缺失数据处理方法研究综述

熊中敏,郭怀宇,吴月欣
上海海洋大学 信息学院,上海 201306

摘 要: 大数据时代,数据爆炸式的增长,数据获取变得更容易的同时数据缺失现象也更加普遍。数据的缺失极大地降低了数据的实用性。数据缺失问题的处理成为大数据处理的热点研究课题。介绍了数据缺失问题的研究意义和国内外研究现状。系统地分析了造成数据缺失的原因,对数据缺失问题进行了分类。对近年来国内外缺失数据处理方法进行了综述,总结了各自优缺点、适用范围、效果评价指标。重点阐述了回归填充、聚类填充等填充方法。对缺失数据处理方法领域进行了总结与展望。

关键词: 缺失数据;缺失分类;填充方法;方法比较;效果评价
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.2101-0187

Review of Missing Data Processing Methods

XIONG Zhongmin, GUO Huaiyu, WU Yuexin
School of Information, Shanghai Ocean University, Shanghai 201306, China

Abstract: In the era of big data, with the explosive growth of data, data acquisition has become easier and data missing has become more common. The lack of data greatly reduces the usefulness of the data and the handling of data missing has become a hot research topic in big data processing. The article first introduces the research significance of the problem of missing data and the current research status at home and abroad, then analyzes the reasons for the missing data systematically, classifies the problem of missing data, and reviews the methods of processing missing data at home and abroad in recent years, summarizes their respective advantages and disadvantages, scope of application, and effect evaluation indicators which focus on the regression filling, cluster filling and other filling methods. Finally, it summarizes and looks forward to the field of missing data processing methods.

Key words: missing data; missing classification; filling method; method comparison; effect evaluation

随着科技信息技术的日新月异,各研究领域对于数据的收集、存储以及处理技术等已经基本成熟。日复一日的海量数据收集工作,使得各领域积累了大规模的数据存储量。据统计,全球各领域数据总量正以每年40%左右的增速大规模增加,到2020年全球数据总量规模可达到40 ZB[1-2]。大数据时代的到来,对于各领域有效地利用大数据提出更高的要求,特别是数据挖掘领域,数据的质量决定着是否能在海量复杂的数据中挖掘出有价值的知识[3-5]。因此面对鱼龙混杂的海量数据,如何通过预处理等手段提高数据的可用性变成急需解决的重要问题。针对不同质量问题的数据采取适合的预处理手段可以改善数据的整体质量。目前,数据缺失作为数据质量问题的重要因素之一,变得难以避免。无论是现实调查还是科学研究中,大部分数据集都出现缺失问题,极大地影响了后续研究工作的准确性。不论是忽略还是简单删除这些缺失数据都会使得原数据集信息量发生改变,最终影响实验工作的进行。因此缺失数据填充方法成为目前的研究热点,如何针对不同的缺失类型进行有效填充是接下来的研究重点。本文介绍了缺失数据处理方法的国内外研究现状,整理了数据缺失原因并进行分类,系统地对最新国内外数据缺失的处理方法进行综述对比,并介了绍常用的数据填充效果评价方法,最后对未来研究发展趋势做出了展望。

1 国内外研究现状

数据预处理中数据缺失问题一直是需要克服的困

基金项目: 国家自然科学基金(41501419);上海市地方院校能力建设项目(19050502100)。
作者简介: 熊中敏(1971—),男,博士后,副教授,CCF会员,主要研究方向为大数据处理、数据仓库与数据挖掘,E-mail:zmxiong@shou.edu.cn;郭怀宇(1996—),男,硕士研究生,主要研究方向为数据仓库与数据挖掘、大数据分析;吴月欣(1994—),女,硕士研究生,主要研究方向为推荐系统、信息处理。
收稿日期: 2021-01-11 **修回日期:** 2021-04-28 **文章编号:** 1002-8331(2021)14-0027-12

难,为此国内外学者对缺失问题进行了深刻研究,贡献出了许多的经验成果[6-8]。本文在查阅大量国内外文献的基础上,对数据缺失问题的国内外的研究现状总结如下。

1.1 国外研究现状

20世纪前期国外就已经开始对数据质量问题进行研究[9-10]。最早是Bowky在1915年对某项调查结果的误差来源进行了研究分析,提出了数据缺失问题。后来,Deming对调查误差进行了各种因素分析,进一步总结了数据缺失问题,其中包括因无回答造成的缺失。20世纪40年代末,数据缺失问题的研究掀起热潮,专家学者提出了各类缺失值的解决方法。这些方法可以大致分为两类:事前防范和事后处理。前者一般是通过大量收集来提高数据的完整度,但这种方法无法保证一定能收集到完整数据。后者通常是在已有数据的基础上进行处理,使其完备化。因此事后处理更符合数据缺失处理的研究方向,也更受欢迎。1940年,Deming和Stephan提出抽样概率的倒数加权法解决单元缺失情况[11]。1949年,Politz和Simmons对这种加权法进行了改进提出了经典的PolitzSimmons调整法[12]。

著名统计学家Yates因实验数据缺失过多无法完成数据分析而提出填补缺失值的方法[13],该方法在对方差分析中表现出不错的效果[14]。然后填充方法成为研究热潮,随后出现了均值填充、回归填充、聚类填充、热卡填补、多重填充等许多经典方法。在前人方法的基本理论基础上,各领域学者结合自身领域数据特点,进行深入研究后提出大量改进方法。1977年,Dempster等人提出期望极大化算法(Expectation Maximization),该方法成为缺失数据处理领域的一个重要里程碑[15],此后许多方法都在它基础上进行研究更新。1978年Rubin提出多重填补方法[16],这又是一次重大突破,相比单值填补,该方法表现出更好的填充效果[17]。1984年Kalton等人根据热卡填补法的思想,提出最近邻填补方法[18],该方法的填充思想影响了后来许多算法。

进入21世纪,数据缺失的处理方法已经走向成熟,鲜少有全新的填充思想被提出,大部分都是基于当下领域的改进及应用。如2003年,Batista等人对监督学习的四种缺失数据处理方法进行了分析比较,证明了k最近邻填补算法在填补手段上的性能优越性[19]。如2018年Zakaria等人利用环境温度和湿度的监测数据来评估四种填补方法(均值填充、回归填充、多重填充和最近邻填充)[20]。2019年Little等人对缺失数据的最新统计处理方法进行了前面分析,并提供了实际应用信息[21]。

1.2 国内研究现状

国内学者对缺失值处理方法的研究相对比较晚,基本上都是在国外已有的先进理论上进行改进、对比完善,大多缺乏原创性理论。如2000年,金勇进等人通过模拟实验对几种缺失值填补方法进行比较,发现均值填充更符合真值,而随机回归填补更能保持样本分布[22]。2009年,金勇进等人出版的《缺失数据的统计处理》,详细地讨论了各类缺失问题以及解决办法[23]。

2010年,邓银燕通过仿真实验研究讨论了数据填充方面的主要方法性能,其中包括均值填充、随机填充、期望最大化(EM)填充、线性回归模型填充、多重填充等方法[24]。实验表明不同方法对于不同缺失率的数据填充效果不尽相同。2014年罗永峰等人根据钢结构检测数据缺失的形成机制,提出基于最小二乘原理以回归分析理论为基础的填充方法[25]。2020年杨弘等人针对混合型缺失数据比较了一些缺失数据处理方法的特点以及在实际应用中的评价效果[26]。后来许多国内学者在经典填充方法的基础上,根据自身领域数据特点设计出各种相适应的填充算法。

