ISSN 2277-3061

Comparative Study of Three Imputation Methods to Treat Missing Values

Rahul Singhai
IIPS, Devi Ahilya Vishwavidyalaya, Indore, India
singhai_rahul@hotmail.com

ABSTRACT

One relevant problem in data preprocessing is the presence of missing data that leads the poor quality of patterns, extracted after mining. Imputation is one of the widely used procedures that replace the missing values in a data set by some probable values. The advantage of this approach is that the missing data treatment is independent of the learning algorithm used. This allows the user to select the most suitable imputation method for each situation. This paper analyzes the various imputation methods proposed in the field of statistics with respect to data mining. A comparative analysis of three different imputation approaches which can be used to impute missing attribute values in data mining are given that shows the most promising method. An artificial input data (of numeric type) file of 1000 records is used to investigate the performance of these methods. For testing the significance of these methods Z-test approach were used.

Indexing terms/Keywords

Knowledge Discovery In database; Data mining; Imputation methods; Sampling. Attribute missing values; Data preprocessing.

Academic Discipline And Sub-Disciplines

Computer Science & Applications

SUBJECT CLASSIFICATION

Data Mining, Data Warehousing, Data Preprocessing, DBMS.

TYPE (METHOD/APPROACH)

Provide examples of relevant research types, methods, and approaches for this field: E.g., Historical Inquiry; Quasi-Experimental; Literary Analysis; Survey/Interview, In this paper three different imputation methods that used in statistics are proposed to treat missing values problem in data mining environment. Sampling approach is used to obtain the reduced representation of large data set. At the end the Z-test method is used to examine the performance of proposed imputation methods.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 11 No 7
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 2277-3061

1. INTRODUCTION

Missing value imputation is an actual yet challenging issue confronted in data mining. Missing values may generate bias and affect the quality of the KDD process. In most cases, data set attributes are not independent from each other. Thus, through the identification of relationships among attributes, missing values can be determined. Imputation is a method to impute missing values in attribute data set. The objective of this work is to compare & contrast the performance of four different imputation methods proposed in statistics in a large database, so that the best suitable methods could be proposed in data mining.

2. Imputation Methods For Missing Data Treatment Using Auxiliary Information:

Imputation is a technique used computation of missing values in the sample obtained as consequence of a survey procedure. In literature, several imputation techniques are described, some of them are better over others. Rubin (1976) addressed three concepts: MAR (missing at random), OAR (observed at random) and PD (parametric distribution). In what follows MCAR (missing completely at random) is used. Let $\bar{Y} = N^{-1}\sum_{i=1}^{N} Y_i$ be the mean of a finite data set under consideration for estimation. A simple random sample *S* without replacement (SRSWOR), of size *n* is drawn from data set $\Omega = \{1,2,...,N\}$ to estimate $\bar{Y}$. The sample *S* of *n* units contains *r* responding units $(r < n)$ forming a set *R* and (*n* – *r*) non-responding with the sub-space (*n* – *r*) having symbol R^C in the space. The attribute *Y* is of main interest and *X* an auxiliary attribute correlated with *Y*. For every unit $i \in R$, the value y_i is observed available. However, for the units $i \in R^C$, the y_i values are missing and imputed values are to be derived. The i^{th} value x_i of auxiliary e is used as a source of imputation for missing data when $i \in R^C$. This is to assume that for sample *S*, the data $x_S = \{x_i : i \in S\}$ are known and $S = R \cup R^C$.

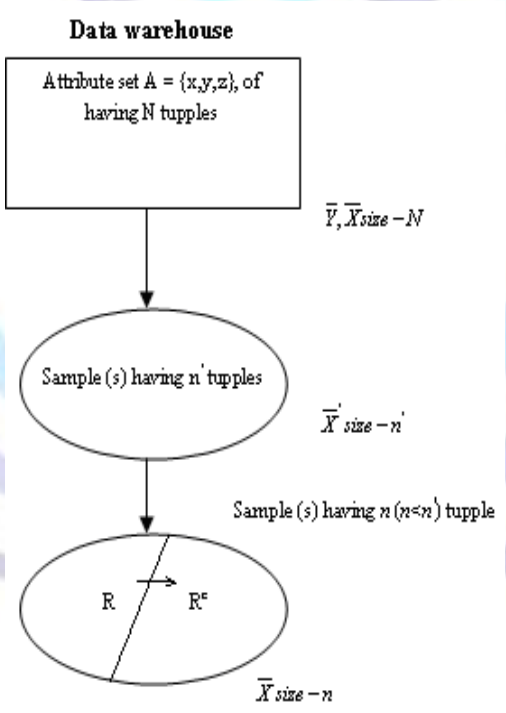

Fig 1: the diagrammatic representation of MCAR sampling procedure.

