A Distance-Threshold kNN Method for Imputing Medical Data Missing Values

Ching-Hsue Cheng and Hao-Hsuan Huang

***Abstract*—In medical filed, missing data often existed, which maybe result in the bias of research results. Therefore, this study proposes a new imputation method, that is a nearest neighborhood method based on distance threshold to impute missing value. The proposed imputation method has two merits: (1) utilize distance threshold to adjust the optimal nearest neighborhood for estimating missing values, (2) the proposed method compares with other imputation methods in medical data missing values. This study collected the stroke dataset from the International Stroke Trial (IST) to verify the proposed method, the result shows that the proposed method is better than other imputation methods, it means that the proposed method can be effectively utilized in practical medical dataset.**

***Index Terms*—Missing value, imputation, nearest neighborhood, stroke disease.**

I. Introduction

The rapidly development of information technology, the most important is to extract the valuable information in big data. Based on big data analysis, we can get available information, such as customer attribute analysis, customer behavior analysis, sales forecasting, etc. However, in practice, the data collected from real world usually has missing value; there are some reasons that can lead incomplete data such as human error in data input, problems of data transfer, and inaccurate measurement and so on.

In previous work for dealing with missing values, there are several common methods such as replacing the missing value with zero [1], average of other complete value (Average imputation), and average of same class value (Class Mean Imputation) [2], etc. After replacing the missing value, it usually had better accuracy than original datasets. Hence, dealing with missing values is an important issue in data mining and machine learning fields, and these methods are usually called imputation methods [1].

In medicine, breast cancer was the most prevalent cancer in women from 2014 [3], and cancer was the main cause of death in Taiwan since 1982 [4]. The breast cancer most occurs in women; especially it is difficult to collect data from women. Because it involving personal privacy that will let relevant information be rarely. In other word, relevant information is rarely, hence, every piece of information is important data source, even noise maybe have some interference effects. Therefore, this study proposed a new imputation method to handle missing value for incomplete medical data.

Because the real-world data usually have missing value, especially, medical data involves personal privacy, and it is difficult to collect complete data. Previous studies do not have considered the stability of model for different missing degrees and missing types. Therefore, this study proposes a new imputation algorithm for obtaining the better accuracy in different missing degree and missing types, and has great effectiveness in medical datasets with missing value. The main contributions of this study are listed as follows:

- Propose a nearest neighborhood imputation (DNNI) algorithm based on distance threshold to estimate missing values, and compare the performance of different missing degrees and missing type.
- Compare the proposed imputation method with popular imputation methods.
- Apply the proposed imputation method to international stroke trial dataset with missing value, and provide some findings.

The rest of the paper is organized as follows. Section 2 is the related researches including imputation techniques in medical domain, kNN algorithm and other popular imputation techniques. Section 3 explains the concept of the proposed method and proposed algorithm. Section 4 describes the experimental framework, environment, datasets, and experiment results. Finally, section 5 is conclusion.

II. Related Work

This section introduced the related literature including missing value in healthcare, concept of missing values, k nearest neighbor technique and common imputation methods.

A. Missing Value in Healthcare

In recent years, the imputation methods are still popular in the medical application. The imputation method is importance in medical, because the analysis results of medical data maybe influence the life of the patient. If there

Manuscript received March 12, 2019; revised June 20, 2019.

Ching-Hsue Cheng and Hao-Hsuan Huang are with Department of Information Management, National Yunlin University of Science & Technology, 123, section 3, University Road, Touliu, Yunlin, 640, Taiwan (e-mail: chcheng@yuntech.edu.tw, liang50641@gmail.com).

doi: 10.18178/jacn.2019.7.1.265

is error in data, physician will make wrong decision. It is very difficult to collect medical data, due to the data involving the patient's privacy, and patients are less likely to make public their personal data. Data with missing values, there are many reasons such as input errors, inaccurate measurement, equipment malfunction, measurement noise, data corruption, etc. In summary, imputation methods are indispensable tool in the medical field.

