JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Special Issue-1 E-ISSN: 2347-2693

A Modified Approach for Missing Values in Data Mining Based on Rough Set Theory, Divided –and-Conquer, Closest Fit Approach Idea

Abhijit Sarkar*
M.E. Student, CSE/IT,
University Institute of Technology,
The University of Burdwan.
Burdwan, India.
abhi.sarkaruit@gmail.com

Kasturi Ghosh
Assistant Professor, CSE/IT,
University Institute of Technology,
The University of Burdwan,
Burdwan, India.
kasturi.dikpati@gmail.com

***Abstract*— Missing data plays a key role in practical fields. How to remove this gap is the main objective of data preprocessing step in data-mining. Many methods such as Statistical and Prediction approaches are generally used for missing data analysis, but unfortunately both approaches have some disadvantages and applicable for serial missing values in column. This paper tries to remove these gaps which are resulting from the two mentioned methods. The proposed algorithm tries to merge up two previously mentioned methods. This modified approach utilizes the potential knowledge and laws suggested by the data in Information System, and some basic mathematical concepts and some concepts from Rough Set Theory. Experimental results show that the proposed algorithm provides better result than the above mentioned two methods.**

Keywords-component; Data mining; missing data; Data preprocessing; Statistical methods; Prediction methods; Rough Set Theory,Serially.

I. INTRODUCTION

As more and more data are being gathered in practical field of technology to get useful knowledge from it is very crucial. Data mining (DM) is the process of analyzing data from different perspectives (pre-existing databases) and summarizing it into useful information. It is generally depended upon the ideal data of the databases. But all data are not ideal due to missing values. Missing values in database is one of the biggest problems in data analysis. But improper data set makes the databases imbalanced. The effects of missing values are reflected on the final results. Generally people have to spend 60% of time on preparing complete data in DM process [6]. This indicates the noise data blocks DM process seriously [1].So it is very crucial to study on efficient data analyzing techniques which will give error free data in time.

Noise data can be divided into two categories: inconsistent and incomplete data. Transforming missing data into specified one is valid to deal with incomplete data. The task for deleting samples with missing data is easy but it wastes resource badly [5]. The probability statistics methods often relay on some statistical hypothesis, which is usually difficult to determine due to the wide state space of data set .The complexity of indiscernibility matrix for long data set is high and its feeling ratio is poor[3]. The advantages of Rough Set (RS), extended value tolerance relation and divide -and-conquer idea are taken in “Rough Set Theory and divide and conquer idea based incomplete Data Analysis Approach” (RSDIDA) [1].But it has some disadvantages too. It is only appropriate for the data which vary in small range. But in case of data which vary in wide ranges and uniform probability distribution for the attributes of those data is also very small, for that case RSDIDA gives poor result. This situation is improved by using statistical approach " A Closest fit approach" [2].But the Closest fit approach has the disadvantage that it has no use serial missing values in columns. In this paper proposed "Modified Rough Set, Divide-and-Conquer and Closest fit based Incomplete Data Analysis Approach" (MRDCIDA) tries to use the benefits of RSDIDA algorithm and the closest fit approach. So the proposed algorithm (MRDCIDA) deals with the prediction approach as well as with statistical approach. It gives the better result than the both approaches [1, 2]. MRDCIDA gives the more accurate results for the missing values in data sets and it can deals missing values serially.

I. ROUGH SET THEORY, EXTENDED MODEL AND STATISTICAL METHOD(CLOSEST FIT APPROACH)

A. Selecting a Related Concept of Rough Set theorey

First, Rough Set Theory was invented by Pawlak in 1982[8].It has an advantage that it needs only the information system based on equivalence relation.

Definition 1: An information system (IS) has four tuples S=<U, A, V , and f>, where $U=\{x_i, |i=1,2,...,n\}$ is a non empty finite set of object, called domain ,$A=\{a_k |k=1,2,...m\}$ is a finite nonempty finite set of attributes, $V=\cup V_a(a€A)$ is the value domain of attribute a, f: $U*A\rightarrow V$ is an function.