2 数据缺失问题

数据质量问题一直是影响实验研究的重要因素之一。而数据缺失问题作为数据质量问题中的关键元素已经普遍存在。例如常见的机器学习领域UCI数据库中,出现数据缺失的数据集已达到40%以上[27-28]。数据缺失的普遍存在已经影响到正常的数据分析及研究。国内外学者开始对数据缺失问题进行深入研究,从产生原因到分类,再到解决办法。本章详细介绍了数据缺失问题产生的原因,根据不同标准对缺失问题进行了分类,为后面处理方法的介绍奠定了基础。

2.1 数据缺失的原因

数据缺失常发生在数据的采集、运输、存储等过程中。如在各领域数据采集中,会存在一些数据无法获取或者人工操作不当而丢失的情况,或者在数据传输、存储等转移过程中发生丢失等等[29-30]。因此对数据缺失原因总结如下:

(1)数据在采集过程中的缺失。客观条件的限制,如历史条件下,设备的局限导致无法获取完整的信息。

(2)数据在运输过程中的缺失。数据的运输转移需要靠人来完成,因此人为操作、判定的失误会导致数据错误或者丢失。

(3)数据在存储过程中的缺失。由于存储介质发生故障及损坏而导致的数据缺失;以及存储过程中对数据进行压缩而导致丢失。

2.2 数据缺失的分类

数据缺失原因的不同产生了不同缺失类型[31],为了能更加有效地应对数据缺失问题,需要对数据缺失类型进行分类,从而能更有针对性地提出解决办法,使得结果更合理准确。本节从缺失模式和缺失机制两个方面对缺失类型进行了分类。

2.2.1 缺失模式分类

因数据缺失而在数据集中产生的缺失结构叫缺失模式[32]。缺失模式可以用来反映数据集中缺失数据之间的关系。目前缺失模式大致分为四种:单变量缺失模式、多变量缺失模式、单调缺失模式、一般缺失模式[33]。

(1)单变量缺失模式

单变量缺失模式是指单属性维度存在缺失值,即所研究数据集中只有一个属性维度存在缺失值,其余属性维度数据完整。

(2)多变量缺失模式

多变量缺失模式是指中多属性维度含有缺失值,即所研究数据集中有一个及以上属性维度存在缺失值。

(3)单调缺失模式

单调缺失模式是指所研究数据集在多属性维度含有缺失值的基础上,缺失数据形成的矩阵进行排列变换后能呈现单调层级模式。

(4)一般缺失模式

一般缺失模式简单点说就是所研究数据集中缺失数据分布在不同属性之间,并且毫无规律可循。这是目前最常见的缺失模式。

2.2.2 缺失机制分类

缺失数据和完整数据之间的关系称为缺失机制[34]。缺失机制的意义在能通过完整数据帮助处理缺失数据。缺失机制大致分为三种:完全随机缺失(MCAR)、随机缺失(MAR)、非随机缺失(NMAR)。

(1)完全随机缺失(Missing Completely At Random,MCAR)

完全随机缺失指数据缺失是随机发生的,与自身属性以及其他属性取值无关。例如研究数学、语文和英语三个属性时,数学属性的缺失与语文和英语两个属性无关,它是完全随机缺失。目前来说,完全随机缺失并不常见。

(2)随机缺失(Missing At Random,MAR)

随机缺失指数据缺失只和完整属性取值有关[35]。例如研究数学和语文两属性时,已知数学属性的缺失和语文属性相关,则可以认为这是随机缺失的。

(3)非随机缺失(Not Missing At Random,NMAR)

非随机缺失指数据缺失不仅与自身取值有关而且与完整属性取值也有关,这种缺失是不可忽略的缺失[36]。由于隐私敏感等问题,隐去某些属性值,这就是非随机缺失。

3 缺失值处理方法

目前对于缺失值的处理方法基本分为三类:删除,填充,不处理[37]。采用什么样的处理方法要因数据集缺失情况以及研究内容而定,本章介绍了目前缺失值处理的几类解决办法,其中详细阐述了数据填充方法以及研究进展。最后总结了各类缺失值处理方法的优缺点以及适用范围。

3.1 简单删除法

最原始的缺失数据处理方法主要有简单删除法,此方法就是将包含缺失值的数据对象、数据属性、成对变量进行删除[38]。

(1)对象删除

对象删除指当数据集中某个研究对象的数据记录中存在丢失时,直接删除该对象。该方法仅适合于缺失对象极小,否则会使得数据集因丢失过多的信息而造成不完整,从而影响后续实验结果的准确性。

(2)属性删除

属性删除指当数据集中某属性存在缺失时就直接删除该属性。这种做法虽然保留了研究对象的个数,但是丢失对象的一些属性信息,若含缺失值的属性过多,就会造成删除过度,后续实验研究将毫无意义。

(3)成对删除

成对删除指配对的两个变量之间,若有一方存在缺失值,就将两个变量同时删除然后再进行相关分析。

综上这类方法操作过程简单,速度快,但很难适用众多领域的缺失数据集。当数据量特别大,缺失对象与数据集中的数据量相比微不足道时,这种方法非常有效,它既解决了数据缺失的问题,又不会影响数据集的信息量以及研究结果。然而,当数据集中缺失数据大量存在时,简单地删除缺失对象以及它所包含的信息就会影响整个数据集的质量,造成数据资源的浪费,丢掉了可能存在的有价值的信息,对后续研究造成影响,使得研究结果无法保证客观性以及结果的正确性。如陈景年在选择性贝叶斯分类算法研究中,为了使朴素贝叶斯分类器的分类效果达到预期目标,选择删除数据集中的冗余属性,使剩余的属性尽可能地满足独立性假设条件,最后达到了预期效果[39]。

3.2 权重法

权重法的使用前提是数据缺失类型为非完全随机缺失情况下,通过logistic或probit等方法将缺失单元的权数分配到完整单元上,从而增大完整单元的权数以减小缺失单元带来的损失。这种方法一般用来处理单元无回答的缺失问题。但是权重法不适合多属性缺失的数据集,因为多属性缺失则会增大计算难度,准确性降低。

3.3 填补

目前针对数据缺失问题国内外学者们提出了多种填补方法,基本上可分为两类:统计学方法和机器学习方法[40]。统计学方法大多是基于数据集本身作出假设,然后利用原数据集对缺失数据进行相应填补。这类方法没有考虑数据对象本身的类别,填充值往往受其他类别对象的影响,填充结果准确性较差,常见的方法有EM(Expectation Maximization)填充算法、回归分析法、多重插补等。机器学习方法,一般是先对缺失数据集进行分类或聚类,然后进行填补。这类方法是随着近年来机

器学习的热潮兴起的。代表性方法有:K最近邻填补、K-means填补、贝叶斯网络等等。其中分类方法以缺失属性为目标进行分类,然后在每个类别内进行填补,但缺失属性过多时容易导致所分类别过多,效率低下;聚类方法则是先将数据对象聚类,划分成多个簇,根据簇内相似对象进行填补,缺失属性的多少不会影响簇的个数,这类方法适用范围广,也是目前研究的热点。本文将现有的填充方法划分成以下几种方法:

(1)人工填写(Filling Manually)

人工填写法就是数据集创造者自身根据自己对数据集的了解自行填充缺失值。这种填充方法对于数据集创造者来说无疑是最快最准确的方法,但是若是数据规模大,缺失数据过多时,不仅费时而且容易出现错误,并且对于其他使用者来说这种方法适用性不大,基本上可行性很低。

