数据科学的科学性与科学问题的分析

朝乐门

引用本文

朝乐门. 数据科学的科学性与科学问题的分析[J]. 计算机科学, 2024, 51(1): 26-34.

CHAO Lemen. Exploring the Scientific Nature and Scientific Questions of Data Science[J]. Computer Science, 2024, 51(1): 26-34.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100121

数据科学的科学性与科学问题的分析

朝乐门
数据工程与知识工程教育部重点实验室(中国人民大学) 北京 100872
中国人民大学信息资源管理学院 北京 100872

摘 要 作为一门新兴的学科领域,数据科学的科学性受到了关注且其科学问题未明确提出。文中从科学研究范式及方法论、可证伪性和可再现性、科学精神及快速迭代以及科学研究纲领及理论体系4个方面探讨了数据科学的“科学性”,并解答了为什么数据科学是一门新兴科学的问题。在此基础上,结合DIKW模型(DIKW Pyramid or Hierarchy)、DMP(Data-Model-Problem)模型、数据科学的统计学和机器学习方法论以及数据科学的流程与活动,提出了数据科学的7个核心科学问题:解释在先还是在后或无、问题对齐数据还是数据对齐问题、更加相信数据还是模型、更加重视性能还是可解释性、如何划分数据、如何用已知数据解决未知数据的问题、人在环路还是人出环路。最后,提出了数据科学研究的4点建议:聚焦数据科学本身的理论研究,推动数据的科学、技术和工程需要进一步分离和专业化,加强人工智能赋能的数据科学的理论与实践以及数据科学学科(Data Science as A Discipline)与学科中的数据科学(Data Science Within A Discipline)的联动。

关键词:数据科学;科学属性;科学问题;DIKW模型
中图分类号 TP391

Exploring the Scientific Nature and Scientific Questions of Data Science

CHAO Lemen
Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing 100872, China
School of Information Resource Management, Renmin University of China, Beijing 100872, China

Abstract As an emerging academic field, the scientific nature of data science has garnered attention, and its scientific questions have not been clearly defined. This paper explores the scientific nature of data science from four aspects: scientific research paradigms and methodologies, falsifiability and reproducibility, scientific spirit and rapid iteration, and scientific research agenda and theoretical framework. It also answers the question of why data science is an emerging science. Building upon this foundation and incorporating concepts such as the DIKW model(data-information-knowledge-wisdom pyramid or hierarchy), the DMP model(data-model-problem model), the statistical and machine learning methodologies of data science, and the processes and activities in data science. This paper presents seven core scientific questions in data science: the precedence of explanation or data, problem alignment with data or data alignment with problems, prioritizing trust in data or models, emphasizing performance or interpretability, data partitioning strategies, solving unknown data problems with known data, and the role of humans within or outside the loop. Finally, four recommendations for data science research are proposed: a focus on theoretical research within data science itself, the further separation and specialization of data science in terms of science, technology, and engineering, strengthening the theory and practice of data science empowered by artificial intelligence, and fostering collaboration between the discipline of data science and data science within other disciplines.

Keywords Data science, Scientific nature, Scientific questions, DIKW model

1 引言

自Tukey于1962年发表论文《数据分析的未来》(“The Future of Data Analysis”),奠定了数据科学领域的发展基础以来,数据科学已历经近60年的发展[1]。随着大数据应用的普及,数据科学受到了前所未有的广泛关注,相关的学术研究、产业应用和人才培养都已取得较大进展。但是,仍有两类核心问题尚未得到解决:一类是数据科学的科学属性体现在哪些方面?数据科学是否是一门科学?另一类是数据科学的核心科学问题到底是什么?数据科学领域应围绕哪些核心问题开展研究?这两类问题的解决对于推动数据科学领域的学术研究具有重要意义。

为此,本文旨在抛砖引玉,探讨数据科学的这两类核心问题。本文第2章结合科学属性的代表性理论,分析数据科学

基金项目:国家自然科学基金(72074214)
This work was supported by the National Natural Science Foundation of China(72074214).
通信作者:朝乐门(chaolemen@ruc.edu.cn)

的4个核心科学属性。第3章从数据科学视角探讨DIKW模型的局限性及其改进建议,并提出数据科学的一个核心科学问题:解释在先还是在后或无?第4章描述了数据科学的三要素模型——DMP模型(数据-模型-问题模型),并提出了数据科学的另外3个核心科学问题:问题对齐数据还是数据对齐问题?更加相信数据还是模型?更加重视性能还是可解释性?在此基础上,第5章对比分析了数据科学的两种不同方法论——统计学与机器学习,并提出了数据科学的两个科学问题:如何划分数据?如何用已知数据解决未知数据的问题?第6章结合数据科学的流程与活动,提出数据科学的第七个科学问题:人在环路还是人出环路?最后,总结全文并展望未来。

2 数据科学的科学属性

通常,科学的主要特征主要体现在从事实推导理论(Deriving Theories from the Facts)、证伪主义(Falsification)、科学研究范式(Research Paradigm)、科学研究纲领(Research Programme)、科学研究方法论(Research Methodology)、可再现性(Reproducibility)以及科学精神(Scientific Spirit)[2-5]。因此,本文从以下4个方面探讨数据科学的科学属性。

2.1 科学研究范式及方法论

第四范式(The Fourth Paradigm),即数据密集型科学发现(Data-intensive Scientific Discovery),是数据科学的主要研究范式。Hey于2007年提出了第四范式的概念,认为第四范式是继经验科学、理论科学和计算科学之后的下一个科学革命[6]。从数据科学角度看,上述4个范式的主要区别在于是否依赖数据以及依赖什么样的数据(见表1)。

1)第一范式(经验科学,Empirical Science):主要依靠经验观察或实验方法收集与生成数据。该研究范式虽然依赖数据,但所依赖的数据为"经验观察或实验数据"——以人的观察获得的数据或实验环境下生成的数据。

2)第二范式(理论科学,Theoretical Science):主要侧重于理论构建、数学建模和逻辑推导的方式研究科学理论。这种研究范式强调的是抽象思维和理论验证,重点在于理论的构建与推导,其研究并不依赖于数据。

