ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 7 , July 2018

A Comparative Study of MissForest and Rough Set Theory for Imputation of Missing Data

Swarnendu Kundu, Tapan Goswami, Bidisha Pyne

P.G. Student, SCOPE, VIT UNIVERSITY, VELLORE,India

ABSTRACT: Handling missing data is one of the biggest challenges for data scientist in this modern era. In order to handle the missing data problem in big dataset various approaches have been introduced, by examining each data set and precise a suitable value in case of missing data. In this paper, comparison among two popular approaches of handling missing data with help of R tool using categorical dataset. One of them is MissForest (MF) which imputes missing data without any parametric measures and another is Rough Set Theory (RST) where imputation of missing data is totally based on rule induction method i.e. creation of decision table using 'if-then' rules is necessary. But the performance analysis indicates the percentage of prediction is much better in case of MissForest algorithm especially for non-medical dataset.

KEYWORDS: Missing Data, MissForest, Roughset

I.INTRODUCTION

A common problem which affects the data quality is the presence of missing data in it. The reasons behind such missing data can be denial of answering certain questionnaires, unexpected demise of a person or fault of equipment and so on. Inclusive to all this, some part of data may be untrue or invalid and the added responsibility is also to discard this false data. However, in this paper a solution to the problem of missing data in real time dataset is analysed with the help of R.

Software implementation:

R is an open source tool, basically used for computational and statistical analysis. Back in history, R is actually an implementation of S plus tool which created under John Chambers at Bell's Laboratory whereas R is the creation of Ross Ihaka and Robert Gentleman in the year 1993. Here, installation of R version 3.2.3 (2015-12-10) in Windows environment is done for missing data imputation as R is also very popular among data miners nowadays.

Packages

Among the list of all the statistical packages available in R, Miss-Forest package has its own unique method of imputing missing data in dataset. Its distinct feature is that it supports mixed data type i.e. it can be used to impute both numerical and categorical data matrix with addition to complex interaction and non-linear relations. This package random forest algorithm to create its decision tree based on observed data to predict missing values of the dataset. This can also estimate imputation error i.e. out-of-bag (OOB) without elaborating cross-validation of test set. Parallel computation is possible in random forest in to order to save computation time.

The methods included in the package can be divided into several categories based on their functionality: discretization, feature selection, instance selection, rule induction and classification based on nearest neighbours done for analysis of data. Deletion Case, Most Common Value Concept, Global Closest Fit, Concept Closest Fit are the additional methods of missing data completion

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 7 , July 2018

Dataset Used

In order to analyse the performance of this said methods MF, categorical dataset such as breast cancer, post-operative and nursery-form data is collect from UCI Repository. And used for further examination of this methods. In order to analyse the performance of this said methods MF, categorical dataset such as breast cancer, post-operative and nursery-form data is collect from UCI Repository. And used for further examination of this methods.

