·综述·

机器学习在因果推断中的应用*

王　星[1,2]　吴亚飞[1,2,3]　方　亚[1,2,3△]

【提　要】 机器学习算法是挖掘数据间隐藏关系的重要手段，因果推断则是证实关联关系是否存在因果关系的重要技术。随着大数据时代的到来，两者的结合越来越密切。本文主要对近年有关机器学习与因果推断的研究进行梳理，分别从机器学习与因果推断概念、两者关系以及机器学习在因果推断中的应用等方面展开论述，并对当前研究面临的困难以及发展方向提出展望。

【关键词】 机器学习　因果推断　因果模型

【中图分类号】 R181.2　　【文献标识码】 A　　DOI　10.11783/j.issn.1002-3674.2024.02.036

机器学习具有强大的模型外线性和非线性预测能力，其侧重于关联，在疾病风险预测、诊断和预后等方面得到广泛应用[1]。因果推断具有良好的推理能力和可解释性，侧重于因果关系。机器学习和因果推断在早期并无太大联系，甚至在某些情况下，二者存在矛盾。进入大数据时代后，由于传统的因果推断方法侧重于二维变量，如果直接应用于高维数据，有可能会影响因果推断的准确性，甚至无法运行，而机器学习善于处理高维数据的特点则为因果推断提供新思路，因此二者的联系逐渐增强。本文主要对近些年有关机器学习与因果推断的研究进行梳理，介绍机器学习和因果推断相关概念、两者之间的联系，并重点归纳机器学习在因果推断各环节的应用，为相关研究者提供借鉴和参考。

机器学习与因果推断简介

1.机器学习概念及其分类

机器学习（machine learning）是一门多领域交叉学科，涵盖了概率学、统计学、数学、计算机科学等，属于人工智能范畴。卡内基梅隆大学的 Tom Mitchell 教授于 1997 年将机器学习定义为：“一个计算机程序在完成任务 T 之后，获得经验 E，其表现效果为 P，如果任务 T 的性能表现，也就是用以衡量的 P，随着 E 的增加而增加，可以称之为学习”[2]。简而言之，机器学习就是利用算法对数据进行分析，学习其内在特征而后应用于外部预测。基于学习方式不同，机器学习可分为监督学习（supervised learning）、无监督学习（unsupervised learning）、半监督学习（semi-supervised learning）和强化学习（reinforcement learning）等。监督学习是通过对已有标签（类别标签或数值标签）的数据进行学习，预测未知数据的标签，包括分类算法和回归算法。无监督学习是指输入数据中无标签，计算机自行观察数据中的特点并输出其潜在规律，学习结果为类别。典型的无监督学习有聚类算法、可视化和降维算法以及关联规则算法等。半监督学习介于无监督和监督学习之间，是用少量的标记数据和大量未标记数据来进行学习，预测未知数据的标签。强化学习是以环境反馈（奖/惩信号）作为输入，不断优化模型，以此获得最大的奖励回报。近些年不断兴起的深度学习技术，是机器学习的一个重要分支，它试图用复杂结构或非线性转换组成的多层神经元对数据进行高层抽象建模，从而实现更准确预测[3-4]。人工智能、机器学习和深度学习的关系可概括为图 1。

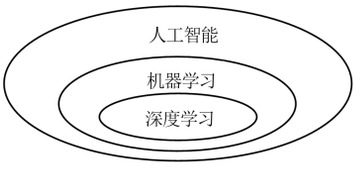

图 1　人工智能、机器学习和深度学习的关系

2.因果推断概念及其类型

因果关系（causal influence）是一种由事物变化关联和时间因素构成的复杂抽象关系，是人类认识世界与解释事物发展变化的重要方式。因果关系必须满足三个基本条件，即：时间顺序、关联关系和因变性[5]。时间顺序是指因必须发生在果前面，如狂犬病毒导致患者死亡，必须是患者先感染狂犬病毒，而后才导致死亡，这是事件发生的先后顺序；关联关系是指因和果存在一定的关联，关联只是因果的前提，因果关系蕴藏于众多关联关系之中，关联不一定是因果关系，如鸡鸣和天亮，但因果关系一定存在关联；因变性是指事件的变化是由事件的变化引起的。因果推断就是判定两个因素之间是否存在真正的因果关系，是科学推论的一种，基于辛普森悖论和因果推断研究方向的不同，因果推断可分为因果学习（causal learning）和因果推理（causal reasoning）[6]，如图 2 所示。因果学习