3)第三范式(计算科学,Computational Science):将计算机模拟与仿真作为主要科学研究方法,侧重用于模拟与仿真复杂系统的行为,其研究虽然依赖数据,但所依赖数据是由计算机程序生成的模拟和仿真数据,并非为人的经验观察或实验数据或真实数据。

4)第四范式(数据密集型科学发现,Data-intensive Scientific Discovery):强调在实际生产、制造、业务或生活中产生的真实数据的分析和洞见,其研究不仅需要依赖数据,而且所依赖的数据为大数据。相对于第一范式的实验数据和第三范式的模拟数据,真实数据往往更具量大、多样、快速、价值密度低等大数据的V's特征。

表1 科学研究的四范式理论的对比

Table 1 Comparison of four paradigm theories in scientific research

	第一范式	第二范式	第三范式	第四范式
名称	经验科学	理论科学	计算科学	数据密集型科学发现
是否依赖数据	是	否	是	是
用什么样的数据	经验观察或实验数据(人的观察获得的数据或实验环境下生成的数据)	不用数据(进行理论构建、数学建模和逻辑推导)	模拟与仿真数据(计算机模拟和仿真的数据)	真实数据(实际业务或生活中产生的大数据)

从方法论角度看,数据科学运用科学方法来探索和分析数据。数据科学中常用的方法论有3种:计算机科学、统计学和数据可视化。数据科学中的数据计算、数据存储、数据管理和数据分析等活动主要采用的是计算机科学领域的方法和技术。数据科学中的数据分析和数据探索通常采用的是统计学方法,包括描述性统计、推断统计和探索性数据分析。数据科学中的数据呈现主要采用的是数据可视化方法。

可见,数据密集型科学发现的研究范式以及计算机科学、统计学和机器学习方法论为主的科学研究方法论的应用,为数据科学提供了科学研究范式,使数据科学具备其他传统科学无法替代的学科定位,使数据科学成为科学体系中一个不可或缺的基础性学科。

2.2 可证伪性与可再现性

数据科学的理论与实践均具备可证伪性(Falsifiability)。数据科学的理论与实践中常用的证伪方法论有:假设检验、模型评价、交叉验证、异常检测、敏感性分析、偏差分析、模型泛化能力分析、模型的信效度检验、模型解释、A/B测试、因果推断等。这些方法论为证伪数据科学的理论和实践提供了基础,使得数据科学成为一个持续改进的科学领域。

数据科学的可再现性(Reproducibility)主要指其他研究者能够使用相同的数据集和分析方法得到"统计意义上(Statistically Significant)"的相似结果,而不一定是完全相同的输出。由于算法在数据划分、超参数调优、变量初始化等方面引入了随机数,因此即使是针对相同的问题、数据和模型,数据科学中的结果也可能因为这些随机因素而有所不同。在数据科学中,信度(Reliability)和效度(Validity)是衡量训练模型、测试结果或数据模型质量的两个关键指标。然而,如何权衡信度和效度(Bias-variance Tradeoff)是数据科学领域研究设计和数据分析中的一个核心命题。分析结果在一定阈值范围内的随机波动不仅不会影响数据科学的可再现性,而且能为数据科学提供灵活性和创造力。

2.3 科学精神及快速迭代

科学精神鼓励科学研究中的质疑、批判和创新。O'Neil等在其著作"Doing Data Science"中提到,她们在哥伦比亚大学开设的"数据科学导论"(Introductionto Data Science)课程的目的是"帮助学生成为批判性思考者、创造性问题解决者

(即使是那些尚未被识别的问题),以及好奇的提问者"[7]。在此基础上,朝乐门进一步提出了数据科学的3C精神——Creative Designing(创造性设计)、Critical Thinking(批判性思考)、CuriousAsking(好奇性提问)[8]。数据科学的3C精神是数据科学的科学精神的具体体现,是数据科学的科学属性的重要表现。

快速迭代是数据科学解决方案的一个典型特性,它并不追求一次性提供完美无缺的终极解决方案。相反,数据科学的方法论承认其解决方案可能存在的局限性和不足,并根据用户的实际使用情况以及相应产生的数据,持续优化所提出的解决方案,进而适应不断变化的需求和环境。在数据科学中,数据科学家可以运用3C精神质疑和批判数据(如数据的来源、质量及统计特征)、模型(如模型的假定、研究假设、泛化能力、鲁棒性、可靠性及可解释性)和问题(如问题的表示、规约、分解、对偶性、解空间、逼近方法及伦理道德风险等)。

2.4 科学研究纲领及理论体系

数据科学的研究纲领(Research Programme)[9]与理论体系已基本形成。朝乐门用"鹰图模型"[8]描述了数据科学的知识体系(见图1),并将数据科学的研究纲领与理论体系划分为4个关键部分:1)数据科学的方法论基础(对应于雄鹰的翅膀和脚),主要包括统计学、机器学习和数据可视化;2)数据科学的核心理论(对应于雄鹰的躯体),包括数据科学基础理论、数据加工、数据计算、数据管理、数据分析以及数据产品开发;3)数据科学的领域应用(对应于雄鹰的头部),聚焦于领域知识导向的问题理解、解决方案及解释方法,为数据科学应用提供了解释和决策依据;4)数据科学中的人文与管理理论(对应于雄鹰的尾部),包括数据伦理、道德、隐私、安全、法律和法规等人文社会科学相关理论,这些是数据科学理论与实践的重要平衡因素。

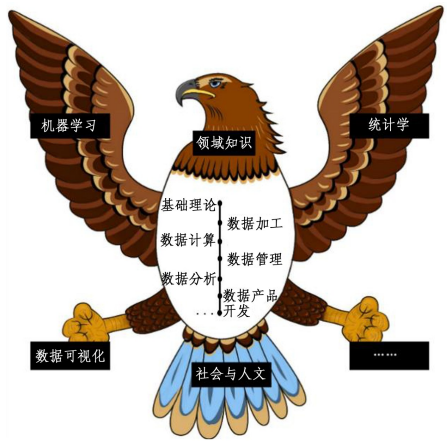

图1 数据科学的知识体系

Fig. 1 Knowledge system of data science

从以上分析可看出,数据科学具备科学研究范式及科学方法论、可证伪性和可再现性、科学精神及快速迭代、研究纲领及理论体系,符合科学的主要特征,是一门新兴科学。

3 DIKW模型的重新认识——解释在先还是在后或无

作为一种描述数据、信息、知识和智慧关系的经典模型,DIKW模型(DIKW pyramid or hierarchy)广泛出现在数据科学相关文献之中,多部数据科学的教材均将其作为数据科学的重要知识点。但是,传统的DIKW模型并不符合数据科学,尤其是第四范式的需求和特征。在数据科学领域,若不对DIKW模型的解释框架和思维方式进行批判与创新,新兴的数据科学理论与实践将难以在既有的框架中得到有效的诠释。此外,DIKW模型的传统认识将限制数据科学理论的创新,使数据科学理论难以突破包括数据挖掘、数据工程和数据管理在内的传统数据相关的理论研究。

