DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5145

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

IMPUTING THE MISSING VALUES IN IOT USING ESTCP MODEL

Mrs. I. Priya Stella Mary
Ph.D Scholar
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli – 2, India

Dr. L. Arockiam
Associate Professor
Department of Computer Science
St. Joseph's College (Autonomous)
Tiruchirappalli – 2, India

Abstract : In Internet of Things, occurrence of missing data is inevitable due to its intrinsic characteristics. This missing data phenomenon occurs due to a variety of reasons such as uneven network communication, synchronization difficulties, untrustworthy sensor devices, environmental aspects and other device malfunctions which often resulted in data incompleteness. A robust approach to missing data is an indispensible component of analysis to promote the perfect explanation of research findings.As the data generated by the IoT devices is usually correlated in space and time, in this paper it is demonstrated experimentally that substituting missing sensor values with spatially and temporally correlated sensor readings using thenovel extended spatial and temporal correlated proximate missing data imputation model (ESTCP)has considerably improved the accuracy than that of the previously proposed STCP model and the existing single imputation and multiple imputation techniques.

Keywords: IoT , imputation, pre-processing

1. INTRODUCTION

Billions of devices such as smartphones, smart wearable devices, smart automobiles etc. are acknowledged and controlled using the Internet producing extraordinary volume of data brings the regime of the Internet of Things (IoT) [2]. This is the prime reason that IoT is considered as the next technological advancement of the Internet [3].

The IoT has copious amount of applications in numerous areas such as smart city,smart logistics, smart healthcare, smart transport systems etc. Even so, there exist numerous blockades that deter the fruition of the Internet of Things technology. One such blockade is the missing data problem. A vigorously fluctuating IoT environment unavoidably creates missing data problem. The calculation of missing values becomes a vital pre-processing step. Removing missing data results in loss of information, it implies a strong approach to missing data is an indispensible component of analysis to promote the perfect explanation of research findings.

In order to create a missing data estimation model, the key reasons for missing data are to be found first and then an examination of missing data mechanism is required. Missing data mechanism can be categorized as missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR). Finally an estimation model is to be built for the IoT depending upon its characteristics [4]. The existing machine learning algorithms don't take into account the characteristics of IoT as well as largely assume that the data is not incomplete so that all the records in the database are filled with valid values. Missing values are a common occurrence in IoT and can have a substantial influence on the inferences that can be drawn from the data [4]. If not imputed appropriately, it would result in imprecise, erratic analytical results. This stresses the novel missing data imputation methods in IoT to handle the data lost during sensing.

In this paper, a novel E-ST-correlated proximate missing data imputation model is proposed to deal with the missing data in the Internet of Things, based on spatial and temporal correlation of the IoT data. The proposed model will work efficiently even though the temporal and spatial correlations among sensed nodes are low. The efficiency of the model proposed in this paper is evaluated via carrying out tests on the real-world IoT datasets and compared with the other prevailing missing data imputation methods. The experimental outcomes proved that the proposed ESTCPmodel can calculate the missing data more precisely.

The rest of this paper is organized as follows. In Section 2, an overview of related works is presented. Section 3 presents the ESTCPmissing data imputation model, in Section 4 the experiments conducted using R tool demonstrate the accuracy of the proposed model in the IoT dataset and Section 5 concludes the paper.

2. RELATED WORKS

Song Gao et al. [5] proposed a missing data imputation algorithm based on least squares support vector machine and particle swarm optimization to deal with missing data problem which reduced the rationality of environmental radiation monitoring. Both the lagged data of the missing sensor node as well as neighbour nodes have been taken to do imputation. The experimental outcomes proved the efficiency of the proposed algorithm in terms of accuracy than neural network model as well as direct LSSVM model.

Chung-Yi Li et al. [6] proposed a novel imputation method that utilized the recommendation system while doing imputation. The proposed model was evaluated using two sensor datasets. Finally, the effect of imputation on the result of data analysis was examined and higher quality prediction model was built to do accurate imputation.Roozbeh Razavi-Far et al. [7] suggested missing data imputation methods in a pre-processing module to handle missing values in the empirical diagnostic systems.

The pre-processing module received and processed the residuals of observers before sending them to a fault classification module. This fault classification module learnt and classified the faults to facilitate imputation. The efficiency of the proposed method was proved in a doubly fed induction generator (DFIG) of a wind turbine.

