郑智泉, 罗忠琛, 常国艳, 等. 数据填补算法的参数优化问题[J]. 智能计算机与应用,2024,14(10):1-11. DOI:10.20169/j.issn.2095-2163.241001

数据填补算法的参数优化问题

郑智泉[1], 罗忠琛[2], 常国艳[3], 张文勇[1], 郑泽鸿[1]
(1 贵州民族大学 数据科学与信息工程学院, 贵阳 550025; 2 贵州医科大学 护理学院, 贵阳 550025;
3 贵州医科大学附属医院, 贵阳 550025)

摘 要: 具体应用场景下,数据填补算法中参数的设定及优化是否合理将直接影响最终的填补效果,如 K 近邻中, k 值选取过大将导致分类模糊, k 值过小可能会导致分类错误;WKNN 虽然在一定程度上缓解了该问题,但如何为不同缺失样本的 k 个近邻构造合理的权重组合又成为问题的核心。本文基于 Bootstrap 方法改进 K 折交叉验证法以优化 k 值;利用高斯函数进行加权,并使用动态调参法解决权重分配问题,基于 R 平台,在 *MAE*、*RMSE*、*MAPE* 准则下对 3 种算法进行评估,分别使用平台内置数据集 Boston 和 UCI 公开数据集 ILPD,对不同缺失机制、不同缺失率、任意缺失列前提下的非完整数据集进行多次填补。实验结果显示: $k = 5$ 时,K 近邻及其改进算法的初期填补效果达到整体最优;本文使用的动态调参法具备一定的通用性;经过 KFCVB、加权、动态调参改进而来的 IWKNN 填补法在不同应用场景下效果最优。

关键词: 缺失机制; K 近邻; 交叉验证; Bootstrap; 动态调参
中图分类号: O212.1 **文献标志码**: A **文章编号**: 2095-2163(2024)10-0001-11

Parameter optimization of data filling algorithm

ZHENG Zhiquan[1], LUO Zhongchen[2], CHANG Guoyan[3], ZHANG Wenyong[1], ZHENG Zehong[1]
(1 School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China;
2 School of Nursing, Guizhou Meidcal University, Guiyang 550025, China;
3 The Affiliated Hospital of Guizhou Medical University, Guiyang 550025, China)

Abstract: In specific application scenarios, whether the parameter setting and optimization in the data filling algorithm are reasonable will directly affect the final filling effect. For example, in K-Nearest Neighbor, too large k value will lead to fuzzy classification, while too small k value may lead to classification errors. Although WKNN can alleviate this problem to a certain extent, how to construct reasonable weight combinations for k neighbors of different missing samples becomes the core of the problem. In this paper, the K-fold Cross Validation method is improved based on Bootstrap method to optimize the k value. The Gaussian function is used for weighting, and the Dynamic Parameter Adjustment method is used to solve the weight allocation problem. Based on R platform, the three algorithms are evaluated under *MAE*, *RMSE* and *MAPE* criteria, and the platform built-in data set Boston and UCI open data set ILPD are used respectively. The incomplete data sets with different missing mechanism, different missing rates and arbitrary missing columns are filled multiple times. The experimental results show that: when $k = 5$, the initial filling effect of K-Nearest Neighbor algorithm and its improved algorithm reaches the overall optimal; The dynamic parameter adjustment method used in this paper has certain universality; IWKNN filling method improved by KFCVB, weighting and dynamic tuning has the best effect in different application scenarios.

Key words: missing mechanism; K-Nearest Neighbor; cross validation; Bootstrap; dynamic parameter adjustment

0 引 言

数据缺失在统计调查与研究中普遍存在,如经济分析[1]、临床数据[2]、问卷调查等领域,这往往会导致统计推断结果不可靠。数据缺失产生的原因有很多,比如技术上的无法获取[3],数据采集设备发

基金项目: 国家自然科学基金地区科学基金项目(82260646);贵州省教育厅高等学校科学研究项目(黔教技[2022]156 号);贵州省科技支撑计划项目(黔科合基础-ZK[2023]一般 151);贵州民族大学基金科研项目(GZMUZK[2023]QN12)。
作者简介: 郑智泉(1990-),男,硕士,实验师,主要研究方向:统计模型与统计计算, Email:278622839@qq.com; 罗忠琛(1990-),女,博士,讲师,主要研究方向:统计学分析,护理学质性研究; 常国艳(1990-),女,硕士,中级统计师,主要研究方向:统计分析与建模; 张文勇(1992-),男,硕士,实验师,主要研究方向:人工智能,大数据分析; 郑泽鸿(1991-),男,博士,实验师,主要研究方向:数据挖掘与分析。
收稿日期: 2023-06-12

生故障,因隐私问题而造成的单元无回答等[4]。数据缺失不仅掩盖了样本集固有的代表性,也可能导致回归估计量发生严重偏差[5]。

不处理、删除、缺失值填补是数据缺失领域的常见处理方法,如在流行病学的研究中,处理缺失值的方法之一便是删除法,这被称为"完整病例分析"[6]。针对不同缺失场景选用合适的处理方法能有效助力后续工作开展[7]。面对繁冗复杂的缺失情况,1987 年,Little 等学者[8]提出了完全随机缺失(Missing Completely at Random,MCAR)、随机缺失(Missing at Random,MAR)、非随机缺失(Not Missing at Random,NMAR),并在后续的研究中[8]进行了更为详实的阐述说明。数据缺失是客观存在的,基于完整数据集进行的统计分析方法,在面临样本量小或缺失率较大时,会出现统计分析方法不可用,或因大量样本数据被迫丢弃而导致的误差增大[9]。综上,缺失值填补在数据缺失领域的作用愈发重要,缺失值填补按填补值个数可分为单值填补和多值填补两类[10]。其中,单值填补的传统方法包括均值填补、中位数填补、回归填补等。随着机器学习算法在各领域的快速发展和应用,基于分类思想的缺失值填补算法成为了近年来数据缺失领域研究的焦点,如装袋法、K 近邻、决策树、随机森林等。由于单值填补会不同程度地改变原始数据的分布,并掩盖数据的不确定性,Rubin 等学者[11-12]提出多重插补思想来应对此问题,单值填补算法均可使用多重插补思想进行改进,这使得填补算法的应用场景、稳定性均得到了显著提升。

针对实际问题,选用合适的填补算法后,仍需要对算法的内部参数进行设定,为了达到最佳的填补效果,参数优化环节往往必不可少。本文基于 KNN 及其改进算法探讨参数优化问题,并尝试对已有算法进行改进。Fix 等学者于 1951 年首次提出最近邻法对缺失值进行填补,其原理是通过选取没有缺失数据的变量作为辅助变量,利用距离函数求取含缺失值样本点与其他完全观测样本点的距离,进而选择最近邻样本点对应的观测值作为插补值即可[12]。最近邻法可以说是 KNN 中 $k=1$ 时的特例,相较于 KNN,其稳定性差是该算法的显著缺点。基于这种局部相似性的理念,2001 年 Troyanskaya 等学者[13]提出了 K 近邻填补,该方法的核心就是为每一个含缺失值的样本点在完整数据集中寻找 K 个近邻,进而分析并生成最终的填补值[14-15]。此后,KNN 及其改进算法在数据缺失领域发展迅速,而 k 值优化[16]、距离函数选取[17]、加权处理、切片或分层处理[18]均是优化或改进 KNN 的重要手段,本文主要基于 KNN 及其改进算法进行研究,致力于解决算法在缺失值填补过程中的 k 值优化、权重组合动态分配、不同数据集下的填补效果等问题。

