- 中山大学重点学科建设成果
- 中国矿物岩石地球化学协会
  大数据与数学地球科学专业委员会推荐

地球科学
大数据挖掘与机器学习

Big Data Mining & Machine Learning in Geoscience

周永章　张良均　张奥多　王　俊◎著

非外借

中山大學出版社
SUN YAT-SEN UNIVERSITY PRESS

- 中山大学重点学科建设成果
- 中国矿物岩石地球化学协会
  大数据与数学地球科学专业委员推荐

地球科学
大数据挖掘与机器学习

Big Data Mining & Machine Learning in Geoscience

周永章　张良均　张奥多　王　俊◎著

中山大學出版社
SUN YAT-SEN UNIVERSITY PRESS
·广州·

版权所有　翻印必究

图书在版编目（CIP）数据

地球科学大数据挖掘与机器学习/周永章，张良均，张奥多，王俊著．—广州：中山大学出版社，2018.9

ISBN 978-7-306-06409-7

Ⅰ．①地…　Ⅱ．①周…　②张…　③张…　④王…　Ⅲ．①地球科学—数据采集—教材　②地球科学—机器学习—教材　Ⅳ．①P-39

中国版本图书馆 CIP 数据核字（2018）第 177989 号

DIQIU KEXUE DASHUJU WAJUE YU JIQI XUEXI

出 版 人：王天琪
策划编辑：曾育林
责任编辑：曾育林
封面设计：曾育林
责任校对：马霄行
责任技编：何雅涛
出版发行：中山大学出版社
电　　话：编辑部 020-84111996，84113349，84111997，84110779
　　发行部 020-84111998，84111981，84111160
地　　址：广州市新港西路 135 号
邮　　编：510275　　传　真：020-84036565
网　　址：http://www.zsup.com.cn　　E-mail:zdcbs@mail.sysu.edu.cn
印 刷 者：广州家联印刷有限公司
规　　格：787mm×1092mm　1/16　17.5 印张　600 千字
版次印次：2018 年 9 月第 1 版　2018 年 9 月第 1 次印刷
定　　价：49.80 元

如发现本书因印装质量影响阅读，请与出版社发行部联系调换

简　介

本书系统地介绍了地球科学大数据挖掘与机器学习的基本框架与原理，重点分析高维数据降维、分类与预测、大图形社区结构识别、无限流数据处理、机器学习及人工智能地质学的建模过程，对必要的应用场景使用 Python 语言给出案例。本书是中山大学研究生试用研究型教材，对运用大数据挖掘技术与机器学习算法解决地球科学问题大有裨益。适合地球科学领域研究生和高年级本科生做教材，也可供科研人员做研究时参考。

序

2007年，图灵奖得主吉姆格瑞发表演讲时指出：大数据已经成为科学研究的第四范式。人类在科学研究的道路上，从经验科学，到理论科学，再到计算科学，到如今的大数据科学，大数据成为第四范式也是必然之路。

在大数据时代，人类的思维方式必然会产生革命性的变革。大数据挖掘特别适合于窥探具有多维性和全面性的现实世界。它可以从很多看似支离破碎的信息中复原一个事物的全貌，并进而能够预测或判断出尚未观察到的事物的现象。

大数据存在于任何行业和领域。通过大数据挖掘，可以发现事物运行和发展的规律。

大数据分析是今后各学科和经济社会领域不可回避的重大课题。美国政府认为大数据是“未来的新石油”，一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制将成为国家间和企业间新的争夺焦点。中国政府于2015年9月印发《促进大数据发展行动纲要》，明确推动大数据发展和应用。中国科学院、复旦大学、中山大学、中国航空航天大学等相继成立了从事数据科学研究的专门机构。

地球科学领域广泛存在大数据。地质调查是地球科学研究获得数据的主要渠道之一。同其他行业和领域一样，地球科学大数据正在以指数形式增长。在这一背景下，地球科学大数据挖掘，日益获得越来越多的地球科学家的重视。超级计算硬件、软件的发展，为地球科学家研究大数据提供了比任何时候都更方便的平台条件。各国长期地质调查和探测取得的海量地质基础调查数据，正在成为超级计算机服务的重点对象之一。