* 基金项目：国家自然科学基金（81973144）

1.厦门大学公共卫生学院（361102）
2.卫生技术评估福建省高校重点实验室
3.厦门大学健康医疗大数据国家研究院
△通信作者：方亚，E-mail：fangya@ xmu. edu. cn

主要侧重于从已有数据中挖掘因果模型或因果关系，数据主要来源于干预数据，但由于随机对照实验（randomized controlled trial，RCT）标准高、难度大，强行施加干预可能会带来伦理学问题，因此也可从观察性数据中进行学习。经典的因果学习如双变量因果关系，它假定变量 X 和变量 Y 非独立且因果关系为单向赋值，使得刺激 X 或 Y 时，Y 或 X 发生相应的变化，如果 X、Y 均未变化，则可能存在变量 Z 为 X、Y 共同的因，双变量因果关系学习详见图 3。而因果推理则是偏向于从已有模型中去分析数据（图 2），确定效果，如利用因果图判定发病率增加（减少）多少。

3.因果推断方法

按照数据来源不同，因果推断方法通常可分为实验（试验）数据的因果推断方法如 RCT，观察数据的因果推断方法：如合成控制法（synthetic control method，SCM）、双重差分法（differences－in－differences，DID）、倾向性得分匹配（propensity score matching，PSM）、断点回归设计（regression discontinuity design，RDD）、工具变量法（instrumental variable analysis，IV）和基于树模型的方法等[7]。不同因果推断方法的原理和优缺点见表 1。