1 算法简介及改进方法

1.1 K 近邻算法

K 近邻(K- Nearest Neighbor,KNN)核心思想是根据测试集中已有的样本信息为待分类样本点在训练集中寻找 $k(k \geqslant 1)$ 个近邻,进而依据被选出的 k 个近邻来判断待分类样本点的类别,如何选取近邻和选取几个近邻是 KNN 算法的核心问题。

针对如何选取近邻点的问题,可以通过度量样本点之间的距离来解决。距离度量公式有很多,常用的有欧式距离、马氏距离、曼哈顿距离等,本文选取曼哈顿距离公式进行度量,公式如下:

$$D_{(M_{\alpha}^{\mathrm{mis}},M_{\gamma}^{\mathrm{obs}})}=\sum\left|M_{\alpha,\varphi}^{\mathrm{mis}}-M_{\gamma,\varphi}^{\mathrm{obs}}\right|,$$
$$\alpha\in[1,n_{\mathrm{mis}}],\ \gamma\in[1,n-n_{\mathrm{mis}}],\ \varphi\in[[1,m]-J] \tag{1}$$

令 M 表示待分析样本集,$M_{ij}(i\in[1,n],j\in[1,m])$ 表示 M 中第 i 行第 j 列元素;n 表示样本点数量;m 表示变量总数。式(1)中,$M_{\alpha}^{\mathrm{mis}}$ 表示样本集 M 中含有缺失值的样本点集合(又称测试集)的第 α 个样本点;$M_{\gamma}^{\mathrm{obs}}$ 表示样本集 M 中不含有缺失值的样本点集合(又称训练集)的第 γ 个样本点;$J(J\subseteq[1,m])$ 表示 M 中含有缺失值的观测变量的 j 值集合;φ 表示不包含缺失值的变量对应的 j 值;$D_{(M_{\alpha}^{\mathrm{mis}},M_{\gamma}^{\mathrm{obs}})}$ 表示 $M_{\alpha}^{\mathrm{mis}}$ 和 $M_{\gamma}^{\mathrm{obs}}$ 两个样本点之间的距离大小,值越小表示 2 个样本点越相似。

KNN 具有简单直观的优点,然而,算法需要为测试集中每个 $M_{\alpha}^{\mathrm{mis}}$ 寻找 k 个近邻,每一次计算均需要遍历整个训练集,算法执行时间较长,尤其是训练集样本或测试集样本数量庞大时。此外,KNN 中 k 值过大,可能会导致分类结果模糊,而 k 值过小,可能会导致分类结果出错,且针对不同的 M,k 值最优解往往也不相同。综上,考虑具体问题中 k 值优化问题是 KNN 填补算法的核心。

1.2 交叉验证法及其改进

交叉验证法在机器学习中常被用于检验模型算法的性能,基本思想是对样本进行随机分组,并多次交替进行训练与检验,以得到较为客观的评价算法的性能指标。常见的交叉验证法有留一交叉验证法(LOOCV)和 K 折交叉验证法(K-fold CV)两种[19]。

其中，K-fold CV 主要是将样本随机分成 K 个大小基本一致的组，每次实验过程中选择其中一组作为测试集，剩下的 $K-1$ 组作为训练集，实验需要重复执行 K 次。在实证分析环节，许多学者采用预先指定 1 个或 2 个缺失变量的方式进行实验[20]，这降低了随机模拟的可能性，正因如此，现有的 K-fold CV 能够满足优化 k 值的需要。在现实问题中，样本集中任何观测指标下均有可能出现缺失值，考虑填补算法应具备的适用性特点，本文基于不同的缺失机制，在不同缺失率、任意观测变量含缺失值的情况下进行实证分析，这极大增加了随机模拟的不确定性，采用已有 K-Fold CV 对 k 值进行优化不能完全符合预期，为此，本文基于 Bootstrap 思想对 K-fold CV 进行改进，提出 KFCVB(K-fold CV based on Bootstrap)方法。

采用 KFCVB 对 KNN 中 k 值选取进行优化，其核心思想仍是先对训练集数据进行随机排序[21]；然后对随机排序后的数据进行等额分组，每组样本量可以依据测试集样本数量与训练集样本数量的比值来确定；进一步地，在计算过程中设定 k 的取值范围，使 k 按既定规则逐步增大，然后基于具体的 k 值，采用有放回的抽样方式随机选取一组子样本作为测试集数据，将其余组子样本作为训练集数据进行计算，并对所得评价结果求平均值；最后通过最优结果来确定 k 的取值。具体实现步骤如下：

(1)将训练集样本进行随机排序得到 M_o，并将 M_o 划分为 z 个规模近似的子样本集 $M_o^1, M_o^2, M_o^3, \cdots, M_o^z$，其中 $z = g\left(\frac{1}{p}\right)$，这里 p 为缺失率，$g(\cdot)$ 为向下取整函数；

(2)随机选取 $M_o^\beta(\beta \in (1,z))$ 作为测试集样本，当前 M_o^β 中样本点数量为 n_k，M_o 中的其余子样本集作为训练集样本。这里，β 的值由如下公式随机产生：

$$\beta = f(1:z,1) \tag{2}$$

其中，$f(\cdot)$ 表示等概率抽样函数。X_J 表示 M_o^β 中含有缺失值的观测变量的集合，$J(J \subseteq [1,m])$ 表示 X_J 中列号集合，考虑到本文实证分析过程是基于任意缺失变量的缘故，为不失一般性，单次程序运行的初始缺失列集合 J 由如下公式随机生成：

$$J = f\left(1:m, f\left(1:g\left(\frac{m}{2}\right),1\right)\right) \tag{3}$$

此外，程序运行中 M_o^β 的每一个样本点真实缺失列集合 $\tau(\tau \subseteq J)$ 由如下随机生成：

$$\tau = f(1:h(J), f(1:h(J),1)) \tag{4}$$

其中，$h(\cdot)$ 表示计算向量长度函数。

(3)此时，M_o^β 中样本点数量为 n_k，且每一个样本点对应的缺失列为 τ，将 $M_{o(\alpha,\tau)}^\beta(\alpha \in [1,n_k])$ 的对应观测值按行保存为 $y_{(\alpha,\tau)}$，此后对其做删除处理，采用预设的 k 值作为当前 KNN 的参数对 M_o^β 中的缺失值进行填补，得到填补值 $\hat{y}_{(\alpha,\tau)}$，其中 $M_{o(\alpha,\tau)}^\beta$ 表示 M_o^β 中第 α 个样本点的第 τ 列值的集合；

(4)研究令 $MAE_{CV} = \{MAE_{CV}^1, MAE_{CV}^2, \cdots, MAE_{CV}^\lambda, \cdots, MAE_{CV}^N\}$，其中 MAE_{CV}^λ 表示第 λ 次运算得到的均方误差，$MAE_{CV}^\lambda = \frac{1}{\Phi}\sum |\hat{y}_{(\alpha,\tau)} - y_{(\alpha,\tau)}|, \lambda \in (1,N)$，这里，$N$ 表示交叉验证中程序执行总次数，Φ 表示 M_o^β 中观测值缺失总数；

(5)令 $\overline{MAE_{CV}} = mean(MAE_{CV})$，$MAE\_SD_{CV} = sd(MAE_{CV})$，其中 $mean(\cdot)$ 表示求均值函数，$sd(\cdot)$ 表示求标准差函数；

