Imputing the Missing Values in IoT using FRBIM

I. Priya Stella Mary

Abstract: ***The Internet of Things (IoT) is the new-fangled communication paradigm in which the internet is stretched out from the virtual world to intermingle with the objects in the physical world. It unleashes a new dimension of services but at the same time, colossal challenges have to be conquered to reap the full benefits of the IoT. One such challenge is missing data imputation in Internet of Things. The presence of missing values hampers the subsequent processes such as prediction, control, decision making etc. due to the dependency of these processes on complete information. In this paper, a novel FRBIM (Fuzzy Rule-Based Imputation Model) model is proposed to impute missing data based on the characteristics of IoT data to accomplish high accuracy rate. Experimental results have proved that the proposed method has outperformed the existing KNN and AKE imputation model in terms of accuracy.***

Keywords : IoT , imputation, pre-processing

I.INTRODUCTION

The Internet of Things has gained significant attentions in recent years, and can deliver intended results for a wide variety of applications, such as environmental monitoring, health monitoring, smart home etc. Data being sensed and collected from the real world applications are hardly flawless. This results in the occurrence of imperfections in the datasets. One of the most common imperfections occur in the IoT are missing values (MVs).

Datasets with missing values are a major hindrance for various learning algorithms that typically necessitate a complete data set to build the model. The presence of missing values in IoT pose several complications such as analytical errors, lack of efficiency, biases resulting from missing information etc. Eventually, the accuracy and reliability of outcomes from the experiments will be greatly compromised. Therefore it becomes essential to assess missing values in IoT [1].

Certain steps are to be followed before imputing missing data in IoT such as identifying missing data mechanism, learning missing data patterns etc. Finally apposite missing data imputation model for the IoT needs to be built to estimate missing values using the proposed model. Data gathered using IoT devices has several properties such as spatial correlation, temporal correlation, attribute correlation etc. While building the imputation model, all these characteristics should be taken into account to enhance the accuracy, reliability, and stability of missing value imputation efficiently.

Revised Manuscript Received on September 15, 2019

I. Priya Stella Mary, Assistant Professor Department of Computer Applications Bishop Heber College (Autonomous) Tiruchirappalli – 620 017 priyanirmal.bhc@gmail.com

The rest of this paper is systematized as follows. In Section 2, discussion is made on certain missing data imputation techniques. In Section 3, an overview of related works is presented. Section 4 presents the proposed FRBIM (Fuzzy Rule-Based Imputation Model) model, in Section 5 comparative analysis of the proposed FRBIM model is made with the existing imputation methods and Section 6 concludes the paper.

II. MISSING DATA IMPUTATION TECHNIQUES

Though there exist several traditional missing data imputation techniques, most of them are not applicable for handling missing sensor readings. Because these techniques don't take into account the temporal and spatial correlation of data during imputation. For these reasons, traditional missing data imputation techniques cannot be directly applied to the data sensed by the sensor devices [2]. Very few techniques such as WARM (Window Association Rule Mining), AKE (Applying K-nearest Neighbor Estimation) etc. are available to deal with missing sensor readings, but there are some pitfalls in these techniques too. While filling in the missing values, the missing data imputation method should retain the original distribution of the data under consideration. Imputation techniques can be categorized as follows [Roo, 2016]

- Model-based imputation methods
- Model-free imputation methods

Imputation techniques can be divided into

- Single imputation techniques
- Multiple imputation techniques

Single imputation techniques replace missing values with a single plausible value and form a single complete dataset, whereas the multiple imputation techniques replace missing values with multiple plausible values thereby forming different complete datasets.

