Atlantis Highlights in Engineering (AHE), volume 1
International Conference on Science and Technology (ICST 2018)

Analysis of Simple Data Imputation in Disease Dataset

Fetty Tri Anggraeny[1], Intan Yuniar Purbasari, M. Syahrul Munir, Faisal Muttaqin, Eka Prakarsa Mandyarta, Fawwaz Ali Akbar
Informatics Engineering
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Surabaya, Indonesia
[1]fettyanggraeny.if@upnjatim.ac.id

***Abstract*— In the statistical data collection it is very possible that there are variables that do not respond or in other words empty, called missing value, that can cause problems in data analysis. In this research we will analyze some simple imputation technique to solve the missing value problem, are zero imputation, mean imputation median imputation, and random imputation. This study used a Pima Indians, hepatitis and breast cancer Wisconsin dataset from UCI Machine Learning. We also compare with incomplete data removal technique. The application of various simple imputations in the disease dataset can increase the accuracy value when compared to deficient data deletion techniques. And the zero imputation technique shows the best performance compared to other imputation techniques and deficient data removal techniques.**

Keywords—analysis; simple Imputation; disease dataset; fuzzy c-means

I. INTRODUCTION

In the statistical data collection, it is highly possible that there are variables that do not respond or in other words empty, called missing value. Missing value can raise problems in the data analysis, so it needs handling to overcome them. Several ways can be done; the easiest way is to delete incomplete variables or instances. Removal of variable data means ignoring these variable factors in the statistical analysis, while incomplete instance deletion will cause a reduction of the problem instance. Another way to overcome missing value is by imputation techniques.

Imputation is a technique for handling missing values by filling in the position of the missing value with another value. The rule of imputation is to get the predictive value as close as possible to the missing value, in other words, the imputation tries to minimize the value between the missing value and the predicted value of the missing value. There are two types of data imputation techniques, namely simple imputation [1] and an approximation approach. Simple imputation uses general statistical values, e.g., zero values, mean, median, and random values, while the approximation approach uses prediction values based on other values in the same variable. Some approximation approaches include MiFoImpute [2], optimization impute [3], regression [4], nearest neighbors [5], shell neighbor [6].

In disease datasets, many data collection or measurement variables are prone to the emergence of missing values. Some disease datasets found on UCI Machine Learning include diabetes, hepatitis, and cancer.

To find out the effect of each simple imputation method, clustering is performed on a dataset that has no missing value. In this study, we used Fuzzy C-Means Clustering (FCM). The FCM clustering algorithm is a useful tool for clustering real s-dimensional data, but it is not directly applicable to the case of incomplete data [7]. The Fuzzy C-Means (FCM) method is chosen because it is a data grouping technique where the presence of each data point in a cluster is determined by the degree of membership. The basic concept of Fuzzy C-Means is to determine the center of the cluster group and each data has a membership degree for each cluster. Each data is not stated to be an absolute member of a cluster, but has a value of membership degree which states how much the data has similarities to the data in the cluster.

II. METHODOLOGY

A. Methods

This study requires several stages, can be seen in Fig. 1, which is divided into two major phases, namely training and testing. Each dataset will be divided into two groups randomly, 30% test data and 70% training data. Training data will be used to obtain the center value of each cluster, while the test data is used to measure the performance of fuzzy c-means clustering. Before going through FCM, the data goes through the preprocessing stages. At this stage, imputation will be carried out on incomplete data, so there is no data deletion. The imputation carried out in this study includes zero, mean, median, and random imputation. Testing results will be used to analyze the results of using the imputation method.

This research used a Pima Indians, hepatitis and breast cancer Wisconsin dataset from UCI Machine Learning. The distribution of training data and test data for each dataset can be seen in Table 1. The last column provides information on the number of instances with complete data.

The training phase is used to obtain the cluster center value that will be used as a reference for measuring distance to each test data. Distance values state the proximity of data to a cluster. A data is grouped into clusters with the smallest distance value.

B. Dataset

This study used a Pima Indians, hepatitis and breast cancer Wisconsin dataset from UCI Machine Learning. Pima Indians diabetes data consists of 768 instances, with 8 attributes and 1 class variable with 0 negative diabetes and 1 positive diabetes. A list of attributes in the Pima Indians diabetes dataset can be seen in Table 1. In this dataset, incomplete data is found in some attributes that are zero, but not all zero values are said to be incomplete. The zero value in the pregnant attribute can be assumed that this value states that the patient has never given birth. The zero value on the plasma-glucose attribute, diastolic blood pressure, triceps skinfold thickness, and body mass index are attributes of measurement results that should not be zero. If the value is zero, it is assumed that the variable is not measured, so that the instance can be removed. A zero value on the insulin attribute can occur if the object data does not use insulin injections to overcome blood sugar levels.

