基于多重插补和 Logistic 回归的专家缺失信息处理方法研究

李梓祎 [1] 吴昕阳 [2]
LI Ziyi WU Xinyang

摘 要 数据是当前各研究领域研究对象的信息载体，但在理论研究、数据收集、统计判断过程中，常因理论缺失、经验不足、主观判断等导致收集的数据不完整。尤其在存在相关变量的数据集上， 数据的缺失将影响整体规律的判断。因此，一种针对不完整数据集的处理方法，用以实现有效 数据预处理，支撑科学研究显得尤为重要。围绕专家在某评估数据集中缺失数据的现象，区分数据类型，对于数值型数据，采用基于链式方程的多重插补法插补；对于类别型数据，采用 Logistic 回归模型插补，并通过专家自身偏好一致性及专家共识度评价指标，比较插补结果的优劣。

关键词 多重插补；Logistic 回归；专家自身偏好一致性；专家共识度

doi：10.3969/j.issn.1672-9528.2024.08.010

0 引言

本文研究专家缺失信息的插补方法，选取某项目中专家的打分数据作为研究对象。该数据集中专家的打分值为数值型数据，考虑到同一个专家的打分偏好，每一位专家的打分值应具有相关性，因此文章采用基于链式方程的多重插补法（multiple interpolation based on chain equation，MICE）对打分值进行插补。此外，数据集还包括专家对研究对象的类别判断数据。文章采用 Logistic 回归方法，在完成对打分数据插补的基础上建立插补模型，进行分类模型训练，完成类别型数据插补。

1 数据集介绍

本文选取的数据集包含 68 位专家在某项目中对同一研究对象的打分值，所有打分值分为两级指标——一级指标与二级指标。整个数据集的打分值包含 8 个一级指标与 24 个二级指标，其中每个一级指标下面包括三个二级指标。在所有打分值后，专家根据自身打分值确定研究对象的类别数据。具体数据变量的架构如图 1。

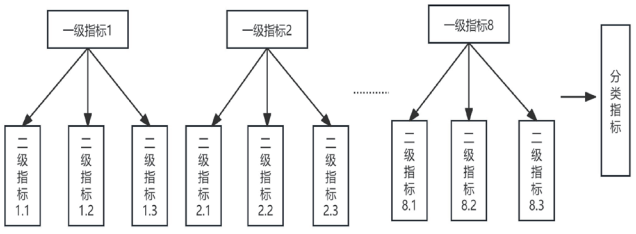

图 1 数据集变量介绍

1.95668 部队 云南昆明 650000
2. 国防科技大学系统工程学院 湖南长沙 410073

专家打分之间是相互独立的，因此同一指标值的缺失应相互独立。但根据专家偏好，同一专家在不同指标下的打分应存在相关性， 特别对同一内容的一级指标与二级指标的分数，相关性应较明显。

为验证上述猜想，这里计算了一级指标与二级指标间偏相关系数，其计算公式为：

$$r_{12,34\ldots(p-1)}=\frac{r_{12,34\ldots(p-1)}-r_{1p,34\ldots(p-1)}r_{2p,34\ldots(p-1)}}{\sqrt{1-r^2_{1p,34\ldots(p-1)}}\sqrt{1-r^2_{2p,34\ldots(p-1)}}} \quad (1)$$

式中：r_{12} 表示 X_1 与 X_2 之间的相关系数，r_{13} 表示 X_1 与 X_3 之间的相关系数，r_{23} 表示 X_2 与 X_3 之间的相关系数，同理可得其他变量的意义。

其中，二级指标是对研究对象某一方面的独立描述，彼此之间理论上无相关关系；一级指标是对 3 个二级指标的综合考虑结果，与 3 个二级指标分别有相关关系。为了验证这一假设，计算出了各一级指标与二级指标的偏相关系数，如表 1 所示。

表 1 一级指标与二级指标偏相关系数

	D1-1	D1-2	D1-3	D1-4
二级指标 -1	0.820	0.740	0.820	0.698
二级指标 -2	0.717	0.809	0.879	0.857
二级指标 -3	0.744	0.786	0.850	0.891
	D2-1	D2-2	D2-3	D2-4
二级指标 -1	0.640	0.722	0.791	0.768
二级指标 -2	0.773	0.815	0.732	0.748
二级指标 -3	0.683	0.701	0.755	0.867