Many imputation techniques have been used in the medical field, such as multiple imputation, k nearest neighbor imputation, and expectation-maximization imputation, etc. [5]-[8]. Ondeck, Fu, Skrip, McLynn, Su and Grauer [5] used multiple imputation on arthroplasty research, they presented the results of compared between the demographics of patients with and without missing preoperative albumin and hematocrit values. Pedro J. García-Laencina , Pedro Henriques Abreu, Miguel Henriques Abreu and Afonoso [9] proposed the research of data imputation by using k nearest neighbor, mode, and expectation-maximization imputation on the 5 – year survival prediction of breast cancer patients with unknown discrete values, their research results of comparing k nearest neighbor imputation with mode and expectation-maximization imputation, the winner was k nearest neighbor imputation. Pombo, Rebelo, Araújo and Viana [10] proposed the Patient Oriented Method of Pain Evaluation System (POMPES), that produced tailored alarms, reports, and clinical guidance based on collected patient-reported data, which was clinical decision support systems (CCDSS), they used linear regression based on a least-squares estimation as imputation technique.

In medical domain, many imputation techniques have been applied, because k nearest neighbor is a simple operation, which is the popular method in data mining and medical domain.

B. Nearest Neighbor Technique

The k Nearest Neighbor (kNN) is the popular method in data mining, the k nearest neighbor algorithm was proposed by Cover and Hart [11]. kNN do not rely on building model during training phase, which classification rules are based on a similarity function between the training instances and the test instances to be classified. The kNN must measure the similarity of a new query instance, which the new instance does not have class label. Then a k nearest neighbors have been found, each neighbor casts a vote on the class to which the query instance is classified. Finally, the query instance is assigned to the class by result of the vote. When find the nearest neighbors, the preferred choice in nearest neighbor classification is to define in the Euclidean distance.

C. Missing Values

In real word, the collected dataset is usually incomplete for research, there are a variety reasons such as data from different source and combining errors, human error and noise in input phase, collect phase, etc. In practice, there are two ways to handle missing values, one is deleting (ignoring) missing values, which is the simplest approach, and it is namely marginalization. The other is replacing the missing values with estimating the missing value, it is namely imputation [1]. Marginalization leads to loss of raw data and will be degrade the quality of research, therefore, not within the scope of the study. And the aim of missing value imputation is to avoid deletion of the instances or attributes, and maintain the integrity of datasets. This study mainly focused on the research of missing value imputation techniques. The next part discussed the common missing imputation techniques and other research on practices.

D. Imputation Methods

In this section introduced some common imputation techniques including Zero Imputation, Average Imputation, Class Mean Imputation [2], and kNN Imputation.

(1) The Zero Imputation (ZI) is the simplest function, which fills the missing values with zero; the function of the zero imputation defined as equation (1):

$$\boldsymbol{ZI}(\boldsymbol{I}(\boldsymbol{i},\boldsymbol{j})) = \begin{cases} \mathbf{0} \ , & \boldsymbol{if}\ \boldsymbol{I}(\boldsymbol{i},\boldsymbol{j})\ \boldsymbol{is\ missing\ value} \\ \boldsymbol{I}(\boldsymbol{i},\boldsymbol{j}) \ , & \boldsymbol{otherwise} \end{cases} \tag{1}$$

(2) The Average Imputation (AI) is to replace the missing value with the averages of the corresponding attribute over the entire dataset; the function of the average imputation defined as equation (2):

$$\boldsymbol{AI}(\boldsymbol{I}(\boldsymbol{i},\boldsymbol{j})) = \begin{cases} \dfrac{\sum_{k=1}^{|S|} \boldsymbol{I}(\boldsymbol{k},\boldsymbol{j})}{|\boldsymbol{S}|} \ , & \boldsymbol{if}\ \boldsymbol{I}(\boldsymbol{i},\boldsymbol{j})\ \boldsymbol{is\ missing\ value} \\ \boldsymbol{I}(\boldsymbol{i},\boldsymbol{j}) \ , \quad \boldsymbol{otherwise} & \end{cases} \tag{2}$$

where S was complete instance, function $I(i, j)$ means i_{th} instance and j_{th} attribute, if $I(i, j)$ is missing value, calculate the average from set S to replace the missing attributes.