III. RELATED WORKS

Sneha Arjun Dhargalkar et al. [3] have proposed an enhanced WARM (Window Association Rule Mining) method to impute missing values in the existing databases so that complete data could be obtained while querying the databases. Two newer forms of the WARM method namely Max-WARM and Pattern-WARM methods have been suggested and implemented. Experiment results proved that the proposed methods reduced the RMSE (Root Mean Square Error) and PCE measures than the existing WARM method and KNN imputation method. Phimmarin Keerin et al. [4] proposed a novel missing data imputation technique called CKNN impute which is an extension to the existing knn imputation method to impute missing values in microarray data by means of local data clustering in order to enhance the quality.

The performance of the proposed missing data imputation technique and other associated methods was assessed using the Normalized Root Mean Square Error (NRMSE) measure. Empirical assessment showed that the proposed CKNN technique outperformed the existing knn method. Kanwal Kiani et al. [5] proposed a novel hybrid missing data imputation method namely K-Nearest Temperature Trends (KNTT) to impute missing weather temperature data. The proposed hybrid missing data imputation technique was gauged using MAE, RMSE, ME accuracy measures and the experimental outcomes demonstrated that the proposed imputation approach has imputed missing values with an error rate less than that of the existing KNN method. Wei Chiet Ku et al. [6] presented a novel data-driven missing data imputation technique that deployed spatial and temporal correlation existing between the traffic flows of multiple road segments to impute missing data. Experiments conducted on the real world datasets, demonstrated the imputation accuracy of the proposed method at different missing data rates. Kumutha et al. [7] proposed a novel missing data imputation method using knn with bagging method to impute missing values. The bagging method fused several learners that used bootstrap samples of the original training dataset. The experimental outcomes established that the proposed method outperformed other imputation methods taken for comparative analysis in terms of distance and density of clusters. The proposed approach has improved the performance of the existing k-NN imputation method using the bagging method. Thanh Le et al. [8] proposed a new missing data imputation algorithm for imputing missing values in the datasets. The proposed method performed efficiently even though the datasets were not uniform. It has been shown that the proposed imputation algorithm outperformed other six popular imputation algorithms on both artificial and real world datasets with anonymous data distribution model. Dan Li et al. [9] proposed an attribute weighted fuzzy c-means algorithm to impute missing values. Attribute weighting was used to distinguish important attributes. The problem of grouping data with missing values was rectified. The experiments conducted have shown the efficiency of the proposed algorithm in imputing missing values so that better clustering results could be obtained than the existing methods taken for comparative analysis.Julie Yu-Chih Liu et al. [10] proposed a novel missing data approach to rectify the problem of missing values in the extended possibility-based fuzzy relational (EPFR) databases. Missing data imputation was proposed using fuzzy functional dependencies and their inference rules. The complexity of filling missing values in the EPFR databases was reduced through the proposed imputation technique. The outcomes of missing data imputation preserved fuzzy functional dependencies in the original data instance.Roozbeh Razavi-Far et al. [11] proposed a fuzzy-neighbourhood density-based clustering technique to impute missing data. The proposed technique was compared with other imputation techniques such as k-means imputation, fuzzy c-means imputation and fuzzy c-means with genetic algorithm imputation. The results obtained proved the efficiency of the proposed imputation technique over other missing data techniques taken for comparative analysis. Youness Riouali et al. [12] evaluated renowned spatial and temporal correlation based missing data imputation techniques such as ARIMA (Autoregressive integrated moving average) and AKE (K-Nearest Neighbour Based Missing Data Estimation Algorithm) to impute missing data, which is an unavoidable problem in wireless sensor networks. These spatial and temporal correlation based imputation models were reviewed to gain insights into the missing data problem in wireless sensor networks and summary of these methods was produced. The efficiency of the each of these models was appraised using the renowned accuracy metric RMSE (Root Mean Square Error). It was found that ARIMA is time consuming due to its requirement to be implemented online than the AKE model which could be used as offline technique for imputation.

