Best Fit Missing Value Imputation (BFMVI) Algorithm for Incomplete Data in the Internet of Things

Benjamin Agbo, Yongrui Qin and Richard Hill
School of Computing and Engineering, University of Huddersfield, U.K.

Keywords: Missing Values, Imputation, Internet of Things (IoT), Best Fit Missing Value Imputation (BFMVI).

Abstract: The noticeable growth in the adoption of Internet of Things (IoT) technologies, has led to the generation of large amounts of data usually from sensor devices. When dealing with massive amounts of data, it is very common to observe databases with large amounts of missing values. This is a challenge for data miners because various methods for data analysis only work well on complete databases. A popular way to deal with this challenge is to fill-in (impute) missing values using adequate estimation techniques. Unfortunately, a good number of existing methods rely on all the observed values in the entire dataset to estimate missing values, which significantly causes unfavourable effects (low accuracy and high complexity) on imputed results. In this paper, we propose a novel imputation technique based on data clustering and a robust selection of adequate imputation equations for each missing datapoint. We evaluate our proposed method using six University of California Irvine (UCI) datasets, and relevant comparison with five recently proposed imputation methods. The results presented showed that the performance of the proposed imputation method is comparable with the Local Similarity Imputation (LSI) technique in terms of imputation accuracy, but is significantly less complex than all the existing methods identified.

1 INTRODUCTION

The Internet of Things (IoT) can be described as a network of multiple devices that can sense, process and share data generated from their surroundings (Singh et al., 2018). The adoption of IoT in various platforms has enabled easy communication and access to a wide range of devices such as sensors, actuators, home appliances, surveillance cameras, vehicles, etc. Therefore, there is a need to deploy more applications that will adapt to the potentially increasing amount and variety of data that will be generated by IoT devices (Agbo et al., 2019).

In order to ensure the usefulness of data generated by IoT devices in various data mining tasks, researchers have attempted to curb the popular challenge of incompleteness associated with sensor generated data. According to (Lata and Chakraverty, 2014), data is often incomplete due to a number of factors such as: human errors, erroneous measurements, communication malfunctions or faulty equipment e.g. sensors. Failure to account for missing data will significantly compromise the validity of findings from a dataset. In general, it could undermine the conclusions of a study by reducing the sample size which introduces bias (Read, 2015). Popular methods that have been used in research to handle the issue of missing data include: list-wise deletion, pair-wise deletion, hot decking, mean imputation and regression imputation. Despite the fact that these methods are straightforward to implement, they may lead to loss of information or introduce bias in the results obtained (Inman et al., 2015). In addition, most imputation methods consider the values of an entire dataset before estimating missing values. This could have unfavorable effects on the imputation process (e.g. high complexity or low accuracy).

One of the leading reasons for handling missing values is to improve the accuracy of clustering and classification tasks (Silva-Ramírez et al., 2015). However, most imputation methods are computationally intensive and therefore, take time to estimate and impute missing values. This may be inconsequential for training processes but it will not be practical to spend much time in estimating values for incomplete instances during clustering or classification tasks. This is most especially true for complex imputation techniques such as Multiple Imputation by Chained Equations (MICE), which rebuilds an imputation structure from every training instance and new instance (Tran

Agbo, B., Qin, Y. and Hill, R.
Best Fit Missing Value Imputation (BFMVI) Algorithm for Incomplete Data in the Internet of Things.
DOI: 10.5220/0009578201300137
In *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security (IoTBDS 2020)*, pages 130-137
ISBN: 978-989-758-426-8

et al., 2018). Although recent literature has shown significant increase in the accuracy of advanced imputation methods, the high costs associated with these methods in various tasks has often raised concerns. Therefore, it has become paramount to address the question of how the computation time of new methods could be reduced without sacrificing their accuracy (Tran et al., 2018).

In recent years, various machine learning (ML) algorithms have been introduced to handle the issue of data incompleteness which often occurs as a result of missing values (Angelov, 2017). These algorithms are designed to handle this issue by imputing the most plausible values in instances with missing values. In contrast to popular statistical methods for filling in missing values, machine learning algorithms use existing data in a dataset to train and develop a model that will be used to impute missing values. Various ML algorithms for imputing missing values have been identified in literature such as probabilistic methods, decision trees, rule based methods etc. (Farhangfar et al., 2008).

