International Journal of Mathematical Archive-5(9), 2014, 27-32
Available online through www.ijma.info ISSN 2229 – 5046

DEVIATION RATIO BALANCE METHOD TO RECOVER MISSING VALUES IN DATA MINING

Sanjay Gaur
Associate Professor & Principal, Advent Institute Management Studies, Udaipur, India.

Mukta Agarwal*
Sr. Lecturer Vidya Bhawan Rural Institute Udaipur, India.

(Received On: 13-08-14; Revised & Accepted On: 25-09-14)

ABSTRACT

In the area of data mining, data preparation is one of the fundamental stage. If there is any missing value in the existing record, it makes hard the data preparation and analysis. Whereas, Data preparation and preprocessing, is an established key pre-requisite of successful data mining with its aim to discover and explore something new form the recorded available facts in a specific database. To overcome the hardship, problem and situation, certain statistical methods and techniques are to be employed during the data preparation. There after we can recover its incompleteness or missing data and reduce ambiguities. In this paper, we introduced a method by which missing attributes values are replaced by the best and closest fit values.

Keywords: *Data Mining, Missing Values, Attribute, Data preparation, Incompleteness, Deviation.*

AMS (2000) Subject Classification: *68T30, 68P20.*

1. INTRODUCTION

Missing values in database is one of the biggest problems faced by researcher in data analysis and further mining applications. This missing values problem provoked imbalanced databases. The effects of these missing values are reflected on the final results. Our prime goal is to achieve the final result in the consolidated form on which we are taking decision.

In this study, a statistical method is discussed which provides an approach to find out pattern to recover or generate missing values from a real imbalanced database with massive missing values. Therefore, the objective of this method is to recover or generate the best fitted value for the missing value and select records completely filled for further applications.

The function of statistical methods has gained stuff in exploring estimation and prediction techniques. Wilks[17] is the pioneer statistician, who has considered estimation of parameters of a normal univariate and bivariate population with missing values. Buck [3] suggested estimation of missing values for use with an electronic computer. Kim and Curry [10] considered the treatment of missing data in their analysis. Rubin [13, 14] explored about inference and missing data and multiple imputations for non-response in the survey. Allison [1, 2] investigated estimates of linear models with incomplete data and on missing data. Smyth [16] and Zhang *et. al* [18] have considered that data preparation is a fundamental stage of data analysis. Chen *et. al* [4] studied and discussed about multiple imputation for missing ordinal data. Qin [12] considered the semi-parametric optimization for missing data imputation. Gaur and Dulawat[6,7] discussed various algorithms which are useful for estimation of missing values also gave univariate analysis by using mean value at the place of missing values for data preparation.

Corresponding author: Mukta Agarwal*
Sr. Lecturer Vidya Bhawan Rural Institute Udaipur, India.

Clark *et. al* [5] proposed a simplest method to handle missing attributes values in which they replaced such values by the most common value in the attribute. Kononenko *et al.* [11] suggest that the most common values of the attribute restricted to the concept is used instead of the most common values for all cases. Gyzymala-Busse [8, 9] give idea that every missing attributes values is replaced by all possible known values. Sharma and Gaur [15] extending the study of closest fit method with gradient middling approach. They also provided global closest fit and concept closest fit method for missing attribute values. The objective of proposed study is to determine the statistical technique which may be significant in the handling of missing attribute values in data mining.

2. FORMULATION OF PROBLEM

The proposed method is based on replacing missing attribute values by the estimated values. The method is suitable for numerical attributes. The method is search of closest fit value which is near to the mean of the attribute and closest to the value of just preceding and succeeding value of the missing values.

In the process of estimation of missing value, we first find out the mean of the attribute with missing values case. The sample mean of the attribute is the most important and often used single statistics is defined as the sum of all the sample values divided by the number of observation in the attribute/ sample and is defined as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where $\bar{X}$ is the mean of the observed values and i is the subscript of attribute X. The mean value is basically calculated by the observed values therefore; here we denote this value as

$$\bar{X}_{obs} = \frac{1}{k}\sum_{i=1}^{k} x_i$$

Here k is the total number of observed values in the attribute

The next stage gets involved in the search of missing 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 preceding value (x_p) and succeeding value (x_s) from the missing value subscript (x_i).

$x_p = value(x_i - 1)$
$x_s = value(x_i + 1)$
Where $x_p \neq x_s \; and \; x_p \; or \; x_s \neq NULL$

At the next stage, after recording the value of just preceding value (x_p) and and succeeding value (x_s) of the missing value subscript, we compute the average of both values$(\bar{X}ps)$.