IV. METHODOLOGY

In the proposed work, besides spatial and temporal correlations, attribute correlation is also considered for doing imputation. 'n' sensor nodes are set up in an area to observe different attributes at the same time. The observation period consists of t time slots. One day is split into 287 time slots. Each time slot comprises of 5 minute period. For example0:00-0:05, 0:05-0:10, 0:10- 0:15 and so on. The collected sensor data in one node can be structured in the following form.

Sensor ID	Time stamp	Attri1	Attri2	--	Lat	Lon

Where sensor ID represents sensor identity number, time stamp represents sampling time and the attributes can be ozone, carbon monoxide and so on. The latitude and longitude co-ordinates are useful for getting location details of the sensor. Let X - be a matrix of sensor data with k attributes collected by ith sensor within t time slots. Owing to data loss in IoT, X is usually an incomplete matrix. The proposed FRBIM (Fuzzy Rule-Based Imputation Model) model which is used to fill in the missing data is described below.

A. The proposed FRBIM (Fuzzy Rule-Based Imputation Model) model

In this section, the imputation model used to impute missing data is summarised. The imputation model involves the following 12 steps that are shown as follows

Input: Sensor Datasets

Output: dataset with no missing values

Process: Imputing missing data

Step (i) Start

Step (ii) Find the sensor dataset with missing values and the percentage of missing values.

Step (iii) Find the existence of attribute correlation

Step (iv) if attribute correlation exists, then impute the sensor with missing data with the values of the correlated attribute else go to step (v)

Step (v) Find the k nearest neighbours based on Euclidean distance

Step (vi) Find the Pearson correlation co-efficient between the sensor with missing values and the k

The workflow of the proposed FRBIM (Fuzzy Rule-Based Imputation Model) model is shown in Fig.1

Fig1. FRBIM imputation model

B. Defining Fuzzy Rule Based System

The proposed FRBIM is modelled by building Linguistic fuzzy rule-based system which finds the top rated neighbouring sensor node to be taken for imputation. Generally, for r variables (X1, X2…..Xr) which can take n values, there will be n·n···n= =nr total possibilities. Since the two linguistic variables 'distance' and 'correlation'(r=2) can take n=5 values each, there will be 5^2=5·5=25 total number of possible fuzzy rules that could be framed. The proposed linguistic fuzzy rule based system is a five stage process and is pictorially represented in the following Fig.2.

Stage (1) Setting up the universe

Stage (2) Defining the Linguistic variables

Stage (3) Defining Fuzzy rules

Stage (4) Building the system

Stage (5) Fuzzy inference and defuzzification

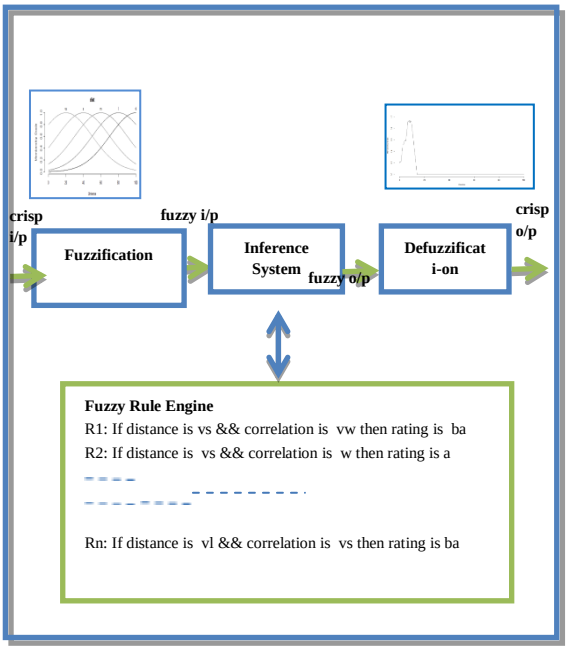

Fig.2. The proposed linguistic fuzzy rule based system

1) Setting up the Universe: The range of the universe is demarcated. Since it is assumed that all n sensor nodes are within the communication range of 100 meters, the range of the universe is between 0 and 100. The inputs for all the variables should fall within this range. If the variables defined are in different ranges, normalization has to be performed.

