ISSN:2231-0711

A Combined Method to Impute Missing Data and Predict Accurate Value of a Target Variable in Supervised Machine Learning

Deepak[1], Ms. Sonia Batra[2]

Department of Computer Science and Engineering
World College of Technology and Management
Gurugram, MDU(Rohtak), Haryana, India

dc124108@gmail.com , sonia807arora@gmail.com

***Abstract*-- An excellent deal of recent method analysis has targeted on two trendy missing data analysis methods: maximum likelihood and multiple imputation. These methods are very good to traditional techniques (e.g. deletion and mean imputation methods etc.) because they require less stringent assumptions and mitigate the pitfalls of traditional techniques. This article explains the new theoretical and practical methods of missing data analyses in big datasets , gives an overview of traditional missing data techniques, predict more accurate value of target variable and provides accessible descriptions of maximum likelihood and multiple imputation. Finally, the paper illustrates ways in which researchers will use intentional, or planned, missing data to enhance their research designs.**

I. Introduction

Missing data are the problematic thing that we face during the training of a model in machine learning. Missing values are representative of the messiness of real world data. There can be a multitude of reasons why they occur — ranging from human errors during data entry, incorrect sensor readings, to software bugs in the data processing pipeline. The normal reaction is frustration. Missing data are probably the most widespread source of errors in your code, and the reason for most of the exception-handling. If you try to remove them, you might reduce the amount of data you have available dramatically — probably the worst that can happen in machine learning. There are three main types of missing data. Missing completely at random (MCAR), Missing at random (MAR) and Not missing at random (NMAR). We use traditional methods mean, median, KNN etc. to impute these missing data which is not a correct way of imputation in modern datasets. Still, often there are hidden patterns in missing data points. Those patterns can provide additional insight in the problem you're trying to solve. One way to handle this problem is to get rid of the observations that have missing data. However, you will risk losing data points with valuable information. A better strategy would be to impute the missing values. In other words, we need to infer those missing values from the existing part of the data. Hence, machine learning techniques can been quite effective in filling these missing data.

A. *Methods to handling Missing Data*:

The basic fundamental steps which are used in handling missing data are given below:

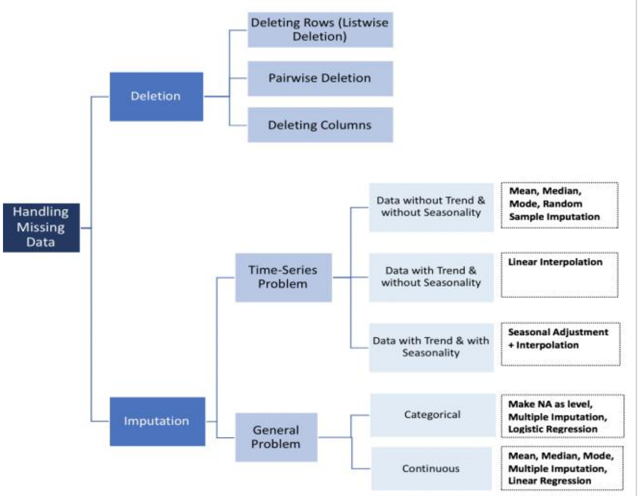
Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

Fig. 1.1: Traditional ways to handle missing data

Imputation vs Removing Data

Before jumping to the methods of data imputation, we have to understand the reason why data goes missing.

Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data

Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations. Note that imputation does not necessarily give better results.

B. Deletion

Listwise:

Listwise deletion (complete-case analysis) removes all data for an observation that has one or more missing values. Particularly if the missing data is limited to a small number of observations, you may just opt to eliminate those cases from the analysis. However in most cases, it is often disadvantageous to use listwise deletion. This is because the assumptions of MCAR (Missing Completely at Random) are typically rare to support. As a result, listwise deletion methods produce biased parameters and estimates.

Pairwise:

Pairwise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. A strength to this technique is that it increases power in your analysis but it has many disadvantages. It assumes that the missing data are MCAR. If you delete pairwise then you'll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult.

