数据缺失及其处理方法综述

晔沙
（华南农业大学资源环境学院，广东广州，510642）

摘要：本文首先系统梳理了数据缺失的概念、产生原因及机制；然后对数据缺失问题常见的处理方法进行综述，比较不同处理方法的优劣以及各自的适应范围；最后详细介绍了数据填充效果评价的常用方法。得出结论：根据不同数据集的特点选择合适的处理方法，才能取得良好的处理效果；指出了随着大数据时代的到来，精准高效的处理海量的数据是数据缺失处理未来的发展方向。

关键词：数据缺失；填充方法；填充效果
DOI:10.16520/j.cnki.1000-8519.2017.18.028

Data deletion and summary of its processing methods

Ye Sha
（College of Resource and Environment, South China Agricultural University, Guangzhou Guangdong, 510642）

Abstract: In this paper, the concept, cause and mechanism of data loss are systematically combed out. Then, the methods of data deletion are summarized, and the advantages and disadvantages of different treatment methods and their respective adaptations are compared. In the end, the method of data filling effect evaluation is introduced in detail. Conclusion: according to the characteristics of different data sets, the proper treatment method can be obtained. It is pointed out that with the advent of the era of big data, the accurate and efficient processing of massive data is the future development direction of data deletion.

Key words: missing data; Filling method; Filling effect

0 引言

数据缺失的现象普遍发生，由于调查失误、数据录入错误、机器损坏等原因容易导致数据收集的不完整，存在数据缺失的现象。缺失数据的存在往往对数据分析和研究推论带来困难，造成分析结果偏差，影响决策准确性。传统的忽略缺失值或者删除缺失记录的方法存在局限性，处理不当时会破坏数据的关联性，影响信息的利用率。如何有效处理缺失数据，充分利用数据信息，准确的反映研究对象的特征，达到研究目的，成为统计研究中的热点与难点问题。本文将系统梳理数据缺失的机制与原因，就当前国内外数据缺失的处理方法进行综述，并介绍数据填充效果评价的常用方法。

1 数据缺失的介绍

1.1 数据缺失的概念

数据缺失是指在数据采集时由于某种原因应该得到而没有得到的数据，它指的是现有数据集中某个或某些属性的值是不完全的。

1.2 数据缺失产生的原因

（1）在数据存储过程中，由于机器的损坏造成数据的存储失败。

（2）在数据录入过程中，人为疏忽造成数据录入失败。

（3）在数据收集过程中，客观条件限制造成的信息无法获取。

1.3 数据缺失的机制

Little 和 Rubin 针对缺失数据定义了 3 种不同的缺失机制。

（1）完全随机缺失（Missing Completely At Random，MCAR）：数据缺失现象完全是随机发生的，和自身或其他变量的取值无关。

（2）随机缺失（Missing At Random, MAR）：数据缺失现象的发生与数据集中其他无缺失变量的取值有关。

（3）非随机缺失（Missing At Non-Random, MANR）。数据缺失的现象不仅和其他变量的取值有关，也和自身的取值有关。

2 数据填充方法

关于缺失数据在国内外研究的比较系统而深入，其处理方法更是被大量研究并应用。综合国内外研究成果，缺失值处理的方法总体来说可以分为三类：删除法、数据填充和不处理。

2.1 删除法

删除法就是将存在缺失信息的记录或者对象删除，从而得到完整的数据集合。常用的删除方法有个案删除、配对删除与列表删除。个案删除就是若任何一个记录含有缺失值，则删除该记录，仅对其余完整数据进行分析。该方法多用于完全随机缺失的数据集。配对删除就是若配对的两个属性之一或者两个都是缺失值时，将其同时删除后再进行分析。该方法多用于因子分析、重复测量设计以及回归相关分析中。列表删除就是若任何一个属性变量含有缺失值，则删除整列属性数据。删除法的优势在简单易行，在被删除的数据量占整个数据集的比例非常小的情况下，是非常有效的。然而删除法的局限性也很大，首先以减少原始数据来换取数据集的完备，会浪费大量有用信息；其次当缺失率较大或者总数据量较少时，删除少量对象就将影响到数据集信息的客观性，导致数据发生偏离，甚至得到错误的结果。

