K-Nearest Neighbor Imputation for Missing Value in Hepatitis Data

Arifin Surya Alianso[a)], Lailis Syafaah, Amrul Faruq

Department of Electrical Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia

[a)]Corresponding author: arifinsuryaali@webmail.umm.ac.id

Abstract. There has been a growing occurrence of errors in a dataset, one of which is the incomplete data on an attribute or commonly acknowledged as a missing value, affecting the results of an analysis conducted for researchers. Attempt to address such issue includes the imputation, a method of filling in the missing value by Replacing the missing value with a possible value based on dataset information. This study aims to deal with missing values in albumin attribute hepatic data by utilizing K-Nearest Neighbor (KNN) imputation, performed by calculating the weight mean estimation for the number of K which has been determined. K is thus the closest observation, where in this study, the K that would be utilized is when K = 3, K = 5, K = 7, K = 9, and K = 15. To determine the accuracy of an imputation, an evaluation is performed by utilizing the Mean Square Error (MSE). Based on the results obtained in this study, the best accuracy of program calculations is obtained when K = 7 and the best MSE is achieved when K = 15.

PRELIMINARY

The common problem that often occurs is incomplete data on a particular variable or attribute. Whereas, quality data is highly expected when conducting data collection and research. Incomplete data is generally acknowledged as missing data or missing value. Missing value represents as the state of an empty value or incomplete value in a data [1]. Missing value is frequently due to accidental system errors, such as unreadable sensors into a database or device that receives input, or probably due to human error such as negligence in data collection [2].

The three common methods to handle missing values include: Case Deletion, Parameter Estimation and Imputation [3]. Case Deletion is conducted by deleting missing data such as Listwise deletion and Pairwise deletion methods [4]. Parameter estimation is performed by replacing the missing value with an estimated value such as the Expectation-Maximization algorithm[5]. The imputation is performed by replacing the missing value with a possible value based on the obtained information in the dataset [6].

In this case, the researcher proposes the application of KNN imputation method to complete an albumin attribute hepatitis data indicating a missing value, since a complete data provides an advantageous information for the researchers. The expected hepatitis data to be completed is the albumin attribute hepatitis data serving as the parameter with the highest gain value from the decision tree [7].

Several studies have proven that utilizing the imputation method in dealing with missing values could improve classification accuracy compared to without using imputation [8], such as in a study entitled "Missing data imputation using statical and machine learning methods in a real breast cancer problem". This study compares imputation methods based on machine learning techniques, which are KNN, Multilayer Perceptron (MLP), Self-Organizing Map (SOM) with imputation based on statistical techniques, such as: multiple imputation, mean, and hot deck. The obtained results indicate that the imputation method based on machine learning techniques presents a higher level of accuracy than that in imputation based on statistical techniques. Of the three machine learning imputation methods, the KNN method is evident to produce the best accuracy.

METHOD

The K-Nearest Neighbor imputation method applies the nearest neighbor technique commonly utilized in the classification process into the imputation process or filling in missing values. The method is conducted by employing the nearest neighbor value of similar attribute to fill in the missing value in another instance. The number of undertaken neighbors is similar depending on the input K value [9].

The K-Nearest Neighbor imputation algorithm is conducted as follows:

1. Determining the value of k, which is the number of desired closest observations.

1st International Conference on Technology, Informatics, and Engineering
AIP Conf. Proc. 2453, 020057-1–020057-5; https://doi.org/10.1063/5.0095625
Published by AIP Publishing. 978-0-7354-4356-3/$30.00

2. Calculating the Euclidian distance between the observation target and observation that does not contain missing values, with can calculated as following in the Equation 1:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{a=1}^{s} (x_a - y_a)^2} \quad (1)$$

where x_a is observation target vector with s variables such $x_a = [x_1, x_2, \ldots, x_s]$, y_a is observation vector that does not contain missing values with s variables such $y_a = [y_1, y_2, \ldots, y_s]$, $d(x, y)$ is a distance between x and y, a is value of the a-variable in 1,2, ..., s.

3. Navigating the observations which have a minimum value of $d(x, y)$.
4. Imputing missing data using the weighted mean of imputation procedure, with utilized in the Equation 2:

$$\hat{x}_a = \frac{1}{W} \sum_{k=1}^{K} w_k \, y_{ka} \quad (2)$$

where $\hat{x}_a$ is imputed value, y_{ka} is the value of the variable at the k-th observation, k = 1,2, …, K, w_k is weight of k-th observation.