$\bar{X} = (x_p + x_s)/2$

Now at the next stage, we have to check the positional value of missing value, and find out that, missing value position is either greater than or lesser than the mid value.
count (x_i) <=mid

If count (x_i) is less then equal to mid then count the positional value of (x_i), now we compute the deviation of missing value from mid, and then subtract the count (x_i) from mid.

$c_{nt} = count(x_i)$

dv = mid - c_{nt}

The deviation value dv is count by subtracting (c_{nt}) from the mid. After that, we compute the ration deviation according the percentage balance which is decided according to the value in the attribute.

$R_d = d_v * 2$

At the next, we compute percentage deviation (P_d) by subtracting (R_d) from 100.

$P_d = (100 - R_d)$

After recording percentage deviation (P_d) we have computed mean percentage deviation(M_{pd}).

M_{p_d}= (P_d * ($\bar{x}_{obs/100}$))

If count (x_i) is greater then equal to mid then count the (x_i), then we compute the deviation of missing value by help of mid in following manner.

$c_{nt} = count(x_i)$

$d_v = (c_{nt} - mid)$

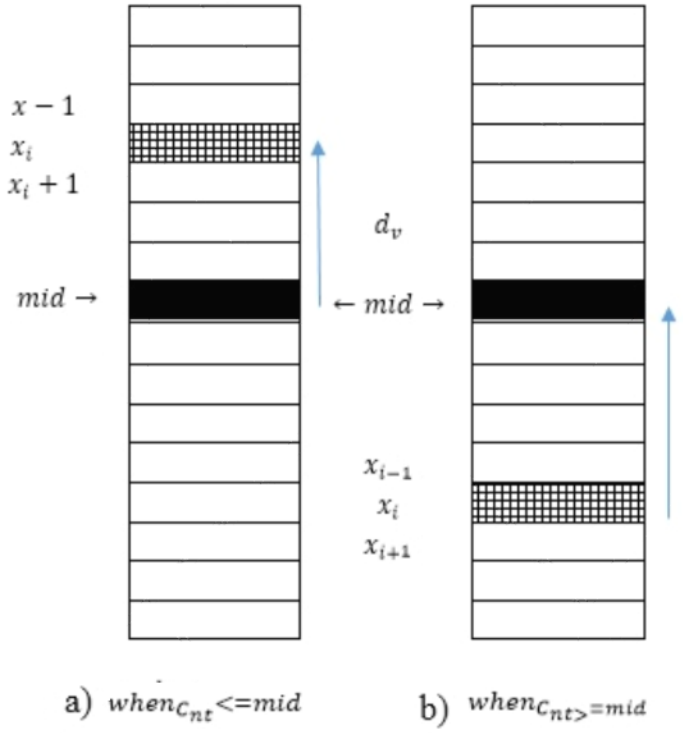
$R_d = (d_v * 2)$

$P_d = (100 + R_d)$

$M_{pd} = (P_d * (\bar{x}_{obs}/100))$

For calculating final estimated value, we get average value of $\bar{x}_{ps}$ and M_{pd}.

x_{est}=(($\bar{X}_{ps} + M_{pd}$)/2)



a) $when_{C_{nt}}$<=mid b) $when_{C_{nt>}=mid}$

3. ALGORITHM

Read X = {x_1,…,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

Calculate $\bar{X}_{obs} = \frac{1}{k}\sum_{i=1}^{k} x_i$ // Calculation of mean of observed values

Read values X = {x_1,…,x_n} // Attribute with observed and missing

For i=1to n, do

If (value (x_i)) == NULL) then
$x_p = value(x_{i-1})$ // Value of preceding of x_i
$x_s = value(x_{i+1})$ // Value of succeeding of x_i
$\bar{x}_{ps} = (x_p + x_s)/2$ // Average of preceding and succeeding
Calculate mid // Calculation of mid value by binary method of data structure

If (count (x_i) <=mid) // Note: - mid is the center value of array.
{
$c_{nt} = count\ (x_i)$ // Position of missing value
$d_v = mid - c_{nt}$ // deviation (distance from mid value)
$R_d = dv * 2$ // Ratio deviation
$P_d = (100 - R_d)$ // Percentage deviation
$M_{pd} = (P_d * \bar{X}_{obs}/100)$ // mean
}
else
{
$c_{nt} = Count(x_i)$
$d_v = (c_{nt} - mid)$
$R_d = (d_v * 2)$
$P_d = (100 + R_d)$
$M_{pd} = (P_d * (\bar{X}_{obs}/100))$
}
$x_{est} = ((\bar{x}_{ps} + M_{pd})/2)$
value (x_i) =x_{est} ,
$i = i + 1$,
repeat Until (i>=n),
Stop.