(2)均值填充(Mean/Mode Completer)

均值填充法就是将现有数据的对应属性均值填充给缺失值,但要注意数据变量需要服从或者近似服从近态分布,否则用该属性下的众数或中位数填充缺失值[41]。简单来说就是先判断缺失值的数据类型,然后根据数据类型采取不同的填充方法,将同属性下其他对象的平均值填充给数值型的缺失值;或采用众数原理将同属性下取值次数最多的值填充给非数值型缺失值。还有一种相似的方法叫分层均值填补,该方法是在填补之前对数据集进行分层,使得相似数据聚集同一层,然后在每层内采取均值填充。以上两种均值填充方法,基本思想是相近的,都采用了均值填充,只不过再具体实现上有所差别。均值填充法是目前填充方法内使用最多,同时基于这种方法延伸最多的方法。但均值填补的缺点是仅仅适合数据规模小,缺失数据少的简单研究,不适应较复杂的分析研究[42-43]。

(3)EM填充(Expectation Maximization Imputation)

20世纪70年代后期,Dempster等人最先提出了EM算法(最大期望算法)[15],该方法经过两个步骤交替进行计算。

第一步是计算期望(E),利用对隐藏变量的现有估计值,计算其最大似然估计值。

第二步是最大化(M),最大化在E步上求得的最大似然值来计算参数的值。M步上找到的参数估计值被用于下一个E步计算中,这个过程不断交替进行。这是当时最有效处理缺失数据的方法。

后来Ghahramani等人对缺失数据进行了研究分析,为了解决因数据稀疏而导致数据最近邻寻找不准确的问题,提出了基于期望最大化的缺失数据处理方法EMI(Expectation Maximization Imputation)[44-45]。EMI算法是一种求解参数最大似然估计的迭代算法[46-47]。与一些传统的填充算法相比,EM算法在数据规模非常大时,它的算法执行非常简单,通过自身稳定的迭代过程找到全局最优解,对缺失数据的填充精度还是比较高的。但该方法通过整个数据集来进行填充,忽略了数据的局部相似性。同时EM算法收敛的速度是无法自身控制的。缺失数据的多少决定了算法速度,数据缺失比例越大,收敛速度也会越慢。还有就是当极大似然函数无法获取时,EMI算法也无法计算。

Rahman等人提出了一种称为模糊期望最大化的缺失值插补的数据预处理新技术(Fuzzy Expectation Maximization Imputation,FEMI)[48]。该算法使用最相似的记录对缺失值进行有根据的猜测。在确定一组最相似的记录时,它考虑了聚类的模糊性质。因此,它将所有记录组(簇)视为相似的,并且具有一定程度的相似性。此外,在基于组估算缺失值的同时,还考虑了属于该组的所有记录的模糊性质,提出了两个级别的模糊性,将记录的隶属度与簇一起使用,以便估算缺失值。该算法相比同类型算法平均值更好,置信区间没有重叠,对于低丢失率和高丢失率几乎都表现良好,但是所提出技术的主要重点是对缺失值的估算,而不是对记录进行聚类,因此该方法不能非常精确地找到最佳簇,K值无法准确给出,需要不断实验,且需要数据集有两个或多个属性来促进FEMI中使用的模糊EMI技术所需的相关性计算。

Ogbeide等人提出了一种基于自适应“期望最大化”方法(Mode-Related Expectation Adaptive Maximization,MEAM)[49],用于缺少观测值的多元数据集,将该新方法与其他填充方法进行比较,显示出一些改进,这种搜索丢失数据的方法是为了从可用数据中获得更好的统计推断。该方法在解决调查观测缺失、无响应或数据缺失等问题时,产生的初始值最接近完整数据集的平均值可减少计算时间。同时MEAM方法属于求解无响应测量中观测缺失的迭代方法,特别是当丢失的数据由于某些条件永远无法恢复时,这种数据清理方法具有减少统计估计误差的优点。但这个方法与EM相比需要额外的步骤。这些附加过程包括从数据集分组和从数据集中选择与加权平均值相关的模式。

Razavi-Far等人提出了一种新的缺失数据填补方法(KNN and Expectation Maximization Imputation,KEMI),该方法基于K最近邻算法用于预计算,而期望最大化算法用于后计算[50]。基本思想为首先使用KNN会自动找到K个最近邻居,然后使用EMI算法来估算缺失的分数。它的优点是可以根据最近的邻居确定初始值,而不是整个数据集。其中基于KNN的技术通常基于记录的相似性找到K个最近的邻居,因此可以忽略特征之间的相关性。然后,使用EM寻找所选样本之间的整体相似度,以估算缺失的分数。KEMI方法不仅关注记录的相似性,而且关注特征之间的相关性。KNN的加入使得它没有太多迭代来估算给定数据集的缺失分数。这意味着KEMI不仅可以提高精度,而且可以提高时间效

率。虽然该方法结合了KNNI和EMI的优点,但是它仍然是基于原始数据内部进行假设,若数据缺失比例较大时,K近邻的选择结果将存在偏差,影响初始值选择。KEMI方法可以处理数值和分类特征,同时可以处理用于混合特征插补的异构数据集。

(4)热卡填充(Hot Deck Imputation,或就近补齐)

根据获取插补值的方法来将热卡插补分为最近距离热卡插补、随机抽样热卡插补、分层热卡插补和序贯热卡插补[51]。但基本思想都是在已有的完整数据中寻找与缺失对象最相似的对象来进行填充,区别就是在寻找最相似对象的具体方法上有所不同。这个方法的缺点也很明显就是如何定义客观的相似性标准来适应不同的数据集。

热卡插补法作为一种单值填充,不论是实践还是研究都应用广泛。与均值填充和其他填充方法相比,对变量经验分布的保持有不错的效果。但是该方法的填充值易受辅助变量所影响,排序变量影响获得的序列,进而填充值也会受影响[52]。

(5)冷卡填充(Cold Deck Imputation)

与热卡填补相比冷卡填补法的填补值不是根据当前的数据集来进行填充的,而是通过历史数据或者其他相关的调查数据来进行匹配填充[53]。这种填充方法存在一定的估计偏差,并不能广泛适用。

(6)回归填充(Regression Imputation)

回归填补基本上是通过完整数据集建立回归方程,然后用回归方程的预测值对缺失数据进行填充。后来提出了效果更好的随机回归填补,该方法在填补过程中给填补值添加了一个随机项,该随机项用来表示预测值的误差影响。随机回归填补法能最大程度地利用数据本身信息,使得预测变量的共线性问题得以解决[54]。回归方程的建立如下:

设Y为缺失变量,$X_j(j=1,2,\cdots,n$是与Y存在线性回归关系的完整变量,那么Y中第i个缺失值的估计值可以表示为:

$$Y_i=a_0+\sum_{j=1}^{n}a_jX_{ji} \tag{1}$$

随机回归填补就是在公式(1)的基础上增加一个随机项,以此来减少预测误差,克服样本分布扭曲的缺陷。此时公式如下:

$$Y_i=a_0+\sum_{j=1}^{n}a_jX_{ji}+\varepsilon_i \tag{2}$$

目前对回归填充法的研究大都是在原始基础上进行相关改进及应用,鲜有突破性进展。

Bashir等人提出一种新算法,用于处理多元时间序列数据集中的缺失数据。这种新方法基于矢量自回归模型,将期望最小化算法与预测误差最小化方法结合在一起,该新算法称为向量自回归插补方法(Vector Auto Regressive Model-Imputation,VAR-IM)[55]。基本思想是先对丢失的数据进行初始猜测,然后进行传统的线性插值估计,然后,通过选择最佳滞后值p来估计$\mathrm{VAR}(p)$模型,最后,通过交替使用EM和PEM算法估计$\mathrm{VAR}(p)$模型的参数,从而提高数据填补的精度。VAR-IM方法为传统的多元时间序列缺失值估算提供了一种有效的替代方法。通过对比显示随着丢失数据量百分比的增加,性能下降的幅度较小。尽管有所改进,该方法仍存在局限性,首先这项研究仅考虑了完全随机缺失数据的情况,也就是说要求数据缺失的原因与观察值和缺失值均无关。其次,VAR-IM方法的有效性要求时间序列应该是固定的。如果丢失数据的百分比很低(例如少于10%),则VAR-IM方法不会优先使用。

Stein等人提出了一种更复杂的方法,即增量属性回归插补(Incremental Attribute Regression Imputation,IARI)[56],它对所有具有缺失值的属性进行优先级排序,然后使用所有没有缺失值或存在缺失值的属性值逐个迭代地"修复"每个属性。已经修复,作为预测指标。此外,目标变量还可以在修复过程中用作预测变量。修复属性是通过构建回归模型并将其用于估计缺失值来实现的。这里使用随机森林算法用于对数值和分类变量进行修复建模。该算法的主要优势是在修复的训练集上训练的最终模型具有更高的准确性,并且可以更准确地估计缺失值。但IARI算法在计算上非常苛刻,它要求建立的随机森林与应修复的属性数量一样多,且算法结果受属性重要度排列以及缺失属性比例影响。通常来说IARI方法在MAR缺失类型中表现较好。

Dzulkalnine等人提出了一种改进的模糊主成分分析-支持向量机-模糊C均值(Fuzzy Principal Component Analysis-Support Vector Machine-Fuzzy C-Means,FPCA-SVM-FCM)的混合填充方法[57]。该方法使用的特征选择方法是模糊主成分分析(FPCA),它在考虑异常值的情况下识别数据集中的相关特征。然后,使用支持向量机对所选特征进行分类并删除不相关的特征。识别出数据集中的重要特征后,然后通过模糊C均值估算缺失的数据。这种方法一定程度上提高了分类以及填充的准确性,减少了时间复杂度。但是如果数据集中存在过多的异常值会降低填补方法的有效性,因为删除过多的异常值,会导致信息不全,影响缺失数据的计算。因此它多适用于MAR类型的缺失值处理。

总的来说回归填补与均值填补相比,效果还是比较好的。但是回归填补和均值填补都没有考虑缺失数据的不确定性,主观增大了变量间的关系[58]。如果样本量过大,回归方程难以准确定义。

(7)聚类填充(Clustering Imputation)

简单的常值填补没有考虑缺失值的偏差,而且容易改变原样本的分布情况。而其他的模型填补需要满足一个模型对应一个缺失属性,当缺失属性增多时效率降低。聚类填充是目前研究使用最广泛的填充方法,该方

法先通过聚类的方式将数据集分类,然后在每一类里进行相似填充。以经典的基于K-means聚类填充算法为例,先将原数据集划分成完整数据集和缺失数据集,在完整数据集上进行聚类,分成K个簇,计算缺失数据每个对象与K个簇中心的相似度,把最相似的簇的属性均值填充给该缺失对象。

近几年来各种聚类填充算法开始涌现,这些聚类填补方法大致可以分成两种。

第一种方法是先聚类缺失数据集中的完整数据来进行分类,然后通过相似度度量将缺失数据对象划分到最相似的簇中,并通过簇内信息进行填补。这类方法的缺点是只考虑缺失数据的局部情况,忽略了整体分布。

比如Raja等人提出了基于粗糙K均值的缺失值填补(Rough K-Means Imputation,RKMI),通过将一个对象放置到一个以上的群集中来解决脆性问题[59]。基于粗糙K均值插补算法,使用下限和上限对象信息代替簇质心,将具有较低值的对象以较低的近似平均值表示,然后使用有关较低的近似值的信息来估算属性值。如果非参考对象存在于较高近似值中,则有关较高近似对象的信息将用于估算缺失值。如果数据集具有较高的方差,则基于粗糙K均值参数的插补可为插补值提供最佳精度。该方法与基于K均值、模糊C均值的填补方法进行了比较,整体性能优于现有方法。该方法虽适用于大型数据集,但中间的K值选择不确定,以及时间复杂度高仍是很大的问题。

对不完整数据进行分类的最流行的方法之一是使用填补以合理的值代替缺失的值。但是,当将分类器应用于新的未知实例时,强大的填补方法会占用大量计算资源。Tran等人提出了整合填补方法,即基于聚类和特征选择的不完整数据填补的新方法[60],通过聚类和特征选择的分类效果来提高效率而又不损失填充准确性。其中聚类用于减少填充使用的实例数量。特征选择用于删除训练数据的冗余和不相关特征,从而大大降低了估算成本,减少了估算时间,大大地提高了效率。由于特征选择会删除不相关特征,所以所提出的方法适用于缺失率不高的大型数据集,如果缺失率过高,聚类精度和特征选择受到影响,那么填充效果也会变差。

Shi等人提出了一种针对不完整数据的改进均值填补聚类算法(K-Means-Improved Mean Imputation,KMIMI)[61],该方法先用无缺失值对象进行聚类,并使用每个聚类的均值属性值分别填充相应的缺失值。采用簇形质心的摄动分析方法,求出最优的填充值。这种方法虽然在一定程度填充准确性有所提升,但也存在局限性,如它要求每个属性在不完整数据集中至少存在一个值。也就是说,一个对象不能缺少所有属性值,并且所有对象也不能缺少相同属性。在大多数情况下,数据集中的缺失率越高,聚类结果的准确性越低,填充性能也会下降。因此这种方法要求缺失率范围在5%至30%之间。

第二种方法是先对缺失数据进行初始化处理或者不处理,如定义缺失数据集的相似度度量,然后根据相似关系对整个数据集进行聚类,最后进行簇内填补。这类方法没有考虑缺失信息带来的误差,容易影响聚类结果,使得聚类过程复杂。

Nikfalazar等人提出一种的新混合填补方法(Decision Trees and Fuzzy Clustering with Iterative Learning,DIFC)[62],以使用混合填补方法来处理MCAR类型的缺失数据。DIFC将决策树和模糊聚类与迭代学习方法结合在一起,其中模糊聚类迭代以从记录中学习新的估计值,这些记录具有由决策树确定的相似属性值。换句话说,所提出的填充方法结合了有监督的机器学习方法(即决策树)和无监督的机器学习方法(即模糊聚类),以迭代的方式来估算缺失值。DIFC填补方法实现了双重分割方法,找到最佳记录来填补缺失值。另外,迭代学习方法提高了估算值的准确性。在每次迭代期间,DIFC使用上一次迭代中的估算值来重新聚类并更新估算值。DIFC方法的性能与丢失率没有显著相关,相反缺失模式是影响DIFC效率的重要因素。虽然DIFC的性能在各种丢失率下均很稳定,但是该方法的计算成本比较高,且适用于缺失值分布均匀的数据集。

