

统计学在人工智能发展中的作用及其应用展望

李志勇¹, 焦微玲²

(1.江苏省社情民意调查中心, 南京 210000; 2.盐城工学院 经济管理学院, 江苏 盐城 224002)

摘要:统计学贯穿于人工智能开发与应用的整个过程。为整合统计理论、模型与方法改进和创新人工智能算法及其结果,并推动人工智能新发展,文章基于数据科学理论与方法深入研究统计学在人工智能发展中的作用及其应用。研究表明,统计学有助于人工智能系统和算法确定研究问题、优化数据收集方案、评价和提高数据质量、推动从关联分析到因果分析、量化不确定性和提高结果的可解释性。

关键词:统计学;人工智能;数据科学;统计模型与方法;作用与应用

中图分类号:C829 **文献标识码:**A **文章编号:**1002-6487(2023)21-0046-06

0 引言

人工智能(Artificial Intelligence, AI)指制造智能机器的科学和工程,强调创造能够在各种新颖和无法预知的情况下模拟、复制和拓展人类智能以有效行动的机器。随着现代人工智能进入快速发展阶段,除伦理、道德、安全、法律、隐私、社会治理规则等问题引起广泛关注外,人工智能在数据获取和处理、因果推断、不确定性、结果可解释性等方面存在的问题也引发大量讨论。统计学提供了大量的发现数据结构、探索数据内涵并进行预测的工具和方法,是分析和量化不确定性最重要的学科之一。数据科学是从数据中提取信息,辅助发现知识并支持决策的一门科学,统计学是数据科学的基础,为数据科学提供了基础理论和方法,人工智能是数据科学的一个应用领域。因此,基于数据科学全流程正确认识统计学在人工智能中的作用与应用,有助于整合统计学模型与方法,改进人工智能算法及其结果,进一步推进人工智能发展,更有助于统计学与人工智能实现融通共进,共同促进数据科学发展。

1 统计学与人工智能的关联性

人工智能是关于知识的学科——怎样表示知识以及怎样获得知识并使用知识的科学,其核心因素是数据、算法和算力。早在20世纪50到60年代,人工智能就开发了包括感知机(Perceptron)算法、反向传播(Backpropagation)算法等第一代数据驱动算法。20世纪90年代人工智能再次得到快速发展,支持向量机(Support Vector Machines)、随机森林(Random Forest)、贝叶斯方法(Bayesian Methods)

等的提出奠定了人工智能机器学习的方法和思路。随着深度置信网络算法(Deep Belief Networks)、生成对抗网络模型(Generative Adversarial Networks, GAN)和深度GAN的提出,人工智能神经网络学习快速发展,人工智能网络结构更具有颠覆性。

人工智能技术及其应用的许多突破源于计算机科学,而统计学在人工智能的整个发展过程中做出了巨大贡献^[1]。统计学促进了人工智能理论研究与实际应用的发展,许多统计理论在人工智能领域有着重要应用,统计思维影响着人机协作数据获取、算法研发和结果分析,统计模型与方法则为人工智能算法奠定了坚实基础并在人工智能系统开发设计、算法完善等方面发挥了重要作用。例如:反向传播和非线性最小二乘法之间关系的实现^[2];径向基函数网络可看作统计学中的非线性回归模型来研究;一致性说明、泛化边界等机器学习方法的理论有效性需要大量的数理统计和概率论基础知识;超级机器学习或宽度学习系统与多元回归、岭回归等线性数据统计方法密切相关;在语音识别以及文本分析和翻译任务中,统计学中擅长语音识别、文本分析和翻译的隐马尔可夫模型得到使用和进一步发展,并取得巨大成功;基于惩罚或鲁棒估计方法改进机器学习算法实现超级机器学习等。

因此,人工智能和统计学有着极强的关联性。目前人工智能培养思维能力的基本途径是机器学习,机器学习是指“通过使用计算机算法自动发现数据中的规律”,2000年开始,随着数据规模的增加,机器学习开始更多地统计方法将统计学方法和算法用于模拟机器智能,通过机器学习实现统计学习。还有学者认为人工智能是统计学的应用,有的学者认为统计学是人工智能的重要基础,而诺贝尔经济学奖获得者 Thomas J. Sargent 则认为人工智能其实就是统计学。

