Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

HFNA: A Hybrid Folding Neighbour Approach to Handle Missing Data

Dr. Aparna Shukla
Department of Computer Science & Engineering
Birla Institute of Technology
Mesra-Ranchi, India

Dr. Suvendu Kanungo
Department of Computer Science & Engineering
Birla Institute of Technology
Mesra-Ranchi, India

Dr. Vandana Bhattacharjee
Department of Computer Science & Engineering
Birla Institute of Technology
Mesra-Ranchi, India

Abstract - **The challenge of missing data is widespread across various domains, impacting the reliability and quality of data-driven analyses and models. Effectively addressing this challenge is imperative to uphold result integrity and facilitate accurate decision-making. The issue of missing data is a recurrent hurdle encountered in numerous research streams throughout the analysis process. Instances of substantial missing data significantly undermine the scientific reliability of causal inferences, underscoring the importance of researchers diligently addressing this concern to ensure the validity of their findings. The occurrence of missing data is influenced by diverse factors when collecting data from heterogeneous database sources, including manual data entry processes, errors in image acquisition equipment, low resolution, and other related aspects. This paper introduces an innovative method called the "Hybrid Folding Neighbour Approach" as a solution to address the challenge of missing data. This approach combines the advantages of multiple imputation techniques with a novel strategy known as neighbour folding. The paper explores various scenarios that arise due to the positioning of missing data points. By considering the location of the missing values, the concept of folding is applied to identify neighboring data points, facilitating the accurate prediction of suitable values for imputation.**

Keywords—Data Mining; Missing Data; Imputation Method; Nearest Neighbour

I. INTRODUCTION

Data Mining involves extracting pertinent subject-specific information from vast datasets that have been aggregated from a variety of disparate sources. Given the multiple amalgamations of data from diverse origins, there is a need to transform the data into coherent and valuable insights. Consequently, Data preprocessing emerges as the preliminary and essential step in the data analysis pipeline, mandated before the commencement of the actual data processing.

Today, the issue of missing data has become a big challenge for researchers, causing incorrect results. Missing data are found in many different areas of research and can lead to poor and inaccurate calculations. Missing data make it hard to explore the data properly and use new data effectively.

Missing data are symbolized as an excerpt of the qualities in the dataset that are either lost or not patrolled. The impurities and presence of missing data value impact greatly on researcher's work. The researchers confronting trouble handling missing which are very important for their research to lead to accurate and correct decision values.

In the realm of statistical analysis of data, decision statements play a crucial role as points of reference for making informed choices. Thus, the absence of data results in the loss of valuable information that significantly impacts the decision-making process. To tackle this issue, a multitude of techniques for filling in missing data have been presented in the literature by various researchers. These techniques generally fall into two categories: those rooted in statistical principles and others based on data mining concepts. Examples of methods involving parameters include the EM method and multiple imputation. While these methods can effectively address missing data, their success heavily relies on the user's selection of an appropriate model and a thorough understanding of the underlying data structure [1].

Data mining techniques often hinge on concepts like rough set methods, decision trees, and nearest-neighbor approaches. On the other hand, statistical methods are founded on principles such as mean, median, and statistical deviation. Among the various available methods, the nearest-neighbor approach stands out as a widely embraced technique for filling in missing data[2]. In the nearest neighbor method, the idea is that the closer the distance, the stronger the connection between items. The central aspect of this approach is the measurement of distance. Various distance measures, like Euclidean distance and Mahalanobis distance, have been employed by different researchers to address the challenge of missing data [3][4].

This paper introduces an approach to managing missing data through the use of an empirical dataset. The method proposed is executed and examined using MATLAB. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

paper concludes by highlighting the assumptions tied to these methodological approaches and offering research recommendations. The organization of the paper is structured into the following sections: (II) A Focus on Missing Data and Its Terminology, (III) Concentration on Missing Data Imputation Methods, (IV) Presentation of the Proposed Algorithm, (V) Outcome Discussion and Future Scope.