冷泳林等人提出基于AP聚类的不完整数据填充算法(Affinity Propagation Imputation,API)[63],该方法改变了传统的先对完整数据聚类的做法,重新定义缺失数据对象间的相似度度量方式,从而直接对缺失数据聚类,最后用同一类对象的属性值填充缺失对象。该方法有效地避免了不同类对象对缺失值的影响,一定程度上提高了填充精度,且对缺失率比较大的数据容忍性比较好,但是它的相似度度量方式选取影响聚类效果,从而影响填充,比较适用属性值连续的数据集。

对于缺失数据集由于大量样本存在缺失值,单一聚类算法无法获得良好的聚类结果,从而填充不准确。为了克服这个问题,Wang等人提出一种基于集成聚类算法的缺失数据填充[64]。在提出的算法中,先用无缺失值对象进行聚类,并使用每个聚类的均值属性值分别填充缺失属性的值。然后应用聚类质心的扰动分析来寻找最优填补。该方法使用集成聚类技术将多个聚类结果组合成一个可能更好的结果,虽然提高了填补精度,但是选择不同的聚类算法会导致具有不同的参数初始化,进而导致不同的聚类填充结果,因此选择聚类算法需要根据数据集情况决定。受到聚类方法的影响该填充方法适用于低丢失率在5%至30%间的大型数据集。

各领域学者针对不同的数据集使用不同的聚类方法和填充方式,效果也各不相同,难以统一标准来比较。这类方法无论在哪种阶段聚类,都会因为数据缺失影响到聚类精度,比较适合处理高维数据集。

以上几种方法都属于单值填补,这类方法填充值是

唯一的,基本上是主观推断填充,操作简单,但没有体现填充值的不确定性,一定程度上改变了原数据集的分布情况,一旦效果不好就会导致研究结果有偏差。

(8)多重填补(Multiple Imputation,MI)

1978年Rubin等人提出多重填补法(Multiple Imputation,MI)[16],并在20世纪90年代初进行了多领域的应用研究[65-66],后经过Schafer[67]和Meng[68]等人的后续研究。已经逐渐形成一个完整的体系。多重填补方法的基本思想是为缺失值推断出多个估计填补值,并产生多个完整数据集进行综合分析,确定最终的估计填充值,这样做考虑了缺失值的不确定性。该方法通过多个估计值来模拟缺失值的实际后验分布[69]。

多重填补认为待填补的值应是随机的,通过已有的值进行预测,估计出待填补的值,然后加上不同的噪声产生多组填补值,最后选取符合依据的填补值[70-71]。多重填充方法的三个步骤如下:

①首先为每个缺失值估计一组可能的填补值,用来反映缺失值的不确定性,并构造多个完整数据集合。

②采用相同的统计方法对这些完整数据集进行计算分析。

③对来自各个完整数据集的结果进行综合分析,通过评分函数选择合适的填补值。

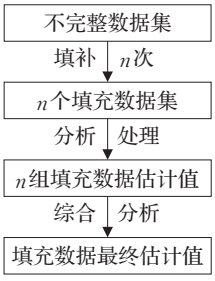

图1 多重填补算法流程

算法流程如图1所示。

在原先填补思想的基础上,许多学者进一步提出来许多相关改进、应用、算法比较。

如大多多重填补的统计文献都集中在无界的连续变量上,Geraci等人提出了一种灵活的基于分位数的估算模型,该模型适用于在单界或双界区间上定义的分布[72]。通过应用一系列具有单个或双重边界范围的变换,可以确保正确支持估算值。仿真研究表明,该方法能够处理偏斜、双峰和异方差性,并且与竞争方法(例如对数正态填补和预测均值匹配)相比具有更好的性能。尽管当有界变量受MAR影响时,它仍然比完整案例分析更有效并可用作预测变量。但是该方法具有随机有效性,且当样本量较小时,基于分位数的估算与其他估算方法相比并不会显视出自身的优势,且计算要求比较高,尤其是从数据估计变换参数时。

Quartagno等人提出基于选择模型的多级填补方法(Full Conditional Specification Multiple Imputation,FCS-MI),该方法将在多重填补的全条件规范框架内使用[73]。具体来说,采用审查的双变量概率模型来描述非随机丢失的二进制变量。该模型的第一个方程式定义了缺失数据机制的回归模型。第二个方程式指定要估算的变量的回归模型。二进制数据的非随机选择是通过两个回归模型的误差项之间的相关性映射的。分层数据结构由两个方程式中的随机截距建模一种新颖且独特的方法来处理假定为非MAR的不完整二进制多级数据。单变量插补方法可以轻松地合并到FCS框架中以处理多变量缺失。但是该方法需要保持簇的数量和簇的大小不变。因为两种量的变化都可能影响到方法的性能。

Gondara等人提出了一种基于超完全深度去噪自动编码器的多重填补模型(Multiple Imputation Using Denoising Autoencoders,MIDA)。提出的模型能够处理不同的数据类型,缺失模式,缺失比例和分布[74]。由于去噪自动编码器在初始化时需要完整的数据,因此该方法在连续变量的情况下最初使用各自的列平均值,而在分类变量的情况下使用最频繁的属性值作为初始化时缺失数据的估计值。提出的模型在MCAR和MNAR的缺失类型下显著优于当前的最新方法。且该方法适用于数据集很大且维度较高。但是该方法要求有足够的完整数据来训练模型,因此缺失比例不宜过大,计算成本较高。

与单值填补相比,多重填补方法保留了完全数据分析法和结合数据收集者知识的能力的优点。并且多重填补方法还表现出另外三个特别重要的优点:一是采取随机抽取的填补方式,使得估计更加有效。二是随机抽取下得出的有效推断是采用直接方式并结合了完全数据推断的,这样做能反映当前模型下因缺失值而产生的附加变异。三是在随机抽取填补下使用完全数据方法,能够对不同模型下无回应的推断敏感性进行研究[75-76]。

多重填补也存在一些缺点:①估计多个填补值比单值填补需要进行更多工作;②存储多重填补数据集的空间需求更大;③多重填补数据集的分析工作花费精力更多。

3.4 不处理

与前两种方法对原数据集进行缺失填充相比,这种方法直接在原数据上直接进行学习[77-78]。最具代表性的方法有贝叶斯网络、人工神经网络、粗糙集方法等。

贝叶斯网络是用来表示变量间连接概率的图形模式[79]。贝叶斯网络需要对当下领域知识熟悉,至少要清楚变量间的依赖关系。因此这种方法对使用者要求比较高。

人工神经网络通过径向基函数等方法能有效解决缺失值问题。但由于神经网络模型知识学习过程复杂难懂,所以应用起来还不尽如人意。所以人工神经网络在缺失值上还有待进一步研究。

粗糙集理论是利用实体间的不可分辨性来描述对象[80]。传统的粗糙集理论主要是针对完整数据集

的。随着粗糙集扩展模型提出[81-82],粗糙集理论开始能有效地应对数据缺失问题,并从缺失数据集上直接进行学习。

4 缺失数据处理方法比较

本文分别从前提、优缺点、适用范围对常见的几类缺失值处理方法以及近三年相关改进方法进行了比较。其中单值填充法的优点是操作简单方便,适合缺失比例不大的数据集。如果缺失比例大于5%,并且缺失类型为随机缺失和非随机缺失,则可以使用多重填补法,虽然工作量比较大,但应对大量缺失值效果更好。如果缺失比例小于5%,缺失类型是完全随机缺失,则可以考虑删除法来解决,这样既不影响数据信息量,效率也高。若数据类型太过复杂,可以考虑使用聚类填补法,通过聚类减少工作量。如表1详细列出了各类缺失数据现有处理方法的对比。表2列出了近三年各类代表性算法的对比情况。