3.1 DIKW模型的局限性

DIKW模型具有层次性和过程性的特点。威尔士大学的Rowley于2007年在其论文"The wisdom hierarchy: representations of the DIKW hierarchy"中提出,DKIW模型是层层递进和不断提炼的过程[10]。但是,DIKW模型的这种层次性和过程性更倾向于以人脑为主的传统数据分析,尤其是学术研究过程,而并不适用于以计算机为主,尤其是计算机自动化决策的应用场景。在数据科学的应用场景中,数据、信息、知识和智慧的关系不一定具备层次性和过程性。

1)数据科学中的数据利用并不一定具备DIKW模型的层次性。数据科学的研究并非一定从数据中提炼出信息、知识或智慧。其原因有两个:(1)并不是所有的数据都包含信息,所以无法提炼出有效信息,更不可能基于这些数据发现新知识或新智慧;(2)数据科学的研究是从大数据中发现有意的洞见或将数据转换为决策,不仅限于将数据转换为信息、知识或智慧后再利用。纽约大学的Provost教授等[11]于2013年明确提出,数据科学的最终目标是改进决策制定,他们认为数据科学不仅支持数据驱动型决策(Data-driven Decision Making),而且与之有交叉与重叠。

2)数据科学中的数据利用并不一定具备DIKW模型的过程性。数据科学并不完全等同于传统的学术研究,并不遵循从数据到信息,再到知识,最终到智慧的递进式认知过程。其主要原因有两个:(1)数据科学强调的是打通数据到决策的通道,即跨越信息、知识和智慧3个层次,直接将数据转换为决策;(2)数据科学强调的是端到端的实时决策[12]。DIKW模型中的知识到智慧需要经过从知识中提炼信息,并将信息转换为知识,再将知识进一步转换为智慧的过程,而这个过程需要较大的时间成本,并不符合数据科学中的实时分析、快速响应、自动决策和敏捷管理的要求。

3.2 DIKW模型的改进

DIKW模型在数据科学中的局限性源自其数据应用范式的单一性,即"解释在先的研究范式(Interpretation-First Paradigm)"。然而,数据科学中的数据应用范式有多种,除了解释在先的研究范式之外,更为广泛使用的是解释在后(Interpretation-Later Paradigm)或无解释的应用范式(Interpretation-Never Paradigm)。

可见,数据科学打破DIKW模型的关键在于改变数据应用研究范式的单一性,重新审视数据的应用与解释之间的内在联系,使数据应用更加符合数据科学理论研究范式和实践应用需求。表2列出了数据科学中的3种数据应用范式的对比。

表 2 数据科学的 3 种数据应用范式的对比

Table 2 Comparison of three data application paradigms in data science

范式名称	是否需要解释	解释与决策的先后顺序	是否与 DIKW 模型一致
解释在先	是	先解释后决策	是
解释在后	是	先决策后解释	否
无解释	否	只决策不解释	否

1)解释在先(Interpretation-First Paradigm):强调的是可解释性,要求基于数据的决策与解决方案必须建立在其可解释性的基础上,适用于"理论驱动型"数据应用。其优点是符合传统 DIKW 模型,易于解释和理解;缺点是实时性和敏捷性差,不适用于自动化应用场景。

2)解释在后或无(Interpretation-Lateror Never Paradigm):强调的是可用性,要求基于数据的决策与解决方案的高可用性,适用于"数据驱动型"数据应用。其优点是实时性和敏捷性高;缺点在于可解释性较差或解释不及时,与传统数据应用模式不同,需要打破 DIKW 模型的认识。

4 数据-模型-问题模型的提出——数据科学的三要素

数据(Data)、模型(Model)、问题(Problem)是数据科学的 3 个核心要素。其中,数据是研究对象,模型是从数据中识别、拟合和训练的模式、模型和规律的统称,而问题是需要解决的业务需求或研究问题。因此,我们可以用数据-模型-问题(DMP)模型来刻画数据科学的 3 个要素及其存在的矛盾,如图 2 所示。

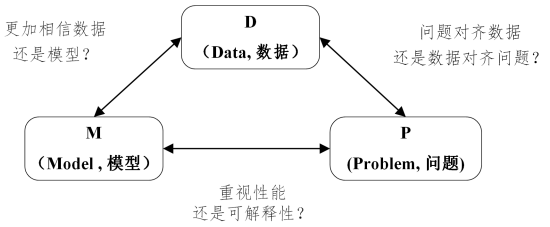

图 2 数据科学的 DPM 模型

Fig. 2 DPM model in data science

DPM 模型主要刻画了数据科学的 3 个要素及其内在联系。数据科学的理论与实践通常围绕以下 3 类工作进行:1)数据的获取、存储、计算、加工与管理;2)模型的训练、评估、部署及持续的迭代和优化;3)问题的理解、定义、描述与解决。然而,数据科学的理论与实践的难点在于如何权衡这 3 个要素之间的动态相互作用,以及它们对最终决策与数据产品的共同影响。

4.1 数据与问题:问题对齐数据还是数据对齐问题

在数据科学的理论研究和实践应用中,选择数据与问题之间正确的对齐方式尤为重要,其关键在于决定研究的出发点:是以已有数据为基础去探寻相关的问题(数据对齐问题),还是以确定的问题为导向去收集或生成所需数据(问题对齐数据)。不同的决策影响着研究的设计方式和成果的有效性。忽略问题与数据之间正确对齐的方向性,往往是数据科学项目失败的一个主要原因。