The second dataset is hepatitis, consisting of 155 instances with 20 attributes. All attributes that have null values are not responded because they are attributes of measurement results. Characteristics of hepatitis dataset attributes can be seen in Table 2.

And the third dataset is Wisconsin breast cancer, consisting of 699 instances with 11 attributes. Only the bare nuclei attribute has a null value and is not responding because it is an attribute of measurement results. The attribute characteristics of breast cancer dataset Wisconsin can be seen in Table 3.

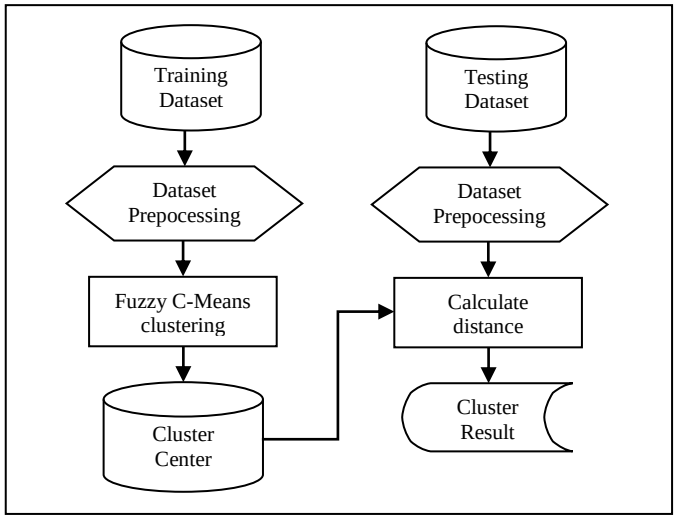

Fig. 1. Research Methodology.

TABLE I. THE DISTRIBUTION OF TRAINING DATA AND TEST DATA

Dataset	All	Training Data	Test Data	Complete Data
Pima Indians diabetes	768	512	256	532
Hepatitis	155	108	47	80
Breast cancer Wisconsin	698	489	209	683

TABLE II. CHARACTERISTICS OF PIMA INDIAN DATASET

Attributes	Value	Mean	Median	Std.Dev.	Zero value
Number of times pregnant	0 – 17	3.8	3	3.4	Not missing value
2-hours OGTT Plasma-Glucose (Mg/dL)	0 – 199	120.9	117	32	Missing value
Diastolic Blood Pressure (Mm Hg)	0 – 122	69.1	72	19.4	Missing value
Tricepts Skin Fold Thickness (mm)	0 – 99	20.5	23	16.0	Missing value
2-hours serum Insulin (Mu U/ml)	0 – 846	79.8	30.5	115.2	Not missing value
Body Mass Index (Kg/m^2)	0.0 – 67.1	32.0	32	7.9	Missing value
Diabetes Pedigree Function	0.0278 – 2.42	0.5	0.3725	0.3	-
Age (years)	21 – 81	33.2	29	11.8	-
Class	0 = Negatif; 1 = Positif	-	-	-	-

TABLE III. CHARACTERISTICS OF HEPATITIS DATASETS

Attributes	Value	Mean	Median	Std. Dev.	Missing value
Age	10-80	41.2	39	12.57	-
Sex	1:Male, female	1.10	1	0.31	-
Steroid	1:No, 2:yes	1.50	2	0.51	yes
Antivirals	1:No, 2:yes	1.85	2	0.36	-
Fatigue	1:No, 2:yes	1.34	1	0.49	yes
Malaise	1:No, 2:yes	1.59	2	0.51	yes
Anorexia	1:No, 2:yes	1.78	2	0.43	yes
Liver big	1:No, 2:yes	1.71	2	0.58	yes
Liver firm	1:No, 2:yes	1.47	2	0.63	yes
Splenpalpable	1:No, 2:yes	1.74	2	0.51	yes
Spiders	1:No, 2:yes	1.61	2	0.55	yes
Ascites	1:No, 2:yes	1.81	2	0.47	yes
Varices	1:No, 2:yes	1.82	2	0.46	yes
Bilirubin	0.39-4.00	1.37	1	1.22	yes
Alk phospate	33-250	85.62	84	62.06	yes
SGOT	13-500	83.68	55	89.53	yes
Albumin	2.1-6.0	3.42	3.9	1.32	yes
Protime	10-90	35.12	35	35.22	yes
Histology	1:No, 2:yes	1.45	1	0.50	-
Class	0:die, 1:live	-	-	-	-

