

doi: 10.14132/j.cnki.1673-5439.2015.05.003

# 大数据与计算可持续性

周绮凤<sup>1</sup> 李涛<sup>2</sup>(1. 厦门大学自动化系 福建 厦门 361005  
2. 南京邮电大学计算机学院 江苏 南京 210023)

**摘要:** 可持续发展是国际社会共同关注的焦点问题。然而,如何从现有的政策驱动落实到技术驱动是切实实现可持续发展的瓶颈问题。近年来,新兴的计算可持续性研究,为解决该难题的一个有效途径和一个新的研究热点。大数据时代的来临为计算可持续性研究带来了机遇,同时也带来了问题复杂性、计算效率、方法可扩展性等新的挑战。针对国内对计算可持续性需求的紧迫性而目前尚未开展相关研究的现状,文中首先介绍了计算可持续性的重点研究内容和主要任务。其次,分析了在大数据时代,运用计算机及信息科学的技术来提高管理和分配自然资源的必要性和有效性,以及计算可持续性研究面临的机遇与挑战,并通过具体案例说明基于数据驱动的计算可持续性在大数据环境下发挥的关键作用。最后,从技术发展和实际应用的角度探讨了未来计算可持续性的一些研究方向。

**关键词:** 大数据; 计算可持续性; 可持续发展; 数据挖掘

**中图分类号:** TP391.4    **文献标志码:** A    **文章编号:** 1673-5439(2015)05-0020-12

## Big data and computational sustainability

ZHOU Qifeng<sup>1</sup>, LI Tao<sup>2</sup>(1. Automation Department, Xiamen University, Xiamen 361005, China  
2. School of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

**Abstract:** The sustainable development has attracted many attention worldwide. However, one of the biggest problems in sustainable development is the practical implementation of existing policies. Computational sustainability, as a new research field, will become an effective tool to address the aforementioned problem. The advent of the big data era brings many research opportunities to computational sustainability, but also poses many challenges, such as computational complexity, modeling complexity and scalability. This paper, systematically introduces the research topics and tasks in computational sustainability, and discusses the use of information technology to improve the effectiveness and the efficiency of managing and allocating natural resources in big data era. Case studies illustrate that data-driven solutions of computational sustainability play key roles in a big data environment. Finally, future research directions of computational sustainability from both the technical and practical application perspectives are discussed.

**Key words:** big data; computational sustainability; sustainable development; data mining

在资源枯竭、环境恶化日益严重的今天,可持续发展已成为世界各国争相研究的一大热点问题。在

我国,淡水、能源等资源短缺,雾霾、干旱等异常天气频发。这些生态问题严重影响了人们的正常生产和

收稿日期: 2015-08-21    本刊网址: <http://nyzr.njupt.edu.cn>

基金项目: 国家自然科学基金(61503313)和江苏省社会安全图像与视频理解重点实验室创新基金(30920140122007)资助项目

通讯作者: 李涛    电话: 025-85866355    E-mail: towerlee@njupt.edu.cn

生活。以近年来频发的雾霾天气为例,2013年1月4次雾霾过程笼罩全国30个省(区、市),在北京,仅有5天不是雾霾天。2014年1月4日,国家减灾办、民政部通报2013年自然灾情,首次将危害健康的雾霾天气纳入。如何改善我们的生活环境,实现可持续的发展和未来,是当前国际社会亟待解决的一个焦点问题。

计算可持续性(Computational Sustainability)是一个新的跨学科研究领域<sup>[1-3]</sup>,其目的是综合应用计算机科学、信息科学、运筹学、应用数学和统计学等多学科交叉技术,平衡环境、经济以及社会需求,以支持可持续的发展。计算可持续性研究涉及能源、生态、经济、环境等众多学科,汇集了计算领域和各种具有悠久传统的可持续性问题,如生态多样性、自然资源管理、生物与环境工程和资源经济学等。

计算可持续性研究的重点是开发计算模型、数学模型及相关方法,以帮助解决一些与可持续发展相关的最具挑战性的问题。然而,随着卫星技术、传感技术的日新月异,我们每天可以采集到的各类环境数据无时无刻不在增加。大数据时代的来临,为计算可持续性研究带来了极大的挑战,同时也带来了新的机遇。一方面,大数据限制了研究者可以使用相对简单的分析技术,已有的构建和优化这些模型的方法,遇到了可扩展性等的挑战;另一方面,大数据蕴含丰富的信息和潜在的知识,给人们研究可持续发展提供了一个以数据为驱动的全新的研究方式,将极大地促进计算可持续性的发展。

目前,数据驱动的计算可持续性已成为一个国际研究热点,各种会议和研讨会正在持续热烈的举行。近年来,在人工智能(Artificial Intelligence, AI)、机器学习(Machine Learning, ML)等国际权威学术会议上,每年都有关于可持续发展的专题研讨,表1列出了近年召开的部分计算可持续性相关的会议。此外,一些国际权威期刊也有关于可持续发展的专刊,如ACM Transactions on Intelligent System and Technology(ACM TIST) Special Issue on Computational Sustainability等。

在大数据环境下,应用人工智能、机器学习、数据挖掘等方法研究可持续发展是一个必然的趋势。同时,可持续发展问题的复杂性、可扩展性以及预测、建模和优化、控制、高性能计算和分布式计算平台等领域的影响,为人工智能、机器学习带来了新的挑战。如何利用现有的技术以及探索开发新的技术以应对这些挑战,成为该领域的研究热点。然而,国

内目前在这方面的研究尚处于起步阶段,鲜有文献等报道。本文针对大数据环境下,计算可持续性研究的现状、面临的机遇与挑战,系统介绍计算可持续性的任务、数据特点、分析方法,并以具体的案例研究,说明在大数据时代,基于数据驱动的方法在解决计算可持续性面临的挑战中发挥的关键作用。

表1 近年召开的人工智能、机器学习与计算可持续性相关会议

会议名称	相关信息
AAAI-15-CompSust	Special Track on Computational Sustainability, during the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30 2015, Austin, Texas, USA
IJCAI-15-CompSust	Special track on AI and Computational Sustainability (AICS), July 25-31 2015, Buenos Aires, Argentina
SOCS-14-CompSust	Symposium on Computational Sustainability, November 26-28 2014, Gölrlitz, Germany
IJCAI-13-CompSust	Special track on AI and Computational Sustainability (AICS), August 3-9 2013, Beijing, China
AAAI-13-CompSust	Special Track on Computational Sustainability and AI, during the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18 2013, Bellevue, Washington, USA
NIPS-13-CompSust	NIPS 2013 Workshop: Machine Learning for Sustainability, December 5-10 2013, Lake Tahoe, Nevada, USA
CROCS-12-CompSust	International Workshop on Constraint Reasoning and Optimization for Computational Sustainability, October 8 2012, Quebec, Canada
AAAI-12-CompSust	Special Track on Computational Sustainability and Artificial Intelligence, during the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26 2012, Toronto
Compsust12 Conference	Third International Conference on Computational Sustainability held July 4-6 2012, Copenhagen, Denmark
Sustainability at CHI 2012	Sustainability Featured Community at The ACM SIGCHI Conference on Human Factors in Computing Systems, May 5-10 2012, Austin, TX
AAAI-11-CompSust	Special Track on Computational Sustainability and Artificial Intelligence during the Twenty-Fifth AAAI Conference on Artificial Intelligence held August 7-11 2011, San Francisco, USA
CompSust10 Conference	2nd International Conference on Computational Sustainability, Massachusetts Institute of Technology (MIT), June 28-30 2010, Cambridge MA, USA

