IJCTA, 9(27), 2016, pp. 1-4 ISSN: 0974-5572

Accentuating the necessity for new-fangled IoT missing data imputation technique

Priya Stella Mary* and L. Arockiam**

ABSTRACT

Missing value imputation is the most common pre-processing task in data mining. IoT generated datasets are largely incomplete. Discarding the rows with missing values will significantly reduce the sample size as well as diminish the power of analysis. Employing an apposite missing value imputation technique would greatly increase the statistical power and yield quality datasets. In this paper, a deep investigation in to existing research works on missing IoT and sensor data imputation has been made; the types and patterns of missing values and prominent missing data imputation tools have been briefly deliberated; It finally becomes obvious that only a new-fangled missing value imputation technique based on the characteristics of IoT data can enrich the accuracy, consistency, and stability of the IoT analytics.

***Keywords**:* IoT, Imputation, analytics.

1. INTRODUCTION

The internet of things is the rapidly growing technology at a breath-taking pace [1]. IoT offers a wide variety of IoT applications such as smart home, smart wearables etc. At the same time these applications pose numerous problems that are to be overwhelmed. One such problem is missing data imputation for the internet of things [2].

The reasons for missing data in the internet of things are plenty. Some of them include sensor faults, intermittent network connection, defective IoT devices, equipment failure, malfunctioning devices etc. Missing data can be a serious obstacle for data analysis. Missing data in the input of predictive model results in poorer outcomes or produces no results at all. Hence finding out missing data can be beneficial and becomes mandatory in IoT to do quality analytics. To perform the missing value imputation, the thorough probe into the missing data mechanisms becomes obligatory [3]. This paper presents an outline of the missing data mechanisms and missing data patterns; investigates existing research works on missing IoT and sensor data imputation; deliberates prominent imputation tools; offers the general missing data imputation model.

2. BACKGROUND

2.1. Missing data mechanism

It is the process by which the data values become incomplete. The missing data imputation model may yield correct inferences under one missing data mechanism whereas the same model may yield incorrect inferences under another missing data mechanism [4]. So gaining knowledge about this mechanism is indispensable for appropriate analysis of missing data. There are three missing data mechanisms namely

1) *Missing Completely at Random (MCAR) :* if the probability that a missing value of a variable is totally random and does not depend on the missing values Y_{mi} as well as the observed variables Y_{ob} in the dataset [4].

* Ph.D Scholar

** Associate Professor, Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli – 2.

2) *Missing at Random (MAR):* if the probability that a missing value of a variable does not depend on the variable itself but depends on all the other observed variables Y_{ob} in the dataset.

3) *Missing not at Random (MNAR):* if the probability that a missing value is not random and depends on the variable that is missing.

2.2. Missing data patterns

There exist three kinds of missing data patterns [6] namely univariate, monotone and arbitrary.

1) *Univariate pattern:* According to this univariate pattern, missingness occurs only in one variable say y. But all other variables x_1, x_2…..x_p are completely observed.

2) *Monotone pattern:* According to this monotone pattern, variables such as x_1, x_2…x_p are ordered in such a way that if x_m is missing then x_{m+1}, x_{m+2}…x_p are missing as well [7].

3) *Arbitrary pattern:* According to this arbitrary pattern, missingness occurs in any variable in a random fashion.

3. RELATED WORKS

XiaoboYan et al. [1] have explored missing data in IoT and proposed three corresponding missing value imputation methods namely Model of missing value imputation based on context and linear mean (MCL), Model of missing value imputation based on binary search (MBS), Model of missing value imputation based on Gaussian mixture model (MGI) based on the type of missing data. Zhipeng Gao et al. [8] have assessed missing values based on the temporal and spatial dimensions by assigning different weights and also proposed Temporal and Spatial Correlation Algorithm(TSCA) to estimate missing data.

Xiaojun Ren et al. [9] have proposed a new estimation model based on a spatial-temporal correlation analysis (STCAM). Kim DJ et al. [10] have proposed the Canonical Correlation based k weighted Angular Similarity (CkWAS) to map the missing data with reference pattern dataset. CesareAlippi et al. [11] have suggested an overall methodology for restructuring missing data based on both temporal and spatial redundancy. Li Peng et al., [12] have presented the density clustering and grey relational analysis methods to impute missing values in medical datasets. But the proposed method is suitable to handle MAR type missing data not NMAR type missing data. JaemunSim et al. [13] have performed an analysis on the characteristics of missing values and missing value imputation methods. Ferrari et al., [14] have described the statistical imputation method used to impute the missing values in the daily precipitation dataset of the State of Parana in Brazil. Then quality control has been done to detect potential errors after the imputation process.

4. GENERAL IOT IMPUTATION MODEL

Data generated by IoT devices are typically time-series data. The collected data usually contain multiple attributes such as time, device id, temperature etc. which are stored in the IoT database. The features that are specific to the IoT data are extracted using feature extraction module and then they are supplied to the chosen missing data imputation model which produces complete dataset. Finally the imputation accuracy is calculated and reported as shown in figure 1.

5. IMPUTING IOTDATA USING SOFTWARE TOOLS

Data generated by divergent IoT devices are not always appropriate to perform analytics. An enrichment step namely pre-processing is indispensible to impute missing values. After the amelioration of datasets with missing values using the potent software tools [15] as illustrated in table 1, IoT analytics could be accomplished. But these software tools should undergo little modifications by taking into account the spatial and temporal characteristics of IoT data.

