International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT22823

An Applied Secant Method for Recovered Missing Mass Values in Data Mining

Dr. Darshanaben Dipakkumar Pandya[1], Dr. Abhijeetsinh Jadeja[2], Dr. Sheshang D. Degadwala[3]

[1]Assistant Professor, Department of Computer Science, Shri C.J Patel College of Computer Studies (BCA), Visnagar, Gujarat, India

[2]Principal(I/C), Department of Computer Science, Shri C.J Patel College of Computer Studies (BCA), Visnagar, Gujarat, India

[3]Head of Computer Department, Sigma Institute of Engineering, Vadodara, Gujarat, India

Article Info

Volume 8, Issue 2

Page Number : 97-103

Publication Issue :

March-April-2022

Article History

Accepted: 01 April 2022

Published: 05 April 2022

ABSTRACT

In data mining, the preparation of complete, quality and real data is a key prerequisite for successful data mining in order to discover something new from data already recorded in a given database. Data preparation for data extraction is a fundamental step in data analysis. Data with missing values complicate both data analysis and application of a new data solution. To overcome this situation, some Numerical techniques must be used during data preparation. With the help of Numerical and technical methods, we can retrieve the incomplete state of missing data in huge sequential values and reduce ambiguities using an applied Secant method. In this article, we present a sequential method by which the values of the missing attribute are replaced by the best adapted value.

Keywords : Data Mining, Missing Sequential Bulk Values, An Applied Secant Method.

I. INTRODUCTION

In database, the missing bulk values are solitary of the major problems faced by data analysis and data mining applications. The effects of these missing bulk values are reflected very much in the final results. Our main goal is to reach the final result in the consolidated form in which we make decisions. There are several forms of missing values in the database, among these, missing bulk values are one of the most difficult to recover cases, despite the only missing value. In this study, one Numerical methods algorithm are introduced and discussed that provide an approach to find models to retrieve missing bulk values from an unbalanced real database with missing values. Therefore, the objective of this study is to uncover missing sequential mass values to recover lost values using the linear approach method and fill them for additional applications.

II. An applied Secant method

The proposed method is based on the replacement of the missing bulk attribute values for the values generated by the linear approximation using secant method. This method is very useful for numeric attributes. In general, this method is the search for a sequential missing bulk value that is very close to the real mean of the attribute and closer to the value than the original value of missing values.

In the process of generating sequential missing mass values for the lost volume value, we first find the first predecessor value and the second predecessor value of the case with missing values. In numerical analysis, an applied secant to find successive approximations to the function. This method is implemented as follows :

The method starts with a function f defined over the real numbers x and an initial value of first Predecessor for missing bulk PredX0 and $PredX_1$ is the second predecessor of the missing bulk..

So, $PredX_0 = K[I]-2$ (First Predecessor value of missing Bulk) (2.1)

And $PredX_1 = K[I]-1$ (second Predecessor value of missing Bulk (2.2)

for a root of the function *f*. If the function satisfies the assumptions made in the formula and the initial guess is close, then a next better approximation is

$$K[x_{i+1}] = f(xi) - K[x_i] * \frac{K[x_i] - K[x_{i-1}]}{f(x_i) - f(x_{i-1})}$$ (Assign estimated value) (2.3)

The process is repeated as

$$K[x_{n+1}] = f(x_n) - K[x_n] * \frac{K[x_n] - K[x_{n-1}]}{f(x_n) - f(x_{n-1})}$$ (Assign estimated value) (2.4)

until a sufficiently accurate value is reached and that $K[I] \neq 0$. where *f* denotes the function *f*. here function ,

$$f(x) = 100x - 2$$ (2.5)

Solving for x_{n+1} gives Now at the next stage, the predecessor value of the missing bulk values considered as $PredX_0$ for the first iteration values and $PredX_1$ second Predecessor value of missing Bulk estimated value may be obtained. This procedure is repeated until all missing bulk values recovered.