1)数据对齐问题(Data-to-Problem Alignment):以现有数据集为基础和出发点,通过探索数据来发现关键信息或潜在模式的过程。数据对齐问题类方法主要应用于数据洞见(Data Insights)类研究,主要采用统计学中的探索性数据分析方法和机器学习中的无监督学习方法。在数据对齐问题类研究中主要存在的风险是数据疏浚(Data Dredging)[13]或 P-hacking(P 值捕猎,亦称 P 值修剪)[14]的出现,即对数据进行反复探索、过渡解读或有意操纵数据、尝试多种数据分析方法并从中挑选等,直到得到具有统计显著性的结果为止。数据疏浚和 P-hacking 的存在会导致数据对齐问题中出现过拟合现象。

2)问题对齐数据(Problem-to-Data Alignment):以一个明确的问题或假设开始,通过获取并分析相关数据来解答问题或检验假设。问题对齐数据类方法主要用于数据试验(Design of Experiments)类研究,主要采用统计学中的假设检验和机器学习中的有监督学习方法。数据试验是数据科学的另一种主要方法论,根据研究假设和问题设计数据试验,并基于数据试验方法检验研究假设或问题是否成立。同时,通过控制实验条件,研究人员可以更好地了解变量之间的因果关系,并得出更加严谨的结论。以 A/B 测试为代表的随机对照试验(Randomized Controlled Trial,RCT)是数据科学应用领域中应用较为广泛的实验设计方法之一。问题对齐数据类方法中存在的主要风险在于数据偏见(Data Bias)的出现,如幸存者偏差(Survivorship Bias)[15]、辛普森悖论(Simpson's Paradox)[16]、伯克森偏差(Berkson's Paradox)[17]。这些偏差可能导致倾向于选取对特定问题有利的数据,扭曲数据理解与分析结果,进而造成错误的解释和决策。因此,数据对齐类研究方法,尤其是数据洞见类实践中需要避免数据偏见的出现,并对数据分析和研究工作进行必要的偏见评估、识别和优化工作。

同时,数据科学应重视"基于数据的证明"与"基于数据的解释"的两类任务的差异性。其中,"基于数据的证明"侧重于如何基于数据进行某一研究假设的信度与效度检验,其关键在于防止过拟合现象的出现。相反,"基于数据的解释"侧重于如何使用数据来解释一个特定研究假设的合理性和有效性,其关键在于提升解释结果的用户体验。可见,这两种方法各有自己的侧重点和局限性,数据科学应避免混淆或曲解。

4.2 数据与模型:更加相信数据还是模型

更加相信数据还是模型是数据科学的另一个核心议题,主要体现在:数据分析到底需要的是多数据还是更好模型的讨论[18-19]、模型的参数的概率与似然的讨论[20]、模型复杂度与过拟合的讨论[21]、模型正则化项及正则化系数的讨论[22],以及贝叶斯主义(Bayesianism)和频率主义(Frequentism)对模型参数的不同立场[23]。在数据科学中到底更加相信数据还是模型,主要取决于数据与模型之间的关系。

1)模型及其训练数据。在模型及其训练数据的关系上,数据科学通常采用的是对模型进行正则化的策略,即采用对模型的损失函数进行正则化的方式定义目标函数,并以目标函数值为依据评估一个模型,其表达式如下:

$$J(\theta)=L(y,f(x;\theta))+\lambda R(\theta)$$

其中，$J(\theta)$是目标函数，数据科学的目标是最小化这个函数；$L(y,f(x;\theta))$是损失函数，它测量了模型预测和观测标签之间的差异；λ是一个超参数，用于平衡损失函数和正则化项的相对贡献，代表的是更加相信数据还是相信模型的信息；$R(\theta)$是正则化项，可以采用 L1 正则化项和 L2 正则化项。

在数据科学中，我们通过优化目标函数 $J(\theta)$与调整模型参数 θ 来训练模型，旨在找到既能最小化预测损失又能避免模型过拟合的方法，进而权衡相信数据和相信模型的关系。总之，数据科学通过惩罚模型的方式权衡训练数据及其训练结果模型的信任关系。

2)模型及其输出数据。在模型及其输出结果的关系上，数据科学通常遵循大概率近似正确(Probably Approximately Correct，PAC)理论，即模型的预测并非绝对无误，而是在一定复杂度限制下，以较大的概率在可接受的误差范围内近似正确。按照 PAC 理论思想[24]，可以用以下公式表示数据科学中的模型(M)及其输出数据($M(x)$)的关系。

$$P(|M(x)-y|<\epsilon)>1-\delta$$

其中，$P()$代表概率；$M(x)$代表模型 M 对于输入 x 的输出结果；y 是对于输入 x 的观测标签或结果；ϵ是可接受的误差阈值；δ 表示概率不确定性的阈值。

PAC 理论对数据科学的主要意义在于提供了一种理解模型及其输出数据之间关系的方法，即模型 M 在一定的置信度 δ 下，能够以一定的精度ϵ逼近未见过的数据 y。这并不要求模型 M 对所有可能的输入 x 都给出完美的预测，只需对大多数情况给出足够好的预测。总之，数据科学遵循 PAC 理论，承认模型输出结果中存在一定的误差或不确定性，进而权衡模型及其输出数据的信任关系。

4.3 问题与模型：更加重视性能还是可解释性

在数据科学中，问题和模型之间的关系主要体现在重视解决问题的性能还是模型的可解释性。通常，算法的性能(Performance)和可解释性(Interpretability)之间存在矛盾，性能高的算法(如深度学习)难以解释，而容易解释的算法(如决策树)则性能低[25]。因此，数据科学中需要权衡两者之间的关系。

1)模型解释与信任。从数据科学的角度来看，模型解释包括模型本身的解释以及模型结果的解释。模型本身的解释方法分为模型相关方法(Model-Specific Methods)和模型无关方法(Model-Agnostic Methods)两种[26]。其中，常用的模型无关解释方法有特征重要性分析、可视化分析、LIME 算法、SHAP 方法以及反事实解释等。模型结果的解释方法属于模型的局部解释(Local Interpretation)，其解释对象为模型对特定样本给出结果。