Under the above setup, some well known imputation methods, that can be used in data mining are given below :

2.1 Mean Method of Imputation :

For y_i define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \bar{y}_r & if \quad i \in R^C \end{cases} \qquad \ldots(2.1)$$

ISSN 2277-3061

Using above, the imputation-based estimator of data set mean $\bar{Y}$ is :

$$\bar{y}_m = \frac{1}{r}\sum_{i\in R}\bar{y}_i = \bar{y}_r \qquad \ldots(2.2)$$

2.2 Ratio Method of Imputation:

The heading for subsubsections should be in Arial11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

For sampled values y_i and x_i define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \hat{b}x_i & if \quad i \in R^C \end{cases} \qquad \ldots(2.3)$$

where $\hat{b} = \sum_{i\in R} y_i \Big/ \sum_{i\in R} x_i$

Using above, the imputation-based estimator of data set mean $\bar{Y}$ is:

$$\bar{y}_S = \frac{1}{n}\sum_{i\in S}\bar{y}_{\bullet i} = \bar{y}_r\left(\frac{\bar{x}_n}{\bar{x}_r}\right) = \bar{y}_{RAT} \qquad \ldots(2.4)$$

where $\bar{y}_r = \frac{1}{r}\sum_{i\in R} y_i$, $\bar{x}_r = \frac{1}{r}\sum_{i\in R} x_i$ and $\bar{x}_n = \frac{1}{n}\sum_{i\in S} x_i$

2.3 Compromised Method of Imputation :

Singh and Horn (2000) proposed compromised imputation procedure

$$y_{\bullet i} = \begin{cases} (n/r)\bar{y}_i + (-\alpha)\hat{b}x_i & if \quad i \in R \\ (-\alpha)\hat{b}x_i & if \quad i \in R^C \end{cases} \qquad \ldots(2.5)$$

where α is a suitably chosen constant, such that the resultant variance of the estimator is minimum. The imputation-based estimator, for this case, is

$$\bar{y}_{COMP} = \left[\alpha\bar{y}_r + (-\alpha)\bar{y}_r\frac{\bar{x}_n}{\bar{x}_r}\right] \qquad \ldots(2.6)$$

Lemma : The bias, m.s.e. and minimum m.s.e. of $\bar{y}_{COMP}$ is [As per Singh and Horn (2000)]:

(i) : $$B(\bar{y}_{COMP}) = (-\alpha)\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}(C_X^2 + \rho C_Y C_X) \qquad \ldots(2.7)$$

(ii) : $$M(\bar{y}_{COMP}) = \left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2 C_Y^2 + \left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2\left[(-\alpha)^2 C_X^2 - 2(-\alpha)\rho C_Y C_X\right] \qquad \ldots(2.8)$$

(iii) : For optimum $\alpha = \left(1-\rho\frac{C_Y}{C_X}\right)$, the minimum m. s. e. of $\bar{y}_{COMP}$ is given by the expression

ISSN 2277-3061

$$M\left(\bar{x}_{COMP}\right)_{\min} = \left[\left(\frac{1}{r}-\frac{1}{N}\right)-\left(\frac{1}{r}-\frac{1}{n}\right)\rho^2\right]S_Y^2 \qquad \ldots(2.9)$$

3 Testing The Significance Of The Difference Between The Means Of Two Large Samples:

Suppose two random samples of n1 and n2 members respectively have been drawn from the same data set of standard deviation σ. since the difference of their means ($\bar{x}_1 \sim \bar{x}_2$) is due to fluctuations of sampling due to the assumption that the samples are independent and drawn from the same data set. The standard error e of the difference of their means is given by $e^2 = e_1^2 + e_2^2$, where e1 and e2 are the standard error of the means of the two samples and are $\frac{\sigma}{\sqrt{n_1}}$ and $\frac{\sigma}{\sqrt{n_2}}$ respectively, so that $e = \sigma\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2}$ …(3.1)

If n1 and n2 be sufficiently large than $\bar{x}_1 \sim \bar{x}_2$ is asymptotically normally distributed with mean zero and standard deviation e. Consequently, if the difference $\bar{x}_1 \sim \bar{x}_2$ exceeds 3e the difference can hardly be accounted for by the fluctuations of sampling and our assumption unlikely to be correct while if difference exceeds 2e, it is regarded as significant at the 5% level of probability.

