International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Multivariate Data Quality Enhancement by Ranked Imputation

Muralidharan Jayaraman, P. Shanmugavadivu

Abstract*: Organizational decisions are based on data-based-analysis and predictions. Effective decisions require accurate predictions, which in-turn depend on the quality of the data. Real time data is prone to inconsistencies, which exhibit negative impacts on the quality of the predictions. This mandates the need for data imputation techniques. This work presents a prediction-based data imputation technique, Rank Based Multivariate Imputation (RBMI) that operates on multivariate data. The proposed model is composed of the ranking phase and the imputation phase. Ranking dictates, the attribute order in which imputation is to be performed. The proposed model utilizes tree-based approach for the actual imputation process. Experiments were performed on Pima, a diabetes dataset. The data was amputed in range between 5% - 30%. The obtained results were compared with existing state-of-the-art models in terms of MAE and MSE levels. The proposed RBMI model exhibits a reduction of 0.03 in MAE levels and 0.001 in MSE levels.

***Keywords*: Data Imputation, Machine Learning, Multivariate, Correlation, Decision Tree.**

I. INTRODUCTION

The highly interconnected world has resulted in a large amount of data being generated online. This data is generated in real-time, representing real-time events [1]. Evolution of storage devices and the lowered storage cost has enabled users to record this data in real-time for analysis. Analysis of such real-time data results in identifying several complex and interesting patterns that enable business decision making [2, 3].

Data mining is the process of performing analysis on data to identify interesting patterns within them. Data mining processes can be categorized as supervised or unsupervised methods [4]. Supervised models train on labelled data and perform predictions, while unsupervised models train on unlabelled data and groups them based on the intrinsic patterns contained in the data. Some common supervised learning models include Classification and Regression [5]. Both the models train on labelled data instances to predict classes for unlabelled instances.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Muralidharan Jayaraman, Research Scholar, Department of Computer Science and Applications, The Gandhigram Rural Institute (Deemed to be University), Dindigul, Tamil Nadu, India. Email: jaymurali@gmail.com
Dr. P. Shanmugavadivu*, Professor, Department of Computer Science and Applications, The Gandhigram Rural Institute (Deemed to be University), Dindigul, Tamil Nadu, India. Email: psvadivu67@gmail.com

It has varied applications ranging in fields like bioinformatics, information retrieval, pattern recognition, image analysis, fraud detection etc.

The major requirement for such applications is clean data. However, not all real-time data are complete [6]. They tend to be laden with several missing values [7]. This is unavoidable, as data for real-time applications can contain human errors during data entry, recording errors by machines such as failure of a certain component or a sensor etc [8].

The major issues with data containing missing values is that data mining models cannot work with missing data. The missing values need to be imputed prior to the training process. Some general methods used for imputation includes, deleting the records with missing values, replacing missing values using statistical measures or replacing missing values using complex imputation techniques [9].

This work presents a prediction-based data imputation model that utilizes a ranking mechanism to improve the efficiency of the prediction process. The proposed Rank based Multivariate Imputation (RBMI) model contains two major phases; the correlation-based attribute ranking mechanism and the tree-based data imputation technique. The first phase analyses the imputation levels to identify the sequence in which the imputation process is to be performed. The second phase performs imputation using prediction models based on the selected attribute. Experiments were performed and results were measured based on MAE and MSE values. Comparisons with existing models in literature indicates highly effective imputation process.

The remainder of this paper is structured as follows; section II presents the related works, section III presents the proposed imputation model, section IV presents and discusses the results and section V concludes the work.

II. LITERATURE REVIEW

Several studies are available in literature focusing on data imputation. However, most of them deal with statistical measures for the imputation process. This section discusses some of the recent contributions in the domain of data imputation.

An analysis of the significance of the pre-processing techniques, specifically data imputation, was proposed by Shobha et al. [10]. This model performs an analytical comparison of existing state-of-the-art works to provide a base for presenting the importance of data imputation as a pre-processing technique when performing data mining. A work presenting the entire workflow of the imputation sequence on incomplete datasets was presented by Brand [11]. This thesis presents a formulation of amputation mechanisms, a discussion of approaches to analyse incomplete data,

list of proposed imputation techniques and ways to validate the imputation efficiency levels. An ensemble-based technique that performs data imputation and prediction was proposed by Shobha et al. [12].

