Frontiers of International Accounting 国际会计前沿, 2025, 14(1), 10-19
Published Online February 2025 in Hans. https://www.hanspub.org/journal/fia
https://doi.org/10.12677/fia.2025.141002

Hans汉斯

基于乳腺癌数据的插补方法比较研究

杨　丹，左俊希

重庆理工大学理学院，重庆

收稿日期：2024年12月6日；录用日期：2025年1月27日；发布日期：2025年2月10日

摘　要

缺失数据一直是数据分析工作中面临的难题之一，缺失数据的存在会导致模型性能的损耗，因此尽可能准确地预测填补缺失的方法变得尤为重要。本文将依托于“威斯康星乳腺癌诊断”数据集进行常见插补方法的比较研究，首先将原始数据按照完全随机缺失机制进行缺失处理，然后通过建立Logistic模型、支持向量机模型两种不同的模型，在不同缺失率(10%、30%)、不同协变量缺失个数(3个、6个)条件下，比较均值插补、KNN插补、多重插补3种不同插补方法的性能。同时，将准确率、F1值、AUC值作为衡量插补效果的量化指标。本文的实验结果表明，支持向量机模型对于乳腺癌数据集的拟合效果明显好于Logistic模型；同时对于所有的插补方法都会随着缺失率和缺失协变量的个数的增加，而性能发生降低。插补性能下降幅度却不相同，多重插补的性能明显更稳定，下降幅度最小，同时多重插补的插补效果综合来看也是最好的。对数据进行多重插补后拟合的Logistic模型和支持向量机模型在缺失率为30%、缺失协变量个数为6个的时候，对应准确率、F1值、AUC值分别为0.894、0.923、0.872和0.923、0.94、0.908。因此得出，基于生成多个数据集来模拟缺失数据不确定性的多重插值，在进行完全随机缺失处理后的“威斯康星乳腺癌诊断”数据集上相较于均值插补和KNN插补，其插补的稳健性和可信度明显更高。

关键词

缺失数据，多重插补，KNN插补，均值插补

A Comparative Study of Interpolation Methods Based on Breast Cancer Data

Dan Yang, Junxi Zuo

College of Science, Chongqing University of Technology, Chongqing

Received: Dec. 6th, 2024; accepted: Jan. 27th, 2025; published: Feb. 10th, 2025

Abstract

Missing data has always been one of the challenges faced in data analysis. The presence of missing

文章引用: 杨丹, 左俊希. 基于乳腺癌数据的插补方法比较研究[J]. 国际会计前沿, 2025, 14(1): 10-19.
DOI: 10.12677/fia.2025.141002

data can lead to a loss of model performance, so it is particularly important to predict and fill in missing data as accurately as possible. This paper will rely on the data set of "Wisconsin Breast Cancer Diagnosis" to carry out a comparative study of common interpolation methods. First, the original data will be deleted according to the complete random deletion mechanism. Then, by establishing two different models, the Logistic model and the support vector machine model, under the conditions of different deletion rates (10%, 30%) and different number of covariate deletions (3, 6), the mean interpolation, KNN interpolation. The performance of three different interpolation methods for multiple interpolation. At the same time, accuracy, F1 value, and AUC value are used as quantitative indicators to measure the interpolation effect. The experimental results in this paper show that the fitting effect of support vector machine model for breast cancer dataset is significantly better than that of Logistic model; At the same time, for all interpolation methods, the performance will decrease with the increase of the missing rate and the number of missing covariates. The decrease in interpolation performance varies, with multiple interpolation showing significantly more stable performance with the smallest decrease. Overall, the interpolation effect of multiple interpolation is also the best. The logistic model and support vector machine model that fit the data after multiple imputation have corresponding accuracy, F1 value, and AUC value of 0.894, 0.923, 0.872, and 0.923, 0.94, and 0.908, respectively, when the missing rate is 30% and the number of missing covariates is 6. Therefore, based on generating multiple data sets to simulate the multiple interpolation of the uncertainty of missing data, the "Wisconsin Breast Cancer Diagnosis" data set after complete random deletion processing is significantly more robust and reliable than mean interpolation and KNN interpolation.