If two independent samples of n1 and n2 members respectively be drawn from different data sets with variances σ_1^2 and σ_2^2 respectively we can examine whether the two data sets from which samples have been drawn differ in mean apart from the difference in dispersion. Since the samples are independent the s.e. e of the difference of their means is given by $e^2 = \left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$ …(3.2)

Assuming that n1 and n2 are large and the two data sets have the same means, the difference of the means of the samples will be normally distributed with mean zero and s.d. e given by (2). If the difference of the means of the samples exceeds 3e, it can hardly be accounted for on the basis of fluctuations of sampling and our assumption that the two data sets have the same mean is almost certainly wrong.

In the above discussion the following assumption have been considered:

1. I assumed that the data set variance σ_1^2, σ_2^2 are known. In practice this is hardly the case and accordingly in the expressions for e, these have to be replaced by their estimated obtained for the samples, viz., by the sample variances s_1^2 and s_2^2 respectively, where $s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(x_i - \bar{x}_j\right)^2 \; j = 1,2$.

2. The above tests are valid for only large samples for two reasons:

(i) The parent data sets may not be normal, though we are assuming that they do not depart strikingly from it. In particular, we assume that the data sets of finite variances. For data sets like Cauchy's where the variance is not finite. the tests would break down completely even for infinitely large samples.

(ii) The data set variances are not known and have to be replaced by their estimates.

3. For normal data sets with known variances, the above tests are valid for all sample sizes.

4. If the hypothesis to be tested is that the data set means are μ and μ', we can carry out the test of significance as above, but in this case

$$z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) \sim \left(\mu - \mu'\right)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad \ldots(3.3)$$

will be asymptotically a standard normal variate for large n_1, n_2.

4. Experimental Analysis:

Table 1: Test of significance between $\bar{Y}$ (without missing) and $\bar{Y}$ (after imputation by mean method)

S.No.	missing %	$\bar{Y}$ (without missing)	S.D. (without)	$\bar{Y}$ (mean)	S.D. (mean)	S.E.	Z-test
1	1.00	42.4850	13.9677	42.5152	13.9622	0.3737	-0.0807
2	2.00	42.4850	13.9677	42.3061	13.7705	0.3724	0.4803
3	4.00	42.4850	13.9677	42.5781	13.8953	0.3732	-0.2495
4	6.00	42.4850	13.9677	42.5638	13.3364	0.3695	-0.2133
5	8.00	42.4850	13.9677	42.5978	13.5484	0.3709	-0.3042
6	10.00	42.4850	13.9677	41.9667	13.2129	0.3687	1.4060
7	12.00	42.4850	13.9677	42.8239	13.0783	0.3677	-0.9215
8	14.00	42.4850	13.9677	42.5523	13.0947	0.3678	-0.1830
9	16.00	42.4850	13.9677	42.6429	12.9134	0.3666	-0.4306
10	18.00	42.4850	13.9677	42.3537	12.7720	0.3656	0.3592
11	20.00	42.4850	13.9677	41.7125	12.2484	0.3621	2.1337
12	22.00	42.4850	13.9677	42.2115	12.2072	0.3618	0.7559
13	24.00	42.4850	13.9677	42.4868	12.1750	0.3615	-0.0051
14	26.00	42.4850	13.9677	42.6284	11.9687	0.3601	-0.3981
15	28.00	42.4850	13.9677	42.6111	11.6830	0.3581	-0.3521
16	30.00	42.4850	13.9677	42.1643	11.1876	0.3546	0.9043
17	32.00	42.4850	13.9677	42.5646	12.0429	0.3606	-0.2208
18	34.00	42.4850	13.9677	42.4091	10.4010	0.3491	0.2175
19	36.00	42.4850	13.9677	42.2188	11.1057	0.3541	0.7520
20	38.00	42.4850	13.9677	41.8710	10.9260	0.3528	1.7405
21	40.00	42.4850	13.9677	44.0750	10.2090	0.3477	-4.5731

Table 2 : Test of significance between $\bar{Y}$ (without missing) and $\bar{Y}$ (after imputation by ratio method)

