[文章编号] 1003－4684(2021)05-0059-05

离散型缺失数据填补法综合比较

袁建裕[1]，闫春艳[1]，叶志伟[1]，杨志勇[2]

(1 湖北工业大学计算机学院，湖北 武汉 430068；2 湖北省公安厅科技信息处，湖北 武汉 430064)

[摘 要] 针对离散型数据填补方法的研究尚不完备的情况，通过改造现有模型，系统地比较和分析了基于众数填补、随机填补、K 最近邻填补、基于自编码器的填补和基于生成对抗网络的填补在离散型数据的填补性能，对在数据预处理阶段选择适合数据集的填补方案具有重要的意义。实验结果显示，不同填补方法的填补结果有较大的差异，进而影响后续分析的准确性。

[关键词] 离散缺失数据；填补方法；方法比较

[中图分类号] TP391.9 [文献标识码] A

据统计，全球各领域的数据总量正在以每年40％的速度快速增长[1-2]。然而由于设备功能限制、设备故障、数据错误或调查中的无回答等因素，数据缺失现象较为普遍，很大程度上影响了数据质量和应用[3]。为解决这一问题，最简单方法就是删除法，即将包含缺失值的数据对象、数据属性、成对变量进行删除。但该方法很难适用不同领域的缺失数据集，且对数据资源造成浪费，无法保证研究结果的客观性和准确性[4]。缺失值填补法是通过现有数据，为缺失值估计一个合理的填补值，从而构造出完整数据集。该方法既保证了数据集的规模，又能对缺失值做出合理的推断，已受到众多科研及从业人员的广泛关注。常用的填补方法大致可分为基于统计学的填补方法和基于机器学习的填补方法[5]。基于统计学的填补方法主要包括均值填补、回归填补、多重填补等，其特点为原理简单、易于实现。基于机器学习的填补方法主要包括 K 最近邻填补法、基于聚类的填补方法、基于神经网络的填补方法等，其优点在于模型通过对属性间关联合理建模，实现了缺失值的有效估计。但上述方法大都适用于连续型数据，并不适用于离散型数据。

本文通过扩展已有模型以适用于离散型数据填补，并对填补结果进行后续分析，系统地比较和分析了众数填补、随机填补、K 最近邻填补、基于自编码器的填补和基于生成对抗网格的填补在离散型数据的填补性能。

1 相关工作

1.1 缺失模式与缺失机制

缺失模式是缺失数据的外在特征，可以按不同的标准进行分类。按照缺失变量个数的不同可分为单变量缺失模式和多变量缺失模式，按缺失结构的不同可分为单调缺失模式和随机模式。缺失机制则是缺失数据的内在特征，Rubin[6] 在其理论研究中把缺失问题归纳为三类，并将数据缺失的概率问题称为数据缺失机制。分别为：

1)完全随机缺失 (Missing Completely at Random，MCAR)，即缺失值完成随机产生，缺失值与已观测数据无关，与未观测数据也无关，并且缺失数据不会使结果产生偏差。

2)随机缺失 (Missing at Random，MAR)，即缺失值与已观测数据存在相关性，缺失值可根据已有数据进行估计。

3)非随机缺失 (Not Missing at Random，NMAR)，即缺失值与未观测数据有关。这一理论规定了填补方法可以提供有效估计的条件。

1.2 离散连续化

离散连续化是特征工程中必不可少的一个环节。在数据挖掘中，许多机器学习模型要求输入变量为连续型，离散型数据需要在预处理阶段转换成连续型，其中，One-hot 编码是用于处理离散数据的常用方法[7]。One-hot 编码有效解决了数据的属性

[收稿日期] 2020－06－06

[基金项目] 福建省大数据管理新技术与知识工程重点实验室开放基金(BD201801)

[第一作者] 袁建裕(1990－)，男，浙江宁海人，湖北工业大学硕士研究生，研究方向为数据挖掘

[通信作者] 叶志伟(1978－)，男，湖北浠水人，工学博士，湖北工业大学教授，研究方向为机器学习和数据挖掘

问题，并在一定程序上扩充了特征空间。以本文使用的数据集 Nursery 中特征“家族情况”为例，其取值集合为 complete，completed，incomplete，foster，相应的 One-hot 编码见表 1。

