International Journal of Advance Foundation and Research in Computer (IJAFRC)
Volume 1, Issue 3, March 2014. ISSN 2348 - 4853

Estimation of Missing Value at Extremes in Data Mining

Dr. Sanjay Gaur
Associate Professor & Principal
Advent Institute of Management Studies, Udaipur INDIA
sanjay.since@gmail.com

A B S T R A C T

Data cleaning is a fundamental stage of the data preparation for data mining. Missing values in the database is a common problem analyst face in the data mining. Missing data are a persistent problem that can cause partiality to unproductive analysis. Missing values at the initial and finish level in the attribute is makes difficult the data analysis and final or consolidated result. It affects loss of accuracy of mediatory result and calculations. By the help of some statistical methods and techniques we can recover missing data and reduce uncertainties. Here, we introduce ratio based closest fit approach by which we can recover missing attribute values at the extremes.

Index Terms: Data mining, Extremes, Attribute, Data cleaning, Incompleteness, Missing Values
MSC(2010) Subject Classification:62-07,62N02,62Q99.

I. INTRODUCTION

Missing values in database is solitary of the biggest problems faced in data analysis and in data mining applications. The belongings of missing values are highly reflected on the final results. The problems get increases when values are missed at the initial or end of the attribute. In this study, ratio based closest fit approach is introduced which find out pattern to generate missing values from a real imbalanced database with missing values at the extremes.

Buck[1] convoluted an estimation method by which missing values will recovered in multivariate data suitable for use with an electronic computer. Chen et al.[2] studied and discussed about multiple imputation for missing ordinal data. Gaur et al.[3,4] and Busse[5] discussed an assortment of algorithms which are useful for estimation of missing values. Kim and Curry[6] well thought-out explain the treatment of missing data in their analysis. Rubin[8] explored about presumption and missing data and various imputations for non-response in the survey. Zhang et al.[10] have considered that data preparation is a fundamental stage of data analysis. The purity of database should be maintains Qin[7] considered the semi-parametric optimization for missing data imputation. Sharma and Gaur[9] give a fabulous algorithm to recover missing values when a block of values in the attribute is missed under the title of contiguous agile approach to manage odd size missing block in data mining.

Although, it may be good option to exclude the missing boundary values during the data analysis. Because boundaries values does not give significant impact on unbalancing the central tendency as well as any other result despite the missing values in the central part of the attribute. But a value missed at the boundary is matter of great consideration which affects the initial of data analysis. The proposed method is search of ratio based closest fit value which is very near to the original value and the values adjustable with other values of the attribute.

II. METHODOLOGY: MISSING VALUE AT THE INITIAL OF ATTRIBUTE

 © 2014, IJAFRC All Rights Reserved

In the process of generation of single value at the beginning of attribute, we introduce ratio based closest fit approach for empty initial subscript. Here, we assume that the relation (Table) has at least two or three attributes. Each attribute in the table has equal number of record. Imagine that there are two attributes X and Y corresponding to attribute year and serial number. Attributes X, have missing values at the boundary. At the beginning read all the table attributes with available (observed) and missing values (missed) case. Here the attribute X has empty subscript or missing values. A scanning pointer applied on the attribute X from first subscript X[1] to the last one X[n]. When search or scan pointer point out the empty subscript of the attribute, which is actually the missing values case in the attribute. The missing value case is pointed by the subscript of the attribute and is denoted by the variable(x_i). In this situation, the first subscript is empty or NULL. Now we have to point out on value (y_i) which is corresponding value of (x_i) in attribute Y. Now, we have to record the succeeding value (x_s) from the missing value subscript (x_i). Similarly, we have to record the succeeding value (y_s) from the subscript (y_i).

