International Journal of Mathematical Archive-3(7), 2012, 2594-2597
IJMA Available online through www.ijma.info ISSN 2229 – 5046

CLOSEST FIT APPROACH TO HANDLE ODD SIZE MISSING BLOCK VALUES

Sanjay Gaur*
Program Coordinator, Faculty of Computer Application, Pacific University, Udaipur, India

(Received on: 29-06-12; Accepted on: 20-07-12)

ABSTRACT

Completeness, quality and real world data preparation is a key pre-requirement for efficient data mining. Database or Table with missing values complicates analysis and data mining. To overcome this situation, certain statistical techniques are required to be employed 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 a method by which odd size missing block values are recovered. Whole study comprises numerical variables of time series data and semi time series data.

Key Words: *Missing Values, Attribute, Data preparation, Incompleteness, Missing Block, Closest fit, Intermediate value.*

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

1. INTRODUCTION

Missing block values in database is solitary of the biggest problems faced in data analysis and in data mining applications. The effects of these missing block values are highly reflected on the final results. Our prime goal is to achieve the final result in the consolidated form on which we are taking decisions. There are various forms of missing values in the database, among those, missing block values case is one of the harder cases to recover, despite the single missing value. In this study, two algorithms of statistical methods are introduced and discussed which provides an approach to find out pattern to recover missing block values from a real world imbalanced database with missing values. Therefore, the objective of this study is to find out closest fit methods to recover missing values and to fill them for further applications.

2. ODD BLOCK FITTING APPROACH

In the proposed method, we first find out the range of block of missing values in the attribute. Here proposed maximum range is 10% of the used dataset. Therefore, maximum three consecutive values may be taken as odd block of missing values.

Now the searches of block missing case in the attribute get start. The first missing value case is pointed by the subscript of the attribute and denoted by the variable (X_i), second and third are denoted from (x_{i+1}) and (x_{i+2}) respectively. Now find average from the values of preceding subscript (X_{i-1}) and succeeding subscript value (X_{i+3}). This average value is replaced at the subscript (x_{i+1}) which is second or centered missing subscript.

x_p = value(x_{i-1})
x_s = value(x_{i+3})

where $x_p \neq x_s$ and $x_p \; or \; x_s \neq \text{NULL}$

value (x_{i+1}) = ($x_p + x_s$) / 2

At the next step calculate average from the values of subscripts (x_{i+1}) and (x_{i-1}), it fill the subscript X_i . Therefore, here the vales of succeeding variable (x_s) get change where preceding (x_p) remain fixed as previous value.

Corresponding author: *Sanjay Gaur**
Program Coordinator, Faculty of Computer Application, Pacific University, Udaipur, India

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

value (x_i) = (x_p + x_s) / 2

Similarly calculate average from the values of subscripts (x_{i+1}) and (x_{i+3}), it fills the subscript X_{i+2}. Thus the equations to fill the value of subscript (x_{i+2}) is formed as:

x_p = value(x_{i+1})
x_s = value(x_{i+3})

