International Journal of Mathematical Archive-2(9), 2011, Page: 1712-1715
IJMA Available online through www.ijma.info ISSN 2229 – 5046

WEIGHTED AVERAGE CLOSEST FIT APPROACH TO HANDLE MISSING VALUES

Sanjay Gaur* and M. S. Dulawat

Department of Mathematics and Statistics, Maharana Bhupal Campus,
Mohanlal Sukhadia University, Udaipur-INDIA

E-mail: sanjay.since@gmail.com, dulawat_ms@rediffmail.com

(Received on: 07-09-11; Accepted on: 21-09-11)

ABSTRACT

Data preparation for data mining is a fundamental stage of data analysis. Data with missing values complicates both the data analysis and final result. It also affects loss of accuracy of mediatory result and calculations. To overcome this situation some sort of applied statistical techniques are required to employ during the data preparation. With the help of statistical methods and techniques, we can recover incompleteness of missing data and reduce ambiguities. In this paper, we introduce weighted average based sequential method by which missing attribute values are recovered.

Key Words: *Missing Values, Attribute, Data preparation, Incompleteness, Data mining, Weighted.*

MSC (2010) Subject Classification: *62-07,62N02, 62Q99.*

1. INTRODUCTION:

Missing values in database is solitary of the biggest problems faced in data analysis and in data mining applications. The effects of these missing values are reflected on the final results. In this study, weighted average based statistical method is introduced and discussed which provides an approach to find out pattern to generate missing values from a real imbalanced database with missing values.

The utility of statistical methods has gained objects in exploring estimation and prediction techniques. Kim and Curry[5] considered the treatment of missing data in their analysis. Rubin[7] explored about inference and missing data and multiple imputations for non-response in the survey. Zhang et. al[8] have considered that data preparation is a fundamental stage of data analysis. Chen et. al[1] studied and discussed about multiple imputation for missing ordinal data. Qin[6] considered the semi-parametric optimization for missing data imputation. Gaur Sanjay and Dulawat M. S.[2,3] and Gyzymala-Busse[4] discussed various algorithms which are useful for estimation of missing values.

2. FORMULATION OF PROBLEM:

The proposed method is based on replacing missing attribute values by the artificially generated values. This method is search of closest fit value which is very close to the original value and the values of just preceding and succeeding value of the missing values.

In the process of generation of weighted closest fit values for missing value place, we first read complete attribute with available (observed) and missing values case. Now search pointer point out the empty cell 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 denoted by the variable x_i.

After pointing missing value case, we have to record the first, second and third proceeding value x_{p1} , x_{p2} and x_{p3} respectively. Now we have to record preceding values in the weighted manner. The first preceding value get 75%, second get 50% and third get 25% weight. Here in the given algorithm it is shown by variable Vp1, Vp2 and Vp3. For equality of all values, the sum of weighted preceding values is divided by the sum of percentage of weight.

Now same process applied to record the succeeding value from the missing value subscript (x_i) First, second and third succeeding values are denoted by x_{s1} , x_{s2} *and* x_{s3} respectively. Now we have to record succeeding values in the weighted manner. The first succeeding value get 75%, second get 50% and third get 25% weight. Here in the given

Corresponding author:** *Sanjay Gaur* *, ***E-mail: *sanjay.since@gmail.com*

algorithm it is shown by variable Vs_1, Vs_2 and Vs_3. For equality of all values, the sum of weighted succeeding values is divided by the sum of percentage of weight.