2)问题表达与求解。问题表述指如何将现实世界的问题转化为算法和模型可以处理的问题，包括确定研究假设、选择适当的数据集、确定分析的方法和工具等。这一步骤对整个项目的成功至关重要，因为一个问题的表述决定了研究的方向和深度；求解问题则涉及到选取合适的模型、设计算法、数据分析、最优化技术等，是实际执行问题解决方案的过程。通常，如何形式化表示问题是问题解决的关键。在数据科学中，问题求解采用的方法包括但不限于：对偶性(Duality)、优化(Optimization)、逼近(Approximation)、启发(Heuristics)、概率推断(Probabilistic Inference)、统计建模(Statistical Modeling)以及计算智能(Computational Intelligence)等。

5 数据科学的方法论——统计学与机器学习两种方法论

在数据科学中，研究数据主要采用的是“数据划分研究方法”——将数据分为已知数据(Seen Data，简称 S-Data)和未知数据(Unseen Data，简称 U-Data)，并依靠 S-Data 解决 U-Data 的问题，如图 3 所示。数据科学的两种基本方法论——统计学和机器学习，分别采用了两种不同的数据划分方法：统计学将数据分为样本(S-Data)和总体(U-Data)，并从样本数据推断总体特性；而机器学习则通常将数据划分为训练集(用于模型训练)、测试集(用于评估模型性能)和验证集(用于模型选择和调参)，其中训练集相当于 S-Data，测试集和验证集可以被视为 U-Data。机器学习的目标是基于 S-data 训练出能够泛化到 U-data 的模型。

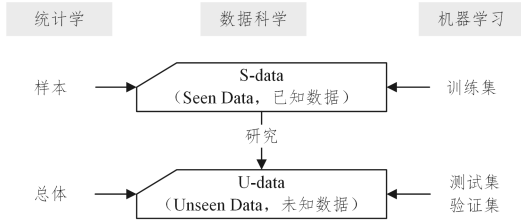

图 3 数据科学中的数据划分方法

Fig. 3 Data splitting methods in data science

5.1 相似性

从方法论角度来看，统计学和机器学习采用的方法均属于“数据划分研究方法”。表 3 列出了统计学、机器学习和数据科学中的术语对照表。当然，每个领域通常都会根据自己的传统和问题背景赋予这些术语特定的含义，表 3 中的术语并非是唯一的，也不要求严格一一对应。

表 3 数据科学、机器学习和统计学术语对照表

Table 3 Comparison of terms in data science, machine learning, and statistics

	数据科学	机器学习	统计学
1	建模(Modeling)	学习(Learn)	拟合(Fit)
2	分析方法(Analytical Method)	算法(Algorithm)	模型(Model)
3	探索性分析(Exploratory Analysis)	无监督学习(Unsupervised Learning)	聚类(Clustering)/密度估计(Density Estimation)
4	预测性分析(Predictive Analysis)	有监督学习(Supervised Learning)	分类(Classification)/回归(Regression)
5	图(Graph)	网络(Network)	模型(Model)
6	模型参数(Model Parameters)	权重(Weights)	参数(Parameters)
7	解释变量(Explanatory Variable)	特征(Feature)	自变量(Independent Variable)
8	响应变量(Response Variable)	目标(Target)	因变量(Dependent Variable)

5.2 差异性

统计学是根据样本推断总体，而机器学习则是寻找可泛

化的预测模式[27]。从数据划分研究方法的角度看,机器学习与统计学的主要区别在于以下 4 个方面。

1)方法论的前提假设。统计学方法通常基于一系列严格的假设,特别是关于数据的概率分布,而机器学习算法则可能不要求这些假设,或对这些假设的依赖程度较低。

2)方法论的实现策略。通常,统计学倾向于采用参数估计和假设检验的策略,旨在基于样本统计量推断总体参数。然而,机器学习更多地使用交叉验证的策略,在训练集上训练模型,在验证集上选择算法和调整参数,最终在测试集上评估模型的预测能力。

3)方法论的侧重点。统计学传统上侧重于模型的选择和假设的验证,机器学习则更多关注于算法的性能和在大数据集上的可扩展性。

4)方法论的理论基础。统计学拥有深厚的理论基础,包括概率论和数理统计,而机器学习则受益于计算机科学、优化理论和人工智能等领域的理论和技术。

当然,统计学和机器学习并不是完全对立的领域,它们之间的界限越来越模糊,越来越多地相互交叉并借鉴对方的技术。机器学习领域广泛采用的正则化技术是由统计学中的 Ridge 回归[28]和 Lasso 回归[29]等方法演化而来的,而统计学界也引入机器学习的算法(如随机森林和支持向量机)来分析数据。统计学中的概率建模为理解什么是学习提供了一个参考框架,已经成为设计能够从经验中获取数据并学习的机器的主要理论和实践方法之一[30]。Van De Schoot 等[31]于 2021 年分析了贝叶斯统计理论在机器学习中的应用,并进一步讨论了先验和后验预测检验的重要性,提出了变分推断以及变量选择方法。统计学和机器学习两个领域会继续相互借鉴和融合式发展,而数据科学将成为两者交叉与融合的主要学科领域。

6 数据科学的流程与活动——人在环路还是人出环路

在数据科学领域,通常采用生命期理论讨论数据科学的基本流程及其活动。Haertel 等[32]于 2022 年通过文献调研分析了 28 种数据科学流程模型,并提出了一个包含业务理解、数据收集、探索和准备、分析、评估、部署和利用 6 个阶段的数据科学生命周期理论。不同的数据科学流程的主要区别在于划分阶段的个数和命名方法不同,而其基本思路都是一致的——数据科学项目并非一步完成,而需要分步完成。其中,哥伦比亚大学 Schutt 教授等[33]于 2014 年提出的数据科学流程(The Data Science Process)不仅体现了数据科学流程与传统数据管理流程的差异性,而且打破了多数已有数据科学流程的线性特征。由于该模型于 2014 年提出,其中的活动及其顺序需要经过必要的调整。图 4 为调整后的数据科学流程图。通过数据化(Datafication)将现实世界中的信息(Information)映射到数据世界中的数据(Data),并采用数据加工(Data Wrangling)将数据转换为算法和模型可以处理的数据模态——规整数据(Tidy Data)。对规整数据进行分析和洞察,对于分析洞察结果(数据模型或预测结果),不仅可以进行数据可视化/故事化后呈现给决策者,也可以作为数据产品进行部署和运维。数据化、数据加工和数据分析与洞察等活动均需要数据存储/管理/计算等底层技术的支持。