II. LITERATURE SURVEY

Grzymala and Ming Hu introduced a new classification technique of LERS. According to them C4.5 is the best approach among all mentioned approaches. Several methods are Most similar data of Attribute, similar Concept Attribute data, C4.5. the process of appointing all available data which is appropriate for particular attribute & having restriction for assigned concept, Method of Ignoring illustration which are consisting of unspecified values of particular attribute. Event covering & LEM-2 is used to handle the unknown data of attributes in terms of special data. It is having more efficiency & less time consuming than rest of classification algorithms & they do not have enough evidence to claim their sentence. [1].Another renowned researcher named Pawlak and his fellow scholar Stowinski analysed Miss-Forest with medical, pharmacology, industry, engineering, control, finance, geology, and social sciences datasets and can to the conclusion that its methods are very much useful in removing the duplicate values in the dataset but its decision table can go wrong on the whole if the dataset is not correct. [2] Another experiment of Jiang Hua in 2008 proves that Miss-Forest can be efficiently used in medical dataset to remove duplicity which can help to access valuable algorithms for searching the set of instruction & models used for diagnosis of medical data. With the help of ILRS they have successfully removed the redundancy and determined the most significant condition attributes for a given data set and with their experiment they also came to know that ILRS does not depends on the size of the dataset. [3]Apart from this, R tool is also previously used by Buuren and Oudshoorn (2011) for analysing the use of MICE software in R for imputing incomplete multivariate data by Fully Conditional Speciation. The extended version of it i.e. MICE v2.0 is proposed which added some new functionality for imputation such as data consisting of Multilevel instruction, choosing predictor which is fully automatic, managing the instructions, imputed data obtained from past processing of particular data. selection of model & extracting with specialization. This function makes this package more powerful though it has a simple architecture and allows easy access of source code through R environment. The practical evidence regarding this package is absent in real time so this functionality is not yet assured. [4]With help of R, Miss-Forest package is being proposed by Stekhoven in 2011 based on random forest. They also pointed out the important feature of this package is it supports mixed variable type data and also it can handle high dimensions, complex interaction and nonlinear data structure [5] Earlier to this Liaw[6] and Breiman[7] introduced random forest in R, using categorical dataset. And concluded random-forest as addition of forest containing automatic tree assigner in such a way that it relays on data of vector which are randomly sampled, individually & with equal partitioning the data of tree in specific forest. [8]Nakayama, Hattori and Ishii covered about rough set theory for data analysis in medical domain. The main drawback of rough set theory is that it mainly supports categorical data type whereas medical data may also contain continuous data types. ID3 like technique is applied to change this variable tokens of data to specific categories information. Then with help of rule extraction technique some rules are found from the simpler attributes. Medical doctors also verified their results from practical point of view. In future they will modify this technology by applying them in more datasets. [9] Another approach of Lagrange Interpolation method is reviewed by L.Sunitha, Dr M.BalRajuJ.Sasikiran for prediction of missing elements in the database. Mentioned approach is widely used for addressing tokens of data with variables in original world appliances. In case dependency is found in the dataset the unknown values can be analysed by the known value. Say for example x is independent but f(x) depends on the value of x. Hence for the given set of values of x, we can find the corresponding f(x) values and then considering thatf(x) is to be found for the value of x by interpolation. Also we can find x for given f(x) values and vice versa. Though this method is not suitable iff(x) is non-uniform in nature.[10] For handling missing data, Mark Fichman and Cummings have put the concept of multiple imputations to be more advantageous over the classical method like list wise deletion, pairwise deletion; unconditional mean imputation, conditional means imputation, maximum likelihood etc. MI not only provides low cost data collection but also higher rate of missing observation. For better probability distribution another method called MCMC (Markov Chain Monte Carlo) is also discussed. This class of algorithm can improve the quality of the task. [11][12] Another idea of Missing Data Imputation Technique (MDIT) is put forwarded by Qinbao Song and Martin Shepperd in 2007 as all the method

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 7 , July 2018

cannot be implemented as every condition, each method has its own rule and context for each case. They discussed about KKD and machine learning approaches with its own advantages and limitations. removing data in order of list & pair, removing various imputation like mean, regression & hot-deck, Expectation Maximization (EM), Raw Maximum Likelihood (RML), Sequential imputation, General Iterative Principal (GIP) Component Analysis (PCA) etc. are the MDITs analyzed here though the practical analysis is to work out. [13] [14]

III. METHODOLOGY

MissForest

This package is derived from random forest algorithm which creates decision tree for imputing missing data based on randomize parameters based on the observed values. This method can be used in both numerical or categorical or mixed type, making this approach more advantageous over the other methods followed. Another significant part of imputation is that is can estimate errors known as (OOB) Out-of-Bag imputation error, without elaboration of cross-validation.

In this method, after each loop of imputation the difference between old and new imputed data matrix is calculated for both numerical and categorical values. And a stopping criterion is defined as the value which stops the loop once this difference become greater than the value. Hence, as soon as stopping criterion is met, this imputation stops and the last imputated matrix is returned as the final one. The numerical error is calculated with help of normalized root mean squared error (NRMSE). NRMSE is can be defined mathematically as follows:

$$\frac{\sqrt{mean\left(\left(A_{old} - A_{imp}\right)^2\right)}}{var(A_{old})}$$

Where A_{old} is original data matrix, A_{imp} is the data matrix after imputation whereas mean and var represent mean and variance computed for numerical missing values only. Another kind of error known as proportion of falsely classified (PFC) is computed in case of categorical missing values only.

Another feature of 'missForest' is that it can be run in parallel. This can be done in two ways. One is to parallelize by creating objects for random forest. This process may take a long time for computation as all objects are in single forest and hence it is best suited for small number of variables in the data. Another way is of parallelizing is to create multiple random forest classifier for different variables. This method is more suitable in case of large number of variables as well as computational time is also not very long. In R, 'missForest' package 'doParallel' instance is used in order to parallel in background.