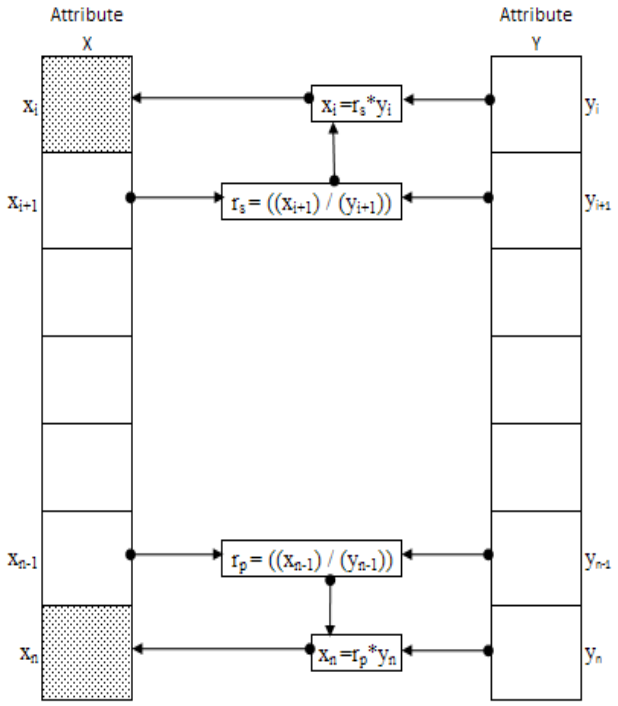

Fig 1: Block Diagram of Ration Based Closest Fit Approach for Extremes

This is corresponding to missing value subscript (x_i).

$$x_s = value\ (x_{i+1}) \quad \& \quad y_s = value\ (y_{i+1})$$

where $x_i \neq x_s$ and $x_s \neq \text{NULL}$, $y_i \neq y_s$ and y_i *and* $y_s \neq \text{NULL}$

Now compute ratio (r_s) between succeeding value (x_s) and (y_s) of the missing value subscript.

$$r_s = x_s / y_s$$

To calculate, estimated value for the boundary missing values subscript (x_i), multiply final ratio (r_s) with the value of (y_i). This estimated value is closest fit value for (x_i) which is replaced by the x_{est} and it is separately computed for each attribute.

$$x_{est} = y_i * r_s$$

 © 2014, IJAFRC All Rights Reserved

ALGORITHM: (FORWARD PASS)

Attribute $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_n\}$

where $X = X_{obs} + X_{mis}$

$X_{obs} = \{x_1, \dots, x_k\}$ // Attribute values observed

$X_{mis} = \{x_{k+1}, \dots, x_n\}$ // Attribute values missing

$Y = Y_{obs} + Y_{mis}$

$Y_{obs} = \{y_1, \dots, y_k\}$ // Attribute values observed

$Y_{mis} = \{y_{k+1}, \dots, y_n\}$ // Attribute values missing

size of (X) = = size of (Y)

Read $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_n\}$ // Read attributes with observed and missing values

For $i = 1$ to n do // Running loop from first to last value of attribute

If (value (x_i) == NULL && (i== 1)) then x_s = value (x_{i+1})

// Value of succeeding of x_i , where $x_i \neq x_s$ and $x_s \neq$ NULL

y_s = value (y_{i+1}) // Value of succeeding of y_i

where $y_i \neq y_s$ and y_i *and* $y_s \neq$ NULL

$r_s = x_s / y_s$

$x_{est} = y_i * r_s$ **// Estimated value**

value (x_i) = x_{est} // Assigning estimated value to missing value place

$i = i + 1$ // increment in counter by one

repeat until($i >= n$)

Stop

ALGORITHM: (FOR BACKWARD PASS)

$r_s = x_s / y_s$

$y_{est} = x_i / r_s$ **// Estimated value**

OR

$r_s = y_s / x_s$

$y_{est} = x_i * r_s$ **// Estimated value**

value (y_i) = y_{est} // Assigning estimated value to missing value place

Note: *Only bold part of forward pass algorithm were replace by the bold part of backward pass algorithm, rest part of forward pass are remain same for backward pass algorithm.*

III. METHODOLOGY: MISSING VALUE AT THE END OF ATTRIBUTE

In the current scene, the last subscript is empty or NULL. Now, we have to point out on value (y_i) which is corresponding the value of (x_i) in attribute Y. We have to record the preceding value (x_p) from the missing value subscript (x_i).Similarly, we have to record the preceding value (y_p) from the subscript (y_i), which is (y_n) now. This is corresponding to missing value subscript (x_n).

$x_p = value\ (x_{n-1})$ & $y_p = value\ (y_{n-1})$

where $x_i \neq x_p$ and $x_p \neq$ NULL , $y_i \neq y_p$ and y_n *and* $y_p \neq$ NULL

Now compute ratio (r_p) between succeeding value(x_p) and (y_p) of the missing value subscript.