C. Preprocessing

Not all instances provide complete information, so it requires treatment to resolve the incompleteness. The easiest way to overcome incomplete data is marginalized (WDS – whole data strategy), this method removes incomplete

TABLE IV. CHARACTERISTICS OF BREAST CANCER WISCONSIN DATASETS

Attributes	Value	Mean	Median	Std. Dev.	Null value
Sample code number id number	-	-	-	-	-
Clump thickness	1-10	4.42	4	2.82	-
Uniformity of cell size	1-10	3.14	1	3.05	-
Uniformity of cell shape	1-10	3.21	1	2.97	-
Marginal adhesion	1-10	2.81	1	2.86	-
Single epithelial cell size	1-10	3.22	2	2.21	-
Bare nuclei	1-10	3.46	1	3.64	yes
Bland chromatin	1-10	3.44	3	2.44	-
Normal nucleoli	1-10	2.87	1	3.05	-
Mitoses	1-10	1.59	1	1.72	-
Class	2 for benign and 4 for malignant	-	-	-	-

Instances so that there is a reduction in instances [8]. Besides deleting incomplete instances, lost data can be replaced with other data with several techniques. There are two techniques, namely simple imputation (zeros, mean, medians, and random values) and approximation approach (MiFoImpute [2], optimization impute [3], regression [4], nearest neighbors [5], shell neighbor [6]). In this study will use a simple imputation technique whose results will be compared to determine the impact of each imputation on the quality of the FCM clustering method.

D. Fuzzy C-Means Clustering

Clustering is the process of grouping data sets into groups/clusters so that objects in one group have similarities and have large differences with objects in other groups [9]. Differences and similarities are measured based on distance parameters. The smaller the distance value, the greater the similarity, and conversely the higher the distance value, the greater the difference. Clustering itself is also called an Unsupervised Classification, because it analyzes data without knowing the data label. The most important characteristic of good clustering results is that an instance is more "similar" to another instance in the same cluster than an instance outside of the cluster. The size of the similarity measure can vary and affect the calculation in determining the members of a cluster.

Reference [10] was introduced the Fuzzy C-Means (FCM) which is based on Dunn's study as in [11] which was the development of K-Means. The FCM is a soft algorithm for clustering fuzzy data in which an object is not only a member of a cluster but a member of many clusters in varying degree of membership as well. In this way, objects located on boundaries of clusters are not forced to fully belong to a certain cluster, but rather they can be a member of many clusters with a partial membership degree between 0 and 1 [12]. Fuzzy C-means (FCM) is a well-known clustering algorithm that shares the dataset into a fuzzy clusters with respect to the distance between the cluster center and the data point, we use Euclidean distance function [13].

The Fuzzy C-Means algorithm is as follows:

1. Input data to be clustered X, in the form of a matrix size nxm, where n is number of data samples, m is attribute data, X_{ij} is an i-th data instance (i = 1, 2, ..., n), and a j-th attributte (j = 1, 2, ..., m)

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_{11}} & \cdots & \boldsymbol{X_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{X_{n1}} & \cdots & \boldsymbol{X_{nm}} \end{bmatrix} \quad (1)$$

2. Specify
   c = Number of cluster,
   w = weighting exponent,
   maxiter = maximum iteration,
   ξ = minimum error,
   P_0 = 0, initial objective function,
   t = 1, initial iteration,
3. Generate random value μ_{ik}, i = 1, 2, ..., n; k = 1, 2, ..., c; as the initial partition matrix elements U. μ_{ik} is a degree of membership which refers to how likely a data can become a member in a cluster. The position and value of the matrix are built randomly. Where membership values are located at intervals of 0.0 to 1.0. At the initial position the U partition matrix is still not accurate as well as the cluster center, so the tendency of data to enter a cluster is not accurate. Calculate the number of each column (attribute):
   $$Q_j = \sum_{k=1}^{c} \mu_{ik}, \quad (2)$$
   Q_j is the total value of the degree of membership of each column, where j = 1, 2, ..., m; then calculated:
   $$\mu_{ik} = \frac{\mu_{ik}}{Q_j}, \quad (3)$$
4. Calculate k-th cluster center, V_{kj}, k = 1, 2, ..., c and j = 1, 2, ..., m:
   $$V_{kj} = \frac{\sum_{j=1}^{m}((\mu_{ik})^w \times X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^w} \quad (4)$$
5. The objective function is used as a looping condition to get the right cluster. So that the tendency of the data to enter the cluster in the final stage is obtained. Calculate the objective value in the t-iteration (P_t):
   $$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^w\right) \quad (5)$$
6. Calculate partition matrix i = 1, 2, ..., n and k = 1, 2, ..., c:
   $$\boldsymbol{\mu_{ik}} = \frac{\left[\sum_{j=1}^{m}(\boldsymbol{X_{ij}} - \boldsymbol{V_{kj}})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(\boldsymbol{X_{ij}} - \boldsymbol{V_{kj}})^2\right]^{\frac{-1}{w-1}}} \quad (6)$$
7. Check the stop condition:
   If **(**$|P_t - P_{t-1}| < \xi$) or ($t >$ *MaxIter* **)** then stop where P_t is the objective function of iteration to t less than the expected