基金项目:全国统计科学研究重点项目(2020LZ16)

作者简介:李志勇(1975—),男,江苏南京人,高级统计师,研究方向:应用统计。

焦微玲(1973—),女,江苏盐城人,博士,教授,研究方向:产业数字化转型。

2 统计学在人工智能中的作用与应用

人工智能本质上是数据驱动的,而统计学是从数据中发现规律的学科,两者都在数据科学中发挥着核心作用。根据 CRISP-DM (Cross Industry Standard Process for Data Mining, 跨行业数据挖掘标准流程) 的规定,数据科学工作主要包括业务理解、数据理解、数据准备、建模、评估和部署六个步骤,每一个步骤都直接关系到结果的有效性和可靠性。统计学在数据测度和评估方面具有极强的专业性,在确定研究目标和问题、设计研究方案、进行数据收集与分析、分析和解释研究结果方面均有独特科学的方案。而人工智能往往更专注于数据分析而较少关注其他步骤,这可能导致在关键问题上产生误导性解释。因此,本文结合现有研究对数据科学的定义和工作步骤,从研究问题、数据收集、数据质量、数据分析、结果解释等方面系统探讨统计学在人工智能中的作用与应用。

2.1 确定研究问题

人工智能的核心是机器学习,根据数据库中的知识发现概念,机器学习从给定的数据或公开数据中提取信息和知识,对经验的依赖性很强,结论的有效性和普适性也因此受到影响。统计学的研究问题从理论问题中衍生出来,实证研究检验的变量具有概念化和可操作性,分析遵循“初步的定性认识—科学的定量分析—理性的定性认识”过程,研究结果服务于特定目标。通过背景分析、确定研究问题进而形成初步的定性认识,将为科学的定量分析奠定良好的基础,从而消除系统性误差、减少随机误差,提高分析推断精度,使得结论更加有效和实用。

确定研究问题还有助于选择更合适的理论框架和方法。统计学具有处理随机性或部分随机性问题和数据的丰富经验,这些数据和问题更加关注概率、误差等方面,具有确定性背景的图论结构数据和问题则更加关注鲁棒性或敏感性。

2.2 设计数据收集方案

2.2.1 人工智能中的数据及其特征

人工智能背景下的数据通常具有以下特征:(1)常规数据。人工智能不会基于特定目标或特定研究问题收集数据,例如出于日常存储或记账目的而收集数据。将这些常规数据用于科学研究是人工智能的典型应用,例如使用医疗账单数据预测医疗事故。这样的数据大多是二手数据,不但存在大量冗余数据,而且数据的逻辑关系模糊。(2)方便数据。被用于科学研究的数据并非随机抽取的样品,而是基于可获取性收集的方便样品,例如在线问卷收集的数据仅仅是由访问该主页并愿意花时间填写问卷的用户提供的。这样的样本数据往往缺乏代表性,可能导致系统性统计误差。(3)高维数据。医疗数据、城市数据、体育数据等往往是高维数据,人工智能具有处理高维数据的先天优势,但是输入高维数据有可能观察到大量特征值,复杂且具有相互作用的非线性关系也经常被用于预测。

高维数据不但会导致不确定性增加,而且还可能面临“维数灾难”,一方面,数据降维会导致稀疏性,即高维空间中的样本变得极度稀疏,造成过度拟合;另一方面,有时即使输入数百万量级的数据,也只有少数变量跟结果相关。较简单的模型无法捕捉高维数据的复杂性,过于复杂的模型和稀疏的训练数据又容易导致训练误差较大,从而影响预测效果。

