DOI:10.13878/j.cnki.jnuist.2019.03.002

顾峥[1]　高阳[1]

第四范式视角下的大数据科学

摘要

随着物联网、云计算技术的飞速发展,大数据及其相关科学成为学术界和工业界关注的热点.本文从范式理论的角度对大数据科学进行剖析,阐述大数据研究与传统研究的联系与区别;从机器学习的视角出发,分析了大数据带来的三大主要挑战及其背后的科学问题.此外,本文介绍了第四范式视角下进行大数据科学研究的洞察力,以及其所具有的积极意义;最后总结并展望未来大数据科学面临的挑战.

关键词

范式理论;大数据;机器学习

中图分类号 TP399

文献标志码 A

收稿日期 2019-05-01

资助项目 国家自然科学基金(61432008);国家重点研发计划重点专项课题(2017YFB0702601)

作者简介

顾峥,男,博士生,主要研究领域为大数据分析、计算机视觉.guzheng@ smail.nju.edu.cn

高阳(通信作者),男,博士,教授,博士生导师,主要研究领域为大数据分析、人工智能.gaoy@ nju.edu.cn

1 计算机软件新技术国家重点实验室(南京大学),南京,210023

0 引言

大数据及其相关概念自提出以来始终是各界关注的焦点,与大数据相关的科学研究蓬勃发展.第四范式是一种基于数据驱动的科学研究范式,被认为是以大数据科学为代表的新型科学研究的准则.大数据科学与第四范式研究之间的关系到底是怎样的?范式理论如何影响大数据科学的发展?研究者又应如何从第四范式的角度重新理解大数据?本文将针对以上几个问题进行初步探讨.

本文第1节介绍大数据的概念及其与第四范式研究的关系,第2节从第四范式的视角对大数据性质研究中关键技术进行介绍,第3节从机器学习的角度介绍大数据研究中的洞察力研究,第4节总结全文,对未来值得关注的研究方向进行探讨.

1 从范式理论到大数据科学

1.1 范式理论

范式(Paradigm)一词最早由美国科学家托马斯·库恩提出.在其代表作《科学革命的结构》一书中,库恩认为科学的发展不是单纯的线性累积,而是存在一种革命性的突变[1].库恩在书中指出,对于在某个常规科学时期的科学共同体,存在一套公认的科学研究模式,包括科学假说、理论、准则和研究方法,作为常规科学赖以运作的理论基础和实践规范,亦称之为范式.

然而,常规科学往往会遇到颠覆科学传统的异常现象,此类异常通常无法与现有研究范式预期保持一致,这促使科学共同体进入非常规的科学研究阶段,通过反思与总结,最终抛弃现有的科学理论,支持另一种与之不相容的理论,完成从一种研究范式到另一种研究范式的转变,从而完成科学的革命,库恩将这个过程称为范式转移(Paradigm Shift).

1.2 第四范式

纵观科学发展史,众多著名的科学转折都是范式转移引发的科学革命.在2007年召开的NRC-CSTB大会上,图灵奖得主、关系型数据库先驱Jim Gray发表了著名的演讲“eScience-A Transformed Scientific Method”,总结人类科学研究经历的4种范式:1)千年前,哥白尼、伽利略、开普勒等人开创了以观测实验为核心的经验主义科学范式;2)几百年前,以牛顿经典力学、麦克斯韦电磁学为代表的理论

主义科学范式,通过理论总结和理性概括的方式进行科学研究;3)几十年前,计算机的发明大大降低了计算的成本,通过模拟复杂现象,仿真实验逐步取代实验,计算主义科学范式成为主流;4)近 10 年来,随着物联网、云计算技术的发展,各类数据呈现爆炸性增长,人们开始关注数据本身蕴含的规律和背后的价值,进而思考:过去人类科学家基于实验、理论和计算进行的科学研究中,数据是作为佐证理论与实验工具.那么,能否以数据为出发点,直接从大量数据中计算得出未知的理论? 这种数据密集型的研究范式,被称为科学研究的第四范式(表 1).

