International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET229218

An Applied N C Differentiation Interpolation technique for improved random Anomalous 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 9, Issue 2
Page Number : 86-92

Publication Issue :
March-April-2022

Article History
Accepted : 01 April 2022
Published: 05 April 2022

ABSTRACT

In data mining, the word “interpolation” refers to interpolating some anonymous information from a given set of known information. The method of interpolation is extensively used as a valuable tool in science and engineering. The predicament is a classical one and dates back to the time of Newton, who needed to solve such a problem in analyzing data on the numerical computations. Numerical applications of interpolation include derivation of computational techniques for numerical differentiation, numerical integration and numerical solutions of differential equations. In this paper a closest fit Application of the formula to numerical data for recovering haphazard Anomalous values in Data Mining has been shown in the case of representing the data on the dataset global carbon dioxide emissions from fossil fuel burning by Fuel Type corresponding as a method of time. The formula is suitable in the situation where the values of the argument are at equal interval.

Keywords : Data mining, Interpolation, Anomalous value, Newton’s central interpolation formula, numerical data

I. INTRODUCTION

Interpolation is the process of calculating the intermediate values of a function from the set of tabulated data values or function. For example, the set of global carbon dioxide emissions from fossil fuel fuels is the time method for global carbon dioxide emissions from fuel type for the last five years 1961,1962,1963,1965 and 1970. the process of calculating the combustion of fossil fuels per type of fuel for the year 1964 is called interpolation. The interpolation process is very interesting and useful for all branches of science, humanities, business and technical branches. There are several methods of interpolation, but Newton provides the most suitable interpolation formulas. Newton interpolation formulas introduced three, known as Newton direct interpolation, Newton interpolation and Newton general interpolation formula.

It is true, interpolate the values of the corresponding variable dependent on an undetermined value of the independent variable by means of an existing interpolation form, if necessary to apply the formula for each separate value for, and, for the time being, the numerical value of the variable of the dependent variable from personal data such as should be done in one of the casinos. To get rid of these repeated numerical calculations a part of the given data, if you can think of an approach that in the representation of the numerical data given for the salvation of lost values.

II. A closest fit Application of Newton Central Difference Interpolation method

The proposed method is based on the replacement of values of abnormal attributes with artificially generated values. This method is very useful for numeric attributes. In general, this method is the search for the nearest adjustment value that is very close to the real mean of the attribute and closer to the value of the previous and next fair value of the missing values. In the process of generating the nearest adjustment values for the position of the anomalous value, therefore, it is possible here to take the values randomly as a table value for the direct interpolation table. Now searches for cases lost in the attribute begin. The first case of missing value is indicated by the subscript X [I], the search for the corresponding missing value of Y [I] is given.

First the searches of 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, first Predecessor value is denoted as PredX0 as X[I]-1. second Predecessor value is denoted as PredX1 as X[I]-2 Now take X0,X1,X2,X3 and their corresponding Y0 ,Y1, Y2,Y3 values from the database. here middle value is given and their corresponding value have to find. When search or scan pointer point out the empty subscript of the attribute, which is actually the missing values case in the attribute. The missing value case is pointed by the subscript of the attribute and is denoted by the variable. We have to find empty or NULL values for Y corresponding to X. Now we have to point out on value which is corresponding value of Y in attribute X. Now, If (X [I] = = NULL) then we have to record the first Preceding and second preceding value as

$$PredX_0 = X[I] - 1 \quad \text{.....(2.1)}$$

$$PredX_1 = X[I] - 2 \quad \text{.....(2.2)}$$

Here H is called the interval of difference , $H = X[PredX_0] – X[PredX_1]$ ....(2.3)

Here P[I] is a array that is used for missing value Y their corresponding X value minus predecessor of missing value in X[I] so , $P[I] = (X[I] – X[PredX_0]) / H$ ....(2.4)

