

Frontiers of Information Technology & Electronic Engineering  
 www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com  
 ISSN 2095-9184 (print); ISSN 2095-9230 (online)  
 E-mail: jzus@zju.edu.cn



观点:

## 人工智能与统计分析\*

郁彬<sup>†1,2</sup>, Karl KUMBIER<sup>1</sup>

<sup>1</sup>加州大学伯克利分校统计系, 美国加利福尼亚州伯克利, 94720

<sup>2</sup>加州大学伯克利分校电气工程与计算机科学系, 美国加利福尼亚州伯克利, 94720

<sup>†</sup>E-mail: binyu@stat.berkeley.edu

投稿 2017-12-7; 录用 2018-1-10; crosscheck 2018-1-28

**摘要:** 人工智能 (artificial intelligence, AI) 本质上是由数据驱动的。在其通过人机协作完成数据生成、算法开发与结果评估的任务中, 需要应用许多统计概念。本文讨论了如何通过数据产生、兴趣问题探究、训练数据代表性和对结果审视等环节 (Population, Question of interest, Representativeness of training data, Scrutiny of results, PQRS) 来解决人机协作的问题。PQRS 的工作流程为融合统计分析的思想与人类输入提供了一个概念框架。这些统计分析的思想包括通过随机化、局部控制以及稳定性的原则来获得算法和结果的可重复性与可解释性。我们讨论了这些原则在自动驾驶、自动医疗以及作者其他合作研究中的应用。

**关键词:** 人工智能; 统计; 人机协作

本文译自 Yu and Kumbier “Artificial intelligence and statistics” (*Front Inform Technol Electron Eng*, 2018 19(1):6-9.

<https://doi.org/10.1631/FITEE.1700813>)

中图分类号: TP391; C8

现代人工智能至少可以追溯到 1943 年, 即人们广泛重视研究神经事件与命题逻辑之间联系 (McCulloch and Pitts, 1943) 的时期。多年来, 人工智能已经发展成为一个将计算机科学、统计学/机器学习、心理学、神经科学、材料科学、机械工程和计算机硬件设计等理念整合并转化的跨学科领域。如今, AI 热潮正在掀起。来自 AI 领域的新思想构成了初创公司和学术部门的核心, 其最新发展也越来越频繁地出现在媒体报道中。这一热潮可部分归

因于现有 AI 产品在全球消费者中获得的成功经验。人工智能的“魔力”可以让人着迷, 像亚马逊 Echo 这样的 AI 新产品, 几乎能够毫不费力地响应并深入解答用户查询。然而, 只要认识到这些详细解答几乎不会超出维基百科的文章内容, 隐藏在人工智能的“魔术”背后的大量人力投入也就不言自明。

亚马逊的 Echo 是一款使用无线网络连接在互联网上搜索信息的智能音响。这些信息通常是人类编辑的文字、录制的言语和创作的音乐。换言之, Echo 的响应源于人机协作: 利用亚马逊研究人员 (使用强大计算与信息处理技术) 设计并测试的计算机算法分析人类生成的数据。类似的, 基于计算机视觉的 AI 产品同样依赖强大的人机协作, 即通过研究人员设计的深度学习算法以及人为构造数据 (例如包含超过 1000 个类别、1400 万张带标签图像的 ImageNet 数据集) 来完成。目前的人工智能研究在

<sup>‡</sup> 通讯作者

\* 本文得到以下项目的资助: the Army Research Office (No. W911NF1710005), the National Science Foundation (Nos. DMS-1613002 and IIS 1741340), the Center for Science of Information, a US National Science Foundation Science and Technology Center (No. CCF-0939370), the National Library of Medicine of the NIH (No. T32LM012417)

ORCID: Bin YU, <http://orcid.org/0000-0002-6803-7964>

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2018

数据生成、算法设计与测试阶段都依赖于人机协作。例如，作者的研究组（Yu Group）与加州大学伯克利分校的 Gallant 神经科学实验室合作，将经过 ImageNet 数据训练的卷积神经网络（convolutional neural networks, CNNs）与回归方法相结合，以表征灵长类动物视觉皮层区域 V4 中神经元的研究。

当前最引人瞩目的人工智能应用包括自动驾驶以及自动医疗，例如使用计算机断层扫描（computed tomography, CT）技术鉴别中风原因（<https://www.newyorker.com/magazine/2017/04/03/ai-versus-md>）。这两个应用很大程度上依赖于计算机视觉算法，而这些算法又依赖于手工生成的数据。Tim Bradshaw 先生在其 2017 年《金融时报》文章（<https://www.ft.com/content/36933cfc-620c-11e7-91a7-502f7ee26895>）中宣称：“事实证明，对人类来说汽车的自动驾驶是项劳动密集型技术。”他接着描述说，大多数自动驾驶公司雇用成千上万的人来标记视频片段，用以训练算法识别行人等障碍。然后，他引用了 Mighty AI 创始人兼首席执行官 Matt Bencke 的话说：“在我看来，AI 从业者共同拥有一个傲慢的盲点，那就是电脑将解决所有问题。”