Dropping Variables:

In my opinion, it is always better to keep data than to discard it. Sometimes you can drop variables if the data is missing for more than 60% observations but only if that variable is insignificant. Having said that, imputation is always a preferred choice over dropping variables.

C. Mean, Median and Mode

Computing the overall mean, median or mode is a very basic imputation method, it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. It is very fast, but has clear disadvantages. One disadvantage is that mean imputation reduces variance in the dataset.

D. Linear Regression

To begin, several predictors of the variable with missing values are identified using a correlation matrix. The best predictors are selected and used as independent variables in a regression equation. The variable with missing data is used as the dependent variable. Cases with complete data for the predictor variables are used to generate the regression equation; the equation is then used to predict missing values for incomplete cases. In an iterative process, values for the missing variable are inserted and then all cases are used to predict the dependent variable. These steps are repeated until there is little difference between the predicted values from one step to the next, that is they converge.

It "theoretically" provides good estimates for missing values. However, there are several disadvantages of this model which tend to outweigh the advantages. First, because the replaced values were predicted from other variables they tend to fit together "too well" and so standard error is deflated. One must also assume that there is a linear relationship between the variables used in the regression equation when there may not be one.

E. Multiple Imputation

Imputation: Impute the missing entries of the incomplete data sets *m* times (*m*=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn't include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carlo (MCMC) simulation. This step results in m complete data sets.

Analysis: Analyze each of the *m* completed data sets.

Pooling: Integrate the *m* analysis results into a final result

Imputation of Categorical Variables

- Mode imputation is one method but it will definitely introduce bias.
- Missing values can be treated as a separate category by itself. We can create another category for the missing values and use them as a different level. This is the simplest method.
- Prediction models: Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test). We can use methods like logistic regression and ANOVA for prediction
- Multiple Imputation

F. KNN (K Nearest Neighbors)

In this method, k neighbors are chosen based on some distance measure and their average is used as an imputation estimate. The method requires the selection of the number of nearest neighbors, and a distance metric. KNN can predict both discrete attributes (the most frequent value among the k nearest neighbors) and continuous attributes (the mean among the k nearest neighbors) The distance metric varies according to the type of data:

Continuous Data: The commonly used distance metrics for continuous data are Euclidean, Manhattan and Cosine

Categorical Data: Hamming distance is generally used in this case. It takes all the categorical attributes and for each, count one if the value is not the same between two points. The Hamming distance is then equal to the number of attributes for which the value was different.

II. PROPOSED WORK

A. Problem Statement

One of the most common problems I everyone faced in Data Cleaning/Exploratory Analysis is handling the missing values. Firstly, understand that there is NO good way to deal with missing data. I have come across different solutions for data imputation depending on the kind of problem — Time series Analysis, ML, Regression etc. and it is difficult to provide a general solution. In this work, I am attempting to summarize the most commonly used methods and trying to find a structural solution for missing data.

B. Research Gaps

After studying lots of blogs, articles and research paper, I identified that there are any proper way or any solid method that can applied on missing data and can fill them with correct value.

Every paper gives their own theory and methods but didn't give the ultimate method to overcome the missing data problem. In this work I found a way to predict the continuous value of a target variable using supervised machine learning algorithm.

C. Objectives:

The main objectives of the study are as follow:

- Apply new methods to impute missing data of a dataset.
- Comparative study of different algorithms used in Supervised machine learning.
- Apply multiple methods to impute missing data to increase prediction accuracy in supervised Machine Learning.

D. Research Methodology

Data Gathering

Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The most critical objective of data collection is ensuring that information-rich and reliable data is collected for statistical analysis so that data-driven decisions can be made for research. Data collection is an important aspect of research. Let's consider an example of a mobile manufacturer, company X, which is launching a new product variant. To conduct research about features, price range, target market, competitor analysis etc. data has to be collected from appropriate sources.

Data Preparation and Analysis

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets.
- Apply mean and KNN both for imputation by increasing the dimension of the dataset.
- Check the accuracy score after training the model.
- Find the difference between traditional ways of imputation and newly methods.
- Conclude the result.