(6)本文将 k 的取值设为 1 到 30，步长为 1，重复执行第(2)步至第(5)步 N 次，寻找最小的 $\overline{MAE_{CV}}, MAE\_SD_{CV}$ 即可。

在交叉验证环节，本文采用 R 语言内置公开数据集 Boston 进行实验，为避免原始数据量纲所带来的影响，本文所有实验均对原始数据集进行标准化处理。首先采用 K-fold CV 对不同缺失率下的 k 值进行优化选取，实验结果如图 1 所示。

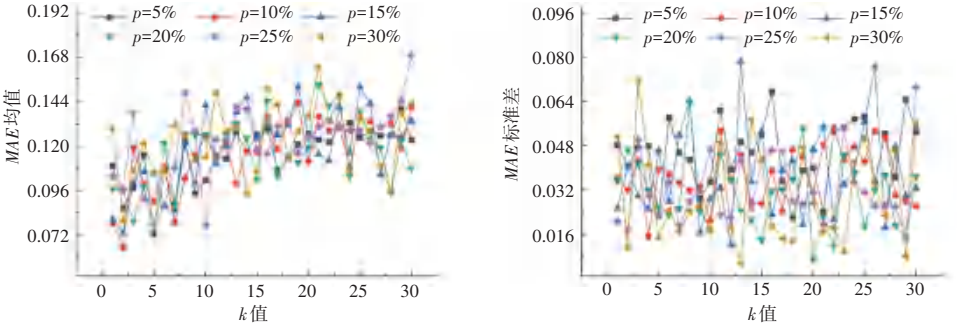


图 1　不同缺失率下 K-fold CV 所得最优 k 值结果示意图

Fig. 1　Schematic diagram of optimal k value obtained by K-fold CV under different deletion rates

图 1 显示,在任意缺失率、任意缺失变量前提下,传统的 K-fold CV 方法并不能为 KNN 中的 k 值提供最优解。这是由于计算机模拟缺失数据集的不确定性增加而导致的,基于此,采用本文提出的 KFCVB 方法继续进行对比实验分析,实验结果如图 2 所示。考虑到不同缺失率对结果产生的影响,在 KFCVB 方法中,本文预设缺失率 p 值为 0.05、0.10、0.15、0.20、0.25、0.30,且 N = 2 000,实验过程均在 MCAR 下进行。

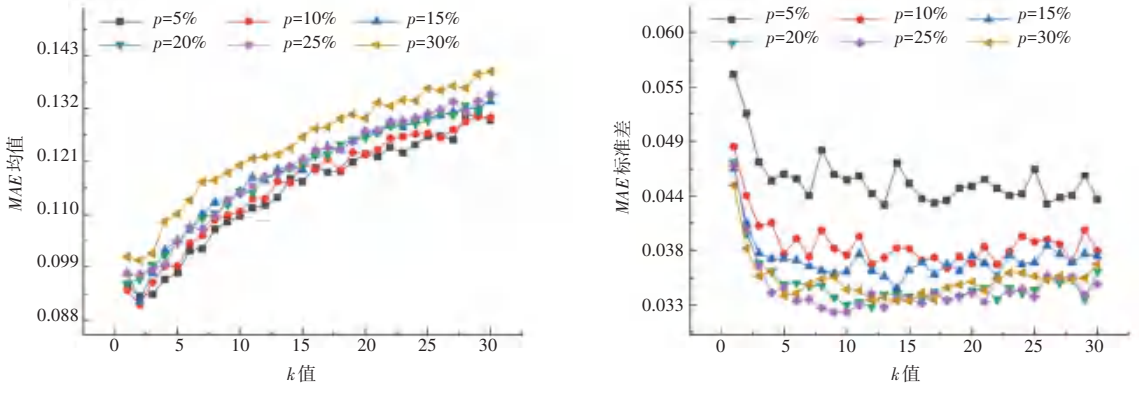


图 2 不同缺失率下 KFCVB 所得最优 k 值结果示意图

Fig. 2 Schematic diagram of optimal k value obtained by KFCVB under different deletion rates

图 2 显示,1)使用 KFCVB 方法对 k 进行优化,在不同缺失率下无法得到固定的最优 k 解,也印证了 k 值会因为具体的样本集、缺失变量、缺失率而发生改变;2)从多次模拟所得的 *MAE* 均值来看,当 $k \leqslant 5$ 时,不同缺失率下的评价结果均值维持在相对较低的水平;k = 2 时 *MAE* 均值最小;结合 *MAE* 标准差进行分析,当 $k \leqslant 5$ 时出现逐步递减的趋势,并在 k = 5 达到该范围的最低水平。综上所述,在后续的实证分析中,本文令 k = 5。

1.3 加权 k 近邻及其改进

WKNN 核心思想是为最邻近待分类样本点的近邻赋予更高权重,通过如下权重函数[21]为每个邻近点求得对应的权重值:

$$\begin{cases} w(D_l) = \dfrac{e^{-(D_l)^2/2\delta^2}}{\sum_{l=1}^{k} e^{-(D_l)^2/2\delta^2}},\ l \in [1,k] \\ \sum_{l=1}^{k} w(D_l) = 1 \end{cases} \tag{5}$$

其中,δ 表示实数常数;D_l 表示第 l 个近邻与待分类样本点之间的距离;k 表示选取的近邻数量。在实际问题中,为每个待分析样本点的 k 个近邻构造合理的权重组合至关重要,这将影响算法最终的填补效果。依据不同的距离 $D_l(l \in [1,k])$ 计算出合理的 $w(D_l)(l \in [1,k])$ 才是算法要解决的核心问题。若固定参数 δ,则不同的 $D_l(l \in [1,k])$ 会对 $w(D_l)(l \in [1,k])$ 的结果产生很大影响。因此,在 WKNN 中,$w(\cdot)$ 要为每一个近邻依据其对应的距离分配合理的权重值,并避免为最近邻的点分配无限接近于 1 的权重值,或为相对较远的一个或几个邻近点分配无限接近于 0 的权重值。若采用固定参数 δ 的方式对实际问题进行求解,必然会出现在 KNN 中 k 值选取不合理所引发的类似问题。因此,本文提出不固定参数 δ,依据实际情况进行动态调节参数 δ 的方式进行问题求解。

1.4 动态调参

对于每一个待填补的样本点,使用 KNN 为其找到 k 个近邻 $M_l^{\text{obs}}(l \in [1,k])$,且每个近邻对应的距离为 $D_l, D_1 \leqslant D_2 \leqslant \cdots \leqslant D_k$。通过式(5)计算对应的权重值 $w(D_l)(l \in [1,k])$,且 $w(D_1) \geqslant w(D_2) \geqslant \cdots \geqslant w(D_k)$。详细的动态调节参数 δ 具体如下。

对同一待分类样本所选出的第 1 个近邻点的权重 $w(D_1)$ 和第 k 个近邻点的权重 $w(D_k)$ 做如下限定,令其满足如下不等式:

$$\frac{w(D_1)}{w(D_k)} \geqslant \xi \tag{6}$$

将式(5)带入不等式(6)可得:

$$\frac{e^{-(D_1)^2/2\delta^2}}{\sum_{l=1}^{k} e^{-(D_l)^2/2\delta^2}} \Big/ \frac{e^{-(D_k)^2/2\delta^2}}{\sum_{l=1}^{k} e^{-(D_l)^2/2\delta^2}} \geqslant \xi, l \in [1,k] \tag{7}$$