表 1 One-hot 编码

原始值	One-hot 编码
complete	[0,0,0,1]
completed	[0,0,1,0]
incomplete	[0,1,0,0]
foster	[1,0,0,0]

1.3 基于统计学的填补方法

1)均值填补：均值填补法是统计方法中应用最为广泛的填补方法[8]，根据待研究数据特征中已观测数据的均值或众数作为缺失值。在进行填补时，若缺失值的数据类型为数值型，则使用缺失值所属特征的均值作为缺失值的填补值，若缺失值的数据类型为非数值型，则使用缺失值所属特征的众数作为缺失值的填补值。另外，可以选择不同的统计量作为填补值，如中位数、修正平均数等。

2)随机填补：随机填补法是社会调查领域中较为常见的缺失值处理方法[9]，统计已有观测数据中该特征各值出现的概率，依概率随机选择一个值作为填补值。显然，出现频次多的特征值作为填补值的概率要大于出现频次少的特征值。

1.4 基于机器学习的填补方法

1)K 最近邻填补：K 近邻填补法的核心是从完整样本中选择与缺失样本距离最近的 K 个完整样本，并将完整样本中的已观测值的加权平均值作为填补值[10]。本文研究基于离散型缺失数据的填补方法，故选取基于信息论中的海明距离[11]作为 K 近邻填补法的距离度量函数，用以统计两个等长字符串对应位置字符不同的个数之和，其定义为：

$$HD = \sum_{i=1}^{m} A_i \otimes B_i$$

其中，A 和 B 为参与计算的两个样本；A_i 和 B_i 为样本 A 和 B 的第 i 个特征，取值为 0 或 1；m 为特征空间的维度。距离越小，样本间的相似度越高，所在样本对应的特征值作为填补值的可靠性也越高。

2)基于自编码器的填补：自编码器是多层感知机模型的一种，其特点是输出层与输入层具有相同数量的神经元，其输入即为模型期望的输出[12]。基于自编码器的填补优点在于只需要训练单个网络的权重，填补速度快，具体步骤：

步骤一：根据缺失数据集确定网络结构。

步骤二：将数据集分成完整数据集 D_{com} 和不完整数据集 D_{miss}。

步骤三：将数据集 D_{com} 作为模型的训练集，确定网络的权重。

步骤四：预填补不完整数据集 D_{miss}，依次将 D_{com} 中的样本作为已训练模型的输入，样本的缺失值以模型输出中的缺失值相应位上的预测值作为填补值。

3)基于生成对抗网络的填补：生成对抗网络由两个网络组成，分别为生成网络和判别网络，通过对抗学习的方式来训练模型的适应性[13]。网络的训练始终处于一种对抗博弈的状态。基于生成对抗网络的填补是通过生成网络与判别网络的相互作用来学习数据的分布，从而预测出缺失值的填补值。其中，生成网络用于模拟和预测样本，而判别网络则用于判定模拟样本与真实样本的差异性。基于生成对抗网络的填补过程主要有三个阶段：判别网络训练、生成网络训练和模型填补。

判别网络训练：首先将数据集中的完整样本标记为真样本，生成器利用随机噪声作为输入生成的样本标记为假样本，将完整样本和生成网络生成的样本输入到判别网络中，以此来训练、更新判别网络的权重。

生成网络训练：首先使用众数填补对不完整数据集进行预填补，将预填补后的样本标记为真样本，再将真样本作为生成网络的输入。然后使用第一阶段训练好的判别器，即固定判别网络，根据模型的损失函数来不断更新生成网络的权重。

