Efficient Missing Data Recovery with Closet Fit: A Scalable Solution for Large-Scale Data Mining

Nidhi S Bhavsar[1], Khushbu Yadav[2], Nita Goswami[3]

Research Scholar, Madhav University, Pindwara, Sirohi, Rajasthan,India[1]
Assistant Professor, Madhav University, Pindwara, Sirohi, Rajasthan,India[2]
Assistant Professor, Monark University, Vahelal, Ahmedabad, Gujarat,India[3]

nidhimscit2011@gmail.com[1], khushbuyadav289@yahoo.com[2], nita.goswami.foet@monarkuni.ac.in[3]

Abstract: Data preparation is a crucial step in data analysis, serving as the foundation for successful data mining. To uncover novel insights from existing databases, it is essential to ensure data completeness, quality, and real-world relevance. However, missing values can hinder analysis and application to new data, necessitating the employment of statistical techniques during data preparation. By leveraging statistical methods, we can address data incompleteness and ambiguity. This paper presents two sequential approaches for imputing missing attribute values, focusing on numerical variables in time series data using the moving average method. A comparative study of both methods is provided, highlighting their effectiveness in recovering missing data.

Keywords: Moving average, chronological, incompleteness, missing values, attribute, and data preparation

---

I. INTRODUCTION

Databases often store information and data in a tabular style. Data sets are essentially the properties of the connected table, whereas records sets are the table's rows. The dataset includes essential information needed for sophisticated reports and queries. The incompleteness or missing values in the dataset directly affect the final reporting. Recognizing and retrieving arbitrarily missing variables remains a critical problem in data mining today. Missing values affect the outcome and are a continuous source of uncertainty. It reduces query accuracy and the ability of authorities to make decisions. It is critical to identify such crises before they impair report preparation and query.

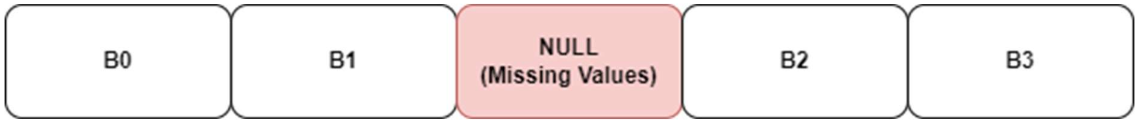
Missing data is a pervasive issue in data mining, hindering the accuracy and reliability of analytical models. Traditional imputation methods often fall short, leading to biased or inaccurate results. To address this challenge, we propose the Closet Fit Algorithm (CFA), a novel approach to recovering missing data. CFA leverages the concept of similarity measures to identify the closest fit for missing values, iteratively refining its estimates to ensure optimal results. By adapting to diverse data distributions and missing value patterns, CFA offers a robust and effective solution for data miners. This paper presents the Closet Fit Algorithm, its methodology, and experimental results demonstrating its superiority over existing imputation techniques.

II. PROPOSED ALGORITHM

This section presents a straightforward numerical approach for approximating missing values in a dataset. We employ the closest fit strategy to recover missing data. First, we identify the attribute elements with missing values. We then logically divide the attribute into two halves: one with missing values and the other with observed values. We focus on finding missing values in the attribute, using two variables, A (year) and B (data set value), which are proportional. Variable A remains constant for other characteristics with missing values, while variable B has varying attributes and random missing values. Notably, variable A has no missing values and serves as the corresponding variable for B.

https://doi.org/10.63766/spujstmr.24.000025

ISSN: 3049-1479 (Online)



$B_0 = B_i - 2$ $B_1 = B_i - 1$ B_i $B_2 = B_i + 1$ $B_3 = B_i + 2$

Fig.1 Show NULL values in Dataset

Let's break down each step of the algorithm to evaluate the expression:

Step 1: Start:

Begin evaluating the expression from the innermost parentheses.

Step 2. Evaluate (B0 + B3):

Add the values of B0 and B3.

Step 3. Evaluate (B0 * 2):

Multiply the value of B0 by 2.

Step 4. Evaluate (B1 + B2):

Add the values of B1 and B2.