S.No.	missing %	$\bar{Y}$ (without missing)	S.D. (without)	$\bar{Y}$ (ratio)	S.D. (ratio)	S.E.	Z-test
1	1.00	42.4850	13.9677	42.5149	13.9622	0.3737	-0.0801
2	2.00	42.4850	13.9677	42.3099	13.7705	0.3724	0.4703
3	4.00	42.4850	13.9677	42.5709	13.8954	0.3732	-0.2302
4	6.00	42.4850	13.9677	42.5542	13.3365	0.3695	-0.1873
5	8.00	42.4850	13.9677	42.5850	13.5485	0.3709	-0.2695
6	10.00	42.4850	13.9677	42.0367	13.2146	0.3687	1.2160
7	12.00	42.4850	13.9677	42.7998	13.0785	0.3677	-0.8559
8	14.00	42.4850	13.9677	42.5684	13.0947	0.3678	-0.2268
9	16.00	42.4850	13.9677	42.6550	12.9135	0.3666	-0.4637
10	18.00	42.4850	13.9677	42.3573	12.7720	0.3656	0.3492
11	20.00	42.4850	13.9677	41.8163	12.2502	0.3621	1.8470
12	22.00	42.4850	13.9677	42.2417	12.2073	0.3618	0.6726

13	24.00	42.4850	13.9677	42.5645	12.1758	0.3615	-0.2200
14	26.00	42.4850	13.9677	42.5171	11.9701	0.3601	-0.0893
15	28.00	42.4850	13.9677	42.7066	11.6841	0.3581	-0.6187
16	30.00	42.4850	13.9677	42.3731	11.1921	0.3547	0.3154
17	32.00	42.4850	13.9677	42.2630	12.0476	0.3607	0.6156
18	34.00	42.4850	13.9677	42.4861	10.4015	0.3491	-0.0031
19	36.00	42.4850	13.9677	42.2182	11.1057	0.3541	0.7534
20	38.00	42.4850	13.9677	42.1737	10.9328	0.3528	0.8824
21	40.00	42.4850	13.9677	43.5799	10.2270	0.3478	-3.1479

Table 3 : Test of significance between $\overline{Y}$ (without missing) and $\overline{Y}$ (after imputation by Compromise method)

S.No.	missing %	$\overline{Y}$ (without missing)	S.D. (without)	$\overline{Y}$ (comp)	S.D. (comp)	S.E.	Z-test
1	1.00	42.4850	13.9677	42.5150	13.9622	0.3737	-0.0802
2	2.00	42.4850	13.9677	42.3090	13.7705	0.3724	0.4727
3	4.00	42.4850	13.9677	42.5726	13.8953	0.3732	-0.2347
4	6.00	42.4850	13.9677	42.5565	13.3365	0.3695	-0.1936
5	8.00	42.4850	13.9677	42.5880	13.5484	0.3709	-0.2778
6	10.00	42.4850	13.9677	42.0214	13.2140	0.3687	1.2575
7	12.00	42.4850	13.9677	42.8051	13.0784	0.3677	-0.8704
8	14.00	42.4850	13.9677	42.5649	13.0947	0.3678	-0.2172
9	16.00	42.4850	13.9677	42.6525	12.9134	0.3666	-0.4568
10	18.00	42.4850	13.9677	42.3565	12.7720	0.3656	0.3514
11	20.00	42.4850	13.9677	41.7926	12.2495	0.3621	1.9125
12	22.00	42.4850	13.9677	42.2353	12.2073	0.3618	0.6902
13	24.00	42.4850	13.9677	42.5487	12.1755	0.3615	-0.1761
14	26.00	42.4850	13.9677	42.5417	11.9696	0.3601	-0.1574
15	28.00	42.4850	13.9677	42.7033	11.6807	0.3581	-0.6097
16	30.00	42.4850	13.9677	42.3252	11.1903	0.3547	0.4507
17	32.00	42.4850	13.9677	42.4589	12.1812	0.3616	0.0721
18	34.00	42.4850	13.9677	42.4701	10.4013	0.3491	0.0426
19	36.00	42.4850	13.9677	42.2184	11.1057	0.3541	0.7531
20	38.00	42.4850	13.9677	42.1205	10.9306	0.3528	1.0330
21	40.00	42.4850	13.9677	43.7278	10.2178	0.3477	-3.5739

5. Conclusions:

This This work analyses the behavior of four imputation methods used for missing data treatment. These methods are analyzed on different percentages of missing data into a common attribute of large data sets. The Ratio method provides very good results, even for training sets having a large amount of missing data. In case of mean method of imputation, only at 24% level of missing data, critical value of z score i.e. 0.0051 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent. In case of ratio method of imputation, only at 34% level of missing data, critical value of z score i.e. 0.0031 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent.

