ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse252952020.pdf
https://doi.org/10.30534/ijatcse/2020/252952020

Data Imputation in Wireless Sensor Networks using Regression Models

A Sheela[1], D Karunkuzhali[2], D Bhuvana Suganthi[3], A Anushya[4], G Ramprabu[5]
[1]Assistant Professor, Department of Computer Science and Engineering, Sri Sairam Engineering College, Chennai, Tamil Nadu, India
[2]Professor, Department of Information Technology, Panimalar Engineering College, Chennai, Tamil Nadu, India
[3]Associate Professor, Department of Electronics and Communication Engineering, BNM Institute of Technology, Bangalore, Karnataka, India
[4]Assistant Professor, Department of Computer Science, St. Jerome's College, Nagercoil, Tamil Nadu, India
[5]Professor, Department of Electronics and Communication Engineering, Bonam Venkata Chalamayya Institute of Technology and Science, Amalapuram, Andhra Pradesh, India

ABSTRACT

Wireless sensor nodes and its inconsistency in reporting sensory data information tend to inaccurate processing at sink. Imputing information of sensors with collocated sensory information is needed to provide intermittent less processing at sink. This work deals with regression model in selecting the nearest sensor based upon nature of dependant variable. Poisson Regression based Imputed Data Information (PRIDI) has been used when there are equidispersion criteria of dependant variable is observed with collocated sensors. Negative Binomial Distribution model using Imputed Data Information (NBDIDI) has been used when over dispersion criteria of dependant variable is observed with collocated sensors. Thus both protocols estimate the imputed information considering the nature of collocated sensor data and type of reported interval information acquired at sink. Simulation comparison has been done with discrete event simulator for proposed protocols in network simulator 2.

Key words: Wireless Sensor Networks, Regression Analysis and Data Imputation.

1. INTRODUCTION

Sensor data acquisition and data imputation is a challenging approach in an undesirable wireless environment and its reporting probabilities to sink. The TinyDB approach minimizes the power consumption by priorly knowing the location coordinates and acquisition cost of data. The limitation of this approach is they discard the tuples when the wok-load is heavy thus causing waste of power in data acquisition [1]. Reliability aspect of sensor nodes and sink has been discussed using two approaches event reliability and query reliability. Event reliability denotes when sensors communicate the desired event to sink. Query reliability denotes the query of sink send to sensors. An Asymmetric reliable transport mechanism has been used to study the classification in essential nodes and non-essential nodes. The protocol uses acknowledgment and negative acknowledgement which increases the overhead of communication [2].

Missing data has been classified into three, namely: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). The MCAR states the missing variable does not depend on exogenous factors. MAR denotes the missing variable might relate to know values. NMAR states the missing variable depends on itself [6]. The taxonomy of missing value imputation has been denoted as either using "statistical methods" or "machine learning algorithms" [13]. The strategy of using collocated sensors to counter the imputation of sensor data imposes burden on its sensing and communication capability causing reliability issues. Aggregated query method to fill incomplete data in its time interval has been discussed. The impact of lower and upper bound of imputed data to the desired query using sum and count methods are compared with ground truth table [14]. The data driven model requires complete data sets to analyse the impact of query which cannot be modelled accurately in wireless environment.

In scenario of missing connectivity among sensors [20] in multi-hop communication sensors suffer from communication void. This accumulates sensed information within a sensor so subsequently checking communication [19] and sensing module are vital in a report interval. In [18] a review of imbalanced data and its associated impact on under sampling and over sampling has been discussed with majority of approaches based on KNN classifier. This work underlies on two mathematical models namely: PRIDI and NBDIDI which relies on information of gathered data to predict the response variable of imputation considering a group of sensors.

The outline of the paper is given as follows. Section 2 discusses the imputation methods in sensor and problem description. Section 3 discusses the proposed system

Regression models. Section 4 deals with the proposed system model and its simulation implementation. Section 5 concludes the overall work and further discussion.

