ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)

Available Online at www.ijetcr.org

Volume 10; Issue 6; November-December; 2022; Page No. 12-16

Tendency Moderate Perspective to Recover Absent Number in Data Mining

Sunil Kumar Paliwal

Associate Professor, Department of Computer Science, SGI, Upali oden, Nathdwara, Rajasthan, India

Conflicts of interest: Nil

Corresponding author: Sunil Kumar Paliwal

Abstract

In this paper, we discussed tendency perspective to find the nearest value of missing value, this missing value present in any numeric database. If missing value can be predicted by any statistical method, then it will help to improve the result in data analysis. Many universities and organization are interested in how to recover missing value which can help to reduce problems in database analysis and used for different purpose. Missing values affect directly on result

In this work, a method has been proposed, by this tendency method. The missing value is replaced by the most probable value. Our main aim was to find out the missing value for incomplete database with many missing values. Tendency perspective is better for number related database.

Keywords: Data mining, Tendency Perspective, Data making, Incomplete Database, Missing numeric value.

I. Introduction

Data mining is to extract or mining, knowledge from the huge amount of data the term data mining use many perspective, statistical as well as mathematical techniques. Data pre-processing and cleaning is the basis step of data mining, which is possible with statistical inference. There have been many achievements in the data mining, data preparation may be more time consuming than data mining and can present equal if not more challenges than data mining Yan et al [1].

Real world data may be incomplete. By finding missing data generates quality database which leads to quality patterns there recover incomplete data mean filling the values missed or reducing ambiguity. In this study tendency perspective based algorithm is introduced and explains which gives an idea to find out a way to recover or produce missing values from a not balanced data base there may be possible missing values are in big amount. Our final aim to find out the nearest value of missing values, due to missing value database analysis faced lot of problem

LITERATURE SURVEY

Clark [2] studied and explained the relation between statistical method and data mining. David [3, 4, 5] considered that the disciplines of statistics and data mining have common aims. Gaur and Dulawat [6, 7, 8, 9] discussed on the various aspects associated with statistical inference and data mining. They also explore the role of statistical methods and tools to fulfill the requirement of data mining to make database pure and complete. Nisbet, Elder and Miner [15] (2009) explain statistical analysis and data mining which deals

with massive and complex data analysis. Buck [10] (1960) suggested a method of estimation of missing values in multivariate data suitable for use with an electronic computer. Hartley and Hocking [14] (1971) made analysis on incomplete data. Little and Rubin [11] (1990, 2002) studied the statistical analysis with missing values; they also considered the data analysis under the title modern method of data analysis. Famili et al.[13] (1997) suggested the method for data preparation and intelligent data analysis. Buhi et al.[12] (2008) gave mechanism for addressing missingness of data. Gaur and Dulawat, focused on the estimation of missing values with various algorithms and their statistical justifications. They developed many algorithms to deal missing values case in the database

PROBLEM STATEMENT AND OBJECTIVE

The main problems in the numeric database arise from missing data, we try to find the nearest value of missing value, this missing value present in any numeric database. The tendency method is based on replacing missing attribute values by the calculated values. This tendency method is find value near to the real value. For the explanation of tendency Moderate perspective first we take one table which has some numeric value then first we hide some value as showing table. Now we try to find this missing value by tendency Moderate method.

PROPOSED SYSTEM

In this perspective of tendency middling, we first read complete data with the observed and missing values. In second step blank Cells is to point out by search pointer of the database this blank cells is actually the missing values case in the database. In this perspective missing value case is pointed by the subscript of the database and denoted by the variable xi. After pointing missing value case, our process is start we take three preceding value which is (x i - 1) , (x i -2), and (x i -3) respectively. This preceding value is written in the tendency Moderate manner. According to tendency perspective we multiply .6 to first value to take 60% weight of first preceding value , then multiply by .45 to take 45% weight of second presiding value and then by .3 to take 30% weight of third preceding value. First preceding is equal to value xp1 second preceding is equal to value xp2 and third preceding is equal to value xp3. For equality of all values, the sum of all weighted preceding values is divided by the weight 1.35.