从表1可以看出不同的缺失数据处理方法有不同

表1 缺失值处理方法比较

处理方法	缺失类型	优点	缺点	适用范围
删除法	完全随机缺失	操作过程简单,速度快	容易丢失有用信息,误差大	数据集缺失比例小于5%
权重法	随机缺失	操作过程简单	稳定性差,误差大	因无回答而造成的缺失问题
人工填充	完全随机缺失	操作过程简单,填充准确性比较高	使用人员要求高	数据集缺失少,使用人员熟悉数据集
均值填充	完全随机缺失	操作过程简单	仅利用了观测到的信息,主观性强,不稳定,有误差	数据规模小,缺失比例小,分布集中
EM填充	完全随机缺失或随机缺失	稳定性好误差较小	基于正态分布假设,收敛的速度一般受缺失数据的多少所影响,不适合高维数据	适用于正态分布或近似正态分布的数据集
热卡填充	随机缺失	操作过程较简单,同均值填补和回归填补相比在保持变量经验分布上效果较好	均方误差公式不明确,填充值易受辅助变量所影响	在同批次收集的数据集间进行
冷卡填充	随机缺失	操作过程简单	填充效果取决于前期数据的质量,存在估计偏差	在不同批次收集的数据集间进行
回归填充	随机缺失	操作过程简单,充分利用了变量之间的关系	没有考虑数据的不确定性,不适合高维数据	适用于正态分布或近似正态分布,且有多个辅助变量的数据集
聚类填充	随机缺失	变量类型要求低,拟合效果好,稳定性高,误差小,适合高维数据	操作过程复杂,时间成本较高	适合任意缺失模式,各种分布类型的数据集
多重填充	随机缺失	保持了原数据集的不确定性,保持了变量之间的关系,稳定性高,误差小	操作过程复杂,只能得到最后的参数估计	适用于有多个辅助变量,缺失率高的数据集
不处理	任意缺失	简单方便	误差大,效果差	缺失比例很小

表2 改进的缺失值处理方法比较

缺失值处理方法	缺失类型	优点	缺点	适用范围
FEMI	随机缺失	准确性高,误差小	填充效果取决于K值的选择	存在两个及以上属性的小规模数据集
MEAM	完全随机缺失或随机缺失	准确性高,误差小	计算过程复杂,成本较高	缺失观测值的多元数据集
kEMI	随机缺失	过程简单,时间复杂度小	缺失比例影响K值选择进而影响初始值选择	可处理数值和分类特征的混合特征数据集以及异构数据集
VAR-IM	完全随机缺失	过程简单,误差小	要求时间序列固定数据缺失的原因与观察值和缺失值均无关	缺失比例大于10%的多元时间序列数据集
IARI	随机缺失	准确性高,误差小	计算复杂,成本高,填充结果受属性重要度排列方式影响	属性值缺失比例小的
RKMI	随机缺失	稳定性高,误差小	过程复杂,时间复杂度高,受K值选择影响	具有较高的方差的大型数据集
KM-IMI	随机缺失	稳定性高,效果好	过程复杂,受缺失比例影响	一个对象不能缺少所有属性值,适合缺失比例在5%~30%的大型数据集
DIFCI	完全随机缺失	准确性高,不受缺失率影响	计算成本高,受缺失模式影响	适合缺失值分布均匀的数据集
API	随机缺失	对缺失率大的数据容忍性好	相似度度量选择影响填充精度	适合属性值连续的数据集
FCS-MI	完全随机缺失或随机缺失	针对特定数据类型 稳定性高,误差小	计算过程复杂,需要保持簇的数量和簇的大小不变	适合二进制多级数据
MIDA	完全随机缺失或随机缺失	能处理不同的数据类型,缺失模式,缺失比例和分布	计算成本高,过程复杂	适合缺失比例不大的高维大型数据集

的适用范围,因此在处理缺失数据时,要根据缺失数据的自身情况,选择最佳的处理方法以求达到最好的效果。

从表1可以看出不同类型的缺失数据处理方法有不同的适用范围,因此在处理缺失数据时,要根据缺失数据的自身情况,选择最佳的处理方法以求达到最好的效果。表2对文中列举的近三年改进方法从优缺点、适用范围作了进一步对比,更直观地了解到目前各领域缺失数据处理方法的多样化。

5 缺失数据填充效果的评价

缺失数据填充效果的评价通常在完整数据集上进行模拟实验。首先以完整数据集为基础,制造几种不同缺失率的缺失数据集。然后用不同的填充方法对缺失数据集进行填充。最后将原始的完整数据集与填充后的数据集进行对比,通过常用的评价指标对数据填充的效果进行评价。本文从参数角度和拟合角度两个指标进行介绍[83]。

参数角度用两种标准衡量填充精度,一是MAD平均绝对离差,该标准用于衡量真实值和填充值两者之间的匹配程度,公式如式(3)所示:

$$MAD=\frac{1}{n}\sum_{i=1}^{n}\left|r_i-e_i\right| \tag{3}$$

第二个标准是RMSE均方根误差,衡量填充值和真实值间平均误差,公式如式(4)所示:

$$RMSE=\left(\frac{1}{n}\sum_{i=1}^{n}\left|r_i-e_i\right|^2\right)^{1/2} \tag{4}$$

其中,n为缺失数值数目,r_i为第i缺失值的真实值,e_i第i个缺失值的填充值,$i=1,2,\cdots,n$两个标准的含义,MAD的值越小,表示真实值和填充值之间的离差越小,两者之间的匹配程度越高,那么填充精度就越高。同理RMSE的值越小,填充值和真实值间平均误差越小,填充精度就会越高[84-85]。拟合角度通常是建立坐标轴,用折线图的形式将真实值和填充值的分布情况做直观的比较。折线图不仅可以反映出真实值和填充值的变化趋势,而且可以从中看出二者的拟合情况对填充效果做出判断。

6 缺失数据处理方法的总结与展望

目前对数据缺失问题的讨论研究已经逐步成熟,对缺失数据的处理涉及到各个研究领域,呈现多元化发展。本文梳理了缺失数据的国内外研究背景、原因以及缺失类型。并介绍了各类处理方法,其中详细阐述了填充方法,对经典的填充方法进行了比较汇总,然后对每类方法的最新改进方法进行了汇总比较,同时从参数角度与拟合角度介绍了数据填充效果的评价常用评价指标。最后作出如下展望:

随着网络科技的发展,各领域的数据采集能力得到提升,数据表现出海量式、高维度性、复杂性、动态性等特征。为了充分发挥各领域数据的价值,缺失数据的处理能力变得至关重要。面对大规模、高维度复杂的缺失数据,国内外学者对现有方法提出许多改进方法,但也存在许多问题。

目前的缺失值填补算法主要是针对MAR缺失机制下的数据集,使用相关的属性值来估计缺失数据的值,但是这些方法都有其自身的缺点,例如线性回归算法基于统计概率和最大期望算法,这些算法必须对数据集中的数据分布有足够的了解。但是对大多数数据集的理解是有限的。基于贝叶斯网络和k邻域算法等数据挖掘类的填充方法也有局限性,使用贝叶斯网络应具有一定的领域知识和数据知识,有必要清楚各种属性之间的依存关系,且直接使用数据集训练贝叶斯网络非常复杂。而面对缺失率很高的情况,KNN相关填充算法所使用的K值并不是真正意义上的K值,影响了后续的填充准确性。