2.2.2 人工智能数据特征对研究结果的影响

数据收集方案是整个研究顺利开展的基础,由于人工智能存在上述数据特征,因此如果缺乏精心设计的数据收集方案,将会影响人工智能系统或算法。

(1)效度。包括内部效度和外部效度,从而影响研究结果的唯一性、普适性和可推广性。效度无疑是非常重要的,人工智能的通常做法是基于大规模 benchmarking 测试数据集验证该算法模型的优越性,由于人工智能程序开发是动态迭代的且周期越来越短,尤其是在与移动应用程序或在线学习系统相关的领域,相应系统、算法和模型的验证难度越来越大,因此需要对其不断地进行有效性验证。

(2)代表性。实践证明并非数据规模越大代表性就越强以及带来的预测率更高,而是可能会带来“假规律”和“伪相关”,导致研究结果失真。例如,深度学习中基于相关性的理论很难高效获取知识,在认知智能方面的效果有限,同时还出现了两个典型悖论:辛普森悖论和伯克森悖论。

(3)模型质量和效果。通常用偏差、方差、欠拟合、过拟合来判断和描述人工智能模型的质量。一般来说,当数据集数据规模足够大且准确时,人工智能模型越复杂,偏差和方差越小,模型拟合效果越好。但是对于有噪声的数据集来说,更高的模型复杂度并不意味着更低的估计误差。由于量化、传感器限制等因素,实际生活中的数据通常是含有噪声的。此外,数据的收集方式,例如时间轴、人工智能中的种族、性别、年龄地位等偏见也会导致结果产生偏差。所以,缺乏良好的数据收集方案可能导致测试数据集的错误较多,人工智能无法捕捉数据潜在分布,结果往往带有某种倾向或偏见^[9]。

(4)稳定性。无论是预测模型还是描述性模型,稳定性都非常重要,主要反映为模型在面对输入数据的微小变动时是否依然能保持判断的准确性,通常用鲁棒性来评价和衡量,鲁棒性的高低直接决定了机器学习模型的泛化能力。当模型缺乏鲁棒性时,在现实中面对纷繁复杂的应用场景时,模型的性能会大打折扣,甚至不可用。

此外,结果还可能面临“可重复性危机”(Replication Crisis)。结果的可复制性是科学方法的重要组成部分,无法复制的研究可能会对许多科学领域产生严重影响。事实上,许多研究,特别是在医学和社会科学领域,其结果是难以或不可能复制的。因此,研究人员开展大量研究来验证以前的发现,以确定这些结论的可靠性。

2.2.3 基于统计学设计数据收集方案的结果

采用统计学的概念、技术、模型、方法等可以优化数据

收集方案,包括样本设计、样本规模和权重确定、数据集限制、抽样误差控制等,从而有助于人工智能系统和算法的优化。

(1)提高模型的有效性和准确性。独立数据检验、外部验证等在人工智能中起着至关重要的作用,但是在许多应用中,最初算法被替换从而导致外部数据评估阶段从未实现。统计学提供了可以验证人工智能模型内部和外部有效性的设计,例如使用灵敏度、特异性、ROC曲线和校准曲线等质量测度统计指标对人工智能模型进行评估,数据生成过程建模、数据集的限制、析因实验法等统计模型、方法和技术则有助于收集和处理人工智能程序所需要的训练数据。

(2)提高样本的准确性和代表性。借助统计技术、模型和方法可以提高人工智能训练样本的表现,一方面控制抽样误差,增加数据样本,通过抽取足量准确样本减少噪声;另一方面加强数据代表性,在确保训练样本量充足的同时保证足够的特征。

(3)优化模型的拟合效果。统计学提供了最小化统计误差的方法、技术和原则,例如统计控制、设计控制等优化设计方法,(部分)随机化、(部分)盲法、匹配等专为数据收集设计的技术,以及偏好风险评估、层别法、边际分析、元分析等不同领域的应用案例。此外,统计学通过对样本的学习进行拟合,从而求得较好的参数集拟合,最小二乘法、无偏估计、有偏估计等统计思想和概念在模型的拟合优化中得到了充分体现和应用。需要注意的是,机器学习高维空间结构处理高维稀疏数据需要大量训练数据,还需要大量统计模型和相应的数学近似值或数值模拟,更需要具有高级统计素养和经验丰富的统计学专家评估训练数据用于人工智能应用程序的可能性、限制条件和数量。