机器学习是应对大数据超常增长、开展大数据信息挖掘的重要选项。它被认为是人工智能的核心，是使计算机具有智能的根本途径，而深度学习恰是一种炙手可热的实现机器学习的技术。因而，它们一起成为当前大数据与数学地球科学研究的重点和热点。尽管依托大数据的人工智能地质学还远不够成熟，但所幸运的是，具有历史使命感的科学家在严肃、认真地探索。

在当今新时代，大数据的广泛应用必将为地质科学的研究增加新的工具，并必将因此改变地质。

试读结束：需要全本请在线购买： www.ertongbook.com

本书在上述认识指引下撰写而成，曾作为中山大学地球科学与工程学院以及国际数学地球科学协会中山大学学生分会（IAMG Student Chapter at SYSU）内部试用教材。2016 年，中国矿物岩石地球化学学会大数据与数学地球科学专业委员会成立，培训新人也是使用本教材。

本书在试教和撰写过程中，大量参考了近几年来学者发表的论文。整个过程得到中国矿物岩石地球化学学会大数据与数学地球科学专业委员会各位委员以及中山大学选修《大数据与数学地球科学》研究生的支持和帮助。谷鸿飞、徐述腾参与了初稿部分章节的撰写。

本书获国家重点研发计划重点专项“深部矿产资源评价理论与方法”项目（2016YFC0600506）、中国地质调查局（12120113067600）、国家自然科学基金（41273040）及广东省地质过程与矿产资源探查实验室开放基金的联合资助，获国际数学地球科学协会（International Association for Mathematical Geology，IAMG）Felix Chayes 奖 7000 美元奖金支持。

中山大学地球科学与工程学院、中山大学国家超级计算机中心、中山大学地球环境与地球资源研究中心、广州泰迪智能科技有限公司、广州高质大数据科技有限公司，以及翟明国、张旗、严光生、肖凡、侯卫生、王正海、王树功、沈文杰、何俊国、吴冲龙、刘刚、朱月琴、郭艳军、成秋明、陈建国、毛先成、路来君、刘洁、刘玉葆、周可法、张雪英、杨永国、高乐、焦守涛、刘艳鹏、张尚佳等对本书的出版给予了不同形式的支持和帮助。

本书适合地球科学领域研究生和高年级本科生做教材，也可供科研人员做研究时参考。

目　录

第1章 绪 论

大数据正在引发地球科学领域的一场深刻的革命，大数据的关键不仅在于数据的大，更在于思维的创新。从数据出发，让数据说话，依靠人工智能方法，让机器学习、深度学习等大数据技术逐步成为必需。大数据作为第四科学范式，研究领域十分宽广，它将改变地球科学家的思维方式，从逻辑思维方式转变为由数据驱动的关联思维方式。

1.1 科学研究第四范式

在科学发展史上，人类经历过四次重要的范式变革。

1.1.1 第一范式

经验科学阶段。在18世纪，科学研究的核心特征是对有限的客观对象进行观察、总结，用归纳法找出其中的科学规律，比如伽利略的物理学定律。归纳法对发生的事件进行总结，形成科学的认识，但它只用于已有规律的认识，本身不会产生新知识。

1.1.2 第二范式

理论科学阶段。从19世纪一直到20世纪中期，科学研究进入理论研究阶段，以演绎法为主。这一阶段，凭借科学家的智慧构建科学理论，并依据理论来解释自然世界。比如相对论、麦克斯韦方程组、量子理论、概率论等。与归纳法不同，演绎法除了用于解释已有事物之外，甚至可以创造新知识。

1.1.3 第三范式

计算科学阶段。自20世纪中期以来，由于客观事物的发展过于复杂，用归纳法和演绎法都难以满足科学研究的需要，人类开始借用计算机的高级运算能力来帮助进行科学计算。这个阶段，主要使用计算机来对复杂事物建模，将大量复杂的单个条件输入计算机，以模拟在多种因素的综合影响下，事物将会发生怎样的变化，比如模拟天气、地震、核试验等。

1.1.4 第四范式

大数据科学阶段。随着IT技术的兴起，人类收集到海量的数据，传统的计算科学已经越来越难以处理海量的数据。为了适应数据量的飞速膨胀，人类需要一种新的研究工具才能更有效地进行科学计算。因此，大数据作为处理海量数据为核心的“第四范式”，应运而生。大数据技术，包括海量数据获取技术、海量数据存储技术、海量数据的计算技术、海量数据的分析技术和数据可视化，成为当前第四范式的主要工具。

