HANDLING MISSING DATA : MICE (A DATA MINING APPLICATION)

Akshaya R[1], Anushree G[2] ,Devaki Sai Mahitha[3],Madhura[4] , Shylaja B[5]

[1] *Final year student ,Department Of Computer Science ,DSATM ,Karanataka , India*
[2] *Final year student, Department Of Computer Science, DSATM , Karanataka ,India*
[3] *Final year student ,Department Of, Computer Science ,DSATM ,Karanataka, India*
[4] *Final year student, Department Of Computer Science, DSATM, Karanataka, India*
[5] *Asst.Professor, Department Of Computer Science, DSATM, Karanataka, India*

ABSTRACT

The aim of this project is to give the complete datasets by imputing the missing data values by multiple imputation techniques. On encountering the missing data values or non-applicable values the system automatically points out the missing values for the particular information or the column and enumerate the mean value and replaces the missed values with the same. This type of different imputations in different types of datasets can be carried out using R package MICE (Multivariate Imputations by Chained Equations). Imputation of categorical data is improved in order to bypass problems caused by perfect prediction.

Keyword : *- Index Terms-multiple imputation, non-applicable, R, MICE, chained equations, categorical data.*

1. INTRODUCTION

We come across many kinds of datasets which consists of huge amount of missing values which are emerged in industrial and research areas. This issue arises due to various reasons such as manual data entry procedures, equipment errors and incorrect quantities. This kind of scenario nurtures many of the problems such as loss of competence, difficulties in handling and examining the data, bias resulting from differences between missing and complete data. Even though many technologies have been introduced for missing data imputation still these problems are unaddressed. This was not considered as a major problem in late 1980's, however, this issue came into the public eye in early 1990's and multiple imputation techniques were developed. This whole issue falls under the domain of data mining which is the process of converting the raw data into useful information. On the other hand, it is very important to catalogue the missing values. There exists three different types of missing data such as, MCAR, MAR, NMAR.

1.1 MCAR

Missing Completely at Random (MCAR): when the scattering of an example having a missing value for an attribute does not depend on either the observed data or the missing data..

1.2 MAR

Missing at Random (MAR): when the scattering ofan example having a missing value for an attribute depends on the observed data but does not depend on the missing data..

1.3 NMAR

Not missing at Random(NMAR): when the scattering of an example having a missing value for an attribute depends on the missing values.

In case of MCAR mode, it is assumed that the scattering of missing and complete data are the same, while for MAR mode, they are different, and the missing data can be projected using the complete data. This paper aims to study and compare the application of multiple imputation techniques as a part of pre processing segment to impute the missing values and to advance the competence.

2. RELATED WORK AND PROPOSED SYSTEM

Multiple imputation primarily based approach like MICE may be a higher way for managing missing information than one imputation as many imputations think about the doubt of lost information. Multiple imputation plan crets m values for 1 missing data . it's complicated to make MICE in sensible condition with a colossal information set because the data miner needs to safeguard and research many datasets rather than 1.during this part, we tend to commit associate degree algorithmic rule Single Centre calculation from Multiple enchained Equation(SICE). it's associate degree expanding of the present MICE algorithmic rule. we've planned 2 different of SICE, particularly SICE Categorical and SICE-Numeric. Following algorithmic SICE-Categorical calculates lost values of categorical attributes like binary or ordinal attributes. For clear knowing, we tend to conjointly gift a step wise diagram of the SICE, that is ok for each categorical and numeric versions. It proceeds the MICE algorithmic rule for user defined m times associate degree adds the ends up in an array. a lost worth is replaced with the most used item of the array. Some scientist like easy strategies while not considering abundant concerning accuracy and a few need additional accurate results for his or her analysis. it's been discovered within the literature that no single methodology is sufficient all told the cases. Their performance depends on the characteristics of the dataset and missing pattern mechanism. it's been seen that the majority analysts like easy and economical strategies. So, efforts are often created in developing such strategies which can handle all types of missingness.