模型填补：对数据集进行预填补，将填补后的数据输入到第二阶段训练好的生成器中，生成器的输出则为缺失值的填补值。

基于生成对抗网络的填补方法的过程见图 1。

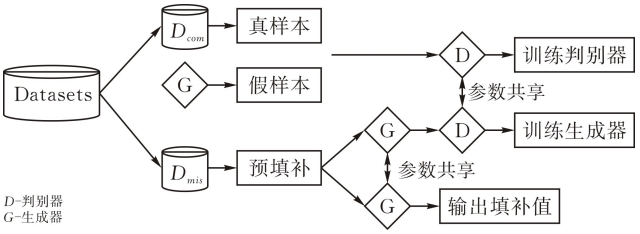

图 1 基于生成对抗网络的填补方法

2 实验设置与结果分析

2.1 实验设置

实验操作平台为 Windows 10 64 位操作系统；处理器为 i5-8265U；运行内存 8G；编程语言为 Python 3.9，主要使用了 NumPy 1.20，Pandas 1.2.4，scikit-learn 0.24，Keras 2.1.1 库。

实验使用的 3 个数据集选自 UCI 机器学习数

据仓库，均为离散型数据集。数据集的详细信息见表 2。

表 2　数据集基本信息

数据集	实例数	特征数	类别数
Balance Scale	625	4	3
Chess	3196	36	2
Nursery	12 960	8	5

2.2　参数设置

本文系统地比较和分析了众数填补、随机填补、K 最近邻填补、基于自编码器的填补和基于生成对抗网络的填补在不同缺失机制、不同缺失比例情况下对离散型数据集进行填补的适用范围和优缺点。其中，缺失比例包括 10%、20%和 30%，缺失机制包括 MCAR、MAR 和 NMAR，缺失模式为多变量随机缺失模式。除了众数填补和随机填补，其它填补方法在填补过程中均使用了 One-hot 编码。为方便表示，各方法在下文图表中分别简称：Mode、Random、KNN、MLP、Encoder 和 GAN。参数设置见表 3。

表 3　参数设置

方法	参数	取值
Mode	None	None
Random	None	None
KNN	k	5
MLP	网络层数	3
	批次个数	1000
	隐层结点数	20
	迭代次数	200
	参数优化器	随机梯度下降
	隐层激活函数	relu(x)
	输出层激活函数	sigmod(x)
Encoder	网络层数	3
	批次个数	200
	隐层结点数	20
	迭代次数	1000
	参数优化器	随机梯度下降
	隐层激活函数	relu(x)
	输出层激活函数	sigmoid(x)
GAN	Dropout	0.2
	relu alpha	0.2
	优化算法	RMSprop
	生成网络层数	5
	判别网络层数	5
	训练次数	100
	批次数	100

2.3　性能指标

本文采用两个性能指标来衡量各填补方法的填补效果，分别为填补准确率和分类准确率。它们的含义如下：

1）填补准确率：所有缺失值的填补值中正确填补的比例。

$$ACC_{impute}=\frac{TP}{TP+TN}$$

2）分类准确率：完整样本组成训练集训练分类模型，本文所使用分类模型为支持向量机（Support Vector Machine，SVM），并采用 10 折交叉验证的方法将填补后样本作为测试集所得到的样本被正确分类的比例。

2.4　结果分析

实验得到的填补准确率和分类准确率的具体数值见表 4，表中加粗项为当前行最优值。接着，从两个不同的维度，即缺失比例和缺失机制，分别对填补准确率和分类准确率取平均值进行可视化（图 2、图 3）。