从宏观层面来看，大数据是一种思维和认知论的革命。大数据开启一次重大的时代转型。

传统的研究，主要依靠有限的调研，再加上经验，然后实现事物和业务决策的判断。然而，随着技术的快速革新，社会在快速发展，经常导致我们的经验往往跟不上事物的发展变化。

大数据，可以全方位地呈现事物的发展轨迹，并能实时动态地呈现事物的发展变化，甚至可以呈现事物各种因素之间的相关关系，找到影响事物的关键因素，进而控制事物的未来发展趋势，做出最正确的业务判断和业务决策。

在大数据时代，人类的思维方式必然会产生革命性的变革。

首先，表现在从追求因果到追求相关性。因果关系，一直是人类探索世界的一种思维方式。探究事物的根本原因，弄明白为什么会发生，这就是因果思维。因果思维看起来可以找到解决事物的根本办法，但却是极其复杂的一种方法。也许有的事物根本就没有因果关系，或者因果关系极其复杂，穷其一生也无法找到。大数据提供了另一种思维方法。与其去寻找为什么(因果关系)，还不如寻找是什么(相关关系)。同时，相关关系也可以作为因果关系的基础。存在因果关系的事物，一定会存在相关关系，通过找到相关的事物，专业人员可以在此基础上进一步去研究因果关系，这样可以缩小因果关系研究的范围，减少因果关系研究的验证成本，从而更快速地发现因果关系。

其次，从追求算法到追求数据。以往当研究的事物数量巨大时，由于计算的量过大，往往会采用随机抽样的方式进行研究，这在过去是切实可行的方法。对于抽样小数据，由于数据量小，为了解决抽样科学性和信息丢失问题，对数据分析算法的要求很高，算法的设计是关键因素，否则会影响最终分析的结果。在大数据时代，当把所有的数据作为分析对象时，相对来说，数据中的所有信息都可以得到。由于数据量大，算法的要求可以相应降低。《大数据时代》的作者断言，大数据的简单计算比小数据的复杂计算更有效。

大数据存在于任何行业和领域。这是因为大数据是对客观世界的量化和记录的结果，是客观事物的规律表现出来的现象，通过对大数据的挖掘，可以发现事物运行和发展的规律。

大数据挖掘特别适合窥探具有多维性和全面性的现实世界。它可以从很多看似支离破碎的信息中复原一个事物的全貌，进而能够预测或判断出尚未观察到的事物的现象。相关性思维作为大数据的核心思维之一，与基于普遍联系的哲学思维不谋而合。当利用大数据方法把影响事物的相关因素找出来，就能够透过事物的现象抓住事物的本质和规律，就能把握事物的发展和变化。

当前，大到国家，中到企业，小到个人，都掀起了一股认识大数据、理解大数据、利用大数据的热潮。国家，把大数据上升为国家战略，尝试用大数据来进行社会管理和经济治理；企业，也把大数据作为战略，尝试用大数据进行企业管理和商业模式的创新，甚至企业的升级和转型；大数据对于个人的生活方式也有重要的改变。

当然，大数据时代，掌握大数据挖掘技术变得尤为关键。此外，面临日益增长的品类繁多的大数据，智能化的处理变得比任何时候都更为迫切，效果也更为有效。

1.2 地球科学数据

在地球科学领域，广泛存在不同类型的数据。这些数据可以是结构化的，如地球化学分析和地球物理探查获得的数据。还有更多的非结构化、半结构化的数据，如古生物、矿物、岩石、矿床、岩心照片，海啸音频、地震视频，构造、遥感光谱图件，标本、野外记录、地质图表等。

还存在另一类数据，无限数据流。它们会随时间不断地有序产生，且产生速度快，数据规模大。如大气观测、地震监测、岩体稳定性监测、水文监测、地球化学监测过程中采集的数据。随着监测密度的提升和技术的进步，海量的监测传感器产生源源不断的时间序列数据，形成无限数据流。

