decomposition for Incomplete Data,"FundamentaInformaticae,* 54(1): *pp.* 1-*16, 2003.*
8. Grzymala-Busse, J. W., Wang, A. Y.: Modified algorithms LEMIandLEM2 for rule induction from datawith missing attribute values, *Proceedings of 5th Workshop on Rough Sets andSoftComputing (RSSC'97) at the 3rd Joint Conference on InformationSciences.* Research Triangle Park (NC, USA), 1997.
9. Weihua Zhou, Wei Zhang, Yunqing Fu. "An incompletedata analysis approach using rough set theory". *IntelligentMechatronicsand Automation.2004,* pp.332-338.
10. *Sanjay Gaur and M.S. Dulawat,*"A *Closest Fit ApproachtoMissing Attribute Values in Data Mining*", *International Journal of Advances in Science and Technology, Vol.* 2, *No.4, 2011,pp.I8-23.*
11. *P. K. Dey*, "Imputation of Missing Data using Fuzzy-Rough Hybridization", *International Journal of Computer Sciences and Engineering, Vol. 6, No. 9, pp. 1-6, 2018.*
12. *P. K. Dey* and S. Mukhopadhyay, "Core reduct based preprocessing approach to incomplete data", *International Journal of Intelligent Engineering and Systems, Vol. 10, No. 5, pp. 19-28, 2017.*

P. K. Dey, "Attribute Reduction with Imputation of Missing Data using Fuzzy-Rough Set", *International Journal of Innovative Technology and Exploring Engineering*, Vol.8, No.11, pp.202--207,. ISSN: 2278-3075,2019.