From the paper, we considered faults that are derived from the stator voltage rotor currents sensors, stator and rotor stator voltage. The builtin system will caryy three major steps. initial, leftover signals likely are reflecting faults in the DFIG system. These systems are produced from tested command inputs and sensor measurements. Second, the pre-processing these module define the location time of faults which is done from accurate fault classification by transforming the leftover signal to feature space. Finally, these processed leftovers aims to map each pattern of feature space to pre-assigned class of faults by slowly applying to defect classification unit that are sequentially put together. Experiments were carried out to compare the outcomes of the proposed scheme with some known strategies for replacing missing values, such as :the concept -limited most common values strategy ,the most common value strategy ,and the delete strategy. experiments have shown that the proposed methods has better results in the number and simplicity of the rules proposed ,execution time and induction time The significant trade in data analysis has been rendered by the model style of assembling massive dataset .the proposed techniques are contrasted favorably with current approaches to various real world dataset taken from machine learning respositories that are conducted on computational data analysis.

3. MISSING DATA HANDLING STRATEGIES

There are quite a lot of approaches for controlling the missing value problems :

i. Delete Strategy: This is uncomplicated approach, however with this approach bags of samples must be gone.

ii. Ignore Value Strategy: Working with this technique, every occurrences with no less than one missing value are tossed out from the dataset

iii. Imputation with K-nearest neighbor algorithm: An examination of four missing data usage approaches for observed studying. with this example created algorithm, consistently we discovered the missing value in a present case, we process the k-nearest neighbors, then assign a quantity on it. For small values, the most mutual value is occupied by all neighbors.

iv. Weighted imputation with K-Nearest Neighbor (WKNNI): This approach selects the cases with the associated values to calculate one, and now it will accredit as KNNI ensures. Though, the predictable quantity at the moment yields into report, the diverse voids near the neighbors, By means of a one-sided average or the maximum recurrent quantity conferring to the gap.

v. Support Vector Machine Imputation: A SVM reversion built method to satisfying missing values in data-Based bright Information and Engineering Schemes. SVM reversion built procedure to seal in lost data, i.e. establish the verdict attributes as the specification aspects and then the specification aspects by way of the conclusion aspects, consequently we will be able to custom SVM reversion to calculate the misplaced order aspect values. In orderliness for that to be finished, we initially choose the instances where there is no missing aspects are found. In the coming step we fix one of the order aspects, certain of those quantities are misplaced, by way of the verdict aspects ,and the verdict aspects by way of the order aspects by inverses. In conclusion, we apply SVM reversion in the direction to guess the result aspect quantities.

vi. Singular Value Decomposition Imputation: Within this procedure, To achieve In this study, the considered faults are derived from the stator voltage, stator and rotor currents sensors. The diagnostic system includes three major steps. First, residual signals which are reflecting faults in the DFIG system are produced from sampled command inputs and sensor measurements. Second, the pre processing module transforms the residual signals to a feature space as required for an accurate fault classification to define the location and time of faults. Finally, the processed residuals are sequentially collected and gradually applied to the fault classification unit which aims to map each pattern of the feature space to a pre-assigned class of faults a set of jointly orthogonal example prototypes that can be explicitly unified to estimate the values of complete points in the data set, we work with precise value disintegration. In order to be completed, we first analyze the MVs through the EM process, and then we decide the basic disintegration of value and achieve the own values.

vii. Local Least Square Imputations: Through this process, a objective case that has lost values isdenoted as a straight blend of comparable cases. Instead of handling every accessible cells within the statistics, not more than one comparable cells basis on a resemblance portion are exploited the process has "local" effect. There are couple of movesin LLSI. Initial move remainsto choose k cells using the L2- standard. And then the next stage is reversion as well as valuation, irrespective of exactly how k cells are elected. A experimental k limit mixture process remains castoff by the instigators.

4. WORKING OF MICE PROCESS IN DETAIL

MICE (Multivariate Imputation via Chained Equations) remains the frequently utilized set by 'R' operators. Generating various assertions by means of associated to a distinct assertion looks after the improbability in misplaced quantities. MICE believes that the data misplaced are ‘Missing at Random’ (MAR), that infers the chance of the lost value varies solitarily on practical quantity as well as it could be calculated by consuming them. Based on stipulating an assertion replica of each variable. For instance: Consider we now have X1, X2….Xk variables. If X1 has lost quantities, then it will be reverted on other changeable quantities from X2 to Xk. The lost quantities in X1 must then be substituted by analytical quantities acquired. Likewise, if X2 has lost quantities, then changeable quantities X1,X3 to XK must be utilized within the estimation replica as individual changeable constants. In coming stages, lost values will be substituted by calculated quantities. From defaulting case, linear regression is utilized to calculate consecutive lost quantities. Logistic regression is utilized for definite lost quantities. As soon as this loop is over, various data sets or collections are created. These data sets or collections varies only in asserted missing values. Usually, it is deemed to be a helpful attempt to develop replicas on these quantity groups or collections discretely and merging their consequences. Accurately, the approaches utilized by set are:

i. PMM (Predictive Mean Matching) – For quantitative changeable constants
ii. Logreg (Logistic Regression) – For Binary changeable constants (through 2 stages)
iii. Polyreg (Bayesian polytomous regression) – For Factor changeable constants (>= 2 stages)
iv. Proportional odds model (ordered, >= 2 stages)

At the commencement of this approach, we were with a data frame which holds lost values for numerous instances. We have to calculate a reversion factor. If there isn’t any lost value, we have to run an ols reversion by using lm() command, in our dataset. However, we need not remove every line that have lost values from the dataset, this will yield significant data and reduce the amount of findings in our data which will impact the statistical significance. At the beginning stage, the mice command generates numerous entire datasets .It counts every lost value to track a

certain distribution, and takes out a reasonable value from this distribution to substitute the lost value. These entire datasets are gathered in an object class called 'mids'. These datasets are duplicates of the primary data frame excluding that lost values are now swapped with values produced by mice. At the end, we combine the 3 factors calculated by the asserted dataset into single last reversion factor, and calculate the difference with the help of 'pool' command. With the hypothesis that reversion factors are attained from a general distribution to get the last factor we need to consider the mean of 3 values. We compute the difference of the predicted factor by calculating between the different variances.

"Predictive mean matching" (PMM): is an appealing manner to perform various assertions for lost data, particularly for asserting numeric variables which isn't conventionally distributed. Contrasted with regular approaches grounded on linear reversion and the normal distribution, PMM generated asserted values that are like original values. If the initial variable is biased, the asserted values will also be biased. If the initial variable is in between the range of 0 and 100, the asserted values will also be in between the range of 0 and 100. And if the original values are distinct, the asserted values will also be distinct. That is for the reason that the asserted values are original values that are pirated from single entities with original data. PMM was developed by Rubin and Little in early 1980's but it has become broadly accessible and to use in real-world lately. In initial stages, it might only be utilized in circumstances where a specific variable had lost data or, more approximately, when the lost data model was constantly increasing. Currently, though, the PMM approach is entrenched in huge number of software packages that instigates a way to deal with various assertion distinctly called as "multiple imputation by chained equations" (MICE), consecutively simplified reversion, or "fully conditional specification" (FCS).