(4)增强算法的稳定性。统计学提供的变量选择方法可以提高人工智能算法的稳定性。模型的复杂程度影响模型的变量选择方法,而变量选择可能会影响模型的稳定性、回归系数的无偏性、P值或置信区间的有效性以及对模型的解释。因此,许多学者强调了稳定性调查的重要作用,并在人工智能算法中引入统计学概念,认为稳定性调查和推断是比较变量选择策略的主要目标参数之一^[4]。

(5)提升模型的鲁棒性。好的数据收集设计方案可以抵消所谓的可复制危机,为了使人工智能系统和算法能更好地应用于真实场景,需要提升模型的鲁棒性,并将鲁棒性作为模型评估的必要维度。

2.3 评价和提高数据质量

2.3.1 数据质量在人工智能中的重要性和影响

数据质量和相关性在所有数据分析中都非常重要,直接关系到研究结果的效度、代表性、偏差、稳定性、可复制性等。人工智能系统或算法的成功从根本上取决于数据质量,将高质量数据与数据思维相结合可以产生新的、应用范围更加广泛的算法^[5]。但人工智能算法从数据湖中提取相关测量数据,即从高维数据中提取相关的影响变量,“维数灾难”不可避免。此外,数据量大不一定就具有代表

性,足够多的数据并不能自动导致理想的预期。

数据质量对人工智能系统和算法的影响主要表现在以下两个方面。

(1)准确性。随着数据收集自动化程度的提高,人工智能能够以经济、高效且简便的方式提高测量精度,但是否能够实现数据质量其他维度的目标还有待验证。如何处理缺失值是一个常见问题,人工智能只能进行正向思维,根据所输入的数据进行学习推断,而且其数据收集方法允许冗余数据存在,虽然对冗余数据可以进行预处理,即通过适当的背景知识使数据集变得完整,但首先要能够背景知识恰当地整合到数据提取过程中。

(2)相关性和及时性。人工智能算法常常先假设数据符合一种模型,例如假设基于隐藏在数据中的模式未来将继续存在,再依据数据样本估计模型的部分参数及统计量,以此了解数据特征。但是在实践中,往往有很多数据并不符合假设的模型分布,从而导致数据分析的结果不理想。

2.3.2 基于统计学方法评价和提高人工智能的数据质量

数据质量是统计的生命力,相关性、准确性和可靠性、及时性和准时性、一致性和可比性、可访问性和清晰度是数据质量的重要维度。统计学提供了评价数据质量和提高数据质量的标准和方法。

(1)诊断与测度数据质量。统计学的重要贡献之一是提出了分布的概念,探索性数据分析提供了一系列将数据经验分布可视化并得出适当度量的工具,这些工具可用于检测异常值、确定典型值和标准值、检测和处理缺失值、纠正输入错误,结合数据存储标准,数据测度过程中的误差可以尽早得到检测和修正。

(2)提供评价标准并评估数据质量。通过上述方法,统计学可以帮助评价数据质量,包括系统化、标准化、记录完整化等。此外,统计调查方法主要关注数据质量,统计调查研究通过数据内部和外部的有效性确保数据质量,这为人工智能算法发展奠定了基础。

(3)处理数据缺失问题。统计学思维同时具有正逆向思维,既可根据已掌握数据推断未知数据,也可对缺失数据进行推断。统计调查中缺失数据的现象较为普遍,统计学家针对非随机缺失、随机缺失、完全随机缺失、无应答偏差、选择偏差、测量误差等情况对缺失值处理进行了深入研究,在缺失数据处理方面积累了大量经验,有许多成熟的处理方法和输入缺失数据的程序,例如插补法及其程序、数据增强方法等,人工智能算法可以根据现有的背景和专业处理缺失数据从而实现数据的完整性。