4. DISCUSSION OF RESULTS

In the Table-A shows the world wide emission of carbon dioxide (co_2) from the consumption of Coal, Oil and Natural Gas respectively for the years 1960 to 2009 are given. The mean emission of carbon dioxide (co_2) due to Coal, Oil and Natural Gas are 2109, 2262 and 879 respectively.

Table-B shows the variables with observed and missing values. Here 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 2101 for Coal, 2231 for Oil and 877 for Natural Gas. It is observed that mean values of incomplete data sets of Table-B are slightly lower than the mean values from all the three variables of Table-A.

The proposed deviation ratio balance method is applied on the data sets of Table-B to fill up the missing values. The outcome of value are shown in Table-C for all three variables which are highlighted by underline. Further, it is observed that the mean values obtained after replacing the missing values by the estimated values in Table-C are quite close to the actual mean as given in Table-A.

5. CONCLUSION

It is quite obvious that, there is no fully perfect technique for handling missing attribute values. The proposed is useful for numerical attribute, having good estimated values near to the mean .This method is suitable for the consolidated report which is generated from the database. Consequently, it is observed that techniques for handling of missing attribute values should be chosen individually or based on the nature and type of data.

Table–A				Table-B				Table-C			
Year	Coal	Oil	Natural Gas	Year	Coal	Oil	Natural Gas	Year	Coal	Oil	Natural Gas
	Million Tons of Carbon				Million Tons of Carbon				Million Tons of Carbon		
1960	1,410	849	235	1960	1,410	849	235	1960	1,410	849	235
1961	1,349	904	254	1961	1,349	904	254	1961	1,349	904	254
1962	1,351	980	277	1962	1,351	980	277	1962	1,351	980	277
1963	1,396	1,052	300	1963		1,052		1963	*1,305*	1,052	405
1964	1,435	1,137	328	1964	1,435	1,137	328	1964	1,435	1,137	328
1965	1,460	1,219	351	1965	1,460	1,219	351	1965	1,460	1,219	351
1966	1,478	1,323	380	1966	1,478	1,323	380	1966	1,478	1,323	380
1967	1,448	1,423	410	1967	1,448		410	1967	1,448	1,454	410
1968	1,448	1,551	446	1968	1,448	1,551		1968	1,448	1,551	514
1969	1,486	1,673	487	1969	1,486	1,673	487	1969	1,486	1,673	487
1970	1,556	1,839	516	1970		1,839	516	1970	1,517	1,839	516
1971	1,559	1,946	554	1971	1,559	1,946	554	1971	1,559	1,946	554
1972	1,576	2,055	583	1972	1,576	2,055	583	1972	1,576	2,055	583
1973	1,581	2,240	608	1973	1,581	2,240	608	1973	1,581	2,240	608
1974	1,579	2,244	618	1974	1,579	2,244		1974	1,579	2,244	659
1975	1,673	2,131	623	1975	1,673	2,131	623	1975	1,673	2,131	623
1976	1,710	2,313	650	1976	1,710	2,313	650	1976	1,710	2,313	650
1977	1,766	2,395	649	1977	1,766		649	1977	1,766	2,135	649
1978	1,793	2,392	677	1978	1,793	2,392	677	1978	1,793	2,392	677
1979	1,887	2,544	719	1979		2,544	719	1979	1,880	2,544	719
1980	1,947	2,422	740	1980	1,947	2,422	740	1980	1,947	2,422	740
1981	1,921	2,289	756	1981	1,921		756	1981	1,921	2,203	756
1982	1,992	2,196	746	1982	1,992	2,196	746	1982	1,992	2,196	746
1983	1,995	2,177	745	1983	1,995	2,177		1983	1,995	2,177	777
1984	2,094	2,202	808	1984		2,202	808	1984	2,116	2,202	808
1985	2,237	2,182	836	1985	2,237	2,182	836	1985	2,237	2,182	836
1986	2,300	2,290	830	1986	2,300		830	1986	2,300	2,242	830
1987	2,364	2,302	893	1987	2,364	2,302	893	1987	2,364	2,302	893
1988	2,414	2,408	936	1988	2,414	2,408	936	1988	2,414	2,408	936
1989	2,457	2,455	972	1989	2,457			1989	2,457	2,454	973
1990	2,409	2,517	1,026	1990	2,409	2,517	1,026	1990	2,409	2,517	1,026
1991	2,341	2,627	1,069	1991		2,627	1,069	1991	2,379	2,627	1,069
1992	2,318	2,506	1,101	1992	2,318	2,506	1,101	1992	2,318	2,506	1,101
1993	2,265	2,537	1,119	1993	2,265	2,537	1,119	1993	2,265	2,537	1,119
1994	2,331	2,562	1,132	1994	2,331	2,562	1,132	1994	2,331	2,562	1,132
1995	2,414	2,586	1,153	1995	2,414			1995	2,414	2,657	1,147
1996	2,451	2,624	1,208	1996		2,624	1,208	1996	2,526	2,624	1,208
1997	2,480	2,707	1,211	1997	2,480	2,707	1,211	1997	2,480	2,707	1,211
1998	2,376	2,763	1,245	1998	2,376	2,763	1,245	1998	2,376	2,763	1,245
1999	2,329	2,716	1,272	1999	2,329	2,716	1,272	1999	2,329	2,716	1,272
2000	2,342	2,831	1,291	2000	2,342	2,831	1,291	2000	2,342	2,831	1,291
2001	2,460	2,842	1,314	2001			1,314	2001	2,614	2,907	1,314
2002	2,487	2,819	1,349	2002	2,487	2,819		2002	2,487	2,819	1,166
2003	2,638	2,928	1,399	2003	2,638	2,928	1,399	2003	2,638	2,928	1,399
2004	2,850	3,032	1,436	2004	2,850	3,032	1,436	2004	2,850	3,032	1,436
2005	3,032	3,079	1,479	2005		3,079	1,479	2005	3,002	3,079	1,479
2006	3,193	3,092	1,527	2006	3,193		1,527	2006	3,193	3,147	1,527
2007	3,295	3,087	1,551	2007	3,295	3,087		2007	3,295	3,087	1,236
2008	3,401	3,079	1,589	2008	3,401	3,079	1,589	2008	3,401	3,079	1,589
2009	3,393	3,019	1,552	2009	3,393	3,019	1,552	2009	3,393	3,019	1,552
Mean	2,109	2,262	879	Mean	2,101	2,231	877	Mean	2,105	2,246	879