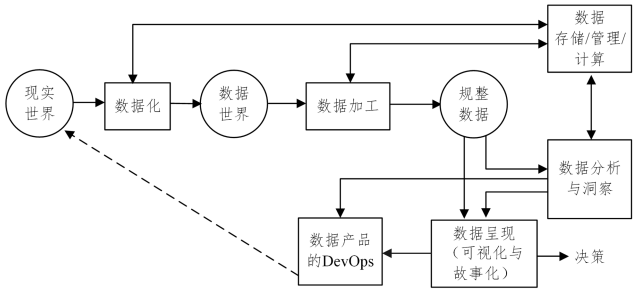

图 4 数据科学的基本流程

Fig. 4 Basic workflow of data science

6.1 数据科学流程的主要活动

1)数据化(Datafication)——样本接近总体

数据化是数据科学中广泛讨论的新术语之一。数据化是对现实世界中的事物,包括人类活动进行数字化和量化记录的过程,其结果是将现实世界映射到数据世界,使数据科学流程的后续活动为通过研究数据世界来解决现实世界的问题。与传统数据采集不同的是,数据化具备采集范围广、实时性强、自动化程度高的特点,且数据对象存在多样性、数据流存在主动性。数据化的实现技术和工具很多,如传感器、物联网、可穿戴设备以及业务系统。

数据化所带来的从现实世界到数据世界的映射,为数据科学通过数据世界的研究以及虚实结合的方法来解决现实世界问题提供了可能。通常,数据化是自动化过程。但是,数据化涉及企业数据安全和个人隐私保护[34]问题时,需要人的参与。数据化的范围、质量、速度与可交互性是数据科学项目成败的主要瓶颈之一。

2)数据计算、存储与管理——数据的一致性与可用性的平衡

在数据科学领域,一致性与可用性的平衡是数据计算、存储和管理的难题之一。大数据的计算与存储领域常用的基础理论包括 CAP 理论、BASE 原则、PACELC 理论(Possibility, Availability, Consistency, Latencyand Cost)、Lambda 架构等,均体现了权衡数据一致性和可用性的必要性,并提供了各自的解决方案。这些理论和技术对于数据科学中设计和管理大规模数据系统,尤其是平衡数据的一致性和可用性具有理论指导作用。

分析型数据(Transactional Data)和事务型数据(Analytical Data)的分离是确保数据一致性和可用性的一种传统做法。在数据管理的早期理论,如数据库和数据仓库中,强调分析型数据和事务型数据,进而得到较高数据一致性和可用性。但是,随着新技术的发展,尤其是数据湖、数据经纬(Data Fabric)[35]的提出,这种传统的分离结构也正在改变,开始尝试用同一平台处理不同数据,以提高组织层次数据的整体一致性和可用性。

3)数据加工(Data Wrangling)——规整数据(Tidy Data)与乱数据(Messy Data)

传统数据分析中强调数据的质量,将数据分为脏数据和干净数据,通过数据清洗将脏数据转换为干净数据。然而,

随着算法的鲁棒性和分析能力的提升，算法可以自动识别和处理数据中的缺失值、异常值、虚假数据等常见的数据质量问题。因此，数据科学所面临的主要矛盾从数据质量转向数据模态。在数据科学中更加强调的是数据模态，将数据分为乱数据和规整数据，通过数据加工(Data Wrangling)将乱数据转换为规整数据，其目的在于将数据转换为算法和模型所需的数据模态要求。

Wickham 等[36]于 2014 年提出了规整数据的 3 个基本原则：每个变量一列、每个观测一行以及每个值一个单元格。数据规整化处理(Data Tidying)的目的是保持数据结构的一致性，降低数据分析的复习度并支持向量化计算[37]。此外，基于人工处理的特征工程(Feature Engineering)[38]、基于自动化处理的表示学习(Representation Learning)[39]以及两者的融合也是数据加工的重要研究课题。

4)数据分析与洞察——相关分析与因果分析

对于相关关系和因果关系，数据科学领域的观点曾发生重要变化。在数据科学研究的早期，更加重视相关关系。例如，Mayer-Schönberger 等[40]于 2013 年在他们的著作“Big data: A revolution that will transform how we live, work, and think”中提到，在大数据时代，相关关系比因果关系更为重要。但是，随着数据科学实践的深入，人们发现基于相关关系的数据科学方法具有显著的局限性——缺乏可解释性和可信任以及难以对其进行干预、控制和优化。因此，学术界开始认识到相关关系并不能满足数据科学的需要，应在相关关系的基础上进一步深入研究因果关系。例如，Pearl 等[41]于 2018 年在他们的著作“The Book of Why: the new science of cause and effect”中提出了因果科学。相关关系能够逼近因果关系的程度、相关关系和因果关系的边界、是否可以利用反事实推断从相关关系中推断出因果关系，以及如何保证大数据分析的结论可信等问题是未来的重点研究方向之一[42]。

从数据分析的角度看，重视因果关系的主要目的是进行处方性数据分析(Prescriptive Analysis)。根据 Gartner 的分析价值扶梯(Gartner's Analytic Value Escalator)模型[43]，数据分析分为描述性分析、诊断性分析、预测性分析和处方性分析。其中，处方性分析位于价值链的顶端，其复杂度和潜在价值均最高。处方性分析不仅预测未来会发生什么，还给出如何通过这些预测来优化未来的建议，主要用于提供问题的最佳解决方案或行动方案。

5)数据呈现——数据可视化与数据故事化

数据呈现是将数据分析与洞见的结果以及数据加工后的规整数据以目标用户容易接受的形式呈现，进而实现决策支持或数据产品提供的目的。数据可视化和数据故事化是数据科学常用的两种数据呈现方法。两者的主要区别在于：数据可视化具有更易于理解、感知和洞见的特征，而数据故事化具备已有记忆、认知和体验的特征[44]。