2. RELATED WORK

Sequence to sequence imputation model (SSIM) has been discussed using a variable length sliding window to detect missing reading in sensors. The SSIM model consists of bidirectional LSTM encoder and unidirectional LSTM decoder. Bidirectional LSTM encoder uses past and future values based time indexes and produces large number of training samples to identify missing data [3]. SSIM does not provide spatial information of sensors which might results loss information in sensing region. Relationship exploiting data packet generated (offered load) and productivity (throughput) has been estimated using linear regression relationship [4]. However, modelling of missing data to enable transfer rate in accordance to fitting model is not considered. In [5], Statistical characterization of energy consumption with time series data to predict the energy and communication void has been discussed with sensor nodes. The reliability of nodes has been much focused on networking nodes for communication rather than data acquisition and missing values [5]. Iterative Imputation Network (IIN) discussion states that latent variable of co-located sensor in the same time period has been used to find the undesirable missing sensor values [7].

The geo-distributed sensors in an IIN topology must have strong correlation among data procurement to produce unbiased results of missing data. Interpreting the missing data of particular sensors is associated localized nearby sensors and its sensing data coefficients. These coefficients are used as input to Deep Neural Network architecture to find the consecutively missing values [8]. A description of wireless sensor denoted with continuous sensing and intermittent connectivity has been discussed using Markovian models. The storage capacity has been estimated in a per node basis [9]. The methodology fails during longer waiting duration in missing connectivity to sink.

Deep Belief Network (DBN) model has been proposed to find the missing values considering the spatial and temporal data values of the nearest sensor. The selection of the nearest node is being achieved with similarity filter [10]. Resolving scalability issues in multiple nodes is missing in the particular work. "Hybrid multiple imputation" [11], the method has been identified for identifying missing data which occurs in arbitrary and monotone manner. The approach first identifies the cluster level information during data acquisition, then pre-processing the data and decision level analysis improves the robustness by avoiding large missing ratio.

Extension of Multivariate Imputations by Chained Equation (MICE) [15], has been used to find high missing rate of data using competition methods instead of linear regression. The completion methods are being selected based on the non linear relationship of data and the associated cycle. Combined approach with machine learning and a consensus model has been proposed for replacing the missing data at edge nodes in IOT networks [16]. The approach provides need of information gathered information location and reporting time coordinates as important metrics in imputation.

2.1 Sensing and Communication Issues in Wireless Sensor Networks

"Probabilistic matrix formulation" (PMF) [12], based approach has been used to identify the missing possibilities of sensed values exploiting sensing possibilities of co-located sensors. Initially, K means clustering is done followed by data normalization then with the help of feature vectors of collocated sensor missing data is identified. Depending on the "Root Mean Square Error" values of recovered matrix the feature vector is being updated to minimize the error values. In [17] physical sensor values are imputed with virtual sensor values considering multi layer perception and genetic algorithm. Kalman filter calculates the faulty values at the associated time and multilayer perception imputes virtual sensor value to missing physical sensor value.

2.2 Problem Description

The challenging issue in obtaining the missing value is to either go with a normal routine transfer of data or to interrupt the process and include an appropriate estimate of missing values. This problem has been demystified in wireless sensor nodes into imputation that occurs during sensing process. The other is imputation that occurs during the communication process. Inferential points in calculating the status of sensor nodes at appropriate interval provides missing coefficient in sensing and communication.

3. PROPOSED SYSTEM

3.1 Motivation

Each sensor has a different follow up time in reporting data to sink. So in case of collocated sensor the redundant information provided matches with the imputed value this has been modelled using Poisson regression model.

3.2 Poisson Regression based Imputed Data Information (PRIDI)

The outcome variable "y" the number of instants sensor report data information to sink. The rate is denoted by equation 1.

$$\text{Rate} = \frac{\text{Count}}{\text{Reporting Time}} \quad (1)$$

Poisson mass function is given by equation (2).

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad (2)$$

The occurrence of the event within the report time is given by λt shown in equation 2.

$$\ln(Rate) - \ln\left(\frac{Y}{t}\right) - b_0 + b_1x_1 + \cdots . b_kx_k \quad (3)$$

The rate equation has been rewritten in equation (3) and equation (4).

$$\ln(Y) - \ln(t) = b_0 + b_1x_1 + \cdots . b_kx_k \quad (4)$$

The dependant variable is denoted using (Y) and independent variable is denoted by using b_0, b_1 and b_2. Poisson distribution is suitable in application where mean is equal to variance. In scenarios where the mean is greater than variance it leads to under-dispersion. In scenarios of mean is less than variance it leads to over-dispersion.

