机器学习用于科研
——拉响“可重复性危机”的警报

代洁琼
武汉美捷登生物科技有限公司

当今的人工智能技术正在迎来一个“春天”。2018 年度的“图灵奖”，AlphaGo 在围棋上的出色表现，以及 AlphaFold2 在蛋白质结构预测上的巨大突破，都激发了科学家们对这一领域的热情，以至于一度让机器学习的研究“火”出了圈。机器学习为人们提供了一种可以模拟、甚至是创造“人类智力”的工具。从生物医学到政治科学领域，研究人员越来越多地使用机器学习这一工具，根据数据建立模型进行预测。但根据新泽西州普林斯顿大学研究人员的报道，许多这类研究中的结论可能有被夸大的“嫌疑”。他们对机器学习领域中所谓的“正在酝酿的可重复性危机”发出了警报。

DOI: 10.14218/MRP.2022.904
通讯作者：代洁琼 Email: wendy_dai2022@163.com

什么是可重复性危机？就是一个实验的研究结果很难在另一个科研团队中重复实现。近数十年来，许多领域都出现了研究结果无法重现的现象，2016 年韩春雨博士在 *Nature Biotechnology*上发表的论文，引发了可重复性这一话题的巨大争议。同年在 *Nature*的一项调查报告中指出有超过 70% 的研究者尝试但未重现出另一位科学家的实验结果，且一半以上的研究者未能重现自己的实验。现在，机器学习也同样面临着研究结果的可重复性危机。

形成可重复性危机的原因

2020 年新冠病毒席卷全球，面对激增的患者人数，人们缺乏精确的检测和治疗方法。也许人工智能可以更早地在肺部图像上检测到这种疾病，并预测哪些患者最有可能患上重病——带着这样的预期，数百项研究如雨后春笋般涌现，它们声称并证明人工智能能够高精度地完成这些任务。但是，英国剑桥大学的一个科研团队对这些总计超过 400 个的模型进行调查，得到一个惊人的结论：每个模型都有致命的缺陷。当实验设计的基本逻辑受到质疑时，可重复性又从何谈起？导致这个现象的根本原因在于研究人员与同行评审并没有完全掌握人工智能这项技术，而现代人工智能是建立在机器学习基础上的。

数据泄露是机器学习应用时最常见的问题。什么是数据泄露？数据泄露存在于机器学习其本身的应用，就是用于训练机器学习算法的数据集中包含了一些将要预测的事物特征，也就是说测试数据中的一些信息泄露到了训练集中。如果不能将这些将要预测的数据和训练数据集分开，则会导致模型在训练数据集和验证数据集都表现得“非常好”，但是在真实世界却表现得“一塌糊涂”。此外，对机器学习算法相关知识的缺乏、对研究数据的了解不足、对研究结果的误判等都是可能造成可重复性危机的因素。

可重复性危机造成的后果

可重复性危机的出现，造成的最直接的后果就是人们无法分辨所观察到的现象，是真实的还是虚构的，亦或纯属巧合的。科学的目的是尽可能准确地建立事实，而当你无法辨别真伪时，研究所得出的结果是否还能称之为科学？并且伴随着人们对于“真、假”的争论，造假的标签也悄然印在发布争议性结果的研究人员、甚至其所处的学科之中，这将会造成一场巨大的信誉危机。正如上个月震惊神经科学领域的阿尔兹海默症研究造假事件，即使被认定造假的图片仅涉及 Aβ*56（并不是该领域的研究主流）这一种 Aβ 寡聚体形态，但依旧对阿尔茨海默症研究领域处于主流地位的 Aβ 假说造成了冲击，人们甚至开始无差别地质疑与其相关的研究结果。

可能的解决方法

人们常常说发现问题后，就要解决问题。普林斯顿大学机器学习研究员 Sayash Kapoor 及其同事提出了可以警惕的八种主要类型的数据泄漏。他们提出的“数据清单”可以帮助科研人员尽早发觉可能存在的数据泄露。英国伯明翰大学临床眼科医生 Xiao Liu 为涉及人工智能的研究制定了报告指南。该指南可协助监管机构辨别研究人员的工作质量（好或者差）。*Nature Computational Science*杂志在发表的文章中指出，机器学习类研究向社会公开提供代码和数据对提升研究方法的可重复性至关重要，其中包括训练、验证和测试模型的代码及数据收集、清理和整理步骤的代码。同时，当我们在实际建模过程中，不知道某个特征和目标变量存在的是因果关系还是相关关系时，可以进行进一步的数据探索，利用相关系数矩阵热力图、特征分布分析、分组箱型图等方法来防止建模中可能发生的数据泄漏。

总结

从过去到未来——机器学习以其可预见性的优势在科学界得到了大量应用。即便是可重复性危机会使其蒙上一层阴霾，但随着研究人员不断提出对危机的应对措施，相信机器学习依旧会是科学界“火热”的应用工具。

参考资料

[1] https://www.nature.com/articles/d41586-022-02035-w.
[2] https://www.hpcwire.com/2019/02/19/machine-learning-reproducability-crisis-science/.
[3] https://blogs.nvidia.com/blog/2019/03/27/how-ai-machine-learning-are-advancing-academic-research/.
[4] https://www.statnews.com/2021/06/02/machine-learning-ai-methodology-research-flaws/.
[5] https://thegradient.pub/independently-reproducible-machine-learning/.
[6] https://machinelearningmastery.com/data-preparation-without-data-leakage/.
[7] https://blog.csdn.net/lomodays207/article/details/87607569
[8] https://zhuanlan.zhihu.com/p/246482947.
[9] https://new.qq.com/omn/20220802/20220802A07Z3V00.html.
[10] https://www.nature.com/articles/s43588-021-00152-6.