Keywords

Missing Data, Multiple Imputation, KNN Interpolation, Mean Interpolation

1. 研究背景及意义

数据在当今社会中扮演着至关重要的角色，它们是我们了解世界、做出决策和解决问题的基础。数据可以帮助我们了解市场趋势、人口统计信息、环境状况、医疗健康和科学研究等方面的信息[1]。在企业和政府机构中，数据可以用于制定战略、优化流程、提高效率和减少成本。在个人生活中，数据可以帮助我们做出更好的决策，例如购买房屋、投资股票、选择医疗保健和规划旅行等。

在过去几年中数据的重要性变得越来越明显，这主要是因为我们现在有更多的数据可用。随着互联网的普及、传感器技术的发展和移动设备的普及，我们现在可以收集和存储比以往任何时候都更多的数据。这些数据可以用于许多不同的目的，例如预测未来趋势、识别模式、发现新的机会和解决问题。同时，高质量的数据可以帮助我们做出更准确的决策，而低质量的数据则可能导致错误的决策。因此，确保数据的准确性、完整性和一致性对于正确使用数据至关重要。我们需要更加重视数据的质量和分析，以确保我们能够正确地使用数据并从中获得最大的价值。

而事实上，数据缺失是实际数据分析中一个常见的问题，在实际应用中，数据可能因为各种主观或客观原因，不可避免地会存在一些数据缺失的情况。例如问卷调查中的未答问题、医疗记录中的遗漏信息、或者实验数据中的测量误差等。如果不处理这些缺失数据，可能会导致分析结果不准确，影响决策的正确性[2]。因此，如何处理缺失数据并尽可能的还原原本真实的数据集变得至关重要。目前对于缺失数据处理方法有很多，例如：均值插补法、众数插补法、热卡填充法、随机森林插补法、多重插补法等等。但如何合理地选择插补缺失数据方法，都没有一个统计的标准，因此在实际处理中，研究者往往通

过经验或尝试多种插值方法，选择最优的。

2. 模型及评价指标介绍

2.1. 模型介绍

2.1.1. 支持向量机

支持向量机(Support Vector Machines, SVM)是针对二分类任务设计的，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 还包括核技巧，这使它成为实质上的非线性分类器[3]。SVM 的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM 的学习算法就是求解凸二次规划的最优化算法。而 SVM 具有根据有限样本找到最优解的能力，能够避免局部极值问题而得到全局最优点和高维特征处理能力。在现实任务中，样本也许在特征空间中不是线性可分的，往往需要将样本从原始空间映射到一个更高维的空间，使得在这个空间内是线性可分的。即非线性分类是建立在线性分类基础上的，故这里需要先对线性分类进行简单的说明。

(1) 线性支持向量机

支持向量机的线性分类的目标是希望在特征空间中找到一个划分超平面，将拥有不同标记的样本分开，并且该划分超平面距离各样本最远，即需找到一组合适的参数 (w,b) 使得[4]