II. MISSING DATA TERMINOLOGY

The issue of missing data is common for various reasons, and it has a notable impact on the conclusions derived from the dataset under analysis. Missing data can emerge on two levels: either at the unit level or at the item level. Unit-level missing data occurs when respondents provide no information, leading to non-response. On the other hand, item-level missing data involves incomplete information being collected from a respondent[7]. Addressing missing data involves three key aspects: the proportion of missing data, the mechanism behind missing data, and the pattern of missing data.

Researchers should consider these three aspects mentioned above before selecting a suitable approach to address the issue of missing data.

A. *Proportion of Missing Data*

This particular aspect of missing data is closely tied to the reliability of statistical conclusions. While numerous methods have been introduced over time, there is still no established threshold for an acceptable percentage of missing data. Moreover, this aspect's influence on research outcomes drawn from the data is relatively minor compared to other factors[6][24].

Consider Table 1, a medical study analyzing patient records. If a significant proportion of patient ages are missing, it might affect the generalizability of conclusions about age-related health trends. However, if a smaller percentage of missing data relates to less crucial information, like a patient's preferred contact method, its impact on research outcomes might be less significant.

TABLE 1: PATIENT RECORDS

Patient Id	Age	Health Trend	Contact Method
01	48	Stable	Email
02	?	Improved	Phone
03	28	Decline	?
04	?	Stable	Phone

In the above Patient record, "?" represents missing values. The example illustrates how missing data in the "Age" column might significantly impact the research's ability to draw age-related health trend conclusions compared to missing data in the "Contact Method" column, which is less crucial for the health trend analysis.

B. *Missing Data Handling Mechanism*

The existence of missing data brings about several challenges. Firstly, it reduces statistical power due to the lack of data, affecting the test's ability to detect true differences. Secondly, estimates can become biased. Thirdly, the representation of samples can be compromised. Fourthly, it can complicate the study's analysis. Each of these issues has the potential to undermine the validity of the experiments and lead to incorrect conclusions. To address this, researchers have introduced various methods to manage missing data and ensure more accurate research outcomes.

Considering samples $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ of size n for each k random variables x_j where $j = 1,2,3,\ldots,k$. Thus, the data point $X = [x_{ij}]$ can be viewed as a matrix where there is a possibility of missing data. Rubin initially categorized and interpreted this matrix into three types of missing data based on postulates that align with the reasons for the data being missing[5]. Based on the missingness mechanism, three categories of missing data are defined: MAR, MCAR, and MNAR[8] [4].

1) *Missing Completely at Random(MCAR)*

Data is considered to be missing completely at random (MCAR) when the absence of observations is entirely independent of both observable and non-observable variables. This means the missing values in the data matrix X are randomly distributed. However, this assumption is often impractical in reality.

Suppose have survey data from a group of people about their favorite colors represented as in Table 2. However, some respondents left their favorite color blank randomly.

TABLE 2: PEOPLE'S FAVORITE COLOR RECORDS

Respondent	Favorite Color
1	Blue
2	Red
3	?
4	Yellow
5	?
6	Green
7	Purple
8	Orange

In this MCAR scenario, "?" represents the missing favorite colors scattered randomly across the respondents. The missing values have no apparent pattern or relation to any other variables. We can simply remove the rows with missing favorite colors without introducing bias since the missingness is not related to any specific characteristic of the respondents.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

2) Missing at Random(MAR)

In situations involving Missing at Random (MAR), the underlying assumption is the absence of discernible patterns. Data is categorized as MAR when there isn't a substantial distinction between the main variable of interest for the researcher and the missing and non-missing values.

In the MAR scenario, the absent value is estimated using existing data, essentially relying on the available information. The crucial aspect of MAR is that the missing value can be reasonably predicted using other variables under investigation. However, it's important to note that while the notion of prediction is valuable, it might not always precisely capture the underlying relationship.

For example-Table 3 represents a study conduted on drug use among a group of individuals, and suspect that income level might affect their willingness to answer questions about drug use. Poorer individuals might be less inclined to respond to such questions, leading to a correlation between drug use and income.

TABLE 3: PARTICIPANT'S DRUG USE RECORDS

Participant	Income Level	Drug Use (Yes/No)
1	High	Yes
2	Low	?
3	Mid	No
4	Low	?
5	Low	No
6	Mid	Yes
7	High	?
8	Low	Yes
9	Mid	?
10	High	?

