ISSN 2618-1584 Vol. 3, Issue 6: 6-10, DOI: 10.25236/FMSR.2021.030602

Classification for the analysis of incomplete medical data

Ximing Ma[1, a], Qi An[2, b]

[1]College of Data Science and Application, Inner Mongolia University of Technology, Huhhot, China
[2]School of Mathematical Sciences, Nanjing Normal Universtiy, Nan Jing, China
[a]2898053657@qq.com, [b]davidknox@foxmail.com

Abstract: *Missing data is common in life. The processing of missing data is the key to classification. Therefore, using the existing reliable data set to complete the missing data is a common and necessary method. These methods have an important impact on dealing with the fuzziness and uncertainty in the data set. Therefore, it is necessary and effective to use accurate data and attribution methods to attribute missing data sets. This paper presents a new missing data classification method. Firstly, the center vector representing each class is calculated by using the training samples. The missing values are then estimated using the center of each class. By comparing the performance of three different interpolation methods in different test data sets, the final results show that the proposed method performs best in general.*

Keywords: *interpolation, missing data, K-nearest neighbors, Decision Tree, Random Forest*

1. Introduction

Missing data has become a common problem in practical applications, especially in the medical field. There are many reasons for the lack of medical research data. For example, patients' test results can not be obtained within the specified time, resulting in the loss of some data[1]. It may also be due to data acquisition equipment failure and other reasons. Once the data is missing, it will affect the accuracy of judgment. Research shows that the more data loss, the lower the performance of classification algorithm[2]. Therefore, classifying these missing data is both important and challenging.

At present, There are also two types of interpolation for missing values: single value interpolation and multi value interpolation[3]. The latter has a huge amount of computation, which makes it a big challenge. For single value interpolation, 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 interpolation, such as zero interpolation (ZI)[4], mean interpolation (MI)[5], k-nearest neighbor interpolation (KNNI)[6].

This paper presents an effective method for missing data classification. Firstly, we use the training samples to calculate the center vector representing each class. Different center classes estimate the missing values. Because the distance from incomplete samples to different class centers is different, we believe that the closer the class center is, the more critical it plays in interpolation. In this paper, we use the weighted average of the class center to estimate the missing value. Then, the incomplete samples after interpolation are classified. This paper classifies and compares the missing data sets with ZI and MI methods.

2. The proposed method

MI is widely used. The mathematical definition of mean is as follows:

$$A_n = \frac{a_1+a_2+\cdots+a_n}{n} \tag{1}$$

It is not difficult to see that the MI needs to use the mean of all data attribute values that are not missing in the data set and the obtained mean interpolate to the missing data. In this method, the sample size remains unchanged, but the variability of data is reduced. Therefore, standard deviation and variance estimates are often underestimated[7].

ISSN 2618-1584 Vol. 3, Issue 6: 6-10, DOI: 10.25236/FMSR.2021.030602

ZI and MI have certain limitations, and there may be large errors. If the interpolation data error is large, the classification effect will be affected. This paper presents a new classification method. The proposed method can effectively avoid the situation that the large difference of attributes in different categories of data set leads to large filling error when using the MI. The realization of the proposed method first requires the mean of each category in the data set, and then calculates the distance from the missing data to the mean of each category. In this paper, the Euclidean distance is used to calculate the distance. Using the obtained distance d_i, according to the formula:

$$F = \frac{\sum_{i=1}^{n} d_i}{d_1} + \frac{\sum_{i=1}^{n} d_i}{d_2} + \cdots \frac{\sum_{i=1}^{n} d_i}{d_n} \tag{2}$$

F is obtained, and the weight of each class is calculated according to F:

$$f_i = \frac{\sum_{i=1}^{n} d_i}{di \times F} \tag{3}$$

According to formula (3), (5), the weighted mean value is obtained:

$$\bar{X} = A_1 \times f_1 + A_2 \times f_2 + \cdots + A_n \times f_n \tag{4}$$

After the interpolation value is calculated by this method, three classifiers are used to classify this part of data. The algorithm flow chart of this method is shown in figure (1).