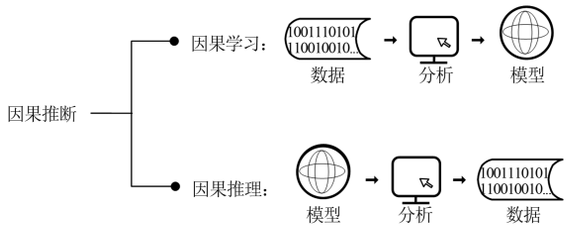

图 2 因果推断类型

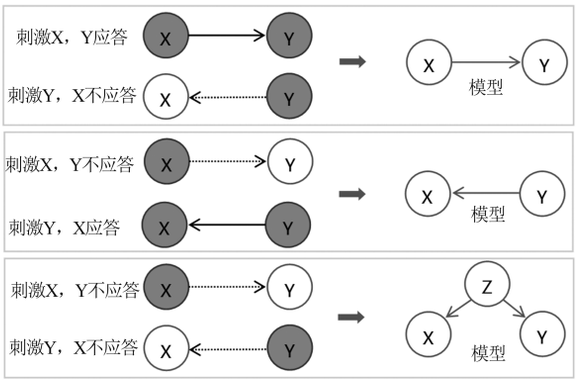

图 3 双变量因果学习

表 1 不同因果推断方法的原理和优缺点

方法	原理	优势	不足
RCT	采用随机分配方法将研究对象分配到实验（试验）组和对照组，然后接受相应的实验措施，在一致的条件或环境中进行同步观测，用客观指标测量和评价实验结果，若差异有统计学意义，则提示存在因果关联。	两组的可比性好；显著性检验合理且统计方法简单；因果推断的金标准。	标准高、难度大、强行施加干预可能会导致伦理学问题。
SCM	基于 Rubin 反事实框架，通过虚拟构造一个“控制组”，即在各方面都与受到干预的处理组一致但未受到干预的组，与处理组进行对比，二者之差即为“处理效应”。	避免政策内生性问题；避免主观选择带来的误差；可以反映每个控制对象对“反事实”事件的贡献，避免过分外推。	“控制组”的构造条件较高；要求干预的期数较大，否则信度较低。
DID	假定实验（暴露）组和对照组在未受到干预前有相同的变化趋势，先计算实验组和对照组在干预前的差值 D_0，然后计算干预后实验组与对照组的差值 D_1，最后计算 D_1-D_0，即为“处理效应”。	可以很大程度上避免内生性问题；模型设置科学，能较为准确地估计出政策效应。	仅适用于重复测量数据；要满足其前提假设，应用范围有限。
PSM	将控制组的个体按照各特性（协变量集中的变量）“距离”相近的方法与处理组中的个体进行匹配，缓解或消除选择偏倚，之后通过计算处理组与对照组的差异，即为“处理效应”。	适合观察性数据的“类随机化”；可以同时调整大量的混杂因素。	要求样本量较大；只能均衡已观测的指标变量。
RDD	构造“断点”，使得个体在该断点之上接受干预（暴露），小于该断点时不接受干预（暴露），以此来构造实验（试验）组和对照组，特别是在连续型变量下，断点附近样本的差别可以很好地反映干预和政策的因果关系。	最接近随机实验（试验）的因果推断方法；能够缓解参数估计的内生性问题。	断点附近的数据要求较高；无法计算平均治疗效应。
IV	利用一些与误差项 ε 无关但与内生性变量高度相关的变量，即工具变量，代替回归模型中的解释变量，以计算“处理效应”。	能够有效地解决内生性问题。	排他性条件难以满足。
树模型	通过训练具有预测因果效应的树，在每个叶子节点进行实验组与对照组的比较，计算“处理效应”。	能够估计非均匀处理效应；适用于复杂数据的因果推断。	计算复杂；目前仍处于研究阶段。

4.机器学习与因果推断的关系

大数据时代，机器学习的蓬勃发展已经极大地引起了学者们的关注，不少学者尝试用机器学习助力因果推断[8-10]。同样，因果推断也可用于机器学习，它能使得机器学习的“黑箱”变得可解释。两者的联系越来越密切。

首先，机器学习与因果推断具有共同的概率论和统计学基础。因果推断的很多统计学方法假定所研究的内容是服从某一概率的随机过程，这个过程称为数据生成过程（data generating process，DGP）。机器学

习与此类似，其也假定 DGP 是一个随机过程，对要学习的数据内部结构或概率分布未知，通过学习数据中某种特定的规律用于外部预测，使算法的泛化性能达到最高。另一方面，机器学习也是从训练数据中学习其特征，之后用于预测未知样本，但必须要指出的是，机器学习算法要求其训练集和测试集的数据特征差异不能过大，也就是意味着训练集能够较好地代表测试集或者更大范围的外部数据，否则其算法的泛化能力将很差，预测意义不大。从这个角度来说机器学习的预测方法只是从样本即训练集来预测更加一般化的“总体”，这与因果推断具有异曲同工之处。尽管在大数据时代，海量的数据使得样本符合总体分布不再是制约因果推断的前提，但大数据仍未改变机器学习从随机抽样推断总体的基本思想[11]。其次，机器学习技术为因果推断扩展了研究领域并补充改进了相关的研究方法。大数据时代产生了许多有价值的半结构化文本和非结构化文本，如电子病例、检查报告、微博等，这些资料在传统因果推断中往往难以得到充分利用，但又具有重要价值。机器学习技术如朴素贝叶斯算法和循环神经网络可以对文本资料进行挖掘提取，可用于构建新的变量如事件影响力指数、舆情指数，然后进行文本回归分析，因而拓宽了因果推断的研究领域。除此之外，机器学习同样能为研究者识别因果推断感兴趣的特征，这些特征随后成为纳入因果模型框架的候选者，然后使用经济学或流行病学的标准方法进行估计[12]。最后，因果推断为机器学习提供了模型优化方向和可解释性的途径。图形处理器（graphics processing unit，GPU）的飞速发展使许多模型能够借助 GPU 训练非常大的数据，并以此取得了惊人的准确度，但也带来了更加复杂的模型和高难的可解释性，理解机器学习模型背后的因果关系有助于提升模型优化的方向和模型可解释性，因果推断中的因果模型和反事实推断能够为其提供解决途径[13-14]。总之，机器学习与因果推断是互有补充、相互促进，随着机器学习和因果推断研究的不断深入，两者的交叉融合可能是未来人工智能重要的发展方向之一[15]。