$r_p = x_p / y_p$

 © 2014, IJAFRC All Rights Reserved

To calculate estimated value for the boundary missing values subscript (x_i) *or* (x_n), multiply final ratio (r_p) with the value of (y_i), which is (y_n).

$$x_{est} = y_i * r_p$$

This estimated value is closest fit value for (x_n) and (x_i) is replaced by x_{est}.

ALGORITHM

Attribute $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_n\}$

where $X = X_{obs} + X_{mis}$

$X_{obs} = \{x_1, \dots, x_k\}$ // Attribute values observed

$X_{mis} = \{x_{k+1}, \dots, x_n\}$ // Attribute values missing

$Y = Y_{obs} + Y_{mis}$

$Y_{obs} = \{y_1, \dots, y_k\}$ // Attribute values observed

$Y_{mis} = \{y_{k+1}, \dots, y_n\}$ // Attribute values missing

Assume that size of (X) = = size of (Y)

Read $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_n\}$ // Attribute with observed and missing values

For *i =1* to *n* do // Running loop from first to last value of attribute

If (value (x_i) == NULL&& (i == n) then

x_p = value (x_{i-1}) // Value of preceding of x_n

where $x_n \neq x_p$ and $x_p \neq$ NULL

y_p = value (y_{i-1}) // Value of preceding of y_n

where $y_n \neq y_p$ and y_n *and* $y_p \neq$ NULL

$r_p = x_p / y_p$

$x_{est} = y_i * r_p$ // Estimated value

value (x_i) = x_{est} // Assigning estimated value to missing value place

i = i + 1 // increment in counter by one

repeat until(*i >=n*)

Stop

IV. DISCUSSION OF RESULTS

Measure of central tendency (mean): Table-1 shows the global carbon dioxide emissions from fossil fuel burning by fuel type coal, oil and natural gas from 1980-2009. The mean of global carbon dioxide emissions due to coal, oil and natural gas are 2484, 2629 and1143 respectively. After missing values at the extremes, the mean calculated from incomplete data sets are 2471 for coal, 2623 for oil and 1143 for natural gas. It is observed that mean values of incomplete data sets are lower than the mean values from the standard dataset.

The proposed ratio based approach method is applied on the data sets of Table 1 to fill up the missing values. It is observed that mean values of coal, oil and natural gas are 2485, 2625 and 1145 respectively. It is considerable that the mean values obtained after replacing the missing values by the proposed approach very close to the actual mean as given.

 © 2014, IJAFRC All Rights Reserved

Standard Deviation: From the analysis of result of standard deviation it is found that after estimation of missing values, the values of standard deviation obtained are very similar to the standard deviation of standard dataset. On the basis of result we can say that proposed algorithm is appropriate for missing values estimation and recovery.

Coefficient of Variation: From the analysis of result of co-efficient of variation (CV) it is found that, after estimation of missing values, the values of co-efficient of variation is not significantly change or slightly decline which shows that the series is uniform now..

Analysis of Variance: We wish to test the hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3$ Against the alternative

H_1: at least two μ's are different (i.e. at least one of the equalities does not hold).

For testing this hypothesis we setup the following analysis of variance for all the variables:

One Way ANOVA (COAL)

Source of Variation	*SS*	*Df*	*MS*	*F*
Between Groups	3593.146497	2	1796.5732	0.011594404
Within Groups	13170899.5	85	154951.76	
Total	13174492.65	87		

Table Value :- F(2, 85) at 5% Level of Significance = 3.0718 , 1% Level of Significance = 4.7865,

One Way ANOVA (OIL)

Source of Variation	*SS*	*Df*	*MS*	*F*
Between Groups	624.8936439	2	312.44682	0.003388721
Within Groups	7837169.573	85	92201.995	
Total	7837794.467	87		

Table Value :- F(2, 85) at 5% Level of Significance = 3.0718 , 1% Level of Significance = 4.7865,

One Way ANOVA (GAS)

Source of Variation	*SS*	*Df*	*MS*	*F*
Between Groups	75.76588506	2	37.882943	0.000525581
Within Groups	6126647.381	85	72078.204	
Total	6126723.147	87		

Table Value :- F(2, 85) at 5% Level of Significance = 3.0718 , 1% Level of Significance = 4.7865,

Decision and Conclusion: Since F (Calculated) < 3.0781 so accept H_0 at 5% level of significance and hence conclude that there is no significant difference among groups of Coal, Oil and Gas regarding mean value.