In case of compromise method of imputation, only at 34% level of missing data, critical value of z score i.e. 0.0426 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent. In case of Ahmed method of imputation, only at 34% level of missing data, critical value of z score i.e. 0.0427 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent.

Although, all the methods are showing approximately correct results at different percentages of missing data but when we compare results of all the methods on same data set, outcome given by ratio method of imputation are more accurate among all. Hence, it may be recommended for imputing the missing values to preprocess the database prior to analysis, so that the quality of the results extracted can be improved.

In future works, the missing data treatment methods canl be analyzed in some other live data sets having attributes other than numeric. Furthermore, in this work missing values were inserted completely at random (MCAR). In a future work, one can analyze the behavior of these methods along with some additional methods when missing values are not randomly distributed. In this case, there is a possibility of creating invalid knowledge. For an effective analysis, we will have to inspect not only the error rate, but also the quality of the knowledge induced by the learning system.

REFERENCES

[1] Cochran, W. G. 2005. Sampling Techniques, John Wiley and Sons, New York.

[2] G. E. A. P. A. Batista and M. C. Monard. K-Nearest Neighbour as Imputation Method 2002. Experimental Results. Technical report, ICMC-USP, ISSN-0103-2569.

[3] Heitjan, D. F. and Basu, S. 1996. Distinguishing 'Missing at random' and 'missing completely at random', The American Statistician, 50, 207-213.

[4] J. W. Grzymala-Busse and M. Hu. A Comparison of Several Approaches to Missing Attribute Values in Data Mining 2000. In RSCTC'2000, pages 340–347.

[5] K. Lakshminarayan, S. A. Harp, and T. Samad. 1999. Imputation of Missing Data in Industrial Databases. Applied Intelligence, 11:259–275.

[6] R. J. Little and D. B. Rubin. 1987. Statistical Analysis with Missing Data. John Wiley and Sons, New York, 1987.

[7] Rao, J. N. K. and Sitter, R. R. 1995. Variance estimation under two-phase sampling with application to imputation for missing data, Biometrica, 82, 453-460.

[8] Reddy, V. N. 1978. A study on the use of prior knowledge on certain population parameters in estimation, Sankhya, C, 40, 29-37.

[9] Rubin, D. B. 1976. Inference and missing data, Biometrica, 63, 581-593.

[10] Shukla, D. 2002. F-T estimator under two-phase sampling, Metron, 59, 1-2, 253-263.

[11] Shukla, D. and Thakur, N. S. 2008. Estimation of mean with imputation of missing data using factor-type estimator, Statistics in Transition, 9, 1, 33-48.

[12] Thakur, N. S., Yadav Kalpana, and Pathak S. 2012. Some imputation methods in double sampling scheme for estimation of population mean, IJMER, Vol.2, Issue.1 Jan-Feb 2012 pp-200-207.

[13] Thakur, N. S., Yadav Kalpana, and Pathak S. 2011.Estimation of mean in presence of missingdata under two-phase sampling scheme, JRSS,Vol 4, issue 2,93-104.

[14] Singh, S. 2009. A new method of imputation in survey sampling, Statistics, Vol. 43, 5 , 499 - 511.

[15] Singh, S. and Horn, S. 2000. Compromised imputation in survey sampling, Metrika, 51, 266-276.

[16] Singh, V. K. and Shukla, D. 1993. An efficient one parameter family of factor - type estimator in sample survey, Metron, 51, 1-2, 139-159.

ISSN 2277-3061

Author' biography with Photo

Dr. Rahul Singhai has obtained M.C.A.degree from H.S. Gour University, Sagar,MP, in 2001 and obtained M.Phil degree in Computer Science from Madurai Kamaraj University, Madurai, Tamilnadu in 2008.The Ph.D degree in Computer Science was awarded in 2011 by Sagar university. He worked as contract lecturer at Deptt. Of Computer Sc, Dr. H.S.G. Central University, sagar for more than 5 years. In July 2009 he joined IIPS, Devi Ahilya University, Indore as permanent Lecturer and Presentally he is serving as Assistant Professor (Senior Scale) at the same department. His area of research are Computer Network, Data mining, DBMS & operating System. He has authored and co-authored 18 research papers in international/national journals and conference proceeding. Currently, he is working on to develop new probability based methods for data preprocessing in data mining. He is the member of various academic & professional bodies/socities.