根据统计学经验，当相关系数 r 的绝对值大于 0.8 时，认为两变量之间有强相关性；当 r 的绝对值介于 0.3 ～ 0.8 之

间时，认为两变量间有弱相关性。由表 1 可知，所有一级指标与对应二级指标间的相关系数都大于 0.6。由皮尔逊相关性检验可知，一级指标与其对应的 3 个二级指标都存在相关关系。综上，由专家打分独立性可知，专家打分数据的缺失值与不完全变量之间是相互独立的；由相关性分析可知，专家打分数据的缺失值依赖于其他完全变量。因此，可以认为专家打分数据是随机缺失的，可以使用多重插补法进行缺失值处理 [1]。

2 插补方法介绍

2.1 基于链式方程的多重插补法（MICE）

MICE 是在 MI 基础上改进的算法，它的基本思想是：在进行插补前，先将每个变量中的缺失值用临时占位符替换（该值来源于该变量可用的非缺失值）。然后对于每一个含有缺失值的变量，插补时考虑除该变量之外的所有变量，通过线性回归模型进行该变量缺失值的插补，这个过程重复多次，每一轮插补都用前一轮插补后生成的数据集。依此类推，插补好每一个含有缺失值的变量 [2]。

具体而言，链式方程插补算法如图 2 所示。

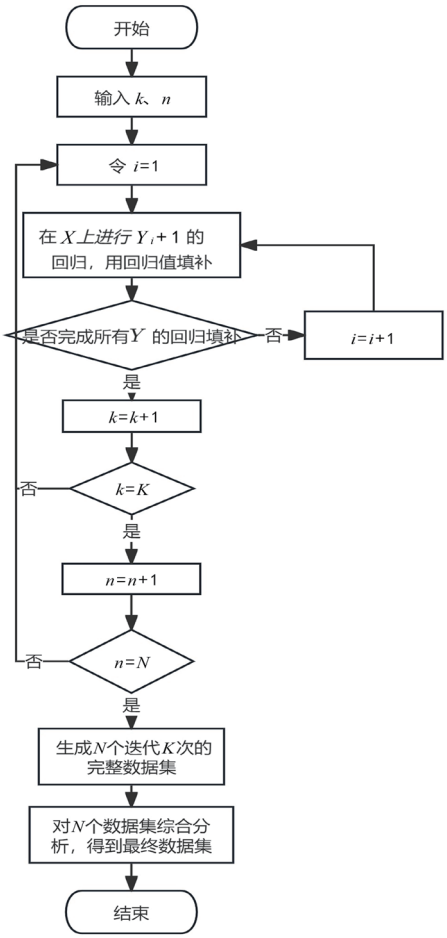

图 2 MICE 流程图

记 x 为无缺失值变量集，$Y_1, Y_2, Y_3, \cdots, Y_m$ 为 m 个含有缺失值的变量。对 $Y_1, Y_2, Y_3, \cdots, Y_m$ 的插补过程一共需要进行 n 次循环，即 n 次迭代（在代码中可以通过调整参数值进行控制）。在第一次循环时，$Y_1, Y_2, Y_3, \cdots, Y_m$ 中还含有缺失值，没有构成完整数据集，因此需要先用临时占位符插补空缺值。MICE 采取的方法是，首先在 Y_1 上做 x 的回归，根据此回归进行 Y_1 缺失值的插补；然后做 Y_2 在 Y_1（包括插补值）和 x 上的回归，同样根据此回归进行 Y_2 的插补；依此类推，将 $Y_3, \cdots, Y_m$ 都进行回归插补。在第一轮迭代结束时，所有缺失值都已经替换为反映数据中观察到的关系的回归预测值，形成了一个不含缺失值的完整数据集。第二轮到第 n 轮的迭代遵循第一轮的过程，区别是此时进行的回归应该包括除去本变量之外的其他所有变量，即每次都用到前一轮迭代的所有变量更新的最新值。在 n 轮迭代结束后，把第 n 轮插补的结果作为最终结果，形成一个新的完整的数据集。为了得到初始设定的 k 个数据集，需要将以上操作重复独立进行 k 次 [3-4]。