2.4 推动从关联分析到因果分析

2.4.1 人工智能的关联分析

30多年前人工智能研究的一个主要挑战是如何对机器进行编程,从而将潜在原因与一组可观察到的特征值相互关联起来,例如通过贝叶斯网络方案解决该问题。随着人工智能的快速发展,现在已经产生了大量可以完成该项任务的算法和方法,例如,被用于机器人、自动驾驶、计算

机辅助检测和诊断系统、药物研究和发现、农业等领域的深度学习可以根据关联分析发现和揭示大量数据中的事物、特征、结构和数据之间的关联。

除了超强的预测能力,人工智能还具有处理大数据集的优异性能,因此被广泛用于观测、处理、分析和记录工业大数据、医疗大数据、零售大数据等。但这些数据并非按照随机研究设计的严格框架收集,即使两个变量之间存在很强的相关性,也并不等同于他们之间存在因果关系。大量研究表明,如果缺乏因果关系考虑,那么从数据中学习到的相关性可能是错误的,相关性学习模型由于泛化能力和稳定性较差容易受到场景变化或数据中异常值的影响,过度依赖数据拟合的结果往往缺乏可解释性。因此,许多学者认为当前的机器学习是对数据和概率分布曲线的拟合,寻找的是变量之间的相关性,变量的内在因果关系被忽略和简化,算法缺乏处理归纳推理、因果推理和过程解释的能力,这不但使科学研究停留在较浅的关联层面,而且还可能导致模型的鲁棒性和可解释性丧失。

2.4.2 因果分析的重要性

贝叶斯网络之父、人工智能领域的先驱 Pearl(2010)^[6]指出因果和关联的区别:关联主要根据观察变量的联合分布来定义相关关系或反映事物间的相互依存关系;而因果不能单独通过计算数据获得,也不能单独从控制观测数据的分布中确定。他提出了因果阶梯论(Pearl Causal Hierarchy, PCH),认为因果推断有相关、干预和反事实三个层级,而机器学习仅实现了第一个层级——关联,应该使用因果推断模型从因果而非单纯的数据角度开展研究,建议将因果框架与机器学习算法相结合,使机器能够得出因果结论并模拟干预。

科学研究的本质是探究变量之间的因果关系,因果关系是一种稳定的机制,不会随环境变化而变化,而且也只有这种稳定的结构才是可解释的。因果关系可以通过因果推断实现,因果推断是指在现象已经发生的情况下推出因果关系结论的过程,是一种通过事物发生的原因推导至某个肯定结果的逻辑方式,在生物医学、经济管理和社会科学中有着广泛应用。因果推断被认为是人工智能领域的一次范式革命,是近年来人工智能领域的研究热点之一,也是当前学术界面临的难题之一。因果推断既可以发现现象背后的深层原因或究其根源,也可以估计定量的因果效应,分析当原因改变时结果变量的响应,从而帮助人们更科学地做出决策干预,更好地解决生活和科学研究中遇到的问题。

2.4.3 统计学推动人工智能理解因果分析

因果推断是统计学的核心问题之一,学者们进行了大量相关探索与研究。例如,1965年流行病学与医学统计学专家 Bradford Hill 提出一套用来判断某事件与其可能结果的因果联系的标准,并对必要证据提出具体要求,正式确立了随机对照方法,建立多项随机对照临床研究的早期研究方法。1974年著名统计学家唐纳德·鲁宾提出反事实框架^[7],阐述了潜在结果的概念,并将其推广到观察性数据中,开启统计学界对因果推断的广泛研究。潜在结果框

架(Potential Outcome Framework)也叫鲁宾因果模型(Rubin Causal Model),其基于潜在结果框架进行因果推断,是偏向计量的统计分析方法。之后几十年,现代社会科学研究人员基于此框架并借助统计学对因果关系的认识,发展出包括实验与准实验、匹配和倾向值匹配、工具变量、倍差法、断点回归等适用于不同情境的因果推断方法。

