Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Is Group Means Imputation Any Better Than Mean Imputation: A Study Using C5.0 Classifier

To cite this article: Faizan U F Khan *et al* 2018 *J. Phys.: Conf. Ser.* **1060** 012014

Is Group Means Imputation Any Better Than Mean Imputation: A Study Using C5.0 Classifier

Faizan U F Khan[1], Kashan U Z Khan[2] and S K Singh[1]

[1] Department of Computer Science and Engineering, IIT (BHU), Varanasi, 221005, India
[2] Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia, New Delhi, 110025, India

fufahad.khan.cse14@iitbhu.ac.in

Abstract. Since most data-driven systems including classifiers require large amounts of complete data, the task of handling missing data has garnered much attention. If one of the variables under study in a dataset has some incomplete values, it is treated as a missing data problem. Various methods in the literature exist for dealing with missing data including complete case analysis, listwise deletion, single imputation and multiple imputations. Out of these, mean imputation remains a favourite for researchers due to its simplicity and ease of use, despite some glaring flaws. In this paper, we compare Mean imputation with a similar single imputation method – Group Means imputation – and present our results on nine real-world datasets with respect to classifier accuracy of the C5.0 classifier on the imputed dataset. We show that while Group Means imputation fares better on training data, the test set accuracies fall in favour of Mean Imputation, which deals with novel data in a much better fashion.

1. Introduction

In recent times, data-driven systems have seen a great boost in terms of application and have had a large impact in the field of computer technology. However, the data initially collected is rarely optimal and need to go through a data preprocessing phase before it can be utilized to mine knowledge. [1][2][3] Data Imputation is a preprocessing method that deals with the inconsistency of missing data. Missing data inconsistency is said to be present in a dataset when one or more variables under study have an absence of values [4].

Missing data can render decision-making systems including neural networks and Support Vector Machines useless as they require complete data to produce output. In such cases, it becomes essential to estimate the missing data which can be fed to the system to extract knowledge. Care should be taken while estimating the missing data as the goal while estimating data should be to make efficient and valid inferences about the sample data rather than to predict, estimate or retrieve missing data or to get the similar output that is with complete data. [5] Data Imputation has emerged as a valid candidate for the problem of missing data. The literature consists of various imputation methods including Mean/Mode Imputation, Group Mean Imputation, Predictive Mean Matching, Regression Imputation and Multiple Imputation using Chained Equations (MICE). [6][7] While many alternatives exist, Mean/Mode Imputation continues to be the most widely abused form of imputation, despite its various flaws. This is due to its simplicity and applicability to all cases, and the fact that it is time optimal.

In this paper, we test the efficiency of Mean Imputation versus a modified version of Mean Imputation – known as Group Mean Imputation – and see if it offers a better and simple alternative to Mean Imputation with respect to classifier accuracy. We use real-world datasets taken from the UCI Machine Learning Repository [8] and artificially introduce missing values to obtain imputed datasets. These datasets are then used to build a C5.0 classifier which is used to evaluate the quality of imputation of the different methods. [9]

2. Preliminaries

The objective of imputation is to replace the missing values by plausible values in order to exploit the information in the recorded variables in the incomplete cases for inference about population parameters. Complete data (Y) can be defined as a function of observed data (Y_{obs}), and missing data (Y_{mis}),

$$Y = (Y_{obs}, Y_{mis}) \tag{1}$$

2.1. Missing data mechanisms

Knowing the reason why the data are missing is beneficial in choosing the right technique to approximate missing data. [10] Generally, we have four missing data mechanisms [7][10], which dictate why the data is missing.

2.1.1. Missingness Completely At Random. This type of missingness does not bias the inferences. The probability of a value of a variable being missing is completely random and unrelated to any other values of other variables, whether missing or observed. There is no observable pattern in missing data and cases with complete data are indistinguishable from cases with incomplete data.

2.1.2. Missingness At Random. In this type of missingness, the cause of the missing data is related to other observed variables but unrelated to the missing values themselves. Also known as the *ignorable case.* We can predict the missing values based on observed values of other variables.

2.1.3. Missingness that depends on unobserved predictors. Missingness is no longer “at random” if it depends on information that has not been recorded and this information also predicts the missing values. This type of missingness must be explicitly modelled, or else bias can occur in the inferences

2.1.4. Missingness that depends on the missing value itself. The instance of the missing data depends on the unobserved data of the variable itself. Particularly hard to handle, it implies that the missing data mechanism is related to the missing values.