3.3 Negative Binomial Distribution model using Imputed Data Information (NBDIDI)

The equation to determine the variance is given using equation 5.

$$\sigma^2 = \mu + \frac{\mu^2}{R} \quad (5)$$

The variance is denoted using σ^2 and mean is denoted using μ and the ancillary parameter is denoted using "R". Report interval is high and the number of instant sensor data available at the sink is high.

The negative binomial distribution has a negative parameter which models the variance in data using a "dispersion parameter" which reports unreported information at the MAC layer reducing communication void.

3.4 Algorithm for forwarding

Step 1: Sensors initialize itself for sensing and communication range in the deployed terrain.

Step 2: Report interval for each sensor and its multi hop path with distance are calculated.

Step 3: The first sensors data information is gathered in a sub terrain area. Assumption is made such that no sensing or communication void occurs initially post deployment.

Step 4: Then Kolmogrov Smirnov test has been to check where the underlying data follows:

Poisson Regression or Negative Binomial Distribution.

$$D = max_{1 \le A \le N}\left[F(A_i) - \frac{i-1}{N}, \frac{i}{N} - F(A_i)\right] \quad (6)$$

Notation of "N" denotes total count of data and $F(A_i)$ denotes frequency of occurrence of data.

Acceptance criteria has been determined by equation 7 where D_a denotes level of significance with default value 0.565.

$$\text{Acceptance criteria} = D < D_a \quad (7)$$

Step 5: Select either Poisson Regression or Negative Binomial Distribution based on sensory Information and report interval.

4. EXPERIMENTAL RESULTS

Network simulator version 2 has been used for development of the proposed protocols. Comparison study of imputation data has been estimated with 10%, 20% and 30% of sensor data imputation as tolerable limits associating it to the simulation time. Significant parameter used in the simulation environment is shown in table 1.

The root mean square error is computed for throughput of predicted model and observed model and plotted against the time period. It is given by equation 8 below notation "p_t" denotes predicted throughput and actual throughput is denoted using "a_t" and "n" denotes instances.

Table 1: Significant Simulation Parameters used for Analysis

Parameter	Value
Number of sensors	50
Number of sinks	4
Terrain	400 m ×300 m
Sensing range	40 m
Communication range	80 m
Packet size	512 B
Initial energy	5 J

Figure 1: Report Interval versus Root Mean Square Error of Throughput

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(p_t - a_t)^2}{n}} \quad (8)$$

The figure 1 shows that RMSE is minimal in a report interval of 500 seconds observed in NBDIDI model when compared to PRIDI. The significant improvement is achieved in term imputed collocated sensor value obtained are subjected to over dispersion matched with NBDIDI.

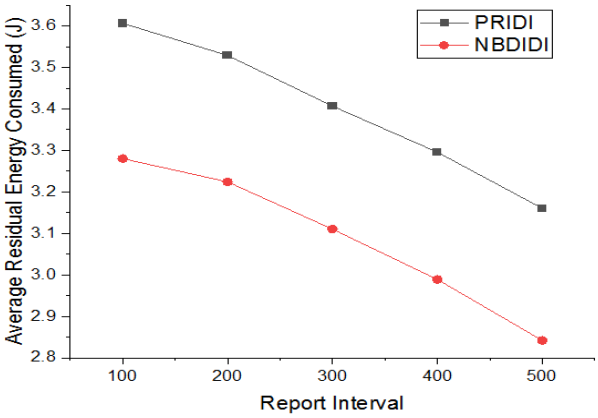

Figure 2: Report Interval versus Average Residual Energy Consumption

The figure 2 shows the average residual energy of nodes divided by the total number of nodes plotted against report interval from 100 seconds to 500 seconds. More energy is being consumed in NBDIDI with more transmissions than the PRIDI model with lesser transmissions.

5. CONCLUSION

The prediction of data to be imputed has been estimated with two count data models and its outcome associated in the previous time intervals. Thus imputing variables are being estimated within the location and report time coordinates considered. Apportion of predicting the missing variable has been estimated with count data models using estimation at sink. Thus the parameter of imputation are considered with nature of reporting at processing centre provides unbiased results in terms of imputation. Further research will deal in nature of bounded delay and application of impute data in sensor actuator networks to analyse the reliability metrics.