2) Defining Linguistic variables: Linguistic variables are made up of linguistic terms with associated degrees of membership. The values for these linguistic variables are words or sentences in a natural language. Fuzzy rules will be constructed based on the linguistic variables. A fuzzy linguistic variable is characterized by [A, T(A), U], A is the name of the variable; T(A) is the term set of A and U is the universe of discourse. All the values of a linguistic variable constitute its term set [Zad, 1975]. For example, the term set of a linguistic variable 'Distance' is defined as

T (Distance) = {Very Short, Short, Medium, Long, Very Long}

The proposed FRBIM model comprises of three linguistic variables namely

- Distance
- Correlation
- Rating

The first linguistic variable 'Distance' has five linguistic values:

- Very short
- Short
- Medium
- Long
- Very Long

The range of the linguistic values are [0-20], [20-40], [40-60], [60-80], [80-100] respectively. This means that if distance is 30, then it is categorized as 'short', if it is 70, it is 'long' and if it is 90, it is 'Very Long'.

The second linguistic variable 'Correlation' has five linguistic values:

- Very Weak
- Weak
- Moderate
- Strong

- Very Strong

The range of the linguistic values are [.00-.19], [.20-.39], [.40-.59], [.60-.79], [.80-1.0] respectively. This means that if the correlation is .15, then it is categorized as 'Very Weak', if it is .70, it is 'Strong' and if it is .95, it is 'Very Strong'.

The third linguistic variable 'Rating' has nine linguistic values range from 0 to 8 as shown in Table- I

Table- I :Linguistic values for Rating

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

3) Defining Fuzzy rules: After defining the variables, the next step is to define the fuzzy rules of the system. A fuzzy rule is expressed as a conditional statement in the following form.

IF p is X

THEN q is Y

Where p and q are linguistic variables; X and Y are linguistic values. In the proposed work, fuzzy rules are defined in the following form.

If (input variable1 is Xi) and (input variable 2 is Yi) and.....

then (output variable is Zi)

The framed 25 fuzzy rules associate the linguistic variables 'distance' and 'correlation' with the linguistic variable 'rating'. The first rule asserts that the rating will be 'BA' (below average) if the distance is very short and the correlation is very weak. The second rule asserts that the rating will be 'A' (average) if the distance is very short and the correlation is weak etc. and finally the 25th rule states that the rating will be 'BA' (below average) if the distance is very long and the correlation is very strong. The rules are reasonably comprehensible as English sentences. Weights are assigned to the linguistic values of the linguistic variables 'distance' and 'correlation'. Based on the weights assigned, ratings are determined as shown in the following Table- II.

Table- II : Weights assignment to the linguistic variables

Corr / Dist	Very Weak 0	Weak 1	Medium 2	Strong 3	Very Strong 4
Very Short 4	4	5	6	7	8
Short 3	3	4	5	6	7
Medium 2	2	3	4	5	6
Long 1	1	2	3	4	5
Very Long 0	0	1	2	3	4

4) Building the system: Once the rules and linguistic variables have been well-defined, the next step is to build the system using the defined linguistic variables and 25 fuzzy rules as shown below. The plot for the linguistic variables of the proposed system is shown in Fig. 3.a, 3.b, 3.c.