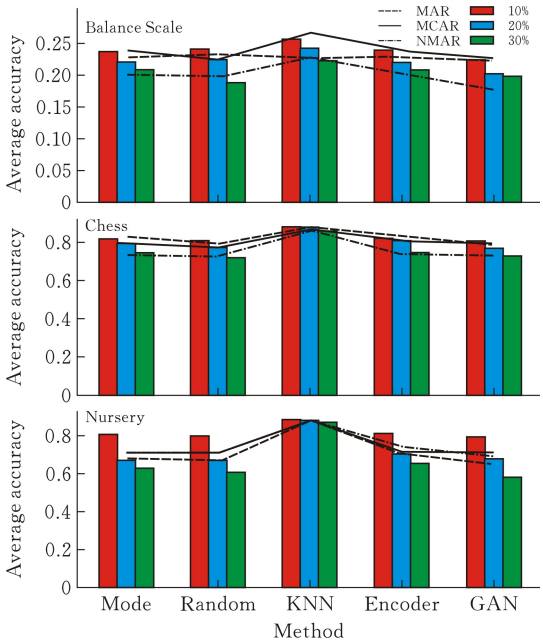

图 2　填补准确率

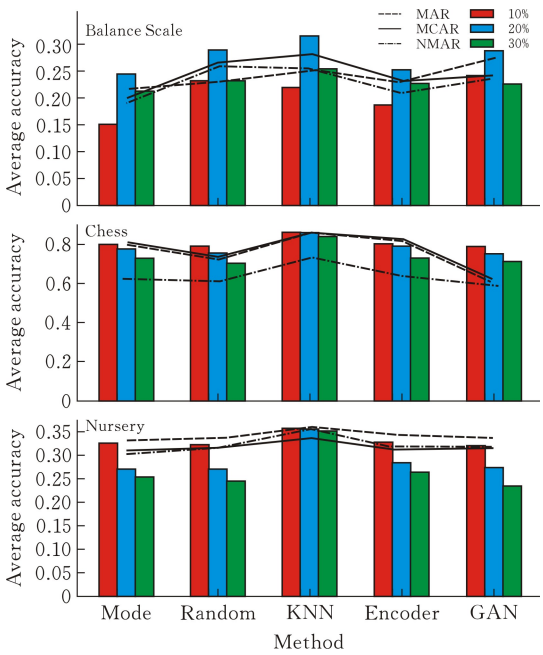

图 3　分类准确率

其中柱形表示在相同缺失比例下填补和分类的平均准确率，折线表示在相同的缺失机制下填补和分类的准确率。

由表 4 知，除了在 Balance Scale 数据集表现得不够稳定，K 最近邻填补在两个性能指标上均要优于其它 4 种填补方法。结合表 2 的数据集信息，不稳定的原因可能有二：其一，数据集实例数较少，找到的 K 个样本与原样本存在较大差异；其二，数据集特征数较少，海明距离不足以度量样本间的相似度。横向对比 Chess 和 Nursery 数据集的特征数，K 最近邻填补在 Chess 数据集上的填补准确率要明显高于在 Nursery 数据集上的填补准确率。所以，K 最近邻填补适合应用于满足一定实例数，特征数较多的数据集的填补工作。从表 5 来看，选择不同的填补方法直接影响着后续分类的结果。基于自编码器的填补和基于生成对抗网络的填补，两者均属于基于神经网络的填补，它们的分类准确率要好于众数填补和随机填补的分类准确率。尽管在填补准确率这一指标上，众数填补和随机填补的准确率部分会好于基于神经网络的填补的准确率，但基于神经网络的填补能够挖掘和提炼数据集所蕴含的潜在信息，所以在分类分析中得到了较高的分类准确率。同样的，选择不同的网络结构会造成分类结果存在差异。由于是处理离散型数据集，图 2 的填补准确率并未如预期随缺失比例的升高而下降，存在不稳定的现象，但填补方法的分类性能随着缺失比例的升高而下降，这也从侧面说明完整样本的数量在分类分析中起到重要的作用。同样的，不同的缺失机制会影响填补方法的性能。总体来看，在 MAR 缺失机制下，各填补方法的性能较好。