(1)控制协变量的不良影响。与随机控制试验不同,协变量效应(Covariate Effect)在随机对照试验中非常重要却常常被忽略。协变量虽然不为实验者所控制,但会影响实验结果,不合理地考虑协变量会导致对因果效应的估计失真,因此在进行随机对照试验研究设计和分析时必须考虑协变量的影响。当随机对照研究出现分组不平衡时,协变量可能带来结果偏倚,此时的协变量变为混杂因素(Confounder),会带来混杂偏倚,导致结论偏倚乃至错误,所以需要正确区分处理因素和混杂因素。有两类统计方法可以用来调整处理分组间潜在或实际存在的不平衡,从而控制协变量带来的偏倚。一是事先控制法,包括完全随机分组、分层随机化、匹配随机化等方法,主要用来强制使处理组间在重要和事先指定的协变量上达到平衡,从而防止实验设计阶段出现不平衡。二是事后控制法,根据协变量的性质和需考虑的协变量数目的不同,采用差值法、协方差分析法、分层分析法、回归分析法等方法对协变量进行校正,从而调整在实验分析阶段出现的不平衡。此外,还可以在评价阶段通过倾向评分匹配、边际结构模型、多元线性回归模型等对混杂因素进行调整和均衡。总之,在观察性数据研究中,识别因果关系比较困难,在进行实验设计和分析评价时,一定要选用正确的统计方法,尽可能地控制和消除混杂因素的影响,改善因果效应估计,提高研究结论的效度和信度。

控制混杂因素、碰撞因素和中介因素对因果效应估计的影响是不同的。与混杂因素会同时影响处理变量和结果变量不同,碰撞因素(Colliders)会同时被处理变量和结果变量影响,对其加以控制会削弱因果关系推理;而中介因素(Mediator)受处理变量影响,同时又影响结果变量,会“阻断”处理变量影响结果变量的路径,控制中介因素会改变对两者之间因果关系的解释,导致只能解释部分因果机制或因果效应估计产生偏差。因此,在数据分析中,特别是在纵向数据分析中,要重视从理论形成视角区分观察数据中协变量与处理和结果变量的关系,从而避免因果效应估计中的偏差。

(2)回答因果问题和模拟干预。因果关系错综复杂,需要考虑的变量可能不计其数,将合适的统计学理论和方法整合到人工智能模型和算法中有助于回答因果问题和模拟干预。现实中最直接、最理想判断干预与效应之间因果效应的估计方法是随机实验。然而由于随机实验不可行、不合伦理或需要很长时间才能完成,有时必须依赖真实世界的观察数据做出因果推论,包括确定因果问题并做出回答。对于应用观察数据的研究人员而言,确定和回答因果问题的一个有效方法是设计可回答该问题的目标实

验方案,包括定义和阐明数据收集标准、分组、因果对比、数据分析计划、因果问题及回答方式等要素,并根据该研究方案进行模拟干预,这也体现了科学设计研究方案的重要性。

(3)提高模型和算法透明度:将统计方法整合用于检测和理解人工智能模型与算法中的因果关系还有助于提高其透明度,从而提高人工智能方法的接受度。已有研究表明,基于因果关系或因果效应的解释比基于概率或统计相关性的解释更有效,也更具说服力。

2.5 量化不确定性

2.5.1 人工智能中的不确定性

由于认知、数据、模型和预测存在不确定性,因此从数据收集、数据预处理、选择模型从数据中学习、选择学习算法训练所需模型,到从所学习的模型中得出推论,人工智能算法的每一个环节都存在固有的不确定性。由于依赖大数据、不确定性评价方法复杂等,因此不确定性问题常常被忽略。事实上为了实现高度准确预测目标,人工智能应用程序可能过于注重底层的性能而忽略中间层和web层的性能,虽然使得程序更灵活,但也让程序的不确定性精确量化变得更加复杂。由于参数的可变性和参数、结构、算法、实验和插值的不确定性,因此人工智能系统做出的预测结果并不完全可靠,在现实中部署时也面临着失败风险。

