中国科学: 信息科学 2015 年 第 45 卷 第 11 期: 1355–1369
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述

大数据挖掘的粒计算理论与方法

梁吉业[①②*], 钱宇华[①②], 李德玉[①②], 胡清华[③]

① 山西大学计算智能与中文信息处理教育部重点实验室, 太原 030006
② 山西大学计算机与信息技术学院, 太原 030006
③ 天津大学计算机科学与技术学院, 天津 300072
* 通信作者. E-mail: ljy@sxu.edu.cn

收稿日期: 2015–05–07; 接受日期: 2015–06–08; 网络出版日期: 2015–09–18
国家自然科学基金 (批准号: 61432011, U1435212, 61322211)、国家重点基础研究发展计划 (973 计划) (批准号: 2013CB329404) 和教育部新世纪人才支持计划 (批准号: NCET-12-1031) 资助项目

摘要 大数据往往呈现出大规模性、多模态性以及快速增长性等特征. 粒计算是智能信息处理领域中大规模复杂问题求解的有效范式. 从推动大数据挖掘研究角度, 本文首先概要地讨论了大数据的特征对可计算性、有效性与高效性提出的 3 大挑战; 其次, 结合粒计算的思维模式特点, 概述了已有研究成果, 分析论述了以粒计算应对大数据挖掘挑战的可行性, 认为粒计算有望为大数据挖掘提供一条极具前途的崭新途径; 最后, 对大数据挖掘的粒计算理论与方法中的若干科学问题进行了梳理与展望, 以期抛引这一领域的学术思考与研究工作.

关键词 大数据 数据挖掘 模式发现 粒计算 信息粒化 多粒度

1 引言

根据维基百科的定义, 大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合. 大数据 = 海量数据 + 复杂类型的数据. 大数据通常来源于以互联网为载体的虚拟社会系统, 或者来源于复杂的工业控制系统、物联网系统、经济与金融系统, 现实社会的各种管理与监控系统, 以及诸如空间探测、大气与地球科学、生物信息学等科学研究领域 [1~3]. 2011 年 IDC 公司发布的《数字宇宙研究报告》称: 全球信息总量每两年就会增长一倍, 2011 年全球被创建和被复制的数据总量为 1.8 ZB. 预计到 2020 年, 全球将总共拥有 35 ZB 的数据量. 大数据在现代信息社会中的数据资源主体地位已成为学术界与企业界的共识. 由于对经济活动与社会发展具有可预见的重要推动作用, 大数据已经进入了世界主要经济体的战略研究计划. 正如美国政府启动的 "Big Data Research and Development Initiative" 计划指出的 "将大力推进大数据的收集、访问、组织和开发利用等相关技术的发展, 提高从海量复杂的数据中提炼信息和获取知识的能力与水平". 从大数据中进行数据挖掘与知识发现是大数据应用的战略问题之一.

引用格式: 梁吉业, 钱宇华, 李德玉, 等. 大数据挖掘的粒计算理论与方法. 中国科学: 信息科学, 2015, 45: 1355–1369, doi: 10.1360/N112015-00092

2 大数据挖掘面临的挑战

2008 年, *Nature*出版的专辑 “Big Data” 从互联网技术、网络经济学、超级计算、环境科学和生物医药等多个方面介绍了大数据带来的挑战 [4]. 2011 年 2 月*Science*杂志刊发了海量数据处理专题, 指出 “倘若能够更有效地组织和利用这些数据, 人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用” [5].

大数据的特征常被总结为 4V, 即 Volume (体量浩大)、Variety (模态繁多)、Velocity (快速增长)、Value (价值巨大但密度很低). 其中, “价值巨大但密度很低” 从大数据的利用潜力和当前分析与处理的技术局限性角度概括了大数据的特征, 是指大数据的价值虽高, 但利用密度很低. 本文重点针对大数据的外在形态特征, 即大数据的规模海量性、多模态性以及快速增长性等对传统的数据挖掘理论、方法与技术在可计算性、有效性与时效性等方面提出的严峻挑战进行研究. 而为应对这些挑战所涌现的新的计算理论、方法与技术将有效推动大数据挖掘的发展与应用.

2.1 大规模性 VS 可计算性

数据规模的急剧膨胀给数据挖掘, 特别是面向互联网的数据挖掘带来了巨大挑战 [6,7]. 为了使巨量数据可计算, 一些新的高性能计算方法逐渐涌现. 2004 年, Google 公司首先推出了以 MapReduce 为代表的非关系数据管理技术, 作为面向大数据分析和处理的并行计算模型, 很快引起了学术界和工业界的广泛关注. 在面向大数据的挖掘技术方面, 国内外学者也进行了一些初步探索. 比如, 针对传统分析软件扩展性差以及 Hadoop 分析功能薄弱的特点, IBM 致力于对 R 和 Hadoop 进行集成 [8]. R 是开源统计分析软件, 通过 R 和 Hadoop 的深度集成, 把并行框架下的计算推向数据. 另有研究者实现了 Weka (类似于 R 的开源机器学习和数据挖掘工具软件) 和 MapReduce 的集成 [9]. 标准版 Weka 工具只能在单机上运行, 并且存在内存不能超越 1 GB 的限制. 经过算法的并行化, 在 MapReduce 集群上, Weka 突破了原有的可处理数据量的限制, 可以轻松地对超过 100 GB 的数据进行分析. 另有开发者发起了 Apache Mahout 项目的研究, 该项目是基于 Hadoop 平台的大规模数据集上的机器学习和数据挖掘开源程序库, 为应用开发者提供了丰富的数据分析功能. 针对频繁模式、分类和聚类等数据挖掘任务, 研究人员也提出了相应的大数据解决方案. 比如, Miliaraki 等 [10] 提出了一种可扩展的在 MapReduce 框架下进行频繁序列模式挖掘的算法, Ene 等 [11] 用 MapReduce 实现了大规模数据下的 K-Center 和 K-Median 聚类方法, Yu 等 [12] 提出了针对线性分类模型的大数据分类方法, Kang 等 [13] 使用 Belief Propagation 算法 (简称 BP) 处理大规模图数据挖掘异常模式. 针对大规模图数据分析, Yang 等 [14] 对基于集群上的大规模图数据管理和局部图的访问特征 (广度优先查询和随机游走等) 进行了研究, 提出了分布式图数据环境和两级划分管理架构.

另一种应对大数据可计算性挑战的思路是使用数据采样技术, 通过采样使数据规模变小, 以便利用现有的技术手段进行数据分析 [15]. 然而, 这一思路可能会遭受两方面的质疑. 一方面, 大数据的混杂性使得抽样所获得的样本未必能反映大数据总体; 另一方面, 普遍认为: 大数据条件下, 基于小样本学习理论的传统数据挖掘、机器学习方法的 “独立同分布假设” 难以保障, 致使样本数据模式能否代表总体数据模式受到质疑.