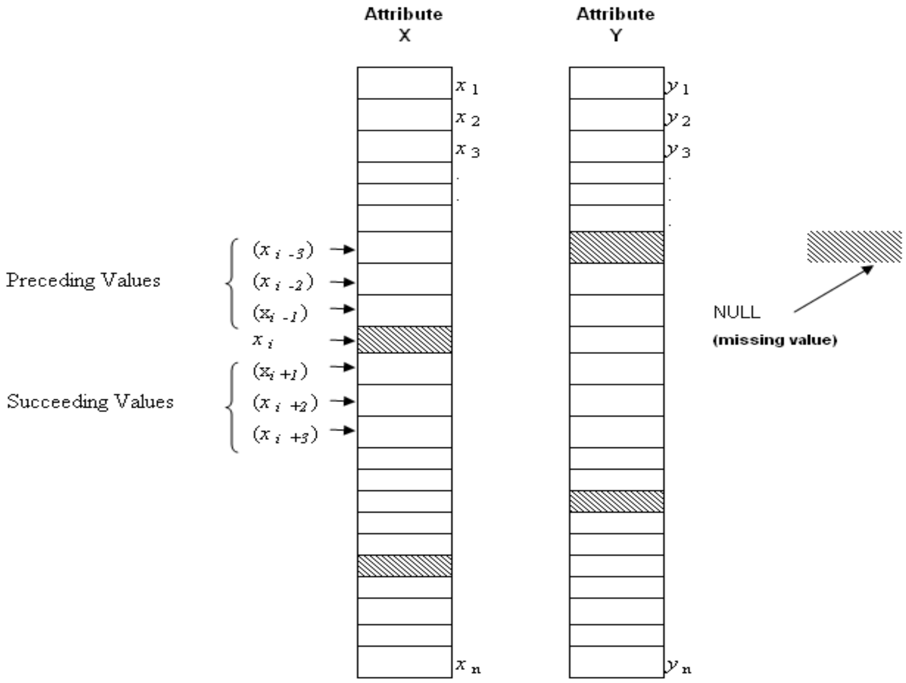

Figure: Block Diagram of Weighted Based Closest Fit Approach

At the third stage, after recording the weighted average values of preceding value (x_p) and succeeding value (x_s) of the missing value subscript, we compute the average of both values ($\bar{x}_{ps}$) which is also estimated value for missing value case.

$$x_{est} = \bar{x}_{ps} - (x_p + x_s)/2$$

Now missing value subscript (x_i) is replaced by the estimated values (x_{est}) and it is separately computed for every missing values subscripts.

3. ALGORITHM:

```
  Attribute      $X = \{x_1, \ldots, x_n\}$
  where          $X = X_{obs} + X_{mis}$
              $X_{obs} = \{x_1, \ldots, x_k\}$        // Attribute values observed
              $X_{mis} = \{x_{k+1}, \ldots, x_n\}$     // Attribute values missing
 Read            $X = \{x_1, \ldots, x_n\}$   // Attribute with observed and missing values
  For  $i = 1$ to $n$  do
        If (value ($x_i$) == NULL) then
              $x_{p1}$ = value ($x_{i\ -1}$)          // Value of first preceding of  $x_i$
              $x_{p2}$ = value ($x_{i\ -2}$)          // Value of second preceding of  $x_i$
              $x_{p3}$ = value ($x_{i\ -3}$)          // Value of third preceding of  $x_i$
              $V_{p1}=x_{p1}*.75$
              $V_{p2}=x_{p2}*.50$
              $V_{p3}=x_{p3}*.25$
              $x_{s1}$ = value ($x_{i\ +1}$)          // Value of first succeeding of  $x_i$
              $x_{s2}$ = value ($x_{i\ +2}$)          // Value of second succeeding of  $x_i$
              $x_{s3}$ = value ($x_{i\ +3}$)          // Value of  third succeeding of  $x_i$
              $V_{s1}=x_{s1}*.75$
              $Vs_2=x_{s2}*.50$
              $V_{s3}=x_{s3}*.25$
              $x_p= ((V_{p1} +V_{p2} +V_{p3})/(.75+.50+.25))$
              $x_s= ((V_{s1} +V_{s2} +V_{s3})/ (.75+.50+.25))$
              $x_{ps} = (x_p + x_s) / 2$                  // Average of preceding and succeeding
              $x_{est} = \bar{x}_{ps}$                  // Estimated value
              value ($x_i$) = $x_{est}$              // Assigning estimated value to missing value place
          $i = i + 1$
        repeat untill($i >=n$)
        Stop
```

4. DISCUSSION OF RESULTS:

Table-A given in appendix shows the world wide emission of carbon dioxide (CO_2) from the consumption of Oil and Natural Gas respectively for the years 1986 to 2009. The mean emission of carbon dioxide (CO_2) due to Oil and Natural Gas are 2725 and 1236 respectively. Table-B shows the variables with observed and missing values. It may be noted that in the planned way 15 % of the values are missing in the random manner for all the variables from Table-A. The means calculated from incomplete data sets are 2722 for Oil and 1231 for Natural Gas.