In this scenario, the missing values "?" in the "Drug Use" column are likely related to the "Income Level" variable. Poorer individuals might be less likely to respond to questions about drug use, creating a situation where there's a connection between drug use and income level.

The above-stated example highlights the possibility of Missing at Random (MAR) data, where the missingness is related to another observed variable (income level). As a result, the correlation between drug use and income could be influenced by the missing data pattern.

3) Missing Not at Random(MNAR)

It is also known as "non-ignorable non-response." Data that doesn't fall into the categories of MCAR or MAR is classified as MNAR. Data is categorized as MNAR when the missing value of a variable is influenced by the reason it's missing, making it non-ignorable.

For example- Suppose we are studying the relationship between employees' job satisfaction and their willingness to disclose their salaries as in Table 4. Employees who are less satisfied might choose not to disclose their salaries, leading to missing salary values.

TABLE 4: EMPLOYEE RECORDS

Employee-Id	Job Satisfaction	Salary
01	4	60000
02	2	?
03	3	45000
04	1	?
05	2	?
06	4	70000
07	3	55000
08	1	?
09	5	85000
010	3	50000

In this MNAR scenario, employees with lower job satisfaction (like Employees 02, 04, 05, and 08) are less likely to disclose their salaries, leading to missing salary values. The missingness is directly related to the unobserved variable (job satisfaction) and is not random.

This example demonstrates a case of MNAR, where the missingness is influenced by an unobserved variable, making it challenging to handle and analyze the data without introducing biases.

C. Pattern of Missing Data

Suppose having a dataset D with n variables denoted as$D = \{Y_1, Y_2, \ldots, Y_n\}$. In general, three types of patterns of missing data are: univariate, monotone, and arbitrary.

A univariate pattern of missing data refers to a situation where certain row have missing values on one or more out of the total n variables[25]. These patterns can involve either complete or partial missing values.

A monotone pattern of missing data is observed when the missing values are organized in a manner where if a particular variable say Y_i is missing, then all subsequent variables $Y_k, k > j$ are also missing [26]. This pattern often arises in longitudinal studies, where if participants drop out at a certain point, all subsequent data points are missing for them in the subsequent measures as well.

A missing data pattern is termed as arbitrary when missing values appear randomly across variables for any participant. This means that any combination of variables might be missing for any given unit.

From a computational perspective, the first two patterns are more manageable than the last one, which is the arbitrary pattern.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

III. MISSING DATA IMPUTATION METHOD

In order to enhance the efficacy of data analysis, it is preferable to replace missing values through imputation rather than discarding observations. Imputation involves substituting suitable missing values and can be conducted at either the unit level or the item level. When a value is replaced for a complete data point, the process is termed 'unit imputation,' while when a value is substituted for a specific component of a data point, the process is known as 'item imputation.' The theory of imputation is continually evolving, necessitating ongoing focus to stay updated with new insights on the subject.

Researchers have adopted a variety of imputation methods to address missing data, spanning from straightforward approaches to more intricate ones. The presence of missing data introduces uncertainty in the data analysis process. As previously mentioned, a range of techniques has been suggested to handle the challenge of missing data, either from a statistical viewpoint or a data mining approach. The process of imputing missing values can occur prior to analysis (Pre-replacing methods like Mean/Mode, Hot Deck, KNN), or it can take place during the analysis itself (Embedded methods like Lazy decision tree, Dynamic path generation) [19].

Numerous statistical imputation methods have been documented in the literature, including the Mean-Mode imputation approach. In this method, the missing data value is replaced with the mean or mode of all the observed values for that specific variable. This technique keeps the dataset size constant and is straightforward to manage [6] [7] [18].

As the mean may not effectively identify outliers, Median imputation methods are more reliable in ensuring robustness. In this scenario, instead of utilizing the mean value, missing data is substituted with the median of all available values for a specific attribute within the class. Another statistical imputation approach is regression-based imputation [17], which differs from the mean imputation strategy. In the regression imputation method, an initial model is created to forecast observed variable values using other variables, and then this model is applied to fill in missing values wherever they appear in the problem domain.