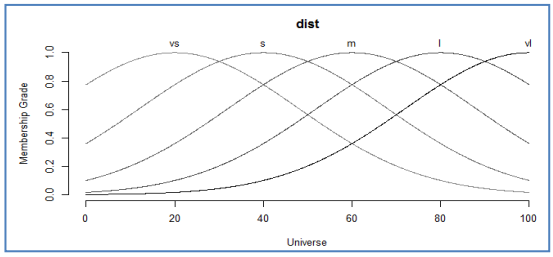

Fig.3.a membership function for the linguistic variable "distance"

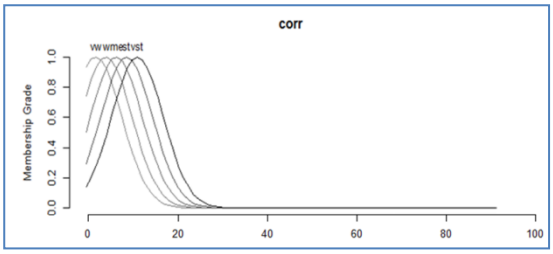

Fig.3.b. membership function for the linguistic variable "correlation"

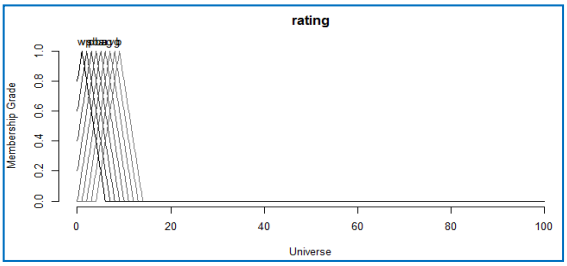

Fig.3.c. membership function for the linguistic variable "rating"

5) Fuzzy inference and defuzzification: Fuzzy inference refers to the process of combining membership functions with the fuzzy rules to obtain the fuzzy output [Yin, 2006]. Using the proposed fuzzy rule based system, the linguistic variable 'rating' is inferred. Fuzzy inference process involves specifying values for the linguistic variables 'distance' and 'correlation' and inferring the linguistic variable 'rating'.

Instance 1

Consider the neighbouring sensor 'S3' whose distance from the sensor with missing values 'S1' is 31 metre and correlation between them is '0. 9577101' then by using these two inputs the fuzzy inference system infers the linguistic variable 'rating' as shown in Fig. 4.a and the defuzzification process produces the defuzzified rating as '7.807779' ≈ '8'.

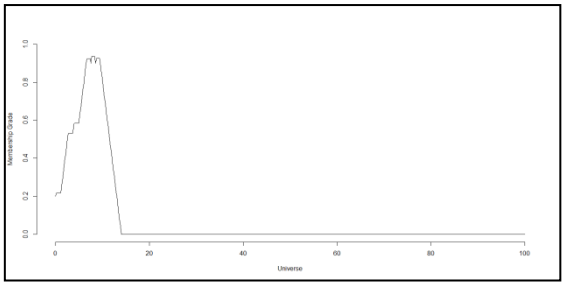

Fig. 4.a. Inferred membership for the linguistic variable 'rating'

Instance 2

Now consider the neighbouring sensor 'S4' whose distance from the sensor with missing values 'S1' is 18 metre and correlation between them is '0. 3455912' then by using these two inputs the fuzzy inference system infers the linguistic variable 'rating' as shown in Fig. 4.b and the defuzzification process produces the defuzzified rating as '5'.

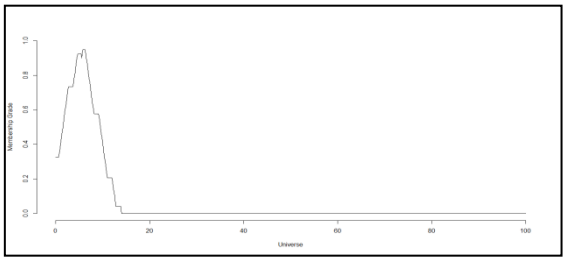

Fig. 4.b. Inferred membership for the linguistic variable 'rating'

Instance 3

Consider another neighbouring sensor 'S5' whose distance from the sensor with missing values 'S1' is 46 metre and correlation between them is '0. 7025665', then by using these two inputs the fuzzy inference system infers the linguistic variable 'rating' as shown in Fig. 4.c and the defuzzification process produces the defuzzified rating as '5'.