(3) The Class Mean Imputation (or Concept Mean Imputation) is a slight modification of AI where replaces the missing value with the average of the attribute over all instances within the same class label as the instance being filled, the function of class mean imputation is defined as equation (3):

$$\boldsymbol{CMI}(\boldsymbol{I}(\boldsymbol{i},\boldsymbol{j}),\boldsymbol{C}) = \begin{cases} \dfrac{\sum_{k=1}^{|T|} \boldsymbol{I}(\boldsymbol{k},\boldsymbol{j})}{|\boldsymbol{T}|} \ , & \boldsymbol{if}\ \boldsymbol{I}(\boldsymbol{i},\boldsymbol{j})\ \boldsymbol{is\ missing\ value} \\ \boldsymbol{I}(\boldsymbol{i},\boldsymbol{j}), \boldsymbol{otherwise} & \end{cases} \tag{3}$$

where T is complete instance, class belongs to $C(i)$, the missing values is replaced by the average from the set T, which is another complete instances and with the same class as label $C(i)$, $C(i)$ means the class label for i_{th} instance.

(4) The kNN algorithm was developed on missing values imputation [12], which is named k Nearest Neighbor

Imputation (kNNI). In kNNI, data have been split into two datasets; one is incomplete data, which contains missing value; the other one is complete data, which without missing data. The missing value of incomplete data was replaced by the average of corresponding attribute of its k nearest neighbors that is instance of complete data. However, this method tends to cover noise and outlier to be part of the predictive value, and then leads to predictive problematic outcomes and affecting the effectiveness of the classification. Furthermore, Troyanskaya *et al*. [13] proposed a weighted k nearest neighbor imputation (WkNNI) which was another imputation method based on k nearest neighbor technique.

III. Proposed Method

The concept of proposed method is to find a set of nearest neighbors and compute the estimated value for replacing the missing value. The set of nearest neighbors is computed by the distance between the instance of missing value and complete data instances. The detailed content of proposed method is described as follow.

The first step is to calculate the weight set of distance between each incomplete data and all the complete data by using equation (4) [14]. Equation (4) is defined as follow:

$$\mathbf{W} = \left[\mathbf{w}_1 \mathbf{w}_2 \cdots \mathbf{w}_m\right]$$

$$= \left[\frac{|\mathbf{X}|}{2\sum_{x\epsilon X}||y_1 - x_1||^{\frac{2}{(q-1)}}}, \frac{|\mathbf{X}|}{2\sum_{x\epsilon X}||y_2 - x_2||^{\frac{2}{(q-1)}}}, \ldots, \frac{|\mathbf{X}|}{2\sum_{x\epsilon X}||y_m - x_\mathrm{m}||^{\frac{2}{(q-1)}}}\right] \quad (4)$$

W is a weight set of all incomplete data, w_m is the m_{th} incomplete data weight, $|X|$ is the complete data of training, y_1 is the first instance of incomplete data, x_1 is the first instance of complete data, *q* is the weighted distance parameters. In other words, *q* can control the pattern of distance between *x* and *y*. Such as, when $q = 2$, the distance between *x* and *y* is the Euclidean distance.