## 1 计算可持续性及其研究现状

### 1.1 计算可持续性主要研究内容

实现可持续发展的关键是如何制定合理的能够平衡环境、经济和社会需求的复杂决策。然而,自然、社会、经济系统的高度复杂性、动态性、不确定性使得实现这一最优或近似最优的决策成为一个巨大的挑战。计算可持续性正是为解决这一挑战而出现的一

个新兴的研究领域 其研究的重点是针对可持续发展问题 开发计算模型、数学模型及相关方法 制定资源的管理和分配决策等。计算可持续性研究涉及面极其广泛 从野生动物保护 生物多样性到社会经济需求平衡 大规模环境布署 以及再生能源的管理等。表 2 列出了目前计算可持续性重点研究的内容。

表 2 计算可持续性主要研究内容

类别	研究内容
气象	气候数据异常值检测,时空数据聚类跟踪,海洋气候数据的分析和跟踪,飓风强度预测等
环境	生态保护 物种分布估计,环境检测,保护区资源规划,多物种分布估计,鸟类迁徙建模等
农业和土地	土地覆盖监测,水资源管理,森林野火管理等
智能电网	电器负载估计,电力检修事件预测,电网可靠性评估,光伏输出功率监测等
经济	节能产品 绿色建筑,节能汽车,节能数据中心等
再生能源	水力发电、风力发电等设施定位,能源发电的最佳组合和存储技术等
社会	社会可持续发展 灾难信息管理,居民环境变迁,人口规模控制与预测等

1.2 计算可持续性主要研究方法

计算可持续性方法涉及机器学习、数据挖掘、统计建模、动态规划、推理优化等多种方法以及这些方法的综合运用(如图 1 所示)。

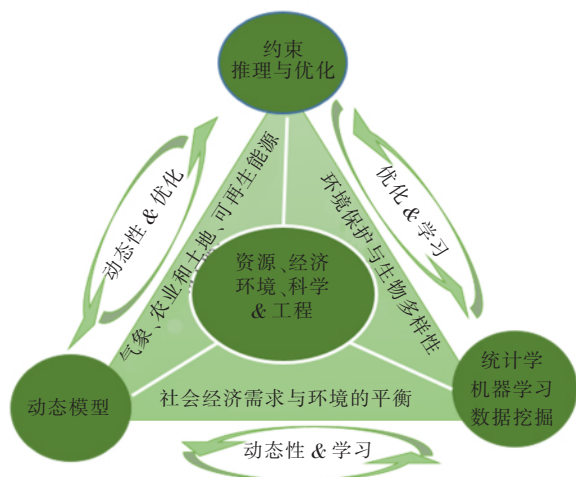


图 1 计算可持续性的主要研究内容及方法<sup>[1]</sup>

以下通过生态环境与社会发展中的两个例子说明计算可持续性的具体研究内容及方法。

(1) 生态环境。生态环境恶化、城市化发展以及过度的土地开发利用,使得自然栖息地减少和破碎 这是导致物种减少和灭绝的主要原因之一。建立保护栖息地(如国家野生动物保护区)等是保护生物多样性的一个策略,但是也会大量消耗政府的

财政预算。

在给定建立保护区的可用资源条件限制下,如何选择最合适的保护区(包括位置、物种、气候、经济条件等)是一个生物可持续发展问题。从数学的角度来看,选址问题本质上是一个多目标优化问题(Multi-objective Optimization Problem, MOP),即在满足预算等一个或多个约束的条件下,选择适宜生品种较多等目标的地区<sup>[4-6]</sup>。与此相关的一个有趣的研究是建立“保护廊道(Conservation Corridors)”,即如何将各个零散的保护地与破碎的自然生物区连接起来,从而产生更高等级的生态效益<sup>[7-8]</sup>。Cornell 大学的计算可持续性研究机构(Institute for Computational Sustainability, ICS)将该问题转化为组合优化中的连接子图(Connection Sub-Graph Problem)问题,研究如何将 Yellowstone, Salmon-Selway 以及 Northern Continental Divide Ecosystems 三大核心生态地区(涵盖 64 个区域)的野生动物栖息地连接起来<sup>[9]</sup>。这一大规模优化问题为当前的计算方法提出了新的需求。他们提出的带约束条件的混合整数优化模型把预算作为一类约束带入了优化问题。结果显示最优解可以大幅减少建立保护廊道的支出,并同时达到最大利用率。

选址和廊道的设计问题的复杂性,随着土地需求时间段不同(例如,申购、保护地役权、拍卖),动态和随机的环境变化,以及考虑多种物种习性而不断增加。例如,保护鸟类栖息地和鸟类走廊设计,需要考虑鸟类迁徙的复杂种群动态,横跨不同的气候,以及地理拓扑结构等。因此,模拟复杂的物种分布和开发保护策略需要新的大规模的随机优化方法。此外,为了获得适当的模型参数,并确定当前物种分布,需要利用机器学习和统计技术分析大量的原始数据<sup>[10-11]</sup>。

(2) 社会发展。社会可持续发展是指在改善人类生活质量的同时不应超越支撑发展的生态系统的负载能力。计算可持续性为我们提供了如何平衡环境、经济和社会可持续发展的各种计算工具。目前在该方面的研究内容主要包括:灾难信息管理,即快速有效应对灾难、饥荒等自然灾害的自动化决策支持系统;生命周期评估,即对规划和建设项目实施后可能造成的环境影响进行分析、预测和评估,提出预防或者减轻不良环境影响的对策和措施;以及居民环境变迁等研究。

如 ICS 的 Chris Barret 研究了非洲社会的经济与贫困、粮食安全及环境负担之间的相互关系<sup>[12]</sup>。

针对东非游牧民的生活环境,研究牧民的迁移模式,及如何建立迁移决策的预测模型。为此,他们基于对家庭、水源、气候等实际数据的分析,利用机器学习的方法来确定模型结构并估计模型的参数。这些模型将帮助决策者预测未来可能的政策干预与环境变化对游牧民族的影响,以改善数万牧民的生计。该问题涉及开发新的技术和方法来处理大规模、动态结构、离散变量等问题,以及开发新的能够同时支持描述型研究以及预测政策分析的计算模型。

## 2 大数据时代的计算可持续性

大数据是信息时代的重要产物,大数据的科学价值和社会价值正在逐渐体现。随着卫星技术、传感技术的广泛应用,海量的生态数据不断涌现。已有的计算可持续性方法面临如何有效处理海量、高维、复杂的生态大数据的难题,而大数据蕴含的丰富的信息和潜在的知识,为计算可持续性研究提供了以数据为驱动的全新的研究方式。大数据环境下,出现在许多领域的可持续发展的计算问题与人工智能和数据挖掘有紧密的相关性,诸如,图模式与概率推理、统计学习、数据与图挖掘、约束与随机优化、不确定性推理、时空数据建模及网络科学等。我们可以利用大数据挖掘技术解决各种具有挑战性的计算可持续性问题。

