地球科学中的大数据分析与挖掘算法手册

Handbook of Big Data Analysis and Mining Algorithms in Earth Sciences

李国庆　刘莹　庞禄申　等 编著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

国之重器出版工程

网络强国建设

学术中国·大数据

地球科学中的大数据分析与挖掘算法手册

Handbook of Big Data Analysis and Mining Algorithms in Earth Sciences

李国庆　刘莹　庞禄申　等　编著

人民邮电出版社

北京

图书在版编目（CIP）数据

地球科学中的大数据分析与挖掘算法手册 / 李国庆等编著. -- 北京 : 人民邮电出版社, 2018.8
（学术中国·大数据）
ISBN 978-7-115-47855-9

Ⅰ. ①地… Ⅱ. ①李… Ⅲ. ①地球科学－数据处理－手册②地球科学－数据采集－手册 Ⅳ. ①P-62

中国版本图书馆CIP数据核字(2018)第151669号

内 容 提 要

本书以数据分析与挖掘思想为主线，深入剖析关联、分类、回归、聚类、序列模式挖掘、深度学习以及异常检测等算法的原理、实现、相似算法、改进思路以及地学案例，具有很强的系统性、完整性以及落地性，可以作为各行业特别是地球科学领域中希望驾驭大数据并发掘其价值的科研人员和工程人员的参考书。读者既可以通过本书系统掌握大数据分析挖掘的思想方法，也可以将其作为算法工具书查阅。

◆ 编　　著　李国庆　刘　莹　庞禄申　等
责任编辑　唐名威
责任印制　杨林杰
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷
◆ 开本：710×1 000　1/16
印张：19.75　　2018年8月第1版
字数：370千字　　2018年8月河北第1次印刷

定价：149.00元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

《国之重器出版工程》
编辑委员会

专家委员会委员（按姓氏笔画排列）：

于　全　中国工程院院士

王少萍　“长江学者奖励计划”特聘教授

王建民　清华大学软件学院院长

王哲荣　中国工程院院士

王　越　中国科学院院士、中国工程院院士

尤肖虎　“长江学者奖励计划”特聘教授

邓宗全　中国工程院院士

甘晓华　中国工程院院士

叶培建　中国科学院院士

朱英富　中国工程院院士

朵英贤　中国工程院院士

邬贺铨　中国工程院院士

刘大响　中国工程院院士

刘怡昕　中国工程院院士

刘韵洁　中国工程院院士

孙逢春　中国工程院院士

苏彦庆　“长江学者奖励计划”特聘教授

苏哲子　中国工程院院士

李伯虎　中国工程院院士

李应红　中国科学院院士

李新亚　国家制造强国建设战略咨询委员会委员、中国机械工业联合会副会长

杨德森　中国工程院院士

张宏科　北京交通大学下一代互联网互联设备国家工程实验室主任

陆建勋　中国工程院院士

陆燕荪　国家制造强国建设战略咨询委员会委员、原机械工业部副部长

陈一坚　中国工程院院士

陈懋章　中国工程院院士

金东寒　中国工程院院士

周立伟　中国工程院院士

郑纬民　中国计算机学会原理事长

郑建华　中国科学院院士

屈贤明　国家制造强国建设战略咨询委员会委员、工业和信息化部智能制造专家咨询委员会副主任

项昌乐　“长江学者奖励计划”特聘教授，中国科协书记处书记，北京理工大学党委副书记、副校长

柳百成　中国工程院院士

闻雪友　中国工程院院士

徐德民　中国工程院院士

唐长红　中国工程院院士

黄卫东　“长江学者奖励计划”特聘教授

黄先祥　中国工程院院士

黄　维　中国科学院院士、西北工业大学常务副校长

董景辰　工业和信息化部智能制造专家咨询委员会委员

焦宗夏　“长江学者奖励计划”特聘教授

《学术中国·大数据》丛书
编辑委员会

《地球科学中的大数据分析与挖掘算法手册》
编 写 组

组　长：

李国庆　刘　莹　庞禄申

成　员：

邹自明　胡晓彦　向　超　吴林志　钟　佳

刘锦怡　崔红元　崔辰州　许允飞　李风朋

张　磊　赵正健　赵　硕

序

大数据正在影响着人类社会和经济活动的模式，同时也正推动着科学的发展。大数据技术与应用已成为继实验、理论和计算模式之后的数据密集型科学范式的典型代表，带来了科研方法论的创新。科学大数据将复杂性、综合性、全球性和信息与通信技术高度集成性等诸多特点融于一身，其研究方法也正在从单一学科向多学科与跨学科方向转变，从自然科学向自然科学与社会科学的充分融合方向过渡，从个人或者小型科研团体向国际科技组织方向发展。另外，科学家不仅通过对广泛的数据进行实时、动态的监测与分析来解决科学问题，更是将数据作为了科学研究的对象和工具，即数据驱动的知识发现。这正是科学大数据的核心价值所在。