Tools

- Anaconda
- jupyter notebook
- Scikit Learn
- Pandas
- Python

III. EXPERIMENTAL RESULTS

A. Dataset Description

On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

The titanic.csv file contains data for 887 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived (SS), their age (AA), their passenger-class (CC), their sex (GG) and the fare they paid (XX).

This is a classic dataset used in many data mining tutorials and demos -- perfect for getting started with exploratory analysis and building binary classification models to predict survival.

Data covers passengers only, not crew.

Features

1. survival - Survival (0 = No; 1 = Yes)
2. class - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
3. name - Name
4. sex - Sex
5. age – Age(missing value data)
6. sibsp - Number of Siblings/Spouses Aboard
7. parch - Number of Parents/Children Aboard
8. ticket - Ticket Number
9. fare - Passenger Fare
10. cabin - Cabin
11. embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
12. boat - Lifeboat (if survived)
13. body - Body number (if did not survive and body was recovered)

B. Results and Discussion

After testing this combine method of KNN and mean for the imputation on missing values on the above dataset we can say that this method is more efficient in the classification problems contains missing features. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Below are few steps of Data Preprocessing. Remember, not all the steps are applicable for each problem, it is highly dependent on the data we are working with, so maybe only a few steps might be required with your dataset.

- Data Quality Assessment
- Feature Aggregation
- Feature Sampling
- Feature Selection
- Dimensionality Reduction
- Feature Encoding
- Feature Scaling

IV. CONCLUSION

This review highlights the inconsistent reporting of missing data in machine learning studies and the continuing use of inappropriate methods to handle missing data in the analysis. This study guidelines as a framework for authors so that the amount of missing data can be handled in appropriate way so that anyone can build a better machine learning model and predict values more accurately.

Future Scope

We can use other methods like linear regression , KNN and statistics. We are increasing the dimension of the dataset which can decrease the time complexity of our model, but it can also increase the accuracy.

So by trying different methods may work to fill missing values. And we can increase the accuracy of our model so that it can predict more accurate than a traditional methods.

REFERENCE

[1] An introduction to modern missing data analyses Amanda N. Baraldi *, Craig K. Enders. Arizona State University, United States Received 19 October 2009; accepted 20 October 2009

[2] Acuna E, Rodriguez C (2004) The treatment of missing values and its effect in the classifier accuracy. In: Banks D et al (eds) Classification, clustering and data mining applications. Springer, Berlin, pp 639–648

[3] Aittokallio T (2009) Dealing with missing values in large-scale studies: microarray data imputation and beyond. Brief Bioinform 11(2):253–264

[4] Armitage EG, Godzien J, Alonso-Herranz V, Lopez-Gonzalvez A, Barbas C (2015) Missing value imputation strategies for metabolomics data. Electrophoresis 36:3050–3060

[5] Aussem A, de Morais SR (2010) A conservative feature subset selection algorithm with missing data. Neurocomputing 73:585–590

[6] Aydilek IB, Arslan A (2012) A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks. Int J Innov Comput Inf Control 8(7):4705–4717

[7] Aydilek IB, Arslan A (2013) A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci 233:25–35

[8] Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J Sch Psychol 48:5–37

[9] Bras LP, Menezes JC (2007) Improving cluster-based missing value estimation of DNA microarray data. Biomol Eng 24:273–282

[10] Brock GN, Shaffer JR, Blakesley RE, Lotz MJ, Tseng GC (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinform 9:12–23

[11] Burgette LF, Reiter JP (2014) Multiple imputation for missing data via sequential regression trees. Am J Epidemiol 172(9):1070–1076

[12] Celton M, Malpertuy A, Lelandais G, de Brevern AG (2010) Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genom 11:15–30

[13] Chen X, Wei Z, Li Z, Liang J, Cai Y, Zhang B (2017) Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation. Knowl Based Syst 132:249–262

[14] Cheng KO, Law NF, Siu WC (2012) Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. Pattern Recogn 45:1281–1289

[15] Chiu C-C, Chan S-Y, Wang C-C, Wu W-S (2013) Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Syst Biol 7:S1

[16] Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage. Allison, P. D. (2002)