Conversely, the data mining-oriented Pre-replaced imputation technique known as KNN exhibits a distinct approach to filling in missing data. KNN establishes a group of K nearest neighbors and subsequently replaces the missing value by calculating the mean of K non-missing values from its neighboring data points [11][20]. FKM is an extension of KNN based on fuzzy k-means clustering [9]. Some other imputation methods are also introduced so far based on Eigenvalues such as singular value decomposition (SVD) [8], and Bayesian principal component analysis (bPCA) [10]. Some of the authors have also developed imputation methods using Support vector machines [21] and many more [13] [22] [23]. A detailed survey of data imputation methods is presented in [27]. Authors in [28] have proposed a new technique for missing data imputation, which is a hybridized approach of single and multiple imputation techniques. They have extended the Multivariate Imputation by Chained Equation (MICE) algorithm. Likewise, other researchers in [29][30] have presented a detailed study on benchmark data imputation methods.

IV. PROPOSED ALGORITHM

This paper presents the '**H**ybrid **F**olding **N**eighbor **A**pproach' (HFNA) as a method for imputing missing values. The HFNA approach put forward in this paper offers various scenarios for replacing missing values based on the position of data points. To impute the missing value, the method calculates the mean of four neighboring values situated adjacent to the missing value in the variable. However, in cases where any of the adjacent neighbors is missing, this scenario only arises when the missing value is positioned at the corner edge of the matrix, referred to as a boundary element.

When dealing with a boundary element, the value that requires imputation is calculated using the folding method. Following the folding concept, elements located at corner edges are considered adjacent to elements at corner edges on the opposite side, achieved through either horizontal or vertical folding.

A. Flow Diagram of the Proposed Approach: HFNA

Fig. 1 illustrates the complete flow diagram of the proposed methodology. This diagram outlines how missing values are imputed within the dataset, emphasizing that various scenarios can arise based on the position of the missing value. The diagram illustrates two potential cases: the missing element may have all adjacent neighbors present, or the missing element might have some adjacent neighbors that are also missing. In the latter scenario, the flow diagram showcases the initial application of the folding concept to identify the missing neighbors. Subsequently, the mean is calculated to impute the missing value.

Through this approach, the missing values are imputed effectively, ensuring its efficiency across various scenarios.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