机器学习在因果推断中的应用

1.缺失值填补

因果推断过程中，数据缺失不可避免，特别是在高维数据的因果推断中，数据缺失可能会对因果推断的准确性造成影响。如在前瞻性队列研究随访过程中会遇到很多缺失数据，且数据类型不同。如果数据是完全随机缺失，如患者的家庭住址，不依赖于其他变量，则不影响推断，但如果数据是随机缺失或非随机缺失，则会影响其推断。对于缺失数据，常用的基于概率分布的统计学处理有：众数填充、均数填充、中位数填充、回归填充等，这些方法虽操作方便，但存在一定的缺陷[16]。机器学习能够从数据中学习其固有特征而后用于预测，并且能够填补一些传统统计学方法无法填补的缺失数据类型如分类变量、文本等，因而能够应用于缺失数据的填充，如 K 近邻算法、支持向量机、随机森林、神经网络等。Justin Y. Lee[17] 等人提出了一种基于 K 近邻的改进算法，用于估算代谢组学数据中代谢物丰度，结果显示该算法能准确地估计代谢组学数据缺失值。张蝉[18] 提出了一种基于支持向量机的数据缺失值填补方法，分别通过支持向量机回归和支持向量机分类方法对连续性变量和类别变量进行缺失值填补，其结果表明缺失率在 45% 时，算法依旧有 91.8% 的填补准确率。张晓琴[19] 等人提出了一种基于随机森林的成分数据缺失值迭代填补法，其结果表明该方法可用于多种类型的数据集且有较高的准确性。此外，李洪飞[20] 利用深度学习中生成对抗网络（generative adversarial nets，GAN）中生成模型和判别模型来填充缺失值，实验结果表明，其在不同缺失数据机制下的因果关系推断性能优于现有算法。

2.个体匹配

匹配是因果推断研究设计和分析时必须要考虑的问题之一，如病例对照研究中病例和对照的选择、随机区组设计中组内个体的匹配等。匹配有 PSM 和协变量匹配。PSM 模型的提出是为了解决观察性研究数据的类随机化，PSM 一般用 logistic 回归虚拟变量（0 或 1）得到倾向分值，有文献表明多元 logistic 回归易产生极端的权值，而假设一个错误的模型，导致性能较差[21]。此外，PSM 模型在样本量较小的情况下无法给出有效的因果推断，利用机器学习技术可以有效解决这一问题。如麦炜琪[22] 利用 GAN 技术学习原有数据集的分布，生成更多分布相似的样本，结合原有样本和生成样本进行倾向得分匹配，从而解决传统 PSM 模型在小样本情况下的局限性，控制选择性误差。协变量匹配是为了控制混杂，利用机器学习技术同样有助于实现组别间的均匀分配，如 Ariel Linden[23] 等人利用最优判别分析（optimal discriminant analysis，ODA）算法来评估协变量平衡，并在匹配策略实施后估计处理效果。与传统方法相比，该算法对偏倚数据或异常值不敏感且能够实现组间的均匀分配。