大数据正在驱动“数字地球”的发展。“数字地球”是国际上 1998 年提出的概念，它将航天航空对地观测技术、地理空间信息技术、计算机网络通信技术等与地球科学高度综合集成，实现模拟地球表层变化、支持政府决策、开展数据共享等重大目标。大数据的诞生与发展为“数字地球”研究注入了新的科学推动力。新一代数字地球是利用海量、多分辨率、多时相、多类型对地观测数据和社会经济数据及其分析算法和模型构建的虚拟地球，是科学大数据的典型学科，而数字地球学科中的数据获取与组织、分析、应用均体现了科学大数据的重要特征。约 20 年前，我们曾将数字地球通俗地解释为“把地球装入计算机”，而在当今大数据时代，我们则可以认为数字地球就是地球大数据。我们可通过在数字地球平台上对海量空间数据和社会经济数据进行高效的组织，从而在更丰富的数据空间进行科学信息挖掘和分析。

地球大数据涉及陆地、大气、海洋、天文、空间等地球系统科学各基础学科以及遥感、导航、地理信息系统、网络、高性能计算、虚拟现实等技术学科。建立基于地球大数据的新型研究范式，需要掌握这些数据的获取、传输、保存、管理、共享、处理和分析等全生命周期的特点，尤其要突破地学大数据的信息挖掘技术。揭示隐藏在海量观测数据、模拟数据和再分析数据中的内在知识是我们利用大数据的根本目标，近20年来数据挖掘方法成为人类驾驭数据的主要途径，但是地学大数据场景下的数据挖掘方法一方面要适应大数据的特征，另一方面还要适应地学数据的特征，需要创造性地继承和发展传统的数据挖掘方法。

李国庆、刘莹等专家针对地球科学和大数据快速发展的趋势，编著了《地球科学中的大数据分析与挖掘算法手册》一书。该书基于地学各学科中数据挖掘方法的使用经验，结合地球科学的具体应用，根据地学数据和信息特点，系统性地对数据挖掘算法进行了梳理，按照“数学方法—算法原理—算法发展—大数据适应性—地学适应性”的思路对算法模型进行比较分析，是一部适时的有前瞻性的著作，对于地学研究人员有重要的参考价值和有益的导向作用。

我有幸先读为快，并向读者推荐该书，有理由相信地学学者和大数据学者可以借此来发展和丰富地球科学领域的大数据挖掘方法，更好地驾驭大数据战略资源，掌握、揭示更多的地球系统科学规律。我同时呼吁大家共同关注地球大数据，让地球大数据成为地球科学发展之光，让地球大数据成为人类认识地球的新钥匙。

郭华东*

2017年4月10日于北京

* 中国科学院院士、俄罗斯科学院外籍院士、发展中国家科学院院士

序二

随着数据采集、存储技术的迅猛发展，以及软件应用技术的日新月异，海量的、复杂的、动态变化的数据——“大数据”已成为我国未来十年的重要科研方向和支柱产业。长期以来，人们常常使用数据挖掘的方法从海量的数据中寻找数据间存在的内在规律、模式、隐藏的知识和价值。然而，在大数据的背景下，已有的这些数据挖掘方法和技术在计算效率、复杂度、可扩展性、适用性等方面都存在较大的局限性，不能满足大数据挖掘的实际应用要求。因此，大数据挖掘是当下最为活跃的研究领域，新算法层出不穷，也是未来计算机信息领域的重要研究方向。

相对于一般的商业、事务、互联网等数据，地学数据更为复杂，涉及地理、天文、空间、大气、海洋、生态、地质等基础学科，包含位置数据、属性数据、空间关系数据等。随着遥感技术的发展，数据的类型也越来越复杂，既有矢量数据，也有栅格数据，还有海量的遥感影像数据。而目前的大数据挖掘研究较少考虑地学数据独有的特征。因此，将计算机科学的大数据挖掘原理、方法、技术与地学数据的特征紧密地结合起来，是一项十分有意义的工作。人们期待着能够从汪洋大海般的地学数据中及时有效地挖掘出有用的、高附加值的知识，为决策服务。

本书的可贵之处在于，这是第一本关于地学大数据挖掘的书籍。撰写团队集多年来的数据挖掘、地学数据分析的智慧与学术界的优秀成果于一体，熔炼成一本算法手册，强调先进性、系统性、可读性、可操作性。本书的可读之处颇多，既包含面向地学大数据的关联规则挖掘、分类、预测、聚类等算法，又包含崭新的深度学习技术、遥感图像识别技术等，并附以多个地学应用大数据的实例等。纸短笔陋，

实难尽述。

通过阅读本书，计算机科学的读者可以了解地学知识、地学数据特点；地球科学的读者可以学习到数据挖掘技术，特别是大数据挖掘的技术；数据挖掘专业的读者可以发现地学大数据的独特魅力；大数据专业的读者可以发现潜在研究空间和发展空间。总之，阅读本书，无论对数据挖掘的学习者，还是对地学研究人员，还是对大数据、数据挖掘软件开发人员都是大有裨益的。