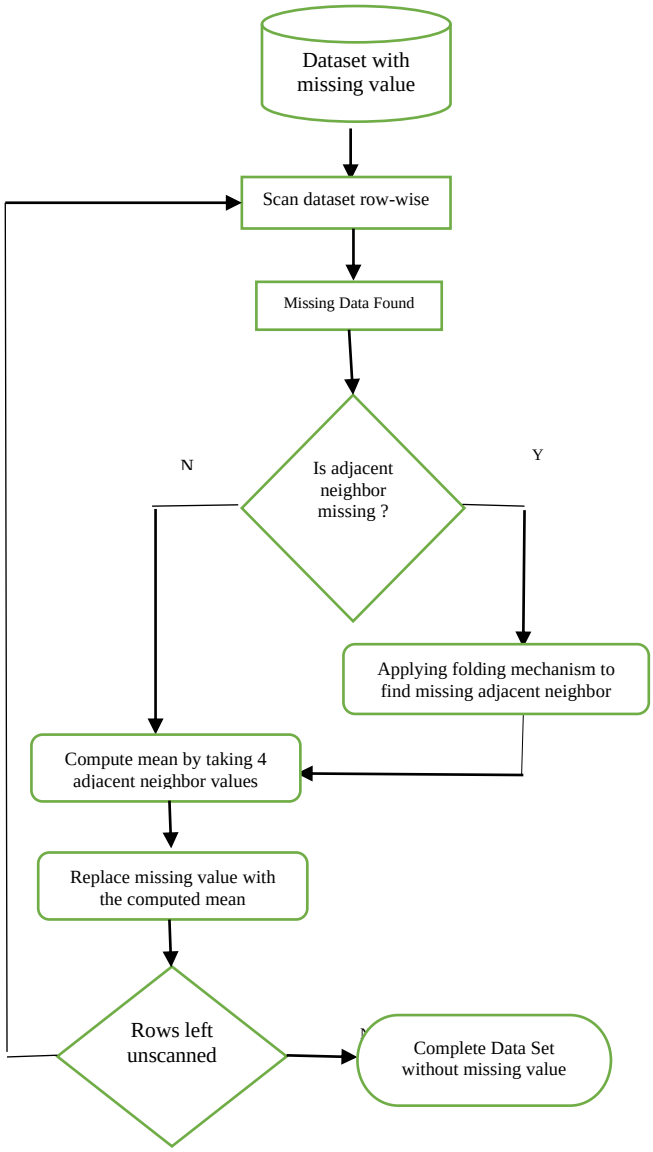

Fig. 1. Flow Chart of Proposed Approach: HFNA

B. Algorithm of Proposed Approach: HFNA

Suppose a data matrix D has a set of n variables denoted as $D = \{d_1, d_2, \ldots, d_n\}$may contain some missing values. In this paper, the author denoted some terms which are used in the proposed algorithm are d_{miss} refer to the missing value of the variable having all non-missing adjacent neighbors and $d_{boundary\_miss}$ refer to the missing value of the variable that does not have all adjacent neighbors with value.

Algorithm 1: Hybrid Folding Neighbor Approach (HFNA)

Input:

$D = \{d_{1m}, d_{2m}, \ldots, d_{nm}\}$ // *set of data items in data matrix* D *of size n* x *m.*

Output:

Imputation of missing values to complete data matrix

Steps:

1. **Initial Phase:** Scan the whole data matrix D row-wise.
2. **Tracing Phase:** Upon encountering a missing value within the data matrix, determine its position. Subsequently, based on the location of the missing data, two distinct cases can be identified:
   i. If the missing data, d_{miss} indicates that all adjacent neighbors are non-missing, i.e., all neighboring elements have values, Algorithm 1(i) is employed to impute this category of missing values.
   ii. If the missing data, $d_{boundary\_miss}$ indicates that some adjacent neighbors are missing, In such case, Algorithm 1(ii) is employed to impute this category of missing values.

Algorithm 1(i): ***Imputing*** d_{miss}

Steps:

1. Let the d_{miss} located at d (i, j), then find the mean of the adjacent neighbor by using equation 1.

$$d_{mean(i,j)} = \frac{1}{4}\sum\left(d_{i,j-1}, d_{i-1,j}, d_{i,j+1}, d_{i+1,j}\right) \quad (1)$$
$$\text{where } i \in N, j \in M$$

2. Replacing the d_{miss} by the computed mean in the respective position of the data matrix i.e., $d_{miss} = d_{mean(i,j)}$

Algorithm 1(ii): ***Imputing*** $d_{boundary\_miss}$ ***by using the concept of folding approach***

$d_{boundary_miss}$is boundary element i.e. it is located in the data matrix on any of the either first or last row edges or may be either first or last column edges. Therefore different cases are to be handled while dealing to impute the boundary element.

Steps:

1. Case 1: $d_{boundary_miss}$element is located at the first-row first column. The following equation is used to predict the value to be filled

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{i,M}, d_{N,j}, d_{i,j+1}, d_{i+1,j}\right) \quad (2)$$
$$where\ i \in N, j \in M$$

2. Case 2: $d_{boundary_miss}$element is located at the first row last column. The following equation is used to predict the value to be filled

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{i,j-1}, d_{N,j}, d_{i,1}, d_{i+1,j}\right) \quad (3)$$
$$\text{where } i \in N, j \in M$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

3. Case 3: $d_{boundary_miss}$element is located at the Last row first column. The following equation is used to predict the value to be filled

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{N,M}, d_{i-1,j}, d_{i,j+1}, d_{1,j}\right) \quad (4)$$
where $i \in N, j \in M$

4. Case 4: $d_{boundary_miss}$element is located at the Last row the Last column. The following equation is used to predict the value to be filled

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{i,j-1}, d_{i-1,j}, d_{N,1}, d_{1,j}\right) \quad (5)$$
where $i \in N, j \in M$

5. Case 5: $d_{boundary_miss}$element is located at the first row in between. Then applying the horizontal folding concept the following equation is used to predict the value to be filled

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{i,j-1}, d_{N,j}, d_{i,j+1}, d_{i+1,j}\right) \quad (6)$$
where $i \in N, j \in M$