对上式进行化简:

$$\frac{e^{-(D_1)^2/2\delta^2}}{e^{-(D_k)^2/2\delta^2}} \geqslant \xi, l \in [1,k] \tag{8}$$

对上式中的参数 δ 进行求解并整理可得如下不等式:

$$\delta \leqslant \sqrt{\frac{(D_k)^2 - (D_1)^2}{2\ln\xi}} \tag{9}$$

利用不等式(9)，通过控制 ξ 值的大小来对加权函数的参数 δ 进行动态调整，以利于适应具体问题中因观测值不同所引起的权重配置不合理的情况。不等式(7)中，ξ 值越大，表明第 1 个近邻点在对缺失值进行填补时所占比重越大；反之 ξ 值越小表明第 1 个近邻点在对缺失值进行填补时所占比重相对越小，且 $w(D_1)$ 和 $w(D_k)$ 的值满足如下不等式：

$$\frac{1}{k} < w(D_1) < 1,\ 0 < w(D_k) < \frac{1}{k} \tag{10}$$

在真实的问题求解过程中，当 $w(D_1)$ 值越接近于 1，此时的 KNNW 填补效果越接近于 $k = 1$ 时的 KNN；当 $w(D_k)$ 值越接近于 $\frac{1}{k}$，此时的 KNNW 填补效果将逐步退化至 KNN。因此，为选定的 k 个邻近点分配合适的权重组合是整个动态调参的核心问题，即 $w(D_1)$ 值不要过大，$w(D_k)$ 值不要过小。

本文基于标准化后的 Boston 数据集，预设 $p = 0.1$，采用前文所述的理论基础，令 $\xi = 1\ 000$，以此对 δ 值进行赋初值，通过遍历的方式寻找令 $w(D_k) \geqslant 0.10, 0.01, 0.001$ 以及 $\xi = 1.1$ 时的 δ 值点，并将上述 δ 值作为权重函数参数，进而采用 KNN、WKNN、IWKNN 进行数据填补，考虑到计算机模拟的随机性，将不同参数下的程序随机执行 1 000 次，并在 *MAE*、*RMSE*、*MAPE* 下对实验结果进行对比分析，结果见表 1。

表 1　不同的权重配比下 3 种填补算法在 3 种评价准则下填补 1 000 次误差结果的均值

Table 1　Under different weight ratio, the mean of three filling algorithms filling 1 000 error results under three evaluation criteria

评价准则	填补方法	ξ	$w(D_k)$		
		1.1	0.1	0.01	0.001
$\overline{MAE}$	KNN	0.099	0.100	0.099	0.100
	WKNN	0.095	0.096	0.095	0.096
	IWKNN	0.098	0.092	0.086	0.089
$\overline{RMSE}$	KNN	0.204	0.210	0.202	0.211
	WKNN	0.198	0.204	0.196	0.206
	IWKNN	0.203	0.201	0.198	0.212
$\overline{MAPE}$	KNN	0.389	0.406	0.404	0.481
	WKNN	0.376	0.393	0.390	0.469
	IWKNN	0.388	0.381	0.357	0.441

表 1 的实验结果显示，第一，当 $\xi = 1.1$ 时，IWKNN 的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 值略微小于 KNN，这表明，在 $w(D_1)$ 和 $w(D_k)$ 的比值逐步接近于 1 时，IWKNN 将逐步退化为 KNN；第二，基于当前实验数据，随着 $w(D_k)$ 值的减小，IWKNN 在 3 种评价准则下的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 值也逐步减小，并在 $w(D_k) = 0.01$ 时达到最佳。针对 $w(D_k)$ 取值不同所带来的算法填补稳定性见表 2。

表 2　不同的权重配比下 3 种填补算法在 3 种评价准则下填补 1 000 次误差结果的方差

Table 2　Under different weight ratio, the variance of three filling algorithms filling 1 000 error results under three evaluation criteria

评价准则	填补方法	ξ	$w(D_k)$		
		1.1	0.1	0.01	0.001
MAE_SD	KNN	0.039	0.040	0.038	0.041
	WKNN	0.039	0.041	0.038	0.041
	IWKNN	0.039	0.041	0.043	0.045
RMSE_SD	KNN	0.138	0.154	0.135	0.145
	WKNN	0.140	0.156	0.137	0.147
	IWKNN	0.138	0.158	0.150	0.158
MAPE_SD	KNN	0.927	1.128	1.191	1.427
	WKNN	0.926	1.125	1.181	1.427
	IWKNN	0.929	1.137	1.254	1.517

表 2 的实验结果显示，第一，当 $\xi = 1.1$ 时，依然出现 IWKNN 与 KNN 结果近似的情形，这加强了表 1 第一点实验结果；第二，随着 $w(D_k)$ 值的减小，IWKNN 的 *MAE_SD*、*RMSE_SD*、*MAPE_SD* 值出现整体变大的趋势，但在 $w(D_k) = 0.01$ 时，IWKNN 的 *RMSE_SD* 值出现下降，对比此时 KNN、IWKNN 的 *RMSE_SD* 结果发现，这种趋势并未改变，这是由于 IWKNN 算法随着 $w(D_k)$ 值的减小而产生的过渡依赖最邻近样本点所带来的弊端。综上论述可知，结合表 1 的实验结果，在本文后续的实证分析环节，将 KNN、WKNN、IWKNN 的参数设定为 $k = 5$，每轮程序运行过程中通过 $\xi = 1\ 000$ 来对 δ 赋初值，并通过 $w(D_k) = 0.01$ 来寻求 δ 的最优解。

2　实验方法

本文实验运行系统环境为：Windows11，软件版本为 R Studio 2023.03.0，R 软件内核为 R4.2.3。为考虑算法填补效果的一般性，本文实验采用 R 内置数据集 Boston 和 UCI 上公开数据集 ILPD。在

MCAR 机制下,本文采用的具体方法是将数据集中的样本进行编号,依据事先设定好的缺失率在 R 平台上生成 n_{mis} 个随机数,进而从原始样本集中取出含缺失值的样本点集合 M^{mis},使用式(3)、式(4)为 M^{mis} 中每一个样本点随机生成具体的缺失列 τ,将 $M^{\text{mis}}_{(\alpha,\tau)}$ 中对应的观测值置为“NA”即可。然而在 MAR 和 NMAR 机制下,计算机模拟数据缺失的过程略有不同。以 MAR 为例,随机缺失表示观测值是否缺失与该样本中已观测到的数据有关,因此,在每次程序运行中,本文采用如下公式生成各缺失变量的参照列 j_R:

$$j_R = f(1:m,1) \tag{11}$$

将 M 中 j_R 列对应的观测值进行升序排列,依照缺失比例的不同对其标注分位点,并将数据划分为若干段,再通过随机数生成法进行数据段的随机选择,进而遍历当前数据段中的所有值,记录该值在原始数据集 M 中的行号,最后从原始样本集中选取出含缺失值的样本点集合 M^{mis}。此时,使用式(3)生成初始缺失列集合 J 时将 j_R 列排除在外,然后使用式(4)为 M^{mis} 中每一个样本点随机生成具体的缺失列 τ,进而将 $M^{\text{mis}}_{(\alpha,\tau)}$ 中对应的观测值置为“NA”即可。模拟 NMAR 机制的情况与 MAR 类似,只需将每一个缺失变量的参考列变更为自身即可,其余步骤不变。