表 4 填补和分类的准确率

数据集	缺失机制	缺失比例	填补准确率					分类准确率				
			Mode	Random	KNN	Encoder	GAN	Mode	Random	KNN	Encoder	GAN
Balance Scale	MAR	**10**	0.156	**0.168**	0.16	0.156	0.232	0.271889	**0.276498**	0.262673	0.271889	0.253456
		20	0.314	0.344	**0.396**	0.326	0.356	0.173228	**0.181102**	0.178478	0.173228	0.173228
		30	0.18	0.184	0.2	0.1987	**0.2373**	0.241453	0.237179	**0.24359**	0.241453	0.241453
	MCAR	**10**	0.136	**0.28**	0.26	0.236	0.232	0.256158	0.256158	**0.285714**	0.256158	0.241379
		20	0.224	0.266	**0.304**	0.224	0.286	0.254386	0.25731	**0.283626**	0.254386	0.251462
		30	0.2413	0.2493	**0.2827**	0.24	0.2093	0.206573	0.164319	**0.232394**	0.206573	0.185446
	NMAR	**10**	0.156	0.252	0.244	0.168	**0.26**	0.181818	0.191388	**0.22488**	0.191388	0.181818
		20	0.196	**0.26**	0.246	0.212	0.224	0.236527	0.236527	**0.266467**	0.236527	0.182635
		30	0.2227	0.2653	**0.2773**	0.2453	0.2293	0.180095	0.165877	**0.194313**	0.180095	0.168246
Chess	MAR	**10**	0.7792	0.702	**0.8675**	0.8096	0.6384	0.798111	0.784337	**0.887839**	0.809524	0.807556
		20	0.7815	0.6982	**0.8555**	0.8078	0.5456	0.85921	0.803285	**0.883848**	0.857646	0.763395
		30	0.8337	0.7681	**0.8699**	0.8446	0.6177	0.829879	0.777865	**0.863121**	0.820884	0.777474
	MCAR	**10**	0.8071	0.7327	**0.8718**	0.8349	0.601	0.854373	0.834191	**0.880095**	0.86664	0.822715
		20	0.8147	0.7349	**0.8655**	0.829	0.6259	0.795855	0.77239	**0.859992**	0.802503	0.801721
		30	0.8115	0.7373	**0.857**	0.8216	0.6272	0.752053	0.721549	**0.855299**	0.762221	0.748925
	NMAR	**10**	0.5776	0.5671	**0.7212**	0.6004	0.5531	0.795257	0.798419	**0.868775**	0.797233	0.793676
		20	0.6324	0.6183	**0.7375**	0.6467	0.6224	0.745405	0.728588	**0.865467**	0.751271	0.737583
		30	0.6685	0.6447	**0.7363**	0.674	0.5859	0.653109	0.642159	**0.840438**	0.659366	0.66054
Nursery	MAR	**10**	0.3471	0.3513	0.2782	0.3471	**0.3582**	0.859441	0.866074	**0.890347**	0.859441	0.872001
		20	0.337	0.3441	**0.4037**	0.3691	0.3436	0.62185	0.592532	**0.873676**	0.690361	0.618558
		30	0.3098	0.3112	0.3954	0.3105	0.3102	0.55969	0.528295	**0.858043**	0.561822	0.46124
	MCAR	**10**	0.3078	0.3143	0.3102	**0.3177**	0.3125	0.760236	0.763034	**0.88218**	0.775405	0.779529
		20	0.3061	0.3158	0.34	0.3074	0.3181	0.703214	0.706654	**0.880039**	0.705794	0.727507
		30	0.3086	0.3165	**0.3623**	0.3093	0.3149	0.668609	0.64599	**0.881371**	0.670881	0.624457
	NMAR	**10**	0.3055	0.3125	0.3	0.3062	**0.3155**	0.784753	0.766023	**0.882353**	0.786509	0.724905
		20	0.2858	0.3194	**0.342**	0.3212	0.3161	0.694114	0.705693	**0.876702**	0.71159	0.685322
		30	0.3105	0.3173	**0.3636**	0.329	0.3164	0.656096	0.647725	**0.871282**	0.712626	0.656884

3 结论

本文以离散型缺失数据集作为研究对象，通过构造不同缺失机制、不同缺失比例的缺失数据集，系统地比较和分析了众数填补、随机填补、K 最近邻填补、基于自编码器的填补和基于生成对抗网络的填补方法的性能，结果表明：