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

value (x_{i+2}) = (x_p + x_s) / 2

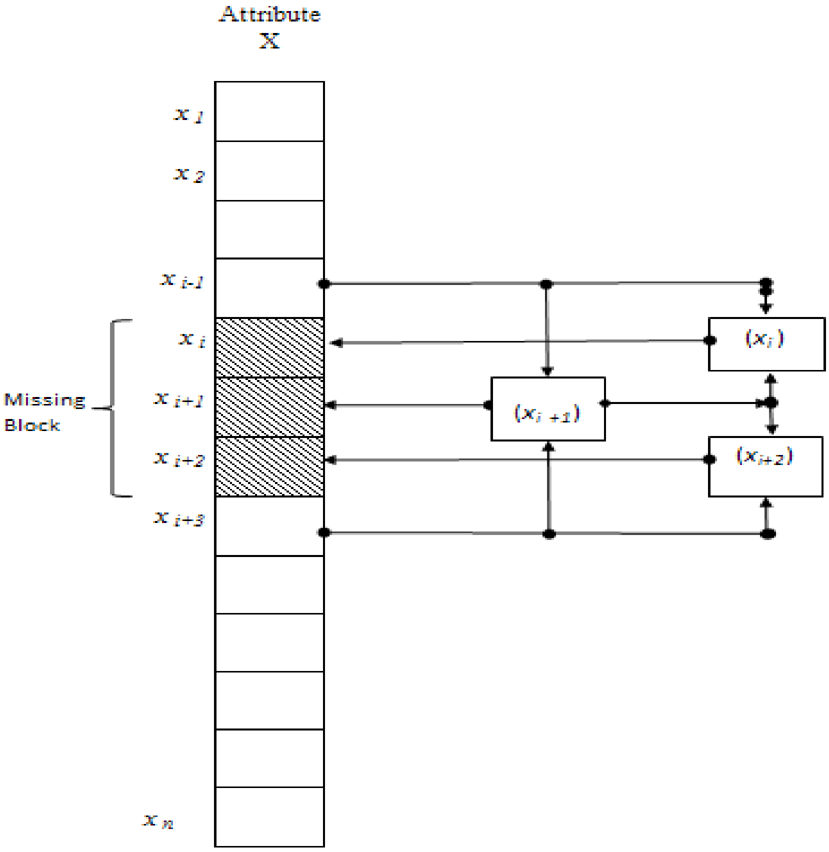

Figure: Block Diagram of Odd Size Block Fitting Approach (Three Values)

3. ALGORITHM

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

For $i = 1$ to n do

If (value (x_i) == NULL && value (x_{i+1})) == NULL&& value (x_{i+2})) == NULL) then

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

x_s = value(x_{i+3}) // Value of succeeding
value (x_{i+1}) = (x_p + x_s) / 2 / Replacing the value for x_{i+1}
(Second (Centered) missed subscript)
x_s = value(x_{i+1})
value (x_i) = (x_p + x_s) / 2 // Replacing the value for xi
(First missed subscript)
x_p = value(x_{i+1})
x_s = value(x_{i+3})
value (x_{i+2}) = (x_p + x_s) / 2 // Replacing the value for x_{i+2}
(Third missed subscript)
endif
i = i + 1
repeat until (i >=n)
Stop

4. DISCUSSION OF RESULTS

Table-1 given in appendix 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. The mean emission of carbon dioxide (CO_2) due to Coal, Oil and Natural Gas are 2109, 2262 and 879 respectively.

It is to be noted that in the planned way odd block of the values are missing in the random manner for all the variables. The means calculated from incomplete data sets are 2097, 2238 for Coal, Oil and 897 for Natural Gas. After recovery of the missing block values the mean of Coal, Oil and Gas are 2107, 2263 and 878 respectively. It is observed that recovered mean values are varying close to means of standard dataset. Same are true for Standard deviation and Coefficient of Variance.

5. CONCLUSION

It is universally known that there is not 100 % efficient technique of handling missing attribute values. The proposed Odd size block fitting approach is useful for numerical attribute, having minor deviation from the mean. The method is appropriate for the consolidated report, also more appropriate and suitable to small size block missing values.

6. 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] Gaur, Sanjay and Dulawat, M.S., Improved Closest fit Techniques to handle missing Attribute values, Journal of Computer and Mathematical Sciences, Vol.-2, No.25, pp. 384-390(2011).

[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).