REFERENCES

1. Madden, S. R., Franklin, M. J., Hellerstein, J. M., & Hong, W. (2005). TinyDB: an acquisitional query processing system for sensor networks. ACM Transactions on database systems (TODS), 30(1), 122-173.
2. Tezcan, N., & Wang, W. (2007). ART: an asymmetric and reliable transport mechanism for wireless sensor networks. International Journal of Sensor Networks, 2(3-4), 188-200.
3. Zhang, Y. F., Thorburn, P. J., Xiang, W., & Fitch, P. (2019). SSIM—A deep learning approach for recovering missing time series sensor data. IEEE Internet of Things Journal, 6(4), 6618-6628.
4. Anand, J. V., & Titus, S. (2014, October). Regression based analysis of effective hydrocast in underwater environment. In TENCON 2014-2014 IEEE Region 10 Conference (pp. 1-6). IEEE.
5. Anand, J. V., & Titus, S. (2017). Energy efficiency analysis of effective hydrocast for underwater communication. International Journal of Acoustics & Vibration, 22(1), 44-50.
6. RJ, A. (1987). Little, DB Rubin, "Statistical analysis with missing data" John Wiley and Sons, New York,.
7. Zhou J & Huang Z. "Recover missing sensor data with iterative imputing network," in Proc. Workshops 32nd Artif. Intell. (AAAI) Conf., New Orleans, LA, USA, Feb. 2018, pp. 209–215.
8. Liu, F., Li, H., & Yang, Z. (2018). Estimation method based on deep neural network for consecutively missing sensor data. Radioelectronics and Communications Systems, 61(6), 258-266.
9. Sydora, C., Jung, J., & Nikolaidis, I. (2019, October). A Study of Simple Partially-Recovered Sensor Data Imputation Methods. In 2019 15th International Conference on Network and Service Management (CNSM) (pp. 1-7). IEEE.
10. Du, J., Chen, H., & Zhang, W. (2019). A deep learning method for data recovery in sensor networks using effective spatio-temporal correlation data. Sensor Review. 39 (2), pp.208-217.
11. Lin, J., Li, N., Alam, M. A., & Ma, Y. (2020). Data-driven missing data imputation in cluster monitoring system based on deep neural network. Applied Intelligence, 50(3), 860-877.
12. Fekade, B., Maksymyuk, T., Kyryk, M., & Jo, M. (2017). Probabilistic recovery of incomplete sensed data in IoT. IEEE Internet of Things Journal, 5(4), 2282-2292.
13. García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. Neural Computing and Applications, 19(2), 263-282.
14. Zhang, A. Z., Li, J. Z., & Gao, H. (2019). Interval Estimation for Aggregate Queries on Incomplete Data. Journal of Computer Science and Technology, 34(6), 1203-1216.
15. Ratolojanahary, R., Ngouna, R. H., Medjaher, K., Junca-Bourié, J., Dauriac, F., & Sebilo, M. (2019). Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. Expert Systems with Applications, 131, 299-307.
16. Kolomvatsos, K., Papadopoulou, P., Anagnostopoulos, C., & Hadjiefthymiades, S. (2019, September). A Spatio-Temporal Data Imputation Model for Supporting Analytics at the Edge. In Conference on e-Business, e-Services and e-Society (pp. 138-150). Springer, Cham.
17. Matusowsky, M., Ramotsoela, D. T., & Abu-Mahfouz, A. M. (2020). Data Imputation in Wireless Sensor Networks Using a Machine Learning-Based Virtual Sensor. Journal of Sensor and Actuator Networks, 9(2), 25.
18. Patel, H., Singh Rajput, D., Thippa Reddy, G., Iwendi, C., Kashif Bashir, A., & Jo, O. (2020). A review on classification of imbalanced data for wireless sensor networks. International Journal of Distributed Sensor Networks, 16(4), 1550147720916404.

19. Shyamala Bharathi, P., Venkatasamy, C.V., Latha, K., Shyamala, K., Ramprabu, G, "Wireless sensor networks based intelligent localization in hostile environment for energy conservation", International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(4), pp. 4783-4787.
20. Vijayan, K., Ramprabu, G., Selvakumara Samy, S., Rajeswari, M., "Cascading Model in Underwater Wireless Sensors using Routing Policy for State Transitions", Microprocessors and Microsystems, 2020, 79, 103298.