$$\max_{w,b}\ \min_{i}\frac{2}{\|w\|}\left|w^T x_i+b\right|$$

$$s.t.\ y_i h(x_i)=1,\ i=1,2,\cdots m$$

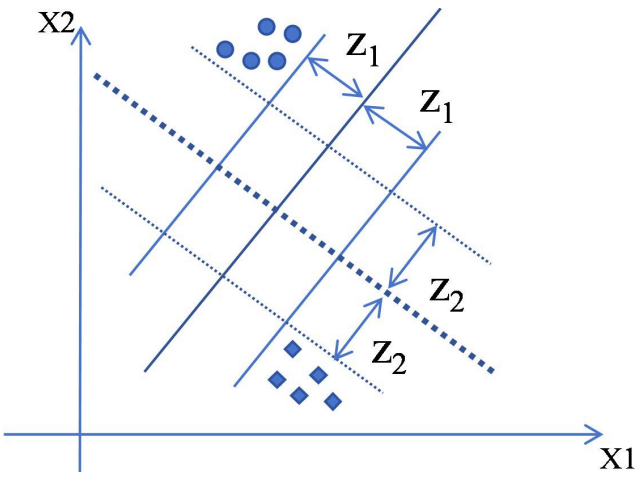

Figure 1. Division of hyperplanes for support vector machines
图 1. 支持向量机超平面的划分

上图 1 描述了两种线性支持向量机分类器(图中虚线和实线)，途中虚线对应的分类器，期间隔为 $2z_2$，实线对应的分类器间隔为 $2z_1$，由于 $2z_2>2z_1$，故最终选择虚线对应的分类器。

(2) 非线性支持向量机

由于数据在原始特征空间 R^d 中不是线性可分的，故支持向量机希望通过一个映射：$\varphi: R^d \to R^{\tilde{d}}$，使得数据在新的特征空间 R^d 是线性可分的[5]。令 $\varphi(x)$ 表示将样本 x 映射到 R^d 中的特征向量，参数 w 的维度相应变为 $\tilde{d}$，则非线性支持向量机可表示为如下形式：

$$\max_{w,b}\ \min_{i}\frac{2}{\|w\|}\left|w^{T}\phi\left(x_{i}\right)+b\right|$$

$$s.t.\ y_{i}h\left(x_{i}\right)=1,\ i=1,2,\cdots m$$

2.1.2. Logistic 模型

逻辑回归是一种用于分类的统计学习方法，它是一种广义线性模型，用于预测二元变量的输出。在逻辑回归中，我们使用一个称为逻辑函数的函数，将一个或多个自变量映射到一个介于 0 和 1 之间的概率值。这个概率值可以被解释为给定输入特征的情况下，输出为 1 的概率。设因变量 y_i 为二元的分类变量(按常规 0/1 变量取值)，由于因变量期望的特殊性，我们不直接将因变量本身的取值作为回归模型中的因变量，而将 $p\left(X\right)=\Pr\left(Y=1\middle|X\right)$ 作为回归模型中的因变量。现在考虑多元 Logistic 回归的情况，Logistic 函数的形式为：

$$p\left(X\right)=\frac{e^{\beta_0+\beta_1x_1+\cdots\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\cdots\beta_px_p}}$$

这里的 $X=(x_1,\cdots,x_p)$ 是 p 个自变量，β_0 为常数项，$\beta_1,\cdots,\beta_p$ 为回归系数；经过变换可得以下形式：

$$\ln\left(\frac{p\left(X\right)}{1-p\left(X\right)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_px_p$$

等式左边的称为对数发生比(log-odd)或分对数(logit)。

2.2. 模型评价指标

对于分类模型的评价方式和评价指标有很多种，常见的有准确率、精确率、召回率、F1 值、ROC 曲线和 AUC 值、混淆矩阵等，不同的评价指标适用于不同的场景和应用，需要根据具体问题选择合适的指标来评价分类模型的性能。而本文将选取准确率、F1 值、AUC 值，3 个指标来评价最终拟合的分类模型的性能。下面将对以上 3 个评价指标做简要介绍：

(1) 准确率

准确率(Accuracy)：分类正确的样本数占总样本数的比例，是最常用的评价指标之一。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

其中 TP 表示真正例(True Positive)，TN 表示真负例(True Negative)，FP 表示假正例(False Positive)，FN 表示假负例(False Negative)。准确率越高，分类器的性能越好。

(2) F1 值

F1 值(F1-score)：F1 值是精确率和召回率的加权平均值，综合考虑了分类器的对正例的识别能力和对正例的覆盖能力。计算公式如下：

$$F1=\frac{2PR}{P+R}$$

其中 $P=\frac{TP}{TP+FP}$ 表示精度，$R=\frac{TP}{TP+FN}$ 表示召回率。F1 值越高，分类器的性能越好。

(3) AUC 值

AUC (Area Under the Curve)是 ROC 曲线下的面积，是评估二分类模型性能的一种指标。ROC 曲线是将二分类模型的真阳性率(TPR)与假阳性率(FPR)绘制成的曲线，其中 TPR 是指实际为正例的样本中被