SehyunTak et al. [8] proposed a modified k-nearest neighbour method based on spatial and temporal correlations to impute missing values in road networks. Multiple correlated sensor readings were taken into account for doing imputation which increased computational efficiency and imputation accuracy.Shin-Fu Wu et al. [9] proposed a new prediction method based on least squares support vector machine (LSSVM). Time series data as well as local time indexes were sent to LSSVM for performing prediction without imputation. The performance accuracy was compared with other imputation methods. Experimental results proved that the proposed method outperformed other imputation methods taken for comparison.

Liang Ze Wong et al. [10] proposed a missing data imputation method based on sparse auto encoder by exploiting spatial characteristics of sensor data. The auto encoder was modified to handle missing values and the proposed model was tested on a sensor test bed in Spain. The proposed method extracted important features from the dataset with large number of missing values and used the extracted features to accurately impute the missing data entries.YuanYuan Li et al. [11] developed a new imputation technique to deal with missing data problem in wireless sensor networks. Unsupervised fuzzy ART neural network was deployed to define missing data patterns on the sensor network and then missing data values were estimated based on spatial and temporal correlations technique. It was proved that the proposed technique outperformed the other nine imputation methods taken for comparative analysis.

Zhipeng Gao et al. [12] proposed a temporal and spatial correlation algorithm to deal with missing data problem in sensor networks. The sensed data were stored as time-series data and the samples were taken from the related time-series data to appraise missing values from the spatial and temporal dimensions to which weights were assigned. Simulations performed on the datasets taken, overtook the existing imputation methods in terms of accuracy.Shah Atiqur Rahman et al. [1] proposed an imputation technique called fuzzy logic k nearest neighbour that used time lagged correlations to impute missing data. The proposed technique is the combination of two other imputation techniques viz. k-NN and the Fourier transform. The efficiency of the proposed method was proved even when the percentage of missing values is high.Adrian Chong et al. [13] performed comparative analysis of five missing data imputation methodsviz. mean imputation, support vector machines(SVM), weighted k-NN, linear regression and replacing missing values with zero.Feature selection was done based on correlation which improved the performance of the imputation method. Based on the percentage of missing data, the strength of each method was also verified.It was established that the linear regression, kNN and SVM assessed missing data precisely than replacing with zero or mean imputation techniques.

3. METHODOLOGY

The statistical tool used to implement the proposed model is R. The air pollution dataset deployed for this research has been taken from city pulse IoT datasets collections. Four datasets have been taken and each comprises of time-series data with 17569 instances. The sampling frequency of each air quality sensor is 5 minutes. The dataset comprises of basic attributes such as time, sensor id to uniquely identify the sensors and observation attributes such as Ozone, Particulate Matter, Carbon Monoxide, Sulphur dioxide, Nitrogen dioxide. From the original dataset, the extracted target dataset comprises of the basic attributes time, sensor id, longitude, latitude to uniquely identify the sensors and only one observation attribute namely Carbon monoxide.

Using the data gathered, the performance of the proposed ESTCP missing data imputation model was appraised at different percentages (5%, 10% and 15%) of missing values.The STCP model proposed earlier will work efficiently and produce accurate results only if the correlation among the neighbouring sensors is high. In case if the correlation is less among the neighbouring nodes, the STCP model will produce less accurate results. To overcome this drawback in the STCP model, the extended STCP model is proposed to impute missing data using lagged variables even though the correlation among the neighbouring sensors readings is low. The new ESTCP model will impute missing values in time series data with lagged correlations. Thus, including lagged correlations in ESTCP model permits more accurate imputation of missing values when data have temporal correlations.

Since observations taken at time ‘t’ are likely to be correlated with observations taken at times ‘t−1’, than observations at t−2, t-3, t-4, t-5 and so on [1]. If data including lagged variables of 1 time step (t−1) is missing then , data including consecutive lagged variables up to 4 time steps (t−2,t−3 andt−4,t-5) will be taken and mean will be found.An extension to STCP with time lagged correlations has been proposed in this paper with the assumption that correlations may persist for a period of time in a sensor. The workflow of STCP and ESTCP is shown in the following Fig1.