2.2 Logistic 回归方法

Logistic 回归也称为对数几率回归，是机器学习领域最为常见的一种模型方法，尤其是在研究因变量为二分类或多分类的分类问题时，常作为主要分析研究方法。Logistic 的主要思想基于两种常用的函数：线性回归函数与逻辑函数 [5]。

2.2.1 线性回归函数

作为普通线性函数的扩展，Logistic 回归的主要框架也是线性回归函数：

$$y = a^{\mathrm{T}}x+b \tag{2}$$

式中：x 为自变量，y 为因变量，a 为常数项，不同的 a 值反映了 x 对 y 的不同贡献程度。

2.2.2 逻辑函数（Sigmoid 函数）

Sigmoid 函数常被用作神经网络的输出函数，将输入变量映射到 0 ～ 1 之间，其公式为：

$$S(x) = \frac{1}{1+e^{-x}} \tag{3}$$

Sigmoid 函数的图形如 S 曲线，见图 3。

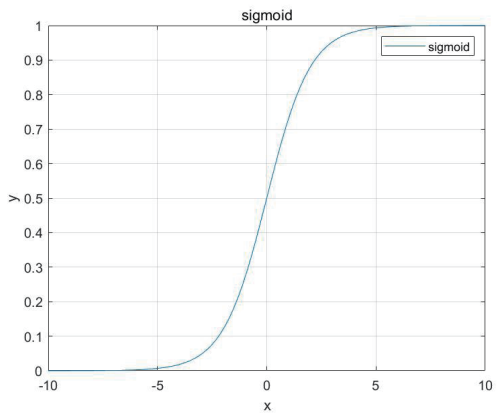

图 3 Sigmoid 函数图像

Logistic 函数是将以上两个函数结合起来，把因变量为 y 的线性回归表达式作为自变量代入 Sigmoid 函数中：

$$S(y)=\frac{1}{1+e^{-y}}=\frac{1}{1+e^{-(a^Tx+b)}} \tag{4}$$

两边取对数求解，得到：

$$\ln\frac{y}{1-y}=a^Tx+b \tag{5}$$

$$P(Y=1|X)=\frac{1}{1+e^{-(a^Tx+b)}} \tag{6}$$

$$P(Y=0|X)=\frac{1}{1+e^{(a^Tx+b)}} \tag{7}$$

这里就将线性回归的结果从 (- ∞ ,+ ∞) 映射到了 (0,1)，目标函数的取值空间发生了变化。由公式 $S(y)=\frac{1}{1+e^{-y}}$ 可知，当 $y \geqslant 0$ 时，$S \geqslant 0.5$；当 $y \leqslant 0$ 时，$S < 0.5$。将计算出的概率值与 0.5 对比，Logistic 回归就可以用来处理二分类问题。当面对多分类问题时，具体思想与二分类问题类似：假设每个样本对应且仅对应一个标签，将它们属于不同标签的概率看作几何分布，根据二分类的思想，类比得到多分类问题如下：

$$S_\theta(x)=\begin{bmatrix}P(y=1|x;\theta)\\ \vdots \\ P(y=k|x;\theta)\end{bmatrix}=\frac{1}{e_k^Tx}\begin{bmatrix}e_1^Tx\\ \vdots \\ e_k^Tx\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e_j^Tx}\begin{bmatrix}e_1^Tx\\ \vdots \\ e_k^Tx\end{bmatrix} \tag{8}$$

式中：θ 为模型参数，矩阵的每一行可看作该样本在训练时，每一个类别分类器的参数，值越大，表示属于此类别的概率越大。矩阵总共有 k 行，输出的 k 个数就表示样本属于该类的概率，总和为 1。通过以上思想，就能够利用 Logistic 回归处理多分类问题 [6-7]。