### 2.1 对大数据的理解和认识

国内外不同的专家和学者对大数据有不同的理解定义,中国科学院计算技术研究所李国杰院士认为:大数据就是“海量数据”加“复杂数据类型”<sup>[13]</sup>。维基百科对大数据的定义是:大数据是由于规模、复杂性、实时而导致的无法在一定时间内用常规软件工具对其进行获取、存贮、搜索、分享、分析、可视化的数据集<sup>[14]</sup>。Gartner 咨询公司给出的定义是:大数据是需要新处理模式具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产<sup>[15]</sup>。而互联网数据中心将大数据定义为:为更经济地从高频率的、大容量的、不同结构和类型的数据中获取价值而设计的新一代架构和技术<sup>[16]</sup>。

笔者认为,大数据的核心和本质是应用、算法、数据和平台4个要素的有机结合(如图2所示)。首先,大数据是应用驱动的,大数据来源于实践,海量数据产生于实际应用中。其次,对数据的有效管理和存储是实现大数据分析的基础。此外,挖掘大数据所蕴含的有用信息,需要开发分析和解决问题

的相关数据挖掘和机器学习算法。算法的设计和开发要以具体的应用数据为驱动,同时也要在实际问题中得到应用和验证,而算法的实现需要高效的处理平台。这个思想是对上述大数据的理解和认识的一个综合与凝练,体现了大数据的本质和核心。因此,本文在此框架下,进一步探讨大数据时代的计算可持续性数据的特征、算法设计及平台构建所面临的机遇与挑战,以及未来的发展方向。

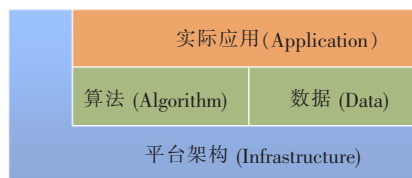


图2 大数据架构

### 2.2 计算可持续性数据及其特征

计算可持续性涉及众多学科和广泛的应用领域,其数据规模大、结构复杂、形式多样,本文以可持续发展中典型的生态数据为例,介绍计算可持续性数据及其特点。

(1) 广泛的数据来源。与传统的地球科学利用数学模型模拟自然现象的方法不同,由数据驱动的数据挖掘提供了全新的研究方法。Dietterich 等人<sup>[17]</sup>提出了生态信息学(Ecosystem Informatics)概念,强调现在的生态学也可以用类似生物信息学的方法来研究,通过收集大量数据来学习分析预测模型。

目前已有很多研究机构、部门和学者搜集了大量可供研究的生态数据集<sup>[18]</sup>,例如:

① 第三届 SensorKDD 研讨会(3rd International Workshop on Knowledge Discovery from Sensor Data)<sup>[19]</sup>提供了50年近2万个地点的气温和降水量时空数据,数据量达到了几个GB。这届研讨会中的挑战比赛旨在利用日益普遍的传感器,急剧增加的监测数据,寻找出气候上的突变或者缓慢的偏移。

② 明尼苏达大学气候研究小组的 Expeditions 研究项目一直在使用数据挖掘的方法研究气候变化<sup>[20]</sup>。他们从数据中发现气候的微小变化,改善传统大气科学的模拟模型,提供了了解复杂生态系统的新途径。

③ 美国国家海洋与大气管理局 NOAA(National Oceanic and Atmospheric Administration)的地球系统研究实验室(Earth System Research Laboratory)公布了数百个数据集。例如,比较常用的数据集 NCEP/NCAR ReanalysisI 包括了 192 \* 94 格点从 1948 年

至今的气压、气温、降水量等遥感数据。MODIS 中分辨率光谱成像仪数据是从美国航空航天局 NASA (National Aeronautics and Space Administration) 的 Terra 和 Aqua 卫星上得到<sup>[21]</sup>。这两颗卫星每 1~2 天观测地球,从多个不同的分辨率下收集 36 大类的数据。

④ 地面观测和预报项目 TOPS (Terrestrial Observation and Prediction System) 目标是建立一个对各种传感器数据的异构数据分析处理的统一基础框架,并且建立短时预报和预测系统。传感器数据可以来自卫星、数据、地面传感器等等。所使用的数据存储在 NASA 的 Earth Exchange 网站<sup>[22]</sup>。不仅包括了 MODIS 卫星数据还包括了地面探测器的数据如生物量、碳存储等观测数据。

⑤ 哥伦比亚大学国际气候和社会研究所 (IRI International Research Institute for Climate and Society) 的 IRI/LDEO Climate Data Library 提供了 300 多个地球和气候观测数据<sup>[23]</sup>。比如最新公布的 NOAA NCEP-DOE Reanalysis II 数据集及 GPCP 气候降水量数据产品。IRI 研究所致力于通过科学途径让社会特别是发展中国家了解天气的影响,改善人群居住环境。IRI 的项目包括气候项目、环境监测项目、非洲项目等等。

(2) 复杂的数据特征“4V+4V”。生态大数据形式多样,结构复杂,具有典型的大数据的“4V 特征”<sup>[18]</sup>:

① Volume (大量): 随着大数据时代的到来,可持续发展中所使用的数据也有达到 PB 级别的。如何利用这些大量的数据,要求数据处理算法具有良好的伸缩性。

② Variety (多样): 计算可持续性问题涉及的数据类型繁多。如,文本、气象图像、视频,以及传感器检测数据、地理位置信息等。

③ Velocity (高速): 观测数据具有时效性,每时每刻都需要处理,需要有动态的数据体系。如在灾害管理等问题上,就需要建立实时的分析模型,比如基于流式的数据处理系统。

④ Value (价值): 海量的数据蕴涵巨大的价值,合理利用低密度价值的的数据并对其进行正确、准确的分析,将会带来巨大的价值。

计算可持续性涉及领域的广泛性和数据来源的丰富性使其具有更多更复杂的大数据特征,本文从数据处理的角度总结出如下几个新的性质:

① Variable (变化性): 在不同的场景、不同的研

究目标下数据的结构和意思可能会发生变化,因此,在实际研究中要考虑具体的上下文场景。

② Veracity (真实性): 获取真实、可靠的数据,是保证分析结果准确、有效的前提。如在人口发展问题中,利用真实数据获得的结果才能制定有效的指导方案。

③ Volatility (波动性): 由于数据本身含有噪音及分析流程的不规范性,导致采用不同的算法或不同分析过程与手段会得到变化的或有差异的分析结果。

④ Visualization (可视化): 在大数据环境下,通过数据可视化可以更加直观的阐释数据的意义,帮助理解数据,解释结果。

### 2.3 大数据环境下计算可持续性研究现状

大数据使得已有的研究方式从探求难以捉摸的因果关系,转而关注发现和使用事物的相关关系。数据挖掘等基于数据驱动的研究方法可以帮助解决更加复杂、更大规模的计算可持续性问题。

基于数据驱动的可持续发展的计算任务包括气象、生物、灾害管理、能源等众多领域的研究。目前,越来越多的研究人员开始利用数据挖掘技术解决计算可持续性问题,如下给出一些代表性的研究工作<sup>[18]</sup>。