In this paper, we propose a novel imputation technique which utilizes the similarity between observed values to perform imputation. This is achieved by partitioning an incomplete dataset in the first instance. Then the similar records within cluster are used to estimate the missing values. However, some challenging issues have been identified with the proposed method including how to perform clustering on the incomplete dataset before imputation. To solve this problem, we initially assign distinctive values to replace all the missing values. This reduces the effect of missing values in the datasets and enhances clustering on the incomplete datasets.

We evaluate the performance of our proposed BFMVI technique against existing techniques namely- LSI, FIMUS, FCM, DMI and EMI, on six datasets obtained from University of California Irvine (UCI) machine learning repository.

2 RELATED WORKS

Many research efforts have been channelled towards addressing the issue of data incompleteness by attempting to develop more accurate and reliable imputation techniques. In this section, we will review various related research and recent efforts aimed at addressing this problem.

A framework for the imputation of missing values using co-appearance, correlation and similarity analysis (FIMUS) was proposed by (Rahman and Islam, 2014). The overal idea behind this method is to make educated guesses based on the correlation between attributes, co-appearance of values and the similarity between values that belong to an attribute. Unlike various existing technique, FIMUS can also be used to impute missing categorical variables. To compute co-appearances between values that belong to different attributes, FIMUS first of all summarizes the values of numerical attributes into various categories. For instance, the algorithm groups the values of an attribute A_p into $\sqrt{|A_p|}$ number of categories, where $|A_p|$ is the domain size of A_p. This strategy of grouping is advantageous due to its simplicity. However, it may not always detect natural groups due to the fact that it artificially makes the range of values for each category equal.

Various missing value imputation techniques have approached imputation using clustering schemes such as k-means and FCM. Another technique proposed by (Zhang et al., 2018) approaches imputation firstly by partitioning a dataset into k clusters. This will result in the formation of membership values for items within a particular cluster or cluster centroid. Then, all the missing values are evaluated using the membership degree of objects that fall within the same cluster centroid. The simplicity of this method constitutes a major advantage. However, the accuracy of the FCM imputation may be significantly affected by clustering results in usual situations when the selection of a suitable number of k clusters is challenging for data miners.

The Expectation maximization imputation (EMI), proposed by (Schneider, 2001; Dempster et al., 1977) is one of the most popular missing value imputation techniques identified in literature. To impute missing numerical values, this technique estimates the mean and covariance matrix from observed values in a dataset and iterates until no considerable change is noticed in the values of the imputed data, mean and covariance matrix, from one iteration to another. According to research, the EMI algorithm only works best in datasets with values that are missing at random. The main disadvantage of this method however, is that it relies on the information from other values in the dataset. Therefore, this method is only suitable for datasets with high correlation among attributes (Deb and Liew, 2016).

Another technique used to handle the issue of missing data is the Decision tree based missing value imputation (DMI) algorithm proposed by (Rahman and Islam, 2013). This technique incorporates the decision tree and the EMI algorithm for imputing missing values. The authors argue that attributes within the horizontal partition of a dataset can have higher correlation than the correlation of attributes over the entire dataset. The processes involved in DMI are described below: firstly, it divides the complete dataset (D_{full}) into two smaller datasets with one having incomplete data i.e. with missing values (D_{miss}) and the other, ($D_{complete}$) having complete records. Next, it builds up decision trees based on ($D_{complete}$), taking the attributes with incomplete values in (D_{miss}) as the class attributes. After that step, it further assigns every record having missing values in (D_{miss}) to the leaf it falls on the tree, which takes the attribute having the missing value as the class attribute. Finally, the DMI algorithm employs the EMI algorithm to fill-in missing numerical values and majority of the class values within each leave to impute missing categorical values.

Another method used to handle the issue of missing data is the Local Similarity Imputation (LSI) technique proposed by (Zhao et al., 2018). Here, missing values are estimated using top *k*-nearest neighbours and fast clustering. Firstly, a dataset with missing values is partitioned into clusters, then the most similar records from each cluster are used to estimate the missing values. To enhance the accuracy of clustering, this technique uses a two-layer deep learning algorithm to detect important features within a cluster. Therefore, this will enable the fast clustering algorithm to effectively read important records from a dataset one time. Lastly, the top *k*-nearest neighbour algorithm is used to evaluate and impute missing values in individual clusters.