数据故事化的主要功能有 3 个：数据认知、算法解释和虚实结合。数据故事化的过程通常分为故事建模和故事叙述两个阶段。其中，故事建模阶段的主要方法为 SHAP, LIME, Anchors，以及反事实解释在内的可解释性机器学习技术及分组对照试验；故事叙述阶段的主要方法为数据可视化、自然语言生成、多模态数据生成与合成方法。为了驱动目标用户的行动，数据叙述必须具备清晰、直接、可信、令人记忆犹新和可执行等特点[45]。数据可视化是数据故事化中常用的叙述手段之一。

6)数据产品的 DevOps-开发与运维一体化

数据科学需将分析模型与数据可视化进一步研发成数据产品，并通过 DevOps 提高其开发、测试、交付和运维的质量与效率，以支持数据产品在现实世界中的部署与应用。数据产品的 DevOps 实践包括跨团队协作、自动化流程、持续集成/持续部署、监控与测试、模型服务以及快速迭代。

DevOps 主要强调开发(Development)与运维(Operations)团队之间的沟通、协作、集成、自动化和快速反馈。数据产品的 DevOps 采用 DataOps 和 ModelOps 的理念和技术处理数据科学产品中流程自动化、数据一致性、模型监控以及快速迭代等问题。

6.2 数据科学流程中人的作用

数据科学流程应遵循“至少一次人在环路原则”(At Least Once Human-in-the-Loop Principle, ALO-HITL 原则)——数据科学流程应设计成至少包括一个环节，其中必须有人类的直接参与。这一原则旨在权衡机器和人在数据科学流程中的不同作用，保障决策过程的问责和智能分析的合理性。

数据科学流程的 ALO-HITL 原则的提出依据和原因如下：

1)自动化流程虽然可以提高数据处理的效率和精度，但机器仍缺乏完全理解复杂问题和做出价值、伦理和道德层次的判断的能力[46-47]。

2)数据科学不仅仅是数据和算法的集合，更是一个融合了业务理解、创新思维和策略规划的综合性学科。AI 尚未准备好做出无人监督的决策[48]。

3)人的直觉和经验判断在数据科学的诸多环节(如问题定义、数据理解和模型评估)发挥着不可替代的作用[49-50]。

数据科学流程的 ALO-HITL 原则的实施并不意味着抵制自动化，而是强调人在自动化流程中的主导地位，强调一个完整的数据科学闭环流程不能完全脱离于人的参与。ALO-HITL 原则可增强数据科学流程的鲁棒性，提升解决复杂问题的能力，并确保最终输出的可靠、可解释和可问责。

数据科学流程中多数活动可采取“人出环路(Human-out-of-the-loop, HOOTL)”[51]的方式，实现自动化处理。但是，数据科学流程的关键活动，尤其是需要主观判断和审批类活动需要具备“人在环路中(Human-in-the-loop, HITL)”的特征。

结束语 数据科学不仅是一门新兴科学，而且是一门独立学科。作为新兴科学，它具备科学研究范式及方法论、可证伪性和可再现性、科学精神及快速迭代、研究纲领及理论体系。作为一门独立学科，它具备独立于其基础学科——统计学和计算机科学的研究问题和研究方法，并具备自己特有的新研究视角、研究目的、研究对象和理论体系。但是，数据科学的理论与实践仍处于起步阶段，后续研究需要在以下几个方面进行进一步的深入研究：1)聚焦数据科学本身的理论

研究。目前,关于数据科学的讨论很多,但针对数据科学本身的科学性和科学问题的系统研究较少。本文提出了数据科学的科学性和科学问题,但有待进一步深入研究。2)数据的科学、技术和工程需要进一步分离和专业化。目前,数据科学相关讨论中的科学、技术和工程问题是混合的,不利于数据科学学术理论的快速发展。本文仅讨论了数据科学中的科学问题,但数据科学中的技术和工程问题以及三者之间的联系有待进一步探讨。3)人工智能赋能的数据科学的理论与实践。数据的直接处理者是智能体而不是人类,因此,除了统计学、机器学习和数据可视化,人工智能将成为数据科学的新研究方法。本文虽提及了人工智能在数据科学中的重要性,但并未深入研究两者的联系,相关问题是数据科学未来研究的重点之一。4)数据科学学科(Data Science as a Discipline)和学科中的数据科学(Data Science within a Discipline)的联动发展。目前,数据科学的相关理论和实践分散在不同的学科领域,导致学科中的数据科学发展非常迅速,而数据科学学科的直接研究相对落后。因此,数据科学未来研究需要借鉴学科中的数据科学理论来推动数据科学学科的发展。

参 考 文 献

[1] DONOHO D. 50 years of data science[J]. Journal of Computational and Graphical Statistics,2017,26(4):745-766.

[2] CHALMERS A F. What is this thing called science [M]. Hackett Publishing,2013:1-304.

[3] BAKER M. 1500 scientists lift the lid on reproducibility[J]. Nature,2016,533:452-454.

[4] EDDINGTON A S. Science and the unseen world[M]. Quaker Press,2007:1-56.

[5] FORTUNATO S,BERGSTROM C T,BÖRNER K,et al. Science of science[J]. Science,2018,359(6379):1-7.

[6] HEY T. The fourth paradigm[M]. Washington:Microsoft Research,2009:1-4.

[7] O'NEIL C,SCHUTT R. Doing data science:Straight talk from the frontline[M]. O'Reilly Media,Inc.,2013.

[8] 朝乐门. 数据科学理论与实践(第三版)[M]. 北京:清华大学出版社,2022:20-24.

[9] HARDING S. Can theories be refuted:Essays on the Duhem-Quine thesis[M]. Dordrecht:Reidel Publishing Company,1975:205-259.

[10] ROWLEY J. The wisdom hierarchy:representations of the DIKW hierarchy [J]. Journal of information science, 2007, 33(2):163-180.

[11] PROVOST F,FAWCETT T. Data science and its relationship to big data and data-driven decision making[J]. Big Data,2013,1(1):51-59.

[12] LAKSHMANAN V. Data Science on the Google Cloud Platform (2nd Edition)[M]. O'Reilly Media,Inc.,2022.