表 1　科学研究的 4 种范式

Table 1　Four paradigms for scientific research

研究范式	主要时间	指导思想	典型代表
第一范式:实验科学	16 世纪以前	实验观察、总结规律	哥白尼地心说
第二范式:理论科学	17—19 世纪	简化实验、模型推理	经典力学、电磁学
第三范式:计算科学	20 世纪	模拟实验、仿真计算	量子力学、混沌理论
第四范式:数据科学	21 世纪	数据驱动、计算为辅	大数据科学

从第四范式的角度,任何学科都存在两个进化分支[1]:计算学分支和信息学分支.计算学分支基于现有理论,进行理论演绎,并采用信息技术对假说进行检验,从而发展新的学科理论;而信息学分支则先对实验、设备、档案、文献等各方面的数据进行采集,通过编码的方式存储在信息空间中,通过信息系统进行分析,研究者通过计算机向信息空间提出问题,并由系统给出答案.从这里可以看出计算主义和数据主义的本质区别:计算主义从计算的角度出发,将某一具体学科作为数据的集合,将数据集合作用于计算模型中进行验证;而数据主义从数据的角度出发,不依赖模型和具体假设,甚至不依赖于具体学科,是将计算作用于数据,从而更好地理解数据.

1.3　大数据科学

大数据是一个抽象的科学概念,其提出最早可以追溯到 2001 年,META 集团(现为高德纳)分析师 Doug Laney 在一项报告中指出数据持续增长带来的三大挑战[2]:海量(Volume)、多变(Velocity)、多样(Variety).有研究者在 Doug Laney 对大数据的 3V 定义上进行扩展,提出了大数据的 4V 定义[3],认为大数据除具备以上三种特性之外,还具有不确定性(Veracity).2010 年,Apache 公司将大数据定义为“无法被一般计算机在可接受的时间范围内获取、管理和处理的数据集”.

大数据的出现使之成为与自然资源、人力资源一样重要的战略资源[4].2012 年 3 月 29 日,美国总统科技政策办公室 OSTP(Office of Science and Technology Policy)公布了每年投资两亿美元的“大数据研究计划”;同一天,我国科技部发布的《“十二五”国家科技计划信息技术领域 2013 年度备选项目征集指南》中,把大数据研究列在首位;2014 年,国家自然科学基金委员会公布了有关大数据的重点项目群.据统计,自 2005 年至今,IBM 已投入超过 160 亿美元用于大数据相关的收购[5],此外,包括微软、谷歌、亚马逊等在内的各大公司都启动了自己的大数据项目,这些公司现在已经成为推动大数据产生和发展的最大动力,创造了巨大的社会经济价值[6].

学术界对大数据的关注也在不断持续.2008 年,*Nature* 发表“Big Data”专刊[7],同年发布一系列相关论文[8-10],介绍大数据相关概念和技术.2010 年,*The Economist* 发表专刊“数据,无处不在的数据”[11],从社会与经济学角度介绍数据为社会发展带来的巨大变革.2011 年,*Science* 发表专刊“Dealing with Data”[12],介绍大数据处理中的关键技术.大数据的产生给传统科学研究带了新的机遇和挑战,促使研究者们开始考虑数据科学的问题,进而产生了以大数据为核心的大数据科学.

2　第四范式视角下大数据科学带来的挑战

2.1　大数据的复杂性导致知识表示的困难

大数据在类型、结构、语义、组织和粒度上都具有一定程度的异构性.以医疗领域为例,医院在采集病人的医疗影像时,常采用的采集设备包括 MR、CT、超声等多种仪器,根据采集介质、衡量指标的不同,即使同一个病人的医疗数据也存在不同的数据结构.传统的数据管理和分析系统大都基于关系数据库,其只适用于结构化数据,无法处理半结构化或非结构化的数据.因此,多源异构的数据无法用传统的关系数据库表示.

针对这一问题,需要寻求面向大数据复杂结构的高效知识表示技术.大数据的复杂性本质来自于其4V特征,具体体现在两个方面:一是大数据内部结构的复杂性,在分布、异构、高维的大数据中,数据的特征或属性之间存在复杂关系;二是大数据外部关系的复杂性,不同数据之间存在相互关系,且这种关系随着时间、空间动态变化,更加剧了大数据的复杂性.因此,充分表示和学习大数据中复杂、动态的关系,有助于挖掘出有用的模式和规则,从而帮助计算机分析数据的复杂结构,使用户对数据的理解更具意义.

2.2 流数据的在线分布特性导致学习方法的改变

流数据是指连续、高速、无限的连续数据,其具有以下特点:

1)无限性:数据从数据源不断产生,总量没有上限;

2)动态性:数据分布随时间变化,存在概念漂移;

3)实时性:数据处理需要在一个既定时间内完成.