At the next step make a pass and obtained the difference table. The differentiation's are $y1 – y0$ and $y2 – y1$, $y3 – y2$, ……, $yn – yn–1$ when denoted by dy1, dy2, dy3, ……, dyn are in that order, called the first backward differences. At next step repeat a loop for $J = 1$ to $J < N$ then , Repeat for $I = 0$ to $I < (N-J)$ then Initialize I and J variables with predecessor and successor value so,

$$I \leftarrow PredX_0 \quad \text{.....(2.5)}$$

$$J \leftarrow Succ \quad \text{.....(2.6)}$$

Then , corresponding Y[I][J] is

$$Y[I][J] \leftarrow Y[I+1][J-1] – Y[I][J-1] \quad \text{.....(2.7)}$$

Then make next iterations of I and J so , I = I+1 , J= J+1

At next step, Perform Missing value Recovery using central difference interpolation method. a closest fit Application of Newton central Interpolation method

$$Y[I] \leftarrow Y[I][J] + (Y[I+1][J-2]) + ((Y[I-1][J-2]))/2 * P[I])) + ((P[I]^2 / (2*1)) * Y[I-1][J-3])) \quad \text{.......(2.8)}$$

Display the Y value for the corresponding missing value for X.

The proposed method is based on the replacement of the haphazard haphazard values for the values generated by an application of Newton central Interpolation method. This method is very useful for numeric attributes. In general, this method is the search for haphazard haphazard values that is very close to the real mean of the attribute and closer to the value than the original value of missing values. The below table shows the overall idea of central difference table.

Table 1. Central Difference table for calculating the Y values of the corresponding X values using formula.

X	Y	ΔY	$\Delta^2 Y$	$\Delta^3 Y$	$\Delta^4 Y$
X-2	Y-2				
		Δ Y-2			
X-1	Y-1		Δ^2 Y-2		
		Δ Y-1		Δ^3 Y-2	
X_0	Y_0		Δ^2 Y-1		Δ^4 Y-2
		ΔY_0		Δ^3 Y-1	
X_1	Y_1		$\Delta^2 Y_0$		
		ΔY_1			
X_2	Y_2				

III. An Application for Newton Central Difference Interpolation method algorithm

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

Introduction: Given an array X and Y are of size N, N= 50, this procedure replaces the missing values with the recovered data from data set. Here PredX_0 is the first predecessor of the missing data and PredX_1 is the second predecessor of the missing data. Here two arrays are taken first is X[I] and Y[I][J] is two dimension array which is used for storing differences of table. The variable I is used to index elements from 1 to N in a given data. The variable J is used to index column 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 the variables]

I ← NULL, J ← NULL, N ← 50, H ←NULL, P[i]=NULL,

PredX_0 ← NULL, PredX_1←NULL.

Step 3: [Create a loop for N passes]

Repeat for I = 0 to I < N.

Read X [I] and Y [I][0].

If (X [I] = = NULL) then

$PredX_0$ = X [I] - 1 // First Predecessor value of missing value.

$PredX_1$ = X [I] - 2 // Second Predecessor value of missing value.

And H = X [$PredX_0$] – X [$PredX_1$] // Interval of successor and Predecessor value

P[I] = (X [I] – X[$PredX_0$]) / H // difference of X[I] of missing data and predecessor value.

Step 4: [Make a Pass and Obtained central difference table]

Repeat for J = 1 to J < N then

Repeat for I = 0 to I < (N-J) then

I← $PredX_0$, J←Succ // Initialize I and J variables with predecessor and successor value

Y[I][J] ← Y[I+1][J-1] – Y[I][J-1] then I ← I+1 , J← J+1

Step 5: [Display Central difference table]

Repeat for I = 0 to I < N then

Print X[I] and I ← I+1

Repeat for J = 0 to J < (N-J) then

Print Y[I][J] and J= J+1

Step 6: [Perform Missing value Recovery using central difference interpolation method]

$Y[I] \leftarrow Y[I][J] + (Y[I+1][J-2]) + ((Y[I-1][J-2]))/2 * P[I])) + ((P[I]^2/ (2*1)) * Y[I-1][J-3]))$

Step 7: [Display the Y value for the corresponding missing value for X]

Print Y[i]

Step 8: Finished.

Stop.

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 randomly, the mean calculated from incomplete data sets are 2,125 for coal, 2,257 for oil and 906 for natural gas.