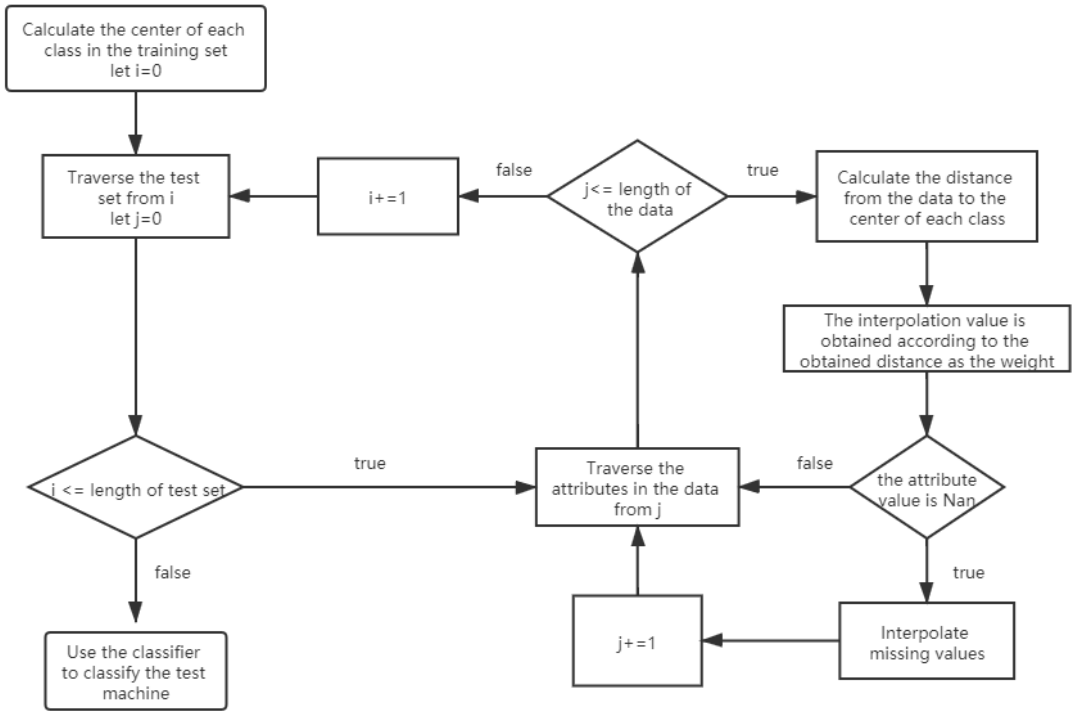

Figure 1: The algorithm flow chart

3. Experimental design

3.1. Datasets

In this paper, three datasets (dataR2, heart_failure_clinical_records_dataset, pd_speech_features) are selected from the UCI machine learning library, the (Prostate_Cancer) dataset is downloaded from kaggle. The performance of ZI, MI, and this paper method is compared. Data sets include number and classification features, whose main features are shown in Table 1.

ISSN 2618-1584 Vol. 3, Issue 6: 6-10, DOI: 10.25236/FMSR.2021.030602

Table 1: The datasets used in our experiments

Dataset	# Instances	#features	#classes
dataR2	116	9	2
heart_failure_clinical_records_dataset	299	12	2
pd_speech_features	756	754	2
Prostate_Cancer	100	8	2

3.2. Missingness mechanism

Data missing mechanisms generally fall into three categories: missing complete at random, missing at random (MAR) and missing not at random[8]. In this paper, missing complete at random is used to test 20 % of the complete data set as a test set and the remaining data as a training set. Each data in the test set is randomly lost greater than or equal to 1 but not completely lost several attribute values to achieve the purpose of missing data.

3.3. Comparison method

The experimental comparison of ZI and MI methods with this paper method. Three methods are used to interpolate missing data and classification. K-nearest neighbors classifier, random forest algorithm and decision tree algorithm are used to classify the processed data. Use Accuracy, precision, Recall and F1-score to compare the classification results.

3.4. Experimental results

K-nearest neighbors[9] classifier was used to conduct five experiments on four datasets using three interpolation methods, and the average value of the experiment was calculated, as shown in figure (2), (3), (4) and (5).