To determine the percentage of imputed results, it is thus necessary to evaluate the imputation of missing data. In this study, Mean Squared Error (MSE) was utilized as an evaluation between program KNN imputation and manual KNN imputation. Smaller MSE value generates smaller prediction error, calculated by the following Equation 3 [10].

$$MSE = \frac{1}{n} \sum_{j=1}^{n} (X_j - F_j)^2 \quad (3)$$

X_j represents the actual data in the "th" period, serving as the predicted data in the "th" period and "n" represents the amount of data.jF_jj

The initial scenario in this study was performed to collect hepatitis data at UCI Machine Learning. The retrieved data has missing values in some attributes. In this study, the data was filled with albumin attributes. Then the imputation process was conducted by using KNN giving the values of K = 3, K = 5, K = 7, K = 9, and K = 15 in programmatic and manual calculations. After obtaining the results of each different K in the program and manual calculations, the step is progressed with the evaluation through using the Mean Square Error (MSE) by comparing the results of the imputed KNN program with the results of the imputed KNN in manual calculations. The process of using KNN is described in Figure 1.

The datasets were obtained from research sites collecting datasets for a long time and these datasets have been widely used in several tests. However, the applied dataset has a missing value. The following contains the source of the retrieved data https://archive.ics.uci.edu/ml/index.php.

In this hepatitis dataset, missing values are navigated. The missing value attribute pursued in this study is albumin. The following table presents the top 10 data from the hepatitis data Table 1.

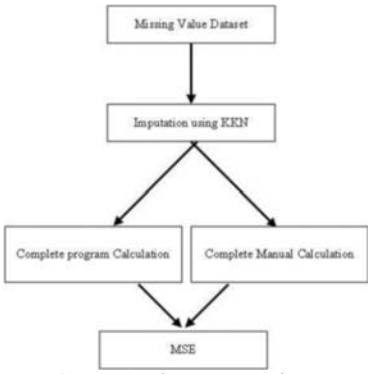

FIGURE 1. Blog System Diagram

TABLE 1. Hepatitis Dataset

Class	2	2	2	2	2	2	1	2	2	2
age	30	50	78	31	34	34	51	21	39	30
Sex	2	1	1	1	1	1	1	1	1	1
Steroids	1	1	2	?	2	2	1	2	2	2
Antivirals	2	2	2	1	2	2	2	2	2	2
Fatigue	2	1	1	2	2	2	1	2	1	2
Malaise	2	2	2	2	2	2	2	2	2	2
Anorexia	2	2	2	2	2	2	1	2	2	2
Big Liver	1	1	2	2	2	2	2	2	2	2
Liver Firm	2	2	2	2	2	2	2	2	1	2
Spleen palpable	2	2	2	2	2	2	1	2	2	2
Spiders	2	2	2	2	2	2	1	2	2	2
Ascites	2	2	2	2	2	2	2	2	2	2
Varices	2	2	2	2	2	2	2	2	2	2
Bilirubin	1.00	0.90	0.70	0.70	1.00	0.90	?	1.00	0.70	1.00
Alk Phosphate	85	135	96	46	?	95	?	?	?	?
Sgot	18	42	32	52	200	28	?	?	48	120
Albumin	4.0	3.5	4.0	4.0	4.0	4.0	?	?	4.4	3.9
Protime	?	?	?	80	?	75	?	?	?	?
Histology	1	1	1	1	1	1	1	1	1	1

RESULTS AND DISCUSSION

Imputation Result

In the hepatitis data utilized in this study, 16 missing values were navigated in the albumin attribute. For the calculation of KNN with manual imputation, the utilized attributes include age and sex. The following Table 2 and Table 3 presents the results of KNN imputation on albumin attribute hepatitis data.

TABLE 2. Results of KNN Imputation Program Calculation on Albumin Attributes

Obs	KNN Program Imputation Results				
	K=3	K=5	K=7	K=9	K=15
7	4.3	4.08	3.91	3.76	3.82
8	4.4	4.3	4.6	4.52	4.26
15	4.1	3.94	3.86	3.86	3.91
32	2.96	3.56	3.53	3.41	3.42
45	3.86	3.68	3.7	4.4	3.62
56	3.76	3.92	4.09	3.96	3.86
57	4.13	4.22	4.16	3.86	3.79
60	3.46	3.44	3.59	3.54	3.71
72	3.43	3.7	3.77	3.86	3.86
87	3.43	3.56	3.53	3.5	3.81
100	3.3	3.58	3.7	3.74	3.64
102	4.1	4.24	4.17	4.23	4.35
108	3.3	3.5	3.44	3.49	3.75
116	4.1	3.78	3.54	3.63	3.69
119	3.93	3.74	3.67	3.51	3.58
123	4.13	4.16	4.14	4.27	4.25

TABLE 3. Results of KNN Imputation Manual Calculation of Albumin Attributes

Obs	KNN Manual Imputation Results				
	K=3	K=5	K=7	K=9	K=15
7	4.27	3.94	3.94	3.71	3.63
8	4.2	4.7	4.39	4.28	4.13
15	3.27	3.4	3.46	3.34	3.65
32	3.93	3.8	3.84	3.66	3.54
45	4.23	3.86	3.77	3.81	3.81
56	3.73	3.92	3.81	3.74	3.69
57	4.23	3.86	3.77	3.81	3.81
60	3.2	3.54	3.5	3.67	3.52
72	4.23	3.86	3.77	3.81	3.81
87	3.2	3.48	3.53	3.53	3.53
100	3.83	3.5	3.43	3.52	3.59
102	4.3	4.4	4.14	4.1	4.25
108	3.87	3.84	3.64	3.47	3.57
116	3.33	3.26	3.5	3.6	3.46
119	3.93	3.56	3.33	3.43	3.51
123	3.47	3.88	3.97	3.97	4.01

Evaluation Result

Missing values were already filled in the hepatitis data albumin attribute from manual or program calculations, continued by calculating the MSE (Mean Squared Error) value. The following Table 4 displays the MSE results of KNN imputation with manual calculation.

TABLE 4. MSE Results of KNN Imputation with Manual Calculation

Total K	MSE
K=3	0.266
K=5	0.079
K=7	0.050
K=9	0.038
K=15	0.030

Table 4 indicates that MSE results of KNN imputation manual calculation indicate the best outcome when K = 15 with an error percentage of 0.030, which value is closest to 0.

CONCLUSION

Based on research on missing data imputation experiments conducted by employing K-Nearest Neighbor imputation on hepatitis data, it is concluded that missing values are overcome by implementing the nearest neighbor value of similar attribute to fill in the missing value in another instance. For the evaluation results using MSE, the best outcome is obtained when K = 15. Further, from the 5 'K' experiments, the best 'K' is obtained when using K=7.

ACKNOWLEDGEDMENT

The researchers would like to extend their gratitude to the Almighty Allah SWT with all His graces and gifts, to my beloved parents who always give encouragement, and the two supervisors, Mrs. Lailis Syafaah and Mr. Amrul Faruq who always provide advice and input to the researchers.

REFERENCES

1. R. Malarvizhi and A. S. Thanamani, "K-nearest neighbor in missing data imputation," *International Journal of Engineering Research and Development*, **5**, 5–7 (2012).

2. H. De Silva and A. S. Perera, “Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data,” in *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 141–146 (2016).
3. R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, **793**. John Wiley & Sons (2019).
4. G. King, J. Honaker, A. Joseph, and K. Scheve, “List-wise deletion is evil: what to do about missing data in political science,” (1998).
5. D. Li, J. Deogun, W. Spaulding, and B. Shuart, “Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method BT - Rough Sets and Current Trends in Computing,” 573–579 (2004).
6. A. Farhangfar, L. Kurgan, and J. Dy, “Impact of imputation of missing values on classification error for discrete data,” *Pattern Recognition*, **41**, 3692–3705 (2008),
7. W. D. Septiani, “Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis,” *None*, **13**, 76–84 (2017),
8. P. J. García-Laencina, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, and M. Verleysen, “K nearest neighbours with mutual information for simultaneous classification and missing data imputation,” *Neurocomputing*, **72**, 1483–1493 (2009),
9. U. Mawarsari, “Imputasi Missing Data Dengan K-nearest Neighbor Danalgoritma Genetika,” *AdMathEdu*, **6**, (2016).