ISSN No. 0976-5697

Volume 4, No. 11, Nov-Dec 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining

Ms. Swati Sharma
Research Scholar, Pacific Institute of Technology,
Udaipur, India
swati.sharma89@yahoo.com

Dr. Sanjay Gaur
Asst. Professor & coordinator, FCA
Pacific University, Udaipur, India
sanjay.since@gmail.com

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.

Keywords: Missing Values, Attribute, Data preparation, Incompleteness, Missing Block, Contiguous Athletic, Intermediate value.

I. 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 data base, among those, missing block values case is one of the harder cases to recover, despite the single missing value. In this study, an algorithm is introduced and discussed which provides an approach to find out pattern to recover odd size missing block values from a real world imbalanced database. Therefore, the objective of this study is to find out contiguous agile methods to recover values for odd size missing block and to fill them for further applications.

II. CONTIGUOUS AGILE APPROACH TO ODD SIZE BLOCK (FIVE VALUES)

In the proposed method, we first find out the range of block of missing values in the attribute. Here proposed maximum range is approx 20% of the used dataset. Therefore, maximum five consecutive values may be taken as odd block of missing values.

Now find the intermediately values of preceding subscript (x_{i-1}) and succeeding subscript value (x_{i-5}). This average value is temporarily hold in the variable (x_{i+2}).

$X_p = Value\,(x_{i-1})$

$X_s = Value\,(x_{i+5})$

Where $x_p \neq x_s$, x_p and $x_s \neq$ NULL

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

Now find out value of preceding temporarily variable between the variable (x_{i+2}) and value of preceding subscript (x_{i-1}). The average value is temporarily held in the intermediately variable ($x_{intr\ p}$).

$X_p = Value\,(x_{i-1})$

$X_s = Value\,(x_{i+2})$

Where $x_p \neq x_s$, x_p and $x_s \neq$ NULL

Value ($x_{intr\ p}$) = (x_p +x_s) /2

At the next stage, find out the value of succeeding temporarily variable (x_{i+2}) and values of succeeding subscript (x_{i+5}). The average value is temporarily held in the intermediately variable ($x_{intr\ s}$).

$X_p = Value\,(x_{i+2})$

$X_s = Value\,(x_{i+5})$

Where $x_p \neq x_s$, x_p and $x_s \neq$ NULL

Value ($x_{intr\ s}$) = (x_p +x_s) /2

Now, with the help of value of preceding subscript (x_{i-1}) and value of temporary preceding variable ($x_{intr\ p}$) , we can find out the value for first missing subscript (x_i). For that, we have to take the average of value of subscript (x_{i-1}) and value of temporary preceding variable ($x_{intr\ p}$) and replace final value in the subscript (x_i).

Value (x_i) = [Value (x_{i-1}) + Value ($x_{intr\ p}$)] /2

Value (x_{i+1}) = [Value (x_{i+2}) + Value ($x_{intr\ p}$)] /2

Value (x_{i+2}) = [Value (x_{i-1}) + Value (x_{i+5})] /2

Value (x_{i+3}) = [Value (x_{i-2}) + Value ($x_{intr\ s}$)] /2

Value (x_{i+4}) = [Value (x_s) + Value ($x_{intr\ s}$)] /2

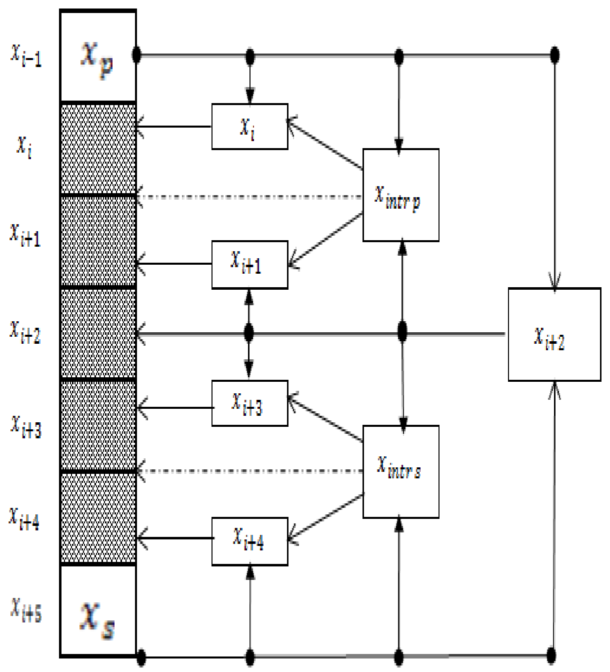

Fig: Block Diagram of contiguous agile approach to recover odd size missing block

III. ALGORITHM

Read X = {x_1, x_2, ____, x_n} // Attribute with observed and missing values

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

X_{obs} = {x_1 , ____, x_k} // Attribute values observed

X_{mis} = {x_{k+1}, ______, 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 && Value (x_{i+3}) == NULL && Value (x_{i+4}) then // Finding the central value(x_{i+2})

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

x_s = Value(x_{i+5}) // Value of succeeding of missing block

// Finding the value between five missing value cases as a central value

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

// Finding value of preceding intermediate ($x_{intr\ p}$) using the intermediate variable value.