Definition 2: Assuming P is a set of equivalence relations on U, if R $\subseteq$P and R$\neq$ NULL, then $\cap$R (the intersection of all equivalence relations of R) is a relation too, called indiscernibility relation on R, marked ind(R) [1].

B. Maintaining the Integrity of Extended models of rough set

Classical RS theory depends on the hypothesis of complete information system and is not directly applicable for incomplete IS because equivalence relations are too strict and has been needed to relax to loose relations, as for tolerance relation [1].

Valued Tolerance Relation: The tolerance relation describes whether two objects are similar, but does not give their similarity degree [4, 7]. For example in Table1, it is seen that every IS contain several attributes. There are mainly two types of attribute: Conditional attributes and Decision attributes. Here conditional attributes are (A_1, A_2, A_3, A_4) and decision attribute is d [1].

Table1.Information System S_0:

U	A_1	A_2	A_3	A_4	d
O_1	2	3	1	0	y
O_2	3	2	3	0	y
O_3	*	3	2	1	n
O_4	2	*	*	2	y
O_5	1	*	*	*	y
O_6	*	3	*	*	n
O_7	2	3	1	*	y

For example in this table1, consider objects O_1,O_2 and O_7 in S_0 ,T(O_6 ,O_1) and T(O_7,O_1) can be resulted by the tolerance relation but it may be expected to get that O_7 is more similar to O_1 than O_6 intuitively ,because there is only one missing value in O_7 and there are three missing values in O_6.

To describe this phenomena , J. Stefanowski proposed valued tolerance relation[4].It based on the uniform probability distribution concept as because its conditional attributes ranges from only{0 to 3},so there is only four kind of possibility can be happened .In that way it can get the value tolerance matrix $T_V\{O_7,O_1\}=1/4$ and it is bigger than $T_v\{O_6,O_1\}=1/64$.It is consistent with the fact that O_7 is more similar to O_1 than O_6. Comparing O_1 and O_7 properly, they are included in the same group as because their every conditional attribute sequence and its values are same. The A_4 of O_7 is missing. But it is to be noted that the missing value will be any value between 0 and 3. That is why its probability is ¼.Comparing between the O_1 and O_6 there is 3 elements are missing and its uniform probability distribution is 1/4 *1/4 *1/4=1/64 [1].

Table2. Valued tolerance relation matrix of S_0

U	O_1	O_2	O_3	O_4	O_5	O_6	O_7
O_1	1	0	0	0	0	1/64	1/4
O_2	0	1	0	0	0	0	0
O_3	0	0	1	0	1/256	1/256	0
O_4	0	0	0	1	0	1/1024	1/64
O_5	0	0	1/256	0	1	1/4096	0
O_6	1/64	0	1/256	1/1024	1/4096	1	1/256
O_7	¼	0	0	1/64	0	1/256	1

Table3.Extended Valued tolerance relation matrix T_{EV} of S_0

U	O_1	O_2	O_3	O_4	O_5	O_6	O_7
O_1	0	0	0	0	0	0	0
O_2	0	0	0	0	0	0	0
O_3	0	0	0	0	1/256	0	0
O_4	0	0	0	0	0	1/1024	1/64
O_5	0	0	1/256	0	0	1/4096	0
O_6	1/64	0	1/256	1/1024	1/4096	0	1/256
O_7	1/4	0	0	1/64	0	0	0

Table2 and Table3 relation depends on MAS (Missing Attribute Set) and MOS (Missing Object Set) [3].

MAS:IS=<U,A,V, f>,if x_i€ U, then the missing attribute set MAS_i of object X_i ,missing object set MOS of IS are defined as :

$$MAS_i=\{k|a_k(X_i)=*,k=1,2,...m\}$$
$$MOS=\{i|\ MAS_i\neq NULL,i=1,2,...n\}$$
When the MAS=0,then MOS=0 also.

It is to be noted that MOS specifies each row and it is based upon some specified rules: T(i,j)=0,MAS_i subset of MAS_j else depends on Uniform Probability Distribution $P_k(i,j)$.