Algorithmic Approach followed in MissForest:

Requirement: A as an mxn matrix, having stopping criteria µ

1. Randomly estimate for existing missing value in A
2. v←vector having attributes arranged according to indices as in matrix A
3. loop as not µ
   - 3.1. A_{old_imp} ← matrix having existing predicted values
   - 3.2. loop as s in k // s is training sample of matrix A
     - 3.2.1. match up a random forest: y(s)obsv ~x(s) obsv
     - 3.2.2. Prediction done based on missing values
     - 3.2.3. A_{new_imp} ← imputed matrix is updated
     - 3.2.4. End of inner loop
   - 3.3. Stopping criteria(µ) is updated
   - 3.4. End of outer loop
4. A_{imp} matrix is updated

Hence, the basic advantages can be point out as follows, one this imputation method can be done for any kind of dataset be it continuous, non linear or mixed type of dataset, it is also suitable for MCAR i.e. missing completely at random type of dataset. Secondly, it is pairwise independent hence it is doesn't delete any row or column with missing data whereas it is used to impute the missing data by observing the rows. And thirdly, it can handle large dataset with

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 7 , July 2018

more missing variable though it may give rise to more imputation error sometimes.

Rough Set Theory

Rough set theory has been broadcast it's utility in machine learning, data mining, and artificial intelligence successfully. Various software tools like ROSE, RSES, and ROSETTA, R etc. are used for implementation of RST. In this paper we have use R tool for further analysis.
Professor Z.Pawlak introduced Rough Set theory (RST) as a very powerful mathematical tool for selecting meaningful patterns from large raw data set. Rough set discover the dependencies among data in data set and removed the duplicate observed variable and created a decision table by using "if-then" rules. It also assesses the minimal property sets.

Rough Set Preliminaries: Rough sets theory gives a strategy of thinking from obscure and loose data. The procedure depends on the presumption that some data is related in some data of the universes of the discourse. Some of the Rough set related terms are presented below-

A. **Information System:**
An Information System is a table, listing attributes of objects. Each row represents objects and each column represents attributes. Information System define as Information_System= (S, A), where S is finite set of objects and A is a finite set of attributes.

B. **Upper Approximation ($\bar{A}(x)$):**
The upper approximation contains all objects which possible to contain within the set S.

C. **Lower Approximation ($\underline{A}(x)$):**
The lower approximation comprises of all objects which most likely to contain within the set S.

D. **Boundary:**
The boundary region of the rough set is the difference between upper approximation and the lower approximation. Boundary Region, B(x) is expressed as
$B(x)=\bar{A}(x) - \underline{A}(x)$.

Data analysis of the Rough set is entirely based on the decision table. The decision Table can be classified into two categories– i) conditional attribute ii) decision attribute respectively. A decision rule has presented on each row attribute with some condition. With the help of the decision table, different types reduction techniques are used in the RST Package in R. The most popular decision rule in Rough set based on *if*(condition) *then(decision class),*where condition are in form on values of decision attributes. One of the popular approach of *if..then* rules is LEM2(Learning from Examples Module,version 2) Algorithm.LEM2 takes input data is a lower approximation or upper approximation. Basically, it computes a local covering and then converts into a rule set.
The available techniques for handling missing data in Rough set packages in R are Deletion Case,Most Common Value Concept, GlobalClosestFit, ConceptClosestFit. In Deletion Case is similar to Listwise deletion. The difference between Listwise deletion and Deletion Case is, Deletion Case deleted entire row based on some certain decision, not directly deleted from original database. So, there is a chance for huge data loss in case of large amount missing data. Most Common value technique is handles missing value by replacing the attribute mean or Common values. The GlobalClosesFit technique is handle missing values by replacing known value of another case which is approximately similar to that of the former. In scanning for the nearest fit case we think about two vectors of attribute, one vector relates to the missing values and another vector is a candidate for the nearest fit. The scanning is led for all the cases, therefore it name GlobalClosesFit. The ClosestFit is same like GlobalClosestFit technique. The original data set is split into smaller data set and compare with the original data set and then the GlobalClosestFit is Utilized for the smaller data sets. All the given approaches are analysed and illustrated in breast cancer data set in R tools.

IV. EXPERIMENTAL RESULTS

From the section of categorical dataset of UCI repository, nursery dataset is collected and the values are missed completely at random, having 5%, 10%, 20% and 35% of missing data and then the result is concluded in the following table.