2.2 数据填充

数据填充就是采用一定的方法，对数据资料的缺失值确定

一合理的估计值，然后替代缺失值以使得数据完整。数据填充的方法相对删除法有相当大的优势，保证了信息的完整性。通常根据对每个缺失值构造估计值个数的不同，将填充方法分为单值填充和多重填充两大类。

2.2.1 单值填充

单值填充是构造出一个估计值来填充缺失值，在完成缺失值的填充后，再对填充后的完整数据集进行相关的数据分析工作。单值填充可分为两类，第一：建模填充。即通过常用的统计模型进行预测，常见方法有均值填充、随机填充、EM 算法、回归填充等。第二。模糊填充。即采用一种算法，根据辅助信息构造相应函数，计算离缺失记录最邻近的记录，以此记录的数据作为填充值。使用该方法时需判断函数的合理性，常见方法有：最近距离填充、热卡填充、冷卡填充等。

(1)建模填充

建模填充的常用方法如下：均值填充，随机填充，回归模型填充，EM 算法填充，加权调整填充。

①均值填充

均值填充法就是用已观测数据的均值作为缺失值的替代值。注意这仅在变量服从或近似服从正态分布的情形下适用，若分布为偏态，则应以中位数或众数替代均值进行填充。最典型方法为总均值填充法和组均值填充法。总均值填充，是用样本总体均数代替该变量所有缺失值。组均值填充法，利用辅助信息，将样本分为若干组，计算各组变量的均值。对于缺失的数据，用其所在组的所有观测数据的均数来填充。均值填充方法通常改变了数据的不确定性，低估填充变量的方差。因此一般情况下均值填充适合比较简单的完全随机缺失的数据集，不适合较复杂的估计。

②随机填充

为了避免均值填充中替代值过于集中，影响数据集不确定性的缺点，随机填充方法有着不错的填充效果。随机填充就是随机抽取已有的观测值以填充数据缺失。随机填充也可分为总随机填充和分层随机填充。若能根据分类变量信息，将样本进行事先分层，再在各层内进行随机抽取，则会取得更好的调整效果。随机填充避免了均值填充扭曲变量分布的特点，使得替代后的分布更接近真实值分布。

③回归模型填充

回归模型填充就是指通过建模，以模型预测值作为缺失值的估计值，其中最典型的就是线性回归填充。其基本思想是通过建立响应变量 Y 关于自变量 Xi(I =1,2,….,m) 的回归模型来预测 Y 的缺失数据，那么第 k 个缺失值的填充值可表示为：

$$z_k = a_0 + \sum_{i=1}^{m} a_i X_{ik} + \varepsilon_k$$

若各变量之间的回归关系比较显著，那么通过回归模型得到的估计值往往更接近于真实值，但构造和评估回归模型费时繁琐，需要对模型进行评价，因此多用于对重要变量缺失值的填充。利用回归模型进行填充时，主观增大了变量的相关关系。当变量不是线性相关或预测变量高度相关时会导致有偏差的估计。

④ EM 算法填充

EM 采用平均值与协方差矩阵对数据缺失属性值不断进行迭代求精。主要用于估计不完整样本的概率密度函数的参数，算法目的是期望值最大化，它的收敛速度与数据缺失率有关，缺失率高则收敛速度慢，反之则相反。相比于传统的填充算法其最大的优点在于大样本条件下，它能非常简单的执行并且能通过稳定的计算步骤找到全局最优解，具有较高的数据填充精度。但此方法没有考虑局部数据的相似性，利用整个数据集对缺失数据进行填充，当数据量大时严重影响算法的执行速度，而且算法初始值的选择不当，将影响算法的收敛速度和稳定性。

⑤加权调整填充

加权调整是当出现缺失单元时，用某种方式把缺失单元的权数分解到非缺失单元（即观测数据） 身上，通过增大样本中有观测数据的权数，以减小由于缺失数据可能对估计量带来的偏差。

几种主要的加权调整方法包括加权组调整法、再抽样调整法、事后分层调整法、迭代调整法、校准法、双重稳健加权法等。加权是一个减少偏差的比较简单的措施，但由于丢弃不完整单位信息，并且没有提供一个内在的方差控制，所以在样本量比较大的时候，易出现错误。