Though these methods show good performance in terms of their imputation accuracy, their huge computation time will reduce their efficiency when dealing with increasing volumes of data.

3 ROBUST BFMVI FOR INCOMPLETE DATA

The structure of our method is represented under two stages: firstly, the incomplete dataset is partitioned into different groups and at the second stage, missing values within each partition is imputed using the BFMVI algorithm.

3.1 Arbitrary Clustering

To partition our datasets, we first of all fill in all missing values with distinctive values. To enable fast execution of our algorithm, we stored the sample of our dataset with pre-imputed records in a array. An arbitrary number (γ) of items were taken from the dataset to form different groups, containing similar records. According to (Zhang et al., 2015), better imputation results could be achieved when similar samples are used to evaluate missing values. However, (Zhao et al., 2018) argued that existing clustering algorithms perform minimally in incomplete datasets due to the fact that missing values pose serious uncertainties and affect the accuracy and usability of existing clustering algorithms. Although, more prospects still remain for the improvement of our clustering approach, the strength of our contribution however, lies in our imputation method.

Algorithm 1: Clustering Algorithm.

```
Input: Dataset with missing values, $D \in X^{n*m}$. Parameter $\gamma, \beta$.
Output: Dataset Clusters and their number $k_i$.
D ← PreImp (dv, D); //initially fill missing values with distinctive value
Arr ← GetValuesIn (D); //get preimputed values of D and store in array
for $i$ = 1 to $l$ do
    [Cluster, γ] ← Partition (Arr [γ], Clusters.β ;// Partition arbitrary values of Arr [γ] into β groups.
end for
Return Clusters and their number $k_i$.
```

3.2 BFMVI based on Arbitrary Clustering

As stated earlier, the first phase of our technique involves partitioning our datasets into groups of items with similar records, then the missing values are estimated using the observed values of records present in each cluster. The strength of our contribution lies in the ability of our model to choose the most suitable imputation method for each missing datapoint. Lets assume [k_1, k_2,…, k_n] to be k clusters generated from the pre-imputed dataset D and [x_1, x_2,…, x_n] is a non-nominal distribution with missing values. In the imputation process, the algorithm develops six imputation results as seen in equation 1-6 and selects a suitable imputation equation for each missing datapoint based on a defined criteria.

Imputation 1: The average value of observed records in each cluster are used to fill in each missing datapoint. Our parameters γ and λ are set to 3 and 0.4 respectively.

$$p_i = \sum_{i=1}^{n_k} \frac{x_i}{n_k} \tag{1}$$

Imputation 2: For each partition with missing values x_{i_k}, missing values are imputed as follows:

$$d_i = \frac{1}{n_k} \sum_{i=1}^{n_k} r_{i_k} \tag{2}$$

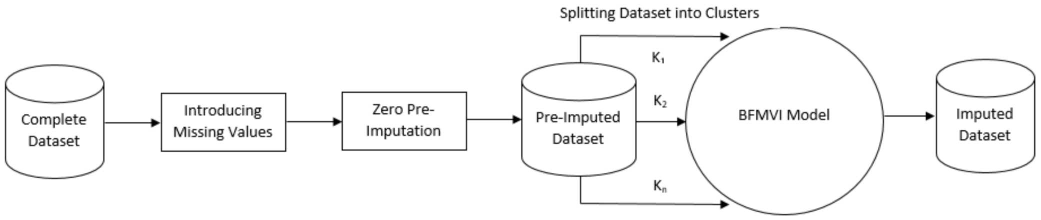

Figure 1: Framework for Imputation Model.

where r_{i_k} is the corresponding mode value of x_{i_k} and n_k represents the distribution size.