3 实验

3.1 MICE 对打分数据的处理

MICE 在进行缺失值插补时，需要在变量间建立回归方程模型，即要求变量间存在相关关系。通过第一节，确定了各一级指标与二级指标间的相关关系。由指标构建的假设已知，8 个一级指标间应该是相互独立的。因此，在使用 MICE 进行该数据集的插补时，将数据集拆分成 8 个子集，每个子集包括一个一级指标与其对应的 3 个二级指标。这样拆分的每个子集里的变量都存在相关性，能够使用 MICE 进行缺失值的插补。

表 2 数据集划分

Dataset1		Dataset2		Dataset3		Dataset4	
D1-1	D1-1-1	D1-2	D1-2-1	D1-3	D1-3-1	D1-4	D1-4-1
	D1-1-2		D1-2-2		D1-3-2		D1-4-2
	D1-1-3		D1-2-3		D1-3-3		D1-4-3

表 2（续）

Dataset5		Dataset6		Dataset7		Dataset8	
D2-1	D2-1-1	D2-2	D2-2-1	D2-3	D2-3-1	D2-4	D2-4-1
	D2-1-2		D2-2-2		D2-3-2		D2-4-2
	D2-1-3		D2-2-3		D2-3-3		D2-4-3

3.2 MICE 与均值、中位数插补方法的效果比较

专家打分数据的缺失值的插补应该满足以下几个要求：（1）尽量符合专家偏好与知识背景，体现各位专家的自身偏好一致性；（2）所有专家对同一指标的打分值，差异不能过大，体现专家较高的共识度。

3.2.1 专家自身偏好一致性

专家自身偏好一致性主要是指专家的打分偏好与知识背景带来的个体差异，主要体现在专家对一级二级指标打分的一致性上，若单独一个专家打分的相关性是比较高的，那么说明这个专家自身的打分是具有一致性的。因此，对于第一个要求，本节使用偏相关系数指标来评价。

表 3 三种插补方法的偏相关系数

均值插补法				
	D1-1	D1-2	D1-3	D1-4
二级指标 -1	0.820	0.713	0.761	0.656
二级指标 -2	0.668	0.750	0.835	0.811
二级指标 -3	0.695	0.748	0.770	0.823
	D2-1	D2-2	D2-3	D2-4
二级指标 -1	0.625	0.656	0.773	0.696
二级指标 -2	0.703	0.716	0.674	0.704
二级指标 -3	0.632	0.668	0.731	0.798
MICE				
	D1-1	D1-2	D1-3	D1-4
二级指标 -1	0.782	0.736	0.797	0.703
二级指标 -2	0.744	0.818	0.869	0.855
二级指标 -3	0.791	0.783	0.855	0.882
	D2-1	D2-2	D2-3	D2-4
二级指标 -1	0.661	0.725	0.789	0.760
二级指标 -2	0.749	0.712	0.734	0.729
二级指标 -3	0.699	0.623	0.738	0.865
中位数插补				
	D1-1	D1-2	D1-3	D1-4
二级指标 -1	0.811	0.704	0.748	0.639
二级指标 -2	0.669	0.740	0.828	0.797
二级指标 -3	0.693	0.737	0.757	0.802
	D2-1	D2-2	D2-3	D2-4
二级指标 -1	0.616	0.660	0.761	0.710
二级指标 -2	0.712	0.718	0.666	0.697
二级指标 -3	0.627	0.659	0.724	0.802

由表 3 可知，大部分情况下 MICE 插补后数据集中一级指标与二级指标的偏相关系数值高于均值插补与中位数插补，MICE 插补方法可以较好地保留数据间的相关性。此外，为对比其差异水平，本文计算了均值插补、中位数插补与 MICE 插补的所有偏相关系数的均值与方差。

表 4 表明，三种方法中 MICE 插补后偏相关系数的均值最大，说明指标间的相关关系总体保持得最好；MICE 插补后方差最大，说明指标间的相关性差异最大，更有利于区分指标间的重要程度。综上，MICE 插补的偏好一致性水平显著高于其他两种插补方法。