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 7 , July 2018

Missing Percent	Missing Number of Instances	Imputed Number of Instances	Error Approximation	
			NRSME	PFC
5%	18	18	NaN	0.4522784
10%	36	36	NaN	0.4464866
20%	72	72	NaN	0.3594403
35%	127	127	NaN	0.2855766
		Approximated ERROR	NaN	**0.3859455**

Table 1. Concluding table for nursery dataset having 360 instances(small sized dataset)

In this experiment, NRSME (i.e. Normalized Root Mean Squared Error) is represented as NaN as it is used to analyse t he errors for continuous dataset. But PFC i.e. Proportion of Falsely Classified is strictly used for categorical imputation error is approximated as 0.3859455 which closer to 0 and since the good performance of missForest leads to close valu e to 0 and bad performance leads to a value near to 1. Hence, we can conclude that performance measured in this case i s good for small sized dataset of nursery dataset.

Missing Percent	Missing Number of Instances	Imputed Number of Instances	Error Approximation	
			NRSME	PFC
5%	40	40	NaN	0.4847551
10%	81	81	NaN	0.4387297
20%	162	162	NaN	0.4120841
35%	256	256	NaN	0.3334944
		Approximate ERROR	NaN	**0.4172657**

Table 2. Concluding table for post-operative dataset having 810 instances (medium sized dataset)

Our experiment is firmly based on categorical variable database and hence the NRSME i.e. Normalized Root Mean Sq uared Error is not applicable here and so it is represented as NaN here whereas PFC i.e. Proportion of Falsely Classified is strictly used for categorical imputation error is approximated as 0.4172657 which closer to 0 and since the good perf ormance of missForest leads to close value to 0 and bad performance leads to a value near to 1. Hence, we can conclud e that performance measured in this case is not as good for medium sized dataset of post-operative.

Another similar analysis is done based on large scaled dataset of Autism collected from UCI repository

Missing Percent	Missing Number of Instances	Imputed Number of Instances	Error Approximation	
			NRSME	PFC
5%	143	143	0.9276798	0.4067929
10%	286	286	0.9661748	0.3912070
20%	572	572	0.8612434	0.3647761
35%	1001	1001	0.7043827	0.2654939
		Approximated ERROR:		**0.3570675**

Table 3. Concluding table for Autism dataset having 2860 instances (large scale dataset)

Similarly, in the case of large scale dataset of Autism we can conclude that this experiment is firmly based on categoric al variable database and so NRSME i.e. Normalized Root Mean Squared Error is not applicable here and it is represent ed as NaN here whereas PFC i.e. Proportion of Falsely Classified is strictly used for categorical imputation error is appr

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 7 , July 2018

oximated as 0.3570675 which is much more closer to 0 and since the good performance of missForest leads to close val ue to 0 and bad performance leads to a value near to 1. Hence, we can conclude that performance measured in this case is better for large sized dataset of breast cancer.

Dataset	Scale of Dataset	Approximated Error
Nursery	Small scaled	0.3859455
Post operative	Medium scaled	0.4172657
Autism	Large scaled	0.3570675

Table 4. *Approximated error of above three datasets*

Analysis is done with three types of dataset and the PFC is compared among them, with that we can conclude that miss forest is suitable for all types of dataset but more accuracy is gained with large scale dataset with hundred percent impu tations as the approximated error for large scaled dataset like breast cancer is closer to 0.

Rough Set Analysis: There are four important methods of rough set which can be used for missing data imputation are as follows:

i) Most common value:

This concept is used in predicting missing values by simply allocating the most common value of the colu mn limited to a concept. In case of numerical or continuous data attribute, the mean of the values are assig ned as a substitute of most common value.

ii) Deletion cases

The concept is used for treatment of missing values by deleting those occurrences denoted by their NA va lues. It is to be kept in account that the output of the function is val.NA which holds the indices of missing values.

iii) Global closest fit

The global closes fit method is purely based on substitution of missing attribute value by the known value in another similar case that is nearest to the case of the missing value attribute. The search is done for com paring two vectors of attribute values, if any of the vector found to be an equivalent case with a missing at tribute value, then this vector is said to be the closest fit. In addition to this search is conducted for all case s, hence the term ‘global’ is added with the name.

iv) Concept closest fit

This method is more likely to the global closest fit method. The only distinction is that it is in its initial sta ges first split the original dataset into small sized dataset, and each of such small dataset matches to the co ncept derived from original dataset. Or in other words, every of these small data set is established from on e of the actual concepts, by limited cases of the concept.