Now for succeeding value from the missing value, .First succeeding, second succeeding and third succeeding values are equal to value (x i +1), value (x i +2) and value (x i + 3) respectively. we have to take 60% of first succeeding value ,45% weight of second succeeding value, and 30%weight of third succeeding value. for equality of all value the sum of all weighted succeeding values is divided by the weight 1.35. then after finding xp and xs in next step we calculate $\bar{X}$ps .

There $\bar{X}$ps = Xest

A. Algorithm -

Attribute X = {x1,…., xn}
Where. $X = X_{obs} + X_{mis}$
Xobs = {x1,….,x k}. // Attribute values observed
$X_{mi} = \{x_k+1, \dots, x_n\}$ **// Attribute values missing**
X = {x1,…., xn} // Attribute with observed and missing values For i =1to n do
If (value (xi) == NULL) then

xp1 = value (xi -1) // Value of first preceding of xi
xp2 = value (xi -2) // Value of second preceding of xi
xp3 = value (xi -3) // Value of third preceding of xi Vp1=xp1*.60
Vp2=xp2*.45 Vp3=xp3*.30
xs1 = value (xi +1) // Value of first succeeding of xi xs2 = value (xi +2) // Value of second succeeding of xi xs3 = value (xi +3) // Value of third succeeding of xi Vs1=xs1*.60
V_s2=xs2*.45

Vs3=xs3*.30
xp= ((Vp1 +Vp2 +Vp3)/(.60+.45+.30)) xs= ((Vs1 +Vs2 +Vs3)/ (.60+.45+.30))
$\bar{X}$ps = (xp+ xs) / 2 // Average of preceding and succeeding
xest = $\bar{x}_{ps}$ // Estimated value

value (xi) = xest // Assigning estimated value to missing value place
i = i+ 1 repeat untill (i>=n)
Stop

RESULTS AND ANALYSIS

TABLE (1) is taken as numerical database .the mean of Coke, Grapico, and Pepsi are 1,761, 6,122 and 6,274 respectively. **Table (P)** shows the variables with missing values and observed value .in this we hide 15 % of the values in the random manner for all the variables from Table P. The mean of are Table **(Q**)are less than the mean values from **Table (P).** Tendency perspective is applied on the data sets of **Table (Q)** to fill up the missing values. These recover values are shown in **TABLE (R)** for all the variables which is shown in **Table (R)**. It is observed that mean values of Coke, Grapico and Pepsi are 1760, 6122 and 6278 respectively. It is considerable that the mean values obtained after replacing the missing values by the Tendency Moderate perspective values very close to the **TABLE (P)** mean value.

Table - [1]: Tendency Moderate Perspective to Recover Missing Values

Table (P)				Table (Q)				Table (R)			
Real Value				Missing Values(15% approx)				Table With Estimated Values			
Year	Coke	Grapico	Pepsi	Year	Coke	Grapico	pepsi	Year	coke	Grapico	pepsi
Million Gallons				Million Gallons				Million Gallons			
1980	1,465	6,063	5,876	1980	1,465	6,063	5,876	1980	1,465	6,063	5,876
1981	1,456	6,020	5,772	1981	1,456	6,020	5,772	1981	1,456	6,020	5,772
1982	1,409	5,908	5,809	1982	1,409	5,908	5,809	1982	1,409	5,908	5,809
1983	1,428	5,946	5,950	1983		5,946	5,950	1983	1,460	5,946	5,950
1984	1,474	6,020	6,112	1984	1,474		6,112	1984	1,474	6,049	6,112
1985	1,484	6,142	6,323	1985	1,484	6,142		1985	1,484	6,142	6,137
1986	1,503	6,170	6,399	1986	1,503	6,170	6,399	1986	1,503	6,170	6,399
1987	1,482	6,133	6,279	1987		6,133	6,279	1987	1,496	6,133	6,279
1988	1,504	6,175	6,063	1988	1,504	6,175	6,063	1988	1,504	6,175	6,063
1989	1,494	6,229	6,275	1989	1,494			1989	1,494	6,193	6,355
1990	1,517	6,212	6,500	1990	1,517	6,212	6,500	1990	1,517	6,212	6,500