Now,

$$P_k(i,j)=1, \quad \text{when } a_k(x_i)\neq * \wedge a_k(x_j) \neq * \wedge a_k(x_i)= a_k(x_j)$$
$$P_k(i,j)=1/|V_k|, \text{when}(a_k(x_i)\neq * a_k(x_j) \neq *)V(a_k(x_i)\neq * \wedge a_k(x_j) =*)$$
$$P_k(i,j)=1/|V_K|^2, \text{ when } a_k(x_i)=* \wedge a_k(x_j)=*$$
$$P_k(i,j)=0, \quad \text{when } a_k(x_i)\neq * \wedge a_k(x_j) \neq * \wedge a_k(x_i)\neq a_k(x_j)$$

The basic difference between the T_V and T_{EV} is that T_{EV} works more perfectly. For O_3, O_5 and O_6 $T_V(3,3)=1$, $T_V(3,5)=1/256$, $T_V(3,6)=1/256$ and $T_{EV}(3,3)=0$, $T_{EV}(3,5)= 1/256$, $T_{EV}(3,6)=0$.By general observation it can be concluded that O_3 cannot be fulfilled by itself ,besides O_3 cannot be fulfilled by O_6.Therefore,only O_5 can fulfill O_3 and extended model proves this appropriately as $T_{EV}(3,5)= 1/256$[1].

C. Statistical Method(Closest fit Approach)

It is rooted on the concept of replacing missing attribute values by artificially generated values. It is very crucial for numerical attributes. In general this method is search of closest fit value which is very close to for numerical attributes. Smyth [9] and Zhang et al [10] have observed that data preparation is a fundamental stage of data analysis .Clark et al [11] considered a simplest method to handle these missing attributes values in which they replaced such values by the most common value in the attribute. Konkani et.al [12] tell that the most common values of the attribute restricted to the concept is used instead the most values for all case. The objective of study [2] is to determine the statistical technique which may be significant in the handling of missing attribute values. The Closest Fit Approach [2] formulates the missing values like this way:

The sample mean of the attribute is the most important and often used single statistics is defined as the sum of all sample values divided by the number of observation in the attribute/sample and is symbolically defined as:

$$X_{mean}=1/n \sum_{i=1}^{n} x_i$$

where X_{mean} is the observed values and i is the subscript of attribute X. It is an estimation of value of the mean of the population from which the sample is drawn. After that, it will have to record the preceding value(x_p) and succeeding value (x_s) from the missing value subscript(x_i).

$$x_p=\text{value}(x_{i-1})$$
$$x_s=\text{value}(x_{i+1})$$

where $x_p \neq x_s$ and x_p or $x_s \neq$ NULL

after that third stage the value of just x_p and x_s of the missing value subscript , compute the average of both values ($x_{mean(ps)}$ $x_{mean(ps)}=(x_p+x_s)/2$ now the estimated result $x_{est}=(X_{mean}+x_{mean(ps)})/2$.

Table 4(a): A Small Example of closest fit approach [2]

Year	Coal	Oil	Natural Gas
1960	1410	849	235
1961	1349	904	254
1962	1351	980	277
1963	*	1052	590
1964	1435	*	*
1965	1460	1219	351

$X_{mean(Coal)}=1401$, $X_{mean(Oil)}=1001$, $X_{mean(Natural\ Gas)}=341$

Now for the missing value of coal Pre and Post Value are X_p= 1351andX_s=1435.Average of Pre and Post$X_{mean(ps)}$=(1351+1435)/2=1393.

X_{est} for Coal=($X_{mean(Coal)}$+ $X_{mean(ps)}$)/2=(1401+1393)/2=1397

Similarly, X_{est} for Oil=1068 and X_{est} for Natural Gas= 406

Actual given Data set values for Coal, Oil and Natural Gas are 1396, 1137 and 386.Now just think about that, here this data set has no decision attributes values. So these rows are not so similar and do not give the so much good result in place of missing values. If it handles the situation with the rough set extension model and divide conquer approach, then the result will be more accurate.

I. A NEW APPROACH FOR MISSING VALUES IN DATA MINING BASED ON THE ROUGH SET THEOREY, DIVIDE-AND- CONQUER IDEA AND CLOSEST FIT IDEA.

In Previous example, every missing value has preceding value and succeeding value. Now think about missing value which has preceding value and succeeding value which are also unknown .New approach is concerned about serially missing values in data set.