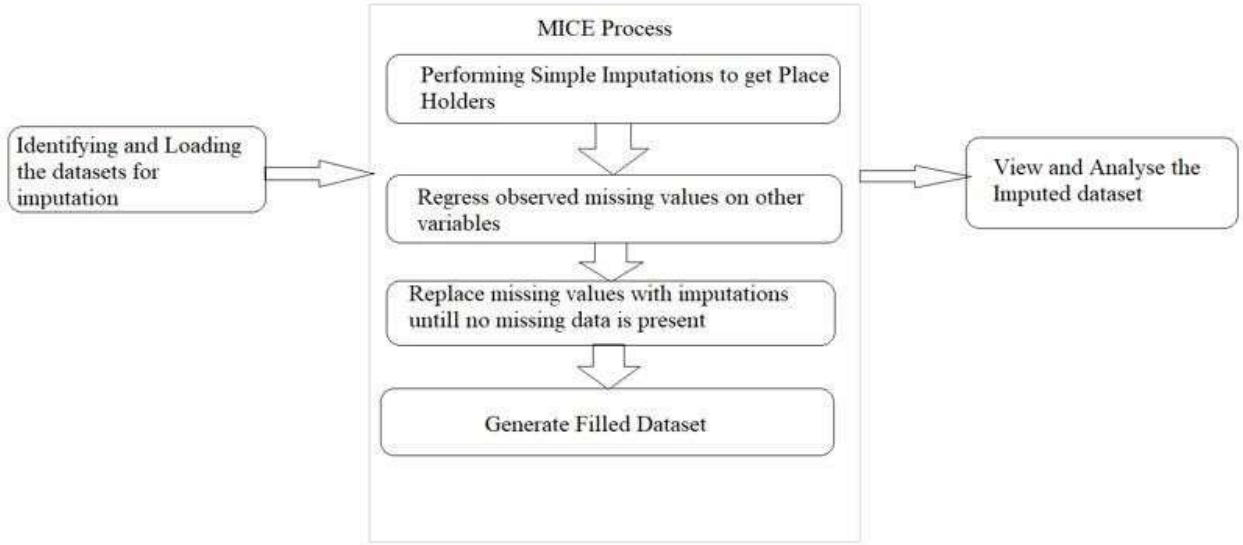

Fig:MICE Process

It is accessible in different statistical packages, with R being one among them. On the other hand ,there are two foremost hazards to PMM, Coming to first hazard, only a few research works have estimated its implementation, so, It is not that well-defined how well it competes with different methods. Coming to the second hazard, minimum of two statistical packages, 'SPSS' and 'Stata', have executed PMM with a general programming that essentially nullifies the technique. If we use any of these packages, we should outweigh the default. PMM is a universal-point approach for lost quantity assertion. One improvement of PMM is assertions are restricted to the practical quantities. PMM must maintain non linear associations likewise at the time of the operational portion of assertion replica stands inaccurate. Let, m stand as a changeable constant with certain lost quantities, and changeable constant n, along no lost data, is utilized to assert m. The algorithm works as the given manner below:

i. For the data which is not lost , linear reversion of m on n is executed, which generates b which is a set of coefficients.
ii. An arbitrary raffle starting with the subsequent analytical scattering of b is completed, which generates a latest group of factors b*.
iii. With the use of b*, projected values for m are produced for every instance.

iv. For the instance with lost m, particular instances are recognized that confined observed m whose estimated quantities are near to the estimated quantity with lost data.
v. Beginning with those near instances, a arbitrary quantity is taken to substitute with the lost value.
vi. To acquire a whole dataset, step-2 to step-5 are reiterated.

5. CONCLUSION

This paper examines the functioning of various assertion process in various circumstances of missing data. The 'principle of fully conditional specification' (FCS) has currently extended broad approval. Reliable software is currently accessible. Different tenders using FCS have turned up, and several more will keep an eye on this. This paper records a large apprise of MICE. FCS has lately stayed approved and fulfilled through SPSS, and was promoted by SPSS as the large procedural progress or development of SPSS figures. Through the years to follow, consideration will move from calculation matters to the query how we can use the procedure in an answerable manner. We should have rules and regulations on how to describe MI, we should a healthier opinion of the threats and disadvantages of the process, we should have combining styles for distinctive allotments, and we should have entry-point data that justify the thought and that exhibits how to utilize the operations in operation. Supposing that all of this occurs, various assertions exploiting FCS will attest to be a abundant accumulation to our numerical implement process.

6. REFERENCES

[1]. O. Mrudula , Dr. A. Mary Sowjanya "ANALYSIS OF MISSING DATA USING MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS (MICE)IN R" Andhra University, Visakhapatnam (India) 2017
[2]. Geeta Chhabra, Vasudha Vashisht "A Review on Missing Data Value Estimation Using Imputation Algorithm" 2017
[3]. Shahidul Islam Khan and Abu Sayed Md Latiful Hoque "SICE: an improved missing data imputation technique" 2020
[4]. Saleh M. Abu-Soud "A Novel Approach for Dealing with Missing Values in Machine Learning Datasets with Discrete Values" 2019
[5]. Eman M. Nejad, Roozbeh Razavi-Far, Q.M. Jonathan Wu, Mehrdad Saif "Multiple Imputation of Missing Residuals for Fault Classification: a Wind Turbine Application" 2015
[6]. Shylaja B, Dr. Saravana Kumar R, "Traditional Versus Modern Missing Data Handling Techniques: An Overview", International Journal of Pure and Applied Mathematics, Volume 118 No. 14 2018, 77-84, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)