1991	1,662	6,242	6,562	1991	1,662	6,242	6,562	1991	1,662	6,242	6,562
1992	1,864	6,233	6,412	1992		6,233	6,412	1992	1,758	6,233	6,412
1993	1,959	6,154	5,845	1993	1,959	6,154		1993	1,959	6,154	5,920
1994	1,942	6,186	5,286	1994	1,942	6,186	5,286	1994	1,942	6,186	5,286
1995	1,907	6,168	5,191	1995	1,907		5,191	1995	1,907	6,187	5,191
1996	1,857	6,215	5,762	1996	1,857	6,215	5,762	1996	1,857	6,215	5,762
1997	1,777	6,181	6,151	1997	1,777	6,181	6,151	1997	1,777	6,181	6,151
1998	2,078	6,148	6,381	1998	2,078	6,148	6,381	1998	2,078	6,148	6,381
1999	2,100	6,191	6,810	1999	2,100	6,191	6,810	1999	2,100	6,191	6,810
2000	2,012	6,141	7,215	2000	2,012		7,215	2000	2,012	6,129	7,215
2001	2,127	6,078	6,699	2001	2,127	6,078		2001	2,127	6,078	6,837
2002	2,038	6,104	6,605	2002		6,104	6,605	2002	2,073	6,104	6,605
2003	1,969	6,069	6,849	2003	1,969	6,069	6,849	2003	1,969	6,069	6,849
2004	2,131	6,034	7,028	2004	2,131	6,034	7,028	2004	2,131	6,034	7,028
2005	2,154	6,018	6,973	2005	2,154	6,018	6,973	2005	2,154	6,018	6,973
Mean	1,761	6,122	6,274		1,772	6,119	6,272		1,760	6,122	6,278
SD	272.96	89.83	507.7. 511.7059		274.644	538.0228	538.023271.		6455	87.65148	

CONCLUSIONS

The use of algorithms to recover missing values would result very closest to original mean values. The variation in mean values from the standard dataset's mean value will deviate zero to two percent maximum, where five percent deviation were on acceptance. In the maximum case results are 99% accurate. This Tendency perspective show still there is no methods which find accurate missing value. This method is appropriate for numeric database. Proposed algorithms may be used at the place of excluding of missing values case.

REFERENCES

1. Yan, X., C. Zhang, and S. Zhang. 2003. Towards databases mining: Pre-processing collected data. Applied Artifi cial Intelligence 17(5–6):545–561
2. Clark, G., Madigan, D., Pregibon, D. and Smyth, P. (1996): Statistical inference and data mining, Communication of ACM, Vol.39.
3. David, J. H. (1998): Data mining: statistics and more?, The American Statistics, Vol. 52, No.2.
4. David, J. H. (1999): Statistics and data mining: interesting disciplines, Department of Mathematics, Imperial College, London, UK, SIGKDD Exploration, Vol.1(1).
5. David, C. H. (2006): The treatment of missing data, University of Vermont.
6. Gaur Sanjay and Dulawat, M. S. (2010,a): "A perception of statistical inference in data mining", International Journal of Computer Science and Communication, 1(2), 653-658
7. Gaur Sanjay and Dulawat, M. S. (2010,b): "Univariate analysis for data preparation in context of missing values", Journal of Computer and Mathematical Sciences, 1(5), 628-635.
8. Sanjay Gaur and M.S. Dulawat, Improved closest fit techniques to handle missing attribute values, Journal of Computer and Mathematical Sciences, 2011, 2(2), 384-390.
9. Sanjay Gaur and M.S. Dulawat, A closest fit perspective to missing attribute values in data mining, International Journal of Advances in Science and Technology, 2011, 2(4), 18-24.
10. Buck, S. F. (1960): 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, 302-306.
11. Little, Roderick and Rubin, D.B. (2002): Statistical Analysis with Missing Data 2nd Edition. Hoboken, NJ: John Wiley & Sons, Inc.
12. Buhi, R., Goodson, P. and Neilands, T. B. (2008): Out of sight, Not out of mind: Strategies for handling missing data, Am. J. Health BehavTM , 32(I), 83-92.
13. Famili, A., Wei-Min, S., Weber, R. and Simoudis, E. (1997): Data pre-processing and intelligent data analysis, Intelligent Data Analysis, 1(1), Elsevier.
14. Hartley, H .O. and Hocking, R. R. (1971): The analysis of incomplete data, Biometrics, 27, 783-823.
15. Nisbet, R., Elder, J. and Miner, G. (2009): Statistical analysis & data mining applications. Academic press, An Imprint of Elsevier, New York