在传统的统计机器学习模型中,数据是定量的、全部可见的,研究者对数据进行多遍扫描,然后建模和计算、部署上线.然而在很多真实场景下,数据批量获取、批量计算的假设是不成立的,同时针对大量的流数据,数据产生是增量式的,如果每次新数据到达都需要重新扫描所有数据,将大大降低处理效率,造成计算资源的严重浪费.进一步来说,由于数据分布随时间不断变化,存生概念漂移的现象,使得机器学习中基本的数据独立同分布假设不再成立.此外,对于PB级别以上的大数据,传统的针对小规模数据的$O(N\log N)$级学习算法在时间效率上将不可接受,传统的易解决问题也有可能变成"难解"问题.

因此,针对以上问题,首先需要从理论层面上回答:在何种情况下,传统的易解决问题会变成大数据难解问题;其次,需要针对大数据的在线、分布特性,估计大数据的计算边界,建立近似非精确、增量式的在线学习理论和方法.

2.3 多实体交互的复杂决策导致推理的低效

在大数据的复杂决策中,数据的产生过程与分析过程不是相互独立的,往往存在决策系统和数据干预者之间的相互博弈,干预者会对数据的产生过程进行干预,从而增加数据分析的困难,使得参与博弈的多个实体之间的相互关系极其复杂、难以刻画,对实体行为的归纳推理也异常困难,博弈结构难以高效学习.

本质上,这种大数据应用是一类存在对抗性对手的复杂博弈问题,需要构造博弈模型,并通过推理算法进行均衡策略的求解.传统的推理技术包括演绎推理、类比推理、归纳推理等.然而,由于大数据本身具有的复杂特性,导致传统的推理方法不再适用,同时,大数据导致博弈具有巨大的策略空间,从而对均衡求解的过程带来困难.

因此,针对复杂决策的归纳博弈推理也是亟待解决的关键问题之一.从学习的角度看,需要对博弈结构进行归纳推理,学习潜在的博弈模型,同时,对博弈行为进行推理,学习数据干预者的行为模型.

2.4 大数据科学的关键技术:从4V到4I

针对大数据海量、多变、多样、不精确的4V特征,大数据研究需要寻求的是适应大数据特征的数据科学基本理论与方法,它需要满足整合性(Integrated)、近似性(Inexact)、增量性(Incremental)、归纳性(Inductive)的4I性质(图1).

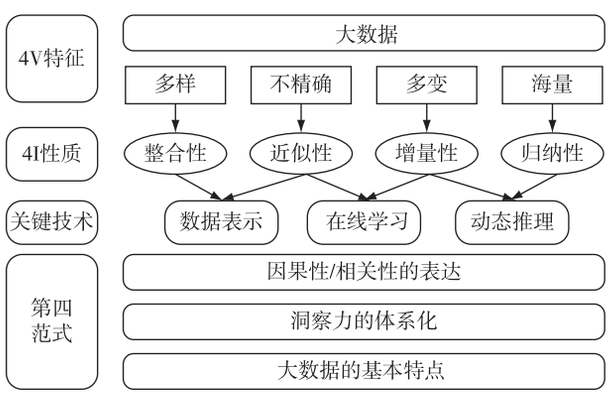

图1 大数据的4V表象到4I本质

Fig. 1 Big data technology: from 4V to 4I

1)近似性是指要将原来的追求绝对精确解转变为追求高效近似解,以应对数据的不精确性特征;

2)增量性是指要将原来的批量式计算方法转变为增量式计算方法,以应对数据快速多变的特性;

3)整合性是指要将原来只能处理单一数据的方法转变为能够处理多源异构数据,从而抓住数据的总体特征,以应对数据的多样性特征;

4)归纳性是指要从观测现象出发,归纳数据之间的相关性,从而抓住数据的本质特征,以应对数据的海量性特征.

3 第四范式视角下的大数据洞察力研究

人对自然事物的认知可以分为三个阶段:观察现象、相关分析、因果分析.人们通过观察发现不同距离的星系光谱波长不同,发现较远星系发出的可见光波长更长,推测出宇宙在不断膨胀,进而推测出宇宙大爆炸的理论.其中,星系距离与波长之间是相关的,而宇宙爆炸则是导致这一系列现象的原因.第四范式理论和大数据的出现,使得人们看待问题、分析问题的方法发生了根本性的变化,对数据的洞察力得到了明显的提升.下面以决策交互数据中的合作与对抗为例,解释第四范式视角下对洞察力的研究与体现.