正确预测为正例的比例，FPR 是指实际为负例的样本中被错误预测为正例的比例。AUC 值的范围在 0 到 1 之间，AUC 值越大，说明模型性能越好，AUC 值为 0.5 时表示模型预测效果等同于随机猜测。

AUC 值的优点在于它不受分类阈值的影响，而且对于不平衡的数据集也有较好的鲁棒性。在实际应用中，AUC 值常用于比较不同模型的性能，选择最优模型。同时，AUC 值也可以用于评估模型在不同阈值下的性能表现，帮助选择最优的分类阈值。

3. 缺失机制及插补方法介绍

3.1. 缺失机制

缺失机制指的是数据缺失的原因和方式。在数据分析中，缺失机制通常分为三种类型：完全随机缺失、随机缺失和非随机缺失，这是由 Robin 等人在 1976 年提出的。开始介绍这 3 种缺失机制之前，我们先给定一些符号设置：

$$\begin{cases} X：\text{表示完成数据集} \\ X_{mis}：\text{表示完整数据集中未观测到的部分} \\ X_{obs}：\text{表示完整数据集中观测到的部分} \\ \delta：\text{当}\delta=1\text{表示目标变量无缺失，}\delta=0\text{表示目标变量存在缺失} \\ \phi：\text{表示未知参数} \end{cases}$$

(1) 完全随机缺失

完全随机缺失(MCAR)是指缺失数据的出现与观测值和缺失值本身无关，缺失数据的出现是随机的[6]。例如，一些数据点在记录时被不小心删除了，这些数据点的缺失是完全随机的。即：

$$P\left(\delta\middle|X,\phi\right)=P\left(\delta,\phi\right)$$

(2) 随机缺失

随机缺失(MAR)是指缺失数据的出现与观测值本身有关，但与缺失值本身无关[7]。例如，一项调查中，女性可能不愿意透露自己的年龄，但是男性却愿意透露，这时候女性的年龄数据就会出现随机缺失。即：

$$P\left(\delta\middle|X,\phi\right)=P\left(\delta\middle|X_{\text{obs}},\phi\right)$$

(3) 非随机缺失

非随机缺失(MNAR)是指缺失数据的出现与观测值和缺失值本身都有关[8]。例如，一项调查中，收入高的人可能不愿意透露自己的收入，这时候收入数据就会出现非随机缺失。即：

$$P\left(\delta\middle|X,\phi\right)=P\left(\delta\middle|X_{\text{obs}},X_{mis},\phi\right)$$

3.2. 插补方法介绍

3.2.1. 均值插补

均值插值法是一种简单的缺失数据处理方法，它的基本思想是用缺失值所在变量的均值来填充缺失值[9]。具体来说，对于一个含有缺失值的变量，我们可以计算出该变量的均值，然后用该均值来替换缺失值。均值插值法的优点是简单易行，计算量小，能够处理大量缺失值。缺点是不考虑其他变量的影响，可能会引入偏差，而且对于分布不均匀的数据，均值插值法的效果可能不佳。

3.2.2. 多重插补

多重插补(Multiple Imputation)是一种常用的缺失数据处理方法，它通过多次模拟来填补缺失数据，并

生成多个完整数据集[10]。每个完整数据集都包含了每一个缺失数据的一个可能的取值，从而反映了缺失数据的不确定性。多重插补的基本思想是：对于每个缺失变量，根据已有的数据和一些随机噪声，生成多个可能的填补值，然后对每个完整数据集进行分析，最后将多个分析结果进行合并。