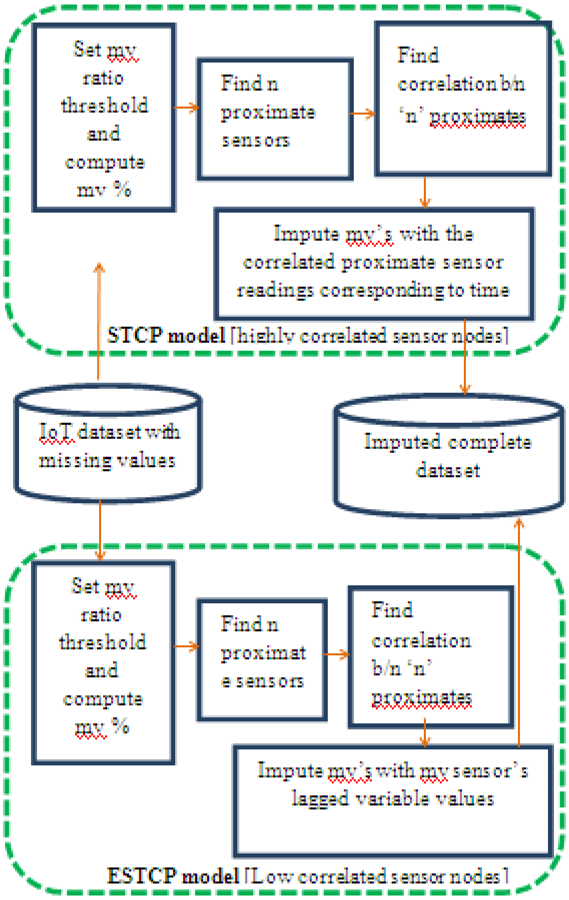
Set mv ratio threshold and compute mv %
Find n proximate sensors
Find correlation b/n 'n' proximates
Impute mv's with the correlated proximate sensor readings corresponding to time
STCP model [highly correlated sensor nodes]
IoT dataset with missing values
Imputed complete dataset
Set mv ratio threshold and compute mv %
Find n proximat e sensors
Find correlation b/n 'n' proximates
Impute mv's with mv sensor's lagged variable values
ESTCP model [Low correlated sensor nodes]

Fig.1 workflow of STCP and ESTCP

A. *ESTCP procedure for IoT missing data imputation*

Step 1: Set the missing value ratio threshold.

Step 2: Calculate the percentage of missing values in the chosen dataset.

Step 3: If the percentage of missing values does not exceed the threshold, then proceed to step 4 otherwise go to step 10.

Step 4: Find the n proximate sensors through Haversine formula using the geographical co-ordinates (spatial correlation).

Step 5: Find the correlation between the sensor with missing values and the n proximate sensors using the Pearson correlation co-efficient.

Step 6: If the correlation is high among n proximate sensors, then go to step 7 otherwise go to step 8.

Step 7: Impute missing sensor data with the correlated proximate sensor readings corresponding to time, got to step 9.

Step 8: Impute missing data using the lagged variables.

Step 9: Output complete dataset.

Step 10: Quantify the accuracy using RMSE (Root Mean Square Error) measure.

Step 11: Exit.

4.RESULTS AND DISCUSSIONS

The statistical tool used to implement the proposed model is R. The four IoT datasets have been taken for experiment and each dataset comprises of time-series data with 17569 instances. The sampling frequency of each air quality sensor is 5 minutes. From the original dataset, the extracted target dataset comprises of the basic attributes time, sensor id, longitude, latitude to uniquely identify the sensors and only one observation attribute namely Carbon monoxide.

First, Missing data threshold is set, though it differs from case to case. Then percentage of missing values is computed. If the percentage of missing values does not exceed the threshold set, then spatial-temporal correlation based imputation will be carried out. Otherwise imputation will be ignored.

Since there were no missing values in the chosen IoT air quality datasets, missing values have been introduced in sensor S1 randomly. Secondly, the 'n' proximate sensors to the Sensor S1 are identified using the Haversine distance formula. It has been found that for n=2, sensors S2 and S3 are closer to S1 than S4. So the proximate sensors to S1 are S2 and S3. Also, it has been found that S3 is geographically closer to S1 than S2. Thirdly, the correlations between the sensor S1 with missing values and the 'n' proximate sensors namely S2 and S3 have been found using the Pearson correlation co-efficient.

For the absolute value of 'r', the strength of the correlation can be found using the guidelines given in [14], which is represented in the following table I.