3. Mean Imputation and Group Mean Imputation

3.1. Mean/Mode Imputation

One of the most common methods of imputation, in this method, the mean of all values within the same attribute is calculated and then imputed in the missing data cells. If the attribute is nominal, mode substitution can be used instead. Mean substitution introduces bias in the parameter estimates as a result of loss in standard deviation, even though the mean of the data is generally preserved. [7] This is because we substitute all missing values with a single measure of central tendency.

$$A_i^j = \sum_{k\in(complete)} A_k^j / n_{|I(complete)|} \tag{2}$$

3.2. Group-Means Imputation

Group means substitution differs slightly from the normal mean/mode substitution. In this method, the missing values are replaced with the group (or class) mean of all known values of that attribute. [11] While the class wise imputation technique will impute multiple values, it will still underestimate the standard deviation and will lead to biased parameter estimates. As in normal Mean Imputation, an alternate measure of central tendency can be used if calculation of mean is not possible.

4. Experiment

The details of the experiment including the datasets used, the ratio and type of missingness and the classifier details are discussed further. The whole experiment was carried out in the R programming language.

4.1. Datasets

We applied the imputation methods on nine different datasets taken from the UCI Machine Learning Repository [8]. The details of the datasets are given in table 1.

Table 1. Datasets used in the experiments.

Dataset Name	Number of Categorical Attributes	Number of Ordinal/Continuous Attributes	Number of Cases
Breast Cancer Wisconsin	1	9	699
Car	7	-	1728
Credit	10	6	690
Cleveland	9	5	303
Ionosphere	1	34	351
Iris	1	4	150
Spambase	1	57	4601
Titanic	3	1	1313
Wine Quality	1	11	4898

For the *Wine Quality* dataset, we converted the continuous *quality* variable to an ordinal variable with three categories depicting the quality of the wine: *"good"* (>7), *"average"* (>4 & <7), *"bad"* (<4).

The whole dataset was divided in the training-test ratio of 70:30. Missing values were artificially introduced in each of the nine different datasets, with three different missing ratios of 30%, 40% and 50%. For each missing ratio, three different incomplete datasets were created, resulting in nine different variations of each dataset. The results were reported after aggregating the results of the imputed datasets to avoid sampling bias.

4.2. C5.0 classifier details

The C5.0 algorithm derives from the C4.5 algorithm, but is faster and less error prone. [9][12] It has some additions over the previous algorithm including generation of smaller decision trees, generation of less error prone rules, winnowing and the ability to incorporate boosting. In our experiments, we generated rule based models and set the number of boosting iterations to 10. Using adaptive boosting, we generate several classifiers rather than just one. When a new case is to be classified, each classifier votes for its predicted class and the votes are counted to determine the final class. We use the evaluation criteria similar to that proposed in [13] in our experiments. The results include the accuracy of the C5.0 classifier on both the training set and the test set. We used the R implementation of the C5.0 decision tree algorithm as provided in the R package *C50.* [14]

5. Results

Table 2. Classification accuracy of the C5.0 classifier with datasets having three different missing ratios. Bold values indicate a higher accuracy value than the other method.

		Training Set		Test Set	
Dataset	Missing Ratio	Mean	Group Means	Mean	Group Means

Breast Cancer Wisconsin	30	99.90	**100.0**	95.71	**96.03**
	40	99.33	**100.0**	**95.56**	95.24
	50	99.60	**100.0**	**96.51**	95.08
Car	30	82.47	**97.37**	72.32	**87.28**
	40	82.27	**97.33**	71.29	**87.86**
	50	82.07	**96.63**	71.29	**87.09**
Cleveland	30	85.43	**97.43**	**84.06**	76.92
	40	87.37	**98.53**	**83.73**	77.65
	50	87.37	**98.40**	**80.52**	77.65
Credit	30	95.43	**100.0**	79.48	**83.57**
	40	93.40	**100.0**	77.65	**82.13**
	50	92.80	**99.83**	77.65	**84.38**
Ionosphere	30	100.0	100.0	**94.66**	90.88
	40	99.73	**100.0**	**92.45**	91.82
	50	100.0	100.0	**95.28**	92.14
Iris	30	89.53	**98.70**	**97.78**	94.82
	40	89.50	**98.70**	**95.56**	91.85
	50	90.80	**98.40**	**94.07**	91.85
Spambase	30	91.43	**99.80**	**92.91**	91.43
	40	90.47	**99.77**	**93.17**	90.30
	50	91.07	**99.80**	**92.88**	90.49
Titanic	30	74.07	**96.30**	78.93	**79.19**
	40	73.37	**95.87**	76.99	**80.20**
	50	75.33	**95.90**	**79.19**	78.43
Wine Quality	30	81.97	**100.0**	**82.02**	75.65
	40	81.73	**99.93**	**82.18**	78.07
	50	84.83	**100.0**	**81.70**	77.80