为了验证算法的有效性,基于 3 种缺失机制,在 MAE、$RMSE$、$MAPE$ 准则下对 KNN、WKNN、IWKNN 算法进行评估,本文设定缺失率 p 由 5%逐步递增至 30%,步长设定为 5%,程序执行次数 $N = 1\,000$。

需要说明的是进行多次填补之后会得到不同的多组不同的评价结果,即:

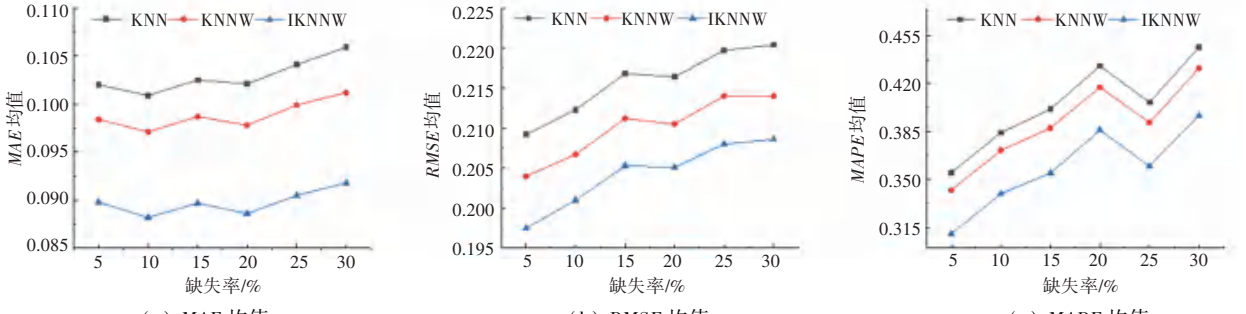
$$MAE = \{MAE_1, MAE_2, \cdots, MAE_\lambda, \cdots, MAE_N\}$$
$$RMSE = \{RMSE_1, RMSE_2, \cdots, RMSE_\lambda, \cdots, RMSE_N\}$$
$$MAPE = \{MAPE_1, MAPE_2, \cdots, MAPE_\lambda, \cdots, MAPE_N\}$$

进而,令 $\overline{MAE} = \frac{1}{N}\sum_{\lambda=1}^{N} MAE_\lambda$,$\overline{RMSE} = \frac{1}{N}\sum_{\lambda=1}^{N} RMSE_\lambda$,$\overline{MAPE} = \frac{1}{N}\sum_{\lambda=1}^{N} MAPE_\lambda$。其中,$N$ 表示实验总次数;λ 表示第 λ 次实验;MAE_λ、$RMSE_\lambda$、$MAPE_\lambda$ 分别表示第 λ 次实验得到的 MAE、$RMSE$、$MAPE$ 评价结果的值。

3 实证分析

3.1 完全随机缺失

在 MCAR 下,KNN、WKNN、IWKNN 基于不同缺失率、不同评价准则所得 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果如图 3 所示。



(a) MAE 均值 (b) $RMSE$ 均值 (c) $MAPE$ 均值

图 3 基于 Boston 数据集,不同缺失率下、3 种填补算法在 MCAR 前提下填补 1 000 次误差结果均值

Fig. 3 Based on the Boston dataset, under different missing rates, the mean of three filling algorithms filling 1 000 error results under the premise of MCAR

图 3 显示,第一,随着缺失率的增大,KNN、WKNN、IWKNN 基于不同评价准则所得 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果均出现增大趋势,这表示随着缺失率的增大,算法的填补效果均出现不同程度的下降;第二,在 MAE、$RMSE$、$MAPE$ 准则下,本文提出的 IWKNN 显著优于其他 2 种填补方法。为了验证不同算法填补的总体效果,采用文献[21]中的公式为 3 种填补算法构造出不同缺失率前提下,基于不同评价准则的 95%置信区间,实验结果见表 3。

表 3 的实验结果显示,第一,在 MAE、$MAPE$ 准则下,IWKNN 在不同缺失率下的置信区间起点最小,而置信区间的终点也基本达到最小,仅在 $MAPE$ 准则下缺失率 $p = 0.05, 0.10$ 时例外;第二,IWKNN 的置信区间长度在不同评价准则、不同缺失率下均不小于其他 2 种算法;第三,在 $RMSE$ 评价准则下,IWKNN 的置信区间起点和终点均大于 WKNN。综上表明,IWKNN 虽然在一定程度上提升了填补效果,但过于依赖最邻近样本点而导致的算法稳定性下降问题凸显。

表 3　不同缺失率下、3 种填补算法在 MCAR 前提下填补 1 000 次误差结果的置信区间及其长度

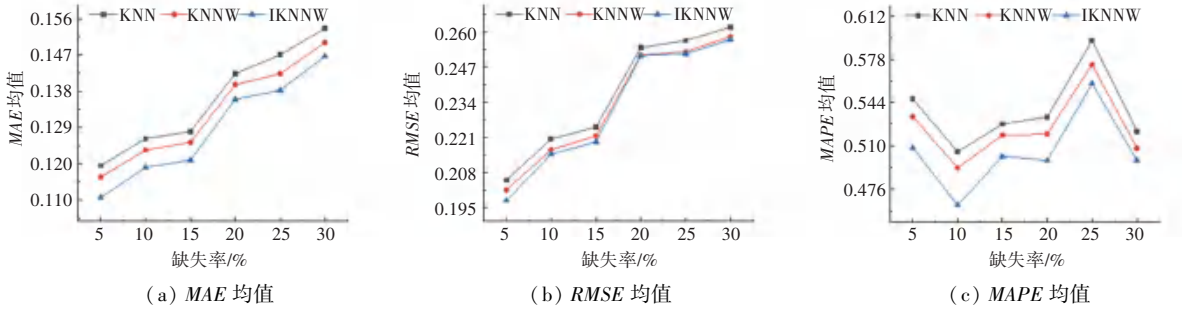
Table 3　Under different missing rates, the confidence interval and length of three filling algorithms filling 1 000 error results under the premise of MCAR

评价准则	填补算法	结果类型	Boston 数据集所含缺失值总体占比(缺失率)/%					
			5	10	15	20	25	30
MAE	KNN	置信区间	[0.096,0.102]	[0.100,0.105]	[0.100,0.105]	[0.099,0.103]	[0.100,0.105]	[0.104,0.109]
		区间长度	0.005	0.005	0.005	0.004	0.005	0.004
	WKNN	置信区间	[0.093,0.098]	[0.096,0.101]	[0.096,0.101]	[0.095,0.099]	[0.096,0.100]	[0.100,0.104]
		区间长度	0.005	0.005	0.005	0.004	0.005	0.004
	IWKNN	置信区间	[0.085,0.091]	[0.088,0.094]	[0.088,0.093]	[0.086,0.091]	[0.087,0.092]	[0.091,0.096]
		区间长度	0.006	0.006	0.005	0.005	0.005	0.005
RMSE	KNN	置信区间	[0.189,0.207]	[0.204,0.222]	[0.206,0.223]	[0.209,0.225]	[0.207,0.223]	[0.219,0.234]
		区间长度	0.018	0.018	0.017	0.017	0.015	0.015
	WKNN	置信区间	[0.184,0.202]	[0.198,0.217]	[0.200,0.218]	[0.203,0.220]	[0.201,0.217]	[0.213,0.228]
		区间长度	0.018	0.018	0.017	0.017	0.016	0.016
	IWKNN	置信区间	[0.186,0.206]	[0.202,0.222]	[0.203,0.222]	[0.204,0.222]	[0.203,0.219]	[0.215,0.233]
		区间长度	0.020	0.020	0.019	0.018	0.017	0.017
MAPE	KNN	置信区间	[0.306,0.514]	[0.383,0.573]	[0.339,0.450]	[0.362,0.463]	[0.381,0.476]	[0.371,0.449]
		区间长度	0.208	0.190	0.110	0.100	0.096	0.079
	WKNN	置信区间	[0.295,0.501]	[0.370,0.560]	[0.327,0.437]	[0.349,0.449]	[0.365,0.460]	[0.355,0.434]
		区间长度	0.206	0.190	0.110	0.100	0.095	0.078
	IWKNN	置信区间	[0.272,0.507]	[0.350,0.565]	[0.299,0.424]	[0.311,0.420]	[0.334,0.438]	[0.321,0.404]
		区间长度	0.236	0.214	0.124	0.109	0.104	0.083

