大数据智能决策

于 洪[1] 何德牛[1] 王国胤[1] 李 劼[2] 谢永芳[3]

摘 要 在全球信息化快速发展的背景下, 大数据已经成为一种战略资源. 各行各业的决策活动在频度、广度及复杂性上较以往有着本质的不同. 决策过程中的不确定性因素增多, 决策分析的难度不断加大. 传统的数据分析方法以及基于人工经验的决策已难以满足大数据时代的决策需求, 大数据驱动的智能决策将成为决策研究的主旋律. 该文结合大数据特性, 对大数据决策的特点进行了归纳, 并从智能决策支持系统、不确定性处理、信息融合、关联分析和增量分析等方面综述了大数据智能决策的研究与发展现状, 讨论了大数据智能决策依然面临的挑战, 并对一些潜在的研究方向进行了展望分析.

关键词 大数据, 智能决策, 不确定性, 信息融合, 关联分析, 增量式学习

引用格式 于洪, 何德牛, 王国胤, 李劼, 谢永芳. 大数据智能决策. 自动化学报, 2020, **46**(5): 878−896

DOI 10.16383/j.aas.c180861

开放科学 (资源服务) 标识码 (OSID):

Big Data for Intelligent Decision Making

YU Hong[1] HE De-Niu[1] WANG Guo-Yin[1] LI Jie[2] XIE Yong-Fang[3]

Abstract As a result of globalization and informatization, big data has become one kind of important strategic resources. Decision-making activities in all walks of life are different from the past in frequency, breadth and complexity. The difficulty of decision analysis is increased due to the increase of uncertainty factors in the decision-making process. The decision analysis methods based on the traditional data analysis methods or manual experiences are gradually unable to meet the needs of decision-making in the era of big data. We think the intelligent decision making methods based on big data driven will become an important solution. This paper presents the characteristics of big data for intelligent decision making in view of analyzing the features of big data. Some recent theoretic studies and applications of intelligent decision-making systems, uncertainty intelligent decision making, methods based on information fusion, methods based on association analysis and incremental learning are reviewed. The paper also points out the future perspectives and potential research points.

Key words Big data, intelligent decision-making, uncertainty, information fusion, association analysis, incremental learning

Citation Yu Hong, He De-Niu, Wang Guo-Yin, Li Jie, Xie Yong-Fang. Big data for intelligent decision making. *Acta Automatica Sinica*, 2020, **46**(5): 878−896

当今社会处于一个信息技术高速发展时期, 数据信息的交互、共享与开放程度持续加快, 使得各行业领域的数据信息呈爆炸式增长. “大数据时代” 如约而至, 并成为当今社会的代名词. 大数据以其蕴藏巨大的经济、社会和科研价值受到社会各界的广泛关注[1]. 2012 年 1 月, 达沃斯世界经济论坛发布的大数据报告 “Big data, big impact: new possibilities for international development” 将大数据列为和货币与黄金同等重要的新经济资产[2]. 2012 年 5 月, 联合国发布的 *Big Data for Development: Challenges & Opportunities* 白皮书指出, 大数据是联合国和各国政府的一个历史性机遇, 利用大数据进行决策, 是提升国家治理能力, 实现治理能力现代化的必然要求, 可以帮助政府更好地参与经济社会的运行与发展[3]. 在科研领域, 大数据正引领数据密集型科学 (Data-intensive science) 的到来, 形成继实验科学、理论科学以及计算科学之后的第四科学范式[4], 有望推动传统科学的假设驱动模式向基于大数据探索的数据密集型方法转变. 在全球信息化快速发展的背景下, 大数据已逐渐成为世界各国的基础性战略

收稿日期 2018-12-29 录用日期 2019-04-11
Manuscript received December 29, 2018; accepted April 11, 2019
国家自然科学基金 (61751312, 61533020, 61876027) 资助
Supported by National Natural Science Foundation of China (61751312, 61533020, 61876027)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 重庆邮电大学计算智能重庆市重点实验室 重庆 400065 2. 中南大学冶金与环境学院 长沙 410083 3. 中南大学信息科学与工程学院 长沙 410083
1. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065 2. School of Metallurgy and Environment, Central South University, Changsha 410083 3. School of Information Science and Engineering, Central South University, Changsha 410083

资源, 运用大数据推动社会经济发展正成为趋势.

现阶段加快发展智能经济、智能服务和智能制造是我国经济增长的内在需求和必然选择. 目前我国处于工业化和信息化的深度融合时期, 我国制造业正处于从价值链的低端向中高端、从中国制造向中国创造转变的关键历史时期, 发展基于大数据的人工智能新技术是实现从制造大国向制造强国迈进的战略举措. 在此背景之下, 国家相继出台了 "'互联网 +' 行动计划" 和 "中国制造 2025" 战略规划, 特别是国务院颁布的 "促进大数据发展行动纲要" 和 "新一代人工智能发展规划" 都将大数据智能作为重点发展方向, 大数据的战略资源地位进一步凸显. 近年来, 以大数据与人工智能技术为基础的 "智能制造[5]" 成为推动大数据从概念到落地的重要模式和手段. 从大数据的供给需求来看, 智能制造的核心要义便是在两化融合的基础上构建智能分析优化系统 "工业大脑", 对大数据进行智能化分析进而实现智能决策.

决策存在于人类一切实践活动当中. 小到一台机器的操作, 大到一个国家的治理, 都离不开决策. 例如, 工业领域的操作优化与资源分配、商业领域的个性化推荐与供应商选择、交通领域的车流控制与路径导航、医疗领域的疾病诊断与治疗策略等都属于决策范畴. 随着社会节奏的持续加快, 来自各领域行业的决策活动在频度、广度及复杂性上较以往都有着本质的提高. 决策问题的不确定性程度随着决策环境的开放程度以及决策资源的变化程度而越来越大. 传统的基于人工经验、直觉及少量数据分析的决策方式已经远不能满足日益个性化、多样化、复杂化的决策需求. 在当前信息开放与交互的经营环境下, 机遇与挑战并存. 如何把握机遇, 这就需要企业或组织具备出色的决策能力. 在这个过程中大数据正扮演着越来越重要的角色.

大数据作为一种重要的信息资产, 可望为人们提供全面的、精准的、实时的商业洞察和决策指导. 杨善林院士等指出, 大数据的价值在于其 "决策有用性", 通过分析、挖掘来发现其中蕴藏的知识, 可以为各种实际应用提供其他资源难以提供的决策支持[6]. 美国应用信息经济学家 Hubbard 认为 "一切皆可量化", 并积极倡导数据化决策[7]. 纽约大学 Provost 教授等认为数据科学的终极目标就是改善决策[8]. 从数据到知识, 从知识到决策, 是当前大数据智能的计算范式[9], 研究大数据的意义就是不断提高 "从数据到决策的能力". 随着大数据技术的发展, 人们传统的决策模式与思维方式正在发生着变革, 基于大数据的决策方式正逐渐成为决策应用与研究领域的主旋律, 大数据决策时代已经到来. 大数据能够突破事物之间隐性因素无法被量化的瓶颈, 充分阐述生产的主客体和生产全过程、全时段的客观状态, 通过智能化分析和预测判断来提高企业的决策能力[10]. 在商业领域, 利用大数据相关分析, 可以更加精准地了解客户的消费行为, 帮助决策者挖掘新的商业模式, 制定商品价格, 实现供应商协同工作, 缓和供需之间的矛盾, 控制预算开支. 例如, 全球零售巨头沃尔玛 (Wal-Mart), 通过对销售交易大数据的知识获取, 成功用于价格策略和推荐活动中的决策支持[11]. 而在工业领域, 为实现智能制造, 每个影响生产决策的因素都可以经过工业大数据的预测, 以直观明了的量化信息形式加以呈现, 方便决策者对制造能力进行整体评估, 进而快速有效地制定各项生产决策, 优化劳动力投入, 避免产能过剩[10]. 目前, 百度的工业大数据监测平台已经应用到汽车、日化等制造行业. 三一重工则利用大数据分析技术为智能工程机械物联网提供决策支持, 推进了制造服务化的步伐. Google 公司旗下的 AlphaGo 以 4:1 的总比分战胜世界围棋冠军李世石同样是大数据决策颇具代表性的案例.

基于大数据的科学决策, 是公共管理、工业制造、医疗健康、金融服务等众多行业领域未来发展的方向和目标. 如何进行大数据的智能分析与科学决策, 实现由数据优势向决策优势的转化, 仍然是当前大数据应用研究中的关键问题. 然而, 对大数据的分析和处理在不同行业和领域均存在着巨大的挑战, 大数据的大体量、高通量、多源异构性和不确定性等对传统的数据处理硬件设备和软件处理方法均构成前所未有的挑战. 目前, 机器学习、数据挖掘及统计理论等传统理论方法已经广泛地应用于大数据分析, 但多数方法是建立在 "独立同分布" 的假设之上, 难以应对大数据的不确定性显著、关联复杂、动态增长、来源和分布广泛等问题, 多数只能挖掘到底层的数据特征, 而对于挖掘高层次的符合人类认知的知识依然无法取得较好的效果, 难以高效地将大数据转化为决策价值. 基于大数据的智能决策是一门集应用性和科研性于一体的学科领域, 目前还存在众多待研究的问题. 大数据智能决策在内涵外延、模型理论、技术方法及实施策略等方面还需要人们继续投入更多的研究与实践.

本文旨在综述大数据决策的特点以及大数据决策技术的发展现状, 分析大数据智能决策面临的问题与挑战, 并对一些潜在研究方向进行展望. 文章结构如下: 第 1 节介绍了大数据的概念及特性, 总结了大数据决策的特点; 第 2 节从智能决策支持系统、基于不确定性分析的智能决策、基于信息融合的智能决策、基于关联分析的智能决策和基于增量分析的智能决策五个方面综述了大数据智能决策的研究与发展现状; 第 3 节讨论了大数据智能决策面临的

挑战与发展趋势; 第 4 节为结束语.

1 大数据决策

1.1 大数据的概念及特性

由于不同领域的大数据在特性上存在差异, 并且人们分析大数据的背景和应用大数据的目的不同, 因此不同的领域专家对大数据的定义也各不相同. 高德纳咨询公司、维基百科、美国国家科学基金会分别从不同的角度给出了大数据的定义. 我国的《工业大数据白皮书 (2019 版)》还对工业大数据进行了定义[12]. 简言之, 大数据就是无法在合理时间内利用现有的数据处理手段进行诸如存储、管理、抓取等分析和处理的数据集合[13].

有关大数据的特性, 业界普遍将其归纳为 4V 特性: 一是数据体量 (Volume) 大, 如一些电商企业日常处理 PB 级别的数据已经常态化; 二是数据类型多样 (Variety), 如在工业大数据中数据类型包含了数值、文本、图片、音频、视频以及传感器信号等; 三是大数据的价值 (Value) 巨大, 但价值密度稀疏, 需要通过分析和挖掘来获取数据当中有价值的信息; 四是大数据的高通量 (Velocity), 它除了指数据高速产生以外, 还意味着数据的采集与分析过程必须迅速及时, 以满足用户 "及时、实时" 的决策需求.

在特定领域, 大数据还有着特有的性质. 如在工业领域, 人们还强调大数据的实时性、闭环性、强关联性、多层面不规则采样性、多时空时间序列性等[14]; 在管理与商业领域, 人们更关注大数据的商用价值, 并提出大数据应用的 5R 模型, 即相关性 (Relevant)、实时性 (Real-time)、真实性 (Realistic)、可靠性 (Reliable)、投资回报 (ROI)[13]. 在科研领域, Wang 等着重分析了大数据的不确定性特征[15]. Wu 等则从大数据的异构 (Heterogeneous)、自治 (Autonomous)、复杂 (Complex)、演化 (Evolving) 四个角度提出了描述大数据特性的 HACE 定理[16].

1.2 大数据决策及其特点

决策是人们为实现某一特定的目标, 在占有一定的信息和经验 (知识) 的基础上, 根据主客观条件的可能性, 提出各种可行方案, 采用一定的科学方法和手段, 对解决问题的方案进行比较、分析和评价, 并最终进行方案选择的全过程. 从本质上来讲, 决策通常是目标驱动的行为, 是目标导向下的问题求解过程, 该过程也广泛地被认为是人类的认知过程. 大数据决策便是以大数据为主要驱动的决策方式. 随着大数据技术的发展, 大数据逐渐成为人们获取对事物和问题更深层次认知的决策资源, 特别是人工智能技术与大数据的深度融合, 为复杂决策的建模和分析提供了强有力的工具.

随着大数据应用越来越多地服务于人们的日常生活, 基于大数据的决策方式将形成其固有的特性和潜在的趋势, 在此我们将它们一并归纳为大数据决策的特点. 在固有特性方面: 大数据的实时产生及动态变化决定了大数据决策的动态性; 大数据的多方位感知意味着通过多源数据的整合可以实现更加全面的决策; 大数据潜在的不确定性也使得决策问题的求解过程呈现不确定性特征. 在潜在趋势方面: 相关分析或将代替因果分析, 成为获取大数据隐含知识更有效的手段; 用户的兴趣偏好在大数据时代将更受关注, 更多的商业决策向满足个性化需求转变. 基于以上理解, 本文对大数据决策的特点进行如下总结:

1) 大数据决策的动态特性

大数据是对事物客观表象和演化规律的抽象表达, 其动态性和增量性是对事物状态的持续反映. 不可否认的是, 人们在决策过程中的每一步行动都将影响事物的发展进程, 并全程由大数据所反映. 此时决策问题的描述以及决策求解的策略都需要跟随动态数据给予及时调整, 通过面向大数据的增量式学习方法实现知识的动态演化与有效积累, 进而反馈到决策执行当中. 大数据决策的动态特性决定了问题的求解过程应该是一个集描述、预测、引导为一体的迭代过程, 该过程须形成一个完整的、闭环的、动态的体系结构. 简要来说, 大数据环境下的决策模型将是一种具备实时反馈的闭环模型, 决策模式将更多地由相对静态的模式或多步骤模式转变为对决策问题动态描述的渐进式求解模式.

2) 大数据决策的全局特性

截至目前, 人们已经开发出多种多样的决策支持系统, 但多数是面向具体领域中的单一生产环节或特定目标下的局部决策问题, 往往无法较好地实现全局决策优化与多目标任务协同. 在信息开放与交互的大数据时代, 大数据的跨视角、跨媒介、跨行业等多源特性创造了信息的交叉、互补与综合运用的条件, 这促使了人们进一步提升问题求解的关联意识和全局意识. 在大数据环境下决策分析会更加注重数据的全方位性, 生产流程的系统性、业务各环节的交互性、多目标问题的协同性. 通过多源异构信息的融合分析, 可以实现不同信源信息对全局决策问题求解的有效协同. 基于大数据的决策系统, 对每个单一问题的决策, 都将以优先考虑整体决策的优化作为前提, 进而为决策者提供企业级、全局性的决策支持.

3) 大数据决策的不确定性特征

一般而言, 决策的不确定性来源于三个方面: 一是决策信息不完整、不确定而导致的决策不确定性; 二是决策信息分析能力不足而导致的决策不确定

性[17]; 三是决策问题过于复杂而难以建模导致的不确定性. 大数据决策的不确定性不外乎以上三个方面. 在信息不完整和不确定方面, 首先, 大数据具有来源和分布广泛、关联关系复杂等特性, 对于多数企业而言, 即便借助各种先进的数据收集手段尽可能地将各种信源数据进行整合, 但仍难以保证信息的全面性和完整性; 其次, 大数据固有的动态特性决定了大数据的分布存在随时间变化的不确定性; 另外, 大数据中普遍存在的噪声与数据缺失现象决定了大数据的不完备、不精确性. 在大数据分析能力方面, 显然现有的大数据分析处理技术还存在着不足, 诸如多源异构数据融合分析、不确定性知识发现及大数据关联分析等方面仍是当前颇具挑战的研究方向. 在决策问题建模方面, 在一些非稳态、强耦合的系统环境下, 建立精确的动态决策模型往往异常困难, 比如流程工业中的操作优化决策. 现阶段面向大数据的决策问题求解, 人们通常使用满意近似解代替精确解, 以此保证问题求解的经济性和高效性. 这种近似求解方式实际上也反映了大数据决策的不确定性特征.

4) 从因果分析向相关分析转变

在过往的数据分析中, 人们往往假设数据的精确性, 并通过反复试验的手段探索事物之间的因果关系. 但在大数据环境下, 数据的精确性难以保证, 数据总体对价值获取的完备性异常重要, 此时用于发现因果关系的反复尝试方法变得异常困难. 从统计学角度看, 变量之间的关系大体可以分两种类型: 函数关系和相关关系, 一般情况下, 数据很难严格地满足函数关系, 而相关关系的要求较为宽松, 在大数据环境下更加容易被接受[18], 并能满足人类的众多决策需求. 该方面的成功案例有 Google 公司的流感预测[19]、啤酒与尿布关联规则的挖掘等. 在面向大数据智能化分析的决策应用中, 相关性分析技术可为正确数据的选择提供必要的判定与依据, 同时将其与其他智能分析方法相结合, 可有效避免对数据独立同分布的假设, 提高数据分析的合理性和认可度.

5) 决策向满足个性化需求转变

在商业和制造业领域, 对用户进行精准营销, 满足用户的个性化需求是提升客户价值和实现企业竞争力的经营准则. 在大数据背景下, 产品和服务的提供以及价值的创造有望更加贴近社会大众的个性化需求. 以互联网大数据为基础, 企业通过舆情分析、情感挖掘等以用户为中心的数据驱动方法, 可以精准挖掘消费者的兴趣与偏好, 做出有针对性的个性化需求预测, 进而为消费者提供专属的个性化产品与服务. 宏观上讲, 大数据可以打通企业和消费者之间的信息主动反馈机制. 社会大众通过意见的表达, 可以迅速转化为商业经营的决策依据, 反向指导产品的设计和制造环节, 实现生产与市场需求的有效对接. 以 Netflix[20] 为代表的推荐系统正是一个基于个性化需求的大数据决策系统. 随着社会化媒体应用的深入, 多元主体参与决策有了更多的便捷性和可能性, 决策过程中价值多元的作用更加明显, 由此传统自上而下的精英决策模型将会改变, 并逐渐形成面向公众与满足用户个性化需求的决策模式.

通过以上有关大数据决策特点的总结, 我们不难发现大数据决策有着相较于传统基于小数据分析决策的诸多不同之处. 更进一步, 大数据决策的特点反应了当前大数据智能决策的研究重点与需求. 大数据决策的不确定性、动态性、全局性以及向相关性分析的转变, 决定了面向大数据的关联分析、不确定性分析、对增量与多源数据的有效利用都将是大数据智能决策研究中的关键内容.

2 大数据智能决策研究现状分析

从静态决策到动态决策、从单人决策到群体决策、从基于小规模数据分析的决策到基于大数据知识发现的决策, 决策理论与方法已经发生了巨大的变化[21], 基于大数据的智能决策逐渐成为新时代决策应用及研究的新生力量. 大数据智能决策就是用智能计算方法对大数据进行智能化分析与处理, 从中抽取结构化的知识, 进而对问题进行求解或对未来做出最优判断的过程. 该过程需要满足大数据决策在不确定性、动态性、全局性以及关联性上的分析需求.

在面向大数据的决策应用中, 关联分析为问题假设的初步分析以及正确数据选择提供必要的判定与依据, 它既是一个重要前提也是一种必要的分析手段; 不确定性是大数据决策的显著特征, 同时也是大数据智能决策研究的重点与难点; 大数据决策的动态性决定了大数据知识动态演化的重要性, 如何有效利用数据的增量性同样是大数据智能决策研究的关键点; 大数据决策追求的全局性, 要求大数据智能决策能够将多源信息进行融合与协同以消除信息孤岛. 需要指出的是, 大数据的关联性、不确定性、增量性和多源性不是相互独立的因素, 四者之间存在着潜在的联系, 在实际应用中可能并发存在, 但从研究的角度出发, 一般很难将上述四种因素的分析同时讨论. 此外, 智能决策支持系统是智能决策分析方法的载体, 随着大数据应用的普及, 智能决策支持系统的发展也是大数据决策领域备受人们关注的研究方向. 结合以上讨论, 本节将从智能决策支持系统、基于不确定性分析的智能决策、基于信息融合的智能决策、基于关联分析的智能决策和基于增量分析的智能决策五个方面展开对大数据智能决策研

究与发展现状的综述分析.

2.1 智能决策支持系统

决策支持是在管理科学和运筹学的基础上发展而来的一门学科, 20 世纪 70 年代, Scott-Morton 提出了决策支持系统 (Decision support system, DSS) 的概念[22]. DSS 是以提高决策有效性为目的, 综合利用大量数据, 有机地结合各种模型, 通过人机交互的方式, 辅助各级决策者实现科学决策的计算机系统. 1980 年, Sprague[23] 将 DSS 设计为由用户接口、数据库管理系统、模型库管理系统三部件集成的两库 (数据库和模型库) 框架. 随着人们对 DSS 研究和应用的深入, DSS 相继引入方法库管理系统、知识库管理系统和推理机并形成四库 (数据库、模型库、方法库和知识库) 框架. 经过几十年的发展, DSS 不断与新技术、新学科相互交叉融合, 并在体系结构、问题处理模式、功能模块集成等方面发生了巨大变化, 其应用也被推广到诸多领域.

智能决策支持系统 (Intelligent decision support system, IDSS) 是由 DSS 不断升级和演化得来. 20 世纪 80 年代, 专家系统 (Expert system, ES) 广泛流行, Bonczek 等[24] 将决策支持系统与专家系统相结合, 充分发挥 DSS 的数值分析能力和 ES 的符号知识的处理能力, 用于解决定量与定性问题以及半结构化、非结构化问题, 有效扩大了 DSS 处理问题的范围. 这种 DSS 与 ES 结合的思想即构成智能决策支持系统的初期模型. 智能决策支持系统利用人工智能和专家系统技术在定性分析和不确定推理上的优势, 以及人类在问题求解中的经验和知识, 为决策问题的求解提供了更加广阔的思路. 近年来, 几乎所有有关决策支持系统的研究都是围绕着人工智能技术的应用而展开的. 人工智能方法已经逐渐渗透到 IDSS 的体系结构、问题求解方法等各个方面. 综合来看, 智能决策系统的研究逐渐由过去的决策部件功能的扩展发展到部件的综合集成, 由过去的定量模型发展到基于知识的智能决策方法[25].

和许多正在发展中的事物一样, 智能决策支持系统是一个发展中的概念. 随着社会的发展, 信息量的激增, 管理、决策日趋复杂, 单纯依靠某一个决策者做出的决策往往不够完善, 于是 Gray 将群决策理论引入 DSS, 提出了群决策支持系统 (Group decision support system, GDSS) 的概念[26], 旨在吸收群体的经验和智慧, 实现群体对决策问题的共同求解. GDSS 为企业的组织决策提供一种开放与协同的决策环境, 达到提高决策质量的目的. GDSS 是智能决策支持系统的一个重要研究方向, 目前分布式环境下的 GDSS 和基于人工智能的群决策方法仍然是该领域的研究热点[27].

传统的 DSS 多采用静态模型, 决策过程需要用户自主选择方法和模型, 系统缺乏主动决策机制. 针对该问题, Manheim 等[28] 最早提出了主动决策支持系统 (Active DSS, ADSS) 的概念, 并给出了相应框架. ADSS 通过建立人类认知模型, 在决策问题求解的不同阶段, 给决策者提供不同的方法选择, 从而形成不同的问题求解路径. ADSS 是基于人类先验知识的, 但其前提假设是系统运行在静态的决策环境下, 因此在实际应用中 ADSS 仍然存在适应性较差的局限性. 不过人们对 ADSS 的研究为自适应决策支持的提出奠定了基础. 为了适应决策环境的变化, Shaw[29] 于 1993 年提出了自适应决策支持系统 (Adaptive decision support system, AdDSS) 框架, 并尝试用机器学习和案例推理等方法从大量历史数据和过往经验中发现与决策问题相关的知识, 以此来使系统具有随时间和决策过程变化调整自身行为的能力. 在此基础之上, 人们对 AdDSS 展开了大量的研究, 包括系统结构自适应、领域知识自适应、用户接口自适应等, 自适应性和自学习能力已经成为智能决策支持系统的一个主要标志.

互联网技术在决策支持领域的应用, 使得决策环境出现了新特点, 即决策分析中的数据不再集中于一个物理位置, 而是分散在不同部门或地区. 在此环境下许多大规模的管理决策活动已不可能或者不便于用集中方式进行, 而分布式决策支持系统 (Distribute decision support system, DDSS) 正是为适应这类决策问题而建立的信息系统. DDSS 将传统集中式 DSS 发展为网络环境下的分布式并行处理的方式[30], 通过网络连接工作平台和分布式数据库、模型库等, 支持分布在各地的 DSS 彼此交互, 从而使他们共同为决策问题求解提供高效及时的决策支持. 在大数据环境下分布式决策支持系统将得到更加广泛的关注, 分布式数据仓库、分布式人工智能、分布式并行化决策已经成为当下决策支持领域的重要研究方向.

随着智能体 (Agent) 在人工智能领域的深入研究, 相关学者将 Agent 技术引入了智能决策支持系统, 特别是多 Agent 理论与技术为分布式决策支持系统的分析、设计和实现提供了新的途径. Bui 和 Lee[31] 将决策支持系统中的 Agent 应具备的能力归纳为: 独立能力、学习能力、协作能力、推理能力、智能性等. 目前, 多 Agent 智能决策支持系统已经成为趋势, 通过加入诸如人机交互 Agent、模型选择 Agent、模型求解 Agent 等可以使决策系统减少对专家的依赖, 实现系统由 “模型驱动” 转为 “问题驱动”, 提高决策系统的整体智能性. Ghadimi 等[32] 提出一种面向供应链可持续供应商选择和订单分配的多 Agent 系统方法, 通过设计数据库 Agent、供

应商 Agent、决策者 Agent 和订单分配 Agent, 有效提高供应商选择和订单分配质量.

随着云计算 (Cloud computing) 技术兴起, 基于云计算的智能决策支持系统成为大数据智能决策支持的一个研究方向. 云计算通过互联网将虚拟化的数据中心和智能用户终端有机地联系起来, 为用户提供了便捷的信息服务环境. 在大数据环境下, 云计算平台可以为大数据的决策分析提供庞大的存储空间和强大的分布式并行计算能力. 决策环境的开放性、决策资源的虚拟化、问题求解的分布式协作性将使得基于云计算的智能决策有着与传统智能决策不同的特征[21]. 随着移动智能设备和移动互联网的普及, 分布式移动云计算环境下智能决策方法成为当前的一个研究热点[33].

随着社会节奏的加快, 企业或组织所面临的内外部环境更加复杂, 业务问题呈现非线性、不确定性、多维化和实时性等特点, 此时继续使用传统 IDSS 工具和利用局部数据进行决策分析的方法已经难以获取高质量的决策效果. 在大数据环境下, 智能决策支持系统应具备大数据的分析处理能力. 通过综合运用互联网、云平台和人工智能技术, 将大数据的采集、存储、管理、分析、共享、可视化等一系列知识发现技术与现有的智能决策支持技术深度融合, 构建形成基于大数据的智能决策支持系统是智能决策应用领域的发展方向. 未来基于大数据的决策支持系统有望具备海量数据汇聚融合能力、快速感知和认知能力、强大的分析与推理能力、自适应与自优化能力, 可以实现复杂业务的自动识别、判断, 并做出前沿性和实时性的决策支持.

2.2 基于不确定性分析的智能决策

不确定性是指客观事物联系与发展过程中无序的、随机的、偶然的、模糊的、粗糙的、近似的属性[34]. 现实世界的多样性、随机性、运动性, 以及人类对事物描述和信息表达的不精确性、模糊性决定了人们所能获取的数据本身存在着较多的不确定性. 而在大数据环境下, 数据的多源、多样、增量及不完备等特点, 加之人们对数据分析处理需求的多样性 (如数据融合等), 使得大数据从宏观上有着相较于传统数据更多的不确定性. 正如 Wang 等指出, 大数据的不确定性不仅存在于大数据本身, 还体现在大数据的处理过程当中[15]. 因此, 关于大数据不确定性信息的表示与处理成为大数据智能决策理论方法研究中不可缺少的一部分. 在不确定性理论方法中模糊集、粗糙集、贝叶斯理论、证据理论等在智能决策方法中都起到了关键作用. 随着大数据应用的增多, 以上方法也逐渐被用于面向大数据不确定性处理的智能决策当中. 本小节将从大数据不确定性处理的角度对相关方法进行回顾和综述.

模糊集于上世纪 60 年代由 Zadeh 提出, 通过隶属度函数表达模糊性概念, 其本身是一种有效的不确定性信息表示与处理方法. 目前模糊集方法已经形成一整套较为完整的理论体系, 包括模糊集、模糊逻辑、模糊系统以及它们的扩展形式[35]. 由于模糊集方法可以在不同信息粒度层次上对不确定性数据进行表示与处理, 因此具有较强的可解释性和可理解性. 模糊集在大数据中的应用, 形成对大数据不确定性的表示与处理的有效手段. 在面向大数据的聚类应用中, 模糊 C-means 算法 (FCM) 已经成为一种常用的软聚类方法. 文献 [36] 将 FCM 应用于机器人触觉感知数据的分析, 解决机器人触觉识别问题. Chang 等[37] 针对高维度数据聚类问题, 提出稀疏正则化 FCM 算法. Di Martino 等[38] 将 FCM 扩展应用于超大型事件数据集中的热点检测, 并进一步提出了一种时空 FCM 方法, 用于面向时空大数据的热点检测与预测问题[39]. 模糊规则分类系统广泛地应用于模式识别和分类任务, 可以为用户提供带有语义标签的可解释分类规则, 降低决策失误的风险. Jindal 等[40] 设计了云环境下的模糊规则分类器, 用于处理多源异构的远程医疗大数据, 实现对病人的远程实时诊断决策. 针对面向大数据的模糊分类, Segatori 等[41] 提出了基于 MapReduce 的分布式模糊决策树 (FDTs) 计算模型. 模糊推理系统还常与神经网络相结合, 以提高决策问题求解的自适应性. 在电力系统控制领域, 文献 [42] 将神经网络与模糊推理系统相结合, 提出三种自适应神经模糊推理系统, 用于太阳能发电企业控制决策中的短时电力预测问题. Jindal 等[43] 针对疾病诊断决策中的分类问题, 提出了用于医疗大数据维度约简的模糊神经分类器方法, 有效提高疾病诊断准确率. 更多有关模糊集在大数据决策方面的研究可以参见文献 [35]. 从现有的基于模糊集方法的大数据决策文献来看, 模糊集方法是适用于大数据不确定性分析的有力工具, 其数据表示的多粒度特性符合人类的认知习惯, 可以满足更多特定领域的大数据决策需求.

粗糙集由波兰数学家 Pawlak 于 1982 年提出. 粗糙集使用具有精确概念的上近似集和下近似集对一个不精确概念/知识进行近似表示与度量, 其独特之处在于不需要主观先验知识, 可以直接对数据进行分析与推理, 并揭示潜在规律. 目前, 粗糙集及其扩展理论已经成为处理不精确、不一致、不完备信息的有力工具, 并广泛用于数据挖掘、知识获取以及各类决策问题的求解. 为满足粗糙集方法的大数据决策分析需求, 已有较多学者从粗糙集的并行化开展了研究. 基于粗糙集的多粒度思想, Qian 等[44] 提出基于 MapReduce 的粗糙集的并行化层次属性约简

方法. Li 等[45] 设计了并行化优势粗糙集近似计算方法. 针对大数据常见的不完备特性, Abdel-Basset 等提出将中性集 (Neutrosophic sets) 和粗糙集相结合的方法来处理智慧城市大数据的不完备性问题[46]. El-Alfy 等基于遗传算法研究了面向决策粗糙集的大规模数据集的并行化属性约简方法, 并成功用于网络入侵检测[47]. Banerjee 等通过粗糙集理论和蚁群算法解决大数据中的不确定性和最优特征抽取分析问题, 提出了面向移动大数据的评价决策分析方法[48]. 针对大规模多模态数据的属性约简问题, Hu 等给出了多核模糊粗糙集方法[49]. 为降低多粒度决策粗糙集在大数据分析中的时间复杂度, 同时使其满足大数据的半监督特性, Qian 等提出了局部多粒度粗糙集方法[50−51].

近年来, 由决策粗糙集发展而来的三支决策理论[52] 成为一种更为一般化且符合人类认知的不确定性决策工具, 正受到越来越多的关注. 在基于 Web 的医疗决策支持系统中, Yao 等将博弈论粗糙集 (GTRS) 用于面向医疗数据的不确定性分析, 通过生成三支决策规则, 提高系统整体决策质量[53]. Yu 等研究了面向多视图数据的不确定性聚类问题, 并提出一种主动三支聚类方法[54]. Zhang 和 Yang 等[55] 基于区间值决策粗糙集提出一种三支群决策模型. 针对现实中有用信息随时间不断增长, Li 等[56] 提出了代价敏感序贯三支决策, 并将其应用于人脸识别. Qian 等[57] 基于多粒度思想, 提出一种更为一般化的多粒度序贯三支决策模型.

基于贝叶斯理论的方法已经在人工智能领域中的不确定性推理、计算机学习等方面取得了许多成果. 对于不同规模大小的贝叶斯网络, 可以分别采用精确推理和近似推理算法进行分析, 并提供决策支持. Lake 等[58] 通过一个基于贝叶斯的 BPL (Bayesian program learning) 模型来建模实现人类层次的概念学习. Sturlaugson 和 Sheppard[59] 研究了连续时间贝叶斯网络中的不确定推理. Abadpour[60] 利用贝叶斯推理构造了模糊可能性聚类算法的目标函数. 胡支军等[61] 研究发现对项目价值事前估计不确定性的贝叶斯建模可以在风险项目投资组合决策中给出更加精确的价值估计. Hao 等[62] 研究了不确定性环境下动态决策中的信息权重确定问题, 提出基于直觉模糊贝叶斯网络的动态属性权重确定方法, 同时构建了面向风险决策问题的动态直觉模糊决策概念框架. 贝叶斯网络同样适用于不完备数据的处理, Feng 等先后提出了不完备数据环境下基于贝叶斯网络的岩爆灾难预测方法[63] 和隧道挤压预测方法[64].

证据理论 (Dempster-shafer theory) 通过引入信任函数, 把不确定与不知道区分开来, 能够在先验概率未知的情况下, 以简单的推理形式, 得到较好的结果. 例如, Zhang 等[65] 采用证据推理方法研究了不确定环境下的多属性决策分析问题. Sun 和 Wang[66] 针对基于属性描述的知识, 通过组合证据来解决多属性融合问题. Troiano 等[67] 应用 D-S 证据理论挖掘用户的偏好信息用于推荐决策. 杜元伟等[68] 将头脑风暴方法中的基本原则引入到主观证据的提取过程之中, 并在此基础上结合证据理论提出了主观证据融合决策方法.

由于专家知识总是有限的, 并且能够以符号逻辑表示并用来推理的知识更为有限, 所以许多专家知识并不是一开始就已经具备, 更多的还是在决策过程中学习得到的. 因此, 人们将人工智能中的仿生方法引入到决策过程中, 并取得了很好的效果. 仿生方法是一类重要的人工智能方法, 能够适应现实环境中普遍的不确定性, 解决那些无法精确定义或建模的决策问题. 神经网络、进化算法、蚁群算法等均被用于对存在大量不确定性信息的学习, 并得到较好的决策效果. 例如, Bukharov 等[69] 基于神经网络和遗传算法构建了一个决策支持系统, 该系统采用区间神经网络来处理不确定数据, 使用遗传算法来选择最重要的输入. Yu 等[70] 结合与或图和粗糙集等方法将蚁群优化算法应用于属性约简、约简选择以及 Web 服务选择中.

此外, 概率推理、赋值代数、连接分析、聚类分析等方法也常常应用于不确定性决策分析中. 上述理论与方法为智能决策问题的求解提供了有力的支持, 但是有关不确定环境下面向复杂大群体决策等方面的求解方法仍然有待进一步的研究.

2.3 基于信息融合的智能决策

多源信息融合是人类所固有的一种基本功能. 人类可以本能地将各种感知器官所探测的信息与先验知识进行综合, 进而对周围的环境和正在发生的事件做出准确的估计. “盲人摸象” 的故事告知我们, 单凭一种感官获得的感知信息, 难以获得对客观事物的全面认知, 而通过对不同度量特征的融合处理可以将多源信息转化成对环境有价值的解释. 多源信息融合就是对人脑综合处理多源信息功能的模拟[71], 以实现自动的或半自动的将不同来源和不同时间点的信息转化为统一表示形式, 进而为人们提供有效决策支持的一系列技术方法[72].

在大数据环境下, 数据的分布式存储与交互式共享会更加普遍, 而具有分布式和分散控制的自治数据源是大数据应用的主要特征之一[16]. 此时, 多源信息融合是提升大数据价值不可或缺的技术手段. 从决策应用的角度来看, 社会经济活动中的企业或组织在决策时需要收集大量的数据, 汇集不同的观

点, 才能制定出符合客观规律的决策. 随着数据获取便利性的增加, 信息的全面性和多源信息的协同作用将更多地被人们关注, 而越来越多的决策任务的开展, 将寻求多源数据甚至是跨平台、跨区域、跨领域数据的参与. 例如, 在城市规划决策中, 政府部门需要结合路网结构、交通流量、城市人口分布以及 POIs 数据进行综合分析[73]; 在医疗诊断中, 有时专家需要将多家医疗机构的诊断结果进行融合分析; 在工业生产过程中, 可以借助火眼图像、槽音频以及其他监控数据来综合判断铝电解槽过热度状态[74]. 多源信息融合对于大数据决策的意义可以归纳为两方面: 一方面, 信息融合有利于进一步挖掘数据价值, 从众多分散、异构的数据源获取隐含价值信息, 丰富决策的内涵; 另一方面, 通过多源数据的交叉引证, 可以降低大数据潜在的噪音、数据缺失、信息不一致和语义模糊等不确定性因素[72], 提高决策的置信度.

简单来说, 信息融合是一种概念框架. 在不同需求和应用场景下, 信息融合所面对的问题不同, 人们提出的模型方法与技术手段也各不相同. 信息融合技术最早以多传感器数据融合 (Multi-sensor data fusion) 的概念出现在军事领域. 上世纪 70 年代美国国防部联合指挥实验室 (Joint Directors of Laboratories) 提出了颇具代表性的 JDL 模型[72], 旨在将来自不同源的数据信息进行多层面的融合处理, 来提高目标识别、身份评估、战况评估和威胁评估的准确性. 在此之后, 信息融合技术不断地被丰富和拓展, 并发展成为涉及信号处理、信息理论、统计学、人工智能、机器学习的多学科研究领域.

从信源之间的关系来看, 学者们把信息融合的类型划分为互补型、竞争型及合作型[75−76]. 互补型中的各信源互不依赖, 各信源感知目标/场景的不同方面, 通过信源融合来获取目标的全局信息; 竞争型中的各信源描述相同目标/场景的同一方面, 多源信息融合用于冗余校准和增强信任; 合作型中各信源之间相互依赖, 从不同角度感知目标, 多源信息融合用于获得全新的信息. 从信息融合的抽象层次来看, 人们常把融合划分为数据层融合、特征层融合及决策层融合[75]. 数据层融合也称作像素层或信号层融合. 由于数据层融合一般面向等价信源的数据[75], 因此其常用融合机制为竞争型. 数据层融合因尽可能多的保持了现场数据, 其具有信息损失小的优点, 但由于要对现场数据进行整体传输和集中处理, 导致其有通信负载大、计算代价高、处理时间长、抗干扰能力差的缺点. 决策层融合也称作语义层融合, 其操作对象是规则或知识. 决策层融合依赖于人们对数据特征意义和关系的理解, 是一种高层次的和更符合人类认知的融合方式. 由于决策层融合不受信源数据形式差异的限制, 使其融合机制也更加灵活, 它可以面向竞争型、合作型和互补型的融合需求. 由于决策层融合传输和处理的是规模较小的知识, 因此其具有通信负载小, 抗干扰能力强, 融合中心计算代价低的优点, 不过在各信源的知识获取阶段仍需花费一定的计算代价且产生一定的信息损失, 使得决策层融合存在信息损失相对较大且整体计算代价不一定会低的问题. 特征层融合的操作对象是从数据中抽取的特征属性, 常用融合机制有竞争型、互补型及合作型, 其优缺点介于数据层融合和决策层融合之间. Gravina 等[75] 总结了不同层次下数据融合对比情况, 见表 1.

表 1 不同层次下数据融合对比情况表

Table 1 Comparison of data fusion under different levels

融合层次	常用融合机制	优缺点			
		处理复杂度	信息损失	性能损失	传输负载
数据层	竞争型	高	否	否	高
特征层	竞争, 互补, 合作	中等	是	是	中等
决策层	竞争, 互补, 合作	低到高	是	是	低

在大数据时代, 信息来源更加广泛, 数据交互更加频繁, 大数据的多源分布现象普遍存在. 随着社会媒体网络、躯体传感网络、智能推荐系统、城市计算等新兴技术领域的崛起, 人们对数据融合技术的需求进一步加大. 然而复杂的大数据环境对信息融合任务的开展构成诸多挑战. 覃雄派等[77] 指出随着大数据的增长, 对大数据进行分析的基本策略是把计算推向数据, 而不是移动大量的数据. 吴信东指出大数据应用的自治数据源和分布式控制的特点使得整合多源数据进而集中式挖掘的方法会因传输代价高昂以及隐私暴露等问题而不可取[16]. 为实现对城市大规模人群聚集事件的有效预测, Huang 等[78] 通过对多源大数据的知识融合, 提出一种基于大数据融合的人群聚集预警方法. Lin 等[79] 基于邻域粒化的方法, 提出一种多信源决策规则表示方法, 进而通过一致性度量原则计算各信源权重, 实现多源决策规则的融合. Zheng[73] 指出大数据时代的信息融合任务会更多地面向跨领域数据. 然而跨领域数据在表示、分布、尺度上普遍存在的模态差异, 这对传统数据层融合方法构成巨大挑战. 虽然已有相关研究工作将深度神经网络 (Deep neural network, DNN) 用于多模态数据的统一特征表示[80−81], 并在一定程度上解决了多源数据特征层融合问题, 但是基于 DNN 的融合方法的效果取决于参数调整的好坏, 最优参数的寻找依然是一项耗时耗力的过程. 另外, 对于 DNN 中间层特征表示依然存在可解释性问题. 针对上述问题, Zheng[73] 提出跨领域大数据融合范式 (如图 1 所示), 即对各个数据源分别进行知识提

取, 在知识层面实现多源信息语义融合. 语义层的信息融合可以大体分为基于多视图的数据融合、基于相似性的数据融合、基于概率依赖的数据融合以及基于迁移学习的数据融合[73].

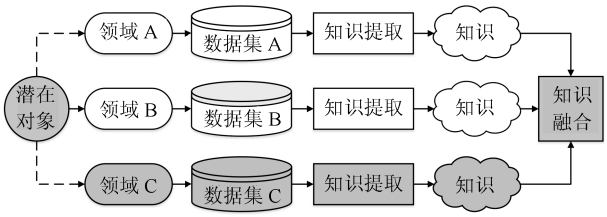

图 1 跨领域大数据融合范式[73]

Fig. 1 The paradigm of cross-domain big data fusion[73]

在大数据多源信息融合任务中, 如何对信源进行评价与选择同样是一项挑战性问题. Xu 等首次提出了使用内部信任度和外部信任度两个指标来评估信源的可靠性方法, 实现对冗余和不可靠信源的过滤, 并通过将原始数据转换为三角模糊信息粒, 实现基于粒计算的多源数据融合[82]. 但上述方法仅适用于多源同构数据集, 难以适应多源异构数据环境. 目前对信源的评价选择问题依然是信息融合领域的一个开放性研究课题. 多源数据信息潜在的不完备、不一致、冲突、语义模糊等不确定性是多源信息融合所要解决的最根本问题, 相关学者已尝试将概率论、粗糙集、模糊集、可能性理论以及 D-S 证据理论等应用到数据融合当中, 并分别在特定领域取得了较好的效果. Khaleghi 等对以上各种融合方法的优缺点做了详细分析, 读者可以参阅文献 [72].

2.4 基于关联分析的智能决策

在现实世界中, 诸多看似没有关系的事物之间其实存在有普遍关联, 而这些普遍关联往往在一些问题求解中起到关键作用. 相关分析便是一种发掘事物之间普遍关联的数据驱动方法. 自 19 世纪 80 年代 Galton 通过研究人类身高遗传问题首次提出 “相关” 概念以来[83], 相关分析便引起人们的关注, 并逐渐成为一种决策分析的重要手段. 作为度量事物之间协同关系和关联关系的有效方法, 大数据的相关分析能够满足人类的众多决策需求. 例如, Google 公司的趋势系统, 通过对互联网搜索数据的关联分析, 实时预测了 2009 年美国 H1N1 流感的爆发[19]. 沃尔玛通过对用户消费数据的关联分析, 发现啤酒与尿布间的关联关系. 需要特别指出的是, 相关关系有别于因果关系. 在大数据时代基于相关关系挖掘的数据分析具有重要的价值. 李国杰院士等指出, 对于简单封闭的系统, 基于小数据的因果关系分析是可行的, 但对于开放复杂的巨系统 (大数据环境), 传统的因果关系分析难以奏效[84]. 首先, 大数据环境下数据结构、数据关系错综复杂且存在很多噪音, 人们很难在变量间建立精确的函数关系并在此基础上探讨因果关系, 寻找因果关系的代价高昂; 其次, 大数据的动态与演化特性, 决定了变量间的因果关系具有时效性, 环境状态稍有变化, 探寻到的因果关系或已失效. 然而相关关系的要求较为宽松, 可以帮助人们更加快捷、高效地发现事物之间的内在关联.

从决策应用的角度来看, 大数据相关性分析对大数据智能决策的推动作用主要体现于以下两个方面. 一方面, 相关性分析技术不仅用于发现变量之间的潜在关联, 而且还用于判定分析变量之间伪相关、假关联. 试想, 通过对一组数据的回归分析, 可以学到一个精度较高的回归模型, 但如果数据之间是伪相关的, 那么学到的模型将导致错误的科学推断及毫无价值的预测结果. 在面向大数据智能化分析的决策应用中, 由于数据混杂且体量大, 如何选择与问题相关且正确的数据来开展分析是一项极为重要的问题. 在该环节, 相关性分析可以为问题假设的初步分析以及正确数据的选择, 提供必要的判定与依据. 在这一方面, 牛津大学 Mayer-Schonberger 教授等也给出了相同的观点: “建立在相关分析法基础上的预测才是大数据的核心”[85]. 另一方面, 在实际应用中, 相关性分析不是一个独立的环节, 而是需要将其与其他模型方法进行有机结合, 进而提高数据分析过程的合理性以及分析结果的认可度. 目前, 较多的数据挖掘与机器学习方法仍建立在数据的独立同分布假设之上, 显然独立同分布只是一种理想假设, 这样的分析结果存在较大的局限性且不能充分反映数据中蕴含的真实知识. 近年来, 为提高数据分析的合理性和准确性, 越来越多的学者将相关分析纳入到智能信息处理当中, 诸如多准则/属性决策[86−87]、分类[88]、聚类[89−90]、多标签学习[91−92] 等, 均取得了较好的效果. 综合来看, 大数据相关分析已经成为大数据智能决策中的一项关键应用技术.

传统相关分析中的相关系数法往往会忽视很多变量间隐含的逻辑关系, 难以对非线性相关关系和非函数相关关系进行准确测量, 这些局限性限制了传统相关分析法在处理大数据问题时的应用范围. 近年来, 相关学者从典型相关分析、基于互信息的相关分析、基于距离的相关分析展开了对非线性相关关系的研究, 此外在伪相关以及时序数据延迟相关方面也取得了较多研究成果. 以上几个方面对大数据相关性分析提供了理论依据, 下述内容是以上几点代表性研究成果的介绍.

目前典型相关分析 (Canonical correlation analysis, CCA) 已经较多地应用在大数据分析当中, 它不仅可以揭示大数据间的关联关系, 还可以提取

大数据中的低维特征. 具有代表性的应用有数据降维[93]、特征融合[94]、数据流挖掘[95]、跨模态检索[96]等. 在典型相关分析的非线性拓展方面, Yin[97] 基于互信息对 CCA 进行了扩展. Lai 和 Fyfe[98] 基于核方法提出了非线性 CCA. Hardoon 等[99] 使用 Kernel 典型相关分析方法来学习图片和问题描述之间的语义表示. 针对传统典型相关分析在大数据 PB 级数据规模时不再适应的情况, 杨静等[100] 提出一种基于云模型的大数据 CCA 方法.

互信息作为相关分析的度量, 其优势在于能有效地刻画变量之间的非线性关系[18], 能够有效探测数据的内在结构和规律, 因此在大数据相关分析中日益受到重视. Reshef 等[101] 通过互信息定义了最大信息系数 (Maximal information coefficient, MIC) 用来衡量两个变量间的相关性, 可以对变量间的非函数相关关系进行有效识别. MIC 被认为具有通用性和均等性, 并适用于大规模的数据集, 但由于其仅针对两个随机变量的相关分析, 因此在实际应用中还存在一定的局限性. Nguyen 等[102] 根据 MIC 方法, 提出了更为一般化的相关分析方法, 即最大相关分析 (Maximal correlation analysis, MAC), 扩展了 MIC 的应用范围, 实现对两组变量之间的非线性相关关系的准确测量.

基于距离的相关系数 (Distance correlation coefficient) 由 Székely 等于 2007 年提出[103], 可以提供比皮尔逊相关系数更多的信息. 基于距离的相关系数从特征函数的距离视角考察了两个随机向量之间的非线性相关关系, 为高维数据的非线性相关分析提供了有效的度量准则. Martínez-Gómez 等[104] 将基于距离的相关系数应用于高维巨量的天体物理数据集中, 用于发现变量之间的非线性关联关系, 从而实现特征的提取, 增强分类及模式识别的效果. Davis 等将基于距离的相关系数用于时间序列分析当中[105]. 基于距离的相关系数从特征函数视角构造相关性度量方法, 不但可以度量非线性相关性, 而且可以度量任意两个不同维度的随机向量的相关性. 但是, 距离相关系数涉及高维向量间的距离计算及矩阵点乘运算, 具有较高的时间复杂度. 如何提高计算效率是基于距离相关系数分析方法的未来研究方向[18].

时序数据的延迟相关性 (Lagged correlation) 是时间序列数据挖掘领域的一个重要研究内容. 延迟相关是时序数据之间普遍存在的现象. 例如, 国际原油价格走势常常会影响到国内成品油的价格行情, 但是这种相关性并不会立即表现出来, 而是存在一定的延迟. 在时间序列的相关性判定中, 既要判断数据之间是否存在时差 (也称作 “时间弯曲”), 又要考虑数据之间是否具有真实的相关性. 曲线排齐法 (Curve registration) 是对延迟序列进行矫正的常用方法. 经典的曲线排齐方法包括位移排齐法、特征点排齐法、连续单调排齐法等. 针对 BRAID 方法 (一种位移排齐法) 在最大延迟相关点较大时准确率不高的问题, 林子雨等[106] 提出了三点预测探查法 (TPFP), 该方法可有效处理最大延迟相关点位置较大的情形, 并可应对延迟突变问题. 姜高霞和王文剑[107] 构造了基于时间序列相关系数特征的相关性判定方法, 并基于光滑广义期望最大化算法提出一种基于相关系数最大化的曲线排齐模型. 针对基于采样的曲线排齐法中均匀采样存在的缺陷, 张文凯等[108] 提出了基于非均匀采样的相关系数最大化曲线排齐方法. 此外动态时间弯曲法 (Dynamic time warping) 也是时下较为流行的时移序列排齐方法[109].

伪相关 (Spurious correlation) 是指不具有相关关系的两组数据却具有较高样本相关系数的一种统计现象. 该现象将产生误导性的统计推断. 关于伪相关的产生原因, 学界普遍认为是由其他未见因素 (共有因素) 的影响而产生. 伪相关的判定问题和如何降低潜在伪相关的影响是相关分析应用中的重要问题, 并且多需要结合数据的背景知识来分析. 在生态系统研究当中, Baldocchi 等[110] 针对冠层光合作用和生态系统呼吸之间可能存在的潜在伪相关性, 通过改变数据汇总和集成的采样方法和时间尺度, 来验证不同采样方法对以上两者之间伪相关度的影响. 在基于元社区结构的物种分类研究当中, Clappe 等[111] 分析了由空间自相关 (独立发生) 引起的物种分布和空间环境之间的伪相关问题, 并基于空间约束空模型 (Spatially-constrained null model) 提出一种新的方差分解方法, 用于从环境数据中校准空间自相关带来的伪相关贡献. Gao 等[112] 提出一种新的两个非独立变量之间伪相关性的判定方法, 通过引入一个 “纯” 伪相关指标, 并将其与伪相关指标进行回归分析, 实现对区域径流悬沙年产量与径流深度之间伪相关性判定, 并进一步分析表明伪相关性受变量易变性的显著影响. 在大数据环境下, 数据的海量性、高维性、动态及不确定性等增加了发现伪相关的难度, 特别是大数据的高维特征将显著增加伪相关的可能性[113], 因此面向大数据的相关分析, 不可一味地追求对数据相关性探寻, 而忽略了对伪相关的分析与判断.

2.5 基于增量分析的智能决策

增量性是大数据的固有特性之一. 现实生活中广泛分布的传感与监控设备、实时互联的社会媒体等都构成了大数据动态增长的在线场景. 基于大数据决策的数据分析, 不单要从历史大数据中获取知

识, 更多的是要对新增数据进行动态知识发现. 传统机器学习方法对历史大数据的挖掘与分析往往是建立在数据隐含规律对未来预测有效性的假设之上, 或假定决策状态始终处于决策模型的闭环之内. 显然现实世界的复杂多变性决定了从历史数据中获取的知识多数只具备历史有效性, 在实用性较强的决策应用领域, 特别是对决策时效性要求较高的工业控制领域和智能交通领域等, 实时动态的增量式知识获取是保证决策质量的必要条件. 近年来随着大数据应用的普及, 更多专家学者开始关注大数据的增量式学习问题. 分类或聚类也是实现决策分析任务的常见方法. 在这类典型的基于机器学习的决策应用中, 增量性主要体现于三个方面: 一是数据样本的增量; 二是样本特征描述信息的增量; 三是类别的增量与数据分布的变化.

在数据样本增量方面. 针对以往增量式学习均假设新增样本是独立且同分布的, Xu 等[114] 研究了依赖采样方法对增量式支持向量机算法的影响, 并提出了一种基于马尔科夫重采样的增量式支持向量机算法 (MR-ISVM), 实现 ISVM 学习效率的显著提高. Gu 等[115] 基于代价敏感铰链损失的支持向量机 (CSHL-SVM) 构建了数据块增量式学习算法, 实现在线场景下的分类模型的快速更新. 粗糙集方法是处理不确定性数据的有效决策工具. 目前已经有专家学者基于粗糙集的决策方法进行了有关增量式知识发现的研究. Chen 等[116] 将变精度粗糙集方法引入集值序信息系统, 研究了变精度集值序信息系统下的近似集增量更新方法. 为应对决策信息系统中数据对象的动态增加问题, Li 等[117] 提出基于优势粗糙集的增量式近似集更新方法, 该方法可以有效解决多准则决策中的动态增量问题. 针对数据样本的增量以及数据中潜在的不确定性, Yu[118] 提出了三支聚类计算框架, 并进一步提出了基于树的增量式三支聚类模型, 该模型为不确定性大数据的增量式聚类计算提供了新思路. Hu 等[119] 通过粗糙集表示聚类问题中数据的不确定性, 形成一套基于粗糙集的增量式模糊聚类集成方法, 实现对不确定性数据的增量式聚类计算.

在样本特征描述信息的增量方面. Hu 等[120] 基于互信息的差异生成策略和特征增量树生长机制提出一种特征增量随机森林 (FIRF) 学习方法, 解决老年人健康护理中因传感器增加形成的数据特征增量问题. Huang 等[121] 在分布式信息系统下基于属性一般化提出了增量式粗糙近似集更新方法. Jing 等[122] 研究了多粒度视角下的知识粒表示方法, 针对大规模动态增量决策信息系统, 提出了多粒度增量式属性约简方法, 有效避免数据增加过程中对等价类的重复计算. 针对层次化多准则分类问题中属性值在不同粒度层次上的动态更新, Luo 等[123] 通过属性值分类对知识粒进行细化和粗化, 实现知识粒的动态特性的形式化表示, 并在此基础上提出了层次化多准则决策系统下的优势粗糙集增量式学习方法. 面向属性增量的聚类算法可以为基于无监督数据的决策活动提供有益帮助, 不过现阶段面向属性增量的聚类研究依然较少.

在类别的增量与数据分布的变化方面. 传统的增量式机器学习方法常假设训练数据和新增数据符合相同的模式, 却较少考虑新数据所属类别的增加与数据分布变化情况, 这使得传统增量式机器学习方法难以适应实际生产中的大数据环境. 现实中的诸多因素会导致模型在学习阶段只能接触到有限的数据类别, 而在测试和实际应用阶段的数据却包含了在学习阶段未曾出现的类别. 该类场景下的学习问题被称作开集学习 (Open-set learning) 问题, 意在寻求对已知类识别的同时, 能有效识别未知新类. Da 等[124] 尝试从无标签数据中获取更多分类信息, 并基于支持向量机的大边缘准则和半监督学习中的低密度分离器技术, 提出了基于无标签数据增广类学习框架及相应的支持向量机方法, 用于开放空间下的样本预测. Ristin 等[125−126] 基于随机森林算法提出了最近类平均森林算法和支持向量机森林算法, 研究了大规模图像分类中数据类别增加的增量式学习问题. Júnior 等[127] 将最近邻分类器扩展应用到开集学习当中, 提出一种开集最近邻方法. 在基于神经网络的图形识别领域, 使用数据集增广技术是应对开集识别问题的一种方法, Neal 等[128] 提出一种反事实图像生成的数据增广方法, 并通过训练后的生成对抗网络生成开集训练样本, 用于对开集图像识别任务的学习. 通过在深度网络中引入新的模型层 OpenMax 并结合元识别 (Meta-recognition) 算法来估计未知新类的概率, Bendale 和 Boult[129] 提出一种深度网络开集识别方法, 实现深度网络对高置信度欺骗图像以及相似于训练样本的对抗图像的识别. 分布外图像检测问题同样可看作是一类特殊的开集学习问题. Liang 等[130] 针对基于神经网络的分布外图像检测问题, 提出了基于神经网络的分布外检测器, 通过控制温度标定并结合添加输入扰动的方法, 增加分布内图像与分布外图像之间的 Softmax 分数间隔, 在无需重新训练网络的情况下, 有效降低分布外样本的误判率. 目前已有的开集学习的研究主要关注了如何检测到新类, 然而如何进一步区分新类同样具有重要的实际意义和研究价值, 目前该方面的研究还较少.

在流式数据和时间序列数据的实时处理任务中, 如何将新增数据的分布变化纳入学习任务当中是一项极其重要的研究工作. 基于概念漂移 (Concept

drift) 的增量式学习方法是应对上述问题的有效途径之一. Ahmad 等[131] 将概念漂移方法用于流式数据的非监督学习当中, 有效提高了在线异常检测的精度. 针对传统 DSS 中的静态数据分析方法在发生概念漂移时无法做出正确决策的问题, Dong 等[132] 研究了数据驱动决策支持系统中的概念漂移问题, 提出一种基于数据分布的概念漂移检测方法, 为数据流提供更好、更精细的经验分布, 使得 DSS 可以在适当的时间调整决策知识以适应不断变化的环境. Lobo 等[133] 使用核密度估计构建了一种进化多样化生成方法, 用于在线学习中概念漂移后学习策略的快速适应.

3 挑战问题与发展趋势

诚然大数据可以为人们带来更加科学全面的决策支持, 但大数据智能决策的应用研究还处于初期阶段, 并仍面临诸多挑战. 在此, 我们讨论大数据智能决策面临的一些问题挑战, 并指出潜在的应对方法或未来的发展趋势.

3.1 大数据多样性带来的挑战

多样性是构成大数据复杂性的主要因素之一, 也是大数据智能决策面临的主要困难. 当一项综合决策需要整合多方面数据时, 不同来源的大数据在类型、分布、频率及密度上可能各不相同, 这对多源大数据融合分析、多源信息协同决策等构成巨大的挑战. 现阶段对于处理大数据的多源异构性, 已经有一些研究成果, 但多数还是面向具体场景和特定一些类型的大数据. 解决多源异构大数据的协同分析问题, 消除信息孤岛进而实现通用性、鲁棒性更好的大数据智能决策, 是目前大数据智能决策的一个关键性研究课题.

多源大数据之间的关系普遍为互补型或合作型, 通过数据层面的融合决策不一定那么有效. 目前在特征层实现异构数据的融合方法中, 有很多基于 DNN 的优秀成果. 然而, 基于 DNN 的方法只克服了多样性中的数据类型多样, 而对于分布、频率等多样性还无法应对. 需要指出的是, 任何决策都是有风险代价的, 数据分析过程的可解释性对于决策者而言至关重要, 然而可解释性却是 DNN 的短板. 基于粒计算的 DNN 可解释性研究可望成为大数据智能分析的一个潜在研究方向.

通过语义层/决策层实现多源数据的综合利用是解决数据异质性较好的方法, 可以有效避免各种异质性问题. 在大数据环境下, 分布式自治数据源是大数据应用的一大特点[16], 去中心化将成为一大趋势. 通过分布式知识获取与协同的方法可以有效实现多源异构数据的协同感知与交互. 所谓协同, 可解释为对不一致信息的冲突分析. 研究基于粗糙集、模糊集和群体智能决策的冲突分析方法如何应用到大数据决策是未来的一个发展方向.

3.2 大数据动态性带来的挑战

日益加快的人、机、物之间的交互活动, 使得数据的快速增长成为大数据显著特性之一. 从决策需求的及时性和准确性来看, 大数据的动态性对现有的增量式机器学习方法构成巨大的挑战. 例如, 在流式数据处理中, 如何在发生概念漂移时及时调整数据分析策略并实现知识库的自适应更新, 仍是一项挑战性的研究任务.

针对大数据动态增量问题, 可以考虑形成一个训练学习、执行预测、漂移检测、漂移理解、漂移自适应的多步骤自适应学习模型. 这类模型的重点和难点在于漂移理解与漂移自适应. 在漂移理解方面可以融入高层次的、符合认知的方法, 可以采用粗糙集、模糊集、商空间等粒计算方法建立不同粒度层次下的漂移认知模型, 实现符合人类认知的层次化概念漂移理解. 针对漂移自适应问题, 可以通过构建有效的知识距离度量方法来度量概念漂移距离与方向, 同时综合运用进化计算与神经网络等方法构建与问题相符的参数自适应模型, 实现对学习模型的演化更新.

3.3 大数据极弱监督性带来的挑战

大数据的快速增长性也决定了大数据的极弱监督性甚至是非监督性. 大数据分类学习中的极弱监督性带来的问题通常表现在两个方面: 一是因标记稀缺而不能正确详尽地反映出整体数据集的特点, 导致学到的学习器泛化能力弱. 二是标记稀缺使得构造多分类器时多样性不足, 导致集成学习不能奏效. 大数据的极弱监督性决定了以聚类算法为特点的无监督学习方法在大数据增量问题上的研究具有巨大的决策应用价值. 不过大数据的增量性不仅体现于数据样本的增加, 还体现于属性的增加. 针对大数据属性增量式聚类问题仍然缺乏有效的方法.

针对大数据的极弱监督性, 可以充分利用多视角信息、相似领域信息、先验知识等, 采用大数据耦合与关联分析、大数据与经验知识相融合等技术增加额外的监督信息. 三支决策[134] 体现了一种渐进决策的思想. 我们可以设计三支聚类模型逐步地、有效地利用少量标签信息或者领域专家知识. 根据数据类型与问题求解需求, 采用合适的粒计算方法构建多粒度聚类分析算法模型, 也可望为大数据属性增量式聚类带来新的解决思路.

3.4 大数据不确定性带来的挑战

不确定性是当前人工智能技术研究中的关键问

题, 同时也是贯穿于大数据智能决策整个过程的核心问题. 目前研究较多的就是获取大数据中的不确定性知识. 然而不确定性的形式众多, 难以用统一的形式化方法表达, 也无法凭单一的技术手段来获取大数据中的不确定性知识. 不确定性知识发现的研究难度大、价值高, 一直是各领域知识发现研究所面临的核心困难问题.

要实现复杂数据中不确定性知识的高效获取, 需要为描述不确定性概念知识提供合适的数学模型, 建立不确定性知识空间中的计算模型, 实现对不确定性知识空间的认知和理解, 进而从数据中高效动态获取满足约束要求的知识. 粒计算[135−136] 是一种基于认知科学的智能信息计算范式, 它适用于近似求解具有不确定性和层次结构的问题, 可以达到对问题的简化、提高问题求解效率等目的. 从多粒度计算的角度来看, 不确定性和确定性是信息在不同粒度层面上的不同表示形式, 在某一层次上的不确定性问题可能是其他层次上的确定性问题[34]. 通过研究大数据在不同粒度层次上的粒度寻优与粒度切换方法可望实现对不确定性信息的有效处理. 此外, 粒计算往往从实际问题的需求出发, 用可行的满意近似解替代精确解, 提高问题求解效率.

3.5 大数据隐私问题带来的挑战

目前大数据隐私保护问题已经被广泛关注[16]. 诸如企业供应链数据、银行交易数据、患者医疗数据、导航用户轨迹数据等均构成了隐私保护的敏感信息范畴. 大数据的应用过程中往往不可避免地触及到敏感数据的传输、交互与分析处理, 特别是在跨平台、跨企业、跨领域数据的决策分析中, 用户隐私数据暴露问题显得尤为突出. 现阶段, 由于缺乏有效的隐私保护手段, 多数拥有数据的企业不愿或不能将数据公开, 这在很大程度上放慢了大数据研究与应用的落地. 目前, 有关大数据应用中的隐私保护还没有标准化的处理手段, 在技术层面和管理层面数据隐私都面临严峻的挑战. 有学者提出通过制定数据访问与分享的隐私保护策略, 比如设置一定的访问资格和权限, 或采用匿名数据的方法[137]. 对于设置数据访问权限的方法, 难点在于对安全认证和访问控制机制的设计以及对用户信用的把握; 而采用匿名数据的方法, 将显著增加数据的不确定性, 为数据分析带来更多困难[7, 16].

隐私信息一般是以最细粒度原始数据的形式存在的[136]. 根据粒计算的观点, 数据是知识在最细粒度上的表现, 知识是数据在不同粒度层次上的抽象[138−139]. 面对复杂大数据, 数据、信息、知识都可以被粒化, 并映射到不同的粒度层次上. 此时的计算单元从原有的最细粒度的 “数据” 转变为具有认知特征的、规模较小的 “知识粒”, 经过粒化后的知识粒隐藏了细节信息, 从而可以实现大数据隐私信息有效保护.

3.6 特例状况带来的挑战

众所周知, 基于机器学习的大数据智能化分析处理方法本质上是建立在对大数据的统计分析基础之上的. 在完全信息环境下, 通过对大数据的智能化分析可以很好地预测、判断大数据已经覆盖的事物状态, 但现实中的决策环境多是开放性的, 事物的状态是千变万化的, 因此即便是经过长期积累的大数据也无法保证信息的完全性. 比如在航天领域中的故障、交通行业中的事故等多数都是特例. 并且由于现实条件的约束人们往往无法通过反复试验的方法来获取覆盖各种特例的大数据进行学习预测, 因此对特例状况的预测和判断是实际应用中的一大挑战.

对于特例状况的学习, 有望借助平行系统和平行学习的方法进行解决. 平行系统的概念是由中科院自动化所王飞跃研究员于 2004 年提出[140], 通过利用大型计算模拟、预测并诱发引导复杂系统现象, 构建一种软件定义的人工系统[141]. 平行学习利用计算实验方法进行预测学习, 通过人工系统, 依据原始 “小数据” 生成大量的人工合成数据. 将人工合成数据与原始的小数据一起构成解决问题的所需的 “大数据”, 通过学习提取, 得到应用于某些具体场景或任务的知识, 进而用于平行控制和平行决策. 平行系统和平行学习可以满足人们对特例状况模拟与预测学习的需求, 在一定程度上可能会解决特例对大数据智能决策带来的挑战.

3.7 大数据认知困难带来的挑战

从本质上讲, 决策活动是人类的一种认知活动, 认知过程是所有决策过程的共性. 现阶段的人工智能技术与机器学习方法对于大数据的处理以及知识的获取多数还处于对事物的感知层面, 如特征提取, 模式识别、预测、回归、聚类等, 它们在实质上都是对事物的分类认知. 然而分类仅是人类的一种低层次认知, 其功能本质在于对事物的区分、辨别与归类. 单纯依靠对事物的分类还不足以构成一项完整的决策. 决策是任务和需求驱动的问题求解过程, 需要决策者在的分类认知的基础之上, 继续赋予研究对象以价值尺度认知或功能偏好认知, 并最终做出选择的全过程. 让机器拥有意识和理解能力才是人工智能最根本的目标, 在这方面人工智能刚走出了决策认知的第一步 (即分类认知), 而偏好认知还多依赖于人的参与. 在实际应用中, 只有不断提高对大数据快速的、完整的认知能力, 才能实现高效及时的大数据智能决策.

陈纯院士指出, 当前大数据智能正从规则的学习推理方法, 到数据驱动的知识挖掘方法, 迈向数据驱动与知识引导的新时代. 将数据驱动的机器学习方法与人类的常识先验与隐式直觉相结合, 可以实现可解释、更鲁棒和更通用的人工智能[9]. 郑南宁院士指出, 由于人类生活环境的高度不确定性和脆弱性以及面临问题的开放性, 任何智能程度的机器都无法完全取代人类, 因此有必要将人类的认知能力或类人认知模型引入人工智能系统, 形成混合增强智能形态[142]. 张钹院士指出, 人类在问题求解中具有天生的知识驱动能力、对不确定性问题的处理优势和对全局整体的感知能力; 传统机器学习具有在数据分析处理中的数据驱动能力、高速计算能力, 二者结合是未来信息处理的发展趋势[143]. 因此, 人机结合的智能形态有望构造出更加有效的认知计算方法. 在今后的大数据智能决策的应用与研究中, 人机结合的增强智能有望实现对大数据更加有效的处理, 并创造出更好的结果.

4 结束语

在全球信息化快速发展的背景下, 大数据以其蕴含的巨大价值正受到社会各界的广泛关注. 发展基于大数据的人工智能新技术, 实现基于大数据的智能决策是推动发展智能经济、智能服务、智能制造的关键手段. 现阶段, 智能决策理论方法在大数据驱动的模式下快速发展, 并逐渐形成一系列围绕多源异构大数据智能化处理的新方法和新趋势. 为了深入了解大数据智能决策的发展现状, 文章对大数据的特性以及大数据决策的特点进行了归纳总结, 并着重从智能决策支持系统的发展、不确定性信息处理、信息融合、关联分析以及增量分析五个方面综述了当前大数据智能决策的发展现状. 最后文章讨论了大数据智能决策仍然面临的问题与挑战, 展望了一些潜在的方法及研究方向. 作为一门快速发展的开放性学科领域, 大数据智能决策在内涵外延、模型理论、技术方法及实施策略等方面还需要人们继续投入更多的研究与实践. 希望本文对大数据智能决策的相关介绍与探讨能够对读者提供有益的借鉴和帮助.

References

1 Big Data. *Nature* [Online], available: http://www.nature.com/news/specials/bigdata/index.html, April 12, 2019.

2 World Economic Forum. Big data, big impact: new possibilities for international development [online], available: http://www3.weforum.org/docs/WEF_TC_MFS_BigData BigImpact_Briefing_2012. pdf, April 12, 2019.

3 United Nations Global Pulse. Big data for development: opportunities and Challenges–White Paper [online], available: http://www.unglobalpulse. org/sites/default/files/BigDataforDevelopment-UNGlobalPulse June2012.pdf, April 12, 2019.

4 Tolle K M, Tansley D S W, Hey A J G. The fourth paradigm: data-intensive scientific discovery. *Proceedings of the IEEE*, 2011, **99**(8): 1334−1337

5 Zhu K P, Joshi S, Wang Q G, Hsi J F Y. Guest editorial special section on big data analytics in intelligent manufacturing. *IEEE Transactions on Industrial Informatics*, 2019, **15**(4): 2382−2385

6 Yang Shan-Lin, Zhou Kai-Le. Management issues in Big Data: the resource-based view of Big Data. *Journal of Management Sciences in China*, 2015, **18**(5): 1−8
(杨善林, 周开乐. 大数据中的管理问题: 基于大数据的资源观. 管理科学学报, 2015, **18**(5): 1−8)

7 Hubbard D W. *How to Measure Anything: finding the Value of "Intangibles" in Business*. New Jersey: Wiley, 2010.

8 Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. *Big Data*, 2013, **1**(1): 51−59

9 Chen Chun, Zhuang Yue-Ting. Big data intelligence: from data to knowledge and decisions, *Fortune World*, 2017, (8): 48−49
(陈纯, 庄越挺. 大数据智能: 从数据到知识与决策. 中国科技财富, 2017, (8): 48−49)

10 Gao Ying-Mai. Industrial big data value mining path. *China Industry Review*, 2015, (2): 21−27
(高婴劢. 工业大数据价值挖掘路径. 中国工业评论, 2015, (2): 21−27)

11 Chen C L P, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Information Sciences*, 2014, **275**(11): 314−347

12 Industrial Big Data White Paper (2019 edition) [online], available: http://www.cesi.cn/201904/4955.html, April 1, 2019 (in Chinese)
(工业大数据白皮书 (2019 版) [online], available: http://www.cesi.cn/201904/4955.html, 2019 年 4 月 1 日)

13 Wu Xin-Dong, He Jin, Lu Ru-Qian, Zheng Nan-Ning. From big data to big knowledge: HACE +BigKE. *Acta Automatica Sinica*, 2016, **42**(7): 965−982
(吴信东, 何进, 陆汝钤, 郑南宁. 从大数据到大知识: HACE + BigKE. 自动化学报, 2016, **42**(7): 965−982)

14 Liu Qiang, Qin Si-Zhao. Perspectives on big data modeling of process industries. *Acta Automatica Sinica*, 2016, **42**(2): 161−171
(刘强, 秦泗钊. 过程工业大数据建模研究展望. 自动化学报, 2016, **42**(2): 161−171)

15 Wang X Z, He Y L. Learning from uncertainty for Big Data: future analytical challenges and strategies. *IEEE Systems, Man, and Cybernetics Magazine*, 2016, **2**(2): 26−31

16 Wu X D, Zhu X Q, Wu G Q, Ding W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(1): 97−107

17 Xie Xin-Shui. Multiple values, big data and coping strategies against decision-making uncertainty. *Journal of Beijing Technology and Business University (Social Sciences)*, 2014, **29**(6): 109−114
(谢新水. 多元价值、大数据与决策不确定性的应对策略. 北京工商大学学报 (社会科学版), 2014, **29**(6): 109−114)

18 Liang Ji-Ye, Feng Chen-Jiao, Song Peng. A survey on correlation analysis of big data. *Chinese Journal of Computers*, 2016, **39**(1): 1−18
(梁吉业, 冯晨娇, 宋鹏. 大数据相关分析综述. 计算机学报, 2016, **39**(1): 1−18)

19 Ginsberg J, Mohebbi M H, Patel R S, Brammer L, Smolinski M S, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*, 2009, **457**(7232): 1012−1014

20 Böttger T, Cuadrado F, Tyson G, Castro I, Uhlig S. Open connect everywhere: a glimpse at the internet ecosystem through the lens of the netflix CDN. *ACM SIGCOMM Computer Communication Review*, 2018, **48**(1): 28−34

21 Luo He, Yang Shan-lin, Ding Shuai. A survey of intelligent decisions in cloud computing. *Journal of Systems Engineering*, 2013, **28**(1): 134−142
(罗贺, 杨善林, 丁帅. 云计算环境下的智能决策研究综述. 系统工程学报, 2013, **28**(1): 134−142)

22 Scott-Morton M S. *Management Decision Systems: Computer Based Support for Decision Making*. Boston: Harvard University, 1971. 30−80

23 Sprague Jr R H. A framework for the development of decision support systems. MIS quarterly, 1980: 1−26

24 Bonczek R H, Holsapple C W, Whinston A B. The evolving roles of models in decision support systems. *Decision Sciences*, 1980, **11**(2): 337−356

25 Ren Ming-Lun, Yang Shan-Lin, Zhu Wei-Dong. Intelligent decision support system: state of art and challenges. *Journal of Systems Engineering*, 2002, **17**(5): 430−440
(任明仑, 杨善林, 朱卫东. 智能决策支持系统: 研究现状与挑战. 系统工程学报, 2002, **17**(5): 430−440)

26 Gray P. Group decision support systems. *Decision Support Systems*, 1987, **3**(3): 233−242

27 Liang D C, Liu D, Kobina A. Three-way group decisions with decision-theoretic rough sets. *Information Sciences*, 2016, **345**: 46−64

28 Manheim M L. An architecture for active DSS. In: Proceedings of the 21st Annual Hawaii International Conference on System Sciences. Kailua-Kona, USA: IEEE, 1988, 3: 356−365

29 Shaw M J. Machine learning methods for intelligent decision support An introduction. *Decision Support Systems*, 1993, **10**(2): 79−83

30 Mayer M K. Future trends in model management systems: parallel and distributed extensions. *Decision Support Systems*, 1998, **22**(4): 325−335

31 Bui T, Lee J. An agent-based framework for building decision support systems. *Decision Support Systems*, 1999, **25**(3): 225−237

32 Ghadimi P, Toosi F G, Heavey C. A multi-agent systems approach for sustainable supplier selection and order allocation in a partnership supply chain. *European Journal of Operational Research*, 2018, **269**(1): 286−301

33 Shi Y, Chen S Z, Xu X. MAGA: a mobility-aware computation offloading decision for distributed mobile cloud computing. *IEEE Internet of Things Journal*, 2018, **5**(1): 164−174

34 Wang Guo-Yin, Zhang Qing-Hua, Ma Xi-Ao, Yang Qing-Shan. Granular computing models for knowledge uncertainty. *Journal of Software*, 2011, **22**(4): 676−694
(王国胤, 张清华, 马希骜, 杨青山. 知识不确定性问题的粒计算模型. 软件学报, 2011, **22**(4): 676−694)

35 Wang H, Xu Z S, Pedrycz W. An overview on the roles of fuzzy set techniques in big data processing: trends, challenges and opportunities. *Knowledge-Based Systems*, 2017, **118**: 15−30

36 Liu C F, Huang W B, Sun F C, Luo M N, Tan C Q. LDS-FCM: A linear dynamical system based fuzzy c-means method for tactile recognition. *IEEE Transactions on Fuzzy Systems*, 2019, **27**(1): 72−83

37 Chang X Y, Wang Q N, Liu Y W, Wang Y. Sparse regularization in fuzzy c-means for high-dimensional data clustering. *IEEE Transactions on Cybernetics*, 2017, **47**(9): 2616−2627

38 Di Martino F, Sessa S. Extended fuzzy C-means hotspot detection method for large and very large event datasets. *Information Sciences*, 2018, **441**: 198−215

39 Di Martino F, Pedrycz W, Sessa S. Spatiotemporal extended fuzzy C-means clustering algorithm for hotspots detection and prediction. *Fuzzy Sets and Systems*, 2018, **340**: 109−126

40 Jindal A, Dua A, Kumar N, Vasilakos A V, Rodrigues J J P C. An efficient fuzzy rule-based big data analytics scheme for providing healthcare-as-a-service. In: Proceedings of the 2017 IEEE International Conference on Communications. Paris, France: IEEE, 2017. 1−6

41 Segatori A, Marcelloni F, Pedrycz W. On distributed fuzzy decision trees for big data. *IEEE Transactions on Fuzzy Systems*, 2018, **26**(1): 174−192

42 Jayawardene I, Venayagamoorthy G K. Comparison of adaptive neuro-fuzzy inference systems and echo state networks for PV power prediction. *Procedia Computer Science*, 2015, **53**: 92−102

43 Jindal A, Dua A, Kumar N, Das A K, Vasilakos A V, Rodrigues J J P C. Providing healthcare-as-a-service using fuzzy rule based big data analytics in cloud computing. *IEEE Journal of Biomedical and Health Informatics*, 2018, **22**(5): 1605−1618

44 Qian J, Lv P, Yue X D, Liu C H, Jing Z J. Hierarchical attribute reduction algorithms for big data using MapReduce. *Knowledge-Based Systems*, 2015, **73**: 18−31

45 Li S Y, Li T R, Zhang Z X, Chen H M, Zhang J B. Parallel computing of approximations in dominance-based rough sets approach. *Knowledge-Based Systems*, 2015, **87**: 102−111

46 Abdel-Basset M, Mohamed M. The role of single valued neutrosophic sets and rough sets in smart city: imperfect and incomplete information systems. *Measurement*, 2018, **124**: 47−55

47 El-Alfy E S M, Alshammari M A. Towards scalable rough set based attribute subset selection for intrusion detection using parallel genetic algorithm in MapReduce. *Simulation Modelling Practice and Theory*, 2016, **64**: 18−29

48 Banerjee S, Badr Y. Evaluating decision analytics from mobile big data using rough set based ant colony. *Mobile Big Data*. Cham: Springer, 2018. 217−231

49 Hu Q H, Zhang L J, Zhou Y C, Pedrycz W. Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets. *IEEE Transactions on Fuzzy Systems*, 2018, **26**(1): 226−238

50 Qian Y H, Liang X Y, Lin G P, Guo Q, Liang J Y. Local multigranulation decision-theoretic rough sets. *International Journal of Approximate Reasoning*, 2017, **82**: 119−137

51 Qian Y H, Liang X Y, Wang Q, Liang J Y, Liu B, Skowron A, et al. Local rough set: a solution to rough data analysis in big data. *International Journal of Approximate Reasoning*, 2018, **97**: 38−63

52 Luo C, Li T R, Huang Y Y, Fujita H. Updating three-way decisions in incomplete multi-scale information systems. *Information Sciences*, 2019, **476**: 274−289

53 Yao J T, Azam N. Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets. *IEEE Transactions on Fuzzy Systems*, 2015, **23**(1): 3−15

54 Yu H, Wang X C, Wang G Y, Zeng X H. An active three-way clustering method via low-rank matrices for multi-view data. *Information Sciences*, 2020, 507: 823−839

55 Zhang H Y, Yang S Y. Three-way group decisions with interval-valued decision-theoretic rough sets based on aggregating inclusion measures. *International Journal of Approximate Reasoning*, 2019, 110: 31−45

56 Li H X, Zhang L B, Huang B, Zhou X Z. Sequential three-way decision and granulation for cost-sensitive face recognition. *Knowledge-Based Systems*, 2016, **91**: 241−251

57 Qian J, Liu C H, Miao D Q, Yue X D. Sequential three-way decisions via multi-granularity. *Information Sciences*, 2020, 507: 606−629.

58 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. *Science*, 2015, **350**(6266): 1332−1338

59 Sturlaugson L, Sheppard J W. Uncertain and negative evidence in continuous time Bayesian networks. *International Journal of Approximate Reasoning*, 2016, **70**: 99−122

60 Abadpour A. Rederivation of the fuzzy-possibilistic clustering objective function through Bayesian inference. *Fuzzy Sets and Systems*, 2016, **305**: 29−53

61 Hu Zhi-Jun, Peng Fei, Li Zhi-Xia. Bayesian evaluation and selection strategies in venture project portfolio decision analysis. *Chinese Journal of Management Science*, 2017, **25**(2): 30−39
(胡支军, 彭飞, 李志霞. 风险项目投资组合决策的贝叶斯评价与选择策略. 中国管理科学, 2017, **25**(2): 30−39)

62 Hao Z N, Xu Z S, Zhao H, Fujita H. A dynamic weight determination approach based on the intuitionistic fuzzy bayesian network and its application to emergency decision making. *IEEE Transactions on Fuzzy Systems*, 2018, **26**(4): 1893−1907

63 Li N, Feng X D, Jimenez R. Predicting rock burst hazard with incomplete data using Bayesian networks. *Tunnelling and Underground Space Technology*, 2017, **61**: 61−70

64 Feng X D, Jimenez R. Predicting tunnel squeezing with incomplete data using Bayesian networks. *Engineering Geology*, 2015, **195**: 214−224

65 Zhang M J, Wang Y M, Li L H, Chen S Q. A general evidential reasoning algorithm for multi-attribute decision analysis under interval uncertainty. *European Journal of Operational Research*, 2017, **257**(3): 1005−1015

66 Sun L, Wang Y Z. A multi-attribute fusion approach extending Dempster-Shafer theory for combinatorial-type evidences. *Expert Systems with Applications*, 2018, **96**: 218−229

67 Troiano L, Rodríguez-Muñiz L J, Díaz I. Discovering user preferences using Dempster-Shafer theory. *Fuzzy Sets and Systems*, 2015, **278**: 98−117

68 Du Yuan-Wei, Duan Wan-Chun, Huang Qing-Hua, Yang Na. Decision making method for integrating subjective evidences based on brain storming principles. *Chinese Journal of Management Science*, 2015, **23**(3): 130−140
(杜元伟, 段万春, 黄庆华, 杨娜. 基于头脑风暴原则的主观证据融合决策方法. 中国管理科学, 2015, **23**(3): 130−140)

69 Bukharov O E, Bogolyubov D P. Development of a decision support system based on neural networks and a genetic algorithm. *Expert Systems with Applications*, 2015, **42**(15−16): 6177−6183

70 Yu H, Zhou Q F, Liu M. A dynamic composite web services selection method with QoS-Aware based on AND/OR graph. *International Journal of Computational Intelligence Systems*, 2014, **7**(4): 660−675

71 Luo Jun-Hai, Wang Zhang-Jing. *Multi-Source Data Fusion and Sensor Management*. Beijing: Tsinghua University Press, 2015
(罗俊海, 王章静. 多源数据融合和传感器管理. 北京: 清华大学出版社, 2015.)

72 Khaleghi B, Khamis A, Karray F O, Razavi S N. Multisensor data fusion: a review of the state-of-the-art. *Information Fusion*, 2013, **14**(1): 28−44

73 Zheng Y. Methodologies for cross-domain data fusion: an overview. *IEEE Transactions on Big Data*, 2015, **1**(1): 16−34

74 Chen Z G, Li Y G, Chen X F, Yang C H, Gui W H. Semantic network based on intuitionistic fuzzy directed hypergraphs and application to aluminum electrolysis cell condition identification. *IEEE Access*, 2017, **5**: 20145−20156

75 Gravina R, Alinia P, Ghasemzadeh H, Fortino G. Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges. *Information Fusion*, 2017, **35**: 68−80

76 Chang N B, Bai K X, Imen S, Chen C F, Gao W. Multisensor satellite image fusion and networking for all-weather environmental monitoring. *IEEE Systems Journal*, 2018, **12**(2): 1341−1357

77 Qin Xiong-Pai, Wang Hui-Ju, Du Xiao-Yong, Wang Shan. Big Data analysis — competition and symbiosis of RDBMS and MapReduce. *Journal of Software*, 2012, **23**(1): 32−45
(覃雄派, 王会举, 杜小勇, 王珊. 大数据分析 —RDBMS 与 MapReduce 的竞争与共生. 软件学报, 2012, **23**(1): 32−45)

78 Huang Z R, Wang P, Zhang F, Gao J X, Schich M. A mobility network approach to identify and anticipate large crowd gatherings. *Transportation Research Part B: Methodological*, 2018, **114**: 147−170

79 Lin Y J, Chen H H, Lin G P, Chen J K, Ma Z M, Li J J. Synthesizing decision rules from multiple information sources: a neighborhood granulation viewpoint. *International Journal of Machine Learning and Cybernetics*, 2018, **9**(11): 1919−1928

80 Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: IMLS, 2014. 595−603

81 Srivastava N, Salakhutdinov R. Multimodal learning with deep boltzmann machines. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA: IEEE, 2012. 2222−2230

82 Xu W H, Yu J H. A novel approach to information fusion in multi-source datasets: a granular computing viewpoint. *Information Sciences*, 2017, **378**: 410−423

83 Galton F. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 1889, **45**(273−279): 135−145

84 Li Guo-Jie, Cheng Xue-Qi. Research status and scientific thinking of Big Data. *Bulletin of Chinese Academy of Sciences*, 2012, **27**(6): 647−657
(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)

85 Mayer-Schonberger V, Cukier K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt, 2013.

86 Ye J. Multicriteria decision-making method using the correlation coefficient under single-valued neutrosophic environment. *International Journal of General Systems*, 2013, **42**(4): 386−394

87 Liao H C, Xu Z S, Zeng X J, Merigó J M. Qualitative decision making with correlation coefficients of hesitant fuzzy linguistic term sets. *Knowledge-Based Systems*, 2015, **76**: 127−138

88 Pei S L, Hu Q H. Partially monotonic decision trees. *Information Sciences*, 2018, **424**: 104−117

89 Yang Y, Ma Z G, Yang Y, Nie F P, Shen H T. Multitask spectral clustering by exploring intertask correlation. *IEEE Transactions on Cybernetics*, 2015, **45**(5): 1083−1094.

90 Wang Y, Lin X M, Wu L, Zhang W J, Zhang Q, Huang X D. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 2015, **24**(11): 3939−3949

91 Ma H F, Jia M H Z, Zhang D, Lin X H. Combining tag correlation and user social relation for microblog recommendation. *Information Sciences*, 2017, **385**: 325−337

92 Zhu Y, Kwok J T, Zhou Z H. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 2018, **30**(6): 1081−1094

93 Chaudhuri K, Kakade S M, Livescu K, Sridharan K. Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM, 2009. 129−136

94 Sun Quan-Sen, Zeng Sheng-Gen, Wang Ping-An, Xia De-Shen. The theory of canonical correlation analysis and its application to feature fusion. *Chinese Journal of Computers*, 2005, **28**(9): 1524−1533
(孙权森, 曾生根, 王平安, 夏德深. 典型相关分析的理论及其在特征融合中的应用. 计算机学报, 2005, **28**(9): 1524−1533)

95 Yang Jing, Li Wen-Ping, Zhang Jian-Pei. A tracking algorithm based on rank two modifications for canonical correlation analysis of multidimensional data streams. *Acta Electronica Sinica*, 2012, **40**(9): 1765−1774
(杨静, 李文平, 张健沛. 基于秩 2 更新的多维数据流典型相关跟踪算法. 电子学报, 2012, **40**(9): 1765−1774)

96 Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet G R G, Levy R, et al. A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM, 2010. 251−260

97 Yin X R. Canonical correlation analysis based on information theory. *Journal of Multivariate Analysis*, 2004, **91**(2): 161−176

98 Lai P L, Fyfe C. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2000, **10**(5): 365−377

99 Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 2004, **16**(12): 2639−2664

100 Yang Jing, Li Wen-Ping, Zhang Jian-Pei. Canonical correlation analysis of big data based on cloud model. *Journal on Communications*, 2013, **34**(10): 121−134
(杨静, 李文平, 张健沛. 大数据典型相关分析的云模型方法. 通信学报, 2013, **34**(10): 121−134)

101 Reshef D N, Reshef Y A, Finucane H K, Grossman S R, McVean G, Turnbaugh P J, et al. Detecting novel associations in large data sets. *Science*, 2011, **334**(6062): 1518−1524

102 Nguyen H V, Müller E, Vreeken J, Efros P, Böhm K. Multivariate maximal correlation analysis. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: W&CP, 2014. 775−783

103 Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 2007, **35**(6): 2769−2794

104 Martínez-Gómez E, Richards M T, Richards D S P. Distance correlation methods for discovering associations in large astrophysical databases. *The Astrophysical Journal*, 2014, **781**(1): 39

105 Davis R A, Matsui M, Mikosch T, Wan P. Applications of distance correlation to time series. *Bernoulli*, 2018, **24**(4A): 3087−3116

106 Lin Zi-Yu, Jiang Yi, Lai Yong-Xuan, Lin Chen. A new algorithm on lagged correlation analysis between time series: TPFP. *Journal of Computer Research and Development*, 2012, **49**(12): 2645−2655
(林子雨, 江弋, 赖永炫, 林琛. 一种新的时间序列延迟相关性分析算法—三点预测探查法. 计算机研究与发展, 2012, **49**(12): 2645−2655)

107 Jiang Gao-Xia, Wang Wen-Jian. Correlation analysis in curve registration of time series. *Journal of Software*, 2014, **25**(9): 2002−2017
(姜高霞, 王文剑. 时序数据曲线排齐的相关性分析方法. 软件学报, 2014, **25**(9): 2002−2017)

108 Zhang Wen-Kai, Wang Wen-Jian, Jiang Gao-Xia. Curve registration method for maximizing correlation coefficient based on non-uniform sampling. *Pattern Recognition and Artificial Intelligence*, 2016, **29**(1): 72−81
(张文凯, 王文剑, 姜高霞. 基于非均匀采样的相关系数最大化曲线排齐方法. 模式识别与人工智能, 2016, **29**(1): 72−81)

109 Zhao J P, Itti L. Shapedtw: shape dynamic time warping. *Pattern Recognition*, 2018, **74**: 171−184

110 Baldocchi D, Sturtevant C, Contributors F. Does day and night sampling reduce spurious correlation between canopy photosynthesis and ecosystem respiration? *Agricultural and Forest Meteorology*, 2015, **207**: 117−126

111 Clappe S, Dray S, Peres-Neto P R. Beyond neutrality: disentangling the effects of species sorting and spurious correlations in community analysis. *Ecology*, 2018, **99**(8): 1737−1747

112 Gao P, Zhang L J. Determining spurious correlation between two variables with common elements: event area-weighted suspended sediment yield and event mean runoff depth. *The Professional Geographer*, 2016, **68**(2): 261−270

113 Altman N, Krzywinski M. Association, correlation and causation. *Nature Methods*, 2015, **12**(10): 899−900

114 Xu J, Xu C, Zou B, Tang Y Y, Peng J T, You X G. New incremental learning algorithm with support vector machines. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(11): 2230−2241

115 Gu B, Quan X, Gu Y H, Sheng V S, Zheng G S. Chunk incremental learning for cost-sensitive hinge loss support vector machine. *Pattern Recognition*, 2018, **83**: 196−208

116 Chen H M, Li T R, Zhang J B. A method for incremental updating approximations based on variable precision set-valued ordered information systems. In: Proceedings of the 2010 IEEE International Conference on Granular Computing. San Jose, USA: IEEE, 2010. 96−101

117 Li S Y, Li T R. Incremental update of approximations in dominance-based rough sets approach under the variation of attribute values. *Information Sciences*, 2015, **294**: 348−361

118 Yu H. Three-way decisions and three-way clustering. In: Proceedings of the 2008 International Joint Conference on Rough Sets. Quy Nhon, Vietnam: Springer, 2018. 13−28

119 Hu J, Li T R, Luo C, Fujita H, Yang Y. Incremental fuzzy cluster ensemble learning based on rough set theory. *Knowledge-Based Systems*, 2017, **132**: 144−155

120 Hu C, Chen Y, Peng X, et al. A novel feature incremental learning method for sensor-based activity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2018, **31**(6): 1038−1050

121 Huang Y Y, Li T R, Luo C, Horng S J. Dynamic updating rough approximations in distributed information systems. In: Proceedings of the 10th International Conference on Intelligent Systems and Knowledge Engineering. Taipei, China: IEEE, 2015. 170−175

122 Jing Y G, Li T R, Fujita H, Yu Z, Wang B. An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view. *Information Sciences*, 2017, **411**: 23−38

123 Luo C, Li T R, Chen H M, Fujita H, Yi Z. Incremental rough set approach for hierarchical multicriteria classification. *Information Sciences*, 2018, **429**: 72−87

124 Da Q, Yu Y, Zhou Z H. Learning with augmented class by exploiting unlabeled data. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec, Canada: AAAI Press, 2014. 1760−1766

125 Ristin M, Guillaumin M, Gall J, Van Gool L. Incremental learning of NCM forests for large-scale image classification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 3654−3661

126 Ristin M, Guillaumin M, Gall J, Van Gool L. Incremental learning of random forests for large-scale image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(3): 490−503

127 Júnior P R M, de Souza R M, de O. Werneck R, Stein B V, Pazinato D V, de Almeida W R, et al. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 2017, **106**(3): 359−386

128 Neal L, Olson M, Fern X L, Wong W K, Li F X. Open set learning with counterfactual images. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 620−635

129 Bendale A, Boult T E. Towards open set deep networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 1563−1572

130 Liang S Y, Li Y X, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks [online], available: https:// arxiv.org/abs/1706.02690, December 20, 2018

131 Ahmad S, Lavin A, Purdy S, Agha Z. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 2017, **262**: 134−147

132 Dong F, Zhang G Q, Lu J, Li K. Fuzzy competence model drift detection for data-driven decision support systems. *Knowledge-Based Systems*, 2018, **143**: 284−294

133 Lobo J L, Del Ser J, Bilbao M N, Perfecto C, Salcedo-Sanz S. DRED: an evolutionary diversity generation method for concept drift adaptation in online learning environments. *Applied Soft Computing*, 2018, **68**: 693−709

134 Yu Hong, Wang Guo-Yin, Li Tian-Rui, Liang Ji-Ye, Miao Duo-Qian, Yao Yi-Yu. *Three-Way Decisions: Methods and Practices for Complex Problem Solving*. Beijing: Science Press, 2015.
(于洪, 王国胤, 李天瑞, 梁吉业, 苗夺谦, 姚一豫. 三支决策: 复杂问题求解方法与实践. 北京: 科学出版社, 2015.)

135 Miao Duo-Qian, Zhang Qing-Hua, Qian Yu-Hua, Liang Ji-Ye, Wang Guo-Yin, Wu Wei-Zhi, et al. From human intelligence to machine implementation model: theories and applications based on granular computing. *CAAI Transactions on Intelligent Systems*, 2016, **11**(6): 743−757
(苗夺谦, 张清华, 钱宇华, 梁吉业, 王国胤, 吴伟志, 等. 从人类智能到机器实现模型 — 粒计算理论与方法. 智能系统学报, 2016, **11**(6): 743−757)

136 Xu Ji, Wang Guo-Yin, Yu Hong. Review of big data processing based on granular computing. *Chinese Journal of Computers*, 2015, **38**(8): 1497−1517
(徐计, 王国胤, 于洪. 基于粒计算的大数据处理. 计算机学报, 2015, **38**(8): 1497−1517)

137 Lee J, Jung J, Park P, Chung S, Cha H. Design of a human-centric de-identification framework for utilizing various clinical research data. *Human-centric Computing and Information Sciences*, 2018, **8**(1): 19

138 Wang G Y, Yang J, Xu J. Granular computing: from granularity optimization to multi-granularity joint problem solving. *Granular Computing*, 2017, **2**(3): 105−120

139 Wang G Y. DGCC: data-driven granular cognitive computing. *Granular Computing*, 2017, **2**(4): 343−355, 514

140 Wang Fei-Yue. Parallel system methods for management and control of complex systems. *Control and Decision*, 2004, **19**(5): 485−489, 514
(王飞跃. 平行系统方法与复杂系统的管理和控制. 控制与决策, 2004, **19**(5): 485−489, 514)

141 Wang Fei-Yue. Software-defined systems and knowledge automation: a parallel paradigm shift from Newton to Merton. *Acta Automatica Sinica*, 2015, **41**(1): 1−8
(王飞跃. 软件定义的系统与知识自动化: 从牛顿到默顿的平行升华. 自动化学报, 2015, **41**(1): 1−8)

142 Zheng N N, Liu Z Y, Ren P J, Ma S T, Yu S Y, Xue J R, et al. Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering*, 2017, **18**(2): 153−179

143 Zhang B, Zhang L. Multi-granular computing in web age. In: Proceedings of the 14th International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Berlin, Heidelberg: Springer, 2013. 11−14

于 洪 重庆邮电大学教授. 主要研究方向为工业大数据分析与处理, 智能决策, 知识发现, 粒计算, 三支聚类, 智能推荐. 本文通信作者.
E-mail: yuhong@cqupt.edu.cn
(**YU Hong** Professor at Chongqing University of Posts and Telecommunications. Her research interest covers industrial big data analysis and processing, intelligent decision making, knowledge discovery, cognitive computing, granular computing, three-way clustering and intelligent recommendation. Corresponding author of this paper.)

何德牛 重庆邮电大学计算智能重庆市重点实验室博士研究生. 主要研究方向为工业大数据智能决策, 三支决策, 粒计算. E-mail: hedeniu@163.com
(**HE De-Niu** Ph. D. candidate at Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications. His research interest covers industrial big data for intelligent decision making, three-way decisions and granular computing.)

王国胤 重庆邮电大学教授, 计算智能重庆市重点实验室主任. 主要研究方向为粒计算, 知识发现, 认知计算, 智能信息处理, 大数据智能.
E-mail: wanggy@ieee.org
(**WANG Guo-Yin** Professor at Chongqing University of Posts and Telecommunications, Dean of Chongqing Key Laboratory of Computational Intelligence. His research interest covers granular computing, knowledge discovery, cognitive computing, intelligent information processing and big data intelligence.)

李 劼 中南大学教授. 主要研究方向为冶金新技术与新材料, 冶金过程计算机仿真优化与智能控制.
E-mail: 13808488404@163.com
(**LI Jie** Professor at Central South University. His research interest covers new technologies and materials for metallurgical, computer simulation optimization and intelligent control of metallurgical process.)

谢永芳 中南大学教授. 主要研究方向为复杂工业过程的建模与控制优化, 分布式鲁棒控制, 知识自动化.
E-mail: yfxie@csu.edu.cn
(**XIE Yong-Fang** Professor at Central South University. His research interest covers modeling and optimal control of complex industrial processes, distributed robust control and knowledge automation.)