Imputation 3: The *log* of d_i is computed and the parameter γ which is set to 3, is multiplied by the resulting value.

$$R = \log d_i(\gamma) \tag{3}$$

Imputation 4: For each missing value within clusters, imputed values are also evaluated by:

$$I = \log p_i(\gamma) \tag{4}$$

Imputation 5: The sum of I and R is computed and their resulting average is used to fill in the missing values within each group.

$$N_{k_i} = \frac{\log d_i(\gamma) + \log p_i(\gamma)}{2} \tag{5}$$

Imputation 6: Finally, our parameter λ is added to the resulting value of N_{k_i} and Missing values within each group is imputed using O_{k_i}.

$$O_{k_i} = (\frac{\log d_i(\gamma) + \log p_i(\gamma)}{2}) + \lambda \tag{6}$$

After computing all the values for the missing instances using equation (1-6), the error between each previous imputation (r_{pre}) and the six imputations (α_{curr}) are estimated using the following equation:

$$err = r_{pre,i} - \alpha_{curr,i} \tag{7}$$

For each missing data point, the value of $r_{pre,i}$ is compared with all the values estimated from equations (1-6). The difference between each $\alpha_{curr,i}$ and the previous imputation $r_{pre,i}$ is computed and the value with the lowest error shows a higher similarity with $r_{pre,i}$ and is used to impute the value for a particular missing data point within a cluster.

Considering further improvement and more applications in dynamic environments, our proposed method would have the potential to handle dynamic changes in a dataset as it selects the most appropriate value for each missing data point.

Algorithm 2: Best Fit Missing Value Imputation.

```
Input: Dataset with missing values, $D \in X^{n*m}$ . Parameters α.
Output: Dataset with Imputed values $P$.
while 1 do
    [Clusters,$k_i$]← Clustering alg (D); // partitioning the incomplete dataset using Algorithm 1.
    for $i$ = 1 to $k_i$ do
        [InData, p] ← GetFromData(Clusters.k) ;// get subsets with incomplete records p;
        for $j$ = 1 to $p$ do
            r1 = mean (InData[j], (Clusters.k);
            r2 = (mode/length)(InData[j],
            r3 = (log(r2)*3))(InData[j], Clusters.k)
            r4 = (log(r1)*3)(InData[j], Clusters.k)
            r5 = r3 + r4/2
            r6 = r5 + 0.4
        end for
        Get set of imputation results $r_{curr}$ of Clusters.k
        $\alpha_{curr}$ ← GetSet ($r_{curr}$) ;// get current set of imputation results
    end for
    Calculate err between previous and current imputations via (5)
    Let $\alpha_\gamma$ = err($\alpha_{curr}$)
    for each $\alpha_{curr}$ do
        if $r_\gamma$ = min $\alpha_\gamma$ then
            $P \leftarrow OutputDataset(D, r_{curr})$ ;// $r_{curr}$ with lowest error is used for imputation
            Stop
        end if
    end for
end while
Return complete dataset $P$;
```

4 EXPERIMENTS AND ANALYSIS

4.1 Experimental Design

To assess the plausibility of our technique against other existing techniques, namely LSI, FIMUS, FCM, DMI and EMI, we used six UCI machine learning datasets with no missing values as ground truth. Then, the missing values were artificially imposed on the

Table 1: Description of Six UCI Datasets.

Dataset	Records	Attributes	Classes
Iris	150	4	3
Pima	768	8	2
Wine	178	13	3
Yeast	1484	9	10
Housing	506	14	Null
Adult	48842	14	Null

Table 2: d_2 and average execution time (sec) of the six imputation techniques on the six UCI datasets (at 3, 6, 9, 12 and 15% missing data).

	Datasets					
Imputation Methods	Iris d$_2$ (*t*)	Pima d$_2$ (*t*)	Wine d$_2$ (*t*)	Yeast d$_2$ (*t*)	Housing d$_2$ (*t*)	Adult d$_2$ (*t*)
BFMVI	0.977 **(0.031)**	0.907 **(0.145)**	**0.959 (0.043)**	0.946 **(0.258)**	0.967 **(0.105)**	0.9657 **(6.89)**
LSI	**0.983** (0.358)	**0.914** (2.439)	0.952 (0.331)	**0.948** (15.557)	**0.983** (1.654)	**0.971** (35.65)
FIMUS	0.966 (1.154)	0.90 (313.248)	0.938 (1.393)	0.854 (1412.75)	0.940 (7.257)	0.954 (1923.35)
FCM	0.964 (0.256)	0.882 (0.874)	0.788 (0.242)	0.929 (13.974)	0.916 (0.301)	0.751 (23.75)
DMI	0.954 (2.683)	0.860 (412.386)	0.864 (12.363)	0.936 (73.146)	0.912 (84.552)	0.881 (103.04)
EMI	0.957 (0.173)	0.848 (1.674)	0.868 (0.549)	0.911 (5.417)	0.905 (2.785)	0.713 (19.78)