TABLE I
STRENGTH OF THE CORRELATION FOR THE ABSOLUTE VALUE OF 'r'

Absolute value of 'r'	**Strength of 'r'**
.00-.19	very weak
.20-.39	weak
.40-.59	moderate
.60-.79	strong
.80-1.0	very strong

The intensity of the correlation between the sensor S1 with missing values and the 'n' proximate sensors namely S2 and S3 using the Pearson correlation co-efficient is shown in the following table II.

TABLE II
CORRELATION COEFFICIENT OF THE SENSORS

Correlation between two sensors	**Correlation co-efficient (r)**	**Strength of 'r'**	**Type of correlation**
S1 and S2	0.331608	weak	positive correlation
S1 and S3	-0.3300029	weak	negative correlation

From the above table, it is clear that the correlations between proximate nodes are low. In this case, the ESTCPmodel for low correlated nodes is deployed for imputation rather than the STCP model which produced high accuracy results when the correlations between proximate nodes were high.

A. the proposed model

The imputation has also been performed using single imputation methods namely mean, median, mode and multiple imputation method namely MICE package in R. Comparative analysis of ESTCP model with other methods has been made which is shown in the following table III.

TABLE III
COMPARISON OF METHODS WITH THE PROPOSED ESTCP

Name	**Impute missing data**	**Lagged-relationships**	**Spatial-temporal correlations**
Mean	yes	no	no
Median	yes	no	no
Mode	yes	no	no
MICE	yes	no	no
STCP	yes	no	yes
ESTCP	yes	yes	yes

Since RMSE (Root Mean Square Error) is the usually deployed accuracy measure for assessing the performances in time series data [14], this measure has been used to compute the accuracy of these statistical techniques and the proposedESTCP missing data imputation model. Since the lower the RMSE, the more accurate is the evaluation [15]; it is shown in the following table IV

TABLE IV
RMSE FOR MEAN, MEDIAN, MODE, MICE AND STCP

RMSE	**5%**	**10%**	**15%**
Mean	12.30203	9.06369	9.068528
Median	12.69646	8.558621	9.380832
Mode	11.61895	8.558621	9.862386
MICE	12.45793	8.927486	8.985173
ESTCP	3.405877	2.983287	2.581989

.It has been established in the following Fig 2.

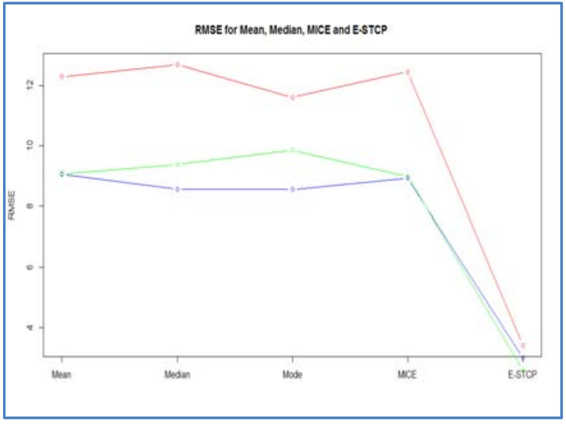

Fig 2

. RMSE for Mean, Median, MICE and proposed ESTCPmodel for 5%, 10%,15% missing values

that the accuracy of the proposed model outperformed the statistical techniques taken for comparative analysis at different percentages (5%, 10% and 15%) of missing values.

Performance of STCP and ESTCP model during high correlation proximate sensing nodes as well as low correlated sensing nodes is shown in the following table V and table VI.

TABLE V
ACCURACY OF STCP AND ESTCP FOR LOW CORRELATED SENSINGNODES

RMSE	**5%**	**10%**	**15%**
STCP	17.79326	14.28986	16.48434
ESTCP	3.405877	2.983287	2.581989

TABLE VI
ACCURACY OF STCP AND ESTCP FOR HIGHLYCORRELATED SENSING NODES

RMSE	**5%**	**10%**	**15%**
STCP	1.843909	3.03315	3.530817
ESTCP	3.405877	2.983287	2.581989

So for highly correlation among sensing nodes, STCP procedure would be invoked and for low correlated sensing nodes ESTCP procedure would be invoked to perform imputation.

5. CONCLUSION

The main contributions of the proposed ESTCP model are two-fold: i) Imputing missing data even though the correlation is low between sensor readings ii) Incorporating time lagged correlations between the variables during imputation.The objective of this research work is to exemplify the importance of imputing missing data in IoT and improve the accuracy. It has been ascertained that imputation performed using the proposed ESTCP imputation model is more accurate than the single imputation methods namely mean, median, mode and multiple imputation using MICE package in R.

