An Applied Divided Difference Interpolation Method for Recover Arbitrarily Missing values in Data Mining

Darshanaben Dipakkumar Pandya¹, Dr. Sanjay Chaudhary², Dr. Sanjay Gaur³

¹Research Scholar, Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan.
²Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan.
³Associate Professor-CSE, Jaipur Engineering College and Research Center (JECRC), Jaipur, India.

Abstract: Data cleansing is a significant step for data research. The values misplaced in the database are an ordinary difficulty faced by data analysts. A value which is misplaced in data mining is repetitive difficulty that can produce errors in data analysis. Arbitrarily missing elements in the dataset create data analysis complex and also influenced to related result. It affects the correctness of the result and intermediary queries. By using numerical techniques, one can improve the absent data and reduce the suspiciousness in the database. The existing paper provides an applied divided difference Interpolation techniques to recuperate the misplaced/missing values.

Keywords: Data mining, missing values, Divided difference Interpolation, Arbitrarily.

I. INTRODUCTION

Usually, Information and data in the database are kept in the tabular arrangement. Data set are essentially attributes of the connected table while the records set are rows of the table. Data in the dataset reside as essential part and are used for advance reports and query. Whereas dataset is imperfect or include values which are missing, it directly has an outcome on the finishing reports. In data mining, arbitrarily missing values recognition and revival is till nowadays very essential problem. Missing values everlastingly reason of uncertainty and it effect on final results. It degrades accurateness of query and deducts decision making capability of authorities. It is essential to determine such crisis before than affecting for report preparation and query. To defeat such circumstances there is necessitate of numerical methods to recuperate the arbitrarily values which is missing.

An Applied Divided Difference Interpolation is numerical technique that can be applied to create non-natural values in connection of accessible data. The current paper is an attempt to produce non-natural value at the position of value which is missing to as recuperation technique. It mechanism as closest fit approach through applied Divided Difference Interpolation to recuperate missing value. This is essentially a request of the idea of Divided Difference Interpolation approach which is used to recuperate the values which is missing.

II. FORMULATION OF PROBLEM

The estimated numerical technique is an easy approach for obtaining arbitrarily missing value in dataset. It gives a way to work in direction of closest fit approach for recovery of missing data. In this, we first look at the complete attribute element for missing value cases. Subsequent to missing and observed values, attribute is separated in two parts as mentioned as observed and missing values. Although both are remaining in the same attribute, it is only logical demarcation.

Now looking for the missing values in attribute and search begin. At this point, we have two variable X and Y in proportion titled as year and data set value. Variable X (year) is fixed for other attributes Y, which have missing values. Attributes for Y are changeable whereas X is stable for present study and Y has missing value. Here randomly missing values are available in the attribute Y. At this point the variable X is corresponding variable of Y, which does not have any missing values.

Construct loop, for i = 1 to i<= n .

X₀ = value(Xi-1)………………………………………………… (2.1)
X₀ previous value from Xi
X₁ = value(Xi+1)…………………………………………………(2.2)
X₁ first succeeding from Xi
X₂ = value(Xi+2)…………………………………………………(2.3)
\[ X_2 = \text{second succeeding from } X_i \]
\[ Y_0 = \text{value}(Y_{i-1}) \] \hspace{1cm} (2.4)
\[ Y_0 = \text{previous value from } Y_i \]
\[ Y_1 = \text{value}(Y_{i+1}) \] \hspace{1cm} (2.5)
\[ Y_i = \text{first succeeding from } Y_i \]
\[ Y_2 = \text{Value}(Y_{i+2}) \] \hspace{1cm} (2.6)
\[ Y_i = \text{second succeeding from } Y_i \]
\[ X = \text{value}(X_i) \] \hspace{1cm} (2.7)

\( X \) is the consequent value from \( Y_i \).

Whereas \( X_0, X_1, X_2, Y_0, Y_1, Y_2, X \neq \text{"NULL"} \).

Now, initialize the variables \( \text{Sum} = 0, \text{Multi}, X, i, j, n \) \hspace{1cm} (2.8)

Now, initialize first two dimensional arrays for difference assign to zero value, therefore \( \text{diff (1)(1)} = 0 \) \hspace{1cm} (2.9)

Here, loop encountered for attribute. Thus for \( j=1 \) to \( n-1 \), the inner loop get activated in ascending order.

\[
\text{for } i=1 \text{ to } (n-j) \text{ then } \text{applied this approach for calculating difference table. Then condition is checked if } (j = 1) \text{ then }
\]
\[
\text{value(diff)(j)(diff)} = \text{value}((Y_{i+1})-\text{value}(Y_0) / \text{value}(X_{i+1})-\text{value}(X_0)) \] \hspace{1cm} (2.10)

otherwise
\[
\text{value(diff)(j)(diff)} = \text{value(diff)(j-1)} / \text{value}(X_{i+j})-\text{value}(X_0) \] \hspace{1cm} (2.11)

then make increment in \( i \) counter, thus \( i = i + 1 \), then inner loop encountered till \( i < (n-j) \).

Here inner loop is closed, after that increment \( j \) loop encountered, thus \( j = j + 1 \), loop is finished till \( j <= n-1 \). Here loop is completed.