This model performs clustering-based imputation and utilizes an ensemble model for the final prediction. A model considering the nature of data as the base prior to the actual imputation process was proposed by Soares et al. [13]. Although several models were developed for performing imputation, they do not however, consider the relationship between data and the data properties. This results in reduced performances when the model is applied new and unseen data. An analysis of this issue is performed by Soares at al. [13]. This work also proposes solutions for overcoming this issue.

A comparison and analysis of standard imputation techniques was proposed by Nanni et al. [14]. This model analyses various imputation techniques under several imputation levels ranging between 10%-50%. The effects of imputation on a classification model was studied by Aisha et al. [15]. The study concluded that SVM based imputation model outperformed the other models in terms of the obtained error levels. A similar study was conducted on a medical dataset by Rahman et al. [16]. An analysis of the imputation models in terms of classification accuracy was performed by Gracia-Laencina et al. [17]. This study utilized both synthetic and real datasets for analysis, hence providing reliability of the study. A data imputation model for genetic sequences was presented by Chen et al. [18]. The proposed Gimpute model creates pipelines to effectively perform imputation on genomes.

A multivariate imputation model based on Gibbs sampler was proposed by Buuren et al. [19]. This model enhances the simple Gibbs based imputation model to provide enhanced results. A Generative Adversarial Network based data imputation model for handling missing data was presented by Yoon et al. [20]. The generator component identifies and imputes the missing values. The discriminator component tries to identify the imputed values. This cycle is continued till the generator imputes the best values based on the data distribution. A Gaussian modelling system for imputing high-dimensional data was proposed by Imani et al. [21]. The model has been designed for healthcare systems operating on huge data. a multiple imputation model to handle systematic and sporadically missing data was proposed by Rigon et al. [22]. This work presents two imputation methods to be used on data. First is an extension of an existing method and the second is a two-stage method that can effectively handle heteroscedastic data.

III. PROPOSED IMPUTATION MODEL (RBMI)

Qualitative data is mandatory for effective decision making. This work presents a prediction-based imputation model (RBMI) that aims to provide accurate imputation of values that can in-turn enable qualitative data for accurate predictions. The proposed RBMI model is composed of two major phases; the correlation-based attribute ranking phase and the tree-based data imputation phase.

A. Correlation based Attribute Ranking

Data imputation is the primary goal of the current work. However, a dataset might contain missing values in multiple columns. Imputation can be performed based on single attributes. This is iteratively performed on every attribute with null values to obtain clean data. In such a scenario, question arises as to the selection order of attributes. This mandates the need for a ranking mechanism.

The order of imputation is obtained by correlation-based ranking of attributes. Consider *n* data instances with *k* classes, composed of *d* dimensions. The feature vectors are given by

$$X_i = [x_{i1}, x_{i2}, \dots x_{id}] \qquad \forall\, i = 1,2, \dots, n \qquad (1)$$

The correlation for an attribute pair x_i and $x_j$is given by

$$r_{x_i x_j} = \frac{\sum_k x_{ik}x_{jk} - n\bar{x}_i\bar{x}_j}{\sqrt{(\sum_k x_i^2 - n\bar{x}_i^2)(\sum_k x_j^2 - n\bar{x}_j^2)}} \qquad (2)$$

Correlation of all the attributes with the Class attribute is identified. The correlation levels determine the ranks, which in-turn determine the order of imputation. Correlation can be positive, negative or zero. The proposed model ignores the magnitude of the correlation level and considers the value alone for the ranking process. Attributes with low correlation levels (correlations values nearing zero) are provided with lower ranks, while attributes with high correlation (correlation values nearing 1 or -1) are provided with higher ranks.

B. Tree based Data Imputation

Data imputation is performed by considering the attributes with missing data as the column to be predicted. The mining algorithm is trained using the available non null records, as prediction models cannot handle missing values in the data. The trained model is then used to predict the missing values contained in the selected attribute.