3.估计个体处理效应

考虑到处理效应的异质性，因果推断不仅关注平均处理效应，而且关注个体处理效应。传统的因果推断在估计处理效应时大多会采用回归方法，其默认处理效应取决于交互项在统计学上具有显著性的变量，但在真实世界中有多个特征变量影响着处理变量和结

果变量，回归结果缺乏统计能力，利用机器学习可以更好地估计个体处理效应且优于传统模型[24]。目前估计个体处理效应的机器学习方法有贝叶斯自适应回归树、反事实随机森林等。Athey[25]等人将机器学习中常用的分类回归树引入到传统的因果识别框架，定义了因果树（causal tree）的概念，用它来考察异质性处理效应。Jiabei Yang[26]等人基于分类和回归树算法的扩展引入了因果交互树（causal interaction tree，CIT）算法，用于在观察性数据中发现具有异质性治疗效果的个体亚组，并用该算法评估了危重患者右心导管置入术的有效性。Watson[27]等人在随机数据中用机器学习方法检测处理效应的异质性，同时能够更加精确地控制 I 型错误，适用于观察性数据中个体处理效应的估计。

4.对因果推断方法的改进

因果推断方法众多，机器学习可以对部分因果推断方法进行改进，以实现更科学的推断。如 Luo[28]等人在评估交通设施连通性对地方矛盾解决的影响时，以冲突等级作为因变量，采用 LASSO 回归筛选协变量，而后以这些协变量作为 DID 模型中的控制变量，以此来降低控制变量的维数，可以有效缓解混杂因素造成的干扰和内生性。同样，由于传统的线性 DID 估计量依赖于平行趋势假设，即在未进行干预的情况下，治疗组和对照组之间的结果差异随时间保持不变，然而在大多数情况下这种假设可能不成立。Chang[29]提出了基于机器学习的 Abadie 半参数双重差分估计量，能够克服传统 DID 方法无法对高维数据进行非参数估计的问题。Biewen[30]等人在随机森林的基础上开发了两阶段最小二乘（two-stage least squares，2SLS）随机森林用于 IV，并用该方法估计生育多个孩子对女性劳动力供应的影响，结果显示优于传统 IV。此外，孟德尔随机化（Mendelian randomization，MR）是借助遗传变异（genetic variation）作为Ⅳ来推断暴露因素与结局之间因果关联的方法，有效避免了反向因果关联和潜在混杂因素导致的偏倚。然而，孟德尔随机化要求有较大的样本量、符合孟德尔遗传规律、IV 与暴露因素有很强的相关性、排除多基因遗传、连锁不平衡和表观遗传学的影响等。其中，针对大样本这一限制，机器学习 GAN 模型可以生成和原有样本近似的新样本以满足分析条件，例如 Lall[31]等人通过 GAN 模型扩充原有样本，并通过局部敏感哈希更新 GAN 生成器的训练程序，以加快样本生成，保证了下游如孟德尔随机化的可行性。Herlands[32]等人开发了一种机器学习算法来自动识别断点，称之为局部断点回归设计，其能够识别任意维度数据中可解释的局部断点，并且可以在无专家指导的情况下计算治疗效果。