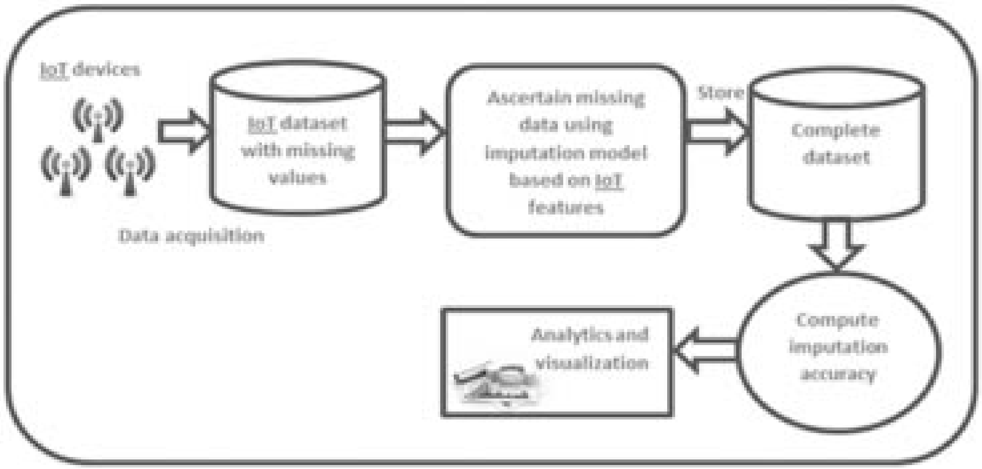

Figure 1: General IoT missing data imputation Model

Table 1
Comparison of missing data imputation software packages

Software Packages	*Specialities*	*Access*
R	Offers multiple packages such as MICE, Amelia, miss Forest, mi etc to do missing value imputation.	Open source
SPSS	Provides an add-on module called “Missing Value Analysis” (MVA) containing many Imputation algorithms to carry out imputation.	Proprietary
STATA	Does not have separate module for missing value imputation but offers a suite of commands to perform imputation.	Proprietary
SAS	Deploys two modules namely PROC MI and PROC MIANALYZE to perform missing value imputation.	Proprietary

6. CONCLUSION

Most of the conventional missing value imputation techniques are not appropriate to handle missing values in heterogeneous IoT data from divergent sources and these techniques are extremely defective and yield biased outcomes. The IoT data is unpredictable in nature, but the existing models are only suitable to predictable MCAR and MAR type missing data but not to unpredictable MNAR type missing data. Also the conventional models don’t take into account the characteristics of IoT data.Eventually, promising IoT missing data imputation techniques are crucial to avoid the perils and pitfalls of existing imputation methodologies.

REFERENCES

[1] Yan X, Xiong W, Hu L, Wang F and Zhao K. “Missing Value Imputation Based on Gaussian Mixture Model for the Internet of Things”, *Mathematical Problems in Engineering*,2015.

[2] Liu Y, Yang Y, Lv X, and Wang L. “A self-learning sensor fault detection framework for industry monitoring IoT”, *Mathematical problems in engineering*, 2013.

[3] Tian, Y., Ou, Y., Reza Karimi, H., Liu, Y.T. and Han, J.Q., “Distributed multitarget probabilistic coverage control algorithm for wireless sensor networks”, *Mathematical Problems in Engineering*, 2014.

[4] Leke C, Marwala T and Paul S. “ Proposition of a Theoretical Model for Missing Data Imputation using Deep Learning and Evolutionary Algorithms”, *arXiv preprint arXiv:1512.01362*. 2015.

[5] Allison PD. Missing data. Sage publications; 2001.

[6] Dong, Y. and Peng, C.Y.J., “Principled missing data methods for researchers”, *SpringerPlus*, **2**, pp.1-17.

[7] Munguía JA. “Comparison of Imputation Methods for Handling Missing Categorical Data with Univariate Pattern”, *Revista de métodoscuantitativospara la economía y la empresa*. **17**, 101-20, 2014.

[8] Gao, Zhipeng, Weijing Cheng, XuesongQiu, and LuomingMeng. “A Missing Sensor Data Estimation Algorithm Based on Temporal and Spatial Correlation”, *International Journal of Distributed Sensor Networks*, 2015.

[9] Ren, Xiaojun, HyonTaiSug, and HoonJae Lee. “A New Estimation Model for Wireless Sensor Networks Based on the Spatial-Temporal Correlation Analysis”, *Journal of information and communication convergence engineering*,**13**, 105-112, 2015.

[10] Kim DJ, Prabhakaran and B. Faulty, “Missing Body Sensor Data Analysis”, *In Healthcare Informatics*, 431-438, 2013.

[11] Alippi C, Boracchi G and Roveri M, “ On-line reconstruction of missing data in sensor/actuator networks by exploiting temporal and spatial redundancy”, *In Neural Networks*, 1-8, 2012.

[12] Peng, Li, Zhang Ting-ting, and Zhang Kai-hui. “Missing Value Imputation Method Based on Density Clustering and Grey Relational Analysis”, *International Journal of Multimedia and Ubiquitous Engineering*,**10**, 133-142, 2015.

[13] Sim, Jaemun, Jonathan Sangyun Lee, and Ohbyung Kwon. “Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications”, *Mathematical Problems in Engineering*, 2015.

[14] Ferrari, Gláucia Tatiana, and Vitor Ozaki. “Missing data imputation of climate datasets: Implications to modeling extreme drought events.” *RevistaBrasileira de Meteorologia*, **29**, 21-28, 2014.

[15] Kropko J, Goodrich B, Gelman A and Hill J. “Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches”, *Political Analysis*,2014.