Data preparation is the first stage of the processing model. All records containing null values are eliminated and the training data is created. This results in a clean training data containing only the actual values. The prediction algorithm is trained using this data.

This work uses Decision Tree as the base algorithm for the prediction process. Supervised training models are based on the data type of the prediction variable. Models predicting categorical values are categorized as Classification models and the models predicting numerical values are categorized as Regression models. Data imputation is performed on multiple attributes, which can contain numerical or categorical values. Hence all numerical attributes are predicted using Decision Tree Regression models, while categorical attributes are predicted using Decision Tree Classifiers. Decision Tree [23] is a graph-based model that represents consequent decisions using graph like structures with nodes representing condition and leaves representing final predictions. The tree structure is created using the training data and decision rules are obtained by traversing through the tree to reach the final predictions. The major advantage of this model is that it is a simple and white-box model and can be easy to walk-through.

The records containing missing values that has been eliminated in the training data creation phase is considered as the test data. However, this data might have the possibility of containing missing values in attributes other than the selected attribute.

Hence all other attributes (other than the attribute to be predicted) containing missing values are imputed with the column based mean values. Certain columns however, might completely contain missing values.

Hence if an aggregate mean is not computable for the column, then data is replaced with appropriate values specified by the user. The prepared test data is then used for prediction. The predicted values are used to replace the missing values contained in the actual data.

This process is performed iteratively on all attributes containing missing values. The sequence of processing is performed based on the descending order of rank. This scheme is intentionally adopted in the proposed model, rather than random attributes for imputation. The attributes with low correlation are imputed first, as they have low significance in predicting the class attribute. A single record might contain multiple missing values, apart from the attribute being operated upon. Hence, this will require temporary substitution of dummy values to complete the imputation process. This process of temporary substitution is safer when done on attributes with low correlation, rather than on attributes with high correlation.

IV. RESULTS AND DISCUSSION

Experiments were performed by implementing the proposed RBMI model in Python. The dataset is multivariate and is composed of 769 instances and 8 attributes. The data does not contain any missing values. Data amputation technique proposed by Rianne et al. [24] is used for amputation, and the resultant dataset is used for the imputation process. Missing values are created at random based on Missing Completely at Random (MCAR) [25] method, with their levels ranging from 5% to 30%. Efficiency of the prediction process is identified by calculating the Mean Absolute Error (MAE) and Mean Squared Error (MSE). Lower values for MAE and MSE indicate effective imputation.

Mean Absolute Error measures the effectiveness of the predictions, and is given by

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - y'_i| \quad (3)$$

Mean Square Error indicates the variability of the predictions, and is given by

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - y'_i)^2 \quad (4)$$

Where y_i and y_i' are the actual and the predicted ratings for the N test reviews.

Comparison of the proposed predictions is performed with imputation using mean and median values, which is are some of the most commonly used imputation methods.

A comparison of the MAE of the proposed RBMI model with the Mean and Median based models is shown in Figure 1. The X and Y axis shows level of missingness in the data and MAE values respectively. It could be observed that the proposed model exhibits lower MAE levels compared to the other models. This exhibits the high correlation in the prediction process, leading to more accurate imputation. A tabulated view of the values is shown in Table 1.