多重插补的优点是可以反映缺失数据的不确定性，同时可以保留样本的大小和分布特征。它的缺点是需要进行多次模拟，计算量较大，而且需要对每个缺失变量进行填补值的生成和分析结果的合并。

3.2.3. K 最近邻插补

K 最近邻插补也是一种十分常见的缺失数据处理方法。他通过比较样本与样本之间的相似程度，来找到最接近缺失样本的 K 个样本，然后综合这 K 个样本的信息来填充对应的缺失值。具体步骤如下[11]：

(1) 计算缺失样本与其他样本的相似度，即计算样本与样本之间的具体，常见的计算距离的公式有：欧氏距离($D=\left|x_i-x_j\right|$)、马氏距离($D=\left(x_i-x_j\right)^T S_{xx}^{-1}\left(x_i-x_j\right)$)。

(2) 对计算出来的与缺失样本的相似度进行排序，然后找出和缺失样本相似度最高的的 K 个样本。

(3) 最后根据这 K 个样本的均值、中位数或众数等信息来填补缺失值。

K 最近邻插补方法的优点是简单易用，且不需要对数据分布进行假设，但也存在一些缺点，比如对于高维数据，K 最近邻方法的计算量会很大，同时对于样本分布不均匀的情况，插补结果可能不够准确。同时，K 值的选取一定程度会影响插补缺失数据的效果，在实际操作中可通过交叉检验等方式来确定 K 值。

4. 缺失数据插补方法的对比实验

4.1. 数据来源

数据来源于 UCI 机器学习数据仓库(UCI Machine Learning Repository)的“威斯康星乳腺癌诊断”数据集，该数据是由威斯康星大学的研究者捐赠的。本文截取了其中部分数据用作研究，最后得到 569 例细胞活检案例，12 个特征变量，其中 diagnosis 表示癌症诊断结果(“M”表示恶性，“B”表示良性)。部分数据如下所示，见表 1：

Table 1. Part of the original data
表 1. 部分原始数据

id	diagnosis	radius	texture	perimeter	area	smoothness	compactness
87139402	B	12.32	12.39	78.85	464.1	0.1028	0.06981
8910251	B	10.6	18.95	69.28	346.4	0.09688	0.1147
862989	B	10.49	19.29	67.41	336.1	0.09989	0.08578
89827	B	11.06	14.96	71.49	373.9	0.1033	0.09097
91485	M	20.59	21.24	137.8	1320	0.1085	0.1644
8711003	B	12.25	17.94	78.27	460.3	0.08654	0.06679
925277	B	14.59	22.68	96.39	657.1	0.08473	0.133
867387	B	15.71	13.93	102	761.7	0.09462	0.09462

4.2. 模型建立

4.2.1. 生成缺失数据集

在获取的乳腺癌数据集中，我们需要将不含缺失的完成数据集进行缺失值处理，目的是为了比较不

同插补方法在处理缺失数据上的效果。通过运用均值插补、KNN 插补、多重插补 3 种不同的插补方法得到不同的数据集，将这些数据集分别拟合 Logistic 模型和支持向量机，并用具体的指标(准确率、F1 值、AUC 值)来量化插补方法的效果。

缺失变量的个数按照 3 个、6 个两个维度进行设置，缺失变量的选择按照随机原则进行选取。对于需要缺失的变量，按照完全随机缺失机制，通过 python 中 miceforest 模块的 ampute_data 函数生成缺失数据，缺失率按照 10%，20%，30%进行设置。实验的具体设置如下，见表 2：

Table 2. Experimental setup details
表 2. 实验设置明细表

参数	变动范围
样本量	569
缺失变量个数	3、6
缺失率	10%、30%
插补方法	均值插补、KNN 插补、多重插补
拟合模型	Logistic 模型、支持向量机

4.2.2. Logistic 模型

根据得到的缺失数据，我将运用不同的插补方法来补全数据，从而得到完成的数据集。对得到的完整数据集，我们将数据集的 70%作为训练集用于拟合模型，将数据集的 30%作为测试集用于评估模型的性能，同时反映数据插补效果。接下来我们将按照缺失协变量个数 3 个、6 个两个维度对数据集建立 Logistic 模型。