多重插补是目前处理缺失数据的高级方法。标准填补过程建立在MAR缺失机制的假设基础上,但是该方法可以处理MCAR和NMAR类型的数据集,尽管在NMAR缺失机制下填补要复杂得多,多重填补也可根据来自可用数据的信息提供关联的无偏和有效估计,即得出的估计类似于从完整数据计算得出的估计。但该方法不仅会影响缺失数据的变量系数估计,还会影响其他完整数据的变量估计。为了使填补效果更加接近实际情况,还可以在数据的来源、变化以及影响因素等多个方面努力,通过提前准备工作尽可能地学习其样本特征,从而有针对性地填补。此外使用机器学习算法需要大量时间进行填补和获取总数据集。在时间要求很高的应用领域(例如医学、金融或制造业)中,可能会感觉到长时间计算所带来的影响。因此在未来可以利用动态编程来加快计算时间。随着数据共享时代的到来用于混合特征填补的异构数据集,似乎是未来研究的一个有价值的方向。

参考文献:

[1] MEMBERS B I G D C.Database resources of the big data center in 2019[J].Nucleic Acids Research,2019,47:8-14.

[2] CARLSON D,CARIN L.Continuing progress of spike sorting in the era of big data[J].Current Opinion in Neurobiology,2019,55:90-96.

[3] CHMIELEWSKI M,KUCKER S C.An Mturk crisis shifts in data quality and the impact on study results[J].Social Psychological and Personality Science,2020,11(4):464-473.

[4] SUNDARARAJAN A,KHAN T,MOGHADASI A,et al.

Survey on synchrophasor data quality and cybersecurity challenges, and evaluation of their interdependencies[J]. Journal of Modern Power Systems and Clean Energy, 2019,7(3):449-467.

[5] MOORE E W G,LANG K M,GRANDFIELD E M.Maximizing data quality and shortening survey time: three-form planned missing data survey design[J]. Psychology of Sport and Exercise,2020,51:101701.

[6] LAI S M,WAN L,ZENG X J.Comparative analysis of multi-fractal data missing processing methods[J]. Applied and Computational Mathematics,2019,8(2):44-49.

[7] YESHA Y.Summary of missing data and its processing methods[J].Electronic Testing,2017(18):65-67.

[8] ZHU J,GE Z,SONG Z,et al.Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data[J].Annual Reviews in Control,2018,46:107-133.

[9] ZHONG H,HU W,PENN J M.Application of multiple imputation in dealing with missing data in agricultural surveys:the case of bmp adoption[J].Journal of Agricultural and Resource Economics,2018,43:78-102.

[10] AZEROUAL O,SAAKE G,WASTL J.Data measurement in research information systems: metrics for the evaluation of data quality[J].Scientometrics,2018,115(3):1271-1290.

[11] STEPHAN F F,DEMING W E,HANSEN M H.The sampling procedure of the 1940 population census[J]. Journal of the American Statistical Association,1940, 35:615-630.

[12] POLITZ A,SIMMONS W.An attempt to get the "not at homes" into the sample without callbacks[J]. Journal of the American Statistical Association,1949,44(245):9-16.

[13] YATES F.The analysis of replicated experiments when the field results are incomplete[J]. Empire Journal of Experimental Agriculture,1933,1(2):129-142.

[14] ALADE O A,SALLEHUDDIN R,RADZI N H M,et al. Missing data characteristics and the choice of imputation technique:an empirical study[C]//International Conference of Reliable Information and Communication Technology.Cham:Springer,2019:88-97.

[15] DEMPSTER A,LAIRD N,RUBIN D.Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society,1977,39(1):1-38.

[16] RUBIN D B.Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse[C]// Proceedings of the Survey Research Methods Section of the American Statistical Association,1978:20-34.

[17] URANGA R,MOLENBERGHS G,ALLENDE S.A multiple regression imputation method with application to sensitivity analysis under intermittent missingness[J].Communications in Statistics-Theory and Methods,2020:1-16.

[18] KALTON G,KISH L.Some efficient random imputation methods[J].Communications in Statistics-Theory and Methods,1984,13(16):1919-1939.

[19] BATISTA G E,MONARD M C.An analysis of four missing data treatment methods for supervised learning[J]. Applied Artificial Intelligence,2003,17(5/6):519-533.

[20] ZAKARIA N A,NOOR N M.Imputation methods for filling missing data in urban air pollution data for Malaysia[J].Urbanism. Arhitectura. Constructii,2018,9(2):159.

[21] LITTLE R J A,RUBIN D B.Statistical analysis with missing data[M].[S.l.]:John Wiley & Sons,2019.

[22] 金勇进,邵君.不同差补方法的比较[J].数理统计与管理, 2000,19(4):50-54.

[23] 金勇进,邵君.缺失数据的统计处理进程[M].北京:中国统计出版社,2009.

[24] 邓银燕.缺失数据的填充方法研究及实证分析[D].西安: 西北大学,2010.

[25] 罗永峰,叶智武,郭小农.钢结构施工过程监测数据缺失机理与处理方法[J].同济大学学报(自然科学版),2014, 42(6):823-829.

[26] 杨弘,田晶,王可,等.混合型缺失数据填补方法比较与应用[J].中国卫生统计,2020,37(3):77-81.

[27] NUGROHO H,SURENDRO K.Missing data problem in predictive analytics[C]//Proceedings of the 2019 8th International Conference on Software and Computer Applications,2019:95-100.

[28] DU J,HU M,ZHANG W.Missing data problem in the monitoring system:a review[J].IEEE Sensors Journal, 2020,20(23):13984-13998.

[29] ZHANG N H.Methodolgical progress note:handling missing data in clinical research[J].Journal of Hospital Medicine,2020,15(4):237-239.

[30] GOMILA R,CLARK C S.Missing data in experiments: challenges and solutions[J].Psychological Methods,2020.

[31] PANDA B S,ADHIKARI R K.A method for classification of missing values using data mining techniques[C]// International Conference on Computer Science, Engineering and Applications(ICCSEA),2020:1-5.

[32] GARCÍA-LAENCINA P J,ABREU P H,ABREU M H, et al.Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values[J].Computers in Biology and Medicine, 2015,59:125-133.

[33] LIN W C,TSAI C F.Missing value imputation:a review and analysis of the literature[J].Artificial Intelligence Review,2020,53(2):1487-1509.

[34] YU L,ZHOU R,CHEN R,et al.Missing data preprocessing in credit classification:one-hot encoding or imputation[J].Emerging Markets Finance and Trade,2020:1-11.

[35] FLETCHER MERCALDO S,BLUME J D.Missing data

and prediction:the pattern submodel[J].Biostatistics,2020,21(2):236-252.

[36] SANTOS M S,PEREIRA R C,COSTA A F,et al.Generating synthetic missing data:a review by missing mechanism[J].IEEE Access,2019,7:11651-11667.

[37] PATIDAR P,TIWARI A.Handling missing value in decision tree algorithm[J].International Journal of Computer Applications,2013,70(13):31-36.

[38] 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.

[39] 陈景年.选择性贝叶斯分类算法研究[D].北京:北京交通大学,2008.

[40] BERTSIMAS D,PAWLOWSKI C,ZHUO Y D.From predictive methods to missing data imputation:an optimization approach[J].The Journal of Machine Learning Research,2017,18(1):7133-7171.