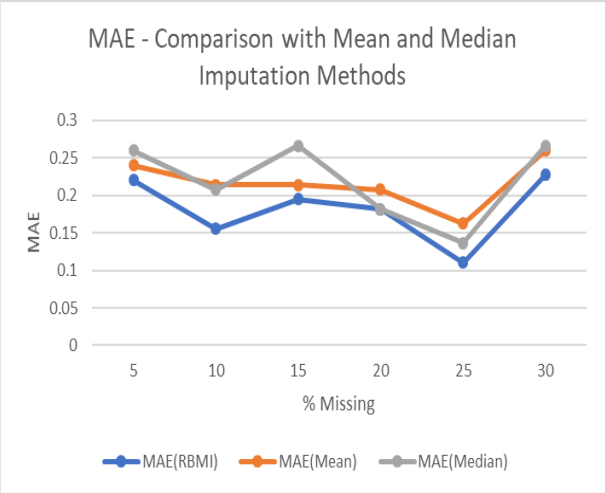

Fig. 1. Comparison of MAE Values with Mean and Median based Imputation Methods

Table 1: MAE Comparison

Proportion of Missing Values	MAE (RBMI)	MAE (Mean)	MAE (Median)
5	0.221	0.240	0.260
10	0.156	0.214	0.208
15	0.195	0.214	0.266
20	0.182	0.208	0.182
25	0.110	0.162	0.136
30	0.227	0.260	0.266

A comparison of the MSE values of the proposed model with Mean and Median based models is shown in Figure 2. A tabulated view of the data is shown in Table 2. It could be observed that the proposed model exhibits lower MSE values, indicating that the model exhibits low magnitude-based variations.

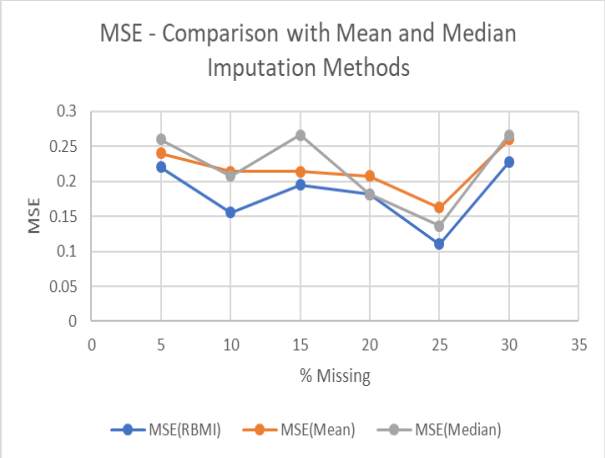

Fig. 2. Comparison of MSE Values with Mean and Median based Imputation Methods

Table 2: MSE Comparison

Proportion of Missing Values	MSE (RBMI)	MSE (Mean)	MSE (Median)
5	0.221	0.240	0.260
10	0.156	0.214	0.208
15	0.195	0.214	0.266
20	0.182	0.208	0.182
25	0.110	0.162	0.136
30	0.227	0.260	0.266

V. COMPARISON WITH STATE-OF-THE-ART

A comparison of the proposed models is performed with a recent imputation model proposed by Shobha et al. (2019) [12]. This is an ensemble-based approach to perform both imputation and prediction. Comparisons are performed by identifying the deviation level of the model with respect to the mean-based imputation technique. A comparison of average deviation levels of MAE and MSE values is shown in Figure 3. Lower values show that the proposed model is almost similar to the mean-based model, while higher deviations shows that the proposed model performs better than the mean-based imputation technique. It could be observed that the proposed RBMI model exhibits lower deviation levels of both MAE and RMSE levels. The RBMI model exhibits a reduction of 0.03 in MAE levels and 0.001 in MSE levels compared to the model proposed by Shobha et al., exhibiting the high efficiency of the proposed approach.

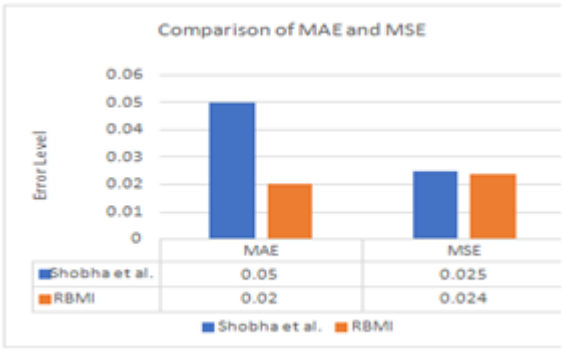

Fig. 3. Comparison of Deviation Levels of MAE and MSE Values

VI. CONCLUSION

The current prediction environments mandate data imputation for the need of qualitative results. This work proposes the Rank Based Multivariate Imputation (RBMI), that can effectively impute data on a multivariate dataset. The model is composed of a ranking stage and a tree-based imputation stage. The ranking stage determines the significance level of attributes to determine the attribute sequence in which the imputation is to be performed. The next stage uses tree-based prediction model; classifier for imputing categorical values and regressor for predicting numerical values. Experimental results on the proposed model indicates low errors (MAE and MSE), indicating the significance of the model.