Table : 1

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

Standard Dataset					Odd Size Block Fitting Approach					
		Standard Data			Missing Values			Recovered Values		
SN	Year	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas
		Million Tons of Carbon			Million Tons of Carbon			Million Tons of Carbon		
1	1960	1,410	849	235	1,410	849	235	1,410	849	235
2	1961	1,349	904	254	1,349	904	254	1,349	904	254
3	1962	1,351	980	277	1,351	980	277	1,351	980	277
4	1963	1,396	1,052	300	1,396	1,052	300	1,396	1,052	300
5	1964	1,435	1,137	328	1,435	1,137	328	1,435	1,137	328
6	1965	1,460	1,219	351	1,460	1,219	351	1,460	1,219	351
7	1966	1,478	1,323	380	1,478	1,323	380	1,478	1,323	380
8	1967	1,448	1,423	410	1,448	1,423	410	1,448	1,423	410
9	1968	1,448	1,551	446	1,448	1,551	446	1,448	1,551	446
10	1969	1,486	1,673	487	1,486	1,673	487	1,486	1,673	487
11	1970	1,556	1,839	516	1,556	1,839	516	1,556	1,839	516
12	1971	1,559	1,946	554	1,559	1,946	554	1,559	1,946	554
13	1972	1,576	2,055	583	1,576	2,055		1,576	2,055	***571***
14	1973	1,581	2,240	608	1,581	2,240		1,581	2,240	***589***
15	1974	1,579	2,244	618	1,579	2,244		1,579	2,244	***606***
16	1975	1,673	2,131	623	1,673	2,131	623	1,673	2,131	623
17	1976	1,710	2,313	650	1,710	2,313	650	1,710	2,313	650
18	1977	1,766	2,395	649	1,766	2,395	649	1,766	2,395	649
19	1978	1,793	2,392	677	1,793	2,392	677	1,793	2,392	677
20	1979	1,887	2,544	719	1,887	2,544	719	1,887	2,544	719
21	1980	1,947	2,422	740	1,947	2,422	740	1,947	2,422	740
22	1981	1,921	2,289	756	1,921	2,289	756	1,921	2,289	756
23	1982	1,992	2,196	746	1,992	2,196	746	1,992	2,196	746
24	1983	1,995	2,177	745	1,995	2,177	745	1,995	2,177	745
25	1984	2,094	2,202	808	2,094	2,202	808	2,094	2,202	808
26	1985	2,237	2,182	836		2,182	836	***2,174***	2,182	836
27	1986	2,300	2,290	830		2,290	830	***2,254***	2,290	830
28	1987	2,364	2,302	893		2,302	893	***2,334***	2,302	893
29	1988	2,414	2,408	936	2,414	2,408	936	2,414	2,408	936
30	1989	2,457	2,455	972	2,457	2,455	972	2,457	2,455	972
31	1990	2,409	2,517	1,026	2,409	2,517	1,026	2,409	2,517	1,026
32	1991	2,341	2,627	1,069	2,341	2,627	1,069	2,341	2,627	1,069
33	1992	2,318	2,506	1,101	2,318	2,506	1,101	2,318	2,506	1,101
34	1993	2,265	2,537	1,119	2,265	2,537	1,119	2,265	2,537	1,119
35	1994	2,331	2,562	1,132	2,331	2,562	1,132	2,331	2,562	1,132
36	1995	2,414	2,586	1,153	2,414		1,153	2,414	***2,612***	1,153
37	1996	2,451	2,624	1,208	2,451		1,208	2,451	***2,663***	1,208
38	1997	2,480	2,707	1,211	2,480		1,211	2,480	***2,713***	1,211
39	1998	2,376	2,763	1,245	2,376	2,763	1,245	2,376	2,763	1,245
40	1999	2,329	2,716	1,272	2,329	2,716	1,272	2,329	2,716	1,272
41	2000	2,342	2,831	1,291	2,342	2,831	1,291	2,342	2,831	1,291
42	2001	2,460	2,842	1,314	2,460	2,842	1,314	2,460	2,842	1,314
43	2002	2,487	2,819	1,349	2,487	2,819	1,349	2,487	2,819	1,349
44	2003	2,638	2,928	1,399	2,638	2,928	1,399	2,638	2,928	1,399
45	2004	2,850	3,032	1,436	2,850	3,032	1,436	2,850	3,032	1,436
46	2005	3,032	3,079	1,479	3,032	3,079	1,479	3,032	3,079	1,479
47	2006	3,193	3,092	1,527	3,193	3,092	1,527	3,193	3,092	1,527
48	2007	3,295	3,087	1,551	3,295	3,087	1,551	3,295	3,087	1,551
49	2008	3,401	3,079	1,589	3,401	3,079	1,589	3,401	3,079	1,589
50	2009	3,393	3,019	1,552	3,393	3,019	1,552	3,393	3,019	1,552
	Mean	2,109	2,262	879	2,097	2,238	897	2,107	2,263	878
	SD	567.87	621.14	400.26	583.78	633.19	406.64	567.11	622.00	400.88
	CV	26.92	27.46	45.54	27.84	28.30	45.35	26.92	27.48	45.65

source: www.earth_policy.org

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