error value or if *t* (number of iterations) is greater than the maximum iteration. If not, then *t*=*t*+1 and repeat to step 4.

III. Result and Discussion

The research was carried out in each dataset where there was a difference in imputation treatment with zero values, mean, median, and random. Each dataset is tested 50 times for each imputation; each trial uses a different group of instances that are randomly selected with 70% training data and 30% test data. The FCM parameters used in each trial are fixed value, weight=2, iteration=100, and minimum error=1e-5. We analyze the accuracy, specificity, and sensitivity.

Table 5 shows the accuracy value of each dataset with various imputation techniques used. Based on these trials, zero imputation techniques have the highest accuracy values in all datasets. Then the median imputation technique has the second highest accuracy value in the dataset of Pima Indians and breast cancer Wisconsin. Trials of datasets after reducing incomplete data also did not show good results.

We also try to analyze cluster using silhouette plot which has index value -1 until +1. If the silhouette index approaches to +1, then the sample is far from the neighbor cluster. If it is negative, the sample is in the wrong cluster. And if it approaches 0, then the sample is close to the boundary of neighboring cluster area. We compare the silhouette between the real cluster and result cluster. In this trial, we use all data on each dataset using zero imputation, and we choose one randomly set of the trial. All the silhouette plot can be seen in Fig. 2 until 4.

See Fig. 2, in Pima Indians dataset there are a lot of data have mapped to wrong real cluster, and FCM clustering fixes it, some data that previously had a negative silhouette index changed to positive. Fig. 3 hepatitis dataset, FCM shows a slight improvement in cluster one with the data being a positive index value and reducing cluster zero with the data changing to a negative value. And Fig. 3 breast cancer dataset, FCM do a good job in this dataset.

IV. Conclusion

The application of various simple imputations in the disease dataset can increase the accuracy value when compared to incomplete data deletion techniques. The zero imputation technique shows the best performance compared to other imputation techniques and incomplete data removal techniques.

TABLE V. The accuracy of Test Data

Dataset	Zero Imp.	Mean Imp.	Median Imp.	Random-value Imp.	All data
Pima Indians	0.65	0.51	0.54	0.51	0.43
Hepatitis	0.87	0.40	0.38	0.45	0.5
Breast Cancer	0.76	0.40	0.58	0.45	0.42

For the next research, the imputation approach approximation will be implemented. So the value of the missing value is more in line with the overall data.

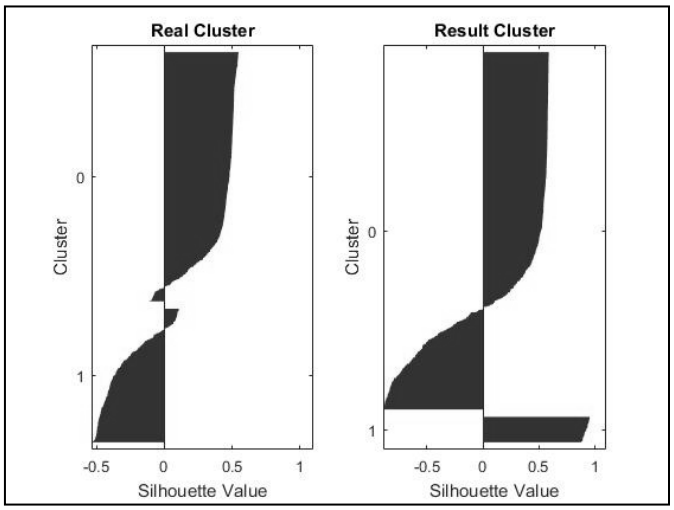

Fig. 2. Pima indians silhouette.