事实上, MapReduce 是在大规模非结构化数据的管理层面为人们提供了一种并行处理架构. 而在大数据数据分析与挖掘层面遭遇的可计算性挑战方面, 尽管已有一些工作, 但还处于借用 MapReduce 对数据进行管理的阶段, 还没进入面向数据挖掘任务、针对大数据本身研究其拆分理论与方法, 以应对可计算性挑战的阶段.

2.2 多模态性 VS 有效性

多模态是大数据的另一个显著特点. 当前, 数据采集方式、手段的多样性一方面为人们提供了从不同视角观测自然系统、工业系统、社会系统中复杂现象的可能性, 另一方面也使得观测对象的数据描述呈现出多模态特征. 比如, 在医疗检测中, 提供的心电、脑电、超声、X 射线、电子计算机断层扫描 (CT)、磁共振成像 (MRI)、正电子发射断层扫描 (PET)、单光子发射断层成像 (SPECT) 及功能磁共振成像 (FMRI) 等多种模态信息是互为差异、互相补充的, 对不同模态信息进行适当的融合成为临床诊断和疾病治疗的迫切需求. 在天体物理研究中, 太阳大气成像仪记录了太阳内部结构和磁场结构、太阳的极紫外线辐射、太阳盘面的数个不同波长紫外线和极紫外线影像, 一天生成的数据将近 2 T. 太阳物理学家需要从如此海量的多模态数据中发现太阳活动的物理规律, 以揭示空间天气的形成机理, 并建立可靠的预报模型.

当前, 在一些领域已经开展了多模态数据分析的探索性研究 [16~19], 主要思路是: 将分别从不同模态的数据中提取的特征合并成一个更大的特征空间, 然后在这个特征空间中进行数据分析与挖掘. 现有方法属于特征层面融合后的分析思路, 其有效性依赖于根据先验知识提取的特征, 难以推广到先验知识匮乏的前沿探索领域. 现有方法的主要局限性表现在以下 3 个方面. 其一, 所获取的数据模式 (知识) 表现出高度非线性特点, 难以被用户理解; 其二, 难以对带有分支、层次、网络等结构的复杂问题进行数据层面的建模; 其三, 融合仅限于特征层面, 还没有深入到知识和推理层面.

如何充分利用大数据的多模态性, 发展面向复杂问题求解, 能从数据、特征、知识、推理等不同层面体现融合思想, 具有分层递阶、分而治之特点的高效挖掘理论与方法是多模态大数据分析的主要挑战.

2.3 增长性 VS 时效性

大数据的又一个显著特点是数据量随着时间快速积累、迅速增长, 人们可以充分利用历史数据和新增数据分析对象的状态、预测事件的发展趋势. 许多实际应用领域的数据挖掘任务具有较高的时效性要求. 比如, 在客户购买行为模式的分析中, 电子商务平台上的数据几乎每时每刻都在动态增加和更新, 决策者需要及时掌握客户行为的模式和消费趋势, 以便更精准地投放广告进行商品推荐. 在股票投资决策中, 股票交易数据在开盘期间实时累积, 如何及时做出优化组合投资决策对降低投资风险、提高收益率至关重要. 在社会网络中, 大量的节点上的状态在不断发生变化, 节点与节点之间的链接情况也在不断发生变化, 这给面向社会网络的数据挖掘的实时性要求带来了挑战. OLAP (online analytical processing) 正是为了契合这种数据分析的时效性需求而被提出的. 然而, 在大数据时代, 数据的增长不仅仅体现在其快速性上, 而且体现在分析所需时间段内数据增量的大规模性, 这使得传统的单增量或小规模批增量机器学习技术的局限性凸显, 大数据呼唤更高效的在线数据分析技术. 最主要的挑战在于: 基于大规模批增量数据的模式更新机制以及高效算法.

综合上述分析可以看出, 大数据的大规模性、多模态性与快速增长性给大数据挖掘提出的挑战是多方面的、多层面的. 衍生出的问题既具有领域相关性, 又具有多学科交叉性. 为此, 需要在现有研究成果的基础上, 以全新的视角发展大数据挖掘的新理论与新方法, 推动大数据学科的发展与应用.

3 粒计算 —— 大数据挖掘的新途径

粒计算是专门研究基于粒结构的思维方式、问题求解方法、信息处理模式的理论、方法、技术和

工具的学科, 是当前智能信息处理领域中一种新的计算范式. 从人工智能角度来看, 粒计算是模拟人类思考和解决大规模复杂问题的自然模式, 从实际问题的需要出发, 用可行的满意近似解替代精确解, 达到对问题的简化、提高问题求解效率等目的. 从数据分析与处理层面看, 粒计算通过将复杂数据进行信息粒化, 用信息粒代替样本作为计算的基本单元, 可大大提高计算效率. 粒计算主要包括数据粒化、多粒度模式发现与融合、多粒度/跨粒度推理等核心研究内容. 大数据的表现性态、大数据挖掘面临的挑战、基于大数据的复杂问题建模与粒计算框架的契合之处主要表现在以下 3 个方面.

3.1 大数据经常具有多层次/多粒度特性

1990 年, 我国著名科学家钱学森先生在其论文《一个科学新领域 —— 开放的复杂巨系统及其方法》[20] 中就指出: “只有一个层次或没有层次结构的事物称为简单的系统, 而子系统种类很多且有层次结构, 它们之间关联关系又很复杂的系统称为复杂巨系统. 任何一个复杂系统都是一个具有层次结构的系统”. Friedman 等 [21] 在*Science*上发表的论文认为在诸如复杂细胞网络、蛋白质互作用网络等生物大数据中都广泛存在着多层次、多尺度特性. Clauset 等 [22] 在*Nature*上发表的论文也指出, 在复杂社会网络中也存在天然的层次结构. Ahn 等 [23] 则专门研究了大数据的多尺度复杂性. 著名社会网络科学家 Watts [24] 在其提出的小世界网络研究中, 也指出网络中嵌套的诸多社区内部也满足小世界网络的要求. 大数据往往来自于对复杂的自然/人工巨系统的观测记录, 或者由人类社会系统借助网络自主产生. 这就意味着, 反映复杂巨系统形态及运动规律的大数据必然隐含着由这些系统所决定的局部与整体关系, 以及复杂的层次结构, 即数据的多粒度/多层次特性.

3.2 挖掘任务通常呈现多层次/多粒度特性

数据挖掘总是面向实际应用的, 即使面对同一个数据集, 用户需求的多层次/多粒度特性也决定了挖掘任务的多层次/多粒度特性. 比如, 在金融大数据领域, 决策任务可能是面向国家层面、区域层面, 或者是地方层面的, 甚至是面向某个银行的; 也可能是面向不同种类的存款、贷款, 或理财产品. 这就使得挖掘任务可能同时面向不同层面、不同方面. 挖掘任务的多层次/多粒度特性必然要求数据挖掘工具不仅能够从不同视角探索大数据不同层面隐含的模式, 而且还能够进行复杂有效的融合、自动的跳转, 以及便捷的定制.

3.3 大数据挖掘要求算法具有高效近似求解性

在 2012 年出版的大数据著作《大数据时代: 生活、工作与思维的大变革》[25] 中指出: “大数据意味着所有数据”. 大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合. 因此, 大数据挖掘首先要解决 “大数据能算的问题”, 这就要求对大数据进行合理的分解, 即大数据集的粒化, 然后采用并行处理策略, MapReduce 正是基于这种策略在大数据管理方面的实践结果.

基于大数据的复杂问题建模往往具有极其复杂的结构, 这就要求大数据挖掘算法能够按照任务的要求自动地或人机交互地从大数据中抽取与组织出具有多层次/多局部特征的结构, 并能在这种复杂结构上进行推理, 以达到挖掘的预期目标.

大数据挖掘算法的高效近似求解特性, 主要来自于用户对挖掘过程、挖掘结果的时效性要求, 大数据的巨量增长性对在线挖掘技术提出了严峻挑战. 与传统的小数据集上的挖掘与学习不同, 大数据的混杂性、不确定性, 以及高噪声对 “独立同分布假设” 的破坏使得追求问题的最优/精确解变得几乎

不可能, 迫使我们转向寻找问题的满意近似解. 另一方面, 满意近似解在很多环境下已能很好地满足实际应用的需要, 无需一味追求问题的最优/精确解.

综上分析可知, 从隐含于大数据中的结构特征, 大数据挖掘任务的类型特征, 到大数据挖掘算法的性能特征, 综合这些角度, 大数据挖掘的计算框架与粒计算所蕴含的计算范式具有高度契合性. 鉴于这一认识, 可以推测: 粒计算将为大数据挖掘提供一条极具前途的崭新途径.

4 现状分析

早在 1979 年, 美国著名控制论专家 Zadeh [26] 就首次提出了模糊信息粒化问题. 他认为, 人类认知能力可概括为粒化 (granulation, 全体分解为部分)、组织 (organization, 部分集成为整体) 和因果 (causation, 因果的关联) 3 个主要特征. 1985 年, Hobbs [27] 提出了粒度 (granularity) 的概念. 在 20 世纪 90 年代初, 我国的张钹和张铃 [28,29] 在其专著《问题求解理论及应用》中特别指出 "人类智能的一个公认特点, 就是人们能从极不相同的粒度上观察和分析同一问题. 人们不仅能在不同粒度世界上进行问题的求解, 而且能够很快地从一个粒度世界跳到另一个粒度世界, 往返自如, 毫无困难". 这种处理不同粒度世界的能力, 正是人类问题求解的强有力的表现. Yager 和 Filev [30] 进一步指出 "人们已形成了一个关于世界的粒度观点, ……, 在此观点下, 人类的观察、度量、概念化和推理都是在粒度意义下进行的". 这些观点都认为, 粒化作为人类认知的重要特征之一, 对复杂数据的知识发现具有重要作用. 1997 年, Zadeh [31] 第一次提出了粒计算 (granular computing) 的概念. 随后国际上许多不同领域的学者都开始关注和研究这个问题, 其逐渐形成了智能信息处理中一个新的研究方向.

自粒计算这一概念提出以来, 大量关于粒计算研究的学术论文相继发表, 在国际上形成了专门的研究群体. 近年来, 国际上两个系列会议 "IEEE International Conference on Granular Computing" 与 "International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing" 每年举办一次. 在国内, 2001 年以来, 粒计算的研究成果集中在每年举办的 "中国粗糙集与软计算学术会议" 上报道和交流. 由于粒计算在国内的迅速发展, 2007 年开始每年举办 "中国粒计算学术会议". 短短 十余年的发展已经凸显出它对信息科学特别是对计算机科学的作用和影响. 粒计算已逐渐成为数据分析领域的新分支, 在数据粒化、多粒度模式发现以及粒度推理等方面的研究已经取得了一些重要的进展.

4.1 数据粒化研究进展

数据粒化是基于粒计算的数据分析的基础, 是按照给定的粒化策略将复杂数据分解为信息粒的过程. 根据不同的数据建模目标和用户需求, 可以采用多种多样的粒化策略. 单纯依赖数据的常用粒化策略大多可以归结为基于数据二元关系的粒化策略, 其本质是将满足预先定义的二元关系的两个数据样本分配到同一个数据粒中. 诸多粒化策略通过使用等价关系、相似关系、极大相似关系、模糊等价关系、模糊相似关系、邻域关系、优势关系等二元关系可将数据粒化为相应的二元粒结构 [32~39]. 再如, 图数据中的连通分支, 极大全连通子图、各种路及圈, 以及树中的子树、链等. 基于聚类的粒化策略尽管本质上也是依赖于二元关系, 但它是在目标函数引导下, 通过一个迭代寻优过程学习得到隐含于数据中的簇团结构 [40]. 聚类可以被理解为数据簇团结构的发现方法, 也可以被理解为一种数据粒化策略. 聚类粒化具有很强的数据针对性, 如针对高维数据, 其代表性方法有基于子空间的聚类 [41]、联合聚类 [42] 以及基于超图的聚类 [43] 等; 针对复杂数据, Brendan [44] 在*Science*发表了一个基于消息传递的聚类算法, 成功应用于人脸图像聚类、基因外显子发现、手稿中心句识别以及最优航线搜索等

方面; 针对多模态数据, Ahmad 等 [45] 提出了一种数值型和符号型并存的多模态数据的 K-Means 算法. Huang [46] 提出了用于解决符号数据聚类的 K-Modes 算法, 并将其与 K-Means 算法相融合用于多模态数据聚类. 此外, 在图像处理领域里, 有一类通过提取图像本身的纹理、边缘、灰度值等特征, 并将其作为多模态特征来进行图像数据的粒化方法 [18,19,47].

目前的数据粒化策略与方法很少考虑适应大数据的可计算性需求, 多是从单一模态特征出发, 在不同模态特征之间设置权重参数或者简单地对结果进行集成, 并没有从本质上进行融合, 不能够保证其语义一致性或语义相关性, 无法有效解决具有多模态特征的数据协同粒化问题.

4.2 多粒度模式发现研究进展

多粒度模式发现与融合是粒计算框架下复杂问题求解的内在逻辑要求. 所谓多粒度, 可以是多个数据子集、表示空间的多个子空间、多个不同的模态变量集、也可以是问题求解过程中的多个局部或中间结果, 它对应于问题的多个角度、多个局部或多个层次. 为了获得整体数据集或问题的全局解, 需要对多个单一粒度上发现的模式进行融合. 尽管没有使用多粒度这一术语, 国内外学者已经针对医学图像分析、网络、视频语义分析、标注和检索、情感识别等领域的多模态问题开展了融合研究, 主要考虑从不同模态的数据中分别提取特征, 构成多模态特征空间, 发展具有多模态特征的模式发现理论与方法. 目前的研究主要集中在 3 个方面: 基于多核学习的多模态数据分类 [48], 基于多字典协同表达的多模态数据建模 [49] 和基于深度学习的多模态数据融合 [50]. 比如, 在多模态视频挖掘中, Hershey [51] 将视觉和音频从视频中提取出来, 利用参数模型的方法对音频信号在视频画面中对应的发生区域进行了挖掘. Darrel 等 [52] 则提出了一种基于参数模型的新方法. 这些多模态融合方法主要分为乘积融合、线性融合、非线性融合等. 在多模态 Web 挖掘中, 网页上丰富的图片、音频、视频以及文本等多种模态信息构成了典型的多模态数据挖掘问题. 一些学者提出了语义网方法来描述单词和图片之间的相关性, 并利用文本信息帮助进行图像的检索. 多模态图像检索系统 iFind [53] 提出了一种同时利用文本和图像视觉信息的相关反馈算法. 胡清华 [54] 系统地研究了数值型数据和符号型数据并存的多模态数据的粗糙集方法, 并将粗糙集方法推广应用到了太空天气预报、风力发电故障识别等领域. Hwang 等 [55] 研究了多模态数据的聚类问题, 将图像数据中的纹理、灰度值、线条等提取出来作为多个模态来研究. Wang 等 [56] 将网页中的图片和图片周围文字分离成两类事物, 通过两者之间的关联矩阵将同一事物中的相似度传播到另一类事物中, 得到了更为精确的图片之间的相似度. Qian 等 [57] 系统提出了多粒度粗糙集理论, 通过挖掘不同粒度下的隐含模式来对目标概念与目标决策进行融合学习, 用于更加高效合理地进行决策. 然而, 目前的研究多集中在基于聚类的多粒度结构发现以及基于表示空间的多粒度多模态分类问题, 还很少考虑基于多粒度的分类、回归和相关关系分析等数据分析任务, 尤其是多模态数据的分层多级的分类回归分析较少有人涉及. 此外, 多个粒化结构之间的关系研究也鲜见报道, 它也应是多粒度理论与方法中的重要研究内容.

4.3 粒计算推理研究进展

推理是人类智能中的重要能力之一. 推理是一种形式逻辑, 是用于研究人们思维形式、规律以及逻辑方法的科学. 推理的作用是可从已知的知识得到未知的知识. 粒计算推理指的是利用已知的信息粒或粒空间进行演绎的逻辑方法.

在粒计算领域中, 已经有一些关于粒计算推理的研究. Yao [58] 对粒计算的基本问题、基本方法进行了系列研究, 通过采用决策逻辑语言 (DL- 语言) 来描述论域的粒度, 构建粒度世界的逻辑框架; 将

邻域系统、区间分析、粗糙集理论和粒计算进行融合, 对粒计算中的信息粒化和概念近似问题进行研究; 利用所有划分构成的格研究了一致分类学习问题. 刘清等 [59] 基于粗糙集研究了用于逻辑推理的决策规则粒和粒语言. Thiele [60] 于 1998 年发表了 "词计算理论的语义模型" 和 Zadeh [26] 发表的论文促进了词计算理论的发展, 这些理论旨在解决利用自然语言, 进行模糊推理和判断, 以实现模糊智能控制. 在这些研究中, 不同层面的知识可以通过不同程度的模糊信息粒来刻画, 然后利用模糊逻辑进行推理和计算, 它对于复杂信息系统的模糊推理和控制尤其重要. 针对复杂生物网络, Daphne 等 [61] 提出了概率图模型利用特征之间潜在的相关性来研究粒度之间的推理原理, 已经形成了数据分析领域中一种重要的研究方法. Friedman [21] 用概率图模型对细胞网络进行了推理, 研究了不同粒度层次意义下的细胞网络结构. Andrew 等 [62] 从多层次、多结构角度出发研究了数据分析中的回归问题. Fan 等 [63] 则利用多结构分类思想对多层次图像语义挖掘问题进行了深入研究. 1990 年, 张钹和张铃 [28,29] 提出了商空间理论, 专门研究不同粒度之间的关系、合成、综合、分解和推理, 其最重要的性质是同态原则, 即保真原理 (或保假原理).

尽管在粒推理方面已经取得了一些有益的研究成果, 然而已有方法主要讨论单一粒度下的粒化推理问题, 很少有人考虑多粒度、跨粒度的推理, 特别是缺乏关于多模态数据粒推理的有效方法, 而多粒度、跨粒度推理恰恰是解决大规模复杂决策任务的重要手段.

4.4 基于粒计算的高性能算法研究进展

在粒计算中, 采用信息粒代替样本作为基本的运算单位, 用可行的满意近似解代替精确解, 可达到设计高性能算法的目的.

近年来, 采用粒计算的思想来求解大数据问题已有一些初步的探索. Ye 等 [64] 通过对数据空间和特征空间的粒化, 并利用集成学习技术实现了大规模数据的聚类分析. Chang 等 [65] 提出了一种利用决策树思想的大数据分解方法, 然后在每个分解的数据粒上分别学习 SVM 分类器, 极大提高了 SVM 的学习效率. Gopal 等 [66] 则利用了数据类别之间的层级关系, 并给出了一种相应的 Bayesian 模型来提升其泛化性能. Miao 等 [67] 利用 MapReduce 中的数据分解原理提出了一种可并行计算的属性约简方法. Liang 等 [68] 通过拆分原始大数据集为多个易于处理的信息粒, 通过求解和融合每个信息粒上面的特征选择结果, 给出了一种高效的大数据特征选择算法. Qian 等 [69] 利用信息粒构造了粗糙集的正向近似, 并基于此提出了特征选择加速器, 可用于加速一大类前向贪婪搜索的特征选择算法. Chen 等 [70] 在一篇关于大数据分析的综述性论文中专门将粒计算作为一种应对大数据挑战的潜在技术, 指出: 不同的信息粒度隐含着不同的特征和模式, 可更加有效地去设计机器学习和数据挖掘算法.

对于基于粒计算的高效数据挖掘研究而言, 这方面的工作仅仅是进行了初步的尝试. 其面临的挑战主要体现在两个方面: (1) 如何合理地进行信息粒化, 以确保算法所得到的解是有效的; (2) 如何平衡算法的效率与求解结果的精度, 以高效地获得可行的满意近似解.

5 几个科学问题

大数据的大规模、高维、多模态、多源异构、快速增长等特征对信息粒化、模式发现与融合、推理等粒计算理论与方法的核心要素提出了严峻挑战, 涉及一些重要的科学问题.

5.1 信息粒化

数据粒化就是将数据进行分解, 拟或还需将分解的局部数据按照分析的要求重新组织. 它可能是按照数据的某些自然属性简单分解, 也可能是按照问题求解所基于的框架、理论、方法和技术的特点对局部数据的内涵要求进行数据分解与组织. 粒化是基于粒计算框架求解复杂问题的基础. 大数据的规模性主要体现在样本规模的海量性和特征规模的高维性两个方面; 而大数据的复杂性是其表征的巨系统复杂性在数据层面的反映, 主要表现在数据的多源性、多模态性、混杂性以及多数据输出源本身的结构复杂性. 大数据的大规模性、复杂性对大数据挖掘的可计算、复杂问题求解、挖掘结果的理解与应用提出了巨大挑战, 这些挑战要求我们对大数据从一个、多个或者从某个具有结构的参照框架等视角进行粒化.

5.1.1 大规模数据的信息粒化

对于大数据来说, 数据的整体是海量的, 而整体更显意义, 所以如何拆分大数据而保持数据的整体特性是一个十分重要的问题. 简单随机抽样不适合解决此类问题. 从数据粒化的角度出发, 按照某种策略, 将大数据集拆分为若干个小数据集, 通过对每一个小数据集进行推断, 然后将小数据集上的推断融合形成一个整体推断, 并使得其能反映大数据集的某些整体性质. 拆分的本质困难在于如何分解以体现数据的某些 (某种, 某类) 整体特征. 初步的研究表明: 对于大规模数据的粒化可能需要遵守 3 个准则: 近似性、传递性与遍历性 [68]. 近似性指的是粒化后的每个信息粒中样本的分布要与整体数据的分布尽可能一致; 传递性指的是每个信息粒之间的隐含的模式信息要具有传递性; 遍历性指的是原始大数据集上的所有样本要尽可能的被使用. 如何设计满足这 3 个准则的信息粒化方法将是非常有意义的研究.

5.1.2 多源异构多模态数据的信息粒化

大数据可能是从不同物理地址上的数据源获得的异构数据, 兼或具有多模态特征, 其典型代表是广泛分布于互联网上的社会化媒体数据. 如微博数据中就包含了微博用户的性别、年龄、职业、社会网络关系、自然文本、图像、视频、音频等. 对此类数据, 即使考察同一属性下不同物理空间上的同一类对象, 也不能期望其具有同分布特性. 对符号、数值、时间序列、图像、文本和社会网络结构等单一模态, 基于符号和数值特征构造的等价关系、相似关系或者序关系, 可以形成对象的划分、覆盖或者嵌套的粒化结构; 基于分级聚类方法可以产生数据的分级粒化结构; 基于主题模型、图像语义和概念本体结构, 可以建立起多层语义粒化结构; 基于独立个体的社会网络结构及其社区分析算法, 可以建立由通讯关系、关注关系等信息获得社群粒化结构. 然而, 大数据的异构多模态的特征描述自然且必然地蕴涵十分丰富的语意, 如何分别挖掘与融合异构多模态数据中的语义信息, 并基于此揭示、建立和评价大数据的多尺度粒结构是一个亟待解决的问题.

5.1.3 高维数据的信息粒化

探测数据在特征空间中是否具有块分布现象, 并发现各数据块对应的特征子空间, 对数据应用业务中的客户关系管理、分类设计用户兴趣模型、构建样本相关关系、构建基于内容的精准推荐模型等具有重要的作用. 从数据挖掘的角度讲, 发现数据块对应的特征子空间也可以在一定程度上消除高维性引起的数据稀疏问题.

面临高维或超高维数据挖掘任务时, 密切关联于任务的特征子空间可能不止一个. 比如在医疗数据挖掘中, 数据会呈现出对应于某些特征子空间的块状分布特点; 再如, 高维数据的特征约简理论也证明: 特征空间的约简可能会有多个.

特征视角下的子空间数据粒化问题实际上是部分数据的相关特征子空间的发现问题. 子空间聚类技术可将样本聚类与发现簇所对应的特征子空间发现融合在同一个优化过程中. 然而, 面临的挑战在于: (1) 如果数据的维度非常高、数据表示异常稀疏, 基于子空间的信息粒化应当如何保证其有效性? (2) 由于关联于数据块的特征子空间可能不唯一, 那么如何评估这些粒化结果以便选择?

5.1.4 增长性数据的信息粒化

在大数据时代, 数据的增长不仅仅体现在其快速性上, 而且体现在分析所需时间段内数据增量的大规模性.

这引起的挑战在于: 传统的单增量或小规模批增量信息粒化技术的具有明显的局限性, 如何设计更高效的在线信息粒化技术? 此外, 众所周知, 很多数据的分布通常会随着时间的推移发生变化, 那么对于这类增长性数据应当如何有效信息粒化?

5.2 多粒度模式发现与融合

多粒度模式发现 (融合) 是大数据超大规模性、多模态性、混杂性特征的自然要求, 也是粒计算框架下复杂问题求解的内在逻辑要求. 大数据粒化后, 样本视角下每个同质数据粒上的模式发现问题尽管可以用现有的方法解决, 但全局数据上的模式发现需要在融合策略指导下进行. 对多个异质数据粒 (对应于不同特征子空间所抽象出的概念层次、不同模态以及参照框架下的数据粒化结果) 上的模式发现问题, 自然地要求保证发现过程的语义一致性或语义关联性, 从而保证全局融合结果的正确性和强可解释性.

5.2.1 多粒度聚类

聚类、分类是两种最重要的数据挖掘技术, 它们分别被用于探测数据的抱团性, 以及决策空间对特征空间的依赖性. 数据的这两类特性也分别被称为数据的聚类模式和分类模式.

当大数据被基于样本视角粒化后, 数据的模式发现只能在局部数据粒上进行, 然后再融合成大数据整体上的模式. 对于聚类任务而言, 通过对局部信息粒内的样本空间进行聚类分析后, 关键在于如何有效地融合为整体数据上的模式. 当面对高维数据时, 这个问题本质上是如何融合多个粒空间下发现的模式为整体上的模式. 这对于多模态数据、增长型数据以及多源异构数据而言, 本质上都是如何将局部粒上的模式发现结果有效地融合为整体数据的模式发现结果 (通常需要考虑信息粒与模式的语义一致性). 这类多粒度聚类方法也都需要一套有效性验证方法来保证模式发现的合理性与准确性.

5.2.2 多粒度分类

多粒度分类暗含着两层含义. 一类是通过对大数据的多种策略信息粒化后, 在得到的 m 个粒空间中, 把其中每个信息粒看作抽象的样本来学习一个分类器, 然后对这 m 个分类器进行 (选择性) 集成学习. 另一类是对大数据本身进行拆分 (通常是覆盖, 可参照 5.1.1 中的粒化准则), 形成互相联系的信息粒集, 然后从每个信息粒上学习分类器, 然后把这些分类器融合为一个整体上的分类器. 如何将局部数据粒上的分类模式进行聚合, 作为整体数据集上的分类模式是这类方法成功的关键.

5.2.3 多粒度相关关系发现

大数据相关性分析是支撑大数据应用最重要的方面之一. 与小数据时代不同, 大数据的混杂性迫使我们在无数据分布先验假设的条件下进行分析, 数据相关性分析不可能再沿用 "主观假设、观测采样、统计分析、获得结论" 的分析方法, 而更多的是采用探测性分析方法, 这一趋势从较早的关联规则挖掘就开始了. 数据的相关分两个方面, 一种是特征 (变量) 间相关关系, 典型地, 如一元或多元回归分析, 它强调的是变量在数据集上呈现的带有某种不确定性的函数关系; 另一种是样本 (对象) 间的相关关系, 如关联规则, 它强调的是具有某种性质的对象集间依赖关系、同现关系等. 传统上, 统计学中的 Pearson 相关系数等刻画的是两个随机变量的函数依赖关系 (表现为带有随机扰动的函数曲线). 对大数据来讲, 相关关系并不仅仅意味着函数关系, 它还可能是特征间带有多种不确定性的依赖关系、分布相似性、或者特征的某些取值的同现关系等. 2011 年发表于*Science*上的论文 (Detecting novel associations in large data sets) 研究了大规模数据集中两个实变量间的相关关系 [71]. 大数据多粒度相关关系包括几种基本含义, 一种是探测特征子集 (可能存在多个) 与响应变量的相关性; 另一种是探测数据子集 (可能多个), 在该数据子集上变量 (组) 之间的相关性; 另外, 对多模态数据, 还需要研究不同模态特征之间的相关关系.

5.2.4 多粒度数据模式更新

自然系统和社会系统中的现象往往具有时变性, 在数据挖掘领域, 这种时变性表现为隐含于数据中模式的演化或突变, 其分析的数据基础表现为分时段的数据变化性. 对大数据而言, 这种数据变化包括样本规模的增长性、维度规模的增长性以及数据取值变化的动态性. 探索大数据条件下的数据模式更新技术对于大数据应用具有重要意义. 我们认为模式的变化是由于其数据分布约束的变化而引起的, 这种数据分布变化之于模式的变化意味着紧密关联于模式的特征子空间的变化. 对于不断变化的大数据而言, 每次对全体数据重新进行处理可能不是一个高效的策略, 利用已有的分析结果动态更新模式发现结果是一种可行策略. 这种策略的核心挑战在于如何设计高效的增量式算法.

5.3 多粒度推理

不同于传统机器学习解决单一的决策任务, 基于大数据的决策往往涉及复杂大系统的多任务多目标的分级决策. 基于决策树、神经网络或者支持向量机的决策模型难以有效表示和建模多粒度的复杂决策任务.

5.3.1 多粒度不确定性推理的表示

不确定性推理是人类智能的重要体现, 是人工智能领域的重要研究问题之一.

不确定性推理已经有很多种模型, 比如模糊推理、概率推理、Bayes 推理、证据理论与粒化不确定性推理 [72] 等. 图模型被广泛应用决策过程表示和推理, 尤其是概率图模型被广泛应用于表示变量间的随机依赖关系, 并且发展出了能够表示序贯决策过程变量依赖性的 Markov 随机场、条件随机场以及分级多层 Bayes 网络等具有强大表示能力的模型, 在许多领域取得了成功应用. 基于大规模异构数据的推理必然是一种依赖于变量结构信息、事件发展的时空信息和同时混杂了多种不确定信息的决策过程. 目前的概率图模型尚不足以表示如此复杂的模型, 而 Pawlak 提出的决策流图也无法表示此类过程. 构建能够同时表示多种不确定性、能够刻画时空依赖特性、具有多尺度结构的不确定推理模型可能是解决大规模复杂决策的基石和有效途径.

5.3.2 多粒度推理策略与转换机制

大规模复杂问题的求解往往受到多种资源的约束, 如特定环境下的信息缺失、计算资源有限或者费用不足等, 用户只需要在多类资源约束条件下获得在某一置信度阈值之上的近似解. 细粒度的信息获取和推理往往意味着需要更多的计算资源、信息采集代价和求解时间, 而粗粒度的推理则面临着决策不够精细以及决策置信度下降等问题. 因此, 在实际决策过程中, 往往需要在决策代价和决策粒度之间进行折中, 建立最优的损失函数. 资源约束条件下的多粒度推理机制显然是大规模复杂决策的核心问题, 但相应的粒度选择和切换问题还存在许多理论空白. 其中在多个不同粒度之间的切换与推理是其挑战问题之一.

5.3.3 人机协同的多粒度推理

大数据挖掘的一个重要目的是用户对数据的理解, 而且是多粒度多视角的理解, 对于科学大数据建模而言, 这一点尤其重要, 在一定意义上决定了挖掘算法的实用性, 自动建模算法只是领域专家理解数据的工具和推理的辅助手段. 因此开发用户友好的、甚至人在回路 (man-in-loop) 的推理模型就显得极为重要. 粒计算契合了数据理解的需求, 以人们容易理解的信息粒子 (基本概念) 作为计算单元和推理的原子概念, 以图模型为基本表示工具实现多粒度信息结构提取和推理, 此计算范式容易将人们的先验信息引入到模型结构中, 设计人在回路的、人机协同的建模和复杂问题求解机制. 人机协同的多粒度推理将为大数据环境下的复杂决策任务提供快速高效的求解策略, 也为不同层次决策者提供大数据的多粒度理解和多粒度推理的灵活机制与算法.

6 结语

本文首先讨论了大数据的大规模性、多模态性以及增长性在可计算性、有效性与高效性等 3 方面给大数据挖掘提出的严峻挑战; 接着论证了大数据挖掘的特点与粒计算理念的高度契合性, 认为粒计算有望为大数据挖掘提供一条极具前途的崭新途径; 通过对研究现状的详细分析, 最后也指出了面向大数据挖掘的粒计算理论与方法在未来值得关注的一些重要问题, 包括信息粒化、多粒度模式发现与多粒度推理等方面. 本文主要是对大数据背景下粒计算理论与方法的未来研究进行了一些粗浅的思考, 希望为大数据挖掘的粒计算理论与方法体系的构建起到积极的推动作用.

参考文献

1 Li G J. Sciencific value of big data research. Commun CCF, 2012, 8: 8–15 [李国杰. 大数据研究的科学价值. 中国计算机学会通讯, 2012, 8: 8–15]

2 Wang Y Z, Jin X L, Cheng X Q. Network big data: present and future. Chin J Comput, 2013, 35: 1125–1138 [王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望. 计算机学报, 2013, 35: 1125–1138]

3 Meng X F, Li Y, Zhu J H. Social computing in the era of big data: opportunities and challenges. J Comput Res Develop, 2013, 50: 2483–2491 [孟小峰, 李勇, 祝建华. 社会计算: 大数据时代的机遇与挑战. 计算机研究与发展, 2013, 50: 2483–2491]

4 Lynch C, Goldston D, Howe D, et al. Big data. Nature, 2008, 455: 1–136

5 Science Staff. Dealing with data. Science, 2011, 331: 639–806

6 Zhang C S. Challenges in machine learning. Sci Sin Inform, 2013, 43: 1612–1623 [张长水. 机器学习面临的挑战. 中国科学: 信息科学, 2013, 43: 1612–1623]

7 Wu X D, Zhu X Q, Wu G Q, et al. Data mining with big data. IEEE Trans Knowl Data Eng, 2014, 26: 97–107

8 Das S, Sismanis Y, Beyer K S, et al. Ricardo: intergrating R and Hadoop. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, New York, 2010. 987–998
9 Wegener D, Mock M, Adranale D, et al. Toolkit-based high-performance data mining of large data on MapReduce clusters. In: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, Washington, 2009. 296–301
10 Iris M, Klaus B, Rainer G, et al. Mind the gap: large-scale frequent sequence mining. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, 2013. 797–808
11 Alina E, Sungjin I, Benjamin M. Fast clustering using MapReduce. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2011. 681–689
12 Yu H F, Hsieh C J, Chang K W, et al. Large linear classification when data cannot fit in memory. ACM Trans Knowl Discov Data, 2012, 5: 1–23
13 Kang U, Meeder B, Papalexakis E E, et al. HEigen: spectral analysis for billion-scale graphs. IEEE Trans Knowl Data Eng, 2014, 26: 350–362
14 Yang S Q, Yan X F, Zong B, et al. Towards effective partition management for large graphs. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, New York, 2012. 517–528
15 Kleiner A, Talwalkar A, Sarkar P, et al. The big data bootstrap. In: Proceedings of 29th International Conference on Machine Learning, Edinburgh, 2012. 1759–1766
16 Friston K J. Modalities, modes, and models in functional neuroimaging. Science, 2009, 326: 399–403
17 Zhang D, Shen D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage, 2012, 59: 895–907
18 Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Recogn Mach Intell, 2009, 31: 39–58
19 Guo Z, Zhang Z F, Xing E. Enhanced max margin learning on multimodal data mining in a multimedia database. In: Proceeding of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2007. 340–349
20 Qian X S, Yu J Y, Dai R W. A new discipline of science-open complex giant system and its methodology. Nature Magazine, 1990, 13: 3–11 [钱学森, 于景元, 戴汝为. 一个科学新领域 —— 开放的复杂巨系统及其方法论. 自然杂志, 1990, 13: 3–11]
21 Firedman N. Inferring cellular networks using probabilistic graphical models. Science, 2004, 303: 799–805
22 Claruset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks. Nature, 2008, 453: 98–101
23 Ahn Y Y, Bagrow J P, Lehmannn S. Link communities reveal multiscale complexity in networks. Nature, 2010, 466: 761–765
24 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393: 440–442
25 Sheng Y Y, Zhou T. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Zhejiang: Zhejiang People's Publishing House, 2012 [盛杨燕, 周涛. 大数据时代: 生活、工作与思维的大变革. 浙江: 浙江人民出版社, 2012]
26 Zadeh L. Fuzzy sets and information granularity. In: Gupta N, Ragade R, Yager R, eds. Advances in Fuzzy Set Theory and Application. Amsterdam: North-Holland, 1979. 111–127
27 Hobbs J R. Granularity. In: Proceedings of IJCAI, Los Angeles, 1985. 432–435
28 Zhang B, Zhang L. Theory and Applications of Problem Solving. North-Holland: Elsevier Science Publishers, 1992
29 Zhang L, Zhang B. Problem Solving Theory and Application. Beijing: Tsinghua University Publishing House, 2007 [张铃, 张钹. 问题求解理论及应用. 北京: 清华大学出版社, 2007]
30 Yager R R, Filev D. Operations for granular computing: mixing words with numbers. In: Proceedings of 1998 IEEE International Conference on Fuzzy Systems, Anchorage, 1998. 123–128
31 Zadeh L. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst, 1997, 90: 111–127
32 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publishers, 1991
33 Kryszkiewicz M. Rule in incomplete information systems. Inform Sci, 1999, 113: 271–292
34 Leung L, Li D Y. Maximal consistent block techniques for rule acquisition in incomplete information systems. Inform

Sci, 2003, 153: 85–106
35 Zhu W, Wang F Y. On three types of covering-based rough sets. IEEE Trans Knowl Data Eng, 2007, 19: 1131–1144
36 Miao D Q, Gao C, Zhang N, et al. Diverse reduct subspaces based co-training for partially labeled data. Int J Approx Reason, 2011, 52: 1103–1117
37 Greco S, Matarazzo B, Slowinski R. Rough sets theory for multicriteria decision analysis. Eur J Oper Res, 2001, 129: 1–47
38 Wang G Y, Ma X A, Yu H. Monotonic uncertainty measures for attribute reduction in probabilistic rough set model. Int J Approx Reason, 2015, 59: 41–67
39 Hu Q H, Xie Z X, Yu D R. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recogn, 2007, 40: 3509–3521
40 Aggarwal C C, Reddy C K. Data Clustering: Algorithms and Applications. London: Chapman and Hall/CRC, 2014
41 Kriegel H P, KrÄger P, Zimek A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data, 2009, 3: 1–58
42 Zhang L J, Chen C, Bu J J, et al. Locally discriminative coclustering. IEEE Trans Knowl Data Eng, 2012, 24: 1025–1035
43 Schaeffer S E. Graph clustering. Comput Sci Rev, 2007, 1: 27–64
44 Brendan J F, Delbert D. Clustering by passing messages between data points. Science, 2007, 315: 972–976
45 Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng, 2007, 62: 503–527
46 Huang Z X. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowl Discov, 1998, 2: 283–304
47 Hsu C C, Chen C L, Su Y W. Hierarchical clustering of mixed data based on distance hierarchy. Inform Sci, 2007, 177: 4474–4492
48 Bucak S S, Jin R, Jain A K. Multiple kernel learning for visual object recognition: a review. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 1354–1369
49 Yang M, Zhang L, Zhang D, et al. Relaxed collaborative representation for pattern classification. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, 2012. 2224–2231
50 Wu P C, Hoi S C H, Xia H, et al. Online multimodel deep similarity learning with application to image retrieval. In: Proceedings of the 21st ACM International Conference on Multimedia (MM2013), New York, 2013. 153–162
51 Hershey J, Movellan J. Using audio-visual synchrony to locate sounds. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001
52 John W F, William T, Darrell T, et al. Learning joint statistical models for audio-visual fusion and segregation. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2000
53 Lu Y, Hu C H, Zhu X Q, et al. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: Proceedings of the 8th ACM international conference on Multimedia, New York, 2000. 31–37
54 Hu Q H, Yu D R. The Application of Rough Calculation. Beijing: Scientific Publishing House, 2012 [胡清华, 于达仁. 应用粗糙计算. 北京: 科学出版社, 2012]
55 Hwang S, Grauman K. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int J Comput Vision, 2012, 100: 134–153
56 Wang X J, Ma W Y, Xue G R, et al. Multi-model similarity propagation and its application for web image retrieval. In: Proceedings of the 12th annual ACM International Conference on Multimedia, New York, 2004. 944–951
57 Qian Y H, Liang J Y, Dang C Y. Incomplete multigranulation rough set. IEEE Trans Syst Man Cybern-Part A, 2010, 40: 420–431
58 Yao Y Y. A generalized decision logic language for granular computing. In: Proceedings of IEEE International Conference on Fuzzy Systems, Honolulu, 2002. 12–17
59 Liu Q, Liu Q. Granules and applications of granular computing in logical reasoning. J Comput Res Develop, 2004, 41: 546–551 [刘清, 刘群. 粒及粒计算在逻辑推理中的应用. 计算机研究与发展, 2004, 41: 546–551]
60 Thiele H. On semantic models for investigating computing with words. In: 2nd International Conference on Knowledge Based Intelligent Electronic Systems, Adelaide, 1998. 32–98
61 Daphne K, Nir F. Probabilistic Graphical Models: Principles and Techniques. Cambridge: MIT Press, 2009

62 Andrew G, Jennifer H. Data Analysis Using Regression and Multilevel/ Hierarchical Models. Cambridge: Cambridge University Press, 2007
63 Fan J, Gao Y, Luo H, et al. Mining multilevel image semantics via hierarchical classification. IEEE Trans Multimedia, 2008, 10: 167–187
64 Ye Y M, Wu Q Y, Huang Z X, et al. Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recogn, 2013, 46: 769–787
65 Chang F, Guo C Y, Lin X R. Tree decomposition for large-scale problems. J Mach Learn Res, 2010, 11: 2935–2972
66 Gopal S, Yang Y M, Bai B, et al. Bayesian model for large-scale hierarchical classification. In: Advances in Neural Information Processing Systems, South Lake Tahoe, 2012. 2420–2428
67 Qian J, Miao D Q, Zhang Z H, et al. Parallel attribute reduction algorithms using MapReduce. Inform Sci, 2014, 279: 671–690
68 Liang J Y, Wang F, Dang C Y, et al. An efficient rough feature selection algorithm with a multi-granulation view. Int J Approx Reason, 2012, 53: 912–926
69 Qian Y H, Liang J Y, Pedrycz W, et al. Positive approximation: an accelerator for attribute reduction in rough set theory. Artif Intell, 2010, 174: 597–618
70 Chen C L, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inform Sci, 2014, 275: 314–347
71 Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets. Science, 2011, 334: 1518–1524
72 Liang Y J, Qian Y H. Information granules and entropy theory in information systems. Sci China Ser E-Inf Sci, 2008, 38: 2048–2065 [梁吉业, 钱宇华. 信息系统中的信息粒与熵理论. 中国科学 E 辑: 信息科学, 2008, 38: 2048–2065]

Theory and method of granular computing for big data mining

LIANG JiYe[1,2*], QIAN YuHua[1,2], LI DeYu[1,2] & HU QingHua[3]

1 *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006*, China*;
2 *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China*;
3 *School of Computer Science and Technology, Tianjin University, Tianjin* 300072*, China*
*E-mail: ljy@sxu.edu.cn

Abstract The external form of big data often presents large-scale, multiple modal, and growth characteristics. In this paper, we discuss and analyze the challenges in data mining from the viewpoint of big data; these challenges include computability, effectiveness, and efficiency. Granular computing is an effective method for solving complex problems for intelligent information processing. By analyzing the feasibility of large data analysis based on granular computing, we argue that granular computing shows great promise as a new way for data mining in the context of big data. We also analyze several important problems in data mining based on granular computing, and the results will lead to further interpretations and developments in the field of big data mining.

Keywords big data, data mining, pattern discovery, granular computing, information granulation, multigranulation

LIANG JiYe was born in 1962. He received the Ph.D. degree in Information Science from Xi'an Jiaotong University. He also has a B.S. in computational mathematics from Xi'an Jiaotong University. He is a professor at the School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education at Shanxi University. His research interests include artificial intelligence, granular computing, data mining, and knowledge discovery.

LI DeYu was born in 1965. He received the M.S. degree from Shanxi University in 1998 and the Ph.D. degree from Xi'an Jiaotong University in 2002. He is a professor at the School of Computer and Information Technology and Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education at Shanxi University. His current research interests include rough set theory, granular computing, data mining, and knowledge discovery.

QIAN YuHua was born in 1976. He received the M.S. degree and the Ph.D. degree in Computers with Applications at Shanxi University in 2005 and 2011, respectively. He is a professor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He is actively pursuing research in pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He is best known for multigranulation rough sets in learning from categorical data and granular computing.

HU QingHua was born in 1976. He received his B.S., M.E., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. He started working with the Harbin Institute of Technology from 2006, and was a post-doctoral fellow with the Hong Kong Polytechnic University from 2009 to 2011. In 2012, he joined Tianjin University. At present, he is a full professor at the School of Computer Science and Technology, Tianjin University, and the Head of Department of Computer and Information Technology and the Leader of Lab of Pattern Analysis and Computational Intelligence (PANDIT). His research interests are focused on pattern analysis from data and machine learning for classification and regression.