datasets in order to test the accuracy of the six imputation techniques. Since the original values of the datasets are known, we can easily evaluate the accuracy of the missing data imputation techniques by observing how close the imputed values are to the original (Zhao et al., 2018). Each of the UCI datasets were then regenerated into five unique data sets with different percentages of missing values: 3%, 6%, 9%, 12% and 15% respectively on each dataset.

The six imputation methods are then used to fill in the different percentages of missing values in each dataset. For the purpose of our simulation, we used a dimensionality reduction technique called Principal Component Analysis (PCA) to reduce interrelated components, thereby retaining the variation of values present in each dataset. This led to the generation of new sets of uncorrelated records called principal components, which were used to simulate the different percentages of missing data. The criteria that are used to quantify the performance of the imputation methods are RMSE and d_2. We further computed the execution time for each technique to evaluate their performance in resource constraint scenarios. From equation 8 and 9, N represents the number of values missing. P_i and O_i are the respective imputed and actual values of the ith missing values, and $\bar{O}$ represents the average of the actual values. The RMSE value can range from 0 to ∞, with a lower value indicating better imputing performance. The value of d_2 can range from 0 to 1 with a higher value indicating better resemblance (Zhao et al., 2018).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{8}$$

$$d_2 = 1 - [\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2}] \tag{9}$$

4.2 Results and Analysis

Figure 2-7 presents the accuracy of our BFMVI technique against LSI, FIMUS, FCM, DMI and EMI techniques on iris, wine, boston housing, yeast, pima and adult datasets in terms of their RMSE for 5 missing data ratios.

Table 2 further shows the index of agreement (d_2) and execution time (in seconds) for the six imputation techniques on the six UCI datasets.

From the results, it can be observed that the proposed method shows a low error rate and good imputation accuracy but does not completely outperform the LSI technique. However, it shows the best performance in terms of execution time compared to the five other methods. Although, the popular EMI technique considers the entire instances in a dataset before performing imputation, it still has the lowest accuracy among all the six methods that were tested.

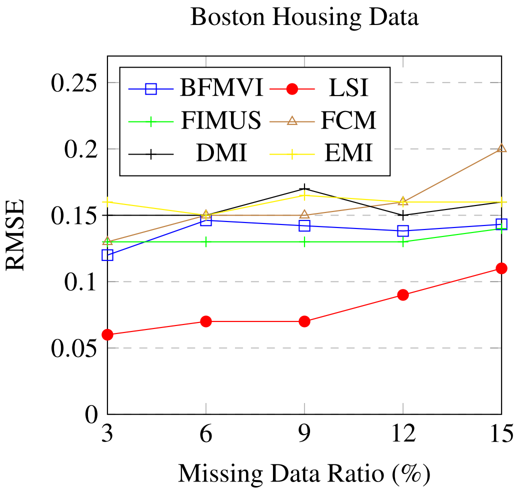

Figure 2: RMSE of imputation methods on Housing Data.

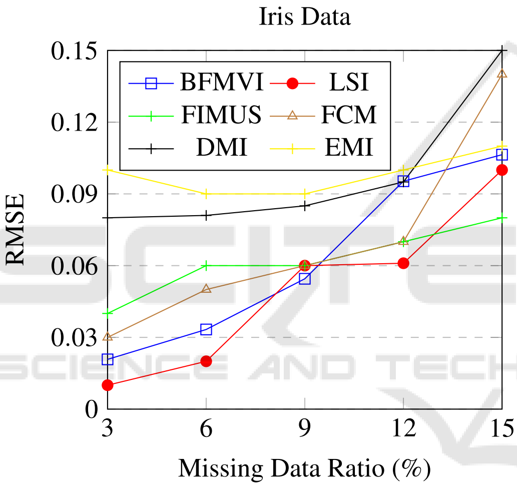

Figure 3: RMSE of imputation methods on Iris Data.