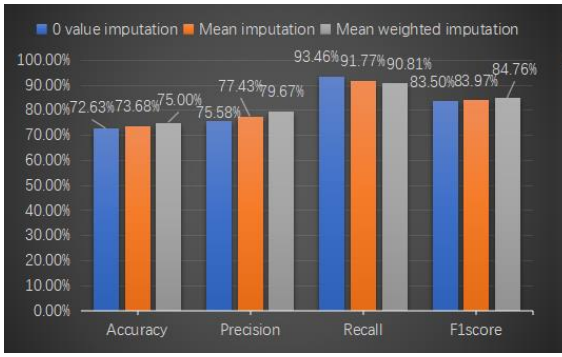

Figure 2: pd_speech_features

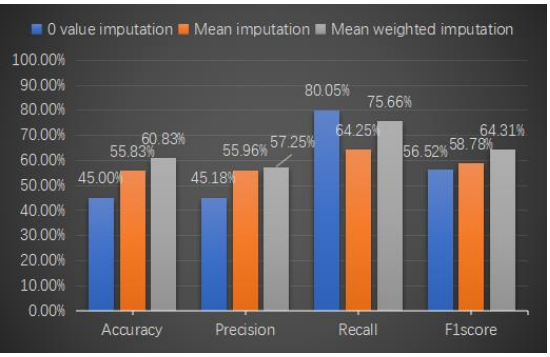

Figure 3: dataR2

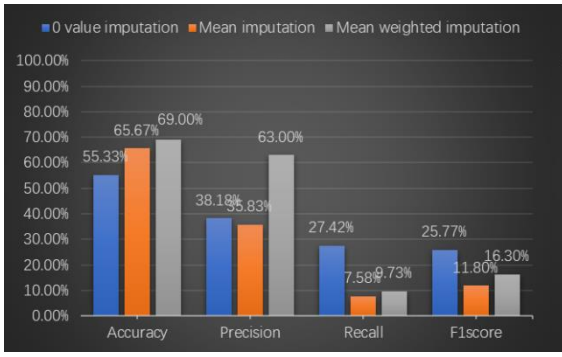

Figure 4: heart_failure_clinical_records_dataset

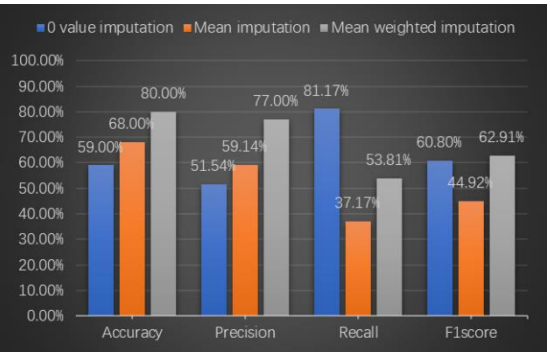

Figure 5: Prostate_Cance

Decision tree[10] classifier was used to conduct five experiments on four datasets using three interpolation methods, and the average value of the experiment was calculated, as shown in Figure (6), (7), (8) and (9).