Step 5. Evaluate 3(B1 + B2):

Multiply the result from step 4 by 3.

Step 6. Add (B0 + B3) and (B0 * 2):

Add the results from steps 2 and 3.

Step 7. Add 3(B1 + B2) to the result:

Add the result from step 5 to the result from step 6.

Step 8. Add the results from steps 2, 3, and 5:

Combine the results from steps 2, 3, and 5.

Step 9. Divide by 10:

Divide the final result by 10.

Step 10. End:

The final result is the evaluated expression.

ISSN: 3049-1479 (Online)

TABLE I

A CLOSET FIT ALGORITHM

Population increase 1950 -2022							
A Closet fit Algorithm to Recover Missing Data							
	Actual DataSet			Missing DataSet			Recovered DataSet	
Sr.No.	Year (A)	Million People (B)		Year (A)	Missing Data (B)		Year (A)	Recovered Data (B)
1	1950	16.9		1950	16.9		1950	16.9
2	1951	17.3		1951	17.3		1951	17.3
3	1952	17.7		1952	17.7		1952	17.7
4	1953	18.2		1953			1953	18.0
5	1954	18.6		1954	18.6		1954	18.6
6	1955	19.1		1955	19.1		1955	19.1
7	1956	19.6		1956	19.6		1956	19.6
8	1957	20.1		1957	20.1		1957	20.1
9	1958	20.6		1958			1958	20.40308
10	1959	21.1		1959	21.1		1959	21.1
11	1960	21.7		1960	21.7		1960	21.7
12	1961	22.3		1961	22.3		1961	22.3
13	1962	22.9		1962	22.9		1962	22.9
14	1963	23.5		1963	23.5		1963	23.5
15	1964	24.2		1964	24.2		1964	24.2
16	1965	24.9		1965			1965	24.66231
17	1966	25.6		1966	25.6		1966	25.6
18	1967	26.4		1967	26.4		1967	26.4
19	1968	27.2		1968	27.2		1968	27.2
20	1969	28.0		1969	28.0		1969	28.0
21	1970	28.8		1970	28.8		1970	28.8
22	1971	29.7		1971	29.7		1971	29.7
23	1972	30.6		1972	30.6		1972	30.6
24	1973	31.5		1973	31.5		1973	31.5
25	1974	32.5		1974	32.5		1974	32.5
26	1975	33.5		1975	33.5		1975	33.5
27	1976	34.4		1976	34.4		1976	34.4
28	1977	35.4		1977	35.4		1977	35.4
29	1978	36.5		1978			1978	36.17176
30	1979	37.7		1979	37.7		1979	37.7
31	1980	39.1		1980	39.1		1980	39.1
32	1981	40.8		1981	40.8		1981	40.8
33	1982	42.6		1982	42.6		1982	42.6
34	1983	44.6		1983	44.6		1983	44.6
35	1984	46.6		1984	46.6		1984	46.6
36	1985	48.7		1985			1985	47.86945
37	1986	50.8		1986	50.8		1986	50.8
38	1987	52.8		1987	52.8		1987	52.8
39	1988	54.8		1988	54.8		1988	54.8
40	1989	56.7		1989	56.7		1989	56.7