5.在因果模型中的应用

因果模型是表示单个系统或群体内因果关系的数学模型，它有助于从统计数据中推断因果关系。因果推断常用因果模型有两个：一个是著名统计学家 Donald Rubin 教授在 1978 年提出的鲁宾因果模型（Rubin causal model，RCM），另一个是 Judea Pearl 教授在 1995 年提出的因果图模型（causal diagram）。RCM 的核心是比较同一个研究对象在接受干预（实验组）和不接受干预（对照组）时的结果差异，认为这一结果差异就是接受干预相对于不接受干预所导致的效果。因果图是一个有向图，它显示了因果模型中变量间的因果关系。因果图包括一组变量（或节点），每个节点通过箭头连接到一个或多个对其具有因果效应的其他节点，箭头描绘了因果的方向。因果图包括因果环图（causal loop diagrams），有向无环图（directed acyclic graphs，DAG）和鱼骨图（ishikawa diagrams）[33]。但目前因果图无法实现自学习机制且推理的先验知识完全由该领域的专家所提供。机器学习能利用已有数据学习因果图的结构与参数，进而克服上述缺点。如石庆喜[34]用期望最大化（expectation-maximum，EM）算法进行在线因果图参数和结构的学习，并给出了一种学习因果图结构的在线修改与学习算法，较好地解决了因果图知识获取的重要问题。Kang[35]等人在探索新型冠状病毒肺炎严重程度与环境条件（如气候因素和空气质量）的潜在因果关系时，采用集成学习算法 XGBoost 分析时间序列数据，然后筛选了部分特征用于构建 DAG，较好地解决了先验知识有限或无法获取的问题。

6.反事实推断

反事实推断是指对已发生的事件进行否定并做出与事实相反的假设，反事实是因果推断重要的一环，机器学习能够直接用于反事实推断。Samii[36]等人用机器学习中的集成学习方法从回顾性数据中估计因果效应，该方法能够克服传统方法如回归建模或匹配的限制以及在估计因果效应时使用较多协变量容易产生偏差或效率低下的问题。Schuler[37]等人使用目标最大似然估计（targeted maximum likelihood estimation，TMLE）并基于反事实来直接估计因果治疗效果，TMLE 的优点是其双鲁棒性的特点能够保证估计平均治疗效应时无偏差，它用暴露概率的估计值来更新模型而后计算治疗组与对照组之间的反事实差异。Chin[38]用 bootstrap 和机器学习重采样方法进行统计推断，在模拟实验中，该方法在去偏估计方面优于现有的基于邻域暴露模型的逆倾向加权估计方法。Averitt[39]等人提出了一种基于 GAN 的模型，称之为反事实 χ-GAN，该模型能够最大限度地减少治疗效果的抽样误差，通过学习特征平衡权重来进行反事实

推断，其结果表明该模型可作为无法实施 RCT 时估计因果效应的替代方法。

小结与展望

本文主要介绍了机器学习和因果推断的概念、类型，阐述了二者之间的关系，并从多个角度介绍了机器学习在因果推断中的应用。因果推断是揭示事物间联系的重要手段，机器学习作为目前重要的人工智能方法之一，在因果推断的各个环节得到了越来越多的应用，如数据填充、个体匹配、估计个体处理效应、因果推断方法的改进和反事实推断等。由于 RCT 实验的标准高、成本昂贵和伦理学等问题，如何充分发挥观察性资料在因果推断中的作用是当前研究的热点和难点。观察性资料因果推断的关键在于如何均匀分配和控制所有的混杂因素，使其达到类随机化或无限接近随机化。目前，因果推断仍有几个难点需要解决。

1.高维数据的因果推断

利用传统的因果推断方法在高维数据中学习因果网络结构和提高学习准确率是目前研究的难点。如在合成控制法中，虚拟控制组很难在高维数据中完美构造，而深度学习能够拟合原有数据，生成最优相似数据，未来可考虑将二者结合。倾向性得分匹配中，如果变量过多，个体的匹配可以考虑用机器学习模型来输出样本间的相似性，进而精准匹配。此外，在因果模型中，高维数据的因果模型推断也可以引入机器学习算法，可以考虑通过递归的方式将因果图一分为二，降低维度，学习其局部结构，并自底向上逐步整合成全局结构，解决先验知识有限或无法获取的问题。

2.不完全观察数据上的隐变量检测

不完全观察数据会导致虚假的因果关系，如吸烟会导致黄牙和肺癌，但如果吸烟没有被检测到或观察到，利用因果推断方法，我们往往会得到黄牙和肺癌之间虚假的因果关系。事实上，未观察到的吸烟才是两者的共同原因。针对这一情况，机器学习可以尝试从大量临床数据中挖掘相关因素，构建风险因素知识图谱，通过查询来发现隐变量。