1)在不同缺失机制、不同缺失比例的情况下，K 最近邻填补的整体填补效果要优于其它填补方法。

其中，K 近邻则适用于处理特征较多且取值范围广的离散型数据集。在具体实施过程中，针对特定的数据集，选择合适的距离度量函数也至关重要。

2）缺失机制对填补效果的影响较为显著，不同填补方法在不同缺失机制下的填补效果差别较大。

3）从不同评价方式来看，数据特征和缺失机制对填补准确率影响较大，进而影响分类准确率。

综上所述，对于离散型缺失数据集的填补工作，构造一个具有普适性的填补模型是相对困难的，应该保持科学谨慎的态度合理地选择填补模型，并结合实际问题加以分析和应用。

[参 考 文 献]

[1] BIG Data Center Members. Database resources of the BIG data center in 2019[J]. Nucleic Acids Res, 2019, 47(1):8-14.

[2] CARLSON, DAVID, LAWRENCE CARIN. Continuing progress of spike sorting in the era of big data[J]. Current opinion in neurobiology, 2019, 55(4):90-96.

[3] ZHANG, ZHONGHENG. Missing data imputation: focusing on single imputation[J]. Annals of translational medicine, 2016, 4(1):9.

[4] 熊中敏，郭怀宇.缺失数据处理方法研究综述[J/OL].计算机工程与应用:1-13[2021-06-03].http://kns.cnki.net/kcms/detail/11.2127.TP.20210508.1003.004.html.

[5] 陈娟，王献雨，罗玲玲，崔晶晶.缺失值填补效果：机器学习与统计学习的比较[J].统计与决策，2020，36(17):28-32.

[6] RUBIN, DONALD B. Inference and missing data[J]. Biometrika,1976,63(3):581-592.

[7] ALAYA M Z, BUSSY S, S GAÏFFAS, et al. Binarsity: a penalization for one-hot encoded features[J]. Journal of Machine Learning Research, 2017, 20:1-34.

[8] YAGYANATH R. Multivariate imputation for missing data handling a case study on small and large data sets [J]. International Journal of Human Computing Studies, 2020, 2(1):5-11.

[9] 薛洁，吴霞，姚雨萌.我国五大热门城市住房分享发展现状分析——基于爱彼迎中国平台数据[J].杭州电子科技大学学报(社会科学版)，2019，15(3):26-32.

[10] AL-ZOUBI A, TATAS K, KYRIACOU C. Design space exploration of the KNN imputation on FPGA [C]. 2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST). IEEE, 2018:1-4.

[11] TAHERI R, GHAHRAMANI M, JAVIDAN R, et al. Similarity-based Android malware detection using Hamming distance of static binary features[J]. Future Generation Computer Systems, 2020, 105: 230-247.

[12] GU S, KELLY B, XIU D. Autoencoder asset pricing models[J]. Journal of Econometrics, 2021, 222(1): 429-450.

[13] LUO Y, CAI X, ZHANG Y, et al. Multivariate time series imputation with generative adversarial networks [C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 1603-1614.

Comprehensive Comparison of Methods for Imputing Discrete Missing Data

YUAN Jianyu[1], YAN Chunyan[1], YE Zhiwei[1], YANG Zhiyong[2]

(1 *School of Computer Science, Hubei Univ. of Tech., Wuhan* 430068, *China*;
2 *Hubei Provincial Public Security Department, Wuhan* 430064, *China*)

Abstract: In the process of data mining, the widespread problem of missing data would largely affect the data quality and the robustness of analyses, and ultimately lead to biased decision-making. The commonly used imputation methods are mainly targeted at continuous data, most of which are not suitable for discrete data. However, there still lacks comprehensive research on discrete data imputation methods both at home and abroad. To this end, we have systematically estimated the discrete data imputation performance of several methods, including mode-based filling, random filling, K nearest neighbor filling, auto-encoder-based filling, and generative confrontation based filling, by modifying the existing models to fit discrete data. The results indicated that the performances varied largely among different filling methods, which in turn affects the accuracy of subsequent analyses. Therefore, it is crucial to choose suitable imputation scheme according to different data sets during the data preprocessing stage.

Keywords: missing data; data imputation; discrete data; method comparison

[责任编校：张岩芳]