(2)模糊填充

常见的方法有模糊填充的方法有最近距离填充、热卡填充、冷卡填充等。

最近距离填充法是利用辅助信息，定义一个测量单元间距离的函数，筛选出离缺失单元最近距离的回答单元，以该回答单元作为填充值。用于定义赋值单位的距离函数可以有很多类型，马氏距离就是其中一种。由于距离函数的类型不同，用最近距离函数得到的填充值具有伪随机性，并且距离函数的是否合理也必须检验，这给最近距离填充法的使用增加了难度。

热卡填充法首先对数据进行分层，然后在每层中按照某种顺序对单元排序，对有数据缺失的单元利用最近顺序的数据进行填充。该方法存在的问题是填充值的选择是由辅助变量决定的，利用不同的辅助变量进行分层排序将会得到不同的序列，对于某一个缺失值来说可能采用的填充值也会不同。并且，在大型数据集中，难以界定变量相似性，在填充数据时容易导致缺乏准确性。因此，应该选择与研究变量性质高度相关的排序变量，使得排列位置相邻的单位在研究性质上也相近。

冷卡填充法是相对于热卡填充而言的，指的是填充值不是从当前的调查获取。而是从以往的调查和历史数据中获取的。

最小迭代算法(IBLLS)将数据集分别在水平与竖直方向按照 k 个最相似记录与 k 个最相似属性的思想进行分片，然后分别在每个分片子集中对缺失数据进行填充。此方法充分考虑了数据之间的相似性，提高了数据填充的精度。然而数据集较大时，相似属性与相似记录的划分需要大量的计算时间，计算和存储的时间成本较高，并且 k 值的选择，也是难以解决的问题。

另外支持向量机、遗传算法和神经网络模型的数据填充算法。这些方法使用多种算法与模型相结合的方式进行数据填充，具有较高的填充精度，但是时间复杂度较高，不适合大数据量的处理。

2.2.2 多重填充

单值填充所填充的数据都是唯一的，所以填充后的数据集不能表现出原有数据集的不确定性，降低了抽样误差，对样本分布存在不同程度的扭曲，进而导致结果产生较大的偏差，使得总体推断有失偏颇。于是多重填充的思想应运而生，并逐渐完善形成了比较系统的理论体系。

近十年里，逐渐形成了两种最常见的多重填充的方法，分别是联合模型法和全条件定义法，其主要思想是：(1)填充：为每个缺失值构造一套可能的估计值，这些值反映了缺失模型的不确定性，这样就形成若干个完整数据集。(2)分析：对每个完整数据集分别使用相同的方法分析。(3)合并：综合不同填充数据集的分析结果，把这个分析结果作为缺失值的估计值。多重填充方法的实质是一种模拟方法，模拟不同条件下的估计值的分布，反映了缺失模型的不确定性，以此来提高估计的有效性和可靠性。

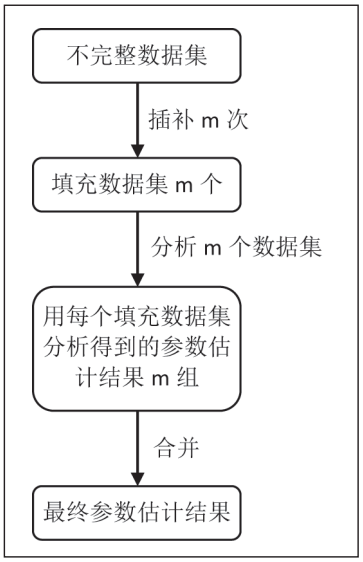

图 1　多重填充处理流程图

多重填充方法与单值填充的方法相比优势明显，第一：多重填充方法将辅助信息合理的利用起来，通过提供多个替代值的方式，保持了原数据集的不确定性。第二：多重填充方法能够尽可能接近真实情况下去模拟缺失数据的分布，在这样的条件下能够尽可能的保持变量之间的原始关系。

2.3 不处理

数据填充的过程中，从一定的程度上改变了原始数据的分布特点或者属性之间内在的联系，如果填充方法不当，则会引入新的噪声，导致错的结果。因此，不对缺失数据做任何处理，直接在缺失数据集上进行数据挖掘将最大程度保证数据的原始性。但这种方法的缺陷在于没有先验知识的情况下，进行数据处理，处理时长增加，处理精度也难以保证。