2.5.2 人工智能的不确定性量化

人工智能的不确定性量化大致包括两类:一类是解决不确定性的正向传播问题的方法,例如:蒙特卡罗模拟、重要性抽样、自适应抽样等基于模拟的方法,使用学习代理模型或完全贝叶斯方法等基于代理的通用方法,泰勒级数、摄动法等基于局部展开的方法,Neumann展开、正交或Karhunen-Loeve展开(KLE)等基于函数展开的方法,一阶可靠性方法(FORM)、二阶可靠性方法(SORM)等基于最可能点(MPP)的方法,全因子数值积分(FFNI)、降维(DR)等基于数值积分的方法。另一类是解决反向不确定性问题的方法,例如:解决偏差校正和参数校准问题的贝叶斯框架、模块化贝叶斯方法、针对模型和差异函数的高斯过程建模、用于计算未知参数的后验分布等。

由于这些方法有的需要严格执行才能得到精确的置信区间,有的基于较为理想的假设条件开发,因此还需要在实践中验证其可操作性和理论有效性。此外,选择特定的不确定性量化方法受到基础模型、学习任务类型、数据特征、学习模型的透明度、最终目标等诸多因素影响,如果使用不当,可能因产生较差的不确定性估计而误导用户。当然,如果沟通不畅,即使是高度准确的不确定性估计也可能产生误导。

2.5.3 基于统计学的不确定性量化

学者们普遍认为综合数据分析无法打破不确定性原理,需要对不确定性进行量化,从而提高估计和预测的有效性和可靠性,防止数据欺骗^[9]。虽然将人工智能算法嵌入统计模型可能影响其灵活度,但是可以通过指定有效的预测区间和置信区间更好地实现不确定性量化或提高结

果的可解释性。例如:在生存分析(time-to-event analyses)中将Kaplan-Meier或Aalen-Johansen非参数估计和机器学习重采样方法结合起来构造出累积发生率函数(cumulative incidence functions, CIF)的同时置信带;在时间序列预测中,将人工神经网络与ARIMA模型混合或者基于多阶层结构有助于得到更好的解释。

研究表明,通过统计学的不确定性量化方法也可以提高人工智能算法的可解释性和效度。例如:在人工智能背景下假设特定的概率统计模型或依存结构并开展鲁棒性调查、渐近一致性或(有限)误差界证明等数学研究,对(随机)模拟设计和易于解释的辅助统计模型进行探索和阐述,对准确性、灵活性、特异性等人工智能算法的质量标准进行详细分析和评价。

2.6 解释结果

2.6.1 可解释性在人工智能中的重要性

由于许多人工智能程序和算法的估计参数难以解释,输入的数据和输出的答案之间存在“黑箱”,因此人工智能往往缺乏透明度和可解释性。可解释性衡量的是用户可以理解模型预测背后原因的程度,从而影响用户对人工智能的信任,并带来算法歧视、算法安全、算法责任等问题。因此,需要进一步研究以增加其可解释性,进而帮助研究人员更有效地理解人工智能做出的决策,及时对决策偏差进行纠正,提升算法的可信任性、安全性和性能。可解释的人工智能不仅有助于促进最终用户的信任、模型可审计性和人工智能的高效使用,还能降低使用人工智能所面临的合规性、法律、安全和声誉问题。

2.6.2 基于统计学提高结果的可解释性

研究表明,统计学习方法和人工智能学习方法相辅相成,与不确定性量化一样,可解释性提高也高度依赖统计技术。使用统计模型作为辅助可以使人工智能算法更具解释力,例如回归算法、决策树、分类规则等传统的相对简单的统计模型不仅能够更好地描述人工智能算法的重要模式,而且能帮助人工智能实现高度的可解释性;图像模型、支持向量机、随机森林等集成算法以及神经网络等不仅可以实现可解释性,还可以用于不确定性量化,从而提高算法的准确性。

影响人工智能算法可解释性的一个重要因素是模型的复杂度。一般来说,简单模型具有更高的透明度,复杂模型可以提高性能,复杂度越高,准确度越高,但同时也意味着模型越难以解释。因此,制定算法解释权和义务等规则、减少计算量、降低模型复杂度已成为人工智能不可或缺的一部分。模型复杂度和可解释性的权衡方法也是统计学面临的问题,规则化、降低模型复杂度使模型具有更好的可解释性是现代统计学的基本准则,有许多方法可供借鉴,例如经验贝叶斯、估计量收缩处理等方法。