III. An applied Secant method algorithm

The intended method is based on replacing missing attribute values by an applied Newton Raphson method. This method is very much helpful for numerical attributes. In general, this method is search of missing values and after searching its value is replaced by recovered value of the attribute in sequential missing bulk.

Introduction: Given an array K of size N, this procedure replaces the missing values with the recovered data from data set. Here X_0 is the first predecessor and X_1 is the second predecessor of the missing bulk. F(x)= 100x-2 is the initial function. The variable I is used to index elements from 1 to N in a given data. Following are the steps of the algorithm in detail:

Step 1: Select a dataset on which Missing values recovery is to be performed from the database.

Step 2: Initialize

X_0 , X_1, N , I ← NULL.

Step 3: Create a loop for N passes

Repeat through step 8 for I = 1, 2… N.

Step 4: Perform Missing value Recovery Process from database.

do

If (K [I] = = NULL) then

$K[PredX_0] = K[I] - 2$ //first Predecessor value of missing Bulk

$K[PredX_1] = K[I] - 1$ // second Predecessor value of missing Bulk

Step 5: Initialize the function

$f(x) = 100x - 2$ // Initialize the function

Step 6: Apply secant formula

$$K[x_{i+1}] = f(xi) - K[x_i] * \frac{K[x_i] - K[x_{i-1}]}{f(x_i) - f(x_{i-1})}$$ // Assign estimated value.

Step 7: Make iterations of each pass.

I = I +1. // Iterations

Step 8: Iteration is to be performed till condition is satisfied.

Repeat until (K[I] ≠ NULL)

Step 9: Finished.

Stop.