本文阐明了数据缺失三种处理方法删除法、数据填充、不处理的理论，分析不同方法的优劣，并讨论了各自的适用范围。结果显示，不同的填充方法各有其适应的环境。因此在数据缺失处理时应充分了解数据特点，挖掘辅助信息，确定填充方法的可行性，降低决策风险。

表 1　数据缺失处理方法

方法			优点	缺点	适用范围
删除法			简单易行	浪费大量信息 易导致数据偏离	缺失数据 所占比例小
数据填充	单值填充	建模填充	计算稳定 填充精度较高	模型构造 较为繁琐	数据分布 特性明确
		模糊填充	充分利用辅助信息	难以界定 变量相似性	变量之间 关联性强
	多重填充		保持原数据集的不确定性； 保持变量之间的原始关系	处理过程复杂	总数据量较小
不处理			简单易行	处理效果较差	缺失数据 所占比例小

3 数据填充的效果评价

为模拟分析比较不同方法的填充效果及适用条件，数据填充的实验通常由以下步骤展开。首先，根据原始的完整数据集建立适当的数学模型。其次，以完整数据集为基础，模拟不同缺失率的完全随机缺失数据集，分别采用不同的数据填充方法对不同缺失率的数据集进行填充。最后将填充后的数据集与原始数据集进行比较，对数据填充的效果进行评价。数据填充效果的评价通常从参数角度与拟合角度两个方面进行考察。

3.1 参数角度

主要考察填充后数据集与原始数据集相关参数，比较对数据集的模拟情况以及对真实值模拟情况。

(1)数据集模拟

均值：$\bar{x}=\frac{\sum x_i}{n}$ $\bar{x}=\frac{\sum x_i}{n}$

标准误：$\sigma=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$

标准方差：$s^2=\frac{1}{n}(x_i-\bar{x})^2$

其中 $\bar{x}$ 为平均值，x_i 表示第 i 个数据，n 为总数据量。均值、标准误、标准方差在统计中常用来表示数据的集中趋势和离散程度，在一定程度上可以反应数据集整体分布情况，并描述数据集的变异程度。

(2)真实值模拟

平均绝对离差与标准平均离差平方和：

$$MAD=\sum_i|\hat{y}_i(miss)-y_i|/n_0$$

$$RMSD=\sqrt{\left[\sum_i\frac{(\hat{y}_i(miss)-y_i)^2}{n_0}\right]}$$

其中，$\hat{y}_i$ Y $miss$ Y 表示第 i 个缺失值的估计值，y_i 是对应的真实值，n_0 为缺失值总数。平均绝对离差与标准平均离差平方和反映了估计值相对于真实值的偏离程度，其值越小，表明估计值与真实值的偏离程度越小。

后验差检验：

$c=s_2/s_1$，其中 S1 为原始数据方差，S2 为残差方差。后验差比值反映了残差相对于标准偏差偏离的程度，后验差比值越

（下转第 60 页）

3 系统实现

多媒体教室端的核心设备——中央控制器主机在选型上需对技术成熟、最佳性价比原则予以遵循，可采用基于 TCP/IP 协议网络的中央控制器。该系统由网络中控套机（含主机与手动按键面板）、网络总控服务器（主控制室配备、网络管控配备）共同组成。通电后，系统会自动与远程管控中心的总控服务器建立连接，利用总控服务器与配套的管控系统软件，远程、双向监控与管理各多媒体教室。在主控制室，管理人员可利用远控软件开启教室中的电子讲台设备，同时，连接其他相关设备的控制。

3.1 远程开启 / 关闭多媒体教室

利用远程管理系统中的网络控制功能模块，启动多媒体教室，在课前让多媒体设备进入开机状态；同时，网络控制模块还具有关闭多媒体教室的功能，在教师课后忘记关闭设备之时，多媒体教室管理人员可以利用远程管理系统执行快速、远程关闭操作。