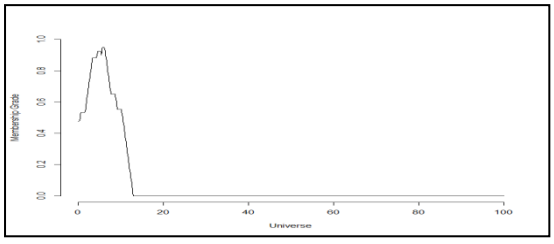

Fig. 4.c. Inferred membership for the linguistic variable 'rating'

C. Sorting correlating neighbours as per the ratings assigned by fuzzy system

The proposed FRBIM model assigns ratings to the correlated neighbours which are sorted in the descending order as shown in the following Table- III. It has been found that sensor node 'S3' is the highly correlated nearest neighbour than the sensor nodes 'S4' and 'S5'.

Table- III :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

D. Imputing missing sensor data with the top rated neighbour sensor readings

The sensor 'S1' with missing data is imputed using the readings of the top rated correlated neighbouring node 'S3' corresponding to time and complete dataset is produced.

V. COMPARATIVE ANALYSIS OF THE PROPOSED FRBIM MODEL WITH THE EXISTING IMPUTATION METHODS

The imputation has also been performed using the KNN and AKE imputation methods. Since RMSE (Root Mean Square Error) is the typically used accuracy measure for evaluating the performances in time series data, this measure has been used to compute the accuracy of the KNN, AKE imputation techniques and the proposed FRBIM missing data imputation model.

The KNN imputation method finds a set of k-nearest neighbours to the sensor with missing data and then replaces the missing data by averaging (non-missing) values of its neighbours. The AKE imputation method adopts the linear regression model to define spatial and temporal correlations of sensor data among neighbouring sensor nodes to estimate missing data jointly rather than individually.

Since the sensor datasets taken for experimentation don't contain any missing values, they are artificially introduced into the dataset. The performance of the proposed FRBIM imputation model was evaluated at different percentages (5%, 10% and 15%) of missing values. The RMSE is computed and outcomes are demonstrated in the following Table- IV

Table- IV:RMSE for FRBIM and KNN imputation

RMSE	5% of mvs	10% of mvs	15% of mvs
FRBIM	1.843909	2.213594	2.32379
KNN Imputation	2.614147	3.962558	3.400378
AKE	2.008796	3.000112	2.980765

It has been proved that the accuracy of the proposed model FRBIM model outperformed the KNN and AKE missing data imputation techniques taken for comparative analysis at different percentages (5%, 10% and 15%) of missing values. It has also been graphically displayed in the following Fig. 5 that the accuracy of the proposed FRBIM model is higher than the KNN and AKE missing data imputation techniques for varying percentages of missing values.

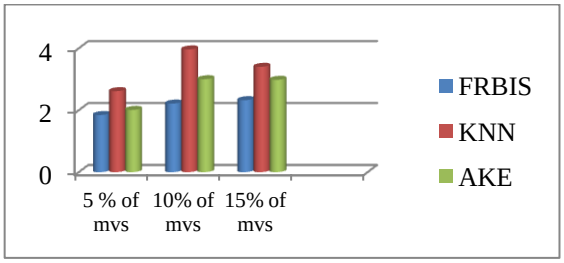

Fig.5 RMSE for the proposed FRBIM model, KNN and AKE imputation

VI. 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. Also the conventional models don't take into account the characteristics and unpredicTable- nature of IoT data. Ultimately, promising IoT missing data imputation model is crucial to avoid the perils and pitfalls of the existing imputation methodologies. The proposed FRBIM (Fuzzy Rule-Based Imputation Model) model accomplishes high accuracy rate than the conventional KNN and AKE imputation techniques. In the experimentation, it has been proved that the accuracy rate of the FRBIM model is much better than the existing KNN and AKE imputation model. Thus the proposed FRBIM model outperforms the KNN and AKE imputation methods apparently.