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 2,109, 2,259 and 875 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	2153.033359	2	1076.517	0.003231	0.996774	3.060292
Within Groups	46981137.13	141	333199.6			
Total	46983290.16	143				

Table 1 Value :- F(2, 141) 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	26340.47	2	13170.24	0.032878	0.967664	3.060292
Within Groups	56481620	141	400578.9			
Total	56507961	143				

Table 2 Value :- F(2, 141) 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	9139.001	2	4569.5	0.027803	0.972585	3.060292
Within Groups	23173403	141	164350.4			
Total	23182542	143				

Table 3 Value :- F(2, 141) 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 A Suit Approach of Newton Central Interpolation method for haphazard Anomalous values of data. 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	___	1351	980	262
4	1963	1396	1,052	300	1396	1,052	300	1396	1,052	300
5	1964	1435	1,137	328	1435	___	328	1435	1,147	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	___	1,423	410	1430	1,423	410
9	1968	1448	1,551	446	1448	1,551	___	1448	1,551	427
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	583	1576	2,055	583
14	1973	1581	2,240	608	___	2,240	608	1562	2,240	608
15	1974	1579	2,244	618	1579	2,244	618	1579	2,244	618
16	1975	1673	2,131	623	1673	2,131	623	1673	2,131	623
17	1976	1710	2,313	650	1710	2,313	650	1710	2,313	650
18	1977	1766	2,395	649	1766	___	___	1766	2,210	623
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	___	2,289	756	1932	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	2,182	836	2237	2,182	836
27	1986	2300	2,290	830	2300	___	830	2300	2,322	830
28	1987	2364	2,302	893	2364	2,302	893	2364	2,302	893
29	1988	2414	2,408	936	2414	2,408	936	2414	2,408	846
30	1989	2457	2,455	972	2457	2,455	972	2457	2,455	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	2,385	2,586	1,184
37	1996	2,451	2,624	1,208	2,451	2,624	1,208	2,451	2,624	1,208
38	1997	2,480	2,707	1,211	2,480	2,707	1,211	2,480	2,707	1,211
39	1998	2,376	2,763	1,245	2,376	___	1,245	2,376	2,633	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	____	2,819	1,349	**2,520**	2,819	1,349
44	**2003**	2,638	2,928	1,399	2,638	____	1,399	2,638	**3,055**	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	**1,465**
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	2,125	2,257	906	2,109	2,259	875
S.D		567.89	621.13	400.27	580.06	620.20	396.00	568.79	621.73	399.92
C.V		0.27	0.27	0.46	0.27	0.27	0.44	0.27	0.28	0.46

V. CONCLUSION

In this work the problem of detecting haphazard Anomalous values in streams of data has been addressed. In general, there is no universal and absolute technique for managing the values of missing attributes. The closest fitting method proposed is useful for the numerical attribute, with a deviation lower than the average. This is the best way to recover haphazard Anomalous values from the database. Accordingly, it is noted that the techniques for managing the values of Anomalous attributes must be chosen individually or according to the nature and type of data.

VI. REFERENCES

[1]. Numerical Methods for Scientists and Engineers by Richard Hamming, Second Edition, Dovers Publications.

[2]. M.K. Jain, S.R.K. Iyengar, R.K. Jain, “Numerical Methods for Scientific and Engineering Computation”, New Age International Publishers.Applied Numerical Methods, 3rd Edition by Steven C. Chapra, Raymond P. Canale, Tata McGraw Hill Education.

[3]. David B. Thompson, “Numerical Methods 101 – Convergence of Numerical Models”. USGS Staff, University of Nebraska – Lincoln.

[4]. SIAM Journal on Numerical Analysis- Developing and Analyzing Numerical Methods. ISSN 1095 – 7170.

[5]. Fundamentals of Numerical Methods and statistical techniques by Anshuman and Jaspal. ISBN: 81 – 89510-32-0.

[6]. T.E. Simos ,“ New closed Newton Cotes type formulae as multilayer and symplectic Integrators” , THE JOURNAL OF CHEMICAL PHYSICS 104108, P 133, 2010.

[7]. T.E. Simos , “ New Stable Closed Newton -Cotes Trigonometrically Fitted Formulae for Long -Time Integration” , Hindawi Publishing Corporation Abstract and Applied Analysis Volume 2012, Article ID 182536, 15 pages.

[8]. Nasrin Akter Ripa, “investigation of Newton’s Forward Interpolation Formula ”, International Journal of Computer Science & Emerging Technologies (E-ISSN:2044-6004)12 Volume 1, Issue 4, December 2010.

Cite this article as :

Dr. Darshanaben Dipakkumar Pandya, Dr. Abhijeetsinh Jadeja, Dr. Sheshang D. Degadwala, "An Applied N C Differentiation Interpolation technique for improved random Anomalous values in Data Mining", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 2, pp. 86-92, March-April 2022. Available at doi : https://doi.org/10.32628/IJSRSET229218
Journal URL : https://ijsrset.com/IJSRSET229218