3.1 从因果关系到相关关系

与传统科学研究不同,大数据科学的核心问题不再是对数据因果关系的追求,而是对相关关系的追求[13].

相关性的一个典型例子是商品推荐.电商网站通过收集用户浏览、点击的商品,由系统生成个性化推荐,从而实现对不同用户群体的精准投放.沃尔玛公司通过数据分析发现,每当季节性飓风来临之前,不仅手电筒的销售量增加了,蛋挞的销售量也随之提升,因此当季节性风暴来临时,沃尔玛将蛋挞放在靠近飓风用品的位置,从而增加销售量.在这个过程中,系统不需要知道人们“为什么”对某一类信息感兴趣,只需要知道人们感兴趣的“是什么”,这种洞察力足以重塑包括电子商务在内的许多行业.

从中不难发现,追求相关性并不是一种“退而求其次”的策略,相反,得益于大数据的支撑,原本无法被洞察和挖掘的相关信息能够被用于数据分析和预测,这种相关性作为大数据内部的某种特征客观存在,并帮助研究者更好地捕捉规律、预测未来.

3.2 从相关关系到合作对抗

数据间的相关关系能够用于分析和预测,而行为实体间的相关关系反过来影响数据的产生.在许多场景中,不同个体之前存在多种复杂的合作竞争关系.数据干预者会针对数据决策系统的学习模型,对数据做出相应的修改,以改变数据的特征,从而影响其他数据观察者的决策,不断往复,形成一个决策闭环.

决策数据是一种典型的交互式数据,它是在决策者与决策系统的不断交互中产生的.这种交互体现在多个方面:第一,决策者需要根据决策系统提供的信息作出决策,这种决策的出发点往往是最大化自身的收获,然而在许多场景下,决策者的决策依据对于观察者来说是未知的;第二,决策的过程都是多方交互、持续干预的共同结果,这使得决策数据与流数据一样存在时序性的特点;第三,交互的种类可以是合作,也可以是竞争,甚至两者同时存在,并且对于观察者而言,不同决策者之间的合作竞争关系也可能是未知的.

由于决策数据存在合作对抗的特点,如何从环境中的已观测数据中进行决策是十分困难的.强化学习是一种基于环境行动和最大化预期利益的机器学习方法,通过不断与环境交互从而学习一个回报最大的策略.在任何一个决策系统中,决策的目的都是使决策者获利最大化,因此我们可以对这个过程进行抽象,并利用强化学习的思想进行建模.在不断与环境进行交互反馈的过程中,干预者策略最终会收敛到最优策略,从而实现收益最大化.

3.3 合作对抗场景中的相似性迁移

洞察力不仅体现在数据的相关性,更能够体现在数据的其他层面.人类之所以能够从已有现象总结规律并加以运用,其核心在于举一反三的能力,而其本质上是对数据在不同层次相似性的洞察能力.一个会骑自行车的人,比一个不会骑自行车的人更容易掌握摩托车的驾驶,这是因为两个任务之间存在较大的相似性,骑自行车的知识能够被用于解决骑摩托车这个任务.

决策的过程实际上是多个决策者相互博弈的过程,因此,如果能够定义博弈结构的相似性,就能够将已有经验的决策知识进行迁移[14],从而帮助决策.以强化学习中的均衡迁移问题为例[15],对于一个已知存在纳什均衡的博弈场景 G,如果能够将其博弈过程进行迁移,得到一个与之相似的博弈 G',当然,迁移将不可避免地带来求解上的偏差,目前已有相关证明[16].G 的纳什均衡 p 可以作为 G' 的近似纳什均衡解,从而以相对较小的计算代价快速学习到一个良好的博弈策略.