石　勇*

2017 年 4 月 10 日于北京

* 发展中国家科学院院士

前 言

数据挖掘是自动地从海量的数据中找到新颖的、有价值的、隐藏的知识的技术，是近十几年来信息科学领域最活跃、最重要的一个交叉学科。研究成果包括关联规则算法、分类模型、聚类模型、顺序模式算法、异常值检测算法、推荐系统等。如今，数据挖掘技术已在商业智能、客户关系管理、互联网、基因工程、科学数据分析、地理信息系统、安全监控、军事国防等领域得到成功应用。

然而，由于大数据具有“3V”特征，即数据量极大（通常指TB级以上）、类型复杂多样、数据变化快且实时性强，已有的数据挖掘算法和技术正在面临挑战。首先，由于很多算法的复杂度是非线性的，处理“大数据”的速度无法满足用户的需要。其次，目前已有的复杂类型数据的挖掘方法，例如图挖掘、空间数据挖掘、文本挖掘、网页挖掘等，所能处理的数据类型十分有限，特别是高维数据，挖掘的准确性和效率有明显的局限性。另外，对于实时数据流应用，例如流媒体数据、实时监控数据等，目前的数据流挖掘技术还不能满足其实时性要求。

与一般数据相比，地球科学数据更为复杂。它涉及地理、天文、空间、大气、海洋、生态、地质等基础学科，包含位置数据、属性数据、空间实体关系数据等，具有空间性、时间性、高维性、海量性、复杂性、不确定性。目前的大数据挖掘研究较少考虑地球科学数据的独有特征。而且，随着遥感技术的发展，除了矢量数据、栅格数据，还出现了海量的遥感影像数据。尽管以深度卷积神经网为代表的深度学习模型已在图像识别和分类应用中获得成功，但鉴于遥感影像比普通图像分辨率低、

噪声多、数据量更大，目前还处于研究的起步阶段。

国际科学数据委员会（CODATA）中国委员会汇集了我国一大批科学数据领域的专家，为大家提供了很好的学术交流平台。参与本书撰写的几个专家都是CODATA中国委员会的活跃分子：中国科学院大学的刘莹教授团队长期从事大数据挖掘算法理论研究；中国科学院国家天文台的崔辰州研究员团队长期进行天文大数据研究；国家空间科学中心的邹自明研究员团队长期从事空间科学大数据研究；中国科学院遥感与数字地球研究所的李国庆研究员团队长期在地球观测大数据方面开展研究。近年他们都主持和参加了国家相关部门组织的与地球科学大数据相关的大量研究项目，在天文、空间、遥感、地理、地质、生态等领域初步进行了大数据研究，开展了地球科学大数据相关的数据管理、共享、分析、挖掘工作，并为公众和科学家提供了多种大数据服务。在科研交流中，我们有一个共同的困惑和压力，那就是缺少适合于地球科学研究的大数据分析和挖掘算法，特别是刚刚开始进行地学大数据研究的学者往往困惑于该如何下手选择算法，现在已出版的书籍和文献往往无法说清楚哪些算法是适合于大数据挖掘的。编著这本书就是我们尝试回答这个问题的一个探索。

本书不是特定科研项目的产出，而是撰写团队对于过往工作的总结和反思。在撰写任务分工上，刘莹团队（成员包括向超、吴林志、刘锦怡、崔红元）负责关联规则、*K*-最近邻分类、基于层次的聚类、基于网格的聚类、序列模式挖掘、卷积神经网络以及自动编码器算法的撰写，邹自明团队（成员包括胡晓彦、钟佳）负责决策树分类、贝叶斯分类、粗糙集分类、神经网络分类、支撑向量机、线性回归、*K* 均值、*K*-medoids 以及基于密度的聚类算法的撰写，李国庆团队（成员包括庞禄申、李风朋）负责集成学习、逻辑回归、深度信念网以及异常检测算法的撰写，崔辰州团队（成员包括许允飞、张磊）参与了决策树、支撑向量机以及 *K*-means 分类算法的撰写，中电科海洋信息技术研究院有限公司的赵正建、赵硕参与了 Bagging 算法的撰写，中国科学院遥感与数字地球研究所的刘鹏副研究员对全部算法进行了严格的学术审查，中国科学院遥感与数字地球研究所的庞禄申博士研究生为本书的统稿和撰写组织做了大量工作，在中国科学院遥感与数字地球研究所进行客座研究的中国地质大学的李风朋同学为本书的成书做了大量辅助性工作。

本书的出版得到了国家重点研发计划“地球观测与导航”专项“地球资源环境动态监测技术”项目之课题四“多源遥感监测数据在线融合及协同分析云平台

（编号：2016YFB0501504）”以及中国科学院数字地球重点实验室主任基金（编号：1108000001）的支持。特别感谢中国科学院院士郭华东和中国科学院大学的石勇研究员对这项工作的指导，并为本书作序。在本书的撰写过程中，得到了许多专家的帮助和支持，无法一一列举，在此谨向他们表示诚挚的感谢。

我们希望在论述中可以尽可能覆盖到目前本领域的主要研究进展，所以本书引用了大量国内外学者的研究成果，书中都一一进行了标注，在此一并表示感谢。由于作者水平有限，部分内容和作者观点可能有不妥之处，恳请读者批评指正。

李国庆　刘　莹

2017 年 4 月 3 日于北京

目 录