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

x_s = Value(x_{i+2}) // Value of central intermediate variable

// Value between (x_{i+2}) and preceding

value ($x_{intr\ p}$) = (x_p + x_s) /2

// Finding value of succeeding intermediate ($x_{intr\ s}$) using the intermediate variable value

x_p = Value(x_{i+2}) // Value of central intermediate variable

x_s = Value(x_{i+5}) // Value of succeeding of block

Value ($x_{intr\ s}$) = (x_p + x_s) /2

// Intermediate value between (x_{i+2}) and preceding

// Finding value of (x_i), first missing value subscript

Value (x_i) = [Value (x_{i-1}) + Value ($x_{intr\ p}$)] /2

// Replacing the value (x_i)

// Finding value of (x_{i+1}), second missing value subscript

Value (x_{i+1}) = [Value ($x_{intr\ p}$) + Value (x_{i+2})] /2

// Replacing the value (x_{i+1})

// Finding value of (x_{i+2}), third missing value subscript

Value (x_{i+2}) = [Value (x_{i-1}) + Value (x_{i+5})] /2

// Replacing the value (x_{i+2})

// Finding value of (x_{i+3}), fourth missing value subscript

Value (x_{i+3}) = [Value (x_{i+2}) + Value ($x_{intr\,s}$)] /2
// Replacing the value (x_{i+3})

// Finding value of (x_{i+4}), Fifth missing value subscript

Value (x_{i+4}) = [Value ($x_{intr\,s}$) + Value (x_{i+5})] /2
// Replacing the value (x_{i+4})

Endif
i = i + 1
repeat until (i >=n)
Stop

IV. DISCUSSION OF RESULTS

Table-A given in appendix shows the Beverage Consumption in the United States, 1980-2005. The mean of Beverage consumption in the United States due to Tea, Milk, and coffee are 1961, 6322 and 6474 respectively.
It is observed that mean values of incomplete data sets of Table-B are deviated from original mean values both the variables of Table-A.
The proposed contiguous agile method is applied on the data sets of Table-B to fill up the missing values. These contiguous middling values are shown in Table-C for all the variables which are highlighted. It is observed that contiguous middling values of Tea, Milk and Coffee are 1975, 6326 and 6456 respectively. It is considerable that the middling values obtained after replacing the missing values by the proposed contiguous agile values in Table-C are 99% close to the actual mean value as given in Table-A. It is observed that recovered mean values are varying close to means of standard dataset. Same may true for Standard deviation and Coefficient of Variance

V. CONCLUSION

It is universally known that there is not 100 % efficient technique of managing missing block attribute values. The proposed contiguous agile approach for missing block values 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.

VI. 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: Bevarage Consumption in USA 1980-2005

Beverage Consumption in the United States, 1980-2005

Table A			
Original Value			
Year	Tea	Milk	Coffee
Million Gallons			
1980	1,665	6,263	6,076
1981	1,656	6,220	5,972
1982	1,609	6,108	6,009
1983	1,628	6,146	6,150
1984	1,674	6,220	6,312
1985	1,684	6,342	6,523
1986	1,703	6,370	6,599
1987	1,682	6,333	6,479
1988	1,704	6,375	6,263
1989	1,694	6,429	6,475
1990	1,717	6,412	6,700
1991	1,862	6,442	6,762
1992	2,064	6,433	6,612
1993	2,159	6,354	6,045
1994	2,142	6,386	5,486
1995	2,107	6,368	5,391
1996	2,057	6,415	5,962
1997	1,977	6,381	6,351
1998	2,278	6,348	6,581
1999	2,300	6,391	7,010
2000	2,212	6,341	7,415
2001	2,327	6,278	6,899
2002	2,238	6,304	6,805
2003	2,169	6,269	7,049
2004	2,331	6,234	7,228
2005	2,354	6,218	7,173
Mean	1,961	6,322	6,474

Table B			
Missing Values(20% approx)			
Year	Tea	Milk	Coffee
Million Gallons			
1980	1,665	6,263	6,076
1981	1,656	6,220	5,972
1982	1,609	6,108	
1983	1,628	6,146	
1984	1,674	6,220	
1985	1,684	6,342	
1986		6,370	
1987		6,333	6,479
1988		6,375	6,263
1989		6,429	6,475
1990		6,412	6,700
1991	1,862	6,442	6,762
1992	2,064		6,612
1993	2,159		6,045
1994	2,142		5,486
1995	2,107		5,391
1996	2,057		5,962
1997	1,977	6,381	6,351
1998	2,278	6,348	6,581
1999	2,300	6,391	7,010
2000	2,212	6,341	7,415
2001	2,327	6,278	6,899
2002	2,238	6,304	6,805
2003	2,169	6,269	7,049
2004	2,331	6,234	7,228
2005	2,354	6,218	7,173
	1,634	5,093	5,259

Table C			
Table With Estimated Values			
Year	Tea	Milk	Coffee
Million Gallons			
1980	1,665	6,263	6,076
1981	1,656	6,220	5,972
1982	1,609	6,108	6,035
1983	1,628	6,146	6,162
1984	1,674	6,220	6,225
1985	1,684	6,342	6,289
1986	1,706	6,370	6,416
1987	1,751	6,333	6,479
1988	1,773	6,375	6,263
1989	1,795	6,429	6,475
1990	1,840	6,412	6,700
1991	1,862	6,442	6,762
1992	2,064	6,434	6,612
1993	2,159	6,419	6,045
1994	2,142	6,411	5,486
1995	2,107	6,404	5,391
1996	2,057	6,389	5,962
1997	1,977	6,381	6,351
1998	2,278	6,348	6,581
1999	2,300	6,391	7,010
2000	2,212	6,341	7,415
2001	2,327	6,278	6,899
2002	2,238	6,304	6,805
2003	2,169	6,269	7,049
2004	2,331	6,234	7,228
2005	2,354	6,218	7,173
	1,975	6,326	6,456