The second step utilizes *W* weight set to compute the weighted distance between *x* and *y*, the calculated method is defined as Equation (5).

$$\boldsymbol{D}_{ij} = [\boldsymbol{d}_{i1}, \boldsymbol{d}_{i2,} \ldots \boldsymbol{d}_{in}]$$
$$= [(\mathrm{w}_{\boldsymbol{i}}(\boldsymbol{y}_i - \boldsymbol{x}_1)^2), (\mathrm{w}_{\boldsymbol{i}}(\boldsymbol{y}_i - \boldsymbol{x}_2)^2), \ldots, (\mathrm{w}_{\boldsymbol{i}}(\boldsymbol{y}_i - \boldsymbol{x}_n)^2)] \quad (5)$$

The D_{ij} represents i_{th} instance and its j_{th} attribute, which is the set of weighted distance between *x* and *y*, d_{ij} is the weighted distance between y_i and x_i. And w_i is the weight of the y_i incomplete data, y_i is the ith incomplete data, x_i is the i_{th} complete data. The set of weighted distance D_{ij} is reference points in this study, such as, if d_{ij} is less than the self-defined threshold, x_i will be added to the set of nearest neighbors.

Third step, after all of D_{ij} has been computed, the proposed method employed some central tendencies to calculate the estimated missing value, and these central tendencies include average, median, and geometric mean. And then this study compares the result of imputation with these three central tendencies under accuracy. The average central tendency is defined as equation (6).

$$\boldsymbol{M}_{\boldsymbol{a}}(\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{X}) = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3 + \cdots \boldsymbol{x}_n}{\boldsymbol{n}} \quad (6)$$

where (i, j) represents i_{th} instance of j_{th} attribute, *X* denotes the set of nearest neighbors; n is the number of nearest neighbors set.

A. Proposed Algorithm

For easily understanding the steps of the proposed method, the proposed procedure included data collection, data preprocessing, imputation, and classification & evaluation, and the detailed step are described as step 1 to step 4.

1) Data collection

This study collected the real international stroke trial dataset [15]. In real international stroke trial medical experiment, the collected dataset already contains missing values, which missing data type is belong to MNAR missing data type. Therefore, we directly impute the missing value by proposed method and compare with other methods.

2) Data preprocessing

In medical experiment, the collected dataset already has missing values, we directly impute the missing value and compare with other methods. In addition, we delete some unrelated attributes in patient death of stroke, that is, collected date, comment, and other death. The reason of deleting attributes is the three attributes unrelated with research objective for predicting patient death of stroke. Finally, we delete the instance of patient who is alive, and distinguish the remained data into two classes (died of stroke/ died of other).

3) Data imputation

In data Imputation phase, the missing value is replaced by estimated value of the proposed imputation techniques. In experiment, this study selected the most common imputation techniques such as, ZI, AI, CMI, MI and kNNI to compare with proposed imputation method. This study proposed a nearest neighborhood imputation based on distance threshold method. The following will introduce the proposed imputation method in detail.

- Before imputation, dataset must be normalized in domain [0, 1]. Because the proposed method is based on distance computation, and the normalized data in same scale can avoid different ranges of attribute for affecting distance calculation. Then, dataset was spilt into two data, one is the incomplete data that contains the missing values, the other is the complete data without the missing values.

- In imputation stage, first calculate the weight set of distance between each incomplete data and all the complete data by using equation (4), the weights of all incomplete data have been stored in the set of weight W. Next, employ equation (5) to calculate a weighted distance d_{ij}. Then, we pre-defined a threshold, if d_{ij} less than the threshold, which added to set of reference points. Lastly, use the reference points to calculate the estimated value by using equation (6). Moreover, we can directly replace the missing value with the estimated value, because this study employs classifier to evaluate their accuracy, the estimate value need not de-normalize to original data. The proposed imputation algorithm is shown in Algorithm 1.

4) *Classification and evaluation*

In classification and evaluation phase, this step compares the performance of proposed imputation with the listing methods, five classifiers: J48 [16], KNN [11], MLP [17], SVM [18], and RF [19] are employed to classify the imputed dataset for accuracy comparison. Accuracy is popular metric to evaluate the performance of classifiers. This study utilizes training accuracy as evaluation indicators because it shows the integrity of data after imputed data, and accuracy is defined as equation (7).

$$\boldsymbol{accuracy} = \frac{\boldsymbol{True\ Positive + True\ negative}}{\boldsymbol{All\ of\ the\ classify\ instances}} \tag{7}$$