Now, initialize first value of missing value subscript to \( \text{Sum} \) using \( \text{Sum} = Y_0 \) \hspace{1cm} (2.12)

Here, loop encountered for attribute. Thus for \( i=1 \) to \( n \), the inner loop get activated in ascending order. Now, initialize \( \text{Multiplier} \) variable to 1 using \( \text{Multi} = 1 \) \hspace{1cm} (2.13)

For \( j=0 \) to \( i-1 \) then sub loop is created for calculating estimated value. Then subtract value of \( X \) from the value of \( X_0 \) assign it to \( \text{Multi} \) variable.

\[
\text{Multi} = \text{Multi} * ((X - \text{value}(X_0)) \] \hspace{1cm} (2.14)

Then assign \( \text{Multi} \) value to the value of \( \text{value(diff)(j)(diff)} \) and finally it added to \( \text{Sum} \) variable and assigned final value to \( \text{Sum} \).

\[
\text{Sum} = \text{Sum} + \text{value(diff)(j)(diff)} * \text{Multi} \] \hspace{1cm} (2.15)

Then make increment in \( j \) counter, thus \( j = j + 1 \), then inner loop encountered till \( j <= i-1 \). Then second inner loop closed.

Then make increment in \( i \) counter, thus \( i = i + 1 \), then inner loop encountered till \( i <= n \). Then loop closed.

After these process estimated value is obtained \( Y_{est} = \text{Sum} \). Assigning estimated value to missing value place.

\[
\text{value}(Y_i) = Y_{est} \] \hspace{1cm} (2.16)

Assigning estimated value to missing value place. Then encounter loop \( i, i = i + 1 \). Here main loop get finished.

### III. ALGORITHM

Attribute \( X = \{X_1, \ldots, X_n\}, Y = \{Y_1, \ldots, Y_n\} \)

Where \( X = X_{obs} + X_{mis} \)

\( X_{obs} = \{X_1, \ldots, X_k\} \) // Observed Attribute values

\( X_{mis} = \{X_{k+1}, \ldots, X_n\} \) // Missing Attribute values

\( Y = Y_{obs} + Y_{mis} \)

\( Y_{obs} = \{Y_1, \ldots, Y_k\} \) // Observed Attribute values

\( Y_{mis} = \{Y_{k+1}, \ldots, Y_n\} \) // Attribute values missing

array(Y) = = array(X)

Read \( X = \{X_1, \ldots, X_n\}, Y = \{Y_1, \ldots, Y_n\} \) // missing data place detection.

for \( i=1 \) to \( n \), do // initialization of loop
If (value (Yi) = NULL) then
X0 = value(Xi-1) // preceding of Xi.
X1 = value(Xi+1) // first succeeding from Xi.
X2 = value(Xi+2) // second succeeding from Xi.
Y0 = value(Yi-1) // preceding of Yi.
Y1 = value(Yi+1) // first succeeding from Yi.
Y2 = Value(Yi+2) // second succeeding from Yi.
X = value(Xi) // corresponding value of missing value of Yi.
where X0, X1, X2, Y0, Y1, Y2, X ≠ "NULL"
Sum = 0, Multi, i, j, n // Initialize the variables.

diff(1)(1) = 0 // Initialize first two dimensional array.
for j = 1 to n-1, do // create loop
for i = 1 to (n-j) do // create sub loop
if (j = 1)
value(diffi)(diffj) = value((Yi+1)-value(Y0)) / value(Xi+1)-value(X0)) // calculating difference table.
else
value(diffi)(diffj) = value(diffi+1)(diffj)-value(diffi)(diffj+1) / value(Xi+j)-value(X0)) // calculating difference table
i = i + 1 // increase the i counter
endfor // second inner loop closed.
j = j + 1 // increase in j loop
repeat-until (j <= n-1),
end for //loop closed.
Sum = Y0 // initialize first value of missing value subscript.
for i = 1 to n, do // create loop
Multi = 1
for j = 0 to i-1 do // create sub loop
Multi = Multi * (X - value(X0))
Sum = Sum + value(diffi)(diffj) * Multi
j = j + 1 // increase in j loop
endfor // second inner loop closed.
i = i + 1 // increase the i counter
repeat-until (i <= n), end for // loop finish
Yest = Sum // predicted value

value (Yi) = Yest
i = i+1
repeat-until (i <= n),
endfor
stop.
The below TABLE I indicates the Deviation Anomaly Method using real database. The real data set is taken from www.earth_policy.com
| Year | Coal | Oil | Natural Gas | Coal | Oil | Natural Gas | Coal | Oil | Natural Gas |
|------|------|-----|-------------|------|-----|-------------|------|-----|-------------|
| 2003 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 2002 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1999 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1998 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1997 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1996 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1995 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1994 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1993 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1992 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1991 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1990 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1989 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1988 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1987 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1986 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1985 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1984 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1983 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1982 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1981 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1980 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1979 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1978 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1977 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1976 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1975 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1974 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1973 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1972 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1971 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1970 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1969 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1968 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1967 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1966 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1965 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1964 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1963 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1962 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |
| 1961 | 425  | 410 | 395         | 426  | 410 | 395         | 427  | 410 | 395         |
| 1960 | 428  | 410 | 395         | 428  | 410 | 395         | 428  | 410 | 395         |