(1) 设置缺失协变量个数为 3 个

Table 3. Relevant evaluation indicators for logistic models with 3 missing variables
表 3. 3 个变量缺失下 Logistics 模型的相关评价指标

缺失率							
10%				30%			
插补方法	准确率	F1	AUC	插补方法	准确率	F1	AUC
均值	0.895	0.923	0.866	均值	0.861	0.896	0.839
KNN	0.921	0.943	0.897	KNN	0.892	0.926	0.876
多重插补	0.915	0.934	0.887	多重插补	0.912	0.929	0.879

见表 3，通过以上分析结果，可以发现当缺失协变量个数为 3 个的时候，随着缺失率的增加 3 种插值方法对应拟合的 Logistic 模型的精确度、F1 值、AUC 都有一定程度的降低。在缺失率为 10%的时候，KNN 的效果最好，对应准确率、F1 值、AUC 分别为 0.921、0.943、0.897。在缺失率为 30%的时候，多重插补的效果最好，对应准确率、F1 值、AUC 分别为 0.912、0.929、0.879。

综合两种不同的缺失率得出的结果来看，可以发现多重插补的效果较为稳定，虽然准确率、F1 值、AUC 都有一定程度的下降，但是下降的幅度很小。而均值插补和 KNN 插补的效果虽然在缺失率较小的时候也还不错，甚至 KNN 的插补效果还好于多重插补。但是我们可以明显的看出均值插补和 KNN 插补的稳定性明显不够，由于均值插补是根据每个缺失协变量对应的均值去插补每一个缺失值，因此当缺失率较高的时候，均值插补就显得略微粗糙了；而 KNN 插值在缺失比例较高的时候，样本之间的距离计算会变得不可靠，导致插补的效果变差。因此，对比起来多重插补的稳定性更好。

(2) 设置缺失协变量个数为：6 个

Table 4. Relevant evaluation indicators for Logistics models with 6 missing variables
表 4. 6 个变量缺失下 Logistics 模型的相关评价指标

缺失率							
10%				30%			
插补方法	准确率	F1	AUC	插补方法	准确率	F1	AUC
均值	0.885	0.910	0.866	均值	0.842	0.866	0.831
KNN	0.871	0.905	0.839	KNN	0.868	0.906	0.825
多重插补	0.912	0.935	0.899	多重插补	0.894	0.923	0.872

见表 4，当我们将缺失协变量的个数设置为 6 个数的时候，可以发现所有插补方法的效果在一定程度上都出现了下降。因此可以得出，缺失变量的个数一定程度决定了还原真实数据的难度，缺失变量个数越多，对于插补方法的要求也更高。同时，根据上述结果可以看出多重插补综合来看的效果显然更好，而且更稳定。虽然随着缺失率的增加，准确率、F1、AUC 都出现了下降，但和均值插补和 KNN 插补相比，下降幅度明显更小。下面给出了多重插补在缺失协变量个数为 6 个，缺失率为 30%的插补值分布、插补值均值的诊断图，如下：

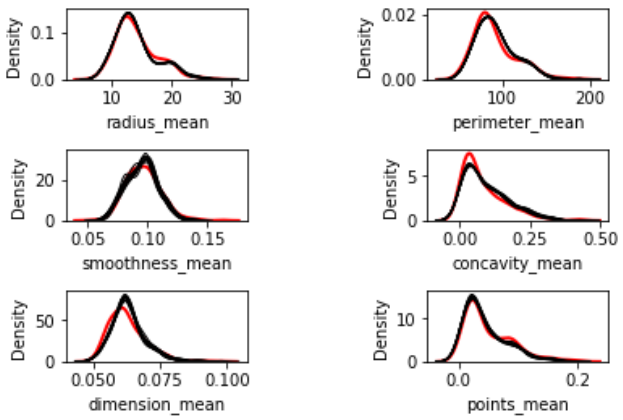

Figure 2. Multiple interpolation of the distribution of each variable
图 2. 多重插值各变量的分布图