REFERENCES

1. 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 , doi: http://dx.doi.org/10.1155/2015/548605, 2015 , pp.1-8.
2. [Yir, 2013] Dong, Yiran, and Chao-Ying Joanne Peng,"Principled missing data methods for researchers" SpringerPlus, Vol.2, No. 1, 2013, pp.1-17.
3. S. A. Dhargalkar and A. U. Bapat, "Determining missing values in dimension incomplete databases using spatial-temporal correlation techniques", In Advanced Communication Control and Computing Technologies (ICACCCT), IEEE International Conference, doi={10.1109/ICACCCT.2014.7019157}, 2014, pp. 601-606.
4. P. Keerin and W. Kurutach and T. Boongoen, "Cluster-based KNN missing value imputation for DNA microarray data", In Systems, Man, and Cybernetics (SMC), IEEE International Conference, doi={10.1109/ICSMC.2012.6377764} , ISSN : 1062-922X, 2012, pp. 445-450.
5. Kiani, Kanwal, and Khalid Saleem, "K-Nearest Temperature Trends: A Method for Weather Temperature Data Imputation", In the Proceedings of 2017 ACM International Conference on Information System and Data Mining, doi={10.1145/3077584.3077592}, ISBN: 978-1-4503-4833-1, 2017, pp. 23-27.
6. W. C. Ku and G. R. Jagadeesh and A. Prakash and T. Srikanthan, "A clustering-ased approach for data-driven imputation of missing traffic data", IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS), doi={10.1109/FISTS.2016.7552320}, 2016, pp. 1-6.
7. Kumutha, V., and S. Palaniammal, "An enhanced approach on handling missing values using bagging k-NN imputation", IEEE International Conference on Computer Communication and Informatics (ICCCI), doi={10.1109/ICCCI.2013.6466301}, 2013, pp. 1-8.
8. T. Le, T. Altman and K. J. Gardiner, "Probability-based imputation method for fuzzy cluster analysis of gene expression microarray data", In ninth IEEE International Conference on Information Technology: New Generations (ITNG), doi={10.1109/ITNG.2012.159},2012, pp. 42-47.
9. Li, Dan, and Chongquan Zhong, "An Attribute Weighted Fuzzy c-Means Algorithm for Incomplete Datasets Based on Statistical Imputation", seventh IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), doi={10.1109/IHMSC.2015.128}, vol. 1, 2015, pp. 407-410.
10. Liu, Julie Yu-Chih, and Chiung-Hua Huang, "Handling missing data in extended possibility-based fuzzy relational databases", Third International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA), doi={10.1109/IBICA.2012.39}, 2012, pp. 57-62.
11. Razavi-Far, Roozbeh, and Mehrdad Saif, "Imputation of missing data using fuzzy neighborhood density-based clustering", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), doi={10.1109/FUZZ-IEEE.2016.7737913}, 2016, pp. 1834-1841.
12. Riouali, Youness, Laila Benhlima, and Slimane Bah, "A benchmark for spatial and temporal correlation based data prediction in wireless sensor networks", 10th International Conference on Intelligent Systems: Theories and Applications (SITA), doi={10.1109/SITA.2015.7358441}, 2015, pp. 1-7.

AUTHOR'S PROFILE

Dr. I.Priya Stella Mary - is an Assistant Professor at the Computer Applications Department in Bishop Heber College, Trichy. She has received her Ph.D degree from St. Joseph's College (Autonomous), Tiruchirappalli and Masters in Computer Applications from Bharathidasan University, Tiruchirappalli, India. She has two years of industrial experience. She has published many papers in the national and international journals. She has presented papers in the national conferences, seminars and international conferences. She has also attended various workshops on IoT analytics, Big Data analytics and Data mining and also served as a resource person in the short courses, seminars conducted. She has cleared NET examination. Her research interests are data mining and IoT.