Aro-Nims: A New-Fangled Framework for IoT Data Pre-processing

Mrs. I. Priya Stella Mary[1], Dr. L. Arockiam [2]
[1]*Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli – 2*
[2] *Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli – 2.*
e-mail: priyastellamary123@gmail.com[1], larockiam@yahoo.co.in[2]

Abstract*:* Evolving technologies in recent years and major augmentations to the Internet and computing systems have facilitated the communication between divergent devices much simpler. According to various projections, around 25-50 billion devices are anticipated to be linked to the Internet by 2020. This leads to the rise of the advanced concept of Internet of Things (IoT). Billions of IoT devices will be linked to the internet and so handling such a massive amount of data generated by these devices is undoubtedly a matter of great concern. Since the data, being emitted by these devices are often incomplete, this makes analysis process cumbersome. As a result, preparing incomplete data to an unwavering and consistent form is a vital pre-processing step in IoT that can massively enable successive handling of the data simpler. Though there exist several traditional pre-processing frameworks to pre-process data, most of them are not appropriate for pre-processing IoT data, because these frameworks don't consider the temporal and spatial correlation of IoT data during pre-processing. To overcome this drawback, a new-fangled IoT data pre-processing framework is proposed in this paper by carefully considering all the characteristics of IoT data.

Index Terms- IoT , framework, pre-processing

1. INTRODUCTION

IoT has provided the physical inanimate world a "digital nervous system". With the influx of this technology [1], several industries viz. manufacturing, healthcare, transport etc., have multiple ways of increasing profit from the IoT sector. This evolving hyper-connected IoT environment will sense and transfer data, thereby increasing the demand for IoT data pre-processing. Raw data are being generated at an increasing rate by a multitude of IoT devices deployed at various physical areas that include Smart homes, Smart cities, Healthcare, Retail, Agriculture, Transportation, Manufacturing and more [10]. All these interconnected IoT devices spewing zettabytes of data that is almost literally unimaginable [9]. Since IoT data will come from different sources with varying formats and structures. Data may need to be pre-processed to handle noise, missing data, outliers and inconsistencies [2]. A soaring data volume inevitably leads to a growing demand for IoT data pre-processing. Pre-processing such a massive amount of data and extracting valuable business information from it, becomes the biggest challenge.

Spatial and temporal correlation is the noteworthy and exclusive characteristic of IoT data that can be exploited to considerably enrich accuracy of analysis. Existing pre-processing techniques don't consider the heterogeneous nature of IoT data and so the deployment of these techniques in the IoT environment to perform data preparation tasks produces biased, inaccurate or misleading results. More fundamentally, the nature of the errors in IoT data cannot be easily corrected by traditional pre-processing solutions which don't consider the temporal or spatial aspects of data. Overcoming these loopholes inherent in the existing pre-processing system and capitalizing on them is the need of the hour to spearhead new pre-processing techniques. In this paper, a unified framework Aro-Nims pre-processing framework comprises of three pre-processing models is proposed to pre-process IoT data by carefully considering the spatial, temporal, attribute and time-lagged correlations so as to enhance accuracy of subsequent analysis.

The rest of this paper is organized as follows. In Section 2, an overview of related works is presented. Section 3 presents the new-fangled Aro-Nims IoT data pre-processing framework, in Section 4 the experiments conducted using R and MongoDB tools demonstrate the accuracy of the proposed framework in the IoT dataset. Section 5 presents the results and discussion and Section 6 concludes the paper.

2. RELATED WORKS

Monidipa Das et al., [3] proposed a deep-learning-based-framework to impute missing data to enable accurate analysis with remote sensing time series data. The missing data were imputed based on the preceding and subsequent timestamps. During imputation, the causality constraint in spatial and temporal analysis was preserved. In the proposed framework, a group of forecasting modules was constructed based on the observed time-series data. The coupling between the forecasting modules was done with the support of dummy data. These dummy data were predicted using the previous part of the time-series data. Later, the dummy data were increasingly enriched in an iterative method. Each forecasting module in the group was based on the Deep-STEP, a modified deep stacking network learning approach. The proposed work has also been validated through a case study on predicting missing data in vegetation time-series data. Comparative performance study established the efficiency of the proposed framework.