Analysis of all the above methods are analysed and the following result is depicted as follows in post operative dataset:

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 7 , July 2018

Analysis Methods	5% Missing Data		10% Missing Data		20% Missing Data		35% Missing Data	
	Imputed values	NA value	Imputed values	NA value	Imputed values	NA value	Imputed values	NA value
Most common value	40	0	81	0	162	0	283	0
Deletion cases	0	40	0	81	0	162	0	283
Global closest fit	39	1	71	10	140	22	204	79
Concept closest fit	27	13	58	23	89	51	174	109

Table 5. Result of analysis of 4 methods of rough set theory

Comparative Table:

	Rough Set (Global Closest Fit)		Miss Forest		
	Percentage of Missing after prediction	Percentage of prediction of total missing data	Percentage of Missing after prediction	Percentage of prediction of total missing data	PFC (error)
5% missing data	2.564	97.436	0	100	0.4847551
10% missing data	12.345	87.655	0	100	0.4387297
20% missing data	13.580	86.420	0	100	0.4120841
*35% missing data	27.915	72.085	0	100	0.3334944

Table 6. Comparison of global closest fit and missForest after imputation

	Rough Set (Concept Closest Fit)		MissForest		
	Percentage of Missing after prediction	Percentage of prediction of total missing data	Percentage of Missing after prediction	Percentage of prediction of total missing data	PFC (error)
5% missing data	32.5	67.5	0	100	0.4847551
10% missing data	28.39	71.61	0	100	0.4387297
20% missing data	31.48	68.52	0	100	0.4120841
35% missing data	38.51	61.49	0	100	0.3334944

Table 7. *Comparison of concept closest fit and missForest after imputation*

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 7 , July 2018

The above two tables represents a comparison of rough set methods and missForest, which clearly represent MF attains hundred percent imputation of the missing data along with some computed error (known as PFC) but in the case of rough set theory, hundred percent imputation is only attainable for MostCommonValue method whereas the inverse result is obtained by performing DeletionCases. But the scenario is different in case of GlobalClosestFit and ConceptClosestFit. Though high percentage of missing value is predicted in case of global closest fit than that of concept closest fit.

V.CONCLUSION AND FUTURE WORK

In this paper, the analysis of the behaviour of the two methods is being done for treatment of missing data; the MissForest imputes data without any parameters whereas the rough set theory imputation techniques are totally based on decision table. These methods are performed by inserting different percent of missing data in all the attributes of three datasets, shows some potential results. The MissForest method have provided hundred percent imputation power even in case of large amount of missing data but this kind of prediction power can be harmful or useless in case of medical domain analysis as this imputation technique may fail to approximate the original (missing) value. Whereas, the rough set theory techniques used decision table for every method of imputation of missing data, which can be very useful mainly in case of medical dataset like breast cancer or post operative. It should be noted that all the performance in this work is based on missing data completely at random, and then the result is analysed. In the future work, the erroneous analysis of rough set theory and its originality of prediction of missing value can be done with some more datasets. And also the behaviour of these methods to be analysed when missing data is are distributed following certain pattern. So, in future apart from this error rate, the quality of knowledge gained after imputation should be analysed specially in medical field.

REFERENCES

[1] Grzymala-Busse, J. W., & Hu, M. (2000, October). A comparison of several approaches to missing attribute values in data mining. In *Rough sets and current trends in computing* (pp. 378-385). Springer Berlin Heidelberg

[2] Pawlak, Zdzisaw, and Roman Sowinski. "Rough set approach to multi-attribute decision analysis." *European Journal of Operational Research* 72.3 (1994): 443-459.

[3] Hua, Jiang. "A Knowledge Acquisition Model of Inconsistent Medical Data Based on Rough Sets Theory." Intelligent Information Technology Application, 2008. IITA'08. Second International Symposium on. Vol. 1. IEEE, 2008.

[4] Azar, Ahmad Taher, Nidhal Bouaynaya, and Robi Polikar. "Inductive learning based on rough set theory for medical decision making." *Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on*. IEEE, 2015.

[5] Buuren, S., &Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*(3).

[6] Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28.1 (2012): 112-118.

[7] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[8] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

[9] Nakayama, Hirotaka, Yuichi Hattori, and Renichi Ishii. "Rule extraction based on rough set theory and its application to medical data analysis." Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on. Vol. 5. IEEE, 1999.

[10] Sunitha, L., BalRaju, M., &Sasikiran, J. (2013). Data Mining: Estimation of Missing Values Using Lagrange Interpolation Technique. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),2(4), pp-1579.

[11] Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. Organizational Research Methods, 6(3), 282-308

[12] Nielsen, Søren Feodor. "Proper and improper multiple imputation." International Statistical Review 71.3 (2003): 593-607.

[13] Song, Q., &Shepperd, M. (2007). Missing data imputation techniques. International journal of business intelligence and data mining, 2(3), 261-291.

[14] Finch, W. Holmes. "Imputation methods for missing categorical questionnaire data: a comparison of approaches." Journal of Data Science 8.3 (2010): 361-378.