3.2 远程监管多媒体教室计算机

在接口处连接显示屏，让管理人员在主控制室实现对教师授课内容的查看。出现状况时，主控室的管理人员无需赶赴现场便可及时接管并控制多媒体教室端的计算机。

3.3 远程监测设备使用状态

利用中央控制系统，管理人员可以对多媒体设备的使用情况进行实时监测，随时掌握各项设备已使用的时间、使用次数以及使用状态等信息，尤其是对设备中易耗品的使用进行实时与实地监测，当使用量达到临界值之时，系统自动开启报警功能。

3.4 管理设备资产

智能化与动态化管理全部多媒体设备资源应是高校多媒体教室中央控制系统的重要任务之一。集控系统对多媒体教室中的所有投影仪、教师机等设备软硬件信息进行准确的搜集，据此生成 Excel 资产报表，让学校资产管理部门收集多媒体教室、设备等资产信息。一旦计算机软硬件配置出现问题，或投影仪灯泡发生损坏，系统会向主控制室发出警报，保证管理人员对设备资产使用情况的实时监视，更好地实现资产的管理工作。

4 结语

多媒体中央控制系统的设计与应用能够向高校多媒体教师的使用与管理创造极大的便利，在提供有力的技术保障于高校教学活动开展的同时，进一步提高管理人员的工作效率与水平，减小工作负荷。教师、教育工作者以及相关管理人员应在日常的多媒体教学应用中培养观察意识，进行多媒体中央控制系统更多应用、问题解决方法以及发展走向的持续性探索，保证多媒体技术长期朝向信息化与智能化方向发展，不断优化教学环境、提高教学质量。

参考文献

[1] 刘和连 . 多媒体教室网络智能化中央控制系统的设计与建设 [J]. 中国医学教育技术，2015(2)：157-160.
[2] 胡彦玲 . 高校多媒体教室中央控制系统的研究与设计 [D]. 华东师范大学，2010.

（上接第 67 页）

小，表明估计值偏离真实值的程度越小，就越接近真实值。

3.2 拟合角度

以纵轴为数值，横轴为按顺序排列的估计值与真实值，作估计值的分布折线图，从图形上以直观的形式与对应真实值的分布折线图做比较。折线图不仅可以反映缺失量、数值的变化情况，也能显示估计值与真实值变化的趋势，以此了解估计值对原始数据集的拟合情况。

4 总结与展望

当前在调查研究领域，有关缺失数据的理论探讨、实务处理均逐步趋向成熟，而且有关缺失数据填充的应用范围也逐渐涉及到各个研究领域，方法越来越多元化。本文梳理了数据缺失的概念、原因及机制。对常见处理缺失值问题方法进行综述，比较不同处理方法的优劣以及各自的适应范围。从参数角度与分布角度详细介绍了数据填充效果评价常用方法。

（1）填充方法有各自的适应范围。目前常用的填充方法有三种：EM 算法、多重填充、加权调整填充，然而这些方法也有其适应范围，实际选择填充方法时应充分了解原始数据的背景，根据不同填充方法的特点， 挖掘辅助信息，确定填充方法的可行性，降低决策风险。

（2）数据填充效果的评价需从多角度考虑。从参数角度可以反映估计值对数据集及真实值的拟合情况，从分布角度可以更直观显示变化趋势。实际评价时需从多个角度考虑，结合多种方法，得到更全面的结论。

（3）物联网、社交网络及电子商务技术的兴起和发展，数据正以前所唯有的速度增长，在物联网领域、电子商务领域，每天产生的数据量达到了 TB 级别。随着大数据时代的到来，精确高效的处理来源多、数据结构复杂、时空特性与动态连续性强的海量数据，将是数据缺失处理未来的发展方向。

参考文献

[1]Rubin D B. Inference and missing data[J]. Biometrika,1976,63(3):581-592.
[2] 金勇进 . 调查中的数据缺失及处理（ⅰ）——缺失数据及其影响 [J]. 数理统计与管理，2001,20(1):59-62.
[3] 花琳琳 . 不同缺失值处理技术的模拟比较 [D]. 郑州大学，2012.
[4] 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，2005,21(6):683-685.
[5]Little R J A, Rubin D B. Statistical Analysis with Missing Data, 2nd Edition[J].2014.
[6] 沈琳，陈千红，谭红专 . 缺失数据的识别与处理 [J]. 中南大学学报（医学版），2013(12):1289-1294.
[7]Molenberghs G, Kenward M G. Missing data in clinical studies[M].2007.