Weiwei Shi et al., [4] proposed a framework to enhance the accuracy of datasets by imputing missing values. The proposed framework combined two methodologies namely optimizing support vector machine (OSVM) and refining SVM (RSVM). Data training methods were placed on top of the OSVM engine. First the generated data were pre-processed through standardization and then the conventional SVM was trained to acquire a maiden prediction model. Next by using this prediction model, a new training dataset was obtained to attain optimized SVM predictors. Lastly the imputed missing data using optimized SVM were supplied as an input to the conventional SVM and the refined SVM was finally achieved. The proposed OR_MLF framework was tested on missing data prediction of power transformers in power grid system. An investigational study demonstrated that the predictors based on the proposed framework accomplished lower mean square error than conventional ones. It was also proved that the proposed framework was good at predicting the missing data in power grid system.

Collins Leke et al., [5] presented a novel framework to do missing data imputation in high dimensional datasets. Deep Learning technique was employed in association with a swarm intelligence algorithm. The proposed approach combined the benefits of deep auto encoder networks and the Bat algorithm to estimate the missing data. Five metrics namely standard error, correlation co-efficient, relative prediction accuracy, the signal-noise ratio and the global deviation were used to assess the performance of the proposed approach. The performance of the proposed framework was experimentally verified and compared with other prevailing missing data imputation approaches on an off-line dataset. But the drawback of the proposed missing data imputation framework was the computational time required to estimate the missing data that could be rectified by parallelizing the approximation process. The outcomes of the experimentation proved the potential of the proposed missing data imputation framework in improving accuracy of the imputation but with marginally lengthier execution times.

Ke Zhang et al., [6] presented a novel Local Distance-based Outlier Factor framework to quantify the outliers in distributed datasets. This factor has used the location of the object with respect to its neighbours to define object deviation from its neighbours. First the LDOF threshold has been set up and then the top-n outlier detection method was employed. Two properties of the factor namely the lower bound and false-detection probability were analysed and the technique for choosing k- values has been recommended. The parameter setting was simplified and the top-n method was deployed to easily select k- values. Theoretical bounds on the factor were defined for false detection probability. Two real world datasets and one synthetic dataset were taken to prove the efficiency of the presented method in detecting outliers accurately. The proposed framework has detected the outliers with high precision even though the size of the neighbourhood is high. It was proved experimentally that the proposed method has outperformed the classical k nearest neighbour algorithm as well as Location outlier factor based outlier detection algorithm.

Kun Niu et al., [7] have presented a novel outlier detection framework based on clustering to detect outliers in wireless sensor networks. During the clustering process, k-means clustering algorithm was used. The time slot for sampling the data was defined first and then the rational number of clusters was received for all timeslots of sensor nodes. Maximum cluster and minimum cluster were obtained after the clustering process. Depending upon the distances of the centre of the two clusters, time slots were allotted. The proposed framework has detected the latent outliers through the clustering labels of time slots. The proposed framework was compared with the existing distributed anomaly detection framework for wireless sensor network to appraise its ability. False positive rate and false negative rate were the two criterions used to determine the efficiency of the proposed framework. Experimental outcomes proved the effectiveness and the strength of the proposed framework on real world wireless sensor data sets.

Aymen Abid et al., [8] have proposed an outlier detector framework based on the distance between the current reading and its neighbours' readings. The assessment was done by introducing random values into the Intel Berkley lab real world database. Continuous tests were performed and at each test, the size of the learning window was increased to have more outlier readings. Outliers were detected through the proposed detector with an accuracy rate of 89% and low false alarm rate was an average of ten percentages. It was found that the detection rate of the detector could even attain 100%. The proposed detector framework has detected outliers with high detection rate, low false alarm rate and great accuracy.

3. METHODOLOGY

The proposed IoT data pre-processing framework is the foremost and indispensable step towards obtaining final complete dataset. After the application of the pre-processing models in the proposed Aro-Nims IoT data pre-processing framework, the final data set obtained, can be considered as a trustworthy and appropriate source for analytics applied afterwards. The Proposed Aro-Nims framework comprises of three models namely STCPOD (Spatial and temporal correlated Outlier Detection) model, STCP (Spatial and Temporal Correlated Proximate Model) and FRBIS (Fuzzy Rule-Based Imputation System) model as shown in the Figure 1.