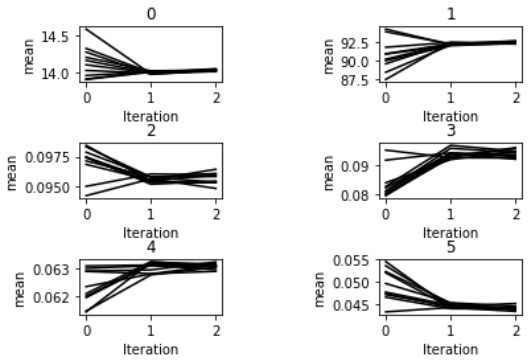

Figure 3. Plot of the degree of convergence of the means of the variables
图 3. 各变量均值的收敛程度图

在图 2 中，红色的线代表的是原始数据的核密度曲线，黑色的线是多重插补中设置的 10 个数据集，对应插补后的数据的核密度曲线，由于缺失了 6 个协变量，因此对应有 6 个核密度图。从图 2 中，可以发现插补后的数据和原始数据基本吻合，说明插补效果不错。从图 3 中，可以发现对于每个缺失变量，其插补的平均值，都在迭代一次以后就基本收敛，说明插补的数据较为稳定。

4.2.3. 支持向量机

在上一节我们运用 Logistic 模型，对插补后的乳腺癌数据集进行了建模，同时评估各种插补方法的效果，接下来我们将用支持向量机对插补后的数据集进行建模。同样的将 70%的数据集作为训练集，将 30%的数据集作为测试集，按缺失协变量个数为 3 个和 6 个两个维度下，探究不同插值方法在不同缺失率下的效果。对于支持向量机参数的选取，我们利用随机搜索的方法寻找出最优参数，根据随机搜索的结果，发现下面拟合的支持向量机模型都是采用的线性核的形式，具体实验结果如下：

(1) 设置缺失协变量个数为：3 个

Table 5. Relevant evaluation metrics for support vector machines with 3 missing variables
表 5. 3 个变量缺失下支持向量机的相关评价指标

缺失率							
10%				30%			
插补方法	准确率	F1	AUC	插补方法	准确率	F1	AUC
均值	0.923	0.943	0.904	均值	0.885	0.916	0.866
KNN	0.921	0.943	0.897	KNN	0.894	0.917	0.881
多重插补	0.935	0.949	0.929	多重插补	0.912	0.932	0.895

见表 5，在缺失协变量个数为 3 个的时候，不管缺失率为 10%还是 30%多重插值的效果都好于均值插补和 KNN 插补，同时也更加稳定。在缺失率为 10%的时候，多重插补后拟合的支持向量机精确率、F1 值、AUC 值分别为 0.935、0.949、0.929；在缺失率为 30%的时候，多重插补后拟合的支持向量机模型精确率、F1 值、AUC 值分别为 0.912、0.932、0.895。

(2) 设置缺失变量个数为：6 个

Table 6. Relevant evaluation metrics for support vector machines with 6 missing variables
表 6. 6 个变量缺失下支持向量机的相关评价指标

缺失率							
10%				30%			
插补方法	准确率	F1	AUC	插补方法	准确率	F1	AUC
均值	0.912	0.935	0.887	均值	0.851	0.876	0.841
KNN	0.935	0.951	0.916	KNN	0.894	0.924	0.866
多重插补	0.938	0.954	0.926	多重插补	0.923	0.94	0.908

见表 6，在缺失协变量个数为 6 个的时候，同样可以发现模型的拟合效果相较于缺失变量为 3 个的时候都有一定程度的降低，同时随着随机缺失率的提高，所有插补的效果都存在一定的较低。3 种不同的插值方法中，依然是多重插补的效果最好。在缺失率为 10%和 30%的时候，多重插补的精确率、F1 值、AUC 值分别为 0.938、0.954、0.926 和 0.923、0.94、0.908。同时，我们发现不管是在缺失变量为 3 个还