41	1990	58.4		1990	58.4		1990	58.4
42	1991	59.9		1991	59.9		1991	59.9
43	1992	61.2		1992	61.2		1992	61.2
44	1993	62.4		1993	62.4		1993	62.4
45	1994	63.5		1994	63.5		1994	63.5
46	1995	64.6		1995	64.6		1995	64.6
47	1996	65.8		1996	65.8		1996	65.8
48	1997	66.9		1997	66.9		1997	66.9
49	1998	68.1		1998			1998	67.62396
50	1999	69.2		1999	69.2		1999	69.2
51	2000	70.3		2000	70.3		2000	70.3
52	2001	71.4		2001	71.4		2001	71.4
53	2002	72.4		2002	72.4		2002	72.4
54	2003	73.4		2003	73.4		2003	73.4
55	2004	74.4		2004	74.4		2004	74.4
56	2005	75.4		2005	75.4		2005	75.4
57	2006	76.4		2006	76.4		2006	76.4
58	2007	77.5		2007	77.5		2007	77.5
59	2008	78.5		2008	78.5		2008	78.5
60	2009	79.7		2009			2009	79.23882
61	2010	80.8		2010	80.8		2010	80.8
62	2011	82.0		2011	82.0		2011	82.0
63	2012	83.2		2012	83.2		2012	83.2
64	2013	84.5		2013	84.5		2013	84.5
65	2014	85.8		2014	85.8		2014	85.8
66	2015	87.1		2015	87.1		2015	87.1
67	2016	88.4		2016	88.4		2016	88.4
68	2017	89.7		2017	89.7		2017	89.7
69	2018	91.0		2018			2018	90.45812
70	2019	92.3		2019	92.3		2019	92.3
71	2020	93.5		2020	93.5		2020	93.5
72	2021	94.7		2021	94.7		2021	94.7
73	2022	95.9		2022	95.9		2022	95.9
Mean		**51.8**			**52.2**			**51.8**
S.D.		**25.6**			**25.4**			**25.6**
C.V		**0.5**			**0.5**			**0.5**

Source; http://www.earth-policy.org

III. RESULT AND ANALYSIS

A. Analysis of Mean ($\bar{x}$): According to Table 1, the average value of People Population is 51. 8. In the missing value circumstance, 52.5 is recorded for People Population. After filling in the missing numbers from the derived approximated values, the result is 51.8 for People Population. After estimating the missing value using the proposed method, the values are quite similar to the original value.

B. Standard Deviation: It is observed that after generating missing values using the suggested method, values are extremely similar to the original value, and the standard deviation value is nearly equal to the standard deviation of the original set values.

C. Coefficient of Variation: It was discovered that after estimating missing values using the suggested method, the coefficients of variation were not considerably different from the CV of the original dataset.

ISSN: 3049-1479 (Online)

TABLE II

ANOVA TEST RESULT FOR TABLE I

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	11.29412683	2	5.647063413	0.00879	0.991249	3.03994
Within Groups	131697.4596	205	642.4266324			
Total	131708.7538	207				

IV. CONCLUSION

In general, it is well acknowledged that there is no send percent competent solution to manage all forms of lost values. The estimated technique is significant for numerical values. This method produces an appropriate result for the corresponding report generated by the database. CV and SD results are significant in terms of central tendency. One-way ANOVA tests also produce significant results when the hypothesis is accepted. As a result, the outcomes can be considered statistically significant. Finally, it is claimed that the presented methods are important for small databases with linear type trends.

REFERENCES

[1] Prof. Nidhi S Bhavsar, Dr. Khushbu, Dr. Darshanaben Dipakkumar Pandya , *" A New Clustering Approach for Anomaly Intrusion Detection, International Journal of Scientific Research in Science and Technology(IJSRST),* Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 11, Issue 1, pp.127-134.

[2] Nidhi S Bhavsar, Dr. Khushbu *"Closest fit Approach for A typical Value Revealing and Deciles Range Anomaly Detection Method for Recovering Misplaced value in Data Mining",* International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), Volume 8, Issue 5 September-October-2022 ,p.p-217-222

[3] S. Gaur, M.S. Dulawat, "*A perception of statistical inference in data mining"* Int. J. Comput. Sci. Commun. 1(2), 653–658 (2010)

[4] J.W. Grzymala-Busse, "*Data with missing attribute values: generalization of in-discernibility"*

[5] D.B. Rubin, Inference and missing data. Biometrika 63, 581–592 (1976)

[6] Realtion and rules induction, Transactions of rough sets. Lect. Notesin Comput. Sci. J. Subline, 1, 8–95 (2004). (Springer-Verlag)

[7] 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.

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

[9] P.D. Allison, Estimation of Linear Models with Incomplete data, Social Methodology (Jossey Bass, San Francisco, 1987), pp. 71–103

[10] L. Chen, M.T. Drane, R.F. Valois, J.W. Drane, Multiple imputation for missing ordinal data. J. Mod. Appl. Stat. Methods 4(1), 288–299 (2005)