6. REFERENCES

[1] Rahman, Shah Atiqur, Yuxiao Huang, Jan Claassen, and Samantha Kleinberg. "Imputation of Missing Values in Time Series with Lagged Correlations", In Data Mining Workshop (ICDMW) IEEE International Conference, 2014, doi:10.1109/ICDMW.2014.110, pp. 753-762.

[2] Vongsingthong, Suwimon, and Sucha Smanchat. "A Review of Data Management in Internet of Things." Asia-Pacific Journal of Science and Technology, Vol. 20, No. 2, 2015, pp. 215-240.

[3] Dave Evans, “The Internet of Things How the Next Evolution of the Internet Is Changing Everything”, Cisco White Paper, 2011, pp. 1- 10.

[4] Yan, Xiaobo, Weiqing Xiong, Liang Hu, Feng Wang, and Kuo Zhao. "Missing value imputation based on Gaussian mixture model for the internet of things." Mathematical Problems in Engineering,Article ID 548605, Vol.15, 2015, pp. 1-8.

[5] Gao, Song, Yaogeng Tang, and Xing Qu. "LSSVM based missing data imputation in nuclear power plant's environmental radiation monitor sensor network." In Advanced Computational Intelligence (ICACI), IEEE Fifth International Conference, 2012,ISBN: 978-1-4673-1744-3, pp. 479-484.

[6] Li, Chung-Yi, Wei-Lun Su, Todd G. McKenzie, Fu-Chun Hsu, Shou-De Lin, Jane Yung-jen Hsu, and Phillip B. Gibbons. "Recommending missing sensor values." In Big Data (Big Data), IEEE International Conference, 2015,ISBN: 978-1-4799-9926-2, pp. 381-390.

[7] Razavi-Far, Roozbeh, and Mehrdad Saif. "Imputation of missing data for diagnosing sensor faults in a wind turbine." In Systems, Man, and Cybernetics (SMC) IEEE International Conference, 2015, DOI 10.1109/SMC.2015.30, ISBN:978-1-4799-8697-2, pp. 99-104.

[8] Tak, Sehyun, Soomin Woo, and Hwasoo Yeo. "Data-driven imputation method for traffic data in sectional units of road links." IEEE Transactions on Intelligent Transportation Systems, Vol 17, No. 6, 2016, pp. 1762-1771.

[9] Wu, Shin-Fu, Chia-Yung Chang, and Shie-Jue Lee. "Time series forecasting with missing values." In Industrial Networks and Intelligent Systems (INISCom), IEEE International Conference, 2015, ISBN: 978-1-63190-022-8, pp. 151-156.

[10] Wong, Liang Ze, Huiling Chen, Shaowei Lin, and Daniel Chongli Chen. "Imputing missing values in sensor networks using sparse data representations",In Proceedings of the 17th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems, 2014, DOI: http://dl.acm.org/citation.cfm?doid=2641798.2641816, pp. 227-230.

[11] Li, YuanYuan, and Lynne E. Parker. "Classification with missing data in a wireless sensor network", In Southeastcon IEEE International Conference, 2008,DOI: 10.1109/SECON.2008.4494352, pp. 533-538.

[12] Gao, Zhipeng, Weijing Cheng, Xuesong Qiu, and Luoming Meng. "A missing sensor data estimation algorithm based on temporal and spatial correlation." International Journal of Distributed Sensor Networks, Article ID 435391,DOI:http://dx.doi.org/10.1155/2015/435391, 2015, pp.1-10.

[13] Lam, Khee Poh, Weili Xu, Omer T. Karaguzel, and Yunjeong Mo. “Imputation Of Missing Values In BuildingSensorData" IBPSA-USA Journal, Vol.6, No. 1,2016 .

[14] Wuensch, K.L., “Straightforward Statistics for the Behavioral Sciences”, Journal of the American Statistical Association, Vol.91, No.436, 1996, pp.1750-1752.

[15] Schmitt, Peter, Jonas Mandel, and Mickael Guedj. "A comparison of six methods for missing data imputation", Journal of Biometrics & Biostatistics, ISSN: 2155-6180, DOI: 10.4172/2155-6180.1000224, Vol.6, No. 1, 2015, pp.1-6.