4 结论

本文从范式理论和机器学习的角度对大数据科学中的主要挑战和科学问题进行梳理,第四范式作为数据密集型科学研究的指导准则,为大数据科学的发展提供了诸多基础,并在气象、环境、医疗、能源等诸多方面取得了很大进展[17].随着移动互联网的发展,第四范式理论也在不断自我完善.基于第四范式的大数据科学不是新瓶旧酒,也非明日黄花.未来

的大数据科学仍存在以下几个方向的挑战:

1)需要完善基于大数据的计算理论研究.目前人类社会仍处于数据加速生产阶段,越来越多的数据将会以更多的形式呈现在人们面前,真正的数据密集型社会即将到来.因此,需要进一步完善和发展大数据相关的计算理论研究,特别是近似计算理论研究.

2)需要寻求与人工智能结合的智能大数据技术.新一代人工智能已在全球范围内蓬勃兴起,作为新一轮产业变革的核心驱动力,正在促进人类生产水平的飞速提高,并加速新一轮科技革命和产业变革.目前的大数据科学主要扮演人工智能的支撑者角色,随着人们生活水平的不断提高,基于大数据的智能融合计算、认知、推理与创造技术仍是未来科学研究的重点突破口.

3)需要构建开放环境的通用大数据平台.在当前的大数据环境中,新一代通信技术已蓄势待发,相信在未来,数据传输的瓶颈效应将大大降低,因此,需要建立以此为支撑的开放通用大数据平台,从而实现大数据下的通用群体智能.

参考文献

References

[1] Kuhn T S.The structure of scientific revolutions[M].University of Chicago Press,2012
[2] Laney D.3D data management:controlling data volume,velocity and variety[J].META Group Research Note,2001,6(70):1
[3] Gantz J,Reinsel D.Extracting value from chaos[J].IDC Iview,2011,1142(2011):1-12
[4] 李国杰.大数据研究的科学价值[J].中国计算机学会通讯,2012,8(9):8-15
LI Guojie.Scientific value on big data research[J].Communications of China Computer Federation,2012,8(9):8-15
[5] Chen M,Mao S,Liu Y.Big data:a survey[J].Mobile Networks and Applications,2014,19(2):171-209.
[6] Oussous A,Benjelloun F Z,Lahcen A A,et al.Big data technologies:a survey[J].Journal of King Saud University-Computer and Information Sciences,2018,30(4):431-448.
[7] Data B.Science in the petabyte era[J].Nature,2008,455(7209):8-9
[8] Lynch C.How do your data grow?[J].Nature,2008,455(7209):28-29
[9] Frankel F,Reid R.Big data:distilling meaning from data[J].Nature,2008,455(7209):30
[10] Howe D,Costanzo M,Fey P,et al.The future of biocuration[J].Nature,2008,455(7209):47-50
[11] Cukier K.Data,data everywhere:a special report on managing information[M].Economist Newspaper,2010
[12] Jonathan T O,Gerald A M,Sandrine B.Special online collection:dealing with data[J].Science,2011,331(6018):639-806
[13] Bryant R,Katz R H,Lazowska E D.Big-data computing:creating revolutionary breakthroughs in commerce,science and society[J].2008
[14] Pan S J,Yang Q.A survey on transfer learning[J].IEEE Transactions on Knowledge and Data Engineering,2009,22(10):1345-1359.
[15] Hu Y J,Gao Y,An B.Accelerating multiagent reinforcement learning by equilibrium transfer[J].IEEE Transactions on Cybernetics,2015,45(7):1289-1302
[16] Claus C,Boutilier C.The dynamics of reinforcement learning in cooperative multiagent systems[J].AAAI/IAAI,1998,1998:746-752
[17] 孟小峰,慈祥.大数据管理:概念、技术与挑战[J].计算机研究与发展,2013,50(1):146-169
MENG Xiaofeng,CI Xiang.Big data management:concepts,techniques and challenges[J].Journal of Computer Research and Development,2013,50(1):146-169

Big data science from the perspective of the fourth paradigm

GU Zheng[1] GAO Yang[1]
1 State Key Laboratory for Novel Software Technology at Nanjing University,Nanjing 210023

Abstract With the rapid development of Internet of things and cloud computing,big data and its related science have become the focus of industry and academia.In this paper,we analyzes big data science from the perspective of paradigm theory and expounds the difference and connection between big data and traditional research.Three major challenges brought by big data are proposed in perspective of machine learning,with the corresponding scientific problems following.In addition,this paper introduces several insights of big data science from the perspective of the forth paradigm and its positive significance.In the end,we summarize and look forward to the challenges of big data science in the future.

Key words paradigm theory; big data; machine learning