3.因果方向判别

因果方向判别是判断两个具有确定因果关系变量间的方向，即判断哪个变量是因，哪个变量是果。传统方法基于统计假设检验，利用两种因果方向之间的差异来进行判断，但是这类方法需要预先对数据的分布类型以及因果机制的类型做出强假设才能实施。因此，也可以尝试通过神经网络将因果方向判别看做二分类问题，在模拟因果数据上进行训练，之后在真实因果数据上测试。

总而言之，机器学习和因果推断的结合有望实现基于观察性资料的因果推断，使其无限逼近因果关系而非关联关系，拓广因果推断的研究领域和方法。同样，因果推断也能为机器学习提供模型优化方向和增强其可解释性，减少机器学习模型拟合时间、增强模型鲁棒性等。因此，研究者不仅需要充分掌握传统因果推断技术，更应适当融合机器学习技术以适应大数据时代的因果推断研究。

参 考 文 献

[1] Xu C, Jackson SA. Machine learning and complex biological data. Genome Biology, 2019, 20(1): 1-4.

[2] Mitchell TM. The discipline of machine learning. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006:87-88.

[3] Berrar D, Dubitzky W. Deep learning in bioinformatics and biomedicine. Briefings in Bioinformatics, 2021, 22(2): 1513-1514.

[4] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444.

[5] 詹思延, 叶冬青主编. 流行病学. 第 8 版. 北京: 人民卫生出版社, 2017, 142-143.

[6] Shanmugam R. Elements of causal inference: foundations and learning algorithms. Journal of Statistical Computation and Simulation, 2018, 88(16): 3248-3248.

[7] 苗旺, 刘春辰, 耿直. 因果推断的统计方法. 中国科学: 数学, 2018, 48(12): 1753-1778.

[8] Luo Y, Peng J, Ma J. When causal inference meets deep learning. Nature Machine Intelligence, 2020, 2(8): 426-427.

[9] Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. Proceedings of the National Academy of Sciences, 2019, 116(52): 27151-27158.

[10] Ghosh S, Boucher C, Bian J, et al. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). Computer Methods and Programs in Biomedicine Update, 2021, 1: 100020.

[11] 洪永淼, 汪寿阳. 大数据、机器学习与统计学: 挑战与机遇. 计量经济学报, 2021, 1(1): 17-35.

[12] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 1974, 66(5): 688-688.

[13] Narendra T, Sankaran A, Vijaykeerthy D, et al. Explaining deep learning models using causal inference. arXiv preprint, 2018 : arXiv: 1811.04376.

[14] Fernández RR, De Diego IM, Aceña V, et al. Random forest explainability using counterfactual sets. Information Fusion, 2020, 63(1): 196-207.

[15] Lin SH, Ikram MA. On the relationship of machine learning with causal inference. European Journal of Epidemiology, 2020, 35(2): 183-185.

[16] 帅平, 李晓松, 周晓华, 等. 缺失数据统计处理方法的研究进展. 中国卫生统计, 2013, 30(1): 135-139+142.

[17] Lee JY, Styczynski MP. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. Metabolomics, 2018, 14(12): 1-12.

[18] 张婵. 一种基于支持向量机的缺失值填补算法. 计算机应用与软件, 2013, 30(5): 226-228.