ISSN 2618-1584 Vol. 3, Issue 6: 6-10, DOI: 10.25236/FMSR.2021.030602

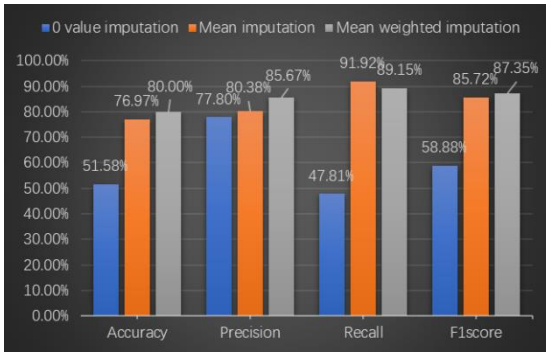

Figure 6: pd_speech_features

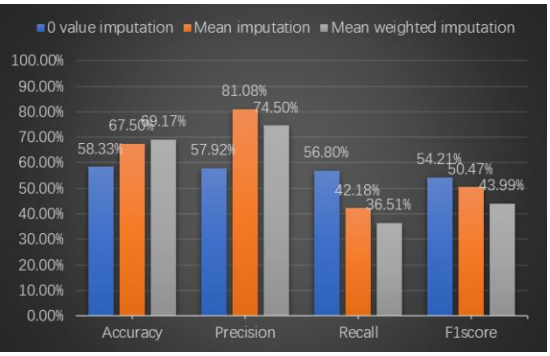

Figure 7: dataR2

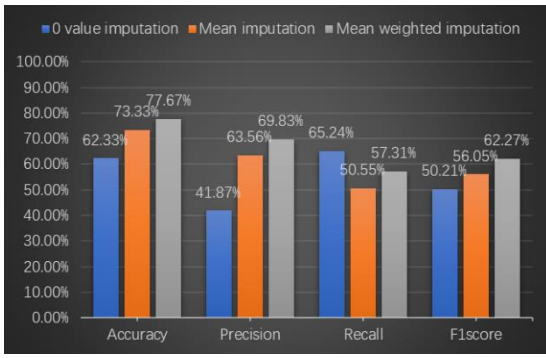

Figure 8: heart_failure_clinical_records_dataset

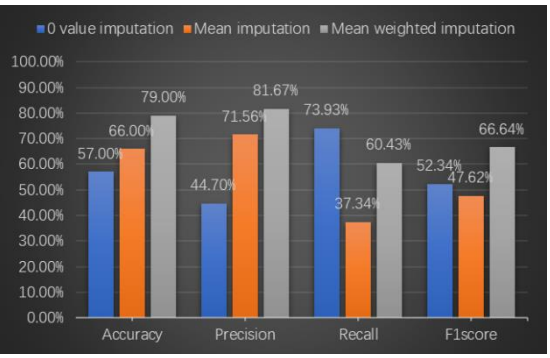

Figure 9: Prostate_Cance

Random forest[11] classifier was used to conduct five experiments on four datasets using three interpolation methods, and the average value of the experiment was calculated, as shown in Figure (10), (11), (12) and (13).