(1) 气象。2001 年,明尼苏达大学 Kumar 等人<sup>[24]</sup>较早用数据挖掘技术开展地球生态数据的研究。生态数据比如湿度、温度、降雨量、海平面高度等变量同时具有时间和空间属性。从时间这个维度来看,这些数据又具有周期变化特征,比如海平面高度数据每年随着季节变化而周期变化。厄尔尼诺现象、温室效应等气候异常现象是偏离正常周期变化的异常特征,而这些异常总是会被正常周期变化掩盖。所以他们在做数据预处理时,用了很多种方法去除正常周期变化,比如奇异特征值分解 (Singular Value Decomposition, SVD)、离散傅里叶变换 (Discrete Fourier Transform, DFT) 和移动平均数 (Moving Average)。文献 [24] 研究了同一个地点的多个变量的时间序列。对这些生态数据中每一个变量,把偏移正常范围的值作为一个事件编制出一系列类似于数据库中的事务。如同对数据库系统的事务进行关联分析,来找出这些气候变量之间的依赖关系。最后,通过聚类方法把数据相似的地区连接起来,找出大范围的气象特征。

地球科学中各种各样的气候指标基本都是手工制定的。随着卫星观测数据的日益丰富,人们开始

从大量的数据中发现其他相关的气候指标。但是传统方法如 PCA(Principal Component Analysis)<sup>[25]</sup>、SVD(Singular Value Decomposition)<sup>[25]</sup> 只能找出特征最强的几个互相正交的变量。一些互相关联的特征较弱的指标被遗漏掉了。Steinbach 等人<sup>[26-27]</sup> 提出了 SNN(Shared Nearest Neighbor) 聚类方法。因为强烈的天气事件一般广泛地发生在一个区域之内,聚类技术可以找出这种具有数据一致性的区域。实验表明 SNN 不要求这些指标互相正交,这样所得到的结果更加容易解释,并且能发掘出强度较弱的指标。SNN 在海水表面温度数据集上找到的指标中有些是已知的,证明了这个方法的有效性,而且结果中有些则是已知指标的变体但具有更好的环境描述能力。

在很多气候数据挖掘的工作中,仅仅使用了单个变量而没有充分考虑世界气候在地点上所形成的复杂关系。Steinhaeuser 等人<sup>[28-29]</sup> 使用了经典的 NCEP/NCAR 数据集,在 50 年 720 多个世界地点上使用了多个气候指标,如气温、湿度、降雨量等。在一个基于统计相关的空间上计算物理地点之间的气候相似性。这样不仅把时间和空间维度考虑在内,也同时利用了多个气候指标。

Bhaduri 等人<sup>[30]</sup> 研究了如何从 PB 级别的高维度数据中寻找离群值。NASA 的 Terra 和 Aqua 卫星升空,收集到的数据从几百 TB 上升到了几十个 PB 数量级。而且这些数据分布式地存储在不同地方,比如 NASA 的分布式主动存储中心。研究者只能研究数据集的一小部分,或者把所有的数据汇集到一起,但是这种方法不仅受到数据大小的限制而且由不同的研究队伍完成。Bhaduri 等人<sup>[30]</sup> 还设计了一个基于 One-class SVM(Support Vector Machine) 的分布式模型用来寻找 MODIS(Moderate Resolution Imaging Spectroradiometer) 卫星数据中的离群值。实验表明这个方法不仅精确度达到了 99%,而且只需要很少的带宽传送分布式数据。

(2) 生态保护。估计物种分布是生态学中的基本问题。Ferrier 等<sup>[31]</sup> 使用标准的物种数据评估传统模型和最近发展的统计模型和机器学习模型。他们所使用的数据覆盖了世界范围的 6 个区域,包括了 200 多种生物,并提供了非常详实的多种模型比较结果。其中,传统的估计模型包括包络模型族(Envelop Model) 中的 BIOCLIM、DOMAIN 和 LIVES,统计学习模型包括广义线性模型 GLM(Generalized Linear Models)、广义可加模型 GAM(Generalized

Additive Models<sup>[32]</sup>)、多元自适应样条回归 MARS(Multivariate Adaptive Regression Splines<sup>[32]</sup>)、最大熵模型(Maximum Entropy Model) 等。多个实验结果利用 ROC 曲线(Receiver Operating Characteristic) 中的 AUC(Area Under Curve) 指标等多个标准进行比较,评估结果充分显示,新近发展的统计学习模型远远优于传统模型。

目前除了大量的传感器和卫星数据之外,还有很多数据是通过人力得到的,比如康奈尔大学的鸟类实验室 ebird 项目<sup>[33]</sup>。在这个项目中,志愿者在野外观察到鸟类活动时提交报告,详细记录了鸟的活动地点时间等信息。有了这些手工记录的数据后,估计物种的分布就是去学习能够判断某个地点是否会出现鸟的模型。人工采集的数据给分析带了很多困难。数据不仅存在很大的抽样偏差,而且有些鸟类不容易被观察到,志愿者的专业知识水平也有很大差别。直接从这些数据中估计鸟类的分布是一个在高维度空间推测概率密度的难题。康奈尔大学研究小组提出的方法回避了这个难点,转而去估计已有的概率密度同需要的概率密度的比值。他们在传统的模型上把回归树模型(Regression Tree) 和 Logistic 回归模型(Logistic Regression) 通过 boosting 的方式结合在一起,得到了非常好的预测结果。

(3) 农业和土地。Ekasingh 等人<sup>[34]</sup> 用决策树模拟了如何在泰国北部地区选择作物种植来达到环境效益和经济利益的平衡。IWRAM(Integrated Water Resource Assessment and Management) 项目同时从经济、环境、社会多个角度出发研究如何管理自然资源。这个项目用问卷调查的方式收集了三个流域上耕地、生产费用、作物的经济效益和农民劳动力等信息。为了科学合理利用水土资源,他们同时收集了农学家的建议。基于这些信息,他们用 C4.5 算法分别为雨季和旱季建立了不同的决策树,用于指导作物的种植。

Boriah 等人<sup>[35]</sup> 展示了如何从大量高维的生态数据中监测土地覆盖变化。通过监测土地表面变迁可以掌握森林砍伐、城市化进程、农业集约化对自然植被的破坏程度。这种土地覆盖变化直接影响到当地和全球气候。他们提出的分析方法利用了数据中时间空间之间的关系,可以处理大规模高分辨率的地球科学数据。文献[35]用美国航空航天局 NASA 的地球观测系统(Earth Observation System, EOS) 卫星数据的中分辨率成像光谱仪 MODIS 得到增强植被指数 EVI(Enhanced Vegetation Index)。这个指标

衡量了土地绿色植被覆盖程度。在加利福尼亚湾区的土地覆盖变化的分析中,EVI 数据包含了从 2000 年 2 月至 2006 年 5 月之间 38 万多地点的数据,找出了高尔夫球场建筑工地,农田开垦,森林大火等事件。

在森林野火的管理方面,烟雾进入大气层的高度是导致野火顺风蔓延的主要因素。Mazzoni 等人<sup>[36]</sup>利用 Terra 卫星上的多角度成像光谱仪 MISR (Multi-Angle Imaging SpectroRadiometer) 和中分辨率成像光谱仪 MODIS 在阿拉斯加和毗邻的加拿大 Yukon 地区约 4 个月的观测数据,建立了一个基于 SVM 的从云层和气溶胶粒子中分辨出野火烟羽、测量烟雾高度等数据的原型系统,为了解烟雾高度同野火和本地气候提供了自动化工具。

