ISSN 2616-5775 Vol. 4, Issue 8: 63-66, DOI: 10.25236/AJCIS.2021.040812

An efficient method to classification with missing data

Haojian Huang

College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang, 150001, China

Abstract: *Missing data is widely existing in life. Processing missing data is essential in classification. Therefore, it is a common and essential method to use the existing reliable data set to impute the missing data. These methods have a significant effect on the processing of ambiguity and uncertainty in the data set. At the same time, the processing of missing data sets widely exists in the fields of noise processing and enhancing system robustness. Therefore, using accurate data and imputation methods to impute missing data sets is essential and effective. In this paper, a new method for classification with missing data is proposed. First of all, the training data set is optimized so that classifiers can get trained well. Then it will be used to estimate missing values with the proposed method. By comparing the Precision, Recall, F_1, and ARI indicators of the classifier in the classification test with different testing data sets by four different imputation methods, the final result shows that the proposed method performs best on the whole.*

Keywords: *Missing data, Imputation, Machine Learning*

1. Introduction

Data classification is an essential task in pattern recognition, widely used in various fields, such as image recognition, video understanding, remote sensing image processing, and so on [1]. At present, many of the existing classification methods assume that the data set is complete, but the data is usually affected by noise or even missing. It will significantly affect the performance of these data classification methods. Therefore, dealing with missing values is an important issue in data mining and machine learning communities.

Based on the process undergone to produce a dataset and the nature of the data itself, there could be three kinds of missing mechanisms: missing completely at random (MCAR) [2], missing at random (MAR) [3], and not missing at random (NMAR) [4]. There are also two types of imputation for missing values: single-value imputation and multi-value imputation. The latter has a huge amount of computation, which makes it a big challenge. For single value imputation, we already have many methods. The easiest way to deal with the missing data is to discard them directly. It will lead to the loss of useful information and have a greater impact on research. Therefore, the more common method is to perform data imputation, such as zero imputation (ZI) [5], mean imputation (MI) [6], k-nearest neighbor imputation (KNNI) [7][8], support vector machine imputation (SVMI) [9], neural network imputation [10], etc. As for ZI and MI, the missing value is replaced by 0 or the average of all known values of the attribute. However, there is a significant disadvantage: estimates may lead to changes in the original distribution, thereby reducing the quality of the estimate and skewing the results. KNNI, one of the most prevalent methods, employs the neighbors to estimate the missing values. [11]. However, they have a drawback. Finding k-nearest neighbor based on global strategy may be affected by some outliers, leading to the increase of estimation error. The original KNN algorithm uses the standard Euclidean distance of nearest neighbors as the criterion for finding neighbors in the multi-attribute sample data set [12]. Actually, this can cause significant errors as it finds neighbors that are of different classes in the whole data set for estimation.

In this paper, an efficient method to classification for missing data is presented. It will estimate the class to which the data with missing values roughly belongs and then impute it with nearest neighbors of the same class, which ensures the imputation is relatively accurate and reliable. What's more, it will optimize the training set to reduce the influence of noise before searching the nearest neighbors. These are the reasons why the proposed method can often perform better in the task of classifying multi-attribute data sets. The proposed and ZI, MI, and KNNI methods were applied to classify data sets with missing data and comparative analysis.

In section 2, the proposed method will be introduced in detail for the improvements of the classic

ISSN 2616-5775 Vol. 4, Issue 8: 63-66, DOI: 10.25236/AJCIS.2021.040812

nearest neighbors imputation method. In the experiment, the performance of the 4 imputation methods are compared and analyzed. Finally, a conclusion will be drawn in the last section.

2. Proposed methods

Unlike the KNN algorithm, the proposed method enhances the understanding of the data set. As we know, the obtained data sets often contain lots of noise, which will make the imputation biased. Therefore, at the start, the proposed method will optimize the training set to reduce noise.

A testing set is considered $X = \{x_1, \ldots, x_n\}$ with missing values and a training data set $Y = \{y_1, \ldots, y_m\}$ is classified with s-dimensional attributes in the class editing framework. By calculating the center of each class in the training set and calculating the standard Euclidean distance between each sample and its class center, the training set is sorted in ascending order in terms of standard Euclidean distance d_c.

$$d_{c_i}^2 = \sum_{k=1}^{n} \frac{(x_{ik} - x_{ck})^2}{\sigma_k} \tag{1}$$

Where x_i is the i^{th} sample in the optimized data set, x_c is the center of class, σ is variance, and k is the k-th attribute.

$$D = \{\ d_{c_1},\quad d_{c_2},\quad \cdots d_{c_n} \big| d_{c_i} < \delta\} \tag{2}$$

Next, a reasonable threshold δ is set to eliminate the samples that have a large standard Euclidean distance with the center. Actually, the δ is determined empirically for the proper reduction of noise. In this way, an optimized data set gets ready.