[19] 张晓琴, 程誉莹. 基于随机森林模型的成分数据缺失值填补法. 应用概率统计, 2017, 33(1): 102-110.
[20] 李洪飞. 高维缺失数据因果推断方法研究. 南华大学, 2020:45-46.
[21] 陈文松, 刘曼, 刘玉秀, 等. 观察性研究中三种控制混杂偏倚的匹配方法比较. 中国卫生统计, 2022, 39(3): 322-328.
[22] 麦炜琪. 因果推断中的 GAN 技术及应用. 华南理工大学, 2019: 49-50.
[23] Linden A, Yarnold PR. Using machine learning to evaluate treatment effects in multiple - group interrupted time series analysis. Journal of Evaluation in Clinical Practice, 2018, 24(4): 740-744.
[24] Hu L, Ji J, Li F. Estimating heterogeneous survival treatment effect in observational data using machine learning. Statistics in Medicine, 2021, 40(21): 4691-4713.
[25] Athey S, Imbens GW. Machine learning methods for estimating heterogeneous causal effects. Stat, 2015, 1050(5): 1-26.
[26] Yang J, Dahabreh IJ, Steingrimsson JA. Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. Biometrics, 2021:78(2):624-635.
[27] Watson JA, Holmes CC. Machine learning analysis plans for randomised controlled trials: detecting treatment effect heterogeneity with strict control of type I error. Trials, 2020, 21(1): 1-10.
[28] Luo J, Wang G, Li G, et al. Transport infrastructure connectivity and conflict resolution: a machine learning analysis. Neural Computing and Applications, 2021: 1-17.
[29] Chang NC. Double/debiased machine learning for difference-in-differences models. The Econometrics Journal, 2020, 23(2): 177-191.
[30] Biewen M, Kugler P. Two-stage least squares random forests with an application to Angrist and Evans(1998). Economics Letters, 2021, 204: 109893.
[31] Lall S, Ray S, Bandyopadhyay S. LSH-GAN enables in-silico generation of cells for small sample high dimensional scRNA-seq data. Commun Biol. 2022,5(1): 577.
[32] Herlands W, McFowland III E, Wilson AG, et al. Automated local regression discontinuity design discovery. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2018: 1512-1520.
[33] Pearl J, Mackenzie D. The book of why: the new science of cause and effect. Basic Books, 2018:11-12.
[34] 石庆喜. 因果图学习与推理算法研究. 重庆大学, 2005:101-102.
[35] Kang Q, Song X, Xin X, et al. Machine Learning-Aided Causal Inference Framework for Environmental Data Analysis: A COVID-19 Case Study. Environmental Science & Technology, 2021, 55(19): 13400-13410.
[36] Samii C, Paler L, Daly SZ. Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia. Political Analysis, 2016, 24(4): 434-456.
[37] Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. American Journal of Epidemiology, 2017, 185(1): 65-73.
[38] Chin A. Regression adjustments for estimating the global treatment effect in experiments with interference. Journal of Causal Inference, 2019, 7(2): 20180026.
[39] Averitt AJ, Vanitchanant N, Ranganath R, et al. The Counterfactual χ-GAN: Finding comparable cohorts in observational health data. Journal of Biomedical Informatics, 2020, 109: 103515.

(责任编辑:郭海强)

(上接第 309 页)

[6] R Core Team(2020). R:A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[7] Drew AL, Jeffrey BL (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, 42(10), 1-29. URL http://www.jstatsoft.org/v42/i10/.
[8] Ma J, Xu L, Jordan MI. Asymptotic convergence rate of the EM algorithm for gaussian mixtures. Neural Comput, 2000, 12(12): 2881-2907.
[9] Dayton CM. Latent Class Scaling Analysis.. Thousand Oaks, CA: SAGE Publications., 1998, 12-43.
[10] Weller BE, Bowen NK, Faubert SJ. Latent Class Analysis:A Guide to Best Practice. Journal of Black Psychology, 2020,46(4):287-311.
[11] Collins LM, Lanza ST. Latent class and latent transition analysis : with applications in the social, behavioral, and health sciences. Hoboken, N.J.:Wiley, 2010, 23-47.
[12] 王培刚, 梁静, 张刚鸣主编. 多元统计分析与 SAS 实现. 武汉:武汉大学出版社, 2020:193-223.
[13] 王孟成主编. 潜变量建模与 Mplus 应用·基础篇. 重庆:重庆大学出版社, 2014:14-40.

(责任编辑:邓 妍)