(4) 智能电网。智能电网、智能交通和绿色计算等工程方法目标是要实现化石燃料等不可再生能源科学合理的利用,延缓消耗速度和匮乏趋势,避免对这些资源的开采而造成的生态破坏,减少碳排放,以及对可再生能源的充分有效利用。美国能源部把智能电网定义为:一个完全自动化的电力输送网络。它能够监视和控制每个用户和每个电网节点,并且保证电厂和终端用户以及整个输配电过程中所有节点之间的信息和电能的双向流动。智能电网对各种工程领域提出了挑战,也对数据挖掘人工智能提出了新的挑战<sup>[37]</sup>。

目前,电网的最大挑战是陈旧的电网设施结构。频繁发生的电力故障表明陈旧的电网基础设施导致电网达不到可靠性要求。根据美国能源部的资料显示,美国大部分电网已经有至少 120 年的历史,电力缺口达到一百兆瓦,5 万人受到影响<sup>[38]</sup>。在数据挖掘和机器学习领域,MIT 的 Rudin 等人 and 爱迪生联合电器公司(Consolidated Edsion)合作开发了智能挖掘学习系统 NOVA(Neutral Online Visualization-aided Autonomic Evaluation Framework)<sup>[39-43]</sup>来挖掘

历史电网数据,预测电网可能出现故障的设备,从而指导电网公司的维护和维修工作。这样电力设备可以在出现故障之前得到保养维修,避免发生级联式的电力中断。NOVA 系统借鉴了信息检索领域的排序技术,按可能发生故障的概率对馈线、电缆、变压器等设备进行排名。该系统使用了 SVM Regression 方法<sup>[42]</sup>对设备的故障平均时间 MTBF(Mean Time Between Failure)进行估计。不仅如此,系统还对电力公司使用改进的方法检修电力设备后所得到的电网情况进行了评估。

## 2.4 实例分析

### 2.4.1 基于数据驱动的气象分析

这里给出一个采用数据挖掘方法分析气象数据的例子。气象数据上的一个典型的数据挖掘应用是找出地球上具有相同气候变化的区域和远距离地区之间的气候关联。数据挖掘中的聚类分析可以自动分割出这些区域。聚类使得在同一个区域中的数据之间相似,不同区域之间的数据尽量不同。

我们采用著名的 NCEP/NCAR Reanalysis 数据,选择来自 NOAA 网站上的海水月平均表面温度数据集(Skin Temperature)作为实验数据<sup>[43]</sup>。用 K-means 方法进行聚类。每一个地点上都有一组随时间变化的表面温度数据。两个地点之间的相似性可以通过相关性得到。同时使用 K-means 方法对海洋区域进行聚类,因为陆地和海洋的表面温度有较大的差异。

实验结果如图 3 所示。图中除了用深色表示陆地之外,每一个不同颜色区域表示不同聚类区域。从图 3(a)中,我们可以看到东太平洋赤道附近的赤道洋流构成的区域,大西洋的墨西哥湾暖流影响的北大西洋区域等。由于位势高度取决于气压数值,因此图 3(b)和图 3(c)显示出一定的相似性。大部分类的形状都是沿着经度的带状区域,位势高度显示出由气压形成的几条带状区域,表明了风带构成的大气环流。

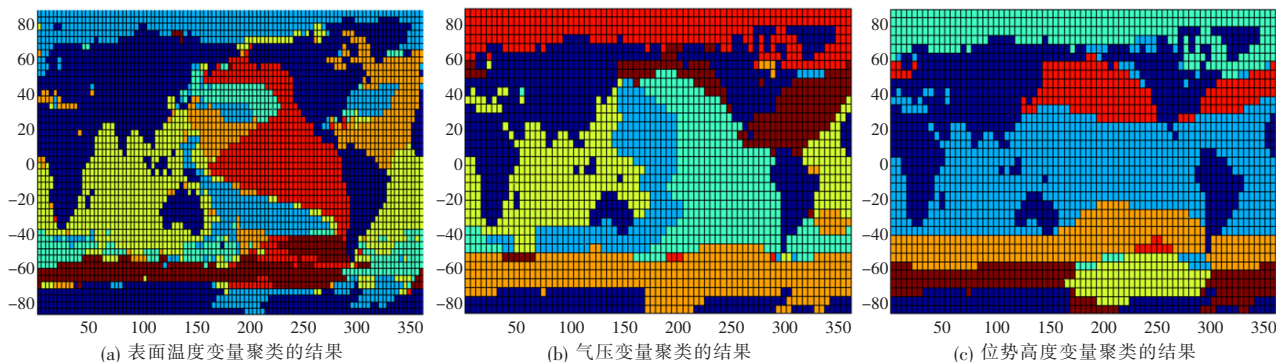


图 3 使用 K-means 对海洋区域聚类的结果

### 2.4.2 基于数据驱动的大规模建筑环境影响评价

建筑区域环境影响评价(Regional Environmental Impact Assessment, REIA)是实现建筑生态可持续发展的重要的前沿性课题<sup>[44]</sup>。该课题是在一定区域内,综合考虑大规模建筑群或建筑规划对环境的影响,旨在解决如下问题:① 给定区域内建筑规划方案,评价该方案对环境的影响;② 给定基本的建筑耗材类型、数量等,确定什么样的建筑形式或系统能满足环保要求。

由于区域级建筑环评规模大、代价高、数据来源多样且关系复杂,仅从建筑生态学的角度已无法实现高效准确的评价,而必须借助相关信息学的方法从大量数据(建筑物及其属性)中寻找潜在的规律(建筑物之间及建筑与环境影响的关系),约减评估规模,从而拓展已有评估方法的应用范围。为降低评估的代价,需要利用建筑属性和专家知识,寻找区域内具有相似环境影响的建筑集合,从而只需分析区域内有限数量的建筑,就可以确定该地区所有建筑物对环境的影响。因此,区域级建筑环评规模约减的核心问题可视为建筑社区发现问题,即寻找建筑区域聚类,该聚类需要符合建筑的自然分布,且表达多个建筑个体具有类似环境影响的特性。

针对区域级建筑环评规模大、建筑形式多样、先验知识不同的问题,我们提出了一种基于聚类分析的区域级建筑环评规模约减方法,即通过建筑社区发现、可移植半监督聚类集成,以及最小子集确定三个步骤的研究,约减评估的数量。