3.2　随机缺失

在 MAR 下，采用相同的实验方法对 3 种填补算法进行评估，实验结果如图 4 所示。

图 4 显示，第一，随着缺失率的增大，KNN、WKNN、IWKNN 基于不同评价准则所得 $\overline{MAE}$、$\overline{RMSE}$ 结果仍出现增大趋势，这与图 3 第一点结论一致，而 $\overline{MAPE}$ 结果始终保持在高位，这表明在 MAR 前提下，随着缺失率的增大，填补算法同样面临失效；第二，在 *MAE*、*RMSE*、*MAPE* 准则下，本文提出的 IWKNN 仍优于其他 2 种填补方法，但优势不再显著，在相同缺失率下对比图 3 的 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果发现，在 MAR 下 3 种算法的实验结果均出现不同程度的增加，这表明 MAR 下，3 种算法的填补效果均受到了影响。同样构造 3 种填补算法在不同缺失率前提下，基于不同评价准则的 95% 置信区间，实验结果见表 4。



(a) *MAE* 均值　(b) *RMSE* 均值　(c) *MAPE* 均值

图 4　基于 Boston 数据集，不同缺失率下、3 种填补算法在 MAR 前提下填补 1 000 次误差结果均值

Fig. 4　Based on the Boston dataset, under different missing rates, the mean of three filling algorithms filling 1 000 error results under the premise of MAR

表 4 不同缺失率下、3 种填补算法在 MAR 前提下填补 1 000 次误差结果的置信区间及其长度

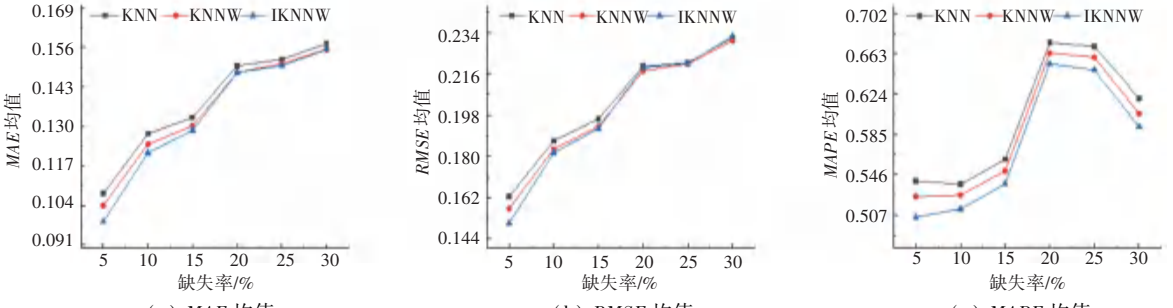
Table 4 Under different missing rates, The confidence interval and length of three filling algorithms filling 1 000 error results under the premise of MAR

评价准则	填补算法	结果类型	Boston 数据集所含缺失值总体占比(缺失率)					
			5%	10%	15%	20%	25%	30%
MAE	KNN	置信区间	[0.107,0.117]	[0.118,0.128]	[0.127,0.138]	[0.133,0.143]	[0.141,0.152]	[0.146,0.159]
		区间长度	0.010	0.010	0.010	0.009	0.011	0.013
	KNNW	置信区间	[0.104,0.114]	[0.116,0.126]	[0.124,0.135]	[0.130,0.140]	[0.136,0.148]	[0.143,0.156]
		区间长度	0.010	0.010	0.010	0.010	0.011	0.013
	IKNNW	置信区间	[0.100,0.111]	[0.114,0.124]	[0.121,0.132]	[0.127,0.137]	[0.133,0.145]	[0.140,0.154]
		区间长度	0.011	0.010	0.011	0.010	0.011	0.014
RMSE	KNN	置信区间	[0.180,0.199]	[0.209,0.230]	[0.222,0.245]	[0.239,0.261]	[0.241,0.264]	[0.253,0.278]
		区间长度	0.019	0.022	0.023	0.022	0.022	0.025
	KNNW	置信区间	[0.176,0.196]	[0.206,0.228]	[0.218,0.242]	[0.235,0.257]	[0.238,0.260]	[0.250,0.275]
		区间长度	0.020	0.022	0.023	0.022	0.023	0.025
	IKNNW	置信区间	[0.179,0.201]	[0.213,0.236]	[0.222,0.246]	[0.241,0.265]	[0.245,0.269]	[0.256,0.281]
		区间长度	0.022	0.023	0.024	0.024	0.024	0.025
MAPE	KNN	置信区间	[0.344,0.616]	[0.450,0.686]	[0.446,0.681]	[0.463,0.623]	[0.493,0.655]	[0.495,0.618]
		区间长度	0.273	0.236	0.235	0.161	0.162	0.123
	KNNW	置信区间	[0.335,0.608]	[0.439,0.673]	[0.433,0.669]	[0.449,0.609]	[0.471,0.634]	[0.483,0.606]
		区间长度	0.272	0.234	0.236	0.160	0.163	0.124
	IKNNW	置信区间	[0.315,0.586]	[0.401,0.611]	[0.413,0.653]	[0.416,0.567]	[0.452,0.616]	[0.467,0.593]
		区间长度	0.271	0.210	0.240	0.151	0.164	0.127

表 4 显示,第一,随着缺失率的增大,3 种算法在不同评价准则下的置信区间起点、终点均出现增大趋势,这印证了随着缺失率的增加,填补算法面临失效的问题;第二,在 *MAE*、*MAPE* 准则下,基于不同的缺失率,IWKNN 的置信区间起点和终点值均小于其他 2 种算法,而对应的置信区间长度却不存在类似的优势;第三,在 *RMSE* 准则下,基于不同的缺失率前提,IWKNN 的置信区间起点、终点、长度几乎全部大于 KNN、WKNN。综上,在 MAR 下,IWKNN 仍能保持良好的填补优势,但算法填补的精度依然有所下降,这与 MACR 下的结论保持一致。特别地,对比分析 MACR 下的实验结果发现,在 MAR 下,IWKNN 的填补优势有所下降。

3.3 非随机缺失

在 NMAR 下采用相同的实验方法对 KNN、WKNN、IWKNN 继续进行评估,实验结果如图 5 所示。



(a) *MAE* 均值 (b) *RMSE* 均值 (c) *MAPE* 均值

图 5 基于 Boston 数据集,不同缺失率下、3 种填补算法在 NMAR 前提下填补 1 000 次误差结果均值

Fig. 5 Based on the Boston dataset, under different missing rates, the mean of three filling algorithms filling 1 000 error results under the premise of NMAR