6. Case 6: $d_{boundary_miss}$element is located at the Last row in between. Then applying the horizontal folding concept suitable value to be imputed is computed by the following equation

$$d_{boundary_{miss(i,j)}} = \frac{1}{4}\sum\left(d_{i,j-1}, d_{i-1,j}, d_{i,j+1}, d_{1,j}\right) \quad (7)$$
where $i \in N, j \in M$

7. Case 7: $d_{boundary_miss}$element is located at the first column in between. Then applying the vertical folding concept suitable value to be imputed is computed by the following equation

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{i,M}, d_{i-1,j}, d_{i,j+1}, d_{i+1,j}\right) \quad (8)$$
where $i \in N, j \in M$

8. Case 8: $d_{boundary_miss}$element is located at the Last column in between. By applying the vertical folding concept the following equation is used to predict the value to be filled

$$d_{boundary\_miss(i,j)} = \frac{1}{4}\sum\left(d_{i,j-1}, d_{i-1,j}, d_{i,1}, d_{i+1,j}\right) \quad (9)$$
where $i \in N, j \in M$

The various scenarios described above are applied for filling in missing values within a data matrix, with the approach depending on where these gaps are situated in the matrix. In Algorithm 1 (ii), both vertical and horizontal strategies are employed to estimate the missing value. This estimation is based on the relevant equation provided when the absent element is positioned at a corner. Conversely, when the missing element resides along an edge, only one of the folding concepts is used to calculate the predicted value that will be inserted.

V. DISCUSSION AND FUTURE SCOPE

In this segment, two distinct sets of datasets were examined, encompassing varying instance sizes: the Fisher-iris and Yeast datasets. The efficacy of the proposed HFNA method was substantiated through a series of experiments conducted on both datasets.

For 30 instances of the Fisher-iris dataset, with three attributes, the data imputation time was 0.03 seconds. For 150 instances with the same set of attributes, the time taken was 0.02 seconds. In the case of the yeast dataset, comprising 20 instances and 17 attributes, the process of data imputation was completed within 0.026 seconds. When dealing with 2884 instances bearing the same attributes, the imputation process was accomplished in 0.54 seconds.

Presently, the issue of missing data has emerged as a fresh challenge across various domains of research. Dealing with this challenge has become a significant hurdle for researchers, as the presence of missing data can potentially distort research outcomes. Numerous imputation techniques have been documented in the literature to address the task of estimating missing data values within variables.

In alignment with this line of thought, this paper introduces a novel approach to imputation. This approach explores diverse scenarios for handling missing values, taking into consideration their specific positions within the data matrix. This new dimension of the imputation method aims to enhance the understanding and management of missing data-related challenges. This paper employs the notion of folding, whether in a horizontal or vertical manner, to identify any absent adjacent neighbors in cases where missing data is detected.

The HFNA method introduced in this paper is showcased through various subsets of the dataset, specifically the Fisher Iris and Yeast datasets. As part of our ongoing work, the impact of data imputation on the performance of classifiers shall be investigated. This we believe would give us more insights about improving our techniques.

REFERENCES

[1] L. Xingyi and Z. Chunhua. “The Handling and challenges of missing data”, [J] Qinzhou University, (6), pp. 25-29, 2008.

[2] L. Xingyi and N. Guocai. “The comparison of several different filling missing values method,” [J] Nanning Teachers College, (3), pp. 148-150, 2007.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 08, August-2023

[3] L. Xingyi, T. Yao, and Z. Chunhua, "Filling missing data method based on the Mahalanobis distance," [J].Microcomputer Information, (9), pp. 225-226, 2010.

[4] L. Xingyi, "Filling missing value algorithm based on Mahalanobis distance and gray analysis", [J].Journal of Computer Applications, (9), pp. 2502-2506, 2009.

[5] D.B. Rubin, "Inference and missind data", Biometrika, vol. 63, no. 3, pp. 581–592, 1976.

[6] RJA Little and D.B. Rubin, "Statistical Analysis with Missing Data", (2nd edn.) New York: John Wiley and Sons, 2002.

[7] D.B. Rubin, "Multiple Imputation for Nonresponse in Survey", John Wiley and Sons, Inc, 1987.