Algorithm 1: The proposed imputation computation

```
        Let C = {$x_1$, $x_2$, …, $x_n$} be a set with n complete instances,
        M = {$y_1$, $y_2$, …,$y_m$} be a set with m incomplete instances.

BEGIN
        Set w = null, 0 < threshold < 1
        FOR i = 1 to m:
                Initialize sum = null
                FOR j = 1 to n:
                                Compute distance from $y_i$ to $x_j$
                                Sum be the total distance of $y_i$ to C
                W(i) = |C|/2 * sum
        FOR k = 1 to m:
                Initialize d = 0.
                FOR l = 1 to n:
                IF(class of $y_k$ and $x_l$ is same)
                                THEN d = $W_k$ * distance of $y_k$ to $x_l$
                                IF d < threshold
                                THEN Include $x_l$ in the set of reference points
        Replace the missing values in $y_k$ as the average of reference points
END

Output: instance without missing values
```

IV. Experiment and Results

The experimental environment is Python (Python 2.7 version) on Intel i7-4710MQ with 2.5 GHz CPU and Windows 10 Operating system, then use five classifiers to verify and compare the proposed imputation with the listing methods. In order to compare the performance of proposed imputation method with the listing methods, this study uses the 10 times average of training accuracy for each classifier, as a comparison indicator. This study uses practical medical dataset from International Stroke Trial [15], which is collected between 1991 and 1996. There are 112 attributes and 19435 instances in original dataset. Firstly, the objective of our research is to predict stroke death by using medical IST dataset, then we delete some attributes, which is irrelevant research objective such as date, and comment, etc. Next, we delete the instance of patient who is alive, and create two classes label (died by stroke/ died by other). Lastly, there are 69 attributes with 4242 instances and 6578 missing value in the IST dataset.

Table I show the result of stroke dataset experiment for different imputation methods, which shows that the proposed imputation method is better than other imputation methods, except MI in kNN classifier. From the average accuracy of five classifiers, we can see that the proposed method is better than other imputation methods.

TABLE I: The Accuracy of Stroke Datasets

	J48	KNN	MLP	RF	SVM	Ave.
AI	92.20	63.44	87.24	70.75	75.78	77.88
CMI	92.20	63.02	89.53	72.16	75.81	78.54
KNN	92.55	65.09	90.05	72.64	76.07	79.28
MI	90.53	**66.48**	95.30	73.51	78.47	81.66
ZI	92.20	64.15	88.21	73.34	75.78	78.74
DNNI	**96.44**	65.56	**95.76**	**88.67**	**78.76**	**85.04**

A. *Finding*

From experimental results, we will summarize some findings as follows:

1) *Threshold setting*

This study utilizes distance threshold to adjust the optimal nearest neighborhood for estimating missing values. After experiments, the optimal threshold of the proposed method is less than two, because the experimental dataset has been normalized in the range [0, 1], only a few of distances between central tendency and instance are greater than two. Moreover, if the distance is greater than two, the instance will be belonged to outlier.

2) *Different imputation methods*

After experiment, the results have shown that the proposed method is better than other imputation methods as Table 8-9. However, when missing degree increases, the accuracy of all imputation methods would decrease; it means that the dataset has reached higher uncertainty. If the numbers of missing values are bigger than the numbers of instances for dataset, even all instances have missing values, we find that the MI imputation cannot implement all instances with missing values.

V. Conclusion

This study has proposed a new imputation method based on distance threshold method for estimating missing values, and this study collected practical medical dataset with missing values, that is, the International Stroke Trial dataset was employed to verify the proposed method in practice. The experimental results show that the proposed method is better than other imputation methods, it means that the proposed method can handling missing values.

In future work, this study will choose eight datasets from UCI datasets to verify proposed imputation method in MAR and MCAR type experiments. Furthermore, feature selection techniques can be combined with the proposed method, which will enhance the performance.

References

[1] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing,* vol. 205, pp. 152-164, 2016.