V. CONCLUSION

This paper shows the universal truth that there is no accurate method of treatment missing attribute values. The proposed approach is an important one for the arithmetical real value having deviation from

 © 2014, IJAFRC All Rights Reserved

the mean due to their presence in the attribute. This approach gives proper result for the consolidated report which is generated from the database. As a result, it is observed that techniques for handling of missing attribute values at the extremes should be fit according to environment and type of data. The method is appropriate for the consolidated report and suitable to small size attribute.

VI. FUTURE SCOPE

Proposed approach provides proper consolidated report using ratio between the relative attributes of the database. It is obvious that values in the parallel/relative attribute or dependent attribute have certain correlations in the database. Furthermore the more work can be undertaken to identify the correlation between the attributes, which in turn shall help in recovery of missing values. One can also laid the emphasis on working upon the said research as a basis and evolve more types of patterns and minutely elaborate on nature and distribution of values in the attribute for filling missing values and its implications. So, there is huge scope underling to be worked upon yet and to identify patterns between attributes in the database to recover missing values.

VII. REFERENCE

[1] Buck, S.F., A method of estimation of missing values in multivariate data suitable for use with an electronic computer, J. Royal Statistical Society, Series B, Vol-2, pp. 302-306(1960).

[2] Chen, L., Drane, M.T., Valois, R.F., and Drane, J.W., Multiple imputation for missing ordinal data, Journal of Modern Applied Statistical Methods, Vol.-4, No.1, pp. 288-299(2005).

[3] Gaur, Sanjay and Dulawat, M.S., Univariate Analysis for Data Preparation in context of Missing Values ,Journal of Computer and Mathematical Sciences, Vol.-1, No. 5, pp. 628-635(2010).

[4] Gaur, Sanjay and Dulawat, M.S., A Closest Fit Approach to Missing Attribute Values in Data Mining,, International Journal of advances in Science and Technology, Vol.-2, issue-4, (2o11).

[5] Grzymala-Busse, J.W., Data with missing attribute values: Generalization of in-discernibility realtion and rules induction, Transactions of Rough Sets, Lecture Notesin Computer Science Journal Subline, Springer-Verlag, 1,8-95 (2004).

[6] Kim, J.O., and Curry, J., The treatment of missing data in multivariate analysis, Social Methods and Research, Vol.-6, pp. 215-240(1977).

[7] Qin, Y.S., Semi-parametric optimization for missing data imputation, Applied Intelligence, Vol.-27, No. 1, pp. 79-88(2007).

[8] Rubin, D.B., Inference and missing data, Biometrika, 63, pp. 581-592(1976).

[9] Sharma, Swati and Gaur, Sanjay, Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining", International Journal Of Advanced Research In Computer Science, Vol.- 4(11), pp 214-217 (2013).

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

 © 2014, IJAFRC All Rights Reserved

Table: 1

Ratio Based Closest Fit Approach For Extremes (Single Missing Values)

Global Carbon Dioxide Emissions from Fossil Fuel Burning by Fuel Type, 1960-2009 (In Million Tones of Carbon)