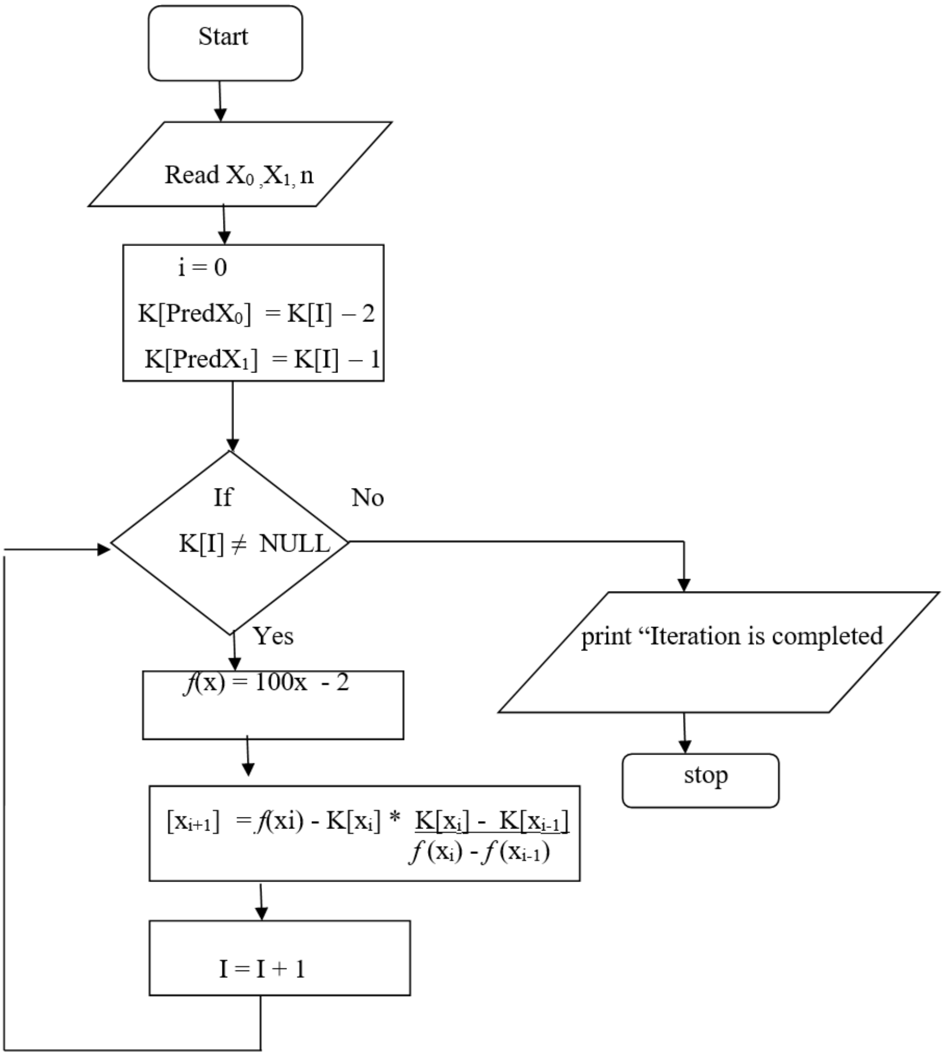

Figure 1 : Flow chart of an applied secant method

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 1960-2009. The mean of global carbon dioxide emissions due to coal, oil and natural gas are 2109, 2262 and 879 respectively. After missing values at the extremes, the mean calculated from incomplete data sets are 2,076 for coal, 2,254 for oil and 908 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 2,095, 2,243 and 871 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.

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

H0: μ1= μ2= μ3 against the alternative

H1: 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)

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	26467.43	2	13233.71	0.040381	0.960434	3.059831
Within Groups	46536284	142	327720.3			
Total	46562752	144				

Table 1 Value: - F (2, 142) at 5% Level of Significance = 3.0718, 1% Level of Significance = 4.7865,

One Way ANOVA (OIL)

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	9257.188	2	4628.594	0.011604	0.988464	3.059831
Within Groups	56639469	142	398869.5			
Total	56648726	144				

Table 2 Value :- F(2, 142) at 5% Level of Significance = 3.0718 , 1% Level of Significance = 4.7865,

One Way ANOVA (NATURAL GAS)

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	35533.76	2	17766.88	0.107829	0.897855	3.059831
Within Groups	23397227	142	164769.2			
Total	23432761	144				

Table 3 Value :- F(2, 142) at 5% Level of Significance = 3.0718 , 1% Level of Significance = 4.7865,

Decision and Conclusion: Since F (Calculated) < 3.0781 so accept H0 at 5% level of significance and

Hence conclude that there is no significant difference among groups of Coal, Oil and Gas regarding Mean value.

Table-4. Table for An applied Secant method for five missing bulk of data (five missing)

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

Standard Data					Missing Values			Recovered Values		
S. N	YEAR	COAL	OIL	NATURL 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	1349	904	254	1349	904	254	1349	904	254
3	1962	1351	980	277	1351	980	277	1351	980	277
4	1963	1396	1,052	300	1396	1,052	300	1396	1,052	300
5	1964	1435	1,137	328	1435	1,137	328	1435	1,137	328
6	1965	1460	1,219	351	1460	1,219	351	1460	1,219	351
7	1966	1478	1,323	380	1478	1,323	380	1478	1,323	380
8	1967	1448	1,423	410	1448	1,423	410	1448	1,423	410
9	1968	1448	1,551	446	1448	1,551	446	1448	1,551	446
10	1969	1486	1,673	487	1486	1,673	487	1486	1,673	487
11	1970	1556	1,839	516	1556	1,839	516	1556	1,839	516
12	1971	1559	1,946	554	1559	1,946	554	1559	1,946	554
13	1972	1576	2,055	583	1576	2,055	____	1576	2,055	**548**
14	1973	1581	2,240	608	1581	2,240	____	1581	2,240	**543**
15	1974	1579	2,244	618	1579	2,244	____	1579	2,244	**538**
16	1975	1673	2,131	623	1673	2,131	____	1673	2,131	**533**
17	1976	1710	2,313	650	1710	2,313	____	1710	2,313	**528**
18	1977	1766	2,395	649	1766	2,395	649	1766	2,395	649
19	1978	1793	2,392	677	1793	2,392	677	1793	2,392	677
20	1979	1887	2,544	719	1887	2,544	719	1887	2,544	719
21	1980	1947	2,422	740	1947	2,422	740	1947	2,422	740
22	1981	1921	2,289	756	1921	2,289	756	1921	2,289	756
23	1982	1992	2,196	746	1992	2,196	746	1992	2,196	746