If missing value is the first element of the column then the preceding value of that element will be the last value of the column just like circular concept.

If missing value is last element of the column then the succeeding value will be first element of the column like circular concept of the column.

TABLE 4(b) Simple example for missing values occurrences in data set

U	X	Y	Z

A	11	*	12
B	*	23	21
C	*	12	55
D	*	34	22
E	20	22	11
F	14	8	*

Now for the Column X, preceding value of serially missing value is 11 and succeeding value is 20.
For Column Y, preceding values of missing value is 8 and succeeding value is 23 like circular concept of column.
For Column Z, preceding value of missing value is 11 and succeeding value is 12.Calculation logics are same as closest fit approach. RSDIDA value must be calculated for these missing values. Then averaging the both values of Closest Fit and RSDIDA ,it will get better result in place of missing values which will appropriate for all kind of missing values occurrences.

A. The Propose Approach (MRDCIDA)

Now Proposed approach (MRDCIDA) just try to modify the RSDIDA approach with the help of above statistical approach as well as basic mathematical concepts .The values which are gained for a particular missing value for any specific dataset by using RSDIDA and Closest Fit approach are averaged. This average result is good for missing value in any kind of data set as because RSDIDA approach is only applicable for data set where data are varied in small ranges and in statistical approach where data are varied in high ranges. Every data set is divided with the help of the decision attribute value which eventually forms a subset. Now every Information Subsystem has same type of rows which are more similar than that total information dataset. Interestingly statistical approach does the great deal as because same rows bear the same type of characteristics.

B. Divide and Conquer idea

In this paper, the divide and conquer idea is adopted [1] to divide original system into some subsystems according to decision attribute value first, and then process every subsystem independently. The divide method is: For $S=<U, A, V, f>$, $A=C\cup\{d\}$, $V_d=\{d_1,d_2,.....,d_L\}$ is the value domain of decision attribute d, divide S into $S=S_1\cup S_2\cup....\cup S_L$ which are the particular subsets. They have the same criteria as rough set has. When every operation is done in every subset, then conquer the whole result to get the whole data set.

D. Modified Rough Set, Divide-and-Conquer and Closest fit based Incomplete Data Analysis Approach(MRDCIDA) :

Use Input: Incomplete Information system
$S^O=<U, A=C\cup D, V, f^O>$
Out Put: Complete Information System
$S=<U, A=C\cup D, V, f>$

Step1: To divide S^O according to decision attribute set D, get $U/D =\{U_1,U_2,.........,U_L\}$ and now every system is viewed as independent decision system, namely S^O_1, S^O_2,....., S^O_L .

Step2: For every decision subsystem $S^O_l=<U_1, A, V_1, f>$
$l€\{1,2,....,L\}$:

2.1 To compute the initial T^0_{EVl}, MAS^0_{il} and MOS^o_l of
S^o_l,$i€\{1,2,...,|U_i|\}$,let r=0

2.2 T^0_{EVl} matrix similarity maximum rows are replaced :
$a_k(O_i^{r+1})= a_k(O_{j'}^{r+1})$,if $a_k(O_i^r)$=* else
$a_k(O_i^{r+1})= a_k(O_i^r)$, $a_k(O_i^r)\neq$*

Then Compute MAS^{0}_{il}, MOS^{o}_{i}
Otherwise, matrix will be same.
Now,
O_{RMV} (Replaced Missing Value)= $a_k(O_i^{r+1})$;
When, any Column has missing value,
$O_{mean}=1/n \sum_{i=1}^{n} x_{i,j}$ here j=column number is fixed for any specified column
O_p=value ($x_{i-1,j}$)
x_s=value($x_{i+1,j}$)
if, $O_p \neq O_s$ and O_p or $O_s \neq$NULL
else, final result will be mean of (O_{RMV}, mean of(mean of(O_{p-1}, O_{s+1}), mean of the column except the missing values))
If, missing values are occurred in first or in last element,
Then column will be treated like circular concept.
after this stage, compute the value of x_p and x_s of the missing value subscript
Compute the average of both values ($x_{mean(ps)}$)
$O_{mean(ps),j}$=(O_p+ O_s)/2
Now the estimated result Oest, j = (O_{mean} + $O_{mean(ps)}$)/2.
Finally, O(i,J)=(O_{RMV}+ $O_{est,j}$)/2.