是 6 个的时候，支持向量机在本文中的乳腺癌数据集的拟合效果综合来说都好于 Logistic 模型，因此用支持向量机建立的模型用来预测乳腺癌的诊断效果更好。

5. 总计与展望

本文通过对乳腺癌数据集进行缺失，来探究不同插值方法的插值性能，并分别拟合 Logistic 模型和支持向量机模型，对插值的效果进行量化的评估。根据本文的实验结果，发现所有插值方法都会随着缺失率的增加和缺失变量的个数的增加性能出现降低。同时，发现多重插补在处理缺失的性能上相较于均值插补和 KNN 插补更好，一方面它使得建立的模型在测试集上的精度更高，另一方面它的插补效果更稳定，不会随着缺失率和缺失变量个数的变化出现陡增或骤减的情况。均值插补是一种常用且简单的插值方法，在缺失率较低的情况，它和其他的插补方法效果区别是不特别大；但正是由于他的插补方式过于粗糙，导致在缺失率以及缺失变量个数增加的时候，他的效果会出现下降。而 KNN 在缺失率降低的时候其实与多重插补的效果相比是不相上下的，但数据缺失率较高的时候，运用 KNN 进行插补会受到一些限制；由于 KNN 算法需要计算样本之间的距离，而缺失率过高的时候，会影响样本间距离的计算，这时候 KNN 的可信度和性能就会降低。而多重插补通过生成多个数据集来模拟缺失数据的不确定性，从而能够提高结果的稳健性和可信度。

数据缺失的情况在各个领域都十分常见，因此作为处理缺失数据的常见方法插补变得越来越重要，如何结合实际数据选择合适的插补方法是一个具有实现意义的话题。而本文中对于插补方法的比较研究中，涉及的插补方法较少，因此后续应该考虑更多种插补方法的比较。同时，对于缺失变量的不同数据类型，对插值方法的选择有没有影响，这些都是值得研究的问题，需要后续进一步地深入研究。

基金项目

重庆理工大学研究生创新项目资助，为重庆理工大学研究生教育高质量发展行动计划资助成果。项目编号：gzlcx20232084；项目类选：校级全额资助一般项目；成果单位：重庆理工大学。

参考文献

[1] 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.

[2] 郑智泉, 陈妍, 王孟孟, 田维琦. 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.

[3] 彭海艳, 李意芝, 孟利军. 基于数据缺失率和缺失模式的多重插补误差研究[J]. 统计与决策, 2022, 38(1): 20-24.

[4] 汤健, 徐雯, 夏恒, 等. 面向城市固废焚烧过程的缺失数据填充及应用[J]. 北京工业大学学报, 2023, 49(4): 435-448.

[5] 费雪, 惠永昌, 吴帮玉, 等. 自监督连续缺失地震数据插值方法[C]. 2022 年中国石油物探学术年会论文集(下册). 2022: 502-505.

[6] 饶珍敏. 删失指标随机缺失数据下两类回归模型的统计推断[D]: [硕士学位论文]. 杭州: 浙江工商大学, 2022.

[7] 赵若男, 苏同生, 宋瑞, 等. 中风队列研究中多重插补法拟合量表缺失数据效果评价[J]. 中国中医药信息杂志, 2022, 29(3): 110-116.

[8] 邓建新, 单路宝, 贺德强, 等. 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.

[9] 陈玉. 基于聚类和半参数 Logistic 的缺失数据插补研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2022.

[10] 方芳. 完全随机缺失机制下 XGBOOST 模型缺失数据插补方法比较研究[D]: [硕士学位论文]. 昆明: 云南大学, 2021.

[11] 张彪, 韩伟, 庞海玉, 等. 完全随机缺失条件下分类随机变量数据缺失插补方法的比较研究[J]. 中国卫生统计, 2015, 32(5): 903-905, 907.