Then an estimation on the class which the sample with missing values belongs is performed. It performs class estimation by calculating the Euclidean distance between samples with missing values and centers of different classes in the complete and reliable data sets. The smaller the Euclidean distance, the more likely the sample is to belong to the class. Therefore, Eq.(1) will be applied. For each sample, the class it belongs will have the minimum d_c with it. After making the best estimation of the class that sample belongs, the imputation will be calculated in terms of the weighted Euclidean distance of its nearest neighbors. By calculating the Euclidean distance between each sample in the training set, several samples with the smallest Euclidean distance are selected as the nearest neighbors of the sample with missing data. After normalizing the standard Euclidean distance of these points, the weight of each neighbor w_i is calculated according to Eq.(3).

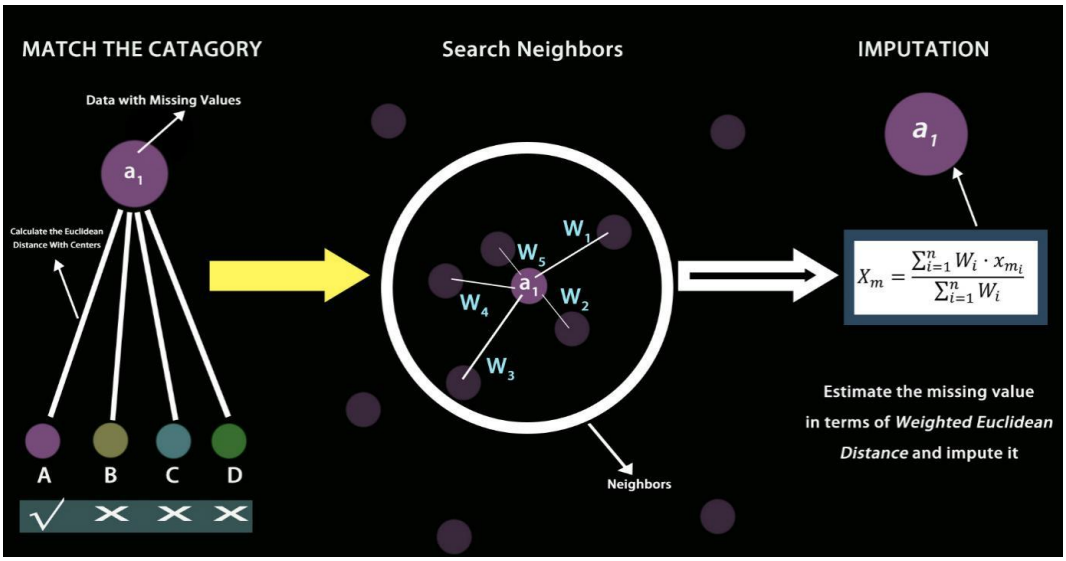

Figure 1: The procedure of imputing

$$w_i = \frac{D_i}{\sum_{j=1}^{n} D_j}, \quad i \le j \tag{3}$$

$$X_m = \frac{\sum_{i=1}^{n} w_i \cdot x_{m_i}}{\sum_{i=1}^{n} w_i} \tag{4}$$

ISSN 2616-5775 Vol. 4, Issue 8: 63-66, DOI: 10.25236/AJCIS.2021.040812

Where X_m is the weighted average of the missing attributes corresponding to the neighbors x_m.

Next, the weighted average of the same attribute as the missing value is calculated, and this value X_m is used as the final imputation that estimates the missing value.

$$Imputation = X_m \tag{5}$$

After that, a complete testing set is ready for classification. A series of classifiers γ can be applied, such as SVM, naive bayes, neural network, decision tree and so on.

3. Experimental Studies

3.1. Compared methods

In a series of comparative experiments, the proposed method algorithm with other conventional imputation algorithms including zero imputation (ZI), mean imputation(MI) and k-nearest neighbor imputation(KNNI).

3.2. Data Sets

The 4 different data sets used in the experiment all come from the UC Irvine Machine Learning Repository. To further test the performance of the proposed method algorithm, Iris and Seeds were designed to randomly lose one attribute value and Heart and Vehicle were designed to randomly lose five attribute values.

3.3. Missing value insertion method

The missing values were inserted in terms of MCAR mechanism. To investigate the performance of the algorithms in high level of loss, we removed more than 25% of the values in data sets.