[41] MAHESWARI K,PRIYA P P A,RAMKUMAR S,et al.Missing data handling by mean imputation method and statistical[C]//EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing,2019:137.

[42] WANG S,LI M,HU N,et al.K-means clustering with incomplete data[J].IEEE Access,2019,7:69162-69171.

[43] KABIR G,TESFAMARIAM S,HEMSING J,et al.Handling incomplete and missing data in water network database using imputation methods[J].Sustainable and Resilient Infrastructure,2019:1-13.

[44] GHAHRAMANI Z,JORDAN M I.Supervised learning from incomplete data via an EM approach[C]//Advances in Neural Information Processing Systems,1994:120-127.

[45] GHAHRAMANI Z,JORDAN M I.Learning from incomplete data[M].Cambridge,MA,USA:Massachusetts Institute of Technology,1994.

[46] YANG H L,PENG J H,XIA B,et al.An improved EM algorithm for remote sensing classification[J].Chinese Science Bulletin,2013,58(9):1060-1071.

[47] RUMALING M I,CHEE F P,DAYOU J,et al.Missing value imputation for PM 10 concentration in Sabah using nearest neighbour method and expectation-maximization algorithm[J].Asian Journal of Atmospheric Environment,2020,14(1):62-72.

[48] RAHMAN M G,ISLAM M Z.Missing value imputation using a fuzzy clustering-based EM approach[J].Knowledge and Information Systems,2016,46(2):389-422.

[49] OGBEIDE E M.A new iterative imputation method based on adaptive expectation maximization[J].Science-Tech Journal,2018,3(1):133-142.

[50] RAZAVI-FAR R,CHENG B,SAIF M,et al.Similarity-learning information-fusion schemes for missing data imputation[J].Knowledge-Based Systems,2020,187:104805.

[51] RAO J N K,SHAO J.Jackknife variance estimation with survey data under hot deck imputation[J].Biometrika,1992,79(4):811-822.

[52] OZTURK A.Accuracy improvement in air-quality forecasting using regressor combination with missing data imputation[J].Computational Intelligence,2021,37(1):226-252.

[53] 金勇进.缺失数据的插补调整[J].数理统计与管理,2001,20(6):47-53.

[54] 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.

[55] BASHIR F,WEI H L.Handling missing data in multivariate time series using a vector auto regressive model-imputation algorithm[J].Neurocomputing,2018,276:23-30.

[56] VAN STEIN B,KOWALCZYK W.An incremental algorithm for repairing training sets with missing values[C]//International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems.Cham:Springer,2016:175-186.

[57] DZULKALNINE M F,SALLEHUDDIN R.Missing data imputation with fuzzy feature selection for diabetes dataset[J].SN Applied Sciences,2019,1(4):1-12.

[58] NIKFALAZAR S,YEH C H,BEDINGFIELD S,et al.A new iterative fuzzy clustering algorithm for multiple imputation of missing data[C]//2017 IEEE International Conference on Fuzzy Systems,2017:1-6.

[59] RAJA P S,THANGAVEL K.Soft clustering based missing value imputation[C]//Annual Convention of the Computer Society of India.Singapore:Springer,2016:119-133.

[60] TRAN C T,ZHANG M,ANDREAE P,et al.Improving performance of classification on incomplete data using feature selection and clustering[J].Applied Soft Computing,2018,73:848-861.

[61] SHI H,WANG P,YANG X,et al.An improved mean imputation clustering algorithm for incomplete data[J].Neural Processing Letters,2020:1-14.

[62] NIKFALAZAR S,YEH C H,BEDINGFIELD S,et al.Missing data imputation using decision trees and fuzzy clustering with iterative learning[J].Knowledge and Information Systems,2020,62(6):2419-2437.

[63] 冷泳林,张清辰,鲁富宇.基于AP聚类的不完整大数据填充[J].计算机工程与应用,2015,51(10):123-127.

[64] WANG P,CHEN X.Three-way ensemble clustering for incomplete data[J].IEEE Access,2020,8:91855-91864.

[65] RUBIN D B.Multiple imputation for nonresponse in surveys[J].Statistical Papers,1990,31(1):180.

[66] RUBIN D B,SCHENKER N.Multiple imputation in healthcare databases:an overview and some applications[J].Statistics in Medicine,1991,10(4):585-598.

[67] SCHAFER J L,OLSEN M K.Multiple imputation for multivariate missing-data problems:a data analyst's perspective[J].Multivariate Behavioral Research,1998,33(4):545-571.

[68] MENG X L.On the rate of convergence of the EC algorithm[J].The Annals of Statistics,1994,22(1):326-339.

[69] HUGHES R A,HERON J,STERNE J A C,et al.Accounting for missing data in statistical analyses:multiple imputation is not always the answer[J].International Journal of Epidemiology,2019,48(4):1294-1304.

[70] MURRAY J S.Multiple imputation:a review of practical and theoretical findings[J].Statistical Science,2018,33(2):142-159.

[71] VAN GINKEL J R,LINTING M,RIPPE R C A,et al.Rebutting existing misconceptions about multiple imputation as a method for handling missing data[J].Journal of Personality Assessment,2020,102(3):297-308.

[72] GERACI M,MCLAIN A.Multiple imputation for bounded variables[J].Psychometrika,2018,83(4):919-940.

[73] QUARTAGNO M,CARPENTER J R.Multiple imputation for discrete data:evaluation of the joint latent normal model[J].Biometrical Journal,2019,61(4):1003-1019.

[74] GONDARA L,WANG K.MIDA:multiple imputation using denoising autoencoders[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining.Cham:Springer,2018:260-272.

[75] STAVSETH M R,CLAUSEN T,RØISLIEN J.How handling missing data may impact conclusions:a comparison of six different imputation methods for categorical questionnaire data[J].SAGE Open Medicine,2019,7.

[76] HUQUE M H,CARLIN J B,SIMPSON J A,et al.A comparison of multiple imputation methods for missing data in longitudinal studies[J].BMC Medical Research Methodology,2018,18(1):168.

[77] ŚMIEJA M,STRUSKI Ł,TABOR J,et al.Processing of missing data by neural networks[C]//Advances in Neural Information Processing Systems,2018:2719-2729.

[78] WEI G C G,TANNER M A.A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms[J].Journal of the American Statistical Association,1990,85(411):699-704.

[79] YE C,WANG H,LU W,et al.Effective Bayesian-network-based missing value imputation enhanced by crowdsourcing[J].Knowledge-Based Systems,2020,190:105199.

[80] 张文修,吴伟志.粗糙集理论介绍和研究综述[J].模糊系统与数学,2000,14(4):1-12.

[81] SUJATHA M,DEVI G L,RAO K S,et al.Rough set theory based missing value imputation[M]//Cognitive science and health bioinformatics.Singapore:Springer,2018:97-106.

[82] PRIETO-CUBIDES J,ARGOTY C.Dealing with missing data using a selection algorithm on rough sets[J].International Journal of Computational Intelligence Systems,2018,11(1):1307-1321.

[83] SPORTISSE A,BOYER C,JOSSE J.Imputation and low-rank estimation with missing not at random data[J].Statistics and Computing,2020,30(6):1629-1643.

[84] WILLMOTT C J,MATSUURA K.Advantages of the mean absolute error over the root mean square error in assessing average model performance[J].Climate Research,2005,30(1):79-82.

[85] CHAI T,DRAXLER R R.Root mean square error or mean absolute error[J].Geoscientific Model Development Discussions,2014,7(1):1525-1534.