具体研究思路如图 4 所示,通过寻找区域内具有相似环境影响的建筑集合,约减评估规模,从而利用已有技术评价该区域内有限数量的建筑,就可以

确定该地区所有建筑物对环境的影响,降低评估代价。

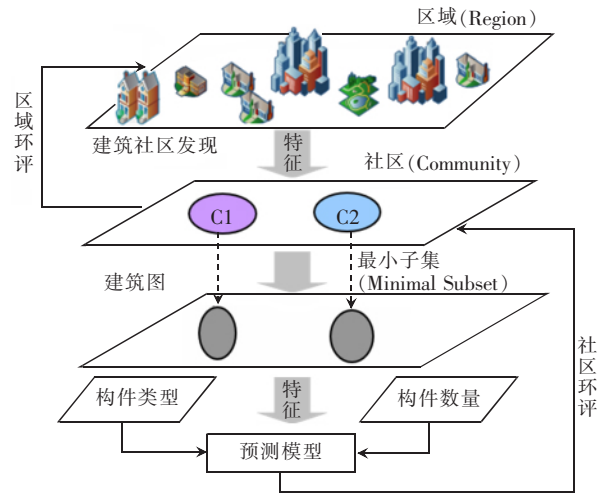


图 4 大数据环境下区域级建筑环评实现模块<sup>[44]</sup>

## 3 计算可持续性研究的挑战与未来研究方向

### 3.1 机遇与挑战并存

目前,包括 ICS 在内的很多研究机构和学者,正在开展各项研究,建立和丰富计算可持续性研究领域,并试图证明计算等方法在可持续发展研究中的有效性。计算可持续性研究是一个机遇与挑战并存的研究领域,其广泛的应用也吸引了各个学科的研究者投入到该领域的研究中。图 5 给出了人工智能杂志 2014 年计算可持续性专刊发表的可持续与人工智能相关的文章主题,可以看到,越来越多的人工智能、机器学习、数据挖掘等方法被应用到城市规划、物种分布、政策制定、健康、农业、交通、能源、智能电网等多种可持续性问题的研究中<sup>[45-50]</sup>。

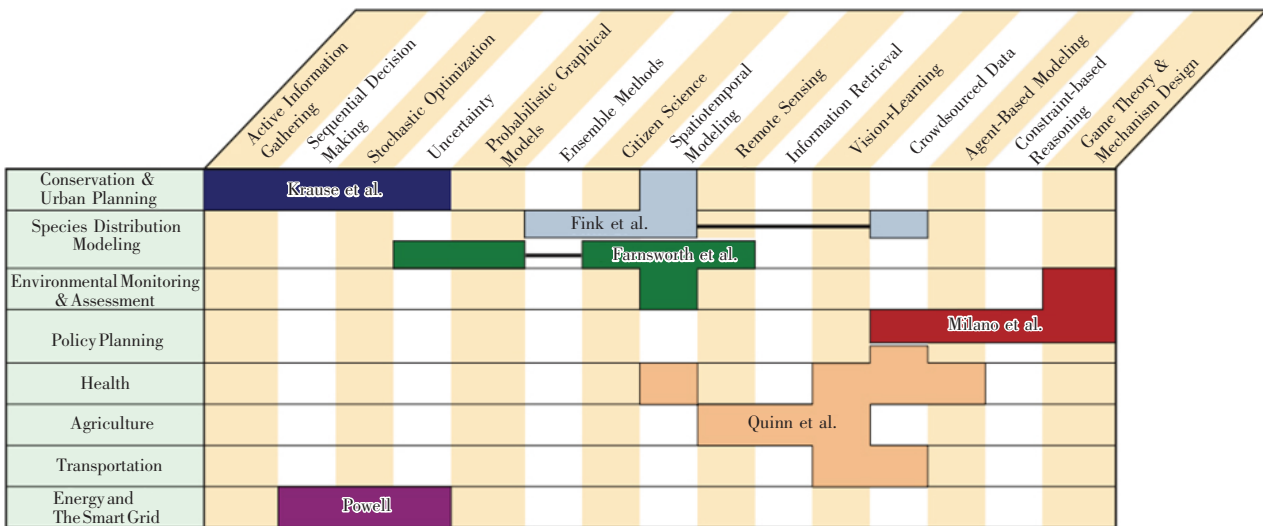


图 5 2014 年 AI 杂志 CompSust 专刊发表的计算可持续性相关研究主题



尽管目前的文献调查表明,可持续发展的关键问题最终可以转化成计算和信息科学领域的决策和优化问题。但是由于计算可持续性涵盖生态环境、自然资源、大气科学、材料科学和生物与环境工程等学科中的不同问题,目前,计算机科学家还未能深入开展计算可持续性研究,该领域尚需构建新的研究方法和研究体系来解决与之相关的问题。

计算可持续性研究的高度跨学科性,数据的大规模、复杂性等问题,也影响和扩展了计算机科学本身的领域界限,需要整合多种计算机领域的技术和应用数学,如约束推理、优化、机器学习、数据挖掘和动力系统。此外,还需要开发计算上可行的,能够解决各种高度相互作用的,以及具有冲突的问题的分析系统。计算可持续性问题的难度与复杂性如图6所示<sup>[1]</sup>。

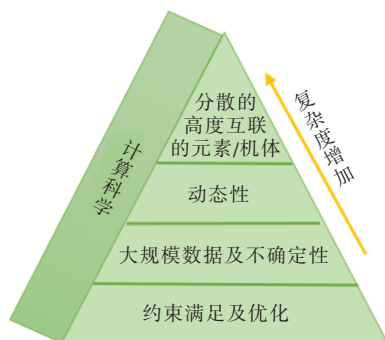


图6 计算可持续性问题的复杂性等级<sup>[1]</sup>

可持续性问题的多维度视角(如经济、社会和环境),以及大容量、高度动态,具有不确定性和相互作用的复杂性质,为其研究带来极大的挑战。传统的计算科学基于最坏情况(Worst-Case)分析建立数学模型等方法无法解决这些挑战。计算可持续性需要从更自然的角度,结合实际问题的不同层次,寻找更合适的解决途径,而不是依赖纯粹的数学抽象的或仿真的方法。

### 3.2 未来研究方向

计算可持续性是实现可持续发展的有效途径,这一研究领域的兴起使得可持续发展从政策驱动落实到技术驱动。让工业界、政府和研究机构对可持续发展从长期的概念架构跨越到技术实施,从想做什么贯彻到如何去做。大数据时代的来临更是让计算可持续性研究充满了无限可能,同时可持续发展实际问题的紧迫性也促使越来越多的研究人员投入到这一新兴研究领域中。

结合本文提出的大数据研究的4个要素,从技

术发展和实际应用的角度看,计算可持续性未来可能的研究方向包括以下几个方面:

(1) 多源数据融合技术的研究。从数据获取手段来看,可持续计算数据来自多种传感器、检测设备;从数据形式来看包括文本、数字、图像等,且这些数据可能具有不同粒度,分布在不同时间段。计算可持续性需要研究多种信息源给出的有用信息的综合、过滤、相关及合成算法。

如在建筑环评问题中,如何找出更切合实际的建筑社区还需要考虑异构社区发现问题,即建筑社区不仅是地理上接近,而且是属性、语义上相关,才对环境具有相似的影响。现有的社区发现算法,大都只利用潜在目标的结构或地理性质,即社会或地域关系,忽略了可用的语义内容的信息,即对象的特质信息。这就需要研究开发新的多源数据融合技术,可以集成从不同的来源获得的不同信息。此外,在气象监测、物种分布、灾难信息管理等计算可持续性问题中都存在多源数据融合问题。

(2) 大数据环境下复杂、动态优化模型的构建及求解。大数据技术还不成熟,面对海量、异构、动态变化的可持续发展大数据,传统的数据处理和分析技术无法应对,现有的计算方法和模型也难以扩展。因此,针对数据本身的复杂性、计算的复杂性及实际应用问题的复杂性,如何扩展现有的计算方法和模型,构建适合大数据分析的动态优化模型是未来的一个研究方向。

(3) 多学科进一步交叉融合技术的探索。计算可持续性研究涉及能源、生态、经济、环境等众多领域,汇集了计算机科学、信息科学、运筹学、应用数学、统计学等多学科技术的交叉融合。事实上,跨学科合作是计算可持续发展的一个重要方面。如探索生态信息学等可能出现的新的交叉学科,通过收集大量数据来学习分析预测模型,同时丰富和扩展计算机学科等的研究范围,具有重要意义。

(4) 计算可持续性数据分析平台的构建。构建高效的数据分析平台可以实现大数据存储、复杂算法执行及验证等任务。此外,统一、规范化的处理平台可以解决数据波动性,实现数据可视化等问题。

平台的构建在解决实际问题中具有重要意义,以我们开发的FIU-Miner(a Fast, Integrated, and User-friendly system to ease data analysis)为例<sup>[51]</sup>,该平台是一个快速集成的和用户友好的分布式数据分

析处理系统。FIU-Miner 允许用户快速配置复杂的数据分析任务,可以帮助用户方便地导入和整合不同的分析程序,并可以有效平衡异构环境中资源利用率和任务的执行。在该平台下已实现的高端制造业生产流程优化等复杂的大数据分析处理任务显著提高了生产效率和经济效益。

因此,构建大规模分布式可持续发展数据处理平台,是实现计算可持续性任务的重要技术支撑。

(5) 基于数据驱动的可替代环保材料的选择。在计算可持续性实际应用问题中,发现可替代材料是一个重要的研究内容,也是实现绿色节能设计的关键。目前,绿色产品设计主要依靠专家的知识经验和经验给出不同的设计方案。可替代材料的选择,需要在产品规划设计时结合环评,给出多种材料选择方案,并找出对环境影响较大的材料的替代产品,从而降低环境污染。大数据环境下,如何从数据分析中直接给出客观、可靠的多种选择,是一个很有实际意义的研究方向。

此外,随着计算可持续性研究的逐渐成熟及应用范围的进一步拓展,未来会有更多的极具吸引力的研究方向出现。计算可持续将为可持续发展提供更加强有力的解决工具,改善人类环境,促进社会的可持续发展。

## 4 结束语

将计算的思想引入可持续发展领域,为可持续性问题研究开辟了新的途径,这个发展过程也将可持续性研究从政策驱动的宏观层面,落实到了项目实施的微观层面。可持续发展的关键问题最终可以转化成计算和信息科学领域的决策和优化问题。计算可持续性这一新兴的研究领域,则为我们提供了社会、经济以及环境需求平衡发展的各种计算方法和工具。

大数据时代的来临为计算可持续性带来了机遇,同时也面临着大规模数据、复杂关系及可扩展方法等新的难题。如何发挥大数据的价值,利用大量易获取的生态数据,建立数据驱动的高效可行的模型和方法,解决能源、社会与经济等的可持续发展,成为计算机科学、信息科学、生态科学等众多领域共同关注的前沿性研究课题。同时,大数据环境下的计算可持续性研究,也将进一步丰富和拓展计算机等学科的发展,为可持续的未来提供重要的技术支撑。

## 参考文献:

- [1] GOMES C P. Computational sustainability: Computational methods for a sustainable environment, economy and society [J]. *The Bridge*, 2009, 39(4): 5-13.
- [2] FRENKEL K A. Computer science meets environmental science [J]. *Communications of the ACM*, 2009, 52(9): 23.
- [3] MURGANTE B, BORRUSO G, LAPUCCI A. Geocomputation, sustainability and environmental planning [M]. Berlin: Springer-Verlag, 2011.
- [4] ANDO A, CAMM J, POLASKY S, et al. Special distributions, land values, and efficient conservation [J]. *Science*, 1998, 279(5359): 2126-2128.
- [5] MOILANEN A, WILSON K, POSSINGHAM H, et al. Spatial conservation prioritization [M]. Oxford: Oxford University Press, 2009.
- [6] POLASKY S, NELSON E, CAMM J, et al. Where to put things? Spatial land management to sustain biodiversity and economic returns [J]. *Biological Conservation*, 2008, 141(6): 1505-1524.
- [7] ONAL H, BRIERS R. Designing a conservation reserve network with minimal fragmentation: A linear integer programming approach [J]. *Environmental Modeling and Assessment*, 2005, 10(3): 193-202.
- [8] WILLIAMS J C, REVELLE C S, LEVIN S A. Spatial attributes and reserve design models: A review [J]. *Environmental Modeling and Assessment*, 2005, 10(3): 163-181.
- [9] CONRAD J, GOMES C, VAN HOEVE W J, et al. Connections in networks: Hardness of feasibility versus optimality [C]// *Proceedings of the 4th International Conference on the Integration of AI and OR Techniques Constraint Programming*, 2007: 16-28.
- [10] DIETTERICH T. Machine learning in ecosystem informatics and sustainability [C]// *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2009: 8-13.
- [11] PHILLIPS S J, DUDÍK M, SCHAPIRE R E. A maximum entropy approach to species distribution modeling [C]// *Proceedings of the 21st International Conference on Machine Learning*, 2004: 83.
- [12] BARRETT C B, LITTLE P, CARTER M. Understanding and reducing persistent poverty in Africa [M]. London: Routledge, 2013.
- [13] 李国杰,程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考 [J]. *中国科学院院刊*, 2012, 27(6): 647-657.  
LI Guojie, CHENG Xueqi. Research status and scientific thinking of big data [J]. *Bulletin of Chinese Academy of Sciences*, 2012, 27(6): 647-657. (in Chinese)
- [14] Wikipedia. Big data [EB/OL]. [2015-07-18]. <https://>