表 4 三种插补法的偏相关系数均值与方差

	均值插补	MICE	中位数插补
均值	0.726	0.767	0.720
方差	0.003 9	0.004 5	0.003 7

3.2.2 专家共识度检验

专家一致也称为专家共识，反映了所有专家对同一个研究对象评价结果的接近程度，它是能否应用专家评价结论的一个重要判别条件。本文采用 Kendall's W 协调系数来评价各种方法插补后的专家共识度。

用 SPSS 工具进行基于 Kendall's W 协调系数的非参数检验，并对比了 3 种插补方法下的 Kendell's W 协调系数 [8]，如表 5 所示。

表 5 Kendall's W 系数 1

	均值插补	MICE	中位数插补
Kendall's W 系数	0.516	0.615	0.523

由表 5 可知，三种插补方法中 MICE 插补的 Kendall's W 系数最高，中位数次之，均值最小。

以上结果表明，基于 MICE 的插补方法既能体现专家自身偏好一致性，又能使得 Kendall's W 系数保持在较高水平，保留专家共识度。基于 MICE 的插补方法适用于专家信息缺失情况的数据插补 [9]。

3.3 Logistic 插补对类别数据的处理

本节使用 Logistic 回归方法对分类变量进行插补，由 3.2 节分析可知，基于 MICE 插补后数据集在专家自身偏好、专家共识度上表现最好，因此将 MICE 插补后的数据集作为本节的待插补数据集。

图 4 为本文的 Logistic 回归训练分类模型并得到分类结果的流程图。首先选取模型的特征变量（因变量）与自变量，再按照类别型数据一列是否为非空，把原数据集划分为数据集 1 与数据集 2（其中数据集 1 为类别型数据列非空）。将数据集 1 分成训练集和测试集，通过调用 Logistic Regression 函数建立 Logistic 回归模型。模型建立好后，用准确率与混淆矩阵来评价模型的优劣。如果模型不能达到预期，则修改随机种子，调整模型，当准确率达到预期时，可看作得到满意模型。最后，将数据集 2 输入分类模型，得到分类结果并填入原始数据集中，完成插补。

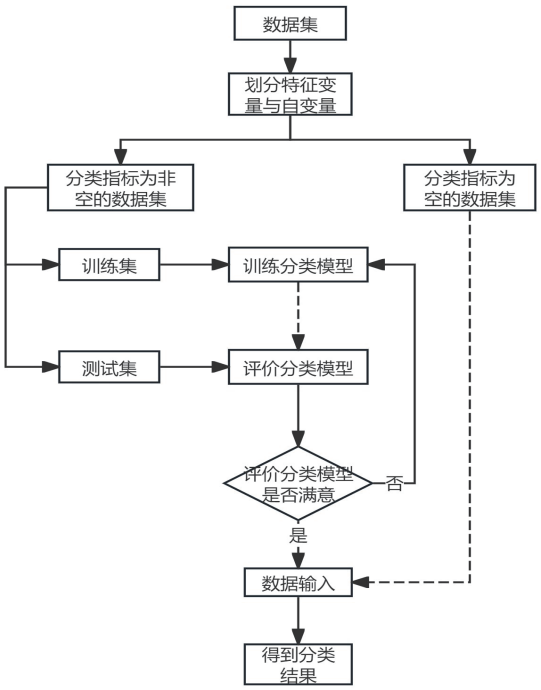

图 4 Logistic 回归流程图

3.4 Logistic 与众数、中位数插补方法的效果比较

对于分类变量，本节不讨论专家自身一致性，根据第二章对专家共识度的介绍，本节同样采用 Kendall's W 系数评价各种方法插补后的专家共识度。三种方法插补后的 Kendall's W 系数如表 6 所示。

表 6 Kendall's W 系数 2

	中位数插补	众数插补	Logistic 回归
Kendall's W 系数	0.567	0.571	0.602

由表 6 可知，Logistic 回归插补的 Kendall's W 系数最高，说明专家共识度最高。在只考虑类别型变量效用的情况下，该方法最适用于专家信息缺失情况的数据插补。

4 结论

专家评估项目中的缺失数据修复工作，对于提升项目评估数据网的完整度、支撑科学研究具有重要意义。本文选取某项目中 68 位专家对同一项目的打分数据为研究对象，区分数值型数据与类别型数据，验证基于链式方程的多重插补法（MICE）可有效修复数值型数据，通过计算偏相关系数及 Kendall's W 系数，证明其在专家自身偏好及专家共识度指