客观存在的地质体与通过地质调查形成的地质文本，是同一个事物的两个方面。客观存在的地质体构成了地质领域一个个不同类型的地质实体，而地质文本是对一定区域范围内地质条件及地质事件的记录，其中，包含大量地质实体且实体类型多样。无论是对地质状况的描述、地质变化的说明还是地质灾害的统计，本质上都是对地质实体、相关附属信息及其之间关系的表达。伴随着传感器、测绘、定位等技术手段的不断发展，文本中对于地质实体的内容描述更加丰富、时空刻画更加精细、更新频率更加迅速。目前，世界各国地质资料馆、地质调查数据库和相关地学文献数据库，提供了多层次、全方位的地质资料信息数据库。大量的地质文本数据，尤其是海量的非结构化或半结构化的照片、视频和地质图，仍默默地“躺”在那里，等待地质文本数据的挖掘。

一般而言，地质科学大数据是一种时空大数据，主要产生于基础地质、矿产地质、水文地质、工程地质、环境地质、灾害地质的调查、勘查和相应的地质科学研究过程中，能源、矿产的开发利用和环境、地灾的监测、防治过程，以及各类天基、空基对地遥感观测活动。获得的途径包括地球物理、地球化学、钻探测井、遥感遥测、传感监测，还可以来自各种拓展应用，如图件编绘、分析计算、模拟仿真、预测评价、智能管控等，通常以文本、图表、声像、标本等多种数据形式存在。如图 1-1 所示。

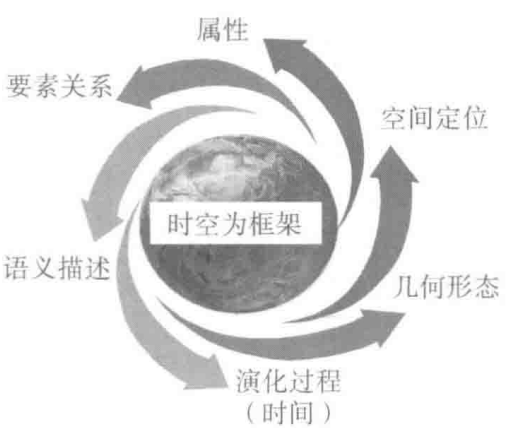

图 1-1 地球科学大数据要素

试读结束：需要全本请在线购买： www.ertongbook.com

地球科学大数据除拥有一般大数据的“4V”共性特征外，亦有自己显著的个性特点，突出体现在其专业背景特点(what、where、when、why、who、whom)。对地球科学领域的不同来源、不同获取方式、不同结构及不同格式的离散数据，开展结构化重建、关联分析、地学建模，加速地学知识的融汇，深化对地球系统的认识和理解，可望引发地球科学研究方式的变革。

在地球科学大数据模型构建中，数据融合是基础性的研究课题。它贯穿于研究对象认知模型、地质时空数据感知模型、地质时空数据分析模型、地质时空数据预测模型及地质时空数据决策模型的研究中。

地球科学大数据研究严重依赖于大数据平台的建设。各类专题的地质时空大数据链组织与实现，有赖于地质时空大数据平台的系统解决方案和整体架构，以及大数据安全存储、索引、调度机制和大数据引擎方法和技术，有赖于管理智能监测、预警与管控数据链的超级计算、云平台等。

地球科学大数据同其他行业和领域一样正在以指数形式增长。在这一背景下，地球科学大数据信息挖掘和人工智能技术，日益获得越来越多的地球科学家的重视(吴冲龙等，2016)。

世界各国实施的“玻璃地球”计划，广泛采取以三维区域地质填图为主导与深部探测计划相结合的方式，应用了大数据理念和处理技术(吴冲龙等，2016)。

2014 年在北京召开了以“中国‘玻璃地球’建设的核心技术及发展战略”为主题的香山科学会议第 491 次学术讨论会。“玻璃地球”旨在利用大数据、物联网、云计算等新一代信息技术，融合、集成和利用各类海量地质数据，构建地球系统和地质勘查系统，提高国家在资源、环境和减灾等领域面临的复杂问题的应对能力，特别是对水资源、环境和地灾的管控和安全保障能力，满足社会需求。