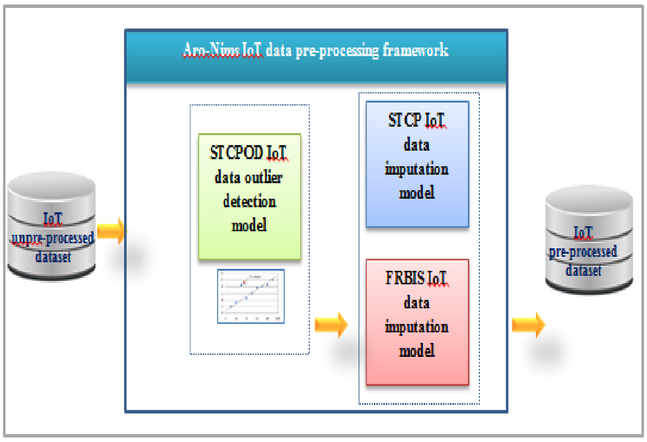
Aro-Nims IoT data pre-processing framework
IoT unpre-processed dataset
STCPOD IoT data outlier detection model
STCP IoT data imputation model
FRBIS IoT data imputation model
IoT pre-processed dataset

Figure 1 Aro-Nims IoT data pre-processing framework

3.1. STCPOD model

The proposed STCPOD (Spatial and temporally correlated proximate for Outlier Detection) model detects outliers in time-series data based on the characteristics of IoT data. It facilitates to attain high detection rate and low false alarm rate. STCPOD model comprises of three Procedures viz. Discover by Self-learning Procedure, Discover by Proximate Sensors Procedure and Outlier Determining Procedure. When a sensor reading is fed into the Discover by Self-learning Algorithm, it determines whether the reading that has been fed is usual or not. When the sensor node generates abnormal data, it can seek the help of its correlated proximate nodes to determine whether the sensor reading is an outlier or not.

Outliers generated by a sensor can be detected through the proposed Discover by proximate sensors algorithm by making comparison with the proximate sensors readings, i.e. an unusual reading is compared with the correlated proximate sensor readings corresponding to time to determine whether it is an outlier or not. The outlier determining algorithm will decide whether the unusual sensor reading is an outlier or not after receiving scores from the correlated proximate sensor nodes.

3.2. STCP model

The proposed STCP (Spatial and temporally correlated proximate) model imputes missing data in time-series data based on the characteristics of IoT data and facilitates IoT data analytics by enhancing the accuracy of imputation. The proposed STCP model is helpful in dealing with inconsistencies to produce complete dataset. As per this model, missing values are imputed either by using proximate sensor nodes readings corresponding to time or by using time-lagged correlations. First, Haversine formula is used to find ‘n’ proximate sensors and then the correlation between the sensor with missing values and the ‘n’ proximate sensors using the Pearson correlation co-efficient is found. Finally, missing sensor data are imputed with the correlated proximate sensor readings corresponding to time.

3.3. FRBIS model

The proposed FRBIS (Fuzzy Rule-Based Imputation System) imputes missing data in time-series data based on the characteristics of IoT data. According to this model, besides spatial and temporal correlations, attribute correlation is also considered for doing imputation. First, if the existence of attribute correlation is found, then the sensor with missing data is imputed with the values of the correlated attribute of the same sensor node otherwise, k nearest neighbours are found based on Euclidean distance. Secondly, the correlation between the sensor with missing values and the k nearest neighbouring sensors is found using the Pearson correlation co-efficient. Lastly, Depending upon the ratings assigned by the defined Fuzzy system, the top rated neighbouring sensor is found to impute missing sensor data corresponding to time.

After pre-processing the IoT data with the proposed framework, simple analytics is performed. It has been demonstrated that the proposed framework enhanced the accuracy of analytics.

4. EXPERIMENTAL SETUP

Devices needed to store the readings of carbon monoxide [11]MQ-7 sensors on cloud (ThingSpeak.com) using wifi modem NodeMCU are explained below.

4.1. Hardware components needed

The following are the hardware components needed as mentioned in Table 1

Table 1. Hardware Components

S.No.	Item	Quantity
1	NodeMCU	4
2	MQ-7 CO gas sensors	4
3	Gas sensor PCB board	4
4	4G- hotspot	1

4.2. Software Components

- ➢ Arduino IDE
- ➢ MongoDB

4.3. Building Circuit

MQ-7 gas sensor is connected with NodeMCU as shown in the following Table 2.

Table 2. NodeMCU Connectivity Details

S.NO.	NodeMCU	MQ7-CO
1.	Vin	VCC
2.	GND	GND
3.	D3	Data Out

After the circuit setup is over, four MQ-7 CO sensors have been deployed at various locations namely hall, kitchen, store room and car parking shed in a home.