		Standard Data			missing values(Coal & Gas)			missing values(Oil)			recovered values(Coal & Gas)			recovered(Oil)		
SN	Year	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas
		Million Tons of Carbon			Million Tons of Carbon			Million Tons of Carbon			Million Tons of Carbon			Million Tons of Carbon		
1	1980	1,947	2,422	740		2,422		1,947		740	**2,032**	2,422	**789**	1,947	**2,319**	740
2	1981	1,921	2,289	756	1,921	2,289	756	1,921	2,289	756	1,921	2,289	756	1,921	2,289	756
3	1982	1,992	2,196	746	1,992	2,196	746	1,992	2,196	746	1,992	2,196	746	1,992	2,196	746
4	1983	1,995	2,177	745	1,995	2,177	745	1,995	2,177	745	1,995	2,177	745	1,995	2,177	745
5	1984	2,094	2,202	808	2,094	2,202	808	2,094	2,202	808	2,094	2,202	808	2,094	2,202	808
6	1985	2,237	2,182	836	2,237	2,182	836	2,237	2,182	836	2,237	2,182	836	2,237	2,182	836
7	1986	2,300	2,290	830	2,300	2,290	830	2,300	2,290	830	2,300	2,290	830	2,300	2,290	830
8	1987	2,364	2,302	893	2,364	2,302	893	2,364	2,302	893	2,364	2,302	893	2,364	2,302	893
9	1988	2,414	2,408	936	2,414	2,408	936	2,414	2,408	936	2,414	2,408	936	2,414	2,408	936
10	1989	2,457	2,455	972	2,457	2,455	972	2,457	2,455	972	2,457	2,455	972	2,457	2,455	972
11	1990	2,409	2,517	1,026	2,409	2,517	1,026	2,409	2,517	1,026	2,409	2,517	1,026	2,409	2,517	1,026
12	1991	2,341	2,627	1,069	2,341	2,627	1,069	2,341	2,627	1,069	2,341	2,627	1,069	2,341	2,627	1,069
13	1992	2,318	2,506	1,101	2,318	2,506	1,101	2,318	2,506	1,101	2,318	2,506	1,101	2,318	2,506	1,101
14	1993	2,265	2,537	1,119	2,265	2,537	1,119	2,265	2,537	1,119	2,265	2,537	1,119	2,265	2,537	1,119
15	1994	2,331	2,562	1,132	2,331	2,562	1,132	2,331	2,562	1,132	2,331	2,562	1,132	2,331	2,562	1,132
16	1995	2,414	2,586	1,153	2,414	2,586	1,153	2,414	2,586	1,153	2,414	2,586	1,153	2,414	2,586	1,153
17	1996	2,451	2,624	1,208	2,451	2,624	1,208	2,451	2,624	1,208	2,451	2,624	1,208	2,451	2,624	1,208
18	1997	2,480	2,707	1,211	2,480	2,707	1,211	2,480	2,707	1,211	2,480	2,707	1,211	2,480	2,707	1,211
19	1998	2,376	2,763	1,245	2,376	2,763	1,245	2,376	2,763	1,245	2,376	2,763	1,245	2,376	2,763	1,245
20	1999	2,329	2,716	1,272	2,329	2,716	1,272	2,329	2,716	1,272	2,329	2,716	1,272	2,329	2,716	1,272
21	2000	2,342	2,831	1,291	2,342	2,831	1,291	2,342	2,831	1,291	2,342	2,831	1,291	2,342	2,831	1,291
22	2001	2,460	2,842	1,314	2,460	2,842	1,314	2,460	2,842	1,314	2,460	2,842	1,314	2,460	2,842	1,314
23	2002	2,487	2,819	1,349	2,487	2,819	1,349	2,487	2,819	1,349	2,487	2,819	1,349	2,487	2,819	1,349
24	2003	2,638	2,928	1,399	2,638	2,928	1,399	2,638	2,928	1,399	2,638	2,928	1,399	2,638	2,928	1,399
25	2004	2,850	3,032	1,436	2,850	3,032	1,436	2,850	3,032	1,436	2,850	3,032	1,436	2,850	3,032	1,436
26	2005	3,032	3,079	1,479	3,032	3,079	1,479	3,032	3,079	1,479	3,032	3,079	1,479	3,032	3,079	1,479
27	2006	3,193	3,092	1,527	3,193	3,092	1,527	3,193	3,092	1,527	3,193	3,092	1,527	3,193	3,092	1,527
28	2007	3,295	3,087	1,551	3,295	3,087	1,551	3,295	3,087	1,551	3,295	3,087	1,551	3,295	3,087	1,551
29	2008	3,401	3,079	1,589	3,401	3,079	1,589	3,401	3,079	1,589	3,401	3,079	1,589	3,401	3,079	1,589
30	2009	3,393	3,019	1,552		3,019		3,393		1,552	**3,334**	3,019	**1,558**	3,393	**3,007**	1,552
	Mean	2,484	2,629	1,143	2,471	2,629	1,143	2,484	2,623	1,143	2,485	2,629	1,145	2,484	2,625	1,143
	SD	407.9	303.0	273.1	370.5	303.0	260.6	407.9	302.3	273.1	399.9	303.0	271.1	407.9	305.5	273.1
	CV	16.421	11.526	23.895	14.996	11.526	22.805	16.421	11.526	23.895	16.094	11.526	23.680	16.421	11.637	23.895

Source: www.earth_policy.org

 © 2014, IJAFRC All Rights Reserved