对 AI 产品来说，对数据的收集和分析进行恰当构建至关重要。这一构建过程的实现采用了包含 PQRS（数据产生、兴趣问题探究、训练数据代表性和审视结果）的统计学框架，并通过人机协作完成。PQRS 工作流程也代表了实现数据驱动型决策的关键步骤。该流程是本文第一作者在伯克利共同创建和教授本科数据科学课程过程中创造的（<http://www.ds100.org/SP17/>）。数据生成（P）反映了生成数据的观测条件。理解 P 有助于识别数据生成过程中的随机性，从而识别数据结果中的不确定性（或错误）。兴趣问题探究（Q）为分析数据提供了背景，并允许人们结合领域知识。训练数据的代表性（R）与 P 密切相关，用于评估可用的训练数据是否提供了有关整体数据的相关信息（相对于所提问题）。针对整体数据是否改变，或者训练和测试数据是否相似思考过程同时明确了 P 和 R。最后，对结果的审视（S）描述了在 PQR 的上下文中评估数据结果或算法输出的过程。

PQRS 工作流程为数据驱动决策（包括自驾车

和自动医疗诊断所需的决策）提供了 4 个具体步骤，用以充分考虑数据分析和算法开发的循环过程。例如，可以从 PQRS 的视角研究动态的天气、交通和施工条件对行人识别的影响。同样，可以使用 PQRS 的步骤来学习患者特征（如年龄、性别和先前的医疗状况）与自动医疗诊断之间的关系。这些步骤的完成需要人类的参与，包括了解问题背景的该领域专家以及能够掌握数据结果的分析师。这一数据结果总会应用在新的个体或情况中。通过构建数据的采集分析过程，可以避免因行文不当而导致的错误答案。而在自动驾驶和自动医疗诊断中，这些错误答案可能是致命的。PQRS 提供了一些有效的概念工具，将人员投入集成到人工智能应用中，从而将人工智能的“魔力”从失败中挽救出来，正面应对动态环境带来的挑战。

PQRS 工作流程最后一部分（S）基于可解释性概念评估数据结果。可解释性有多种形式，包括但不限于算法可解释性（即算法如何将特征映射到结果空间）和领域可解释性（即数据结果在特定问题的上下文中说明什么）。人的参与在这里同样至关重要，因为可解释性必须相对于个体来定义（例如，专家与非专家）。在自动化医疗诊断领域以及更广泛的领域，算法和数据结果的人类可解释性已成为必需。事实上，欧盟“通用数据保护条例”（2016 年）规定用户有权了解算法和数据对结果的解释。因此，自动医疗诊断算法必须向医生和患者提供可解释性。

目前，许多在 AI 产品中广泛使用的监督学习算法无法得到很好解释，例如著名的深度学习算法，即便对其研究人员来说也很难解释，尽管这些算法展现了最先进的预测性能。Yu (2013) 提倡使用稳定性原则来提高可解释性并提高算法和数据结果的可重复性。这一原则在概念上简单易懂，而且在实际应用中很容易使用。它为始于 20 世纪 40 年代的无数研究工作提供了一个统一的标准，为开发新的基于稳定性的方法提供了一个平台。这一原则一方面将知识稳定性的哲学原理与科学的可重复性原理结合起来，另一方面又与统计推断或不确定性评估相联系。稳定性原则的使用要求人为输入同时确定数据和（或）模型的适当扰动以及稳定性度量。例如，深度学习算法对于基于预测的度量是稳定的，但对于依赖于拟合权重的可解释性度量则不是。适当性

是一个重要概念，应该由人类在数据生成过程中和领域知识方面进行仔细判断。

对于与自动医疗诊断相关的算法开发，至少有两种形式的数据扰动看似适当。一种是使用来自训练组中所有患者的所有 CT 扫描的子样本，并研究算法输出相对于不同子样本的稳定性。另一种方法是在扫描中添加少量噪声以查看诊断结果的变化情况。可接受的不稳定水平范围是用户在上下文语意以及与诸如医生等领域专家的合作中必须建立的领域问题。作为不确定性的量度，在将诊断结果传达给病人时，这一问题必须是要考虑到的。

