International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

A Survey on Introduction to Missing Values Imputation in Data Mining

N. Amirtha Gowri
Assistant Professor, Department of IT, Sri Subash Arts and Science College, Pollachi, India

***Abstract*: Many existing, industrial and research data sets contain Missing Values. They are introduced due to various reasons, such as manual data entry procedures, equipment errors and incorrect measurements. Hence, it is usual to find missing data in most of the information sources used. The detection of incomplete data is easy in most cases, looking for Null values in a data set. However, this is not always true, since Missing Values (MVs) can appear with the form of outliers or even wrong data (i.e. out of boundaries). Missing values make it difficult for analysts to perform data analysis. Three types of problems are usually associated with missing values.**

1) Loss of efficiency; 2) Complications in handling and analyzing the data; 3) Bias resulting from differences between missing and complete data.

As little and Rubin stated, there exist three different mechanism of missing data induction.

1) Missing completely at random (MCAR), when the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data. 2) Missing at random (MAR), when the distribution of an example having a missing value for an attribute depends on the observed data, but does not depend on the missing data. 3) Not missing at random (NMAR), when the distribution of an example having a missing value for an attribute depends on the missing values.

***Keywords*: Imputation Methods, Single Imputation–Types of Single Imputation Methods, Conclusion, References.**

1. Introduction

A. Imputation methods

Types of imputation

There are two types of imputation

- Single
- Multiple.

B. Single imputation

Single imputation denotes that the missing value is replaced by a value. In this method the sample size is retrieved. However, the imputed values are assumed to be the real values that would have been observed when the data would have been complete. When we have missing data, this is never the case. We can never be completely certain about imputed values. Single refers to the fact that you come up with a single estimate of the missing value, using one of the seven methods listed below. It's popular because it is conceptually simple and because the resulting sample has the same number of observations as the full data set. Single imputation looks very tempting when list wise deletion eliminates a large portion of the data set.

2. Types of single imputation

- Mean imputation
- Substitution
- Hot deck imputation
- Cold deck imputation
- Regression imputation
- Stochastic regression imputation
- Interpolation and extrapolation

A. Mean imputation

- Simply calculate the mean of the observed values for that variable for all individuals who are non-missing.
- It has the advantage of keeping the same mean and the same sample size, but many, many disadvantages.
- Pretty much every method listed below is better than mean imputation.

B. Substitution

- Impute the value from a new individual who was not selected to be in the sample.
- Find a new subject and use their value instead.

C. Hot deck imputation

- A randomly chosen value from an individual in the sample who has similar values on other variables.
- Find all the sample subjects who are similar on other variables, then randomly choose one of their values on the missing variable.
- One advantage is you are constrained to only possible values. Another is the random component, which adds in some variability. This is important for accurate standard errors.

D. Cold deck imputation

- A systematically chosen value from an individual who has similar values on other variables.
- This is similar to Hot Deck in most ways, but removes the random variation.
- For example choose always the third individual in the same experimental condition and block.

E. Regression imputation

- The predicted value obtained by regressing the missing variable on other variables.
- The predicted value, based on other variables.
- This preserves relationships among variables involved in the imputation model.

F. Stochastic regression imputation

- The predicted value from a regression plus a random residual value.
- This has all the advantages of regression imputation but adds in the advantages of the random component.
- Most multiple imputations are based off of some form of stochastic regression imputation.

G. Interpolation and extrapolation

- An estimated value from other observations from the same individual.
- It usually only works in longitudinal data.
- Use caution, though. Interpolation, for example, might make more sense for a variable like height in children–one that can't go back down over time.

3. Conclusion

Single imputation involves less computation, and provides the dataset with a specific number in place of the missing data. While there is more than one type of single imputation, in general the process involves analyzing the other responses and looking for the most likely (or a set of the most likely) responses the individual would have answered, and then picks one of those possible responses at random and places it in the dataset. When only a little bit of data is missing, single imputation provides a useful enough tool. It fills in the data points well and the variance between the results of the analyses is unlikely to be altered by any significant margin.

References

[1] Analysis of Variance and Covariance
[2] Complex Surveys & Sampling
[3] Count Regression Models
[4] Effect Size Statistics, Power, and Sample Size Calculations
[5] Linear Regression
[6] Logistic Regression
[7] Missing Data
[8] Mixed and Multilevel Models
[9] Principal Component Analysis and Factor Analysis
[10] R
[11] SPSS
[12] Stata
[13] Survival Analysis and Event History Analysis
[14] Introduction to Missing Values Imputation in Data Mining
[15] Imputation Methods
[16] Technical description report
[17] On the suitability of imputation methods for different learning approaches
[18] Quantifying the effects of the imputation methods in the noise and information contained in the data set
[19] Data-sets partitions employed in the papers
[20] Complementary material of the papers
[21] WEB sites devoted to Missing Values
[22] Missing Values Bibliography