国家基金委与新疆维吾尔自治区联合基金“基于大数据的大型矿集区成矿预测”列入 2016 年指南。加拿大 Diagnos 公司在过去 10 年中为不同矿产勘查公司完成了数百个大数据分析、挖掘，进而圈定靶区的项目。这些项目位于加拿大魁北克、安大略、新不伦瑞克、纽芬兰、美国内华达州、多米尼加共和国、墨西哥、布基纳法索和坦桑尼亚共和国等地。2011 年，Diagnos 公司编制了加拿大魁北克西北地区金、铜、银、锌和镍的成矿远景图，覆盖面积 33.09 万平方千米。2012 年便取得了总计 5242 个矿权(占地 2335 平方千米)，覆盖了最有远景和未勘查的目标。已有的三维地质建模软件(如国外的 GOCAD、MVS、MicroStation、Surpac，国内的 QuantyView、GeoView、GeoMo3D、Titan3DM 等)正在得到进一步的优化和功能拓展。中国科学家研发的 3DMine 三维矿业软件通过国土资源部认证。它科学地组织各类矿山信息，将海量异质的矿山信息资源进行全面、高效和有序的管理和整合，运用数据库、三维模型、统计内插值和参数化概念，通过可视化技术、计算机技术和专业相结合，实现矿山重现，并可以快速计算，是自动成图和综合应用的技术平台。

深地资源与找矿靶区遴选已成为近年来矿床研究的重要热点，大数据分析成为其中不可或缺的技术。多元异质大数据集成以及不同学科、不同尺度的数据在三维空间的对比分析是其重要途径。澳大利亚开展了以找矿为目的开展的四维地质填图研究。荷兰建

立了全国1000米以上的3D地层框架模型。加拿大将三维地质填图用于盆地地下水调查。英国建立了全国4个尺度的三维地层框架模型。法国在地质调查等诸多领域开展三维地质建模。德国在北部多个盆地进行跨界三维地质建模。美国针对资源与环境评价开展三维地质框架研究等。

毋庸置疑，大数据研究仍存在一些需要克服的困难。

地球科学家需要探索并建立一个把人类活动与多科学领域无缝整合的模块式科学框架，便于把数据、科学、技术方法和模型组织到恰当的时空尺度中去，实现基于地学时空大数据的知识发现，深化对整个地球系统运转的理解，提升对地质过程的认知程度和对它们开发的决策能力(严光生等，2015)。

大数据处理要求将多源、异构、动态、海量的非(半)结构化数据快速有效地转化为能被分析决策利用的结构化信息(知识)。大数据处理经常面临四大问题：如何有序接纳多源异构、类型繁多的资料？如何高效组织规模海量、时空密集的数据？如何智能提纯结构清晰、关系明确的信息？如何快速驾驭在线实时、自适应强的计算？

大数据涉及数据量规模巨大，目前主流软件工具往往无法在合理的时间内对数据进行接入、管理、处理及挖掘。因此，需要发展新型处理模式，以从高速增长和多样化的海量大数据资源中挖掘优化的流程、智慧的知识和强力的决策。

有学者认为，地球科学大数据分析面临的主要问题有：

(1)如何建立一个多学科整合的模块式科学框架来组织数据、科学、技术和模型。

(2)如何融合监测的动态数据和勘查的静态数据，实现数据与模型的一体化管理。

(3)如何融合多源异质异构的结构化、半结构化和非结构化数据，进行数据挖掘。

(4)如何直接基于大数据进行挖掘、预测和预警，突破参数、模型、模式的限制。

目前，国内地球科学大数据研究与应用存在的主要困难有：数据来源有限(政府、机构公开数据不多)、数据类型混杂(结构化、非结构化，数字、视频、文本)、数据来源分散(部门分割，数据封锁)、数据质量存疑(存在数据篡改、造假等现象)、数据应用方法不清晰(难以清晰反映地质现状)、数据应用工具缺乏(大数据的应用模型复杂)、缺乏最终解决方案的指引(大数据最终产品匮乏)。

地质调查是地球科学研究获得数据的主要渠道之一。地质调查大数据分析，需要充分利用新一代信息技术，更新当前的大数据处理环境，着重进行大数据的智能分析与深度挖掘，由此建立大数据驱动的成矿远景图件。在大数据处理方法上，需要建立基于统一基础地理空间的多源数据集成与管理系统，将地质、构造、矿点、地球物理、地球化学、遥感钻孔等各类数据整合到统一的数据库中，利用云计算、大数据等方法，对多源综合数据进行集成、展示、分析和挖掘。

成矿与找矿模型是大数据理念和技术应用的重要领域。成矿与找矿研究将更充分地利用与“矿”有关的各种数据，包括在一定的地质历史时期或构造运动阶段，在一定的地质构造单元及构造部位，与一定的地质成矿作用有关的时间、空间、成因及矿床产状的数据，还包括庞大的矿床成因方面的数据信息，如成矿温度、成矿压力、流体包裹体、同位素、微量元素等矿床地球化学数据。