图 5 表明，在 NMAR 下，随着缺失率的增大，KNN、WKNN、IWKNN 基于不同评价准则所得 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 结果出现增大趋势并未改变，且对比分析图 3、图 4 发现 $\overline{MAE}$、$\overline{RMSE}$、$\overline{MAPE}$ 值进一步大于 MAR 下的结果。这表明，填补算法在 NMAR 下或将变得不再适用；第二，相较于 KNN、WKNN，IWKNN 的填补优势只在缺失率较低时存在，随着缺失率的增大，3 种算法的填补效果接近一致。不同缺失率下、3 种填补算法在 NMAR 前提下填补 1 000 次误差结果的置信区间及其长度见表 5。

表 5　不同缺失率下、3 种填补算法在 NMAR 前提下填补 1000 次误差结果的置信区间及其长度

Table 5　Under different missing rates, The confidence interval and length of three filling algorithms filling 1000 error results under the premise of NMAR

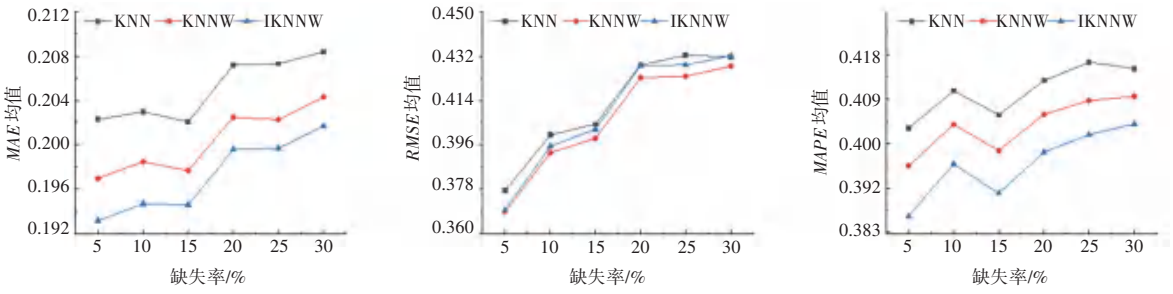
评价准则	填补算法	结果类型	Boston 数据集所含缺失值总体占比(缺失率)/%					
			5	10	15	20	25	30
MAE	KNN	置信区间	[0.110,0.123]	[0.116,0.127]	[0.136,0.151]	[0.145,0.160]	[0.147,0.161]	[0.153,0.166]
		区间长度	0.013	0.010	0.014	0.015	0.015	0.013
	WKNN	置信区间	[0.106,0.119]	[0.114,0.124]	[0.133,0.148]	[0.143,0.158]	[0.143,0.162]	[0.151,0.164]
		区间长度	0.013	0.010	0.014	0.015	0.019	0.013
	IWKNN	置信区间	[0.103,0.117]	[0.114,0.124]	[0.135,0.150]	[0.146,0.165]	[0.146,0.166]	[0.155,0.168]
		区间长度	0.013	0.011	0.015	0.018	0.020	0.012
RMSE	KNN	置信区间	[0.162,0.181]	[0.171,0.186]	[0.203,0.224]	[0.213,0.234]	[0.213,0.232]	[0.225,0.243]
		区间长度	0.019	0.015	0.021	0.021	0.019	0.018
	WKNN	置信区间	[0.157,0.176]	[0.168,0.183]	[0.199,0.220]	[0.211,0.232]	[0.210,0.234]	[0.223,0.241]
		区间长度	0.019	0.015	0.021	0.021	0.024	0.018
	IWKNN	置信区间	[0.159,0.179]	[0.174,0.189]	[0.206,0.229]	[0.220,0.247]	[0.218,0.244]	[0.233,0.252]
		区间长度	0.020	0.016	0.023	0.027	0.026	0.019
MAPE	KNN	置信区间	[0.225,0.484]	[0.297,0.606]	[0.461,0.827]	[0.525,0.831]	[0.567,0.870]	[0.498,0.726]
		区间长度	0.258	0.309	0.366	0.307	0.303	0.228
	WKNN	置信区间	[0.217,0.472]	[0.288,0.596]	[0.452,0.817]	[0.515,0.820]	[0.557,0.859]	[0.485,0.712]
		区间长度	0.255	0.308	0.365	0.306	0.302	0.227
	IWKNN	置信区间	[0.200,0.462]	[0.281,0.597]	[0.446,0.810]	[0.511,0.816]	[0.549,0.849]	[0.478,0.704]
		区间长度	0.263	0.316	0.364	0.305	0.300	0.226

基于上述结论来分析表 5 的实验结果发现，第一，在 *MAE* 准则下，当 $p = 0.05$ 时，IWKNN 的置信区间起点、终点、区间长度相较于 KNN、WKNN 依然具备微弱的优势，当 $p \geqslant 0.1$ 时，IWKNN 的填补效果弱于 WKNN，甚至弱于 KNN，而在 *MAPE* 准则下，却出现相反的结果；第二，在 RMSE 准则下，IWKNN 在不同缺失率下的置信区间起点、终点、长度均大于其他 2 种算法。综上，在 NMAR 下，文中提到的填补算法面临失效，这印证了图 5 的结论；此外，由于 NMAR 下观测值的缺失与自身相关的缘故导致了 IWKNN 算法在结构上过渡依赖最邻近样本点的缺陷被放大。

3.4　其他数据集上的比较分析

为了保证不同数据集上参数优化方法对填补算法的改进效果，使用 UCI 上公开数据集 ILPD (https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29) 继续进行对比分析，基于相同实验方法对 KNN、WKNN、IWKNN 再次进行评估，实验结果如图 6~图 8 所示。

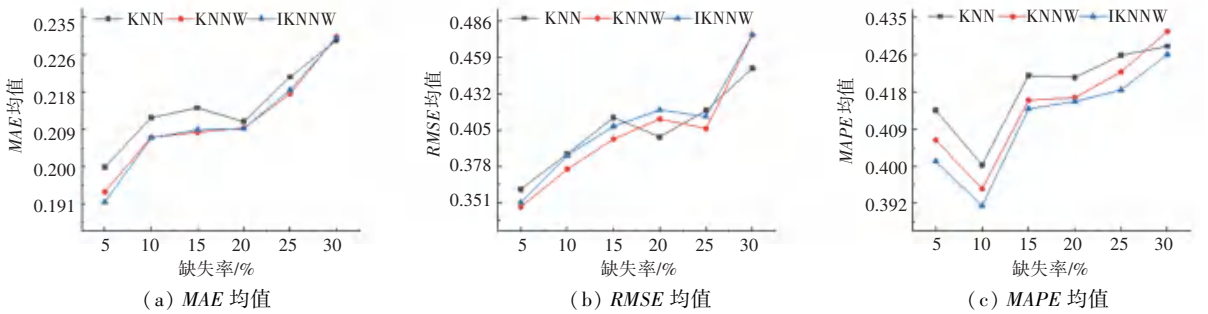
图 6~图 8 的实验结果显示：第一，在 MACR 下，基于 *MAE*、*MAPE* 准则，IWKNN 的填补优势依然显著，而在 *RMSE* 准则下弱于 WKNN，这与表 3 的实验结论基本一致；第二，在 MAR、NMAR 下，基于 *MAE*、*MAPE* 准则，IWKNN 仍然具备一定的填补优势，而在 *RMSE* 准则下，IWKNN 的结果接近或大于其他 2 种算法，这与表 4、表 5 的实验结论基本一致；第三，对比不同缺失机制、不同缺失率下的 $\overline{MAPE}$ 结果来看，NMAR 下，3 种算法填补效果下降严重，这与表 5 的实验结论基本一致。