The major advantage of the RBMI model is that it works on data with multiple imputations. i.e. the model has been designed to handle records that contain multiple missing values. The ranking mechanism enables initially imputing attributes with low significance, followed by imputation of attributes with high significance. Hence the imputed data exhibits very low variance when compared with the actual value. The model is designed to be generic, hence can be operated upon any type of data.

REFERENCES

1. Condoluci, Massimo, Giuseppe Araniti, Toktam Mahmoodi, and Mischa Dohler. "Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications." IEEE Access 4, Pages: 5555-5569, 2016.
2. Fekade, Berihun, TarasMaksymyuk, MaryanKyryk, and Minho Jo. "Probabilistic recovery of incomplete sensed data in IoT." IEEE Internet of Things Journal, Volume 5, Issue 4, Pages: 2282-2292, 2017.
3. Fisher, Paul S., Jimmy James Ii, JinsukBaek, and Cheonshik Kim. "Mining intelligent solution to compensate missing data context of medical IoT devices." Personal and Ubiquitous Computing, Volume 22, Issue 1, Pages: 219-224, 2018.
4. Hand, David J. "Data Mining." *Encyclopedia of Environmetrics, Volume* 2, 2006.
5. Breiman, Leo. *Classification and regression trees*. Routledge, 2017.
6. Van Buuren, Stef. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.
7. Zhang, Zhongheng. "Missing data imputation: focusing on single imputation." Annals of translational medicine, Volume 4, Issue 1 2016.
8. Pedersen, Alma B., Ellen M. Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R. Kristensen, Tra My Pham, Lars Pedersen, and Irene Petersen. "Missing data and multiple imputation in clinical epidemiological research." *Clinical epidemiology*, *Volume* 9, 2017.
9. Quinteros, María Elisa, Siyao Lu, Carola Blazquez, Juan Pablo Cárdenas-R, Ximena Ossa, Juana-María Delgado-Saborit, Roy M. Harrison, and Pablo Ruiz-Rudolph. "Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile." Atmospheric environment, Volume 200, Pages: 40-49, 2019.
10. Shobha, K., and S. Nickolas. "Analysis of importance of pre-processing in prediction of hypertension." CSI Transactions on ICT, Volume 6, Issue 2, Pages: 209-214, 2018.
11. Brand, Jaap JPL. Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. 1999.
12. Shobha, K., and S. Nickolas. "Imputation of Multivariate Attribute Values in Big Data." In Smart Intelligent Computing and Applications, Pages: 53-60. Springer, Singapore, 2019.
13. Soares, JastinPompeu, Miriam Seoane Santos, Pedro Henriques Abreu, Hélder Araújo, and João Santos. "Exploring the effects of data distribution in missing data imputation." In International Symposium on Intelligent Data Analysis, Pages: 251-263. Springer, Cham, 2018.
14. Nanni, Loris, Alessandra Lumini, and Sheryl Brahnam. "A classifier ensemble approach for the missing feature problem." Artificial intelligence in medicine, Volume 55, Issue 1, Pages: 37-50, 2012.
15. Aisha, Nazziwa, Mohd Bakri Adam, and ShamarinaShohaimi. "Effect of missing value methods on Bayesian network classification of hepatitis data." Int. J. Comput. Sci. Telecommun, Volume 4, Issue 6, Pages: 8-12, 2013.
16. Rahman, M. Mostafizur, and Darryl N. Davis. "Fuzzy unordered rules induction algorithm used as missing value imputation methods for k-mean clustering on real cardiovascular data." Lect Notes EngComput Sci, Volume 2197, Issue 1, Pages: 391-394, 2012.
17. García-Laencina, Pedro J., José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. "Pattern classification with missing data: a review." Neural Computing and Applications, Volume 19, Issue 2, Pages: 263-282, 2010.
18. Chen, Junfang, Dietmar Lippold, Josef Frank, William Rayner, Andreas Meyer-Lindenberg, and Emanuel Schwarz. "Gimpute: an efficient genetic data imputation pipeline." Bioinformatics, Volume 35, Issue 8, Pages: 1433-1435, 2018.