4.4. Installation Locations

The sensors have been installed at various locations such as hall, kitchen, Store room, Car parking shed in a home as shown in Figure 2 a, 2 b, 2 c, 2 d.

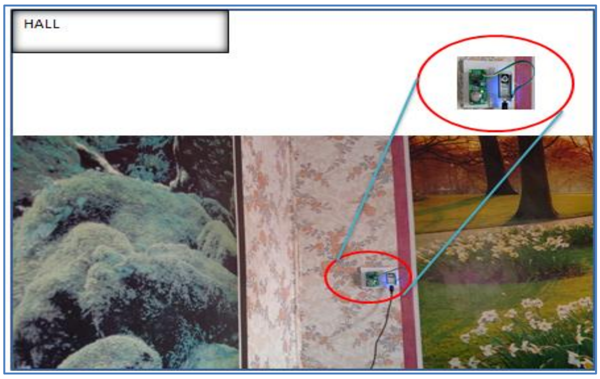

Figure.2 a Sensor ‘S4’ fixed at Hall

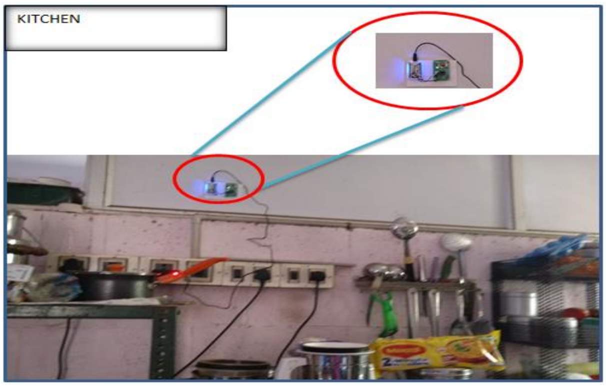

Figure.2 b Sensor ‘S3’ fixed at Kitchen

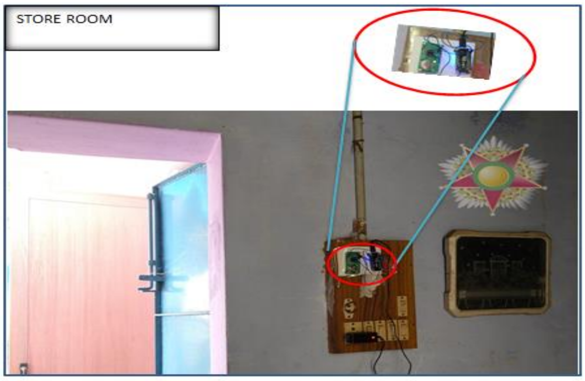

Figure.2 c Sensor ‘S1’ fixed at Store Room

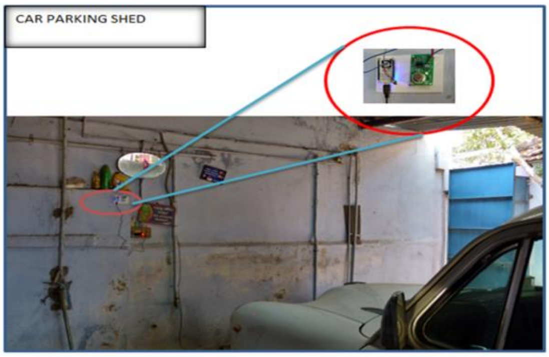

Figure.2 d Sensor ‘S2’ fixed at Car parking shed

Once the sensors’ setup has been completed, the next step is to send the readings sensed by the four CO sensors to Thingspeak.com

5. RESULTS AND DISCUSSIONS

Pre-processing of logged IoT data from Thingspeak.com is done using the proposed Aro-Nims IoT data pre-processing framework. According to the proposed framework, two types of data cleaning techniques are applied on the sensed IoT data.

- Outlier detection technique (STCPOD)
- Missing data imputation technique (FRBIS)

For this particular case study, FRBIS missing data imputation technique is used to do imputation rather than STCP missing data imputation technique. Since, the distance between the sensors installed at various locations in a home is relatively small, Euclidean distance measure deployed in FRBIS imputation model is enough to compute the distance between sensors rather than Euclidean Distance formula deployed in STCP model.

