iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 Issue: V Month of publication: May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5220

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

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

Darshanaben Dipakkumar Pandya[1], Dr. Sanjay Chaudhary[2], Dr. Sanjay Gaur[3]
[1]*Research Scholar, Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan.*
[2]*Department of Computer Science, Madhav University, Pindwara, Sirohi, Rajasthan.*
[3]*Associate 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 .

X_0 = value(Xi-1)………….…………………………………..…... (2.1)

X_0 previous value from Xi

X_1 = value(Xi+1)…………………………………….………..…….(2.2)

X_1 first succeeding from Xi

X_2 = value(Xi+2)……………………………….…………………(2.3)

X_2 second succeeding from Xi

Y_0 = value(Yi-1)……………….………………….……………….(2.4)

Y_0 previous value from Yi

Y_1 = value(Yi+1)……………….……………………………..….(2.5)

Y_1 first succeeding from Yi

Y_2 = Value(Yi+2)………………………………………………...(2.6)

Y_2 second succeeding from Yi

X = value(Xi) ……………………………….…………….....(2.7)

X is the consequent value from Yi.

Whereas X_0 , X_1, X_2, Y_0 , Y_1, Y2, X ≠ „NULL‟.

Now, initialize the variables Sum =0 , Multi , X , i , j ,n ………………...……(2.8)

Now, initialize first two dimensional arrays for difference assign to zero value, therefore

diff (1)(1) = 0……………………………………………………....(2.9)

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

for i=1 to (n-j) then applied this approach for calculating difference table. Then condition is checked if (j = =1) then

value($diff_i$)($diff_j$) = value((Y_{i+1})- value(Y_0) / value(X_{i+1})- value(X_0))..….(2.10)

otherwise

value($diff_i$)($diff_j$)= value($diff_{i+1}$)($diff_{j-1}$)- value($diff_i$)($diff_{j-1}$) /

value(X_{i+j})- value(X_0)) ……………………(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 Sum using

Sum = Y_0 ………………………………….(2.12)

Here, loop encountered for attribute. Thus for i=1 to n, the inner loop get activated n

Ascending order. Now , initialize Multiplication variable to 1 using

Multi = 1……………………………...... (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 Multi variable.

Multi = Multi * ((X - value(X_0))…………………..(2.14)

Then assign Multi value to the value of value($diff_i$)($diff_j$) and finally it added to Sum variable and assigned final value to Sum.

Sum = Sum + value($diff_i$)($diff_j$) * Multi…………….(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 Yest = Sum. Assigning estimated value to missing value place.

value (Yi) = Y_{est} …………………………………….……………..(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 = {X1 , ……, Xn }, Y = { Y1 , ……, Yn }

Where X = Xobs + Xmis

Xobs = { X1 , ……, Xk} // Observed Attribute values

Xmis = { Xk+1 , ……, Xn} // Missing Attribute values

Y = Yobs + Ymis

Yobs = { Y1 , ……, Yk} // Observed Attribute values

Ymis = { Yk+1 , ……, Yn} // Attribute values missing

array(Y) = = array(X)

Read X = { X1 , ……, Xn }, Y = { Y1 , ……, Yn } // missing data place detection.

for i=1 to n, do // initialization of loop

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

If (value (Yi) = = NULL) then
X_0 = value(Xi-1) //preceding of Xi.
X_1 = value(Xi+1) // first succeeding from Xi.
X_2 = value(Xi+2) //second succeeding from Xi.
Y_0 = value(Yi-1) // preceding of Yi.
Y_1 = value(Yi+1) // first succeeding from Yi.
Y_2 = Value(Yi+2) // second succeeding from Yi.
X = value(Xi) // corresponding value of missing value of Yi.
where X_0 , X_1, X_2, Y_0 , Y_1, Y2, X ≠ „NULL"
Sum =0 , Multi , X , 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($diff_i$)($diff_j$) = value((Y_{i+1})- value(Y_0) / value(X_{i+1})- value(X_0))
// calculating difference table.
else
value($diff_i$)($diff_j$)= value($diff_{i+1}$)($diff_{j-1}$)- value($diff_i$)($diff_{j-1}$) /
value(X_{i+j})- value(X_0))
// 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 = Y_0 // 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(X_0))
Sum = Sum + value($diff_i$)($diff_j$) * 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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

TABLE-1

An Applied Divided Difference Interpolation Method

Global Carbon Dioxide Emission from Fossil burning by Fuel Type 1960-2009 (Carbon Emission in Million Tones)

Standard Dataset					Missing Value Dataset			Recovered Dataset		
SN	Year	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas	Coal	Oil	Natural Gas
1	1960	1,410	849	235	1,410	849	235	1,410	849	235
2	1961	1349	904	254	1349	___	254	1349	913	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	___	1435	1,137	325
6	1965	1460	1,219	351	1460	1,219	351	1460	1,219	351
7	1966	1478	1,323	380	___	1,323	380	1440	1,323	380
8	1967	1448	1,423	410	1448	___	410	1448	1,434	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	___	1556	1,839	523
12	1971	1559	1,946	554	1559	1,946	554	1559	1,946	554
13	1972	1576	2,055	583	___	2,055	583	1574	2,055	583
14	1973	1581	2,240	608	1581	___	608	1581	2,219	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	___	1710	2,313	631
18	1977	1766	2,395	649	1766	2,395	649	1766	2,395	649
19	1978	1793	2,392	677	___	2,392	677	1827	2,392	677
20	1979	1887	2,544	719	1887	___	719	1887	2,456	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	___	1992	2,196	727
24	1983	1995	2,177	745	1995	2,177	745	1995	2,177	745
25	1984	2094	2,202	808	___	2,202	808	2135	2,202	808
26	1985	2237	2,182	836	2237	___	836	2237	2,257	836
27	1986	2300	2,290	830	2300	2,290	830	2300	2,290	830
28	1987	2364	2,302	893	2364	2,302	893	2364	2,302	893
29	1988	2414	2,408	936	2414	2,408	___	2414	2,408	928
30	1989	2457	2,455	972	2457	2,455	972	2457	2,455	972
31	1990	2409	2,517	1,026	___	2,517	1,026	2387	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	___	2,331	2,562	1,081
36	1995	2,414	2,586	1,153	2,414	2,586	1,153	2,414	2,586	1,153
37	1996	2,451	2,624	1,208	___	2,624	1,208	2504	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	2,763	1,245	2,376	2,763	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	2,487	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	2,129	2,307	901	2,111	2,261	877
	S.D	567.89	621.13	400.27	586.60	606.41	410.80	568.93	619.65	399.97
	C.V	0.27	0.27	0.46	0.28	0.26	0.46	0.27	0.27	0.46