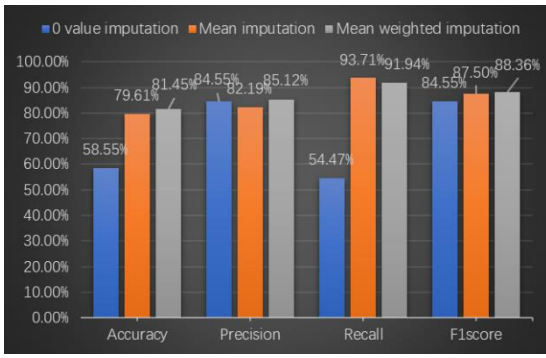

Figure 10: pd_speech_features

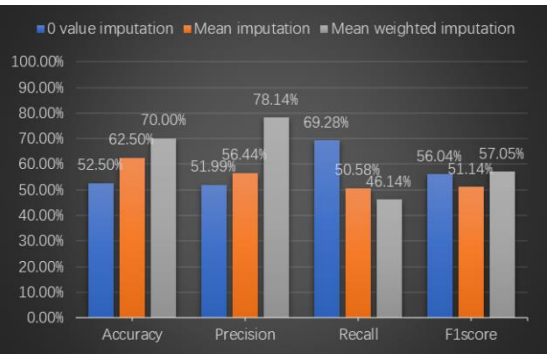

Figure 11: dataR2

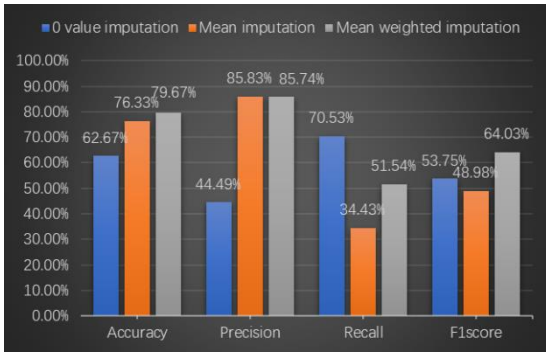

Figure 12: heart_failure_clinical_records_dataset

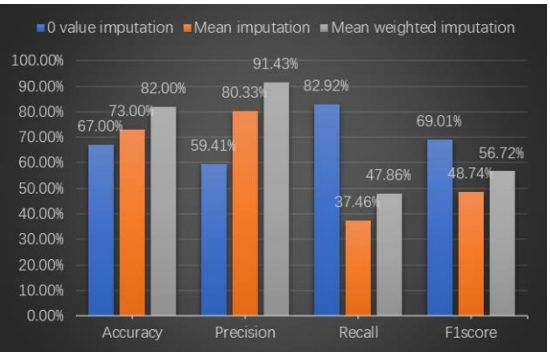

Figure 13: Prostate_Cance

The above results show that when using the same data set and classifier, the accuracy of this method is higher than the other two methods. Compared with K-nearest neighbors algorithm and decision tree algorithm, random forest algorithm has better performance.

ISSN 2618-1584 Vol. 3, Issue 6: 6-10, DOI: 10.25236/FMSR.2021.030602

4. Conclusion

Through the analysis of the experimental results, it can be seen that the accuracy of the proposed method is higher than that of Zi and Mi methods. The accuracy has also achieved good results. Compared with Zi method and MI method, the method proposed in this paper has more advantages.

References

[1] J. Venugopalan, N. Chanani, K. Maher and M. D. Wang, "Novel Data Imputation for Multiple Types of Missing Data in Intensive Care Units," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 3, pp. 1243-1250, May 2019, doi: 10.1109/JBHI.2018.2883606.

[2] E. T. Capariño, A. M. Sison and R. P. Medina, "Application of the Modified Imputation Method to Missing Data to Increase Classification Performance," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019, pp. 134-139, doi: 10.1109/CCOMS.2019. 8821632.

[3] B. Xiang, F. Yan, T. Wu, W. Xia, J. Hu and L. Shen, "An Improved Multiple Imputation Method Based on Chained Equations for Distributed Photovoltaic Systems," 2020 IEEE 6th International Conference on Computer and Communications (ICCC), 2020, pp. 2001-2005, doi: 10.1109/ICCC51575.2020. 9345230.

[4] Ichikawa, M., Hosono, A., Tamai, Y., Watanabe, M., Shibata, K., Tsujimura, S., . . . Suzuki, S. (2019). Handling missing data in an FFQ: Multiple imputation and nutrient intake estimates. Public Health Nutrition, 22(8), 1351-1360. doi:10.1017 /S1368980019000168

[5] Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 2013; 3: e002847. doi: 10.1136/bmjopen-2013-002847

[6] Lin Qiao, Ran Ran, He Wu, Qiaoni Zhou, Sai Liu, and Yunfei Liu. 2018. Imputation Method of Missing Values for Dissolved Gas Analysis Data Based on Iterative KNN and XGBoost. In Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2018). Association for Computing Machinery, New York, NY, USA, Article 11, 1–7. DOI: https: //doi.org/10.1145/3302425.3302447

[7] T. Duy Le, R. Beuran and Y. Tan, "Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare," 2018 10th International Conference on Knowledge and Systems Engineering (KSE), 2018, pp. 247-251, doi: 10. 1109/KSE.2018.8573344.

[8] Zhao Yang. Statistical inference for missing data mechanisms [J]. Statistics in Medicine, 2020, 39(28): 4325-4333.

[9] M. Dixit, R. Sharma, S. Shaikh and K. Muley, "Internet Traffic Detection using Naïve Bayes and K-Nearest Neighbors (KNN) algorithm," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1153-1157, doi: 10.1109/ICCS45141.2019.9065655.

[10] HAN Cheng-cheng, ZENG Si-tao, LIN Qiang, CAO Yong-chun, MAN Zheng-xing. Decision Tree Based Steraming Data Classification Algorithm: A Survey [J]. Journal of Northwest Minzu University (Natural Science), 2020, 41(02): 20-30.

[11] LV Hong-yan, FENG Qian. A review of random forests algorithm [J]. Journal of the Hebei Academy of sciences, 2019, 36(03): 37-41.