Source: world_policy.com
IV. DISCUSSION OF RESULTS

1) Analysis [mean]: According to Table 1 the average value of carbon emissions from coal oil and Natural Gas are 2109, 2262 and 879 respectively. In the missing value condition values are recorded as 2,129 for coal and 2,307 for oil and 901 for Natural Gas. After filling of missing values from the calculated estimated values the results are 2,111 for coal, 2,261 for oil and 877 Natural Gas for respectively. Here, it is found that after estimation of missing value by proposed method, values are very close to original value.

2) Standard Deviation: Here, it is originate that later than generation of missing value by proposed method, values are very close to original value and value of the standard deviation are almost equal to the standard deviation of original set values.

3) Coefficient of Variation: it is found that after estimation of missing value by proposed method, values of the coefficient of variation are not very or we can say CV are similar to CV of original dataset.

4) Analysis of Variance: We wish to test the hypothesis
H0: μ1 = μ2 = μ3 against the alternative
H1: at least two μ different

For testing the hypothesis following arrangement have been done:

A. ANOVA Test Result for Coal

| Source of Variation | SS     | df | MS            | F       | P-value | F crit |
|---------------------|--------|----|---------------|---------|---------|--------|
| Between Groups      | 10811.8| 2  | 5405.902      | 0.016406| 0.983729| 3.060292|
| Within Groups       | 46459366| 141| 329499.1      |         |         |        |
| Total               | 46470178| 143|               |         |         |        |

Observed value at 5% Level of Significance = .0164, the F critical value is 3.06, so hypothesis / assumption is accepted.

B. ANOVA Test Result for Oil

| Source of Variation | SS       | df | MS            | F       | P-value | F crit |
|---------------------|----------|----|---------------|---------|---------|--------|
| Between Groups      | 62654.52107| 2  | 31327.26      | 0.082533| 0.920825| 3.059831|
| Within Groups       | 53898979.64| 142| 379570.3      |         |         |        |
| Total               | 53961634.17| 144|               |         |         |        |

Observed value at 5% Level of Significance = .0825, the F critical value is 3.06, so hypothesis / assumption is accepted.

C. ANOVA Test Result for Natural Gas

| Source of Variation | SS     | df | MS            | F       | P-value | F crit |
|---------------------|--------|----|---------------|---------|---------|--------|
| Between Groups      | 16088.51| 2  | 8044.254      | 0.049431| 0.951787| 3.060292|
| Within Groups       | 22945834| 141| 162736.4      |         |         |        |
| Total               | 22961922| 143|               |         |         |        |

Observed value at 5% Level of Significance = .04943, the F critical value is 3.06, so hypothesis / assumption is accepted.

1) Decision and Conclusion: Given that F (Observed /Calculated) < 3.06 for Coal, Oil and Natural gas ANOVA (One way) test. In case hypotheses are accepted in all cases, therefore it is considerable that, no significant difference found between groups regarding mean value.
IV. CONCLUSIONS

In common, it is commonly recognized that there is no send percent competent method to handle all types of misplaced values. The estimated approach is important for the numeral values. This approach gives suitable result for the related report created by the database. in accordance with amount of central tendency, CV and SD result are important. One way ANOVA test also provides considerable result with acceptance of hypothesis. So it can be said that the outcome are statistically important. In conclusion it can be believed that proposed methods are important for small database which contains of linear type trends in the dataset.

REFERENCES

[1] Sharma, Swati and Gaur, Sanjay, Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining”, International Journal Of Advanced Research In Computer Science, Vol.- 4(11), pp 214-217 (2013).
[2] Rubin, D.B., Inference and missing data, Biometrika, 63, pp. 581-592 (1976).
[3] Darshanaben Dipak Kumar Pandya, Dr. Sanjay Gaur, “Inliers Detection and Recovered Missing value in Data Mining”, International Journal of Emerging Technology and Advanced Engineering, Volume 8, Special Issue4, pp.1-6, April 2018.
[4] 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.
[5] 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).
[6] Darshanaben Dipak Kumar Pandya, Dr. Sanjay Gaur, Detection of Anomalous value in Data Mining, Kalpa Publications in Engineering, Volume 2, pp.1-6, 2018.
[7] Kim, J.O., and Curry, J., The treatment of missing data in multivariate analysis, Social Methodology and Research, Vol.-6, pp. 215-240(1977).
[8] Darshanaben Dipak Kumar Pandya, Dr. Sanjay Gaur, “Closest Fit Approach for Pattern Designing to Recovered Anomalous Values in Data Mining”, International Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 308 - 312, 2018.
[9] 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.
[10] Allison, P.D., Estimation of linear models with incomplete data, Social Methodology, San Francisco:Jossey Bass, pp.71-103 (1987).
[11] Allison, P.D., Missing data, Thousand Oaks CA: Sage publication, 2001
[12] 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, 1,8-95 (2004).