软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(11):3010−3025 [doi: 10.13328/j.cnki.jos.004900]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大数据分析中的计算智能研究现状与展望*

郭 平[1,2], 王 可[1], 罗阿理[3], 薛明志[4]

[1](北京理工大学 计算机学院,北京 100081)

[2](北京师范大学 图形图像与模式识别实验室,北京 100875)

[3](中国科学院 国家天文台 光学天文重点实验室,北京 100012)

[4](商丘师范学院 数学与信息科学学院,河南 商丘 476000)

通讯作者: 郭平, E-mail: pguo@bit.edu.cn, http://igpr.bnu.edu.cn

摘 要: 随着产业界和科学界数据量的爆炸式增长,大数据技术和应用吸引了众多的关注.如何分析大数据,充分挖掘大数据的潜在价值,成为需要深入探讨的科学问题.计算智能是科学研究和工程实践中解决复杂问题的有效手段,是人工智能和信息科学的重要研究方向,应用计算智能方法进行大数据分析具有巨大的潜力.对大数据分析中的计算智能方法进行综述,结合大数据的特征,讨论了大数据分析中计算智能研究存在的问题和进一步的研究方向,阐述了数据源共享问题,并建议利用以天文学为代表的数据密集型基础科研领域的数据开展大数据分析研究.

关键词: 大数据;计算智能;大数据分析;天文大数据

中图法分类号: TP18

中文引用格式: 郭平,王可,罗阿理,薛明志.大数据分析中的计算智能研究现状与展望.软件学报,2015,26(11):3010−3025. http://www.jos.org.cn/1000-9825/4900.htm

英文引用格式: Guo P, Wang K, Luo AL, Xue MZ. Computational intelligence for big data analysis: Current status and future prospect. Ruan Jian Xue Bao/Journal of Software, 2015,26(11):3010−3025 (in Chinese). http://www.jos.org.cn/1000-9825/4900.htm

Computational Intelligence for Big Data Analysis: Current Status and Future Prospect

GUO Ping[1,2], WANG Ke[1], LUO A-Li[3], XUE Ming-Zhi[4]

[1](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[2](Image Processing and Pattern Recognition Laboratory, Beijing Normal University, Beijing 100875, China)

[3](Key Laboratory of Optical Astronomy, National Astronomical Observatories, The Chinese Academy of Sciences, Beijing 100012, China)

[4](School of Mathematics and Information Sciences, Shangqiu Normal University, Shangqiu 476000, China)

Abstract: Big data and its real-world applications have attracted a lot of attention with the explosive growth of data volumes not only in the academic but also in industrial. Big data analysis aimed at mining the potential value of big data has become a popular research topic. Computational intelligence (CI) which is an important research direction of artificial intelligence and information science has been shown to be promising to solve complex problems in scientific research and engineering. CI techniques are expected to provide powerful tools for addressing challenges in big data analytics. This paper surveys the related CI techniques, analyzes the grand challenges brought forth by big data from big data analysis perspectives, and discusses the possible research directions in the future of the big data era. Further, it proposes to conduct the research of big data analysis on scientific big data such as astronomy big data.

Key words: big data; computational intelligence; big data analysis; astronomy big data

* 基金项目: 国家自然科学基金(61375045); 国家自然科学基金委员会-中国科学院天文联合基金(U1531242); 北京市自然科学基金(4142030); 河南省基础与前沿技术研究计划(122300410275)

收稿时间: 2015-05-29; 修改时间: 2015-07-14, 2015-08-11; 定稿时间: 2015-08-26

随着互联网、移动智能终端、物联网等信息与通信技术的迅猛发展,以及计算机存储和计算能力的不断提升,各种数据的爆炸式增长和持续获取成为可能,大数据时代悄然而至.2012 年 3 月,美国政府发布“大数据研究和发展计划”,这标志着美国已经将大数据上升到国家信息发展战略的高度.2014 年 8 月,科技部基础研究司组织召开“大数据科学问题”研讨会,围绕大数据研究布局、中国大数据发展战略、国外大数据研究框架与重点、大数据研究关键科学问题等展开研讨.2015 年初,国务院下发《关于促进云计算创新发展培育信息产业新业态的意见》,明确指出:将着力突破大数据挖掘分析等关键技术,推动大数据挖掘、分析、应用和服务,开展基于云计算的大数据应用示范,支持政府机构和企业创新大数据服务模式.大数据中隐藏着巨大的经济、科学、社会及军事价值,有关大数据的讨论逐渐成为产业界、学术界、政府部门等各界共同关注的热点.大数据正渗透到社会生活的各个方面,将导致人们的工作和生活方式发生巨大变革.

大数据的潜在价值是真实而巨大的,为了充分挖掘大数据的价值,必须解决一系列技术问题,这些问题包括数据采集、信息抽取和清理、数据集成、数据分析以及解释和部署[1].这些问题涉及数据获取、数据存储和管理、数据分析、数据可视化、应用服务、信息共享、数据安全和隐私保护、大规模并行计算、流计算、云计算等多层面的信息技术,需要计算机软、硬件的综合解决方案.

计算智能是人工智能发展的新阶段,是受到大自然智慧和人类智慧的启发而设计出的一类解决复杂问题方法的统称[2,3].与传统的人工智能相比,计算智能的最大特点是不需要建立问题本身的精确(数学或逻辑)模型,不依赖于知识表示,而是在观测数据上直接对输入信息进行处理.这一特点非常适合于解决大数据分析中那些由于难以建立有效的形式化模型而用传统技术难以解决,甚至无法解决的问题.近年来,计算智能理论与技术发展迅速,在图像处理、模式识别、知识获取、经济管理、生物医学、智能控制等许多领域都得到了广泛应用,取得了一系列令人鼓舞的研究成果.同时,大数据也给计算智能发展带来新的挑战与机遇.

本文结合大数据的特征,针对大数据的分析方法,从人工神经网络、模糊系统、演化计算和群体智能这 3 个方面梳理大数据环境下计算智能的相关研究,总结大数据分析中计算智能面临的主要问题;在此基础上,给出进一步深入研究的方向,并阐述了数据源共享与数据密集型科学问题.

1 大数据概述

大数据其实并不是近来出现的新事物,早在 1971 年,智利政府就部署实施了一个名为“Project Cybersyn”的决策支持系统(http://en.wikipedia.org/wiki/Project_Cybersyn).该系统将数据从各地工厂传输到位于圣地亚哥的运营中心,通过分析经济领域中的统计数据,监督生产状况,预估宏观经济走势,发现经济生活中的不稳定因素.这种模式类似于当下的大数据分析,只不过当时的计算机软、硬件还远远不能满足实际需求.1996 年,SGI 公司的首席科学家 John Mashey 在其题为《Big Data…and the Next Wave of InfraStress》的演讲中首次提出了 Big Data 这一概念.时至今日,大数据的定义有了多种不同的描述.直观地,大数据是指数据量达到 PB 级甚至 EB 级的大规模数据.“大”是大数据最直观、最重要的特征,但远非全部.通常认为,大数据是指无法在可容忍的时间内用传统的方法和软、硬件平台对其进行感知、获取、管理、处理和可视化的数据集合;大数据是指需要新的处理模式才能从中获取更强的决策力、洞察力和流程优化能力的海量、高速和(或)多样化的信息资产[4,5].如今,所谓的大数据已经发展为一个较为宽泛的概念,是包括数据及其采集、处理、分析、解释等在内的一系列相关技术、方法、手段的统称.Mayer-Schönberger 等人认为:大数据是人们在大规模数据的基础上可以做到的事情,而这些事情在小规模数据上是无法完成的;大数据是人们获得新的认知、创造新价值的源泉;大数据还是改变市场、组织机构以及政府与公民关系的方法[6].如今,大数据已经不仅仅指数据本身,也不仅仅是一种工具,而是一种战略、世界观和文化,是要大力推广和树立的“数据文化”(http://www.ccf.org.cn/sites/ccf/ccfziliao.jsp?contentId=2774793649105).

META 集团的分析师 Doug Laney 最初将大数据的特征总结为量度(volume)、异度(variety)和速度(velocity)这 3 个维度,即后来流行的 3V[7].IBM 公司在其与牛津大学共同发布的白皮书《Analytics: The real-world use of big data》中又补充了精度(veracity),形成了关于大数据特征的 4V 描述(http://www-03.ibm.com/systems/hu/

resources/the_real_word_use_of_big_data.pdf).后来,从不同的应用视角和需求出发,对 4V 中的第 4 个 V 又出现了价值(value)、多变性(variability)、粘度(viscosity)、邻近性(vicinity)、模糊性(vague)等多种不同的解释,形成了 3+*x*V 的描述.不同领域的大数据应用各有其自身的特点,但量度(volume)、异度(variety)和速度(velocity)一般被视作大数据应具备的 3 个基本维度.这 3 个基本维度贯穿于大数据生命周期中的各个阶段,对信息技术提出了巨大的挑战.

2 大数据分析中的计算智能方法

计算智能一般被认为是在人工神经网络、模糊系统、演化计算这 3 个主要分支发展相对成熟的基础上,通过相互之间的有机融合而形成的新的科学方法[2].计算智能方法的特点决定了其在大数据分析中具有巨大的应用潜力:

- 首先,大数据混杂多样(variety)、多变(variability)的特点决定了模型驱动的方法存在本质上的局限性,因为面对海量、复杂的大数据,往往难以根据先验知识建立精确的模型.演化计算、群体智能等计算智能方法不依赖于知识,不需要对问题进行精确建模而在数据上直接进行分析和处理的特点非常适于进行大数据分析.大数据分析往往伴随着环境的变化,这源于系统本身以及用户需求、目标等主客观因素的变化.传统方法往往难以适应环境的变化,导致算法失效.而以遗传算法为代表的演化算法在代与代之间维持潜在解的种群,并能够根据环境不断优化种群的适应度,因此更容易适应环境的变化.
- 其次,精度(veracity)是大数据的一个重要维度,对不确定性的处理和管理的需求源于数据采集手段、系统状态变化和自然环境等随机因素的干扰,同时也源于大数据固有的不确定性.因此,对不确定和概率数据的挖掘已成为当前大数据分析中的重要问题[8].模糊逻辑、粗糙集等计算智能方法能够有效处理数据中的不完全、不精确或者不确定性,增强了分析结果的客观性和可解释性.
- 最后,大数据的规模和复杂性意味着大数据分析需要巨大的计算时空开销,可能无法在可接受的时间内得到精确解.计算智能方法具有启发式特征,通过模拟人类和其他生物体的智慧求解问题,具有高度的自组织、自适应性、泛化和抽象的能力,可以快速近似求解一些 NP 难的问题,比如组合优化问题,为大规模复杂问题的求解提供了有效手段.

下面结合大数据的特点,从人工神经网络、模糊系统、演化计算和群体智能等几个方面,总结和梳理大数据环境下对计算智能方法的相关研究.

2.1 人工神经网络

人工神经网络是一种模仿动物神经系统行为特征进行分布式并行信息处理的数学模型,具有高度的非线性映射能力、良好的容错性、自适应能力以及分布存储等优良特性,是一类重要的计算智能方法.神经网络不需要具备数据集概率分布的任何先验知识,与统计学方法相比,其限制条件更少.

对于多数大数据应用而言,例如设备传感器、社交网络、搜索引擎以及时域天文学等,数据是持续产生并不断变化的,因此无法像批量学习算法那样从历史数据中构建无偏训练集.此外,数据的规模和产生速度也使得数据无法一次性载入内存.解决这一问题需要发展在线学习算法,每次仅根据一个样本更新目标函数.感知器(perceptron)是一种经典的在线学习模型,也是人工神经网络中的一种典型结构.对于任意一个训练样本,感知器根据预测结果的正确与否来更新连接权重:如果预测结果正确,则权重保持不变;否则,根据输入样本的特征向量及正确的标记更新连接权重.理论上,这种更新策略产生的分类错误率与标准化后的全部训练样本到最优平面最短距离的平方成反比.目前,基于感知器的在线学习算法主要包括投票(voted)感知器算法[9,10]、均值(averaged)感知器算法[11]、权重多数(weight majority)感知器算法[12]、被动主动(passive aggressive,简称 PA)感知器算法[13−15]、置信度权重(confidence-weighted)感知器算法[16−18]、核(kernel)感知器算法[19].

在大数据环境下,人们生产和采集数据的能力日益增强,手段愈发丰富,这将导致数据在规模增大的同时,属性(维度)也随之增长.这样的高维数据会带来两个问题:首先,对于特定的应用而言,一般不需要关注数据的全部属性(维度),原始数据中包含的大量冗余信息和噪声反而会隐藏其中的有价值信息;其次,高维数据严重影响

算法的性能,一些在低维特征空间中有效的算法,在超过 30 维的特征空间中将出现性能退化[20].解决这一问题需要对原始数据进行约简,实现化繁为简、化难为易,滤除冗余特征.神经网络是进行数据约简的有效方法之一.一种有代表性的基于神经网络的数据约简方法是由 Castellano 和 Fanelli[21]提出的,通过度量输入特征与输出结果的相关程度,发现并过滤掉冗余的、次要的特征.换句话说,就是在精度允许的范围内,对已经训练好的网络结构进行修剪,删掉一些无关的输入节点,得到一个约简的网络结构,从而得到对应的特征子集.自组织映射(self-organization mapping,简称 SOM)是一种竞争学习神经网络,其具有拓扑保形的特性,最终输出的模式空间能够反映原始模式空间的分布,因此也用于把高维模式映射到低维模式,从而实现数据的约简[22].但 SOM 需要在全部映射中寻找最佳映射,计算代价大,尤其当数据规模较大时,计算效率较低.

为此,Sagheer 等人[23]对 SOM 的计算效率进行了优化,提出了一种快速 SOM,其主要思想是:原始数据的低阶映射或主成分(principal component)往往能够反映出数据分布的显著规律,如果能够用这样的子空间代替原始特征空间,SOM 就可以在特征子空间中寻找最优映射,从而减少计算的代价,以便于将 SOM 推广到大数据集中.Hinton 等人[24]提出了自编码器(autoencoder)的深度网络,通过中间层来重构高维输入信号,训练一个多隐层的神经网络来学习对数据更本质的刻画,将高维信号转化为低维信号,取得了明显优于传统方法的约简效果.Le 等人[25]提出了一种直接从海量未标记数据中抽取高层特征的无监督学习框架,构建了用于特定类别图像高层特征识别的大规模神经网络.该网络在大规模集群上进行训练,采用并行的策略和异步随机梯度下降(asynchronous SGD)的优化方法,其图像识别的准确性有了大幅度提升.

Hinton 等人[24]的论文引发了深度学习的研究热潮.深度学习[26]是受认知神经科学中人类视觉系统分层信息处理机制的启发而提出的一类多层神经网络学习算法,深度学习旨在建立一种类似于人脑信息处理机制的多层神经网络,通过逐层组合低层特征来获得更抽象的高层特征表达,以发现复杂数据内在的分布式特征表示.与人工构造特征的方式相比,深度学习直接从大数据中学习特征,能够更深刻地刻画出海量数据中蕴藏的丰富信息.虽然深度学习的思想并非近来才出现,但是计算机计算能力的提升使得训练大规模深度神经网络成为可能;同时,数据规模的增长有利于复杂的深度神经网络规避过拟合的风险[27],因此近几年来,深度学习以其强大的学习能力在图像识别[28]、语音识别[29]、自然语言理解[30]等诸多应用领域取得了令人振奋的成果.通过深度神经网络来构建大数据分析模型,能够深刻揭示海量数据中丰富而复杂的信息,能够对未来做出更精准的预测.虽然深度学习目前仍然面临理论研究、建模、工程实现等方面的问题[31],但其无疑是大数据智能分析的有效解决方案之一.

2.2 模糊系统

在大数据应用中,数据受到采集设备的精度、系统状态变化的随机性和非线性、自然环境等不可控因素的干扰,导致获得的数据普遍存在模糊性.除了采集过程中引入的模糊性之外,实际应用中的数据往往还具有固有的模糊性,例如在电商网站、服务点评类网站、社交网络中,用户根据自己的主观感受表达倾向、发表评论,这些信息很难简单地以“好”、“坏”、“喜欢”、“不喜欢”的二值逻辑进行描述,通常要考虑其中蕴含的不完全、不精确或者不确定性,进而用语言进行更为详细的模糊概念的刻画.模糊系统研究的是一种模糊性现象,这种模糊性是由于事物之间差异的中间过渡性引起的划分上的不确定性,它弥合了二值逻辑中“非此即彼”的精确性与现实世界之间的鸿沟,使得概念的外延具有一种不分明性,增强了推理结果的可解释性,是一种已经得到广泛应用的计算智能方法,对于定性或以语言变量描述和分析大数据具有巨大的应用潜力和实用价值.

模糊聚类作为一种非监督学习方法,可用于发现数据中隐含的未知模式.在大数据环境下,模糊聚类算法主要面临可扩展性的问题,即,算法的在大数据集上的时空效率及准度.和许多其他问题一样,提高聚类算法可扩展性的主要策略可归纳为采样[32]、在线处理[33]和分布式并行计算[34].Havens 等人[32]研究了大数据环境下的模糊 c 均值(fuzzy c means,简称 FCM)聚类算法,提出了 3 种新的 FCM 算法,包括随机采样并扩展的 FCM 算法(a simple random sampling plus extension FCM,简称 rseFCM)、单次遍历核 FCM 算法(single-pass kernel FCM,简称 spkFCM)和在线核 FCM 算法(online kernel FCM,简称 okFCM),通过实验分析了这些算法在时空复杂度、速度、准确率等方面的性能,并与同类算法和小数据集上的 FCM 算法进行了对比.研究还归纳总结了不同的 FCM 算

法的适用情景,并给出了算法选择上的建议.对于基于核的模糊聚类算法,如何选择核、如何确定其适用场景,也是需要深入研究的问题.

目前的在线模糊聚类算法大都是采用批量在线的方式,不能真正满足大数据流中逐个处理数据的需求.Wang等人[35]提出了一种基于随机梯度下降的模糊聚类算法(SGFC),虽然实现了根据单个数据样本更新簇中心和隶属度矩阵,但这种方法容易受到簇中心初始化和噪声的影响,因此又进行了批量梯度和重复遍历的折中.此外,聚类算法的核心目标是发现数据中的未知规律,而仅以已知的标记数据来评价算法的有效性是不够全面的.如何评估聚类算法在大数据环境下的有效性,也是一个开放问题.

Yang 等人[36]提出了一种核 FCM 聚类与支持向量机相结合的模糊分类算法.该算法首先利用核 FCM 算法分别对训练集的正负样本子集进行聚类;然后,分别在正负样本子集中选择彼此距离最远的两个簇组成新的训练集;最后,在此训练集上训练支持向量机并得到一个模糊分类器.该算法通过对训练集进行预处理,有效地降低了噪声和离群点对支持向量机性能的影响,这对于大数据中普遍存在的缺值、错误以及不一致等现象具有积极的意义.然而,该算法的计算复杂度较高,计算代价随着问题规模呈指数增长,尚不能直接推广到大规模数据集中.针对这一问题,是否可以将该分类算法与大规模数据集上的 FCM 聚类算法以及 SVM[32,37−40]相结合,采用在线处理、随机采样的方式,发展出一种能够有效地应对噪声和离群点的高可扩展性的模糊分类器,是一个值得继续研究的问题.

从大数据中发现相关性关系具有重要的研究意义和应用价值.已发现的相关性规则又可以作为分类规则对未知数据进行预估,起到增强决策力和洞察力的作用.在现实世界中,需要使用模糊逻辑来软化相关性规则的边界和规则匹配的条件,进而形成一类问题,即,基于模糊规则的分类问题(fuzzy rule-based classification).传统的模糊相关性规则挖掘算法是针对小数据集设计的,无法直接应用于大数据中.为此,Mangalampalli 等人[41]提出了一种名为 FAR-HD 的快速模糊相关性规则挖掘算法.该算法在高维大规模数据集上取得了明显优于之前算法的速度.在基于模糊规则的分类问题中,随着数据规模以及复杂性的增长,分类规则集的搜索空间呈指数增长,导致算法出现可扩展性的问题.Alcala-Fdez 等人[42]提出了一种针对高维数据集的基于模糊规则的分类方法.该方法包括 3 个步骤:首先,通过搜索树搜索模糊相关性规则;然后,对得到的规则集采用样本权重的方案进行压缩;最后,利用遗传算法进行规则筛选和调优,得到进一步约简后的规则子集.这种方法有效压缩了模糊分类规则集的规模,而且最终的规则前件包含更少的变量,降低了分类过程中的计算复杂度,提高了模型的可扩展性.但在该研究所使用的所有实验数据集中,最大变量个数和样本总数分别为 90 和 19 020,这样规模的数据集与实际应用中的大数据还存在不小的差距,因此,该方法是否足以应对大数据有待进一步的考证.

2.3 演化计算和群体智能

以遗传算法(genetic algorithm,简称 GA)为代表的演化计算和以粒子群优化(particle swarm optimization,简称 PSO)、蚁群优化(ant colony optimization,简称 ACO)等为代表的群体智能算法是解决复杂优化问题的常用方法.这类智能算法的主要意义在于:一方面,可以快速近似求解一些难解的问题,比如 NP 难问题;另一方面,还可用于约简问题的规模,从而解决那些由于数据量太大而不易解决的问题.

遗传算法具有对噪声不敏感、不需要先验知识等优势以及隐含的并行性[43,44],已经广泛用于解决复杂优化问题.另外,遗传算法还是进行数据约简的有效手段[45−49].在这类问题中,如果把特征组合看作一个染色体对其进行编码,并引入可以反映特征组合质量的适应度函数,就能通过选择、交叉和变异的遗传算子,高效地找出特征子集.此外,遗传算法还被用于确定复杂系统输入与输出之间的映射,即,所谓的基于遗传算法的机器学习(genetics-based machine learning,简称 GBML).

Aggarwal 等人[49]将遗传算法应用于高维空间中离群点的检测.在该方法中,离群点被视为原始特征空间的某个低维投影中,密度极低的局部区域内所包含的数据点.为了挖掘这些低维空间的低密度区域(异常模式),必须过滤掉冗余特征,找出能够凸显异常模式的特征子集.该方法首先对数据按照属性值进行网格化,将高维数据空间划分为等大小的“数据立方”;然后,以每个“数据立方”中的实际点数和期望点数的差来刻画稀疏程度(适应度函数);最后,迭代使用遗传算子,直至得到满意解.该方法通过进化计算完成数据的约简,消除了“维灾”的影响,

但其假设各维度之间相互独立并且每个维度都符合均匀分布,而实际数据很难满足这些的假设,因此,应当进一步研究维度之间的相关性和概率分布对算法性能的影响.

为了将 GBML 应用于大数据环境中,必须使其能够在可接受的时间范围内完成对巨量数据的处理.为此,Bacardit 等人[44]对大规模数据集中基于遗传算法的机器学习的改进策略进行了归纳和总结,将其分为 4 大类,即,软件的方法、硬件加速技术、并行计算模型以及以 Apache Hadoop(http://hadoop.apache.org/)为代表的数据密集型计算模型.其中,软件的方法不涉及额外的硬件资源,是最为廉价和灵活的一类方法,包括窗口机制(windowing mechanism)、利用数据固有规律(exploiting regularity)、混合方法(hybrid method)和适应度函数替代(fitness surrogate)这 4 大类方法.

粒子群优化算法是一种模拟鸟群、鱼群等生物群社会行为的群体智能算法,其不易受问题的规模和非线性的影响,是一种应用广泛的高效优化技术[50].与遗传算法相比,粒子群优化的原理更简单,算法实现相对容易,收敛速度快,适于求解大数据环境下的复杂优化问题[51].然而在大数据应用中,数据往往具有高维的特征,而随着数据维度的增加,粒子群优化算法的性能会急剧退化,难以直接应用于大数据应用中.采用分而治之的策略是处理高维数据集上粒子群优化问题的直接思路.Li 等人[52]基于分治的思想,提出了一种可用于大规模高维数据空间优化问题的协同演化 PSO 算法.该算法使用动态的随机分组策略,将高维解空间划分为大小可变的低维子空间.而在此前,Yang 等人[53]已经从理论上证明了随机分组的策略可以增加相关变量被划分至同一子分量(subcomponent)的概率,并已经将其应用于大规模协同进化计算中.这一随机分组策略在高维优化问题中具有明显的优势.此外,该算法分别在个体最优和环拓扑邻域最优位置的周围,以柯西或高斯分布随机更新粒子的位置.这种位置更新策略提高了算法的搜索能力.在后续的研究中,Omidvar 等人[54]研究了如何使分组更加智能而不是简单地采用随机策略,提出了一种名为差分分组(differential grouping)的自动分组策略,使得不同分组间的变量之间的相互依赖度最小化.上述分治的策略关键在于如何“分”及如何“合”.虽然这类方法为解决高维粒子群优化问题提供了直接思路,但在面对不可分的问题时,这种策略仍然束手无策.

蚁群优化(ant colony optimization,简称 ACO)[55−57]和粒子群优化[58,59]等群智能算法还为大规模数据的约简提供了有效手段.例如,Aghdam 等人[56]将 ACO 应用于文本特征的约简,将特征表示为图中的节点,使用分类器的分类性能和特征子集的大小作为启发式信息来更新信息素,原始的特征约简就转化为如何让蚁群在图中找到满足优化终止条件的最短路径(特征子集)的问题;Wang 等人[58]提出了一种粗糙集与粒子群优化相结合的方法,通过粒子群优化求解粗糙集最小约简(reduct)的 NP 难问题.从相关研究来看,将多种方法结合起来的混合方法往往可以取得比单一技术更好的性能和约简效果[48,60].

增强决策力和流程优化能力是大数据分析的主要目的之一,而科学研究与工程实践中许多决策问题本质上是最优化问题.在大数据环境中,优化问题不可避免地涉及到更多的决策变量和优化目标,形成更为复杂的多目标优化问题.这些问题中,不同的优化目标往往相互制约,相互冲突,可能某一个解对于其中一个特定的优化目标来说是较好的,但对于其他优化目标而言却是很差的,因此,多目标优化问题实际上就是要进行协调和折中以寻找一个解的集合,即,Pareto 最优解集.目前,绝大部分多目标优化方法通过逼近 Pareto 前沿(Parero front)进行优化.这类方法大都采用基于演化计算的启发式(meta-heuristic)搜索技术[61].与传统方法相比,这类算法不需要目标函数的梯度信息,不受目标函数的形式和性质(如连续性、凹凸性等)的限制,可以用于优化任意形式的目标函数,优化过程独立于具体的应用领域,不易受到具体领域的约束,因此,其具有比传统优化方法更广泛的应用范围.多目标优化已成为演化计算的一个重要研究方向.关于这方面的研究已有大量综述文章和专著[62−67].此外,PSO[51]、ACO[68]等计算智能方法也已被应用于多目标优化领域.

2.4 小 结

计算智能的研究在国内外得到了广泛的关注,其理论和方法也处于不断发展和完善的过程中.经过多年的发展,针对大规模数据的计算智能研究已取得了一定的发展,并逐渐在实际应用中发挥作用.表 1 列举了大数据分析中的主要计算智能方法.我们从大数据的特征出发,分析了这些特征给数据分析带来的具体问题,包括在线学习、数据约简、深度学习、可扩展算法、鲁棒算法、协同演化、多目标优化这 7 个问题,归纳了解决这些问

题所涉及的计算智能方法及其所属的类别,并列举了一些实例.这些实例包含了对数据规模、算法运行环境的说明,其中的“数据规模”根据具体问题的不同,采用数据量(样本个数)和数据维度(变量个数)两种不同的度量标准.

Table 1 Summary of computational intelligence methods in big data analytics

表 1 大数据分析中的计算智能方法

大数据特征及挑战	问题	方法	类别	实例	数据规模	运行环境
数据持续产生,不断变化,无法用传统的批处理方式(variability, velocity);	在线学习	感知器	人工神经网络	投票感知器[9]	70 000	单核 SGI MIPS R10000 CPU (194MHZ)
				均值感知器[11]	240 000	–
				权重多数感知器[12]	–	–
				被动主动感知器[14]	252 800 275	–
				置信度权重感知器[17]	581 012	–
				核感知器[19]	199 523	–
数据维度高,包含冗余属性(volume)	数据约简	神经网络	人工神经网络	基于网络结构约简的特征选择[21]	290	–
				FSOM[23]	41 386	–
		混合方法	人工神经网络、演化计算	SAGA[48]	10 000 维	Intel Pentium D CPU (3.40GHz);2GB RAM
		遗传算法、随机投影	演化计算	基于遗传算法的异常检测[49]	100 000	233MHZ CPU; 100MB RAM
传统的浅层学习无法揭示大数据复杂的规律,大量无标记、多模态数据难以直接利用(variety)	深度学习	深度神经网络	人工神经网络	深度神经网络[24]	804 414	–
				无监督特征学习[25]	10 000 000	1 000 台计算机 (16 000 核)的集群
数据规模大,计算时间长(volume)	可扩展算法	随机采样、增量处理	模糊系统	rseFCM[32] spkFCM[32] okFCM[32]	5 000 000 000	普通 PC
		随机梯度下降、增量处理		SGFC[35]	70 000	四核 Inter I5-2400 CPU;24GB RAM
		分治、数据压缩		FAR-HD[41]	651 425	AMD Athlon X2 4200+ CPU;2GB RAM
		数据压缩		FARC-HD[42]	19 020	四核 Pentium Core 2 CPU (2.5GHz);4GB RAM
数据中存在噪声、错误及缺值(veracity)	鲁棒算法	数据预处理	模糊系统	KFCM-FSVM[36]	6 435	双核 Intel Xeon CPU (2GHZ);4GB RAM
数据维度高,算法性能退化(volume)	协同演化	分治	群体智能	CCPSO[52] CCPSO2[52] DECC-DG[54] CBCC-DG[54]	1 000 维	–
复杂数据空间多重假设(variety)	多目标优化	遗传算法、粒子群优化等	演化计算、群体智能	MODPSO[51]	–	Inter(R) Celeron(R)M CPU 520(1.6GHz), 512MB RAM

2.5 存在的问题和进一步的研究方向

计算智能方法能够处理非确定性的复杂问题,这使其非常适于处理多变的(variability)、多样的(variety)大数据.大数据的特点也给计算智能带来了新挑战,许多针对小数据上的计算智能方法不能直接应用于海量(volume)、高速(velocity)的大数据.从数据分析的角度来看,大数据主要带来以下问题:

1) 数据规模的膨胀使得算法的时空开销随之增长,原本在小数据集上可以被接受的计算复杂度,在大数据集上变得不可接受;
2) 对于多数大数据应用而言,数据是持续产生并不断变化的,数据无法直接载入计算机的主存储器,无法保存全部的历史样本,也无法像传统的批量算法那样从历史数据中构建无偏训练集;
3) 在大数据环境下,人们生产和采集数据的能力日益增强,手段愈发丰富,这将导致数据在规模增大的同时,属性数量也随之增长,数据呈现出高维、稀疏和复杂的特点.

大数据对计算智能的发展提出了新的机遇和挑战.针对大数据分析,我们认为,未来的研究方向应该包括以下一些内容:

1) 提高算法的可扩展性.这里的可扩展性主要指如何处理更大规模的业务,即,当问题规模扩大时,算法或模型在时间、空间上的性能以及结果的质量.由于很多计算智能方法的研究是先于大数据开展的,因此这些方法并非针对大数据分析,也缺少这些方法在大数据分析中性能的相关报道.将原本针对小数据集的计算智能方法移植到大数据集上,是有待研究的重要问题.解决这一问题需要提高算法的可扩展性,常见策略可以归纳为 4 类:在线优化的方法[69−71]、随机化算法[72−75]、基于哈希的方法[76−81]以及通过大规模计算集群实现分布式并行计算[82].应该进一步研究如何将这些常见策略与计算智能方法相结合,如何发展具有高可扩展性的计算智能新方法.
2) 采用分而治之的策略,实现原始问题的化大为小、化难为易、化繁为简.分而治之是处理大规模复杂问题的直接、可行的策略,其关键问题在于如何对问题进行抽象和划分,如何由子问题的解推演出全局解.
3) 粒计算的理论和模型使得计算机可以从多层次、多角度分析和解决问题,实现多层次、多粒度间灵活自如地切换,为复杂问题的求解提供了新的思路[83].粒计算对于大数据处理中面临的主要挑战有着十分积极的作用.作为一种计算范式,粒计算已经在智能信息处理领域发挥了重要的作用,但将其应用于大数据分析还处于起步阶段.
4) 大数据分析中是否一定要使用原始数据集中的全部数据?如果不是,如何发展更为有效和丰富的手段找到和问题相关的数据子集,是值得研究的问题.在一些应用中,对数据进行采样,将应用于小数据集的传统方法应用于大数据的子集上,以牺牲部分准确率为代价降低时空开销,通过分析大数据的子集来揭示大数据中蕴含的规律是一种可行的策略[84−86].
5) 发展在线分析算法.对于无法一次性载入内存的大数据集,每次以一定的概率分布随机选择一个样本作为输入,输出结果仅基于当前样本进行更新;对于持续不断产生的流数据实现在线处理.针对在线算法,应关注以下问题:多个数据源并存时,如何有效地对结果进行融合;当数据分布发生变化时,如何保证数据分析的稳定性;如何降低噪声、离群点、缺值对算法性能的影响;如何满足部分应用中对算法实时性的要求[87,88].
6) 在许多大数据应用场景中,数据蕴含的规律并不是一成不变的,而是不断发展演化的.例如,热点事件的互联网讨论声量会随着时间的推移发生变化,在事件的产生、传播、变化和消亡各个阶段呈现出不同的规律.因此,大数据分析不仅要对某一时刻的数据进行准确的预测,而且要能揭示出数据的动态模式,从而可以更好地挖掘和利用数据中的价值.针对大数据的分析,要关注数据演化的机制是什么、数据的演化是否存在一般性的假设等.
7) 考虑如何利用不同数据源包含的冗余信息,例如天文学和遥感观测上的多波段交叉;如何利用和分析大数据中存在的大量弱约束规则,例如人与人之间的社交网络群组关系、事件之间的空间位置关系、时间先后关系、触发关系等.
8) 大数据的价值密度较低,而往往其中的异常模式具有较高的价值,例如视频监控中的异常现象、日志数据中的错误或故障记录、金融业务中的欺诈行为、舆情分析中的敏感事件、时域天文学中的激变目标等,因此,需要发展快速、可靠的方法来检测异常模式.

9) 研究如何对高维数据进行能够反映其本质规律的低维刻画;如何从高维的大数据集中过滤掉冗余特征,抽取与问题相关的最优特征组合.在大数据环境下,数据约简是大数据分析的必要环节.与此同时,大数据的特点又对数据约简提出了新的研究课题:首先,大数据的规模对数据约简的计算效率提出了更高的要求,目前,大部分的方法依然是针对小数据集,这些方法应用于大数据中可能会遇到可扩展性的问题;其次,很多数据约简的方法是与分类器相关联的[21,56,89],这些方法以分类器的精度来评价和指导数据约简,因此,本质上还是针对传统的小数据集;最后,目前数据约简的研究以离线方式为主,而针对数据流的在线数据约简是值得关注的问题.

10) 大数据分析要结合领域知识,单纯的数据分析可能会找到很多规律,但这些规律却未必是有实际价值的.只有当数据结合领域知识时,才更容易形成精准的领域模型,从而产生价值.此外,应当建立评价大数据分析结果的可行标准和有效方法,增强结论的可解释性,形成可理解的知识以获取更强的决策力和洞察力.

11) 以 MapReduce[90]为代表的分布式并行计算模型的出现,有力地推动了大数据在产业界的应用和发展.虽然针对大数据的计算模型发展迅速,新的模型和技术层出不穷,但作为 MapReduce 开源实现的 Apache Hadoop 仍是目前较为成熟且应用最为广泛的大数据处理平台.在未来一段时期内,相信仍将是以 Hadoop 为主、多平台共存的基本格局.一些研究人员已经开展了一些针对传统计算智能算法在 MapReduce 框架下实现的研究[91−94].如何设计和实现针对包括 Hadoop 在内的多种平台下的计算智能算法,仍将是未来的重要研究方向.

3 数据源共享问题

大数据分析是领域相关的,研究大数据分析方法首先要解决数据源的问题.目前,大数据的来源主要包括管理信息系统、Web 信息系统、物理信息系统和科学实验系统(http://www.ccf.org.cn/sites/ccf/ccfziliao.jsp?contentId=2774793649105),覆盖互联网与通信业、金融业、制造业、电子政务、健康医疗与制药、公共安全、智慧城市、能源、基础自然科学研究等众多领域.这些领域大多涉及到企业的核心商业利益、国家安全、公民隐私、法律法规等诸多问题,短期内开放与共享是不易实现的,这不利于大数据分析相关研究的广泛开展.基础自然科学研究领域一般不涉及商业利益和公民隐私,其中,数据密集型的科学研究正在从以往的小样本向大数据模式转变,这类数据为大数据分析的研究提供了可以利用的数据源.

在众多基础自然科学研究领域中,天文学更被视为信息爆炸的起源[6].设计于 20 世纪 90 年代的斯隆数字巡天(Sloan digital sky survey,简称 SDSS)标志着天文学步入了大数据时代[95].在建的时域天文学(time domain astronomy)巡天项目大型综合巡天望远镜(large synoptic survey telescope,简称 LSST)在每个观测夜将产生约 30TB 的数据,期望发现 10 万个激变目标,观测周期持续 10 年,最终的图像数据预计将达到约 70PB,星表将达到 10PB~20PB[96].现代天文学的诸多研究领域已经或正在与信息技术紧密结合,成为一项由数据驱动的科研活动,为大数据分析的研究提供了海量的数据素材,同时也拓宽了大数据分析的研究空间.一方面,大型天文巡天项目已经积累了大量的数据,并持续产生海量的新数据,这些数据包括多波段的天文图像、天体光谱、星表以及模拟数据;另一方面,天文大数据来源于科学实验系统,它面向的是基础自然科学研究,不以挖掘商业价值为目的,不会危害国家安全和侵犯个人隐私,完全可以在全球范围内实现数据资源的公开和共享.事实上,国际虚拟天文台联盟(IVOA)(http://www.ivoa.net/)一直致力于制订一套完整的标准以促进天文数据和资源在全球范围内实现共享,是一个非常有代表性的 e-Science 信息技术研究项目[97].

在庆祝创刊 125 周年之际,Science 杂志在 2005 年公布了未来人类将致力于研究解决的 125 个最具挑战性的科学问题,其中排名首位的就是破解宇宙的构成这一科学难题.此外,这些科学问题还包括 9 个方面与天文学密切相关的问题[98].随着天文观测技术的发展,开展大规模天文巡天已成为天文学家研究宇宙的主要形式和有效手段,也是未来天文学家解决上述科学问题的基本途径.目前,国际上已有多个国家开展或筹建包括 SDSS, Pan-STARRS(the panoramic survey telescope and rapid response system),WISE(wide-field infrared survey

explorer),2MASS(two micron all sky survey),Gaia,UKIDSS(UKIRT infrared deep sky survey),CRTS(the catalina real-time transient survey),FIRST(faint images of the radio sky at twenty-centimeters),2df(two-degree-field galaxy redshift survey),郭守敬望远镜(large sky area multi-object fiber spectroscopic telescope,简称 LAMOST),500 米口径球面射电望远镜(five hundred meters aperture spherical telescope,简称 FAST),LSST 在内的众多大规模巡天项目.这些大型天文巡天项目的科学使命与天文学的核心科学问题密切相关.在建的大型巡天项目 LSST 包括探索暗物质和暗能量、观测太阳系内的小行星尤其是对地球具有潜在威胁的小行星、侦测瞬态天文现象以及银河系观测这四大科学使命,其中,通过观测弱引力透镜和超新星探索暗物质和暗能量,对于破解宇宙构成的难题意义重大.我国郭守敬望远镜(LAMOST)(http://www.lamost.org/)的科学目标集中在河外星系的观测、银河系结构和演化以及多波段目标证认这 3 个方面.它对千万数量级的星系、类星体等河外天体以及恒星的光谱进行观测,将在宇宙学模型、宇宙大尺度结构、星系形成和演化、银河系结构与演化以及恒星物理的研究上做出重大贡献.

破解宇宙的结构和演化是一个环环相扣的工程,海量天文数据的分析则是最基础的一环.天文数据分析与计算智能、机器学习、数据挖掘、图像处理等领域密切相关,应用了这些领域的众多方法和技术,其基本的研究问题目前包括分类[99,100]、聚类[101,102]、异常(离群)检测[103,104]、目标识别与跟踪[80,105]、时域数据分析[106]、天文图像和光谱处理[107,108]等.面对爆炸式增长的天文数据,亟需与之相应的大数据分析能力.如何丰富天文数据的分析手段,将计算智能方法应用于天文大数据分析?如何发展针对天文大数据分析的计算智能新方法,从不断增长的海量天文数据中挖掘科学价值,识别其中的已知天体,发现隐藏其中的未知天体、稀有天体和新的天文现象?解决上述问题是天文学家和数据科学家面临的共同课题.

如果以研究大数据分析的技术为目的,而不是挖掘其中的商业价值,在难以获取其他大数据时,可以根据天文学领域的需求,结合计算智能、机器学习、模式识别、系统科学等领域的理论与方法,研究与发展天文大数据的分析技术.这样既可以解决大数据研究面临的数据获取困难的现实问题,推动大数据研究的发展,又可以把大数据分析技术应用于天文学,从大数据的视角重新审视宇宙,形成对宇宙更加深刻的理解.此外,大部分领域的数据获取是一个自然记录的过程,是数据的自然沉淀,例如社交网络上个人发布的日志、图片、音频、视频,电商网站记录的交易、商品和用户的购物行为,路网监控拍摄的视频,设备传感器采集的数据等,这些领域的数据收集成本较低.与之相比,天文数据的获取需要造价昂贵的专业观测设施和大量专业技术人员,获取成本极高,因此必须充分利用已有的历史数据和不断产生的新数据,发展新的方法充分挖掘其中蕴藏的科学价值.

大数据意味着科学研究范式的变革.大数据使得科学研究在经历了实验科学、理论科学、计算科学后,进入了数据密集型科学研究的阶段.微软研究院发布的《The Fourth Paradigm: Data-intensive Scientific Discovery》[109]首次全面地阐述了数据密集型科学研究,为科学研究提供了一种全新的思维与范式.然而,就中国计算机学会发布的《2015 年大数据发展趋势预测》来看,大数据在包括天文学在内的基础自然科学研究领域的应用与其他领域相比,关注度相对不高,发展相对滞后.尽管如此,天文大数据分析已经开始引起学术界、产业界及政府的共同关注.2015 年 5 月 25 日,在国务院总理李克强出访智利期间,中科院国家天文台、智利圣玛利亚理工大学和华为智利分公司三方在智利圣地亚哥签署共建中智天文大数据中心的合作协议,旨在促进中智双方科研人员即时分享在智运行的国际先进观测设备所产生的天文大数据,从而开展最前沿的天文学研究工作.这不仅有利于提升中智双方的天文科研水平,而且将对天文学家和数据科学家深入合作,广泛开展天文大数据分析的相关研究起到带动和示范作用.天文大数据分析作为一个较新的研究领域,存在广阔的发展空间和难得的发展机遇,是大数据分析中值得关注的研究领域之一.

4 结束语

大数据在带来巨大机遇的同时,也给信息技术提出了严峻的挑战.本文结合大数据的特点对大数据分析中计算智能方法的研究进行了归纳和总结,讨论了大数据分析中计算智能存在的问题和未来的研究方向,阐述了数据源共享问题,并建议利用丰富和开放的天文大数据,开展基于计算智能的大数据分析研究.总之,计算智能

在大数据分析中具有巨大的应用潜力,尽管目前已经有了一些探索性的研究工作,但是总体上看,针对大数据分析的计算智能方法的研究还处于起步阶段,尚有诸多问题亟待解决.

References:

[1] Jagadish H, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C. Big data and its technical challenges. Communications of the ACM, 2014,57(7):86−94. [doi: 10.1145/2611567]

[2] Guo P. Computational Intelligence for Software Reliability Engineering. Beijing: Science Press, 2012. 14−19, 234−254 (in Chinese).

[3] Jin Y, Hammer B. Computational intelligence in big data. IEEE Computational Intelligence Magazine, 2014,9(3):12−13. [doi: 10.1109/MCI.2014.2326098]

[4] Li GJ, Cheng XQ. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012,27(6):647−657 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3045.2012.06.001]

[5] Chen CLP, Zhang CY. Data-Intensive applications, challenges, techniques and technologies: A survey on big data. Information Sciences, 2014,275:314−347. [doi: doi:10.1016/j.ins.2014.01.015]

[6] Mayer-Schönberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York: Houghton Mifflin Harcourt, 2013. 2−23.

[7] Zikopoulos P, Eaton C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York: McGraw-Hill Osborne Media, 2011. 5−9.

[8] Pei J. Some new progress in analyzing and mining uncertain and probabilistic data for big data analytics. In: Proc. of the 14th Int'l Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Berlin: Springer-Verlag, 2013. 38−45. [doi: 10.1007/978-3-642- 41218-9_5]

[9] Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Machine Learning, 1999,37(3):277−296. [doi: 10.1023/A:1007662407062]

[10] Aston N, Liddle J, Hu W. Twitter sentiment in data streams with perceptron. Journal of Computer and Communications, 2014, 2(3):11−16. [doi: 10.4236/jcc.2014.23002]

[11] Collins M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proc. of the ACL-02 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002. 1−8. [doi: 10.3115/1118693.1118694]

[12] Littlestone N, Warmuth MK. The weighted majority algorithm. Information and Computation, 1994,108(2):212−261. [doi: 10.1006/inco.1994.1009]

[13] Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y. Online passive-aggressive algorithms. The Journal of Machine Learning Research, 2006,7:551−585.

[14] Blondel M, Kubo Y, Ueda N. Online passive-aggressive algorithms for non-negative matrix factorization and completion. In: Proc. of the 17th Int'l Conf. on Artificial Intelligence and Statistics. Reykjavik: JMLR, 2014. 96−104.

[15] Crammer K, Lee DD. Learning via gaussian herding. In: Proc. of the 24th Annual Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2010. 451−459.

[16] Dredze M, Crammer K, Pereira F. Confidence-Weighted linear classification. In: Proc. of the 25th Int'l Conf. on Machine Learning. Helsinki: ACM Press, 2008. 264−271. [doi: 10.1145/1390156.1390190]

[17] Wang J, Zhao P, Hoi SCH. Exact soft confidence-weighted learning. In: Proc. of the 29th Int'l Conf. on Machine Learning. Edinburgh: Omnipress, 2012. 121−128.

[18] Orabona F, Crammer K. New adaptive algorithms for online classification. In: Proc. of the 24th Annual Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2010. 1840−1848.

[19] Dekel O, Shalev-Shwartz S, Singer Y. The forgetron: A kernel-based perceptron on a budget. SIAM Journal on Computing, 2008, 37(5):1342−1372. [doi: 10.1137/060666998]

[20] Sengamedu SH. Scalable analytics-algorithms and systems. In: Proc. of the 1st Int'l Conf. on Big Data Analytics. New Delhi: Springer-Verlag, 2012. 1−7. [doi: 10.1007/978-3-642-35542-4_1]

[21] Castellano G, Fanelli AM. Variable selection using neural-network models. Neurocomputing, 2000,31(1):1−13. [doi: 10.1016/S0925-2312(99)00146-0]

[22] Rasti J, Monadjemi A, Vafaei A. Color reduction using a multi-stage Kohonen self-organizing map with redundant features. Expert Systems with Applications, 2011,38(10):13188−13197. [doi: 10.1016/j.eswa.2011.04.132]

[23] Sagheer A, Tsuruta N, Taniguchi RI, Arita D, Maeda S. Fast feature extraction approach for multi-dimension feature space problems. In: Proc. of the 18th Int'l Conf. on Pattern Recognition. Hong Kong: IEEE, 2006. 417−420. [doi: 10.1109/ICPR.2006.545]

[24] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006,313(5786):504−507. [doi: 10.1126/science.1127647]

[25] Le QV, Ranzato MA, Monga R, Devin M, Chen K, Corrado GS, Dean Jeff, Ng AY. Building high-level features using large scale unsupervised learning. In: Proc. of the 29th Int'l Conf. on Machine Learning. Edinburgh: ICML, 2012. 81−88.

[26] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009,2(1):1−127. [doi: 10.1561/2200000006]

[27] Zhou ZH, Chawla NV, Jin Y, Williams GJ. Big ata opportunities and challenges: Discussions from data analytics perspectives. IEEE Computational Intelligence Magazine, 2014,9(4):62−74. [doi: 10.1109/MCI.2014.2350953]

[28] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the Advances In Neural Information Processing Systems. Lake Tahoe: Curran Associates, Inc., 2012. 1097−1105.

[29] Dahl GE, Yu D, Deng L, Acero A. Context-Dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. on Audio, Speech, and Language Processing, 2012,20(1):30−42. [doi: 10.1109/TASL.2011.2134090]

[30] Socher R, Lin CC, Ng AY, Manning CD. Parsing natural scenes and natural language with recursive neural networks. In: Proc. of the 28th Int'l Conf. on Machine Learning. Bellevue: Omnipress, 2011. 129−136.

[31] Yu K, Jia L, Chen Y, Xu W. Deep learning: Yesterday, today, and tomorrow. Journal of Computer Research and Development, 2013,50(9):1799−1804 (in Chinese with English abstract).

[32] Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M. Fuzzy *c*-means algorithms for very large data. IEEE Trans. on Fuzzy Systems, 2012,20(6):1130−1146. [doi: 10.1109/TFUZZ.2012.2201485]

[33] Hore P, Hall L, Goldgof D, Cheng W. Online fuzzy *c* means. In: Proc. of the 2008 North American Fuzzy Information Processing Society Conf. New York: IEEE, 2008. 1−5. [doi: 10.1109/NAFIPS.2008.4531233]

[34] Ene A, Im S, Moseley B. Fast clustering using MapReduce. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Diego: ACM Press, 2011. 681−689. [doi: 10.1145/2020408.2020515]

[35] Wang Y, Chen L, Mei JP. Stochastic gradient descent based fuzzy clustering for large data. In: Proc. of the 2014 IEEE Int'l Conf. on Fuzzy Systems. Beijing: IEEE, 2014. 2511−2518. [doi: 10.1109/FUZZ-IEEE.2014.6891755]

[36] Yang X, Zhang G, Lu J, Ma J. A kernel fuzzy *c*-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises. IEEE Trans. on Fuzzy Systems, 2011,19(1):105−115. [doi: 10.1109/TFUZZ.2010.2087382]

[37] Hore P, Hall LO, Goldgof DB. Single pass fuzzy *c* means. In: Proc. of the 2007 IEEE Int'l Conf. on Fuzzy Systems. London: IEEE, 2007. 1−7. [doi: 10.1109/FUZZY.2007.4295372]

[38] Eschrich S, Ke J, Hall LO, Goldgof DB. Fast accurate fuzzy clustering through data reduction. IEEE Trans. on Fuzzy Systems, 2003,11(2):262−270. [doi: 10.1109/TFUZZ.2003.809902]

[39] Hore P, Hall LO, Goldgof DB, Gu Y, Maudsley AA, Darkazanli A. A scalable framework for segmenting magnetic resonance images. Journal of Signal Processing Systems, 2009,54(1-3):183−203. [doi: 10.1007/s11265-008-0243-1]

[40] Krishnan S, Bhattacharyya C, Hariharan R. A randomized algorithm for large scale support vector learning. In: Proc. of the 21st Annual Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2007. 793−800.

[41] Mangalampalli A, Pudi V. FAR-HD: A fast and efficient algorithm for mining fuzzy association rules in large high-dimensional datasets. In: Proc. of the 2013 IEEE Int'l Conf. on Fuzzy Systems. Hyderabad: IEEE, 2013. 1−6. [doi: 10.1109/FUZZ-IEEE.2013.6622333]

[42] Alcala-Fdez J, Alcala R, Herrera F. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans. on Fuzzy Systems, 2011,19(5):857−872. [doi: 10.1109/TFUZZ.2011.2147794]

[43] Holland JH. Outline for a logical theory of adaptive systems. Journal of the ACM, 1962,9(3):297−314. [doi: 10.1145/321127.321128]

[44] Bacardit J, Llorà X. Large-Scale data mining using genetics-based machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2013,3(1):37−61. [doi: 10.1002/widm.1078]

[45] Antonelli M, Ducange P, Marcelloni F. Genetic training instance selection in multiobjective evolutionary fuzzy systems: A coevolutionary approach. IEEE Trans. on Fuzzy Systems, 2012,20(2):276−290. [doi: 10.1109/TFUZZ.2011.2173582]

[46] Oh IS, Lee JS, Moon BR. Hybrid genetic algorithms for feature selection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26(11):1424−1437. [doi: 10.1109/TPAMI.2004.105]

[47] Tsai CF, Eberle W, Chu CY. Genetic algorithms in feature and instance selection. Knowledge-Based Systems, 2013,39:240−247. [doi: 10.1016/j.knosys.2012.11.005]

[48] Gheyas IA, Smith LS. Feature subset selection in large dimensionality domains. Pattern Recognition, 2010,43(1):5−13. [doi: 10.1016/j.patcog.2009.06.009]

[49] Aggarwal CC, Yu PS. An effective and efficient algorithm for high-dimensional outlier detection. The VLDB Journal, 2005,14(2):211−221. [doi: 10.1007/s00778-004-0125-5]

[50] Valle YD, Venayagamoorthy GK, Mohagheghi S, Hernandez JC, Harley RG. Particle swarm optimization: basic concepts, variants and applications in power systems. IEEE Trans. on Evolutionary Computation, 2008,12(2):171−195. [doi: 10.1109/TEVC.2007.896686]

[51] Gong M, Cai Q, Chen X, Ma L. Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition. IEEE Trans. on Evolutionary Computation, 2014,18(1):82−97. [doi: 10.1109/TEVC.2013.2260862]

[52] Li X, Yao X. Cooperatively coevolving particle swarms for large scale optimization. IEEE Trans. on Evolutionary Computation, 2012,16(2):210−224. [doi: 10.1109/TEVC.2011.2112662]

[53] Yang Z, Tang K, Yao X. Large scale evolutionary optimization using cooperative coevolution. Information Sciences, 2008, 178(15):2985−2999. [doi: 10.1016/j.ins.2008.02.017]

[54] Omidvar MN, Li X, Mei Y, Yao X. Cooperative co-evolution with differential grouping for large scale optimization. IEEE Trans. on Evolutionary Computation, 2014,18(3):378−393. [doi: 10.1109/TEVC.2013.2281543]

[55] Chen Y, Miao D, Wang R. A rough set approach to feature selection based on ant colony optimization. Pattern Recognition Letters, 2010,31(3):226−233. [doi: 10.1016/j.patrec.2009.10.013]

[56] Aghdam MH, Ghasem-Aghaee N, Basiri ME. Text feature selection using ant colony optimization. Expert systems with applications, 2009,36(3):6843−6853. [doi: 10.1016/j.eswa.2008.08.022]

[57] Al-Ani A. Feature subset selection using ant colony optimization. Int'l Journal of Computational Intelligence, 2006,2(1):53−58.

[58] Wang X, Yang J, Teng X, Xia W, Jensen R. Feature selection based on rough sets and particle swarm optimization. Pattern Recognition Letters, 2007,28(4):459−471. [doi: 10.1016/j.patrec.2006.09.003]

[59] Liu Y, Wang G, Chen H, Dong H, Zhu X, Wang S. An improved particle swarm optimization for feature selection. Journal of Bionic Engineering, 2011,8(2):191−200. [doi: 10.1016/S1672-6529(11)60020-6]

[60] Ke L, Feng Z, Ren Z. An efficient ant colony optimization approach to attribute reduction in rough set theory. Pattern Recognition Letters, 2008,29(9):1351−1357. [doi: 10.1016/j.patrec.2008.02.006]

[61] Jones DF, Mirrazavi SK, Tamiz M. Multi-Objective meta-heuristics: An overview of the current state-of-the-art. European Journal of Operational Research, 2002,137(1):1−9. [doi: 10.1016/S0377-2217(01)00123-0]

[62] Konak A, Coit DW, Smith AE. Multi-Objective optimization using genetic algorithms: A tutorial. Reliability Engineering & System Safety, 2006,91(9):992−1007. [doi: 10.1016/j.ress.2005.11.018]

[63] Gong MG, Jiao LC, Yang DD, Ma WP. Research on evolutionary multi-objective optimization algorithms. Ruan Jian Xue Bao/Journal of Software, 2009,20(2):271−289 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3483.htm [doi: 10.3724/SP.J.1001.2009.03483]

[64] Coello CAC. Evolutionary multi-objective optimization: A historical view of the field. IEEE Computational Intelligence Magazine, 2006,1(1):28−36. [doi: 10.1109/MCI.2006.1597059]

[65] Marler RT, Arora JS. Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization, 2004,26(6):369−395. [doi: 10.1007/s00158-003-0368-6]

[66] Deb K. Multi-Objective Optimization Using Evolutionary Algorithms. New York: Wiley, 2001.

[67] Jin Y, Branke J. Evolutionary optimization in uncertain environments—A survey. IEEE Trans. on Evolutionary Computation, 2005,9(3):303−317. [doi: 10.1109/TEVC.2005.846356]

[68] Alaya I, Solnon C, Ghédira K. Ant colony optimization for multi-objective optimization problems. In: Proc. of the 19th Int'l Conf. on Tools with Artificial Intelligence. Patras: IEEE, 2007. 450−457. [doi: doi:10.1109/ICTAI.2007.108]

[69] Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics, 1952,23(3):462−466. [doi: 10.1214/aoms/1177729392]

[70] Bottou L. Large-Scale machine learning with stochastic gradient descent. In: Proc. of the 19th Int'l Conf. on Computational Statistics. Paris: Springer-Verlag, 2010. 177−186. [doi: 10.1007/978-3-7908-2604-3_16]

[71] Roux NL, Schmidt M, Bach FR. A stochastic gradient method with an exponential convergence rate for finite training sets. In: Proc. of the 26th Annual Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates, Inc., 2012. 2663−2671.

[72] Paul S, Boutsidis C, Magdon-Ismail M, Drineas P. Random projections for support vector machines. In: Proc. of the 16th Int'l Conf. on Artificial Intelligence and Statistics. Scottsdale: JMLR, 2013. 498−506.

[73] Drineas P, Magdon-Ismail M, Mahoney MW, Woodruff DP. Fast approximation of matrix coherence and statistical leverage. The Journal of Machine Learning Research, 2012,13(1):3475−3506.

[74] Kumar K, Bhattacharya C, Hariharan R. A randomized algorithm for large scale support vector learning. In: Proc. of the 21st Annual Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2007. 793−800.

[75] Rahimi A, Recht B. Random features for large-scale kernel machines. In: Proc. of the 21st Annual Conf. on Neural Information Processing Systems. Vancouve: Curran Associates, Inc., 2007. 1177−1184.

[76] Li P, Shrivastava A, Moore JL, König AC. Hashing algorithms for large-scale learning. In: Proc. of the 25th Annual Conf. on Neural Information Processing Systems. Granada: Curran Associates, Inc., 2011. 2672−2680.

[77] Kulis B, Grauman K. Kernelized locality-sensitive hashing. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(6):1092−1104. [doi: 10.1109/TPAMI.2011.219]

[78] Mu Y, Wright J, Chang SF. Accelerated large scale optimization by concomitant hashing. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer-Verlag, 2012. 414−427. [doi: 10.1007/978-3-642-33718-5_30]

[79] Mu Y, Hua G, Fan W, Chang SF. Hash-SVM: Scalable kernel machines for large-scale visual classification. In: Proc. of the 27th IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 979−986. [doi: 10.1109/CVPR.2014.130]

[80] Wang K, Guo P. An efficient and scalable learning algorithm for near-earth objects detection in astronomy big image data. In: Proc. of 2014 IEEE Int'l Conf. on Systems, Man, and Cybernetics. San Diego: IEEE, 2014. 742−747. [doi: 10.1109/SMC.2014.6973999]

[81] Shrivastava A, Li P. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In: Proc. of the 28th Annual Conf. on Neural Information Processing Systems. Montreal: Curran Associates, Inc., 2014. 2321−2329.

[82] Zhu K, Wang H, Bai H, Li J, Qiu Z, Cui H, Chang EY. Parallelizing support vector machines on distributed computers. In: Proc. of the 21st Annual Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2007. 257−264.

[83] Miao DQ, Wang GY, Liu Q, Lin ZY, Yao YY. Granular Computing: Past, Present and Future. Beijing: Science Press, 2007 (in Chinese).

[84] Choudhury MD, Lin YR, Sundaram H, Candan KS, Xie L, Kelliher A. How does the data sampling strategy impact the discovery of information diffusion in social media? In: Proc. of the 4th Int'l Conf. on Weblogs and Social Media. Washington: The AAAI Press, 2010. 34−41.

[85] García-Pedrajas N, de Haro-García A, Pérez-Rodríguez J. A scalable approach to simultaneous evolutionary instance and feature selection. Information Sciences, 2013,228:150−174. [doi: 10.1016/j.ins.2012.10.006]

[86] Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large databases. Information Systems, 2001,26(1):35−58. [doi: 10.1016/S0306-4379(01)00008-4]

[87] Mardani M, Mateos G, Giannakis GB. Dynamic anomalography: Tracking network anomalies via sparsity and low rank. The Journal of Selected Topics in Signal Processing, 2013,7(1):50−66. [doi: 10.1109/JSTSP.2012.2233193]

[88] Ball NM, Brunner RJ. Data mining and machine learning in astronomy. Int'l Journal of Modern Physics D, 2010,19(7):1049−1106. [doi: 10.1142/S0218271810017160]

[89] Li S, Wu H, Wan D, Zhu J. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. Knowledge Based Systems, 2011,24(1):40−48. [doi: 10.1016/j.knosys.2010.07.003]

[90] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 2008,51(1):107−113. [doi: 10.1145/1327452.1327492]

[91] McNabb AW, Monson CK, Seppi KD. Parallel PSO using mapreduce. In: Proc. of the IEEE Congress on Evolutionary Computation. Singapore: IEEE, 2007. 7−14. [doi: 10.1109/CEC.2007.4424448]

[92] Jin C, Vecchiola C, Buyya R. MRPGA: An extension of MapReduce for parallelizing genetic algorithms. In: Proc. of the 4th Int'l Conf. on e-Science. Indianapolis: IEEE Computer Society, 2008. 214−221. [doi: 10.1109/eScience.2008.78]

[93] Zhang J, Li T, Ruan D, Gao Z, Zhao C. A parallel method for computing rough set approximations. Information Sciences, 2012, 194:209−223. [doi: 10.1016/j.ins.2011.12.036]

[94] Zhang J, Wong JS, Li T, Pan Y. A comparison of parallel large-scale knowledge acquisition using rough set theory on different MapReduce runtime systems. Int'l Journal of Approximate Reasoning, 2014,55(3):896−907. [doi: 10.1016/j.ijar.2013.08.003]

[95] Feigelson ED, Babu GJ. Big data in astronomy. Significance, 2012,9(4):22−25. [doi: 10.1111/j.1740-9713.2012.00587.x]

[96] Borne K. A machine learning classification broker for the LSST transient database. Astronomische Nachrichten, 2008,329(3): 255−258. [doi: 10.1002/asna.200710946]

[97] Zhao YH, Cui CZ. From virtual observatory to astroinformatics. e-Science Technology & Application, 2011,2(3):3−12 (in Chinese with English abstract).

[98] Kennedy D, Norman C. What don't we know? Science, 2005,309(5731):75−75. [doi: 10.1126/science.309.5731.75]

[99] Debosscher J, Sarro LM, Aerts C, Cuypers J, Vandenbussche B, Garrido R, Solano E. Automated supervised classification of variable stars. I. Methodology. Astronomy and AStrophysics, 2007,475(3):1159−1183. [doi: 10.1051/0004-6361:20077638]

[100] Henrion M, Mortlock DJ, Hand DJ, Gandy A. A Bayesian approach to star–galaxy classification. Monthly Notices of the Royal Astronomical Society, 2011,412(4):2286−2302. [doi: 10.1111/j.1365-2966.2010.18055.x]

[101] Ordóñez D, Dafonte C, Arcay B, Manteiga M. HSC: A multi-resolution clustering strategy in self-organizing maps applied to astronomical observations. Applied Soft Computing, 2012,12(1):204−215. [doi: 10.1016/j.asoc.2011.08.052]

[102] Murphy D, Geach J, Bower R. ORCA: The overdense red-sequence cluster algorithm. Monthly Notices of the Royal Astronomical Society, 2012,420(3):1861−1881. [doi: 10.1111/j.1365-2966.2011.19782.x]

[103] Borne KD, Vedachalam A. Surprise detection in multivariate astronomical data. In: Feigelson ED, Babu GJ, eds. Proc. of the Statistical Challenges in Modern Astronomy V. Berlin: Springer-Verlag, 2012. 275−289. [doi: 10.1007/978-1-4614-3520-4_26]

[104] Dutta H, Giannella C, Borne KD, Kargupta H. Distributed top-*k* outlier detection from astronomy catalogs using the DEMAC system. In: Proc. of the 2007 SIAM Int'l Conf. on Data Mining. Philadelphia: SIAM, 2007. 473−478. [doi: 10.1137/1.9781611972771]

[105] Kubica J, Denneau L, Grav T, Heasley J, Jedicke R, Masiero J, Milani A, Moore A, Tholen D, Wainscoat R. Efficient intra-and inter-night linking of asteroid detections using kd-trees. Icarus, 2007,189(1):151−168. [doi: 10.1016/j.icarus.2007.01.008]

[106] Huijse P, Estevez P, Protopapas P, Principe J, Zegers P. Computational intelligence challenges and applications on large-scale astronomical time series databases. IEEE Computational Intelligence Magazine, 2014,9(3):27−39. [doi: 10.1109/MCI.2014.2326100]

[107] Waniak W. Removing cosmic-ray hits from CCD images in real-time mode by means of an artificial neural network. Experimental Astronomy, 2006,21(3):151−168. [doi: 10.1007/s10686-007-9079-0]

[108] Yu J, Yin Q, Guo P, Luo AL. A deconvolution extraction method for 2D multi-object fibre spectroscopy based on the regularized least-squares QR-factorization algorithm. Monthly Notices of the Royal Astronomical Society, 2014,443(2):1381−1389. [doi: 10.1093/mnras/stu1250]

[109] Hey T, Tansley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond: Microsoft Research, 2009.

附中文参考文献:

[2] 郭平.软件可靠性工程中的计算智能方法.北京:科学出版社,2012.14−19,234−254.

[4] 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.中国科学院院刊,2012,27(6):647−657. [doi: 10.3969/j.issn.1000-3045.2012.06.001]

[31] 余凯,贾磊,陈雨强,徐伟.深度学习的昨天、今天和明天.计算机研究与发展,2013,50(9):1799−1804.

[63] 公茂果,焦李成,杨咚咚,马文萍.进化多目标优化算法研究.软件学报,2009,20(2):271−289. http://www.jos.org.cn/1000-9825/3483.htm [doi: 10.3724/SP.J.1001.2009.03483]

[83] 苗夺谦,王国胤,刘清,林早阳,姚一豫.粒计算:过去、现在与展望.北京:科学出版社,2007.

[97] 赵永恒,崔辰州.从虚拟天文台到天文信息学.科研信息化技术与应用,2011,2(3):3−12.

郭平(1957－),男,山西洪洞人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算智能及其在模式识别,图像处理,天文大数据,软件可靠性工程等方面的应用.

王可(1985－),男,博士生,助教,主要研究领域为计算智能,大数据分析.

罗阿理(1969－),男,博士,研究员,博士生导师,主要研究领域为天文大数据处理.

薛明志(1966－),男,博士,教授,主要研究领域为进化计算及其应用,大数据分析.