对于监督学习算法，其预测方法极难理解，可采用稳定性原则进行解释，使人们更容易查看并理解结果。例如，作者的研究小组将稳定原则纳入当前基因组学工作中，以识别候选的调节交互。具体而言，该小组通过迭代随机森林 (iterative random forests, IRF) 算法 (Basu et al., 2018) 稳定随机森林决策路径，以恢复通过当前流行的监督学习方法得到的高阶非线性交互操作。该算法通过决策树的阈值机制，整合了关于生物分子相互作用阈值化现象的领域知识 (Wolpert, 1969)。IRF 凭经验论证了稳定性原理的价值，确定了一组高质量的稳定相互作用，其中 80% 成对相互作用在先前果蝇基因组学实验中报道过，有望为发现系统生物学前沿的三阶或更高阶相互作用而开展的实验工作提供有效指导。需要注意的是，本项目中对结果的审视需要稳定的、可解释的相互作用，以及和人类生成的 wet-lab 数据来评估配对结果的质量。

从稳定性角度出发，因果效应也可以视为数据生成过程的可解释稳定机制。为了帮助医生确定药物治疗方案，使用随机实验 (或 A/B 测试) 将患者分配到治疗和对照组，并评估药物效果。这为用于因果推断的有效数据收集 (Imbens and Rubin, 2015) 提供了统计实验设计的随机化原则。为实现个性化或精确的医疗诊断和治疗，最好找到与所考虑患者类似的较小亚组患者，并对该组进行稳定性分析。这种类型的分析代表了统计实验设计的“局部控制”原则的一个实例，通过考虑与结果相关的患者特征，降低由调节或分组引起的不确定性或变异性。这是一个具有挑战性的命题，因为即使使用“大数据”，也很难找到病人分组的相关维度，而这样的

组织可能很小，估计能力较低。再次，解释算法的输出，以便领域专家对相关问题进行审查，可为决策过程提供协助。

数据驱动型决策是 AI 的核心。这类决策通常依靠人的参与，尤其对于尖端 AI 产品，如亚马逊 Echo、汽车自动驾驶和自动化医疗诊断。这类特定产品对于人为参与的依赖可能会减少。但是，对人为参与的需求还会在新的 AI 应用中得到体现。PQRS 工作流程提供了一种通过简单的统计思想——包括随机化和局部控制的实验设计原则 (Box et al., 2005) 以及稳定性原理——将人的参与纳入 AI 产品的方法。将这些概念集成到分析中，可以有效且高效地收集和使用数据，并获得 AI 算法和数据结果的可解释性和可重复性。作者认为它是人工智能的圣杯，能够再现人类智能尚未明确定义的无意识思维。在未来的 AI 领域，人类和统计学将继续发挥不可缺少的作用。

## 致谢

感谢来自 Bryan Liu 和 Rebecca Barter 的有益意见。

## 参考文献

- Basu S, Kumbier K, Brown JB, et al., 2018. Iterative random forests to discover predictive and stable high-order interactions. *PNAS*, 115(8):1-6. <https://doi.org/10.1073/pnas.1711236115>
- Box GE, Hunter JS, Hunter WG, 2005. *Statistics for Experimenters: Design, Innovation, and Discovery* (2<sup>nd</sup> Ed.). Wiley-Interscience, New York, USA.
- Imbens GW, Rubin DB, 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, UK. <https://doi.org/10.1017/cbo9781139025751>
- McCulloch WS, Pitts W, 1943. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*, 5(4):115-133. <https://doi.org/10.1007/BF02478259>
- Wolpert L, 1969. Positional information and the spatial pattern of cellular differentiation. *J Theor Biol*, 25(1):1-47. [https://doi.org/10.1016/s0022-5193\(69\)80016-0](https://doi.org/10.1016/s0022-5193(69)80016-0)
- Yu B, 2013. Stability. *Bernoulli*, 19(4):1484-1500. <https://doi.org/10.3150/13-bejps14>



Bin Yu 是加州大学伯克利分校统计系和电气工程与计算机科学系的校长讲座教授。她目前致力于解决高维数据问题的统计和机器学习理论、方法和算法研究。她的研究小组与基因组学、神经科学和遥感领域的科学家们合作，从事跨学科研究。

她于 1984 年在北京大学获得数学

学士学位，1987年和1990年在加州大学伯克利分校分别获得统计学硕士和博士学位。她于2009至2012年担任加州大学伯克利分校统计系主任，兼任中国北京大学微软统计与信息技术实验室联合创始主任，并担任北京大学统计科学中心科学咨询委员会主席。

她是美国国家科学院院士和美国艺术与科学院院士。她于2006年成为“古根海姆（Guggenheim）学者”，2011年成为ICIAM特邀演讲嘉宾，2012年成为伯努利协会的“Tukey纪念

讲师”（Tukey Memorial Lecturer）。她于2013–2014年担任数理统计研究所（IMS）主席，2016年担任IMS的Rietz讲师。她是IMS、ASA、AAAS和IEEE会士。

她曾担任NAS数学科学与应用委员会委员，并担任SAMS I咨询委员会联合主席。现任ICERM理事会和IPAM科学顾问委员会委员。曾经或正在担任许多期刊的编委，包括《机器学习研究》（JMLR）、《统计年鉴》（Ann Stat）和《美国统计协会会刊》（JASA）。