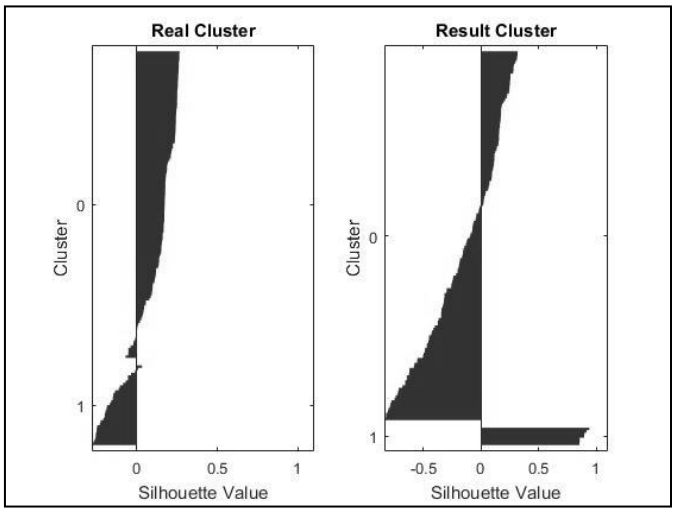

Fig. 3. Hepatitis silhouette.

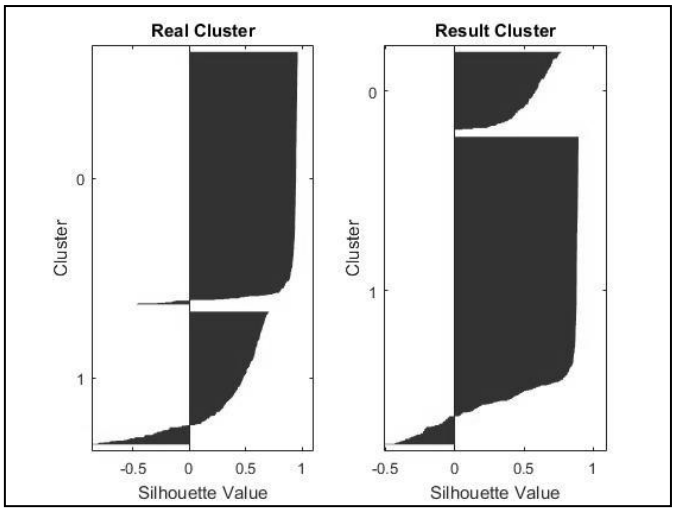

Fig. 4. Breast cancer silhouette.

REFERENCES

[1] D. T. Kadengye, W. Cools, E. Ceulemans, and W. Van den Noortgate, "Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data," Behav. Res. Methods, vol. 44, no. 2, pp. 516–531, Jun. 2012.

[2] C. M. Vastrad, "A Robust Missing Value Imputation Method Mifoimpute For Incomplete Molecular Descriptor Data And Comparative Analysis With Other Missing Value Imputation Methods," Int. J. Comput. Sci. Appl., vol. 3, 2013.

[3] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo, "From Predictive Methods to Missing Data Imputation: An Optimization Approach," 2018.

[4] P. L. Shopbell, M. C. Britton, R. Ebert, K. L. Wagstaff, and V. G. Laidler, "Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy," 2005.

[5] C. Zhang, X. Zhu, J. Zhang, Y. Qin, and S. Zhang, "GBKII: An Imputation Method for Missing Values," in Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1080–1087.

[6] S. Zhang, "Shell-neighbor method and its application in missing data imputation," Appl Intell, vol. 35, pp. 123–133, 2011.

[7] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," IEEE Trans. Syst. Man Cybern. Part B, vol. 31, no. 5, pp. 735–744, 2001.

[8] A. Matyja and K. Siminski, "Comparison of algorithms for clustering incomplete data," Found. Comput. Decis. Sci., vol. 39, no. 2, pp. 107–127, 2014.

[9] J. Han, J. Pei, and M. Kamber, "The Morgan Kaufmann Series in Data Management Systems: Data Mining: Concepts and Techniques (3rd Edition)." 2012.

[10] J. C. Bezdek, "Models for Pattern Recognition," in Pattern Recognition with Fuzzy Objective Function Algorithms, Boston, MA: Springer US, 1981, pp. 1–13.

[11] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," J. Cybern., vol. 3, no. 3, pp. 32–57, Jan. 1973.

[12] Z. Cebeci and F. Yildiz, "3 Zeynel Cebeci, Figen Yildiz: Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures Hungarian Association of Agricultural Informatics European Federation for Information Technology in Agriculture, Food and the Environment Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures I N F O," J. Agric. Informatics, vol. 6, no. 3, pp. 13–23, 2015.

[13] V. Novák, I. Perfilieva, and A. Dvořák, "Insight into fuzzy modeling," unpublished.