It is observed that mean values of incomplete data sets of Table-B are lower than the mean values from both the variables of Table-A.

The proposed weighted closest fit method is applied on the data sets of Table-B to fill up the missing values. These closest fit values are shown in Table-C for both the variables which are highlighted by underline. It is observed that mean values of Oil and Natural Gas are 2725 and 1236 respectively. It is considerable that the mean values obtained after replacing the missing values by the weighted closest fit values in Table-C are 100% close to the actual mean as given in Table-A.

5. CONCLUSION:

It is universal truth that, there is no absolute, technique of treatment missing attribute values. The proposed weighted closest fit method is useful for arithmetical and statistical attribute. Here the recovered missing values are very near to original value having deviation from the mean. This method is added appropriate for the consolidated report which is generated from the database. As a result, it is observed that techniques for handling of missing attribute values should be chosen individually or based on the nature and type of data.

5. REFERENCE:

[1] 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).

[2] Gaur, Sanjay and Dulawat, M. S., Improved closest fit techniques to handle missing attributes values, Journal of Computer and Mathematical Sciences, Vol.-2(2), 384-390 (2011).

[3] 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(4) 18-24 (2011).

[4] 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, Vol-1, pp. 78-95(2004).

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

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

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

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

APPENDIX:

Global Carbon Dioxide Emissions from Fossil Fuel Burning by Fuel Type, 1986-2009

Table –A		
Original Values		
Year	Oil	Natural Gas
1986	2,290	830
1987	2,302	893
1988	2,408	936
1989	2,455	972
1990	2,517	1,026
1991	2,627	1,069
1992	2,506	1,101
1993	2,537	1,119
1994	2,562	1,132
1995	2,586	1,153
1996	2,624	1,208
1997	2,707	1,211
1998	2,763	1,245
1999	2,716	1,272
2000	2,831	1,291
2001	2,842	1,314
2002	2,819	1,349
2003	2,928	1,399
2004	3,032	1,436
2005	3,079	1,479
2006	3,092	1,527
2007	3,087	1,551
2008	3,079	1,589
2009	3,019	1,552

Mean=	2,725	1,236

Table –B		
Missing Values(15% approx)		
Year	Oil	Natural Gas
1986	2,290	830
1987	2,302	893
1988	2,408	936
1989		
1990	2,517	1,026
1991	2,627	1,069
1992	2,506	1,101
1993	2,537	1,119
1994	2,562	1,132
1995		
1996	2,624	1,208
1997	2,707	1,211
1998	2,763	1,245
1999	2,716	1,272
2000	2,831	1,291
2001		1,314
2002	2,819	
2003	2,928	1,399
2004	3,032	1,436
2005	3,079	1,479
2006		1,527
2007	3,087	
2008	3,079	1,589
2009	3,019	1,552

Mean=	2,722	1,231

Table –C		
Table with Estimated Values		
Year	Oil	Natural Gas
1986	2,290	830
1987	2,302	893
1988	2,408	936
1989	**2,453**	**977**
1990	2,517	1,026
1991	2,627	1,069
1992	2,506	1,101
1993	2,537	1,119
1994	2,562	1,132
1995	**2,610**	**1,168**
1996	2,624	1,208
1997	2,707	1,211
1998	2,763	1,245
1999	2,716	1,272
2000	2,831	1,291
2001	**2,836**	1,314
2002	2,819	**1,361**
2003	2,928	1,399
2004	3,032	1,436
2005	3,079	1,479
2006	**3,056**	1,527
2007	3,087	**1,534**
2008	3,079	1,589
2009	3,019	1,552

Mean=	2,725	1,236

Source: www.earth-policy.org

****