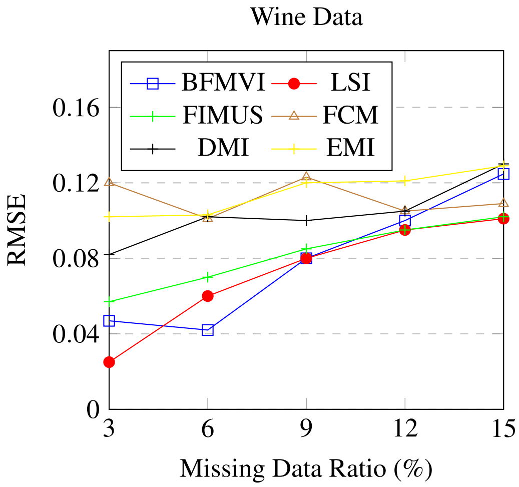

Figure 4: RMSE of imputation methods on Wine Data.

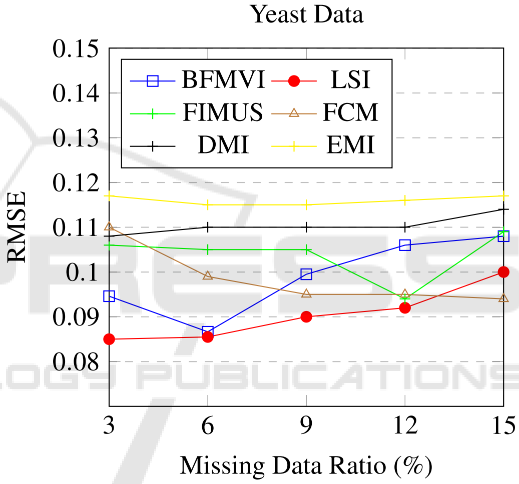

Figure 5: RMSE of imputation methods on Yeast Data.

In contrast, the LSI technique shows the best performance in terms of imputation accuracy but fails to completely outperform FCM, EMI and our method in terms of execution time.

FIMUS is another hybrid method that considers every record in a dataset before imputation. The accuracy of this method is better than FCM DMI and EMI in all five datasets and sometimes outperforms our method when a higher percentage of missingness is observed in a dataset. However, the execution time of this method is poor compared to LSI, FCM, EMI and our method. From our observation, the performance of the execution time reduced significantly when more records were observed (e.g. in the pima and yeast datasets). The DMI and FCM techniques partition the datasets into small groups with similar records which could have a positive effect on the imputation of missing values when closely related records are used to estimate missing records. However, DMI and FCM completely rely on the accuracy of clustering or classification and therefore perform minimally due to clustering or classification inaccuracy.

Overall, the accuracy of our proposed method ran close to the LSI method on five out of six datasets but showed a clear distinction from the LSI method on the boston housing dataset. This was largely influenced by the accuracy of the fast clustering algorithm using a two-layer deep learning algorithm in the LSI method. We will attempt to address these limitations by improving the similarity between records used to estimate these missing values in our proposed method.

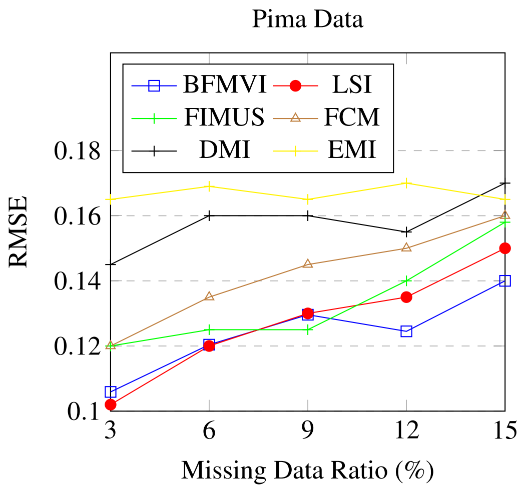

Figure 6: RMSE of imputation methods on Pima Data.

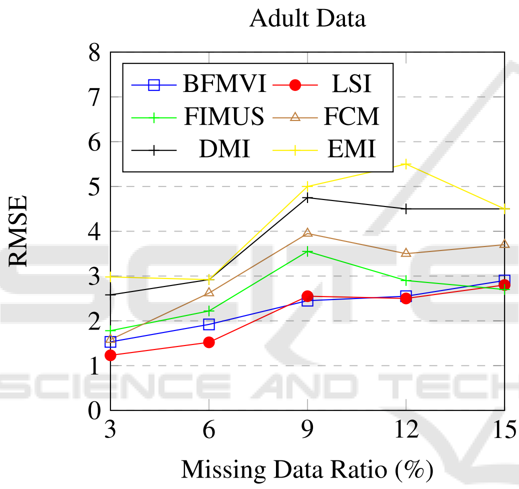

Figure 7: RMSE of imputation methods on Adult Data.

5 CONCLUSION AND FUTURE WORKS

Inferences drawn from various data mining tasks (such as prediction, clustering, classification etc.) can significantly be affected by the presence of missing data. Therefore, to ensure the validity of information drawn from these tasks, the imputation of missing values using adequate techniques is paramount. In this paper, we present a BFMVI technique for handling incomplete static databases. The proposed method first of all fills in missing data points with distinctive values and partitions the pre-imputed dataset using arbitrary values. Secondly, based on the similarity between values in each cluster, missing values are estimated using the BFMVI algorithm. From the experiments, it is observed that our proposed method is less complex that other identified methods and shows considerable performance in terms of imputation accuracy, which makes it a good fit for resource constraint scenarios.

Considering the characteristics of IoT data and its contribution to the big data era, more work still needs to be done with regards to developing robust and less complex algorithms for handling missing values observed in streams of continuously generated data. Our future research will be based on the improvement of the proposed imputation method and its adoption in more dynamic scenarios.

REFERENCES

Agbo, B., Qin, Y., and Hill, R. (2019). Research directions on big iot data processing using distributed ledger technology: A position paper. In *IoTBDS*.

Angelov, B. (2017). Working with missing data in machine learning.

Deb, R. and Liew, A. W.-C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. *Information sciences*, 339:274–289.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Farhangfar, A., Kurgan, L., and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12):3692–3705.

Inman, D., Elmore, R., and Bush, B. (2015). A case study to examine the imputation of missing data to improve clustering analysis of building electrical demand. *Building Services Engineering Research and Technology*, 36(5):628–637.

Lata, K. and Chakraverty, S. (2014). Handling data incompleteness using rough sets on multiple decision systems. In *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pages 1–6. IEEE.

Rahman, M. G. and Islam, M. Z. (2013). Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowledge-Based Systems*, 53:51–65.

Rahman, M. G. and Islam, M. Z. (2014). Fimus: A framework for imputing missing values using co-appearance, correlation and similarity analysis. *Knowledge-Based Systems*, 56:311–327.

Read, S. H. (2015). Applying missing data methods to routine data using the example of a population-based register of patients with diabetes.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of climate*, 14(5):853–871.

Silva-Ramírez, E.-L., Pino-Mejías, R., and López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, 29:65–74.

Singh, M., Singh, A., and Kim, S. (2018). Blockchain: A game changer for securing iot data. In *Internet of Things (WF-IoT), 2018 IEEE 4th World Forum on*, pages 51–55. IEEE.

Tran, C. T., Zhang, M., Andreae, P., Xue, B., and Bui, L. T. (2018). Improving performance of classification on incomplete data using feature selection and clustering. *Knowledge-Based Systems*, 154:1–16.

Zhang, L., Pan, H., Wang, B., Zhang, L., and Fu, Z. (2018). Interval fuzzy c-means approach for incomplete data clustering based on neural networks. *Journal of Internet Technology*, 19(4):1089–1098.

Zhang, Q., Yang, L. T., Chen, Z., and Xia, F. (2015). A high-order possibilistic *c*-means algorithm for clustering incomplete multimedia data. *IEEE Systems Journal*, 11(4):2160–2169.

Zhao, L., Chen, Z., Yang, Z., Hu, Y., and Obaidat, M. S. (2018). Local similarity imputation based on fast clustering for incomplete data in cyber-physical systems. *IEEE Systems Journal*, 12(2):1610–1620.