2.3 Compute MAS^{r+1}_{il}, MOS^{r+1}_{l} ,if $S_l^{r+1} = S_l^{r}$ or, MOS^{o}_{i}=NULL,
Finish the recycle, turn to step2.4, otherwise compute T^{r+1}_{EVl}, $O_{est,j}$, O(i,J);r=r+1
Return to step 2.2
2.4 Mark the decision System S_l^{r+1} as S_l

Step 3 Unite all the S_l,l€{1,2,...,L},get complete information system S.
Step 4: The end

II. EXPERIMENTS AND RESULT ANALYSIS

Hayes-roth and Iris Data Set [13], MRDCIDA is applied. There is no missing data in these data sets initially, so this paper generates some missing values with certain ratio (2% and 5 % respectively) randomly on the conditional attributes of both data sets to satisfy the experiments. As for small example Table5 from Hayes-roth data set it is shown:

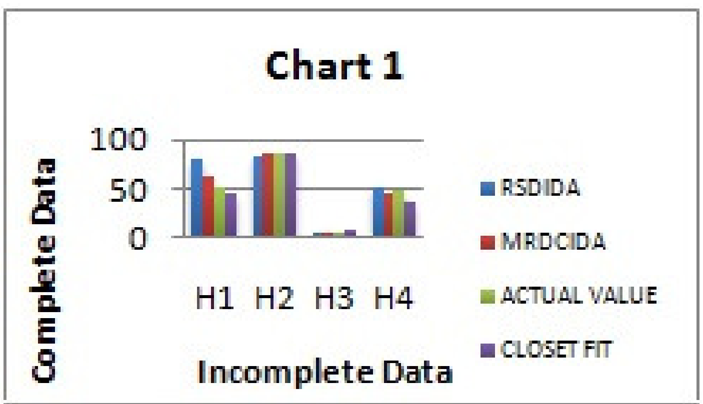
Table 5

Name	Hobby	Age	Educational Level	Marital Status	Class
92	2	1	1	2	1
68	3	3	2	1	1
105	3	2	1	1	1
81	1	2	1	1	1
94	1	1	2	1	1
20	1	*	3	3	1
*	1	*	1	1	1
36	2	*	1	1	1
68	3	3	2	1	1
89	1	2	2	1	1
19	3	2	1	3	1
16	3	2	1	3	1

For the attribute value of first column varies a lot in Table 5. So for missing value replacement, Uniform probability distribution value will be less. In this case Name-column value has the average of 62

By Proposed algorithm:
Here the pre value of missing value=20
Post value of missing value=36
Avg. of pre and post=28
Avg. of column avg. and pre and post avg. =45
In data set it is given 50.Now if it uses RSDIDA, then it will be then 81.
Now average of the both value = (45+81)/2=63 which is much closer to the actual value.
Now for age column it is seen that three missing values has been occurred serially.
Avg. of age column is 2.375.
Avg. of pre and post value of age column is 2.
Avg. of both is 2.185.

For this case RSDIDA for three missing values are 1, 1, 2.Now the final results are other experimental four results for four missing values 1.59, 1.59 and 2.05, and actual results are 1, 2, 2.So it is closed to the actual values. Results of Hayes, Iris are shown under the below charts:



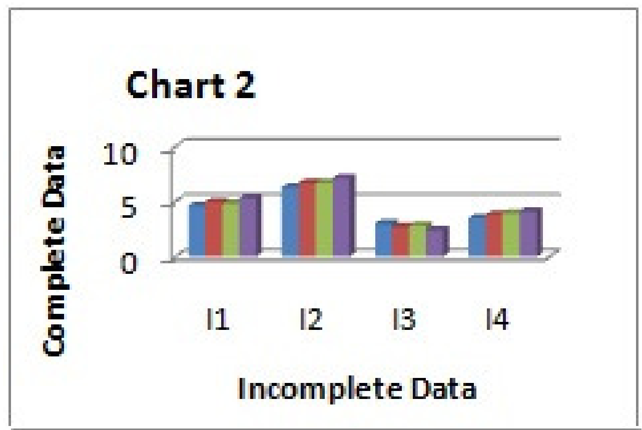


Chart 1 and Chart 2 are shown respectively for Hayes and Iris Data Set. Chart 2 bears the same color style characteristic as chart 1. Here, they are only shown for four missing values. The filling Operations for Hayes Data Set and Iris data set are declared as H1, H2, H3, H4 and I1, I2, I3, I4 respectively. It is to be noted that they denote each missing value operations. Three approaches RSDIDA, MRDCIDA, CLOSEST Fit are applied and their corresponding results are comparatively shown. For the most cases MRDCIDA is better than others two .This way, for other missing values of Hayes data and Iris data set can be calculated by MRDCIDA efficiently. This algorithm is applicable for any types of data set whose decision attributes are known.

III. CONCLUSION

Proposed MRDCIDA shows the greatness of the RS theory on incomplete data. It provides extended valued tolerance relation that describes filling capacity more precisely. The statistical approach and divide-and-conquer idea also do the superior jobs with it. Now when every subset is formed from decision attribute values, then the rows of the subset are same in maximum way. From the experimental results, it is shown that the results of MRDCIDA are much more close to the actual results when some columns of the data set have many missing values serially, then also the result of proposed algorithm is quite accurate and can be adopted as a preprocessing method in data mining. Next work is trying to figure out the missing values when maximum numbers of values are unknown in data set.

REFERENCES

[1] Zaimei Zhang, Renefa Li, Zhongsheng Li,Haiyan Zhang,Gungaxue Yue. "An incomplete Data Analysis approach Based on Rough St Theory and Divide-and-Conquer Idea", Fourth Int' Conf On Fuzzy Systems and Knowledge Discovery(FSKD 2007).

[2] Sanjay Gaur and M.S. Dulawat "A Closest Fit Approach to Missing Attribute Values in Data Mining", International Journal of Advances in Sciences and Technology Vol.2,No.4,2011 .

[3] Weihua Zhou,Wei Zhang,Yunique Fu."An Incomplete data analysis approach using rough set theory", Intelligent Mechatronics and animation.2004,pp.332-338.

[4] Stenfanowski J,Tsoukias A. "On the Extension of Rough Sets Under Incomplete Information". S Zhong, A Skorown, S Ohsuga (Eds).In: Proc. Of the 7th Int'l Workshop on New Directions in Rough Sets, Data Mining, and Granular Soft Computing.Berlin:Springer-verlag,1999,pp.73-81

[5] Jerzy W,Grzymal-Busse,Ming Hu. "A comparison of several approaches to missing attribute values in data mining". In: Proc of the 2nd Int' Conf On Rough Sets and Currents Trends in Computing.Berlin:Springer-Verlag,2000,pp.378-385.

[6] Cios K J.Kurgan L. A. "Trends in data mining and Knowledge

Discovery". In: Knowledge discovery in advanced information systems,Pal,N.R., Jain,L.C., Teodereresku N.eds.Spinger,2002.

[7] Kryszkiewiez M. "Rough set approach to incomplete information Systems". Information Sciences,1998, 112,39-49

[8] Pawlak Z. "Rough Sets". International Journal of Computer and Information Sciences,1982,11(5),pp.341-356.

[9] Symth, P., " Data mining at the interface of computer Science and Statistics", Data mining for Scientific and engineering applications, Department of Information and Computer Science, University of California,CA,92697-3425,Chapter 1,pp.1-20,2001.

[10] Zhang,S., Zhang,C., and Young,Q., "Data Preparation for data mining". Applied Artificial Intelligence,Vol.17,pp.375-381,2003.

[11] Clark, P., and Niblett ,T., "The CN2 induction algorithm", Machine Learning, Vol. 3,pp.261-283,1983.

[12] Konoenko ,I., Bratko, I, and Roskar,E., " Experiments in automatic learning of medical diagnostic rules", Technical Report, Jozef Stefan Institute,LIjubal-jans,Yugoslavia,1984.

[13] http://www.ics.uci.edu/-mlearn/MLRespository.html.