土壤环境地球化学污染监测、模拟、管控与预警也可以深度地应用大数据技术，如

图 1 - 2 所示。有研究项目建议，以高速发展的超大城市为对象，集成、融合行政区内物理空间的土壤污染调查与监测节点海量数据以及网络空间土壤污染相关大数据，建立城市土壤污染基础数据库，并通过数据依时间自动采集、更新和迭代，形成动态监测数据链。然后，开展城市土壤污染大数据分析，揭示主要污染物空间变异规律和土壤地球化学场特征，解析城市土壤污染源；开展基于土壤摄入率的人体健康风险评价和土壤安全等级分区的研究，以及数据链大数据驱动下的城市土壤污染预测预警研究，提供全景式时空透视和预警预测；建立可实际运行的城市土壤智能监测、模拟、管控、预警的技术体系和系列数据模型，以及决策支持系统软件原型。如图 1 - 2 所示。

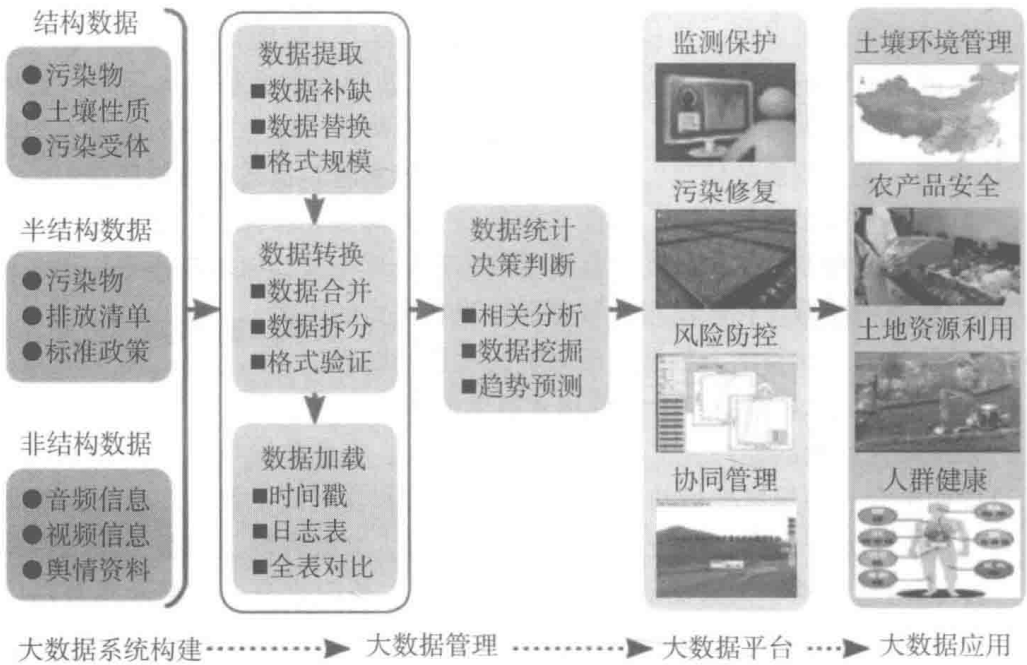

图 1 - 2　土壤环境大数据体系框架

超级计算硬件、软件的发展有力地拓展了大数据的研究空间，提升了大数据挖掘的水平。各国长期地质调查和探测取得的海量地质基础调查数据，将是超级计算机服务的重点对象之一。中山大学“天河二号”超级计算机采用了微异构计算阵列和新型并行编程模型及框架，集高性能计算、大数据分析和云计算于一体，能够支持大数据高吞吐量、高效处理等应用需求，能高效处理普通云计算不能处理的计算密集型问题，并能满足对复杂大数据开展精准、实时分析的需求。基于天河二号计算机的天文、地球科学与环境工程计算应用服务平台已成功落地，并组建有一支相应的超算技术和行业应用队伍。

至今，依托大数据的人工智能地质学还远不够成熟，所幸运的是有科学家在严肃、认真地探索。智能的矿床成因模型和找矿模型，有望成为人工智能研究的亮点。它可能会以地质 - 矿床大数据