5.1. Setting up Threshold

It is significant to set up the proper threshold value based on the context.
Threshold is set to 0.5.
i.e. $\sigma = 0.5$

5.2. Finding the n proximate Sensors through Euclidean Distance Formula (spatial correlation)

The ‘n’ proximate sensors to the Sensor S3 installed at Kitchen, are identified using the Euclidean distance formula. It has been found that for n=2, sensors S1 and S4 are closer to S3 than S2. So the proximate sensors to S3 are S1 and S4. Also, it has been found that S1, S4 are geographically closer to S1 than S2. It is represented in the Table 3

Table 3. Euclidean Distance between Two Sensors

Euclidean Distance between Sensors	Euclidean Distance in metres
S3 to S3	0
S3 to S4	9
S3 to S2	21
S3 to S1	12

5.3. Finding the Pearson Correlation Co-efficient

The Pearson correlation co-efficient between the sensor with outliers and the ‘n’ proximate sensors is calculated, which is shown in the following Table 4. It has been found that the neighbouring sensor node S1 is the highly correlated sensor node than S2 and S4 nodes.

Table 4 Correlation Co-efficient between Sensors

Sensors	Correlation co-efficient (r)
S3 and S1	0.9126915
S3 and S2	0.1403636
S3 and S4	-0.007649843

Similarity measure has been found between the current reading and the preceding reading, until the final reading $S_n(t)$ is reached and then compared with the threshold to detect the presence of outliers. As there were only small percentage of outliers i.e. less than 2% of outliers in the sensed IoT data, 5%, 10%, 15% of outliers have been synthetically introduced.

The performance of the proposed outlier detection model in the proposed framework is assessed using the following metrics

- outlier detection rate
- False alarm rate

The proposed STCPOD model facilitates to attain high detection rate and low false alarm rate. After the detection and elimination of outliers, the FRBIS imputation model is applied to impute missing values. Since the parameters ‘distance’ and ‘correlation’ have already been calculated for the STCPOD model, the proposed FRBIS deploys these two variables to infer the linguistic variable ‘rating’. The proposed FRBIS model assigns ratings to the correlated neighbours which are sorted in the descending order as shown in the following Table 5. It has been found that sensor node ‘S1’ is the highly correlated nearest neighbour than the sensor nodes ‘S2’.

Table 5. Rating Table

Sensors	Distance	Correlation	Rating
S1	12 (S3→S1)	0.9126915	8
S2	21 (S3→S2)	0.1403636	5
S4	9 (S3→S4)	-0.007649843	5

As illustrated in the following Table 6, the rating value ‘best’ is assigned to the sensor ‘S1’ which will be taken for imputing missing values in sensor node ‘S3’ corresponding to time.

Table 6. Rating Scale

Rating	Values
8	Best
7	Very Good
6	Good
5	Average
4	Below average
3	Significantly below average
2	Poor
1	Very Poor
0	Worst

As there were only small percentage of missing values i.e. totally 10 missing values in the sensed IoT data, 5%, 10%, 15% of missing values have been synthetically introduced. The accuracy of the proposed FRBIS model in the proposed framework is assessed using the RMSE metric. Now the complete dataset obtained after eliminating outliers and imputing missing values, is supplied to analytical process.

5.4. Interpretation

After pre-processing the sensed IoT data on CO levels for a period of 2 months and 15 days (i.e. from 2017-12-01 14:48:54 UTC to 2018-02-15 11:30:34) using the proposed Aro-Nims pre-processing framework, it has been found that the mean CO level [16ppm] in the home where sensors have been installed is within the permissible range according to the OSHA guidelines. In case, the sensed IoT data on CO levels for a period of 2 months and 15 days have not been pre-processed, then it has been detected that the mean CO level [150.1333333 ppm] in the home, where sensors have been installed is not within the permissible range according to the OSHA guidelines. As a result, it causes unnecessary emotional and physical distress to family members. This case study demonstrates the importance of employing pre-processing task before doing any such analytics otherwise that will lead to detrimental consequences.

6. CONCLUSION

Conventional pre-processing frameworks are not appropriate for pre-processing heterogeneous IoT data from divergent sources and these frameworks when deployed in the IoT environment to carry out pre-processing tasks, yielded biased outcomes. Also the conventional frameworks don’t take into account the characteristics and unpredictable nature of IoT data. Ultimately, promising Aro-Nims IoT data pre-processing framework is crucial to avoid the jeopardies of existing data pre-processing frameworks. The proposed framework accomplishes high accuracy rate than the

REFERENCES

[1] Mahdavinejad, Mohammad Saeid, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi and Amit P. Sheth, “Machine learning for Internet of Things data analysis: A survey”, Digital Communications and Networks, 2017, pp. 1-57, DOI: 10.1016/j.dcan.2017.10.002.