The experiments showed varying results for both the imputation methods. We found Group Means Imputation to be almost exclusively better than normal Mean Imputation on all datasets when training set results were compared, as shown in table 2. The Group Means method posted training accuracies of above 96% for all datasets, even when the missing data ratio was 50%. However, this was not reflected in the test data accuracies for Group Means Imputation, which dropped significantly to about 77% for the *Wine Quality* and *Cleveland* datasets, where the training set accuracies were greater than 97%.

Mean Imputation on the other hand posted relatively low training set accuracies as compared to Group Means Imputation, but outperformed Group Means Imputation when test set accuracies were considered. It was better in six of the nine datasets for different missingness ratios. Group Means was better with respect to classification accuracy for the *Car*, *Credit* and *Titanic* datasets for all missingness ratios except for 50% missingness in the *Titanic* dataset. We observe that in all these cases, the number of categorical variables was larger than the number of continuous/ordinal variables, which may have been a factor for these observations. This might be due to the fact that we use mode substitution since mean is not defined for categorical variables.

For *Titanic*, *Spambase* and *Iris* datasets, Mean Imputation reported better accuracies with the test set than the training set, generalizing better on unseen data. This was not the case for Group Means which generalized very poorly for novel data.

6. Conclusion

Our study shows that while both Mean Imputation and Group Means Imputation use similar methodology to handle missing data, Mean Imputation is better suited for the task as it generalizes well for novel data. High classification accuracies for training set for Group Means Imputation is misleading and may result in worse results when dealing with novel data.

However, when the number of categorical variables was larger than the number of continuous/ordinal variables, Group Means Imputation outperformed normal Mean/Mode Imputation in both training and test set accuracies. This should be taken into account when selecting the better of the two methods, since Mode substitution due to a large number of categorical variables seems to reduce accuracy. The ratio of missingness had no observable correlation with the accuracies for both the methods. In addition, the size of the dataset was not a factor which may differentiate between the two methods.

References

[1] Larose DT 2014 *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons

[2] Han J, Pei J and Kamber 2011 M. *Data Mining: Concepts and Techniques*. Elsevier

[3] Magnani M 2007 Techniques for dealing with missing data in knowledge discovery tasks. Obtido http://magnanim. web. cs. unibo. it/index. html.

[4] Bannon W 2015 Missing data within a quantitative research study: How to assess it, treat it, and why you should care. *Journal of the American Association of Nurse Practitioners* **27**(4) 230-2.

[5] Chhabra G, Vashisht V and Ranjan J 2017 A comparison of multiple imputation methods for data with missing values *Indian Journal of Science and Technol* **10**(19).

[6] Scheffer J. Dealing with missing data.

[7] Cheema JR 2014 A review of missing data handling methods in education research *Review of Educational Research* **84**(4) 487-508.

[8] Dua, D. and Karra Taniskidou, E. 2017 UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Pandya R and Pandya 2015 J. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning *International Journal of Computer Applications* **117**(16).

[10] Nelwamondo F, and Marwala T 2008 Key issues on computational intelligence techniques for missing data imputation-a review. *In Proc. of world multi conf. on systemics, cybernetics and informatics* **4** 35-40).

[11] Sim J, Lee JS and Kwon O 2015 Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Mathematical Problems in Engineering*.

[12] Quinlan JR 2014 C4. 5: programs for machine learning. Elsevier.

[13] Aljuaid T and Sasi S 2016 Intelligent imputation technique for missing values. *In Advances in Computing, Communications and Informatics (ICACCI)*, (pp. 2441-2445). IEEE

[14] Kuhn M, Weston S, Coulter N and Quinlan R. 2014 C50: C5. 0 decision trees and rule-based models. R package version 0.1. 0-21, URL http://CRAN. R-project. org/package C. 2014;50.