19. Van Buuren, Stef, Jaap PL Brand, Catharina GM Groothuis-Oudshoorn, and Donald B. Rubin. "Fully conditional specification in multivariate imputation." Journal of statistical computation and simulation Volume 76, Issue 12, Pages: 1049-1064, 2006.
20. Yoon, Jinsung, James Jordon, and Mihaela Van Der Schaar. "Gain: Missing data imputation using generative adversarial nets." arXiv preprint arXiv:1806.02920, 2018.
21. Imani, Farhad, Changqing Cheng, Ruimin Chen, and Hui Yang. "Nested gaussian process modeling for high-dimensional data imputation in healthcare systems." In 2018 Institute of Industrial and Systems Engineers Annual Conference and Expo, IISE 2018. 2018.
22. Resche-Rigon, Matthieu, and Ian R. White. "Multiple imputation by chained equations for systematically and sporadically missing multilevel data." Statistical methods in medical research, Volume 27, Issue 6, Pages: 1634-1649, 2018.
23. Quinlan, J. Ross. "Simplifying decision trees." International journal of man-machine studies, Volume 27,Issue 3, Pages: 221-234. 1987.
24. Schouten, Rianne Margaretha, Peter Lugtig, and GerkoVink. "Generating missing values for simulation purposes: a multivariate amputation procedure." Journal of Statistical Computation and Simulation, Volume 88, Issue 15, Pages: 2909-2930, 2018.
25. Lydersen, Stian. "SP0156 MISSING DATA: IS IT ALL THE SAME?" Pages: 48-48, 2019.

AUTHORS PROFILE

Dr. Shanmugavadivu Pichai is currently the Professor of Computer Science and Applications, at Gandhigram Rural Institute (Deemed to be University), and is involved in Teaching, Research, and Extension. She is the Local Coordinator for the University, for the Global Initiative of Academic Networks (GIAN), MHRD, India. Dr. Pichai has 27+ years of Academic experience, and has guided/guiding Research Scholars, and funded-research projects of UGC, DST and ICMR for an outlay of Rs. 150.6 Lakh. Her research areas include Medical Image Analysis, Healthcare Analytics. Parallel Computing, Digital Image Processing and Content-Based Image Retrieval. She has conducted a national conference, training programmes, and workshops, and has delivered 80+ lectures as Keynote Speaker, Chief Guest, and Guest Lecturer. She has edited three volumes of research publications and two national journals, and authored about 100+ research publications. Recipient of Indo-US 21st Century Knowledge Initiative Award 2015. She had been on an international academic assignment in Malaysia and USA. Dr.Pichai holds a Master's degree in Computer Applications from Regional Engineering College, Trichy, Ph.D. in Digital Image Restoration, and MBA.

Mr. Muralidharan Jayaraman has 29+ years of truly global enrichment in Software Services Industry, and has held several customer-centric leadership positions including Head of India Operations, Delivery Lead, and Account Lead. He is currently on the Board of a Technology Startup, that focuses on Cloud and Analytics. As part of his roles at Deutsche Bank, Tata Consultancy Services, Paragon Solutions, and CGI, he has worked in the US, and Singapore, besides India, in multi-cultural environments on global technology projects. Having travelled extensively on business, he has got precious insights into the local flavours of doing business, and the cultural diversity. He holds a Master's Degree in Computer Applications from Regional Engineering College, Trichy, and followed it up by completing Executive General Management Programme at IIM, Bangalore. He currently pursues Ph.D. in part-time, in the domain of Analytics and Healthcare. He is also a certified Corporate Director. As part of giving back to the society in the field of education, he has written several chapters on soft skills which are prescribed as books at universities, and has been giving several invited lectures at colleges and universities. He also works with a few NGOs on improving lives of some underprivileged children.