[2] Salama Assahli, Mohammed Berrada and Driss Chenouni, “Data preprocessing from Internet of Things: Comparative study”, In IEEE International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2017, pp. 1-4. DOI:10.1109/WITS.2017.7934676.

[3] Das, Monidipa and Soumya K. Ghosh, “A Deep-Learning-Based Forecasting Ensemble to Predict Missing Data for Remote Sensing Analysis”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Volume 10, Issue 12, 2017, pp. 5228-5236. DOI: 10.1109/JSTARS.2017.2760202.

[4] Shi, Weiwei, Yongxin Zhu, Jinkui Zhang, Xiang Tao, Gehao Sheng, Yong Lian, Guoxing Wang and Yufeng Chen, "Improving power grid monitoring data quality: An efficient machine learning framework for missing data prediction", In 12thIEEE International Conference on Embedded Software and Systems (ICESS), 2015, pp. 417-422. DOI: 10.1109/HPCC-CSS-ICESS.2015.16.

[5] Leke, Collins, A. R. Ndjiongue, Bhekisipho Twala and Tshilidzi Marwala, "Deep Learning-Bat High-Dimensional Missing Data Estimator", In IEEE International Conference on Systems, Man and Cybernetics (SMC), 2017, pp. 483-488. DOI: 10.1109/SMC.2017.8122652.

[6] Zhang, Ke, Marcus Hutter and Huidong Jin, "A new local distance-based outlier detection approach for scattered real-world data", In Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Volume 5476, 2009, pp. 813-822. DOI: 10.1007/978-3-642-01307-2_84.

[7] Niu, Kun, Fang Zhao and Xiuquan Qiao, "An outlier detection algorithm in wireless sensor network based on clustering", In Communication Technology (ICCT), 15th IEEE International Conference, 2013, pp. 433-437. DOI: 10.1109/ICCT.2013.6820415.

[8] Aymen Abid, Abdennaceur Kachouri and Adel Mahfoudhi, "Anomaly detection through outlier and neighborhood data in Wireless Sensor Networks", In 2nd IEEE International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2016, pp. 26-30. DOI: 10.1109/ATSIP.2016.7523045.

[9] BahaaEldin, El-Shweky, Karim El-Kholy, Mahmoud Abdelghany, Mahmoud Salah, Mohamed Wael, Omar Alsherbini, Yehea Ismail, Khaled Salah and Mohamed AbdelSalam, "Internet of things: A comparative study", In 8th IEEE Annual Conference on Computing and Communication Workshop and Conference (CCWC), 2018, pp. 622-631. DOI: 10.1109/CCWC.2018.8301678.

[10] Ankush B. Pawar and Shashikant Ghumbre, " A survey on IoT applications, security challenges and counter measures ", In IEEE International Conference on Computing, Analytics and Security Trends (CAST), 2016, pp. 294-299. DOI: 10.1109/CAST.2016.7914983 [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.

[11] Laing, C., "Acute carbon monoxide toxicity: the hidden illness you may miss", Nursing, Volume 40, Issue 11, 2010, pp.38-43.

Authors Profile

1. ***I.Priya Stella Mary*** - received her Masters in Computer Applications from Bharathidasan University, Tiruchirappalli, India. At present she is doing Ph.D in the Department of Computer science at St.Joseph's College(Autonomous),Trichy affiliated to Bharathidasan University, India. She has published papers in the national and international journals (Scopus indexed). She has presented papers in the national/international conferences/national seminars. She has also attended various workshops on IoT, Big Data analytics and Data mining and also served as resource person in the short courses conducted. Her research interests are data mining and IoT.

2. ***Dr.L.Arockiam*** - is working as Associate Professor in the Department of Computer Science, St.Joseph's College(Autonomous),Trichirappali, Tamil Nadu, India. He has 28 years of experience in teaching and 20 years of experience in research. He has published more than 284 research articles in the international & National Conferences and Journals. He has also presented articles in the Software Measurement European Forum in Rome. He has chaired many technical sessions and delivered invited talks in international & National Conferences. He has authored 4 books. His research interests are: Cloud Computing, Big Data, and Cognitive Aspects in programming, Data Mining and Mobile Networks. He has been awarded "Best Research Publications in Science" for 2009, 2010 & 2011 and ASDF Global "Best Academic Researcher" Award from ASDF, Pondicherry for the academic year 2012-13 and also the "Best Teacher in College" award for the year 2014, 2015, 2016 and 2017.