(a) *MAE* 均值 (b) *RMSE* 均值 (c) *MAPE* 均值

图 6 基于 ILPD 数据集,不同缺失率下、3 种填补算法在 MCAR 前提下填补 1 000 次误差结果均值

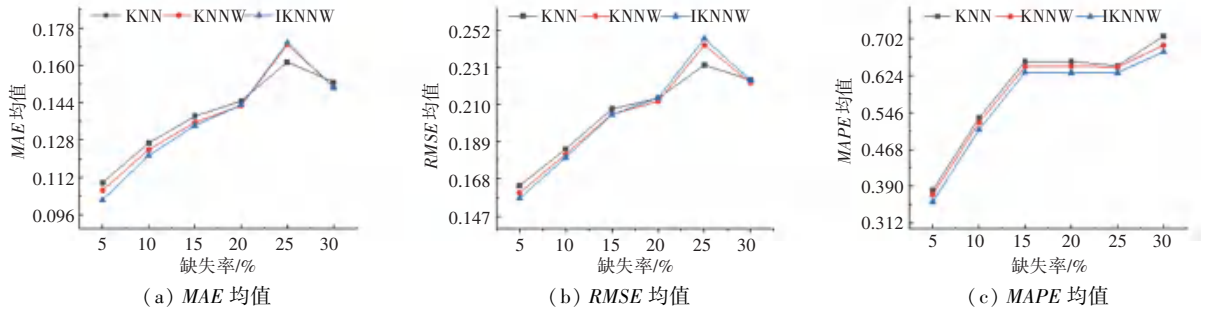
Fig. 6 Based on the ILPD dataset, under different missing rates, the mean of three filling algorithms filling 1 000 error results under the premise of MCAR



(a) *MAE* 均值 (b) *RMSE* 均值 (c) *MAPE* 均值

图 7 基于 ILPD 数据集,不同缺失率下、3 种填补算法在 MAR 前提下填补 1 000 次误差结果均值

Fig. 7 Based on the ILPD dataset, under different missing rates, the mean of three filling algorithms filling 1 000 error results under the premise of MAR



(a) *MAE* 均值 (b) *RMSE* 均值 (c) *MAPE* 均值

图 8 基于 ILPD 数据集,不同缺失率下、3 种填补算法在 NMAR 前提下填补 1 000 次误差结果均值

Fig. 8 Based on the ILPD dataset, under different missing rates, the mean of three filling algorithms filling 1 000 error results under the premise of NMAR

4 结束语

本文的实证分析结果显示:

(1)基于不同缺失率、任意缺失列,传统的交叉验证法将无法给出最优解,而通过 KFCVB 方法对 k 值进行优化,执行多次实验,从多次模拟结果的均值和标准差进行综合考虑,能够获取 k 值的整体最优解,且该方法选取 k 值的过程具有一定的参考价值,也为其他数据填补算法提供了优化路径;

(2)在 MACR 下,基于不同数据集、不同缺失率、任意缺失列的 IWKNN 算法填补优势显著,且在 MAR、NMAR 下,当缺失率较小时,该算法依然能够保持一定的填补优势;

(3)不同缺失机制下,随着缺失率的增大,3 种填补算法的评价结果均出现不同程度的增大趋势,这表明 K 近邻及其改进算法在数据缺失率较大时均面临不适用问题;

(4)基于不同数据集的实验结果可知,动态调参法具备可移植性好、稳定性佳的特点。

综上,在后续的研究中,将依据不同缺失机制的特点来改进参数优化方法,以此提升随机缺失,尤其是非随机缺失下的填补算法的精度,或尝试采用原有数据集的先验分布信息为 NMAR 下的数据填补问题提供解决方案。

参考文献

[1] 鲍晓蕾,高辉,胡良平. 多种填补方法在纵向缺失数据中的比较研究[J]. 中国卫生统计,2016,33(1):45-48.

[2] 李琳,杨红梅,杨日东. 基于临床数据集的缺失值处理方法比较[J]. 中国数字医学,2018,13(4): 8-10,80.

[3] 邓建新,单路宝,贺德强. 缺失数据的处理方法及其发展趋势[J]. 统计与决策,2019,35(23):28-34.

[4] 宋亮,万建洲. 缺失数据插补方法的比较研究[J]. 统计与决策,2020,36(18):10-14.

[5] GRAHAM J W. Missing Data: Analysis and Design[M]. Heidelberg: Springer, 2012.

[6] CHEVRET S, SEAMAN S, RESCHE RIGON M. Multiple imputation: A mature approach to dealing with missing data[J]. Intensive Care Medicine, 2015, 41(2):348-350.

[7] ZHANG Shaodian, KANG Tian, ZHANG Xingting. Speculation detection for Chinese clinical notes:Impacts of word segmentation and embedding models[J]. Journal of Biomedical Informatics, 2016,60:334-341.

[8] LITTLE R,RUBIN D. Statistical analysis with missing data[M]. New York:John Wiley and Sons,2002.

[9] HORTON N J, LAIRD N M. Maximum likelihood analysis of generalized linear models with missing covariates[J]. Statistical Methods in Medical Research,1999,8(1): 37-50.

[10] ALLISON P D. Multiple imputation for missing data: A cautionary tale[J]. Sociological Methods and Research,2000,28(3): 301-309.

[11] RUBIN D,DONALD B. Multiple Imputation for Nonresponse in Surveys[M]. New York: John Wiley & Sones,2002.

[12] RUBIN D, DONALD B. Multiple Imputation After 18+ Years[J]. Journal of the American Statistical Association, 1996, 91(434):473-489.

[13] TROYANSKAYA O, CANTOR M, SHERLOCK G. Missing value estimation methods for DNA microarrays [J]. Bioinformatics,2001,17(6):520-525.

[14] XIAO Jianli. SVM and KNN ensemble learning for traffic incident detection [J]. Physica A: Statistical Mechanics and Its Applications, 2019, 517(C):29-35.

[15] PURWA R, ARCHAN A, KUMARS S. Hybrid prediction model with missing value imputation for medical data[J]. Expert Systems with Applications,2015,42(13):5621-5631.

[16] TUTZ G, RAMZAN S. Improved methods for the imputation of missing data by nearest neighbor methods[J]. Computational Statistics & Data Analysis, 2015,90:84-99.

[17] 杨日东,李琳,陈秋源,等. LKNNI:一种局部 K 近邻插补算法[J]. 中国卫生统计,2019,36(5):780-783.

[18] 陈婉娇. 缺失数据插补方法及其在医学领域的应用研究[D]. 广州:华南理工大学,2019.

[19] 刘佳星,张宏烈,刘艳菊. 基于缺失率的不完整数据填补算法[J]. 统计与决策,2021,37(2):39-41.

[20] 郑智泉,王孟孟,田维琦. 基于加权 K 近邻算法的缺失数据填补研究[J]. 智能计算机与应用,2021,11(11):31-33,42.

[21] 郑智泉,陈妍,王孟孟,等. 不同缺失率下的数据填补算法稳定性研究[J]. 统计与决策,2023,39(8):12-17.