3 统计学推动人工智能发展的展望

人工智能所要研究的是如何让计算机或机器去完成

过去只有人才能做的智能工作,分类是人工智能研究的一个重要分支,目前对于弱人工智能和强人工智能这一分类基本达成共识,并对应形成了弱人工智能观和强人工智能观。弱人工智能是指能够以人类方式思考并执行特定任务的机器、系统或算法等,由于这些机器、系统或算法只能完成程序所赋予的特定任务,无法超越其限制,因此即便是目前人类已经创造的最为复杂且具有超级自学习能力的人工智能系统,也只是某一特定领域的可追踪程序。强人工智能是指能够像人类一样独立思考并高效执行任何智力任务的机器、系统或算法等,这些机器、系统或算法经过适当编程后就是具有精神状态的人脑,强人工智能有可能制造出有知觉和自我意识的、真正能推理和解决问题的智能机器。

缺乏稳定性、透明度和可解释性极大地限制了人工智能模型和算法的落地。实践表明,人工智能已经在近十几年的快速发展中逐渐陷入僵局。机器学习只是数据到概率的一个工具,从概率到真正理解,还需要预测行动的效果、进行反事实想象^[9]。因此,因果推理取代关联推理,教会机器理解问题背后的根源,最终理解现实,成为人工智能进一步发展的关键,“可信任”与“可解释”已成为人工智能发展的主要方向。实现强人工智能的突破口在于因果推断的发展,强人工智能阶段的本质就是基于数据的学习和推理,将统计学应用到人工智能改进中的目标是促进或

支持对模型和数据的解释,并从中获得知识,提高结果的可解释性。

参考文献:

- [1]Weihs C, Ickstadt K. Data Science: The Impact of Statistics [J].International Journal of Data Science and Analytics,2018,6(3).
- [2]Warner B, Misra M. Understanding Neural Networks as Statistical Tools [J].American Statistics,1996,50(4).
- [3]Ntoutsis E, Fafalios P, Gadiraju U, et al. Bias in Data-driven Artificial Intelligence Systems. An Introductory Survey [J]. Data Mining and Knowledge Discovery,2020,10(3).
- [4]Meinshausen N, Bühlmann P. Stability Selection [J].Journal of the Royal Statistical Society: Statistical Methodology,2010,72(4).
- [5]游士兵,徐小婷.统计学方法的发展及其在大数据中的应用[J].统计与决策,2020,(4).
- [6]Pearl J. The Foundations of Causal Inference [J].Sociological Methodology,2010,40(1).
- [7]Rubin D B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies [J].Journal of Educational Psychology, 1974,66(5).
- [8]Dunson D B. Statistics in the Big Data Era: Failures of the Machine [J].Statistics & Probability Letters,2018,(136).
- [9]Pearl J. Causal Inference: History, Perspectives, Adventures, and Unification? [J].Observational Studies,2022,8(2).

(责任编辑/方 思)

The Role of Statistics in AI Development and Its Application Prospect

Li Zhiyong¹, Jiao Weiling²

(1.Social and Public Opinion Survey Center of Jiangsu Province, Nanjing 210000, China; 2.School of Economics and Management, Yancheng Institute of Technology, Yancheng Jiangsu 224002, China)

Abstract: Statistics run through the whole process of artificial intelligence (AI) development and application. In order to integrate statistical theories, models and methods to improve and innovate AI algorithms and their results, as well as to promote the new development in artificial intelligence, this paper develops a systematic study on the role and application of statistics in AI based on the theory and method of data science. The research results show that Statistics helps AI systems and algorithms determine research issues, optimize data collection schemes, evaluate and improve data quality, drive the transition from association to causation, quantify uncertainty, and improve the interpretability of results.

Key words: statistics; artificial intelligence; data science; statistical model and method; function and application