Source: www.earth_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: $\mu 1 = \mu 2 = \mu 3$ against the alternative

H1: at least two μ different

For testing the hypothesis following arrangement have been done:

A. *ANOVA Test Result for Coal*

ANOVA

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*

ANOVA

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*

ANOVA

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 Dipakkumar 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 Dipakkumar 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 Methods and Research, Vol.-6, pp. 215-240(1977).

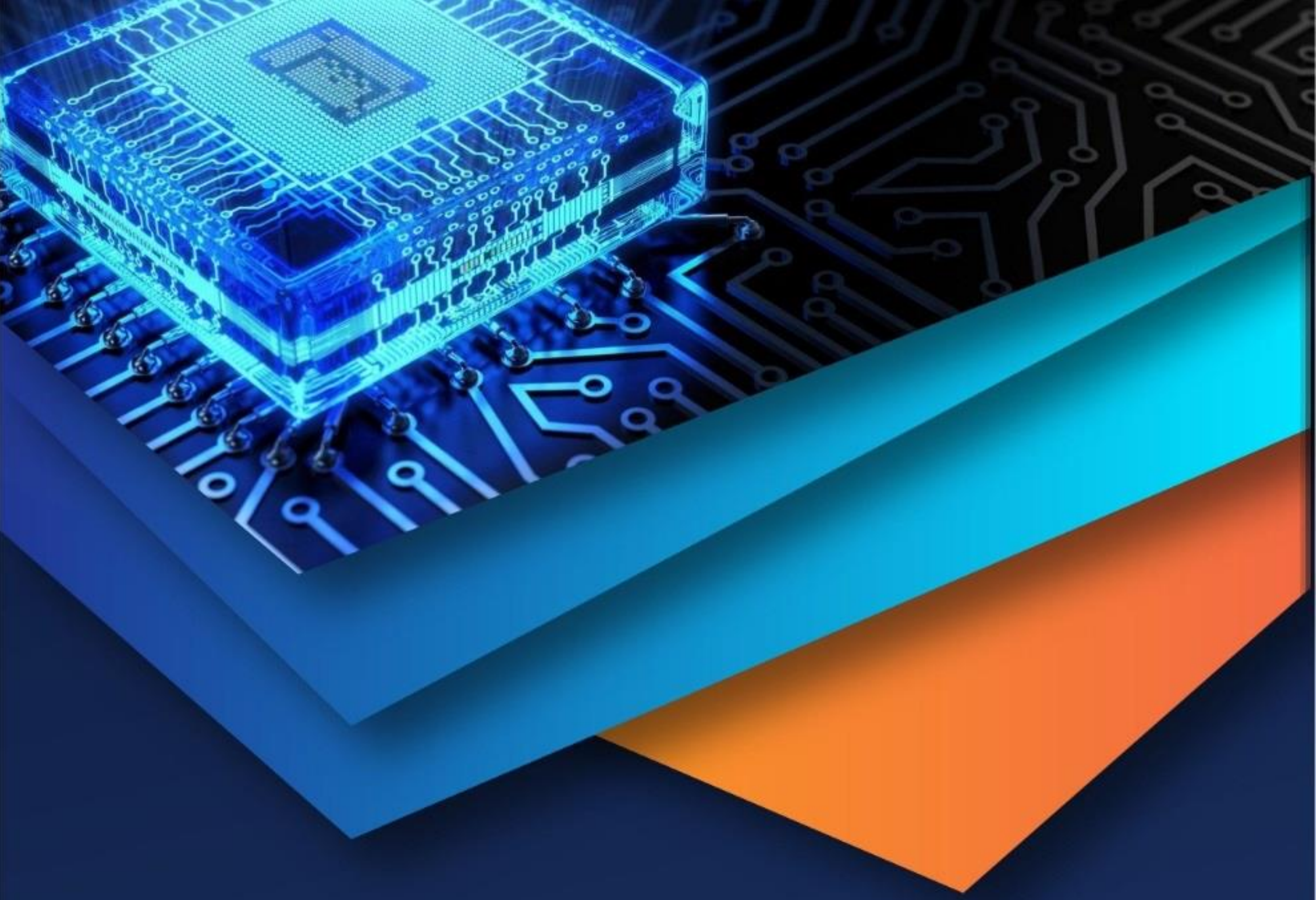
[8] Darshanaben Dipakkumar Pandya, Dr. Sanjay Gaur, "Closest Fit Approach for Pattern Designing to Recovered Anomalous Values in Data Mining", International Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 308 - 312, 2018.

[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).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY