Clustering-Based Hybrid Approach for Multivariate Missing Data Imputation

Aditya Dubey[1]
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India

Akhtar Rasool[2]
Department of Computer Science and Engineering
Maulana Azad National Institute of Technology
Bhopal, India

***Abstract*—In the era of big data, a significant amount of data is produced in many applications areas. However due to various reasons including sensor failures, communication failures, environmental disruptions, and human errors, missing values are found frequently These missing data in the observed data make a challenge for other data mining approaches, requiring the missed data to be handled at the preprocessing stage of data mining. Several approaches for handling the missing data have been proposed in the past. These approaches consider the whole dataset for making a prediction, making the whole imputation approach to be cumbersome. This paper proposes the procedure which makes use of the local similarity structure of the dataset for making an Imputation. The K-means clustering technique along with the weighted KNN makes efficient imputation of the missed value. The results are compared against imputations by mean substitution and Fuzzy C Means (FCM). The proposed imputation technique shows that it performs better than other imputation procedures.**

***Keywords*—Clustering; imputation; KNN; missing at random; multivariate**

I. Introduction

Since the age of big data began, the collection of data from various sources, and the resultant amount of data has risen to the greatest extent [1]. Multivariate datasets are prevalent in several real-world applications, such as electrical system analysis, meteorological or economical strategy planning, security control, and plenty more. In several application areas, multiple sensors are deployed to produce datasets, and they typically have one target to generate the data as activity occurs. For example, in a power grid application several sensors diagnosing the state of power transformers, produce the data by monitoring the state of gases over time [2]. In the era of IoT, a vast number of sensors are utilized for generating the multivariate environmental conditions, for example, the air or water pollution [3]. In biomedical, numerous devices can also be fitted in working areas to track the health and overall well-being of senior citizens, which also ensures that adequate medicine is suggested. Important information and the facts can be obtained deriving these datasets [4]. Preprocessing is one of the steps for analyzing the data. One major issue handled in the preprocessing step is missed value. Unfortunately, the raw dataset generated by the sensor network typically includes missing values due to the rough working conditions or uncontrolled variables such as adverse weather conditions, malfunctions of the infrastructure, or unstable signals. The problem of missing data is quite prevalent in many applications. The incomplete dataset is inadvertently and uncontrolled by the researcher. The outcome is that the data observed cannot be evaluated due to the incompleteness of the datasets.

Several studies have suggested strategies for handling incomplete values in the dataset. The methods for dealing with the missed values may be categorized into three types. The first approach is to disregard the complete record which contains any missing value. Additionally, replace the missed value with zero or mean of the attribute [5]. The major downside of these strategies is that they decrease the efficiency of estimation. By excluding any usable data in those cases having any missing values. This could degrade the expected result. The second is to determine the values that use combinations of the Expectation-Maximization method. The third approach is imputation, which requires the process of completing the incomplete values in the dataset by some potential values, depending on the details in the dataset.

Missing data imputation strategies can be categorized into two types depending on the method of approximating the missing values. The first type, mathematical or predictive methods are used to estimate missing values. These approaches are remarkably simple, replacing each missed value with the mean or mode value of the variable, as well as more complex methods focused on advanced statistical techniques. The second is the imputation based on machine learning that utilizes the dataset knowledge to model the calculation of missing data. It involves a number of methods K nearest neighbor [6], MLP imputation [7], auto-associative neural network imputation [8], SOM imputation [9], recurrent neural network [10] and multi-task networks [12].

The paper is summarized as: Missing value issue is formulated and presented in Section 2. Section 3 describes the literature of the research area by focusing on major imputation procedures. The proposed technique is subsequently implemented in Section 4. In Section 5, the proposed technique is used on the benchmark datasets. The last section includes the conclusion and future work. The analytical results demonstrate that the proposed method works better than other conventional imputation techniques. This research paper has the following contributions:

- The proposed technique considers the local data similarities and introduces a local imputation model that uses a clustering technique for estimating the missing values. In other terms, the complete mechanism includes clustering is performed first and then imputation is done.

	Att_1	Att_2	Att_3	Att_4	Att_5
x_1	?	100	2.2	1	2.7
x_2	1	30	?	0	2.9
x_3	?	90	1.8	1	3.0
x_4	2	30	1.3	0	3.5
x_5	?	20	0.3	0	2.8

Fig. 1. A Dataset Containing Missing Values

- A top KNN distance weighted imputation enhancing the prediction accuracy is utilized to predict the missing value in each cluster.
- The proposed method is implemented and verified using datasets of the UCI dataset repository [11]. For big data analysis, the proposed method is implemented and verified on the MATLAB platform.

II. Problem Formulation

This paper is aimed to focus on making accurate predictions of missing data. For illustrating this problem more specifically, figure 1 depicts a data set having many missing values in some of the attributes $Att_1, Att_2, ..., Att_5$. Let $X = \{x_1, x_2, x_3, ..., x_N\}$ denotes the data collected from N different data sources and j^{th} attribute from i^{th} source can be denoted as X_{ij} for i=1,2,3,...,N, j=1,2,3,...,M. In this paper, the missing value is denoted as '?'. Additionally, to represent whether the value of X is missing or not, an indicator matrix H is used.

$$H_{ij} = \begin{cases} 1, \text{ if value of } X_{ij} \text{ is present} \\ o, \text{ otherwise} \end{cases} \quad (1)$$

III. Literature Review

This section provides a short overview of research on the basic algorithms used for imputation.

Abdella et al.'s research were focused on the use of neural network combined with the genetic algorithms to make imputation of the missing values in the dataset [13]. For training the neural networks Multi-Layer Perceptron and Radial Basis Functions are utilized. Li et al. employed a soft computation-based clustering technique to efficiently handle inaccurate and unconstrained dataset in addition to this fuzzy clustering algorithm handles missed data [14]. Liao et al. provided a fuzzy k-means clustering technique using a sliding window to impute missing data so that the quality of the dataset can be improved [15]. Pelckmans et al. suggested a method not to rebuild missing data but by utilizing support vector machines, the effects of the incomplete dataset over the result and anticipated cost are simulated [16]. The procedure consists of assuming certain models for the covariates of missing data and using the maximum likelihood method to get the predictions for these models. The benefits of this technique are that even if missing data is present among the input variables, classification rules can be derived from the observed values. While the result is that the proposed technique is designed for better classification accuracy rather than improving the imputation accuracy for missing data. Lim et al. suggested a hybrid neural network technique that utilizes the ARTMAP and fuzzy c-means clustering for classifying the patterns utilizing the incomplete training and testing dataset [17]. Fuzzy ARTMAP has the drawback of higher susceptibility to arrange the training data. For fuzzy ARTMAP it is very crucial to select the vigilance parameters because it can be hard to determine the optimal value of the vigilance parameters.

Hathaway et al. proposed a clustering method relying on the dissimilarity of missing data [18]. A benefit of this technique is that for missing data, fuzzy c-means is considered to be a reliable clustering method. The support vector regression technique used by Feng et al. was also used to predict the missing data of DNA microarray gene expression using an orthogonal coding scheme [19]. Comparison research to the earlier established techniques for their imputation, such as KNN and BPCA, showed that the SVR procedure was efficient for imputation. One major benefit of the SVR model is that it takes less time for computation, but the hybrid SVR clustering process produces more sensible outcomes for the dataset having outlier data. Timm et al. observed that the dataset having missing values is an important issue in the data analysis [20]. The class-specific possibility of missing values was implemented to properly disseminate the incomplete data points to the clusters. Farhangfar et al. proposed an extensive analysis of representative imputation methods [21]. They stated that the usage of a low-quality single imputation method resulted in prediction accuracy comparable to the accuracy of utilizing some other advanced imputation method.

Li et al. presumed that missed data are defined in terms of intervals and introduced a novel fuzzy c-means based procedure for handling missing data based on nearest neighbor intervals concept [22]. The drawback of this technique is that the number of clusters is not selected on a theoretical basis, so further procedure is required to examine this issue. Nuovo performed a comparative study on imputation done using fuzzy c-means against the imputation done using case deletion [23]. The methods are compared using a mentally retarded patient dataset in psychological research environmental conditions. The research shows that imputation methods especially the FCM based method provide efficient imputation of data and prevent the deletion of missing data which causes a reduction in the strength of research. The fuzzy c-means imputation makes more improved predictions than that of the regression imputation technique and expectation-maximization technique. One main drawback includes that FCM imputation utilizes a weighting parameter, which values to 2, which requires to be adjusted according to the dataset type. L. Zhao et al. make use of the local similarity imputation procedure for imputing the missing value in the cyber-physical system. The procedure includes the usage of a two-layered stack encoder together with the concept of KNN [24]. Q. Ma et. al in his research addresses the issue of the insufficient complete data subset to make an imputation from its clustered neighbors [25]. Ordered sensitive imputation for clustered missing data makes use of previously imputed value to be used for the next iteration of imputation. Two regularized learning algorithms have been utilized in the method proposed by A. Wang et. al. for imputing the missing value in microarray experiments on gene expression [26]. The RLLSimpute L2 regularized local least square is trained on the target gene and its neighbors so that the missing value can be

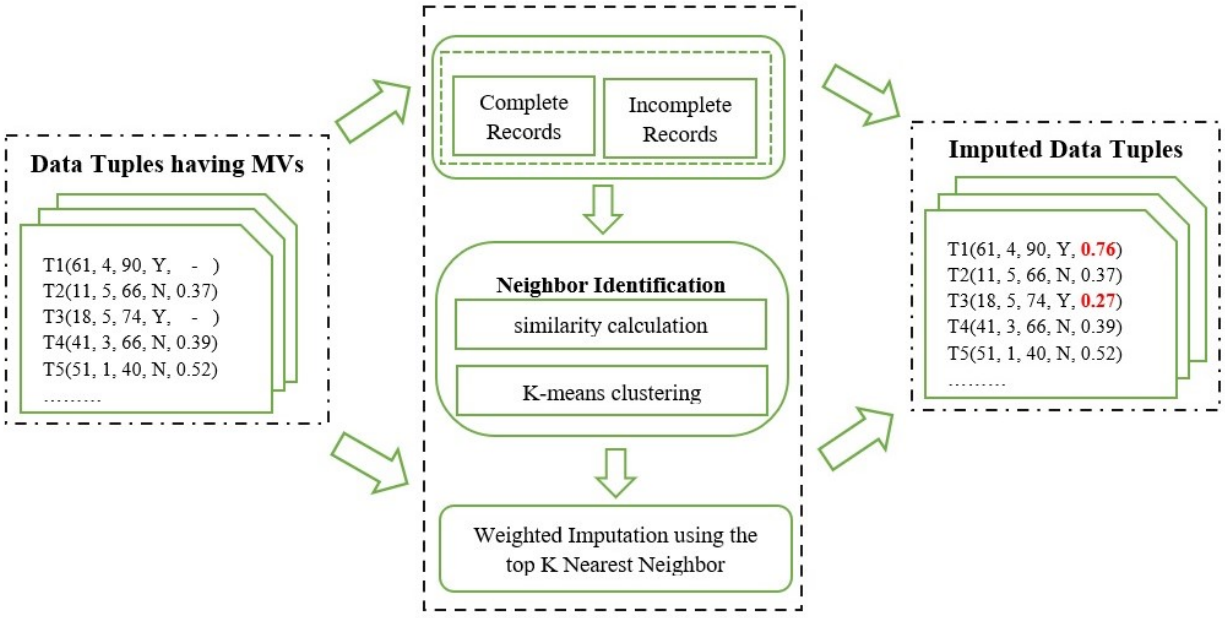


Fig. 2. Block Diagram of Proposed Method

imputed.

A. Types of Missing Data

Missing data have three kinds of missing patterns.

1) Missing completely at random (MCAR)- The missed data does not have any dependability on any other data. In other words, the probability of missing data is equal to all the units [27], [29].
2) Missing at random (MAR) - The missed data depends upon the available data. These available data can be utilized for estimating the missed data.
3) Missing Not at Random (MNAR)- The analysis of missed data relies upon other missing data which causes the missed data to become unpredictable.

IV. Proposed Work

A basic issue in missing data imputation is the retrieval of lost value by utilizing the available dataset information. Clustering is one of the most common data mining methods which arise to fix this problem [28]. The general goal of this clustering is to split available data into multiple desired clusters by recognizing the similarity of objects. The principle is to minimize the intra-cluster similarity and maximize the inter-cluster dissimilarity. Figure 2 describes the imputation procedure, in which the first step is to partition the dataset into complete and incomplete records according to the missing data.

K-means approach is used as a clustering approach consisting of a four-step procedure. The first step is the random selection of a fixed number of cluster centroids. The second step is the assignment of each record to a certain cluster having the closest centroid. The third step is the recalculation of the cluster centroid. The last step states to iteratively repeat the procedure from step second if the algorithm does not reach the termination condition.

The last step in the imputation is to use the cluster information and provide value for each non-reference attribute having an incomplete object. Objects falling in the same cluster are considered as the nearest neighbor of missing values having higher similarity and based on the nearest neighbors missing values are imputed.

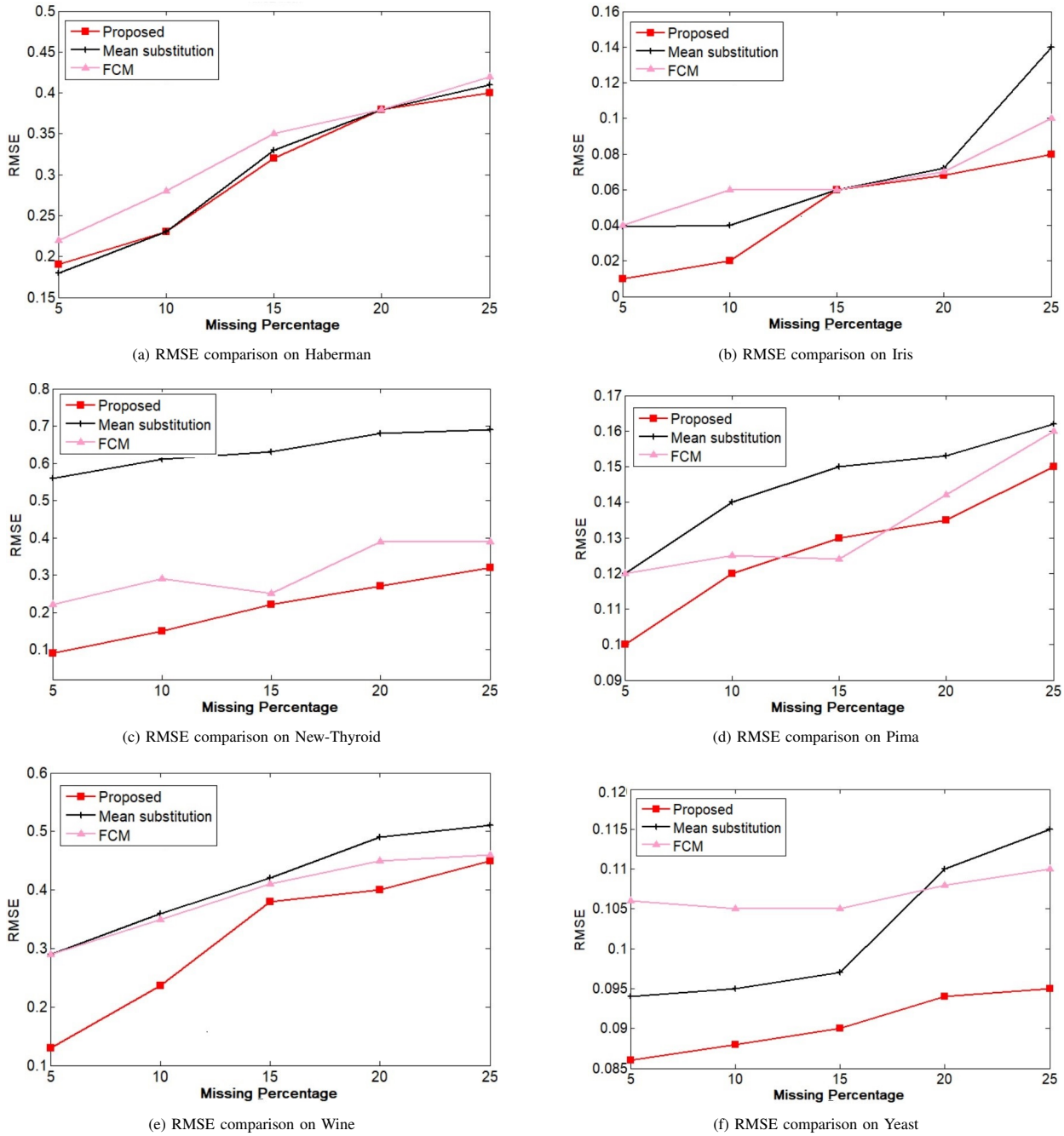
TABLE I. Summary of the Datasets

Name of dataset	Number of Instances	Number of Attributes
Haberman	306	3
Iris	150	4
New-Thyroid	215	5
Pima	768	9
Wine	178	13
Yeast	1484	8

V. Experiments and Analysis

A. Experimental Design

For illustrating the efficacy of the proposed method, extensive experiments are conducted on six UCI datasets [11]. To decide how the technique generalizes, experiments with more datasets, dealing with various numbers, and different types of missing value patterns are required. For performing experiments, some percentage of the data is deleted so that they have 5%, 10%, 15%, 20%, 25% of missing ratio in the dataset. Table I demonstrates the datasets used in this paper.



(a) RMSE comparison on Haberman

(b) RMSE comparison on Iris

(c) RMSE comparison on New-Thyroid

(d) RMSE comparison on Pima

(e) RMSE comparison on Wine

(f) RMSE comparison on Yeast

Fig. 3. RMSE comparison of Mean substitution, FCM and Proposed technique on (a) Haberman (b) Iris (c) New-Thyroid, (d) Pima (e) Wine (f) Yeast

B. Performance Evaluation

Several efficiency procedures have been used for estimating the predictive performance and comparison of distinct models [30]. In every definition that comes ahead, D_i is the actual value, P_i is the predicted value, and error $E_i = D_i - P_i$. Root Mean Square Error (RMSE) represents a root mean square deviation of the predicted values. Since the negative and positive signed errors that have been recorded do not compensate each other, RMSE provides an overview of the error during prediction. RMSE points out that the overall prediction error is greatly influenced by big errors that are considerably more costly than tiny errors. RMSE is also subject to scale shift and data transformations happened at the pre-processing. The less the RMSE, the stronger will be the prognosis. For n number of missing values, RMSE is defined as-

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} Err_i^2} \tag{2}$$

Fig. 3(a)-3(f) represents a comparison of experimental results that imply that the proposed technique is advantageous

in terms of imputed error. Ten different sets of the same missing percentage are created, all three imputation procedures are applied on these sets, thereafter the average of that ten RMSE is calculated. The lower the RMSE, the better will be the imputation. As the missing percentage increases, it becomes a challenge for each technique to make accurate imputation resulting in the increment of RMSE. For the small percentage of missingness, the techniques may have a relatively small difference in their performance but as the missing percentage increase, there exists a clear difference in the performance of each technique. For comparing the performance of the proposed technique, mean substitution and FCM-based imputation are utilized. Experimental results demonstrate that the proposed technique performs much better as compared to two other imputation technique.

VI. Conclusion

In this paper, a local similarity-based hybrid imputation approach utilizing the K-means clustering has been implemented which increases prediction accuracy. Experimental results on natural datasets show that the proposed method dominates over other existing prediction approaches. The analysis of accurate parameters and operational framework for the efficient implementation of the proposed technique has to be done for maximal technique execution. The kind of missing pattern that is MCAR, MAR or NMAR also plays an important role in the imputation procedure. In future, different missing pattern are experimented with the proposed imputation procedure. In addition, more experiments are also required with the larger number of attributes and records. At last, it can be concluded that the proposed technique makes an accurate prediction with higher efficiency.

References

[1] L. A. Kurgan, K. J. Cios, M. Sontag and F. J. Accurso, “Mining the cystic fibrosis data”, Next Generation of Data-Mining Applications, 2005, pp. 415–444.

[2] J. Barnard and X. L. Meng, “Applications of multiple imputation in medical studies: From aids to nhanes”, Stat. Methods Med. Res. vol. 8, no. 1, 1999, pp. 17–36.

[3] K. J. Cios and G. Moore, “Uniqueness of medical data mining”, Artif. Intell. Med., vol. 26, no. 1/2, 2002, pp. 1–24.

[4] A. Dubey and A. Rasool, ”Data Mining based Handling Missing Data”, Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 483-489.

[5] A. Dubey and A. Rasool, ”Time Series Missing Value Prediction: Algorithms and Applications”, Information, Communication and Computing Technology (ICICCT), vol. 1170, 2020, pp. 21-36.

[6] G. Batista and M. C. Monard, “Experimental comparison of k-nearest neighbour and mean or mode imputation methods with the internal strategies used by c4.5 and cn2 to treat missing data”, University of Sao Paulo, 2003, pp. 1–97.

[7] P. Sharpe and R. Solly, “Dealing with missing values in neural network-based diagnostic systems”, Neural Computation Applications, vol. 3, 1995, pp. 73–77.

[8] D. Pyle, “Data preparation for data mining”, morgan kaufmann publishers inc. San Francisco, vol. 22, no. 2, 1999, pp. 115–170.

[9] T. Kohonen, “Self-organizing maps”, Springer notes 3rd edn., 2001, pp. 347–371.

[10] Y. Bengio and F. Gingras, “Recurrent neural networks for missing or asynchronous data”, Adv. Neural Inf. Process Syst., vol. 8, 1995, pp. 395–401.

[11] https://archive.ics.uci.edu/ml/datasets.

[12] T. Kohonen, “Self-organizing maps”, Springer notes 3rd edn., 2001, pp. 347–371.

[13] M. Abdella and T. Marwala, “The use of genetic algorithms and neural networks to approximate missing data in database”, Comput. Inform. vol. 24, 2005, pp. 577–589.

[14] D. Li, J. Deogun, W. Spaulding and B. Shuart, “Towards missing data imputation: a study of fuzzy k-means clustering method”, Rough Sets Curr. Trends Comput., 2004, pp. 573–579.

[15] Z. Liao, X. Lu, T. Yang and H. Wang, “Missing data imputation: a fuzzy k-means clustering algorithm over sliding window”, Fuzzy Syst. Knowled. Discovery, vol. 14, 2009, pp. 133–137.

[16] K. Pelckmans, J. D. Brabanter and J. Suykens, “Handling missing values in support vector machine classifiers”, Neural Networks, vol. 18, 2005, pp. 684–692.

[17] C. P. Lim, J. H. Leong and M. M. Kuan, “A hybrid neural network system for pattern classification tasks with missing features”, IEEE Trans. Pattern Anal., vol. 27, 2005, pp. 648–653.

[18] R. Hathaway and J. Bezdek, “Clustering incomplete relational data using the non-euclidean relational fuzzy c-means algorithm”, Pattern Recogn. Lett., vol. 23, 2002, pp. 151–160.

[19] X. Wang, A. Li, Z. Jiang and H. Feng, “Missing value estimation for dna micro-array gene expression data by support vector regression imputation and orthogonal coding scheme”, Bmc Bioinform., vol. 7, no. 32, 2006, pp. 1–10.

[20] Timm, C. Doring and R. Kruse, “Different approaches to fuzzy clustering of incomplete datasets”, Int. J. Approx. Reason., vol. 35, 2004, pp. 239–249.

[21] A. Farhangfar, L. Kurgan and W. Pedrycz, “A novel framework for imputation of missing values in databases”, IEEE Trans. Syst. Man. Cybernet., vol. 37, no. 5, 2007, pp. 692–709.

[22] D. Li, H. Gu and L. Zhang, “A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data”, Expert Syst. Appl., vol. 37, 2010, pp. 6942–6947.

[23] A. D. Nuovo, “Missing data analysis with fuzzy c-means: a study of its application in a psychological scenario”, Expert Syst. Appl., vol. 38, 2011, pp. 6793–6797.

[24] L. Zhao, Z. Chen, Z. Yang, Y. Hu and M. S. Obaidat, “Local similarity imputation based on fast clustering for incomplete data in cyber-physical systems”, IEEE Systems Journal, vol. 12, no. 2, 2018, pp. 1610–1620.

[25] Q. Ma, Y. Gu, W. C. Lee and G. Yu, “Order-sensitive imputation for clustered missing values”, IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 1, 2019, pp. 166–180.

[26] A. Wang, Y. Chen, N. An, J. Yang, L. Li and L. Jiang, “Micro array missing value imputation: A regularized local learning method”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 3, 2019, pp. 980–993.

[27] Q. Ma, Y. Gu, W. C. Lee and G. Yu, “Order-sensitive imputation for clustered missing values”, IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 1, 2019, pp. 166–180.

[28] A. Ng, M. Jordan and Y. Weiss, “On spectral clustering: Analysis and an algorithm”, Advances in Neural Information Processing Systems, vol. 14, 2002, pp. 849–856.

[29] Mellenbergh, Gideon J, Missing Data, In Counteracting Methodological Errors in Behavioural Research, 2019, pp. 275- 292.

[30] Chai T, Draxler R, Root mean square error (rmse) or mean absolute error (mae), Geosci Model Dev Discuss, 7, 2014, 1525-1534.