基于改进的 TF-IDF 标签权重算法的电商用户画像构建

白雨珂[1] 卢胜男[1]
BAI Yuke LU Shengnan

摘 要 在电商环境中，用户画像构建是为了更好地理解和满足用户需求而进行的重要任务。传统的 TF-IDF 标签权重计算方法无法很好地对标签权重进行调整，为了解决这一问题，提出基于 TF-IDF 算法的改进方法，旨在提高用户画像的准确性和个性化程度。融合相关系数矩阵，对相关性强的标签进行适当降权操作。不同类型的行为对标签信息产生不同的权重，并且标签的权重可能会随着时间的推移而衰减。因此，采用拟合记忆遗忘曲线模拟得到的兴趣遗忘曲线，对用户画像权重进行调优操作。实验结果表明，使用所提出的改进的 TF-IDF 算法构建用户画像的效果得到显著的提升。

关键词 电商；相关系数；标签权重；用户画像；TF-IDF 算法

doi：10.3969/j.issn.1672-9528.2024.08.011

0 引言

在当今数字化时代，随着电商平台交易量的不断增长，个性化推荐和精准营销逐渐成为各大电商企业竞争的关键要素。而实现个性化推荐的关键在于准确把握用户的兴趣和偏好，构建准确的用户画像则成为实现个性化推荐的基础。在这一背景下，基于 TF-IDF[1]（term frequency–inverse document frequency）标签权重算法在电商用户画像构建中发挥着重要作用。

TF-IDF 算法自提出以来，凭借着较高的准确率和召回率被广泛使用，但该算法仍然有很大的改进空间。张雷 [2] 提出 word2vec 结合 TF-IDF 权重计算方法进行个性化产品推荐。Shang 等人 [3] 提出一种标签与用户资源相结合的二分图

1. 西安石油大学 陕西西安 710065

标上优于均值插补与中位数插补；Logistic 回归法可有效修复类别型数据，对比 Kendall's W 系数，证明其在专家共识度上优于众数插补与中位数插补。两种方法针对不同数据类型，均可以有效描述专家群体一致性，MICE 对打分数据的插补能有效体现专家自身偏好性，实现对项目评估中缺失数据的修复工作，提升项目评估网的数据质量。

参考文献：

[1] 向南，豆亚杰，姜江，等 . 基于专家信任网络的不完全信息武器选择决策 [J]. 系统工程理论与实践 ,2021,41(3):759-770.
[2] 王晓静，王学丽 . 抽样调查中的不完全数据处理研究 [J]. 陕西科技大学学报 (自然科学版),2009,27(2):141-144.
[3] 岳勇，田考聪 . 数据缺失及其填补方法综述 [J]. 预防医学情报杂志 ,2005(6):683-685.
[4] 刘凤芹 . 基于链式方程的收入变量缺失值的多重插补 [J]. 统计研究 ,2009,26(1):71-77.
[5] 李锦绣 . 基于 Logistic 回归模型和支持向量机（SVM）模型的多分类研究 [D]. 武汉 : 华中师范大学 ,2014.
[6] 刘展，金勇进，韩显男 . 基于倾向得分匹配的缺失数据插补方法 [J]. 数学的实践与认识 ,2016,46(12):193-201.
[7] 刘燕 . 基于 Logistic 回归的近邻择优插补法 [D]. 天津 : 天津财经大学 ,2013.
[8]Sample Size Charts of Spearman and Kendall Coefficients[J]. Journal of biometrics & biostatistics,2021,12(2):1-7.
[9] YU W Z, WANG Y B. An efficient methodology for hardware Trojan detection based on canonical correlation analysis[J]. Microelectronics journal, 2021,115:105162.1-105162.7.

【作者简介】

李梓祎（2001—），女，贵州铜仁人，本科，助理工程师，研究方向：自动化技术、信息处理。

吴昕阳（1988—），女，山东淄博人，博士，副教授，研究方向：系统建模与综合评估。

（收稿日期：2024-05-13）