

遥感大数据的智能处理：现状与挑战

宋维静¹, 刘 鹏¹, 王力哲¹, 吕 科²

(1. 中国科学院遥感与数字地球研究所, 北京 100094;

2. 中国科学院大学工程管理与信息技术学院, 北京 100049)

摘 要: 随着遥感技术的发展, 遥感数据空间分辨率、时间分辨率、光谱分辨率和辐射分辨率越来越高, 数据类型越来越丰富, 与此同时, 数据量也越来越大。遥感大数据向传统的遥感数据智能处理提出了新的挑战。本文从信息处理的角度论述了遥感大数据的海量、异构、多源的外部特征与高维度、多尺度、非平稳的内部特征, 阐述了遥感大数据智能处理在算法方面所涉及的关键技术, 分析了遥感大数据稀疏表征字典训练与稀疏求解方面的研究现状与存在的问题, 讨论了遥感大数据数据挖掘面临的挑战, 阐述了遥感大数据知识发现数据降维方面现有的解决方法与难题, 并对相应的问题进行了概括性讨论。最后, 总结了遥感大数据智能处理目前的研究阶段与研究意义。

关键词: 遥感大数据; 智能处理; 稀疏表征; 数据挖掘; 知识发现

中图分类号: TP7

文献标识码: A

文章编号: 1674-4969(2014)03-0259-07

引言

遥感数据目前已成为政府部门与科研机构快速、直接获取地物信息的重要手段, 具有无需接触探测目标即可实现对地物进行量测和解释的特点。尤其近些年来, 随着国土资源调查与开发利用的不断深入, 农业、林业等领域的定量遥感监测需求不断提高, 以及地震、地质灾害的频繁发生, 遥感数据在国土资源测绘、资源调查、环境监测、粮食估产及灾害评估等领域都得到了非常广泛的应用。这些现实需求促使遥感数据的获取逐渐向多源化、分布式的方向发展。

多源海量遥感信息的协同处理是依托于多种类型航天、航空遥感平台, 利用可见光、红外及微波等多种探测手段, 多方式获取对地观测数据并进行综合型信息处理的过程。同时, 随着对地

观测领域内多种传感器的发展、海量多源遥感数据获取能力的提高, 专业化的数据获取与多源遥感数据的综合应用呈现出精细化分工与协同式综合处理需求并存的发展局面。一方面, 获取遥感数据的传感器更为专业化与精细化, 传感器的地面分辨率数量级从千米到厘米, 波谱范围从紫外到超长波, 时间间隔从十几天一次到每天几次, 探测深度从几米到几万米。例如, 高分一号卫星的空间分辨率已经达到 2 m, 高光谱数据的波段数达到 200 多个。高分辨率、高动态的新型卫星传感器波段数量多、光谱分辨率高、数据下传速率高、重访周期短、数据量特别大。另一方面, 遥感领域的大型综合型应用需要多源遥感数据的支持, 同时面对不同类型传感器数据的处理、不同数据中心之间数据类型与覆盖区域的互补、综合性信息处理等现实问题。此外, 针对近实时处

收稿日期: 2014-07-24; 修回日期: 2014-08-05

作者简介: 宋维静 (1988-), 女, 博士研究生, 研究方向为空间信息处理。

刘 鹏 (1977-), 男, 副研究员, 研究方向为空间信息处理。

王力哲 (1974-), 男, 研究员, 研究方向为数字地球、空间信息处理与高性能地学计算。E-mail: Lizhe.wang@gmail.com

理的遥感应用,如灾害监测、环境预警等,随着遥感数据量的增多,需要处理的数据量成倍增加。因此,随着遥感数据量的急剧增加,遥感数据的快速智能处理是一个亟须解决的问题。

为实现遥感大数据的快速智能处理,应对复杂的遥感应用,近几年来主要从遥感数据处理平台方面寻求解决方案。美国的 NASA (National Aeronautics and Space Administration) 通过协同多个卫星数据中心,对各卫星地面站所接收的遥感数据进行预处理与深加工,实现遥感数据的快速智能分发与处理^[1]。我国也已基本形成了气象、陆地、海洋等三大类型的遥感卫星数据中心,借助于各个卫星数据接收站,可基本实现对我国疆域的全覆盖监测。我国“863”项目正在搭建多中心架构下多源遥感数据协同处理平台。虽然处理平台能够提高单位时间内遥感数据的吞吐量,但是针对遥感大数据内部组织结构多样、数据结构复杂、尺度不一致等特点,传统的遥感数据处理算法成为限制遥感数据处理效率的主要瓶颈。目前在遥感大数据处理算法与处理流程方面的研究还处于起步阶段,还没有形成大数据的标准处理流程,因此需要进行深入研究。

本文从遥感大数据算法处理角度,综述遥感大数据的特点;描述当前遥感大数据智能处理的主要理论、方法、技术,并讨论遥感大数据背景下所面临的问题,提出遥感大数据智能处理面临的挑战,讨论其可能的解决途径。

1 遥感大数据的特征

遥感大数据作为大数据的典型代表,已成为科学研究的重要途径,但在遥感大数据的机理模型及其在科学发现中的理论与方法等方面仍有待深入研究。遥感大数据不仅具有大数据的“4V”(volume、variety、velocity、value)特征,同时具有更丰富的大数据特征值得深入研究。

1) 外部特征:(1) 海量。遥感大数据具有海量数据特点^{[2]1049-1050},如 Landsat 系列卫星已积累

了近 30 年的遥感数据,此外,随着遥感技术的发展,目前高分辨率、高动态的新型卫星传感器波段数量多、光谱分辨率高、重访周期短、单位时间内获取的数据量特别大。(2) 异构。遥感大数据的异构性一方面表现为系统异构,即数据生产所依赖的业务应用系统等之间的不同,如来自不同的数据中心;另一方面表现为模式异构,即数据在逻辑结构或组织方式上的不同。(3) 多源。遥感大数据的多源性主要表现在数据来源和获取手段多样,既包括分布全球的观测网络实时接收的大量遥感数据,也包括航空拍摄获得的遥感数据,还包括民众用户通过互联网和带有地理信息的手持终端设备提供的个性化信息。

2) 内部特征:(1) 高维度性^{[2]1049-1050}。遥感大数据的数据类型多样化,数据的维度也越来越高,主要分为空间维度、时间维度、光谱维度等。光谱范围由可见光的几个波段发展到包含红外、近红外的上百个波段范围。时间维度方面,随着遥感技术的发展,同一地区形成按时间排序的一系列遥感影像。(2) 多尺度性。多尺度是遥感大数据的重要特征。这是由于遥感数据获取时,针对不同的遥感应用及遥感技术水平的约束,遥感数据在空间上与时间上呈现出多尺度的特征,如空间尺度由几千米到零点几米的范围变化。在不同的观察层次上所遵循的规律和体现的特征不尽相同。(3) 非平稳性。遥感大数据因为具有广泛的获取方式和物理意义,因此从信息理论来说是典型的非平稳信号,即分布参数或者分布律随时间发生变化。非平稳性正是经典遥感数据挖掘与分析理论所忽视的。

2 遥感大数据智能处理

遥感大数据研究的热潮激励着各个领域的科研人员寻找求解大数据的方法。例如,处理平台方面,采用高性能、大规模并行的海量遥感数据处理的算法模型与平台来提高海量遥感数据的处理能力,采用并行文件系统,通过并行处理提高

数据的处理效率。如 2004 年由 Google 公司提出的用来进行数据并行处理的 MapReduce^[3]、马艳等^[4]提出的数据密集型遥感数据并行处理平台均可以提高遥感数据处理效率。智能控制方面,提出通过机器学习、非线性随机的数据缩减技术,挖掘大数据中蕴含的信息。信息论方面,提出通过稀疏表征、流形学习等方法简化数据量与数据维度,使大数据变小后再进行后续研究,如利用稀疏表征的方法进行分类、目标识别与遥感数据融合等研究,利用流形学习的方法进行数据降维研究等。利用数据统计的方法挖掘大数据中蕴含的信息成为网络大数据处理的一种主要手段,如经典的“啤酒与尿布”案例、社交网站的好友分析等。

3 遥感大数据信息挖掘

在遥感大数据信息挖掘领域,同样也是各种新成果不断涌现。下面主要从算法处理角度分别综述各个方面的研究现状。

3.1 遥感大数据稀疏表征

大数据由于海量、高维、动态、异构等特征,使数据处理面临的问题更加复杂和困难,以至于大数据所包含的丰富的信息和复杂的先验知识不能直接在地学分析中利用。我们难以直接让所有的数据参与运算,很多情况下我们需要基于数据所蕴含的本质特征或主要属性来分析和解决问题。为此,我们需要先进的数据表征理论和方法。近年来,稀疏表征在数据获取(如压缩感知)^[5]和数据分析(如深度学习)^[6]方面取得了明显的进展。稀疏表征主要涉及两个方面的研究:基函数(基函数的集合也称为字典)的构建和稀疏分解的方法研究。在非解析超完备的冗余字典构建的研究领域,这两方面的研究有一定的耦合性,因为非解析字典的训练过程往往也涉及稀疏编码。

但是一般来说,字典的构建和稀疏编码大体上仍可以看作稀疏表征研究的两个最重要的方面。

关于构建字典的研究,按照字典的存在形式可以分为解析字典和非解析字典;按照字典对数据的适应性可以分为自适应字典和非自适应字典。解析字典计算速度快,构造和逼近性质都有较为严格的证明。解析字典方面,早期的 Fourier 变换是多分辨率的,但不具备局部特性。wavelet^[7-9]继承了 Fourier 变换多分辨率的特性,同时又有很好的局部特性,但是, wavelet 方向性的缺失导致其在逼近边高维奇异特征的时候不具备明显的稀疏性。curvelet^[10]及其离散形式 contourlet^[11]和高维形式 surfacelet^[12]在继承多分辨率和局部化特性的同时,还具方向性,因此对边缘等高维奇异特征有很好的表征能力,但是它们都不是自适应的。bandelet^[13]是自适应的字典,它是基于数据本身的特征以最终的优化结果自适应的构造基函数。总的来说,解析字典的发展经历了多分辨率、局部化、各向异性和自适应等几个阶段,变得越来越有效。非解析字典的原子没有固定的解析形式,但对数据的适应性更强,一般能够更加稀疏地表征数据。而且,非解析字典是从数据中训练得到,因此从一开始就是自适应的。早期非解析字典构造方法(如 Karhunen-Loeve transform, KLT)只考虑了数据的线性关系,MOD(modification)函数^[14]可能是最早面向稀疏表征的非解析字典训练模式。GPCA(generalized PCA)^[15]是 PCA(personal computer assistant)的扩展,把数据的表征理解为多个子空间的联合。K-SVD(K-singular value decomposition)^[16]在模型上与 MOD 函数类似,通过改进、更新原子的策略,速度、效果都获得了显著提高。显然,对于自适应的非解析字典,难以一次性让所有数据参与运算,所以这些方法面对大数据的时候都有明显的困难。近期提出的非

美国沃尔玛连锁超市的真实案例:尿布与啤酒这两种风马牛不相及的商品居然摆在一起,但这一奇怪的举措居然使尿布和啤酒的销量都大幅增加。

参数贝叶斯推理算法 (nonparametric bayesian dictionary learning, NBDL)^[17]、在线字典学习 (online dictionary learning, ODL)^[18]和约束最小二乘算法 (restricted least squares, RLS)^[19]等方法都可以让数据分批的进入训练的模型, 能够处理规模较大的数据, 但是尚不能处理真正的大数据。

字典构造的研究经历了解析—非解析—非自适应—自适应几个阶段, 越来越向数据的本身特性靠拢, 稀疏表征能力越来越强, 但同时也带来计算量的负担和加剧了字典构建过程的复杂性。遥感大数据具有的海量、多源、异构、高维、非平稳和多尺度等特性对稀疏表征提出了新的要求。解析字典虽然速度快、理论较为严密, 但是对数据的表征能力不强, 而构造自适应的解析字典计算量非常大。非解析字典对数据的表征能力较强, 但是, 遥感大数据各种复杂特性使训练过程面临很多困难。比如, 不能一次让所有数据参与运算, 以及数据表征效率、原子产生的机制、原子的结构和原子规模的控制等方面还远远不能适应大数据。

在稀疏编码方面, 主要需要解决的是如何在保持稀疏性的前提下得到数据的表征系数。基于稀疏约束的不同形式进行分类, 一般可以认为有基于 L_0 范数的计算方法、基于 L_1 范数的计算方法和基于 L_p 范数的计算方法。基于 L_0 范数的优化方法主要是以匹配追踪和正交匹配追踪为代表的贪婪算法类方法 (及其变种或改进方法) 和以迭代硬阈值算法为代表的阈值类方法。基于 L_1 范数的凸优化方法方面, 由于 L_1 范数的优化是凸优化问题, 存在有效而精确的数值解法, 最具有代表性的是基的追踪 (basis pursuit, BP) 算法^[20]、谱梯度投影 (spectral projected gradient L_1 , SPGL1)^[21]等。另外, 较为流行的还有 LASSO (least absolute shrinkage and selection operator)^[22], 在解不是特别稀疏的条件下对 L_1 的凸优化有较为稳定的效果。 L_p 范数优化的松弛类方法方面, 非常典型的求解 L_p 范数优化问题的方法是 (focal

underdetermined system solver, FOCUSS)^[23]。该方法通过逐步修正解的局部能量, 使大多数元素趋于零。FOCUSS 能够处理相对规模较大的数据。另一种比较流行的 L_p 范数的优化方法是迭代复加权 (iterative RLS, IRLS)^[24]。将 L_0 、 L_1 和 L_p 范数的算法进行比较后发现, 在很多情况下这些算法的思想是互相借鉴、互相渗透的, 有的时候甚至没有严格的界限。

当我们面对遥感大数据的时候, 无论是字典的训练还是稀疏系数的求解都面临了新的问题。遥感大数据使传统的稀疏表征方法难以有效地对数据进行建模。传统的字典学习方法一般只需要针对较为单一的数据源, 设定确定的字典规模; 此外, 数据的规模不能太大, 变化不能太剧烈。面对海量、高维、动态、异构、多尺度的遥感大数据, 研究对象的复杂性使原子集合的构建, 以及如何高效稀疏地描述遥感大数据都非常困难。

3.2 遥感大数据挖掘

时空数据挖掘是指从时空数据中提取隐含的但又具有潜在价值的模式和知识的过程, 典型例子包括气象模式的发现^[25]、城市或土地利用演变模式的发现^[26]、地震或飓风等自然灾害的预测^[27]、传染病聚集性的发现^[28]、交通行为预测^[29]。

与传统的时空数据挖掘相比, 针对遥感大数据的数据挖掘具有一定的复杂性: 首先, 遥感大数据海量、高维性的特点使得数据挖掘的搜索空间异常巨大, 可能导致搜索出无意义模式的机会增加, 对传统的时空数据挖掘的理论与方法提出巨大的挑战^[30-35]; 其次, 遥感大数据虽然具有高维性, 但仅仅只有小部分维度与时空模式密切相关, 其余的不相关或相关性很小的维度会产生大量的噪声, 从而掩盖真实的时空模式; 最后, 遥感大数据所涉及的数据类型复杂多样, 数据格式不一, 结构化 (空间数据)、半结构化 (辅助数据) 数据并存, 并且由于遥感大数据信息更新频繁, 数据库可能随时增加、删除或者改变变量, 对时空数

据挖掘所依赖的数据库提出很高的要求^[34-35]。

3.3 遥感大数据知识发现

大数据的表征是为了解决对数据内在的结构特征和存在形式的建模问题。而机器学习的目的则是知识发现。虽然机器学习算法的构建过程仍然没有统一的理论,但是基于机器学习的知识发现过程是确定性与随机性的统一。在整个机器学习的研究领域,一直有两大研究方向:一是研究学习机制或途径;二是研究如何有效利用数据和信息。在大数据的背景下,与这两个方向相关的数据的降维、特征选择、模式分类和知识表达等方面都面临新的问题。

下面讨论大数据降维问题。大量数据资源在给人们带来价值的同时,也带来了很多问题。数据降维研究的目的是找出高维数据中所隐藏的低维结构和本质特征。遥感大数据显然具有很高的维数,并提供了有关客观现象的极其丰富、详细的信息。但是,一方面我们需要基于降维从大量的数据特征中找出其本质的特征,另一方面需要在降维后能保持数据集原有的一些本质特征不变。近年来,流形学习成为数据降维的非常重要的手段。关于流形学习方面,里程碑式的研究是2000年J. B. Tenenbaum等和S. T. Roweis等在《Science》上发表的两篇论文,他们提出了各自的流形学习算法:等距映射(Isometric Mapping, isomap)^[36-37]和局部线性嵌入(local linear embedding, LLE)^[6]。在iosmap和LLE之后,又发展出了许多新的流形学习算法,如拉普拉斯特征映射、海森局部线性嵌入(Hessian-based locally linear embedding, HLLE)、局部切空间排列(local tangent space alignment, TLSA)等^[37-42]。这些流形学习方法都有一些共同的特征:1)在每个样本点附近寻找一个邻域,并找出这个局部邻域的几何结构;2)利用这些收集的局部信息非线性的映射到一个低维空间上。大多数算法之间的不同之处在于收集的邻域的局部信息不同,以及如何利

用这些收集的信息来构造全局的嵌入。

然而,对于遥感大数据,我们难以直接使用传统的方法,尤其是超长时间序列数据分布的复杂性、高维性等因素决定了学习它的本征属性及其在高维空间中分布的内蕴结构是非常困难的。

4 结语

毋庸置疑,遥感大数据已经给我们带来了新的机遇与挑战。与一般的数据相比,遥感大数据具有的海量、异构、多源的外部特征与高维度、多尺度、非平稳的内部特征,使遥感数据智能处理面临的问题更加复杂和困难。因此,遥感大数据智能处理在算法处理方面亟须寻求新的解决途径。

本文对遥感大数据的特征进行了总结与描述,阐述了遥感大数据智能处理在算法方面所涉及的关键技术,综述了遥感大数据稀疏表征、数据挖掘、知识发现方面的研究现状、存在的问题,并对其中的问题进行了概括性讨论。

作为科学大数据的一个典型代表,遥感大数据处理还处于初级阶段^{[2]1051-1052},还有许多问题有待研究。但它作为一个新的研究对象,给遥感信息挖掘带来了新的机遇,为许多遥感应用提出了新的解决思路,其研究成果可能对遥感领域产生重大影响。

参考文献

- [1] 冯筠, 高峰, 黄新宇. 构建天地一体化的全球对地观测系统: 三次国际地球观测峰会与 GEOSS[J]. 地球科学进展, 2005, 20(12): 1327-1333.
- [2] 郭华东, 王力哲, 陈方, 等. 科学大数据与数字地球[J]. 科学通报, 2014, 59(12).
- [3] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM 50th anniversary issue, 2008, 51(1): 107-113.
- [4] Ma Y, Wang L Z, Liu D S, et al. Distributed Data Structure Templates for Data-Intensive Remote Sensing Applications [J]. CCPE, 2013, 25: 1784-1797.
- [5] Donoho D L. Compressed Sensing [J]. IEEE Transaction on Information Theory, 2006, 52(4): 1289-1306.
- [6] Roweis S T, Saul L K. Nonlinear Dimensionality Reduc-

- tion by Locally Linear Embedding [J]. *Science*, 2000, 290(5500): 2323-2326.
- [7] Daubechies B. Orthonormal Bases of Compactly Supported Wavelets [J]. *Comm. Pure Appl. Math*, 1988, 41(7): 909-996.
- [8] Mallat S G. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation [J]. *IEEE Trans. Patt. Anal. Machine Intell.*, 1989, 11(7): 674-693.
- [9] Mallat S, Hwang W L. Singularity Detection and Processing with Wavelets [J]. *IEEE Trans. Inf. Theo.*, 1992, 38(2): 617-643.
- [10] Cande's E J, Demanet L, Donoho D L, et al. Fast Discrete Curvelet Transforms [J]. *Multiscale Modeling & Simulation*, 2006, 5: 861-899.
- [11] Do M N, Vetterli M. The Contourlet Transform: An Efficient Directional Multi-Resolution Image Representation [J]. *IEEE Trans. Image Process*, 2005, 14(12): 2091-2106.
- [12] Lu Y M, Do M N. Multidimensional Directional Filter Banks and Surfacelets [J]. *IEEE Trans. Image Process*, 2007, 16(4): 918-931.
- [13] LePennec E, Mallat S. Sparse Geometric Image Representations with Bandelets [J]. *IEEE Trans. Image Process*, 2005, 14(4): 423-438.
- [14] Lesage S, Gribonval R, Bimbot F, et al. Learning Unions of Orthonormal Bases with Thresholded Singular Value Decomposition [C]// *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. IEEE*, 2005, 5: 293-296.
- [15] Vidal R, Ma Y, Sastry S. Generalized Principal Component Analysis (GPCA) [J]. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, 27(12): 1945-1959.
- [16] Aharon M, Elad M, Bruckstein A M. The K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation [J]. *IEEE Trans. Signal Process*, 2006, 54(11): 4311-4322.
- [17] Zhou M Y, Chen H J, Peisley J, et al. Nonparametric Bayesian Dictionary Learning for Analysis of Noisy and Incomplete Images [J]. *IEEE Trans. Img. Proc.*, 2012, 21(1): 130-144.
- [18] Mairal J, Bach F, Ponce J, et al. Online Dictionary Learning for Sparse Coding. *Journal of Machine Learning Research [C]// ICML'09 Proceedings of the 26th Annual International Conference on Machine Learning*. New York: ACM, 2009: 689-696.
- [19] Karl S, Kjersti E. Recursive Least Squares Dictionary Learning Algorithm [J]. *Trans.Sig. Proc.*, 2010, 58(4): 2121-2130.
- [20] Chen S S, Donoho D L, Saunders M A. Atomic Decomposition by Basis Pursuit [J]. *SIAM Rev.*, 2001, 43: 129-159.
- [21] Figueiredo M A T, Nowak R D, Wright S J. Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems [J]. *IEEE Journal of Selected Topics in Signal Processing*, 2007, 1(4): 586-597.
- [22] Zou H. The Adaptive Lasso and Its Oracle Properties [J]. *Journal of the American statistical association*, 2006, 101(476): 1418-1429.
- [23] Gorodnitsky I F, Rao B D. Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-Weighted Minimum Norm Algorithm [J]. *IEEE Transactions on Signal Processing*, 1997, 45(3): 600-616.
- [24] Daubechies I, Devore R, Fornasier M, et al. Iteratively Reweighted Least Squares Minimization for Sparse Recovery [J]. *Communications on Pure and Applied Mathematics*, 2010, 63(1): 1-38.
- [25] Ganguly A R, Steinhäuser K. Data Mining for Climate Change and Impacts [C]// *Proceeding of the 2008 IEEE International Conference on Data Mining Workshops (ICDMW'08)*. Piscataway, NJ: IEEE, 2008: 385-394.
- [26] Tang J, Wang L, Yao Z. Spatio-Temporal Urban Landscape Change Analysis Using the Markov Chain Model and a Modified Genetic Algorithm [J]. *International Journal of Remote Sensing*, 2007, 28(15): 3255-3271.
- [27] Hsu K C, Li S T. Clustering Spatial-Temporal Precipitation Data Using Wavelet Transform and Self-Organizing Map Neural Network [J]. *Advances in Water Resources*, 2009, 33(2): 190-200.
- [28] 廖一兰, 王劲峰, 杨维中, 等. 传染病多维度聚集性探测方法[J]. *地理学报*, 2012, 67(4): 435-443.
- [29] 徐薇, 黄厚宽, 秦勇. 基于时空数据挖掘的铁路客流预测方法[J]. *北京交通大学学报*, 2004, 28(5): 16-19.
- [30] Roddick J F, Hornsby K, Spiliopoulou M. An Updated Bibliography of Temporal, Spatial, and Spatio-Temporal Data Mining Research [M]// Roddick J F, Hornsby K. *Temporal, Spatial, and Spatio-Temporal Data Mining: Lectures Notes in Computer Science*. Springer Berlin Heidelberg, 2001, 2007: 147-163.
- [31] 宫学庆, 金澈清, 王晓玲, 等. 数据密集型科学与工程: 需求和挑战[J]. *计算机学报*, 2012, 35(8): 1563-1578.
- [32] 周傲英, 周敏奇, 宫学庆. 计算广告: 以数据为核心的 Web 综合应用[J]. *计算机学报*, 2011, 34(10): 1805-1819.
- [33] 舒继武, 郑纬民, 沈美明, 等. 大规模问题数据并行性

- 能的分析[J]. 软件学报, 2000, 11(5): 628-633.
- [34] 王树良, 丁刚毅, 钟鸣. 大数据下的空间数据挖掘思考[J]. 中国电子科学研究院学报, 2013, 8(1): 8-17.
- [35] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯, 2012, 8(9): 8-15.
- [36] Tenenbaum J B, Silva V D, Langford J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science, 2000, 290(5500): 2319-2323.
- [37] 孟德宇, 徐晨, 徐宗本. 基于 Isomap 的流形结构重建方法[J]. 计算机学报, 2010, 33(3): 545-555.
- [38] 詹德川, 周志华. 基于流行学习的多实例回归算法[J]. 计算机学报, 2006, 29(11): 1948-1955.
- [39] 孟德宇, 徐宗本, 戴明伟. 一种新的有监督流形学习方法[J]. 计算机研究与发展, 2007, 42(12): 2072-2077.
- [40] Donoho D L, Grimes C. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data [J]. Proceedings of the National Academy of Sciences, 2003, 100(10): 5591-5596.
- [41] Zhang Z, Zha H. Principal Manifolds and Nonlinear Dimensionality Reduction Via Tangent Space Alignment [J]. Journal of Shanghai University (English Edition), 2004, 8(4): 406-424.
- [42] Hinton G E, Ruslan R S. Reducing the Dimensionality of Data with Neural Networks [J]. Science, 2006, 313(5786): 504-507.

Intelligent Processing of Remote Sensing Big Data: Status and Challenges

Song Weijing¹, Liu Peng¹, Wang Lizhe¹, Lü Ke²

(1. Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China;

2. College of Engineering & Information Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: With the development of remote sensing technology, the spatial resolution, the temporal resolution, the spectral resolution and the radiometric resolution of remote sensing data becomes higher, meanwhile the amount of data is becoming larger. This paper focuses on the new challenges brought by remote sensing big data to the traditional large remote sensing data processing. Firstly, this article discusses the external features and the internal features of remote sensing big data from the perspective of information processing. External features include massive, heterogeneous, multi-source and internal features are high-dimensional, multi-scale, and non-stationary. Secondly, we analyze the research status and problems of large data sparse representation, data mining and knowledge discovery. Finally, we discuss the remote sensing big data intelligent processing problems. As a typical representative of the science of big data, remote sensing big data is still in its infancy; therefore, there are many issues to be studied.

Keywords: remote sensing big data; intelligent processing; sparse representation; data mining; knowledge discovery