[13] LAZER D,KENNEDY R,KING G,et al. The parable of Google Flu:traps in big data analysis[J]. Science,2014,343(6176):1203-1205.

[14] MUNAFÒ M R,NOSEK B A,BISHOP D V M,et al. A manifesto for reproducible science[J]. Nature Human Behaviour,2017,1(1):1-9.

[15] GUPTA P,MACAVANEY S. On survivorship bias in MS MARCO[C]// Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022:2214-2219.

[16] SHARMA R,GARAYEV H,KAUSHIK M,et al. Detecting Simpson's Paradox:A Machine Learning Perspective[C]// International Conference on Database and Expert Systems Applications. Cham: Springer International Publishing, 2022: 323-335.

[17] SHARMA R,KAUSHIK M,PEIOUS S A,et al. Why Not to Trust Big Data:Discussing Statistical Paradoxes[C]// International Conference on Database Systems for Advanced Applications. Cham:Springer International Publishing,2022:50-63.

[18] ZHU X,VONDRICK C,RAMANAN D,et al. Do We Need More Training Data or Better Models for Object Detection [C]// BMVC. 2012:1-11.

[19] JUNQUÉ DE FORTUNY E,MARTENS D,PROVOST F. Predictive modeling with big data:is bigger really better[J]. Big Data,2013,1(4):215-226.

[20] SORENSEN D. Statistical Learning in Genetics:An Introduction Using R[M]. Cham: Springer International Publishing, 2023: 51-75.

[21] ROELOFS R,SHANKAR V,RECHT B,et al. A meta-analysis of overfitting in machine learning[J]. Advances in Neural Information Processing Systems,2019,32:9179-9189.

[22] SMITH S L,DHERIN B,BARRETT D G T,et al. On the origin of implicit regularization in stochastic gradient descent[J]. arXiv:2101.12176,2021.

[23] INCHAUSTI P. Statistical Modeling With R:a dual frequentist and Bayesian approach for life scientists[M]. Oxford University Press,2023.

[24] VALIANT L. A theory of the learnable[J]. Communications of the ACM,1984,27(11):1134-1142.

[25] GUNNING D,STEFIK M,CHOI J,et al. XAI-Explainable artificial intelligence[J]. Science Robotics,2019,4(37):eaay7120.

[26] MOLNAR C. Interpretable Machine Learning:A Guide for Making Black Box Models Explainable(2nd Edtion)[M]:Munich:Creative Commons,2022.

[27] DANILO B,NAOMI A,MARTIN K. Statistics versus machine learning[J]. Nature Methods,2018,15(4):233-234.

[28] HOERL A E,KENNARD R W. Ridge regression:Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1):55-67.

[29] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society Series B:Statistical Methodology,1996,58(1):267-288.

[30] GHAHRAMANI Z. Probabilistic machine learning and artificial intelligence[J]. Nature,2015,521(7553):452-459.

[31] VAN DE SCHOOT R,DEPAOLI S,KING R,et al. Bayesian statistics and modelling[J]. Nature Reviews Methods Primers,2021,1(1):1.

[32] HAERTEL C, POHL M, NAHHAS A, et al. Toward A Lifecycle for Data Science: A Literature Review of Data Science Process Models[C]// PACIS 2022 Proceedings. 2022.
[33] O'NEIL C, SCHUTT R. Doing data science: Straight talk from the frontline[M]. O'Reilly Media Inc., 2013.
[34] MAI J E. Big data privacy: The datafication of personal information[J]. The Information Society, 2016, 32(3): 192-199.
[35] SHARMAV, BALUSAMYB, THOMASJ J, et al. Data Fabric Architectures: Web-Driven Applications[M]. Berlin: Walter de Gruyter GmbH & Co KG, 2023.
[36] WICKHAM H. Tidy Data[J]. Journal of Statistical Software, 2014, 59(10): 1-23.
[37] WICKHAM H, ÇETINKAYA-RUNDEL M, GROLEMUND G. R for Data Science(2nd Edition)[M]. Sebastopol: O'Reilly Media Inc., 2023.
[38] ZHENG A, CASARI A. Feature engineering for machine learning: principles and techniques for data scientists[M]. O'Reilly Media Inc., 2018.
[39] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.
[40] MAYER-SCHÖNBERGER V, CUKIER K. Big data: A revolution that will transform how we live, work, and think[M]. Boston: Houghton Mifflin Harcourt, 2013: 50-61.
[41] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect[M]. New York: Basic Books, 2018.
[42] 程学旗, 梅宏, 赵伟, 等. 数据科学与计算智能: 内涵, 范式与机遇[J]. 中国科学院院刊, 2020, 35(12): 1470-1481.
[43] Gartner, Inc. Gartner's analytic value escalator[OL]. (2012-12-12). https://www.flickr.com/photos/27772229@N07/8267855748/.
[44] 朝乐门. 数据故事化[M]. 北京: 电子工业出版社, 2022: 96-97.
[45] VAUGHAN D. Data Science: The Hard Parts [M]: Boston: O'Reilly Media, Inc., 2024.
[46] STAHL B C. Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies[M]. Springer Nature, 2021.
[47] DE CREMER D, KASPAROV G. AI should augment human intelligence, not replace it[J]. Harvard Business Review, 2021, 18: 1.
[48] MCKENDRICK J, THURAI A. AI Isn't Ready to Make Unsupervised Decisions[OL]. (2022-09-15). https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions.
[49] WU X, XIAO L, SUN Y, et al. A survey of human-in-the-loop for machine learning[J]. Future Generation Computer Systems, 2022, 135: 364-381.
[50] CHEN V, LIAO Q V, WORTMAN VAUGHAN J, et al. Understanding the role of human intuition on reliance in human-AI decision-making with explanations[J]. Proceedings of the ACM on Human-Computer Interaction, 2023, 7(CSCW2): 1-32.
[51] SHAHRIARI B, SWERSKY K, WANG Z, et al. Taking the human out of the loop: A review of Bayesian optimization[J]. Proceedings of the IEEE, 2015, 104(1): 148-175.

CHAO Lemen, born in 1979, Ph.D, professor, is a senior member of CCF(No. 50431S). His main research interests include data science and big data analysis.

(责任编辑: 喻藜)