Source: www.earth-policy.org

6. REFERENCE

1. Allison, P.D., Estimation of linear models with incomplete data, Social Methodology, San Francisco: Jossey Bass, pp. 71-103, 1987.
2. Allison, P.D., Missing data, Thousand Oaks CA: Sage publication, 2001.
3. 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.
4. 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.
5. Clark, P., and Niblett, T., "The CN2 induction algorithm", Machine Learning, Vol. 3, pp. 261-283, 1983.
6. Gaur, Sanjay and Dulawat, M.S.,"A perception of statistical inference in data mining", International Journal of Computer Science and Communication, Vol. 1, No. 2, pp. 653-658, 2010.
7. 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.
8. Grzymala-Busse, J. W., Grzymala-Busse, W.J., and Goodwin, L. K., "A comparison of three closest fit approaches to missing attribute values in preterm birth data", International Journal of Intelligent System, Vol. 17, pp. 125-134, 2002.
9. 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.
10. Kim, J.O., and Curry, J., "The treatment of missing data in multivariate analysis", Social Methods and Research, Vol. 6, pp. 215-240, 1977.
11. Konoenenko, I., Bratko, I., and Roskar, E., "Experiments in automatic learning of medical diagnostic rules", Technical Report, Jozef Stefan Institute, LIjubl-jana, Yugoslavia, 1984.
12. Qin, Y.S., "Semi-parametric optimization for missing data imputation", Applied Intelligence, Vol. 27, No. 1, pp. 79-88, 2007.
13. Rubin, D.B., "Inference and missing data", Biometrika, 63, pp. 581-592, 1976.
14. Rubin, D.B., Multiple imputations for non-response in surveys, John Wiley and Sons, New York, 1987.
15. Sharma, Swati. and Gaur, Sanjay., "Gradient Middling Approach to Recover Missing Values", Journal of Environmental Science, Computer Science and Engineering and Technology, E-ISSN No. 2278-179X, Vol-2, No.3, 854-858, Aug.-2013.
16. Smyth, 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.
17. Wilks, S.S., "Moment and distribution of estimates of population parameters from fragmentary samples", Annals of mathematical Statistics, Vol. 3, pp. 163-165, 1932.
18. Zhang, S., Zhang, C., and Young, Q., "Data preparation for data mining", Applied Artificial Intelligence, Vol. 17, pp. 375-381, 2003.

Source of support: Nil, Conflict of interest: None Declared