- en.wikipedia.org/wiki/Big\_data.
- [15] Gartner. Big data [EB/OL]. [2015-07-20]. <https://www.gartner.com/it-glossary/big-data>.
- [16] IDC. Big data & analytics [EB/OL]. [2015-08-01]. <https://www.idc.com/prodserv/4Pillars/bigdata>.
- [17] DIETTERICH T G. Machine learning in ecosystem informatics and sustainability [C]//Proceedings of the 21st International Joint Conference on Artificial Intelligence. 2009: 8 – 13.
- [18] 李涛. 数据挖掘的应用与实践 [M]. 厦门: 厦门大学出版社 2013.
- LI Tao. Data mining where theory meets practice [M]. Xiamen: Xiamen University Press 2013. (in Chinese)
- [19] KDD. Knowledge discovery from sensor data [EB/OL]. [2015-06-18]. <https://www.onrl.gov/sci/knowledgediscovery/SensorKDD-2009>.
- [20] VIPIN K. Expeditions in computing: Understanding climate change—A data driven approach [EB/OL]. [2015-07-13]. <https://climatechange.cs.umn.edu/>.
- [21] MODIS. Data [EB/OL]. [2015-07-16]. <https://modis.gsfc.nasa.gov/data/>.
- [22] NEX. Datasets [EB/OL]. [2015-07-19]. <https://c3.nasa.gov/nex/resources/>.
- [23] IRI/LDEO. Climate data library [EB/OL]. [2015-08-15]. <https://iridl.ldeo.columbia.edu/index.html>.
- [24] KUMAR V ,STEINBACH M ,TAN P. Mining scientific data: Discovery of patterns in the global climate system [C]//Proceedings of the Annual Meeting of the American Statistical Association. 2001: 1 – 10.
- [25] TAN P N ,STEINBACH M ,KUMAR V. Introduction to data mining [M]. Boston: Addison-Wesley Longman Publishing 2005.
- [26] STEINBACH M ,TAN P ,KUMAR V. Discovery of climate indices using clustering [C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003: 446 – 455.
- [27] ERTÖZ L ,STEINBACH M ,KUMAR V. Finding clusters of different sizes ,shapes and densities in noisy ,high dimensional data [C] // SIAM International Conference on Data Mining( SDM ). 2003.
- [28] STEINHAEUSER K ,CHAWLA N V ,GANGULY A R. An exploration of climate data using complex networks [C]//ACM SIGKDD Explorations Newsletter. 2010 ,12: 25.
- [29] STEINHAEUSER K ,CHAWLA N V ,GANGULY A R. Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science [J]. Statistical Analysis and Data Mining ,2011 ,4 ( 5 ) : 497 – 511.
- [30] BHADURI K ,DAS K ,VOTAVA P. Distributed anomaly detection using satellite data from multiple modalities [C]//NASA Conference on Intelligent Data Understanding. 2010: 109 – 123.
- [31] ANDERSON R P ,DUDÍK M ,FERRIER S ,et al. Novel methods improve prediction of species' distributions from occurrence data [J]. Ecography 2006 ,29( 2 ) : 129 – 151.
- [32] HASTIE T J ,TIBSHIRANI R J ,FRIEDMAN J J H. The elements of statistical learning [M]. Berlin: Springer-Verlag 2009.
- [33] HUTCHINSON R ,LIU L ,DIETTERICH T. Incorporating boosted regression trees into ecological latent variable models [C]//25th AAAI Conference on Artificial Intelligence. 2011: 1343 – 1348.
- [34] EKASINGH B ,NGAMSOMSUK K ,LETCHER R ,et al. A data mining approach to simulating farmers' crop choices for integrated water resources management [J]. Journal of Environmental Management 2005 ,77( 4 ) : 315 – 325.
- [35] BORIAH S ,KUMAR V ,STEINBACH M. Land cover change detection: A case study [C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008: 857 – 865.
- [36] MAZZONI D ,LOGAN J ,DINER D ,et al. A data-mining approach to associating MISR smoke plume heights with MODIS fire measurements [J]. Remote Sensing of Environment 2007 ,107( 1/2 ) : 138 – 148.
- [37] United States Department of Energy. "Grid 2030": A national vision for electricity's second 100 years [EB/OL]. [2015-08-01]. <https://energy.gov/oe/office-electricity-delivery-and-energy-reliability>.
- [38] MASSOUD AMIN S. Electrical grid gets less reliable [J]. IEEE Spectrum ,2011 ,48( 1 ) : 80.
- [39] PASSONNEAU R ,RUDIN C ,RADEVA A ,et al. Treatment effect of repairs to an electrical grid: Leveraging a machine learned model of structure vulnerability [C] // Proceedings of the 17th Annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2011.
- [40] RUDIN C ,PASSONNEAU R J ,RADEVA A ,et al. A process for predicting manhole events in Manhattan [J]. Machine Learning 2010 ,80( 1 ) : 1 – 31.
- [41] RUDIN C ,WALTZ D ,ANDERSON R N ,et al. Machine learning for the New York City power grid [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence 2012 ,34( 2 ) : 328 – 345.
- [42] SHIVASWAMY P K ,CHU W ,JANSCHKE M. A support vector approach to censored targets [C] // the 7th IEEE International Conference on Data Mining. 2007: 655 – 660.
- [43] United States Department of Commerce. NCEP/NCAR reanalysis monthly means and other derived variables [EB/OL]. [2015-07-27]. <https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.derived.surfaceflux.html>.

- [44] ZHOU Q ,ZHOU H ,ZHU Y ,et al. Data-driven solutions for building environmental impact assessment [C]//IEEE International Conference on Semantic Computing(ICSC) . 2015:316 - 319.
- [45] KRAUSE A ,GOLOVIN D ,CONVERSE S. Sequential decision making in computational sustainability via adaptive submodularity [J]. AI Magazine 2014 ,35( 2) : 8 - 18.
- [46] FINK D ,HOCHACHKA W M. Documenting stewardship responsibilities across the annual cycle for birds on US public lands [J]. Ecological Applications ,2015 ,25( 1) : 39 - 51.
- [47] FARNSWORTH A ,SHELDON D ,GEEVARGHESE J , et al. Reconstructing velocities of migrating birds from weather Radar: A case study in computational sustainability [J]. AI Magazine 2014 ,35( 2) : 35 - 48.
- [48] POWELL W B. Energy and uncertainty: Models and algorithms for complex energy systems [J]. AI Magazine , 2014 ,35( 3) : 8 - 21.
- [49] QUINN J A ,FRIAS-MARTINEZ V ,SUBRAMANIAN L. Computational sustainability and artificial intelligence in the developing world [J]. AI Magazine ,2014 ,35( 3) : 36 - 47.
- [50] MILANO M ,O' SULLIVAN B ,GAVANELLI M. Sustainable policy making: A strategic challenge for artificial intelligence [J]. AI Magazine 2014 ,35( 3) : 22 - 35.
- [51] ZENG C ,JIANG Y ,ZHENG L ,et al. Fiu-miner: A fast , integrated and user-friendly system for data mining in distributed environment [C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 1506 - 1509.

## 作者简介:



周绮凤(1976 -),女,吉林松原人。厦门大学自动化系副教授,博士。2007年12月获厦门大学控制理论与控制工程专业博士学位。2014至2015年在美国佛罗里达国际大学(Florida International University, FIU)访学。研究方向为机器学习、数据挖掘及其在可持续发展等领域的应用。



李涛(1975 -),男,四川绵阳人。南京邮电大学计算机学院院长,教授,博士,博士生导师。1995年7月获得福州大学计算机应用专业学士学位,1998年7月获得中国科学院计算机应用专业硕士学位,2000年5月获得 Oklahoma State University 理论数学专业硕士学位,2004年7月获得 University of Rochester 计算机专业博士学位。先后任 Florida International University 计算机学院助理教授、副教授(终身教授)、正教授(Full Professor)、研究生主管(Graduate Program Director)、计算与信息学院数据挖掘实验室主任。自2015年1月起任南京邮电大学计算机学院、软件学院院长。在国际会议及期刊上已发表250多篇文章,出版专著2本,参与编写专著6本。作为项目负责人,承担了美国自然科学基金项目、美国军方实验室科研项目等多项科研项目。2006年获得美国国家自然科学基金委颁发的杰出青年教授奖,2009年获得佛罗里达国际大学最高学术研究奖,2010年获得IBM大规模数据分析创新奖,2014年获得佛罗里达国际大学工程学院杰出导师奖。目前的主要研究方向包括数据挖掘、机器学习、信息检索及生物信息学等。