24	**1983**	1995	2,177	745	1995	2,177	745	1995	2,177	745
25	**1984**	2094	2,202	808	2094	2,202	808	2094	2,202	808
26	**1985**	2237	2,182	836	2237	____	836	2237	2,180	836
27	**1986**	2300	2,290	830	2300	____	830	2300	2,158	830
28	**1987**	2364	2,302	893	2364	____	893	2364	2,136	893
29	**1988**	2414	2,408	936	2414	____	936	2414	2,115	936
30	**1989**	2457	2,455	972	2457	____	972	2457	2,094	972
31	**1990**	2409	2,517	1,026	2409	2,517	1,026	2409	2,517	1,026
32	**1991**	2341	2,627	1,069	2341	2,627	1,069	2341	2,627	1,069
33	**1992**	2318	2,506	1,101	2318	2,506	1,101	2318	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	____	2,586	1,153	2,308	2,586	1,153
37	**1996**	2,451	2,624	1,208	____	2,624	1,208	2,285	2,624	1,208
38	**1997**	2,480	2,707	1,211	____	2,707	1,211	2,262	2,707	1,211
39	**1998**	2,376	2,763	1,245	____	2,763	1,245	2,239	2,763	1,245
40	**1999**	2,329	2,716	1,272	____	2,716	1,272	2,217	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	2,076	2,254	908	2,095	2,243	871
	S.D	567.89	621.13	400.27	589.42	654.25	411.90	561.48	621.08	406.12
	C.V	0.27	0.27	0.46	0.28	0.29	0.45	0.27	0.28	0.47

V. CONCLUSION

In data mining, it is universally known that there is no 100% efficient techniques to handle missing mass values . This document shows the universal truth that there is no precise method of treatment for missing mass values. The proposed approach is important for the real arithmetic value of the particular function used. This approach provides the appropriate result for the consolidated report generated by the database. As a result, it is observed that the techniques for managing the missing bulk attribute values in the missing sequential volume must be adjusted according to the environment and data type. The method is appropriate for the consolidated report and is appropriate for the small size bulk values.

VI. REFERENCES

[1]. Gaur, Sanjay and Dulawat, M.S., closer to the lack of attribute principals of mining approach, International Journal of Advances in Science and Technology, Vol-2, Number-4, (2011).

[2]. S. Ramaswamy, R. Rastogi and K. Shim, "Efficient algorithms for outlining large anomalous values Dataset."In Proceedings of the ACM SIGMOD 2000 International Conference on the management of Data,

volume 29, number 2, pages 427-438, May 2000.

[3]. C. Lu, Chen D., Y. Kou, "Algorithms for the recognition of anomalous spatial values" in Acts of the 3rd IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL 2003.

[4]. Buck, S.F., a lost evaluation method suitable for use with an electronic calculator, J. Royal Statistical Society, Series B, Vol-2, multivariate data values pp. 302-306 (1960).

[5]. Sharma, Swati and Gaur, Sanjay, agile contiguous approach to handle strange bulk format that is missing on data mining, "International Journal of Advanced Research in Computer Science, Vol. 4 (11), pp. 214-217 (2013).

Cite this article as :

Dr. Darshanaben Dipakkumar Pandya, Dr. Abhijeetsinh Jadeja, Dr. Sheshang D. Degadwala, "An Applied Secant Method for Recovered Missing Mass Values in Data Mining ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.97-103, March-April-2022. Available at doi : https://doi.org/10.32628/CSEIT22823
Journal URL : https://ijsrcseit.com/CSEIT22823