[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, " Missing value estimation methods for DNA microarrays, Bioinforma. Oxf. Engl. vol. 17, no.6, pp. 520–525, 2001.

[9] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy k-means clustering method Springer Berlin Heidelberg vol. 3066, pp. 573-579, 2004.

[10] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data",‖ Bioinforma. Oxf. Engl. vol. 19, no. 16, pp. 2088–2096, 2003.

[11] S. G. Liao, Y. Lin, DD. Kang, D. Chandra, J. Bon, N. Kaminski, FC. Sciurba, and GC. Tsenq, "*Missing* value imputation in high-dimensional phenomic data: imputable or not, and how?" BMC Bioinformatics, vol 15, pp. 346-357, 2014.

[12] R. L. Vaishnav, and K. M. Patel, " Analysis of Various Techniques to Handling Missing Value in Dataset", International Journal of Innovative and Emerging Research in Engineering, vol. 2, no. 2, pp. 191-195, 2015.

[13] W, Shahad, Q. Rehman and E. Ahmed, "Missing Data Imputation using Genetic Algorithm for Supervised Learning," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 8, pp. 438-445, 2017

[14] J. Scheffer, "Dealing with Missing Data" Research Letters in the Information and Mathematical Sciences, vol 3, pp. 153-160, 2002.

[15] B. Mehala, P. R. J. Thangaiah, and K. Vivekanandan " Selecting Scalable Algorithms to Deal With Missing Values ". International Journal of Recent Trends in Engineering, vol. 1, no. 2, 2009

[16] A. Sumathi, "Missing Value Imputation Techniques Depth Survey And an Imputation Algorithm to Improve the Efficiency of Imputation". IEEE-Fourth International Conference on Advanced Computing, ICoAC 2012

[17] Y. Kou, C.T. Lu, and D. Chen. "Spatial weighted outlier detection". In Proceedings of the Sixth SIAM International Conference on Data Mining, pp. 613–617, 2006.

[18] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," Journal of Climate, vol. 14, pp. 853-871, 2001.

[19] Y. Fujikawa and T. Ho, "Cluster-based Algorithms for Filling Missing Values", Lecture Notes in Computer Science, vol. 2336, pp. 549-554, 2002.

[20] V. Kumutha and S. Palaniammal, "An Enhanced Approach on Handling Missing Values Using Bagging k-NN Imputation", International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, INDIA, 2013.

[21] T.R.Sivapriya, A.R.Nadira Banu Kamal and V. Thavavel, "Imputation And Classification Of Missing Data Using Least Square Support Vector Machines – A New Approach In Dementia Diagnosis", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, vol. 1, no. 4, pp. 29-34, 2012

[22] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2005.

[23] N. R.Pimplikar, A. Kumar, A. M. Gupta., "Study of Missing Value Imputation Methods – A Comparative Approach", An International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 3, pp. 1487-1491, 2014.

[24] S. Zhang, J. Zhang, X.Zhu, Y. Qin, and C.Zhang, "Missing Value Imputation Based on Data Clustering", Transactions on Computational Science (TCOS), vol. 1, pp. 128-138, 2008.

[25] H. Demirtas, "Flexible imputation of missing data", J Stat Softw, vol. 85, no. 1, pp. 1–5, 2018.

[26] Y. Dong, C-YJ. Peng, "Principled missing data methods for researchers", SpringerPlus, vol. 2, no. 1, 2013.

[27] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, " A survey on missing data in machine learning " *Journal of Big Data* **8**, 140, 2021.

[28] Khan, S.I., Hoque, A.S.M.L. SICE: an improved missing data imputation technique. J Big Data 7, 37 (2020). https://doi.org/10.1186/s40537-020-00313-w

[29] Alruhaymi, A. and Kim, C. (2021) Study on the Missing Data Mechanisms and Imputation Methods. Open Journal of Statistics, 11, 477-492. doi: 10.4236/ojs.2021.114030.

[30] Jäger S, Allhorn A and Bießmann F (2021) A Benchmark for Data Imputation Methods. *Front. Big Data* 4:693674. doi: 10.3389/fdata.2021.693674.