[2] A. R. Donders, G. J. van. der Heijden, T. Stijnen, and K. G. Moons, "Review: A gentle introduction to imputation of missing values," *Journal of clinical epidemiology,* vol. 59, pp. 1087-1091.

[3] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics," *CA: A Cancer Journal for Clinicians,* vol. 64, pp. 9-29, 2014.

[4] Y. S. Lee, C. C. Hsu, S. F. Weng, H. J. Lin, J. J. Wang, S. B. Su, C. C. Huang, and H. R. Guo, "Cancer incidence in physicians: A Taiwan national population-based cohort study," *Medicine,* vol. 94, p. 2079, 2015.

[5] N. T. Ondeck, M. C. Fu, L. A. Skrip, R. P. McLynn, E. P. Su, and J. N. Grauer, "Treatments of missing values in large national data affects conclusions: The impact of multiple imputation on arthroplasty research," *The Journal of Arthroplasty,* vol. 33, pp. 661-667, 2018.

[6] K. Mühlenbruch, O. Kuxhaus, R. di Giuseppe, H. Boeing, C. Weikert, and M. B. Schulze, "Multiple imputation was a valid approach to estimate absolute risk from a prediction model based on caseecohort data," *Journal of Clinical Epidemiology,* vol. 84, pp. 130-141, 2017.

[7] C. K. Enders, "Multiple imputation as a flexible tool for missing data handling in clinical research," *Behaviour Research and Therapy,* vol. 98, pp. 4-18, 2017.

[8] C. O. Galán, F. S. Lasheras, F. J. C. Juez, and A. B. Sánchez, "Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions," *Journal of Computational and Applied Mathematics,* vol. 311, pp. 704-717, 2017.

[9] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Computers in Biology and Medicine,* vol. 59, pp. 125-133, 2015.

[10] N. Pombo, P.Rebelo, P. Araújo, and J. Viana, "Design and evaluation of a decision support system for pain management based on data imputation and statistical models," *Measurement,* vol. 93, pp. 480-489, 2016.

[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory,* vol. 13, pp. 21-27, 1967.

[12] G. E. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence,* vol. 17, pp. 519-533, 2003.

[13] 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," *Bioinformatics,* vol. 17, pp. 520-525, 2001.

[14] M. Sarkar, "Fuzzy-rough nearest neighbor algorithms in classification," *Fuzzy Sets and Systems,* vol. 158, pp. 2134-2152, 2007.

[15] P. A. G. Sandercock, M. Niewada, and A. Członkowska. (2011). The international stroke trial collaborative group. *The International Stroke Trial database, Sandercock et al. Trials*. [Online]. Available: http://www.trialsjournal.com/content/12/1/10

[16] J. R. Quinlan, "C4. 5: Programming for machine learning," *Morgan Kauffmann,* 1993.

[17] S. Mitra and S. K. Pal, "Fuzzy multi-layer perceptron, inferencing and rule generation," *IEEE Transactions on Neural Networks,* vol. 6, pp. 51-63, 1995.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[19] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

Ching-Hsue Cheng is born in Taiwan in 1960. He received his bachelor′s degree in mathematics from the Chinese Military Academy in 1982, his master′s degree in applied mathematics from Chung-Yuan Christian University in 1988, and his Ph.D degree in system engineering and management from the National Defence University in 1994. Now he is a professor of the Information Management Department in the National Yunlin University of Science and Technology. His research is mainly in the field of healthcare quality, medical knowledge discovery fuzzy time series, and soft computing. He has published more than 200 papers (including 134 significant journal papers). His research is mainly in the field of healthcare quality, medical knowledge discovery fuzzy time series, and soft computing.

Hao-Hsuan Huang is born in Taiwan in 1994, he received his bachelor′s degree in information management from National Yunlin University of Science and Technology in 2016, his master′s degree in information management from National Yunlin University of Science and Technology in 2018. He is now working at computer engineer in China Medical University Hospital, and he is in charge of multiple medical information systems. His research interest is information system development and missing value imputation.