3.4. Comparison with other imputation methods

First of all, after recording the various indicators of the experiment, the arithmetic average of the classifier indicators applied by each method was taken, so that a more comprehensive analysis of the performance of these methods can be figured out. Obviously, Table 1 shows that the proposed method performs much better than ZI and MI no matter which dataset it is. And it shows us great robustness even the data sets have more than 25% missing values, which is superior to KNNI.

Table 1: Result

	Method	Precision	Recall	F_1	ARI
Iris	ZI	67%	64%	62%	29%
	MI	69%	64%	63%	32%
	KNNI	96%	95%	95%	86%
	Proposed	**97%**	**97%**	**97%**	**91%**
Seeds	ZI	60%	54%	53%	13%
	MI	85%	83%	82%	60%
	KNNI	90%	90%	90%	73%
	Proposed	**92%**	**90%**	**90%**	**73%**
Vehicle	ZI	61%	58%	57%	4%
	MI	66%	65%	63%	9%
	KNNI	73%	72%	72%	24%
	Proposed	**82%**	**82%**	**82%**	**44%**
Heart	ZI	34%	33%	31%	3%
	MI	53%	52%	51%	17%
	KNNI	62%	62%	59%	34%
	Proposed	**64%**	**63%**	**62%**	**36%**

4. Conclusion

In this study, we firstly affirmed the importance of dealing with missing values and the nearest

ISSN 2616-5775 Vol. 4, Issue 8: 63-66, DOI: 10.25236/AJCIS.2021.040812

neighbors algorithm has a pivotal significance in this field. Then an efficient method to imputation was proposed to improve the performance of the KNNI. It was compared with the mean imputation, zero imputation and KNNI next. And these methods were applied to 4 different data sets. At the beginning with, the data sets were optimized to split for training excellent classifiers. Then all of imputation methods were used to process the data sets with missing values. For the proposed method, we will find it effective to estimate the missing values through classifying the sample into the most likely class and calculating the imputation in terms of the weighted Euclidean distance between the sample and its nearest neighbors. After obtaining the imputed data sets, a series of comparative experiment were performed. Through the comparison of the four indicators of Precision, Recall, F_1, and ARI, the final results show that the proposed method does perform better, and further optimizes the nearest neighbors algorithm in terms of classification accuracy. As the proposed method is an improvement to the method in the field of single-value imputation, researches will be focused on multi-value imputation method in the future.

References

[1] King-Sun Fu and Rosenfeld. Pattern recognition and image processing. IEEE Transactions on Computers, C-25(12):1336–1346, 1976.

[2] M. Nakai, D. G. Chen, K. Nishimura, and Y. Miyamoto. Comparative study of four methods in missing value imputations under missing completely at random mechanism. Open Journal of Statistics, 04(1):27–37, 2014.

[3] Krishnan Bhaskaran and Liam Smeeth. What is the difference between missing completely at random and missing at random? International Journal of Epidemiology, 43(4):1336–1339, 2014.

[4] Y. Kano and K. Takai. Analysis of nmar missing data without specifying missing-data mechanisms in a linear latent variate model - sciencedirect. Journal of Multivariate Analysis, 102(9):1241–1255, 2011.

[5] P Jönsson and C. Wohlin. An evaluation of k-nearest neighbour imputation using likert data. Proceedings of International Symposium on Software Metrics, pages 108–118, 2004.

[6] U. Shrestha, A. Alsadoon, Pwc Prasad, S. A. Aloussi, and O. H. Alsadoon. Supervised machine learning for early predicting the sepsis patient: modified mean imputation and modified chi-square feature selection. Multimedia Tools and Applications, 80(11):1–24, 2021.

[7] Dongdong Zhao A, Xiaoyi Hu A, Shengwu Xiong A, Jing Tian A, Jianwen Xiang A, Jing Zhou B, and Huanhuan Li C. K-means clustering and knn classification based on negative databases. Applied Soft Computing, 110(1):107732, 2021.

[8] Z. Ma, Z. Liu, Y. Zhang, L. Song, and J. He. Credal transfer learning with multi-estimation for missing data. IEEE Access, pp(8):70316–70328, 2020.

[9] A. Mathur and G. M. Foody. Multiclass and binary svm classification: Implications for training and classification users. IEEE Geoscience and Remote Sensing Letters, 5(2):241–245, 2008.

[10] Cao Feilong and Zhang Yongquan. Neural network interpolation and approximation in distance space. Acta Mathematica, 51(001):91–98, 2008.

[11] Gustavo E. A. P. A. Batista and Maria Carolina Monard. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6):519–533, 2003.

[12] Gustavo Batista and Maria-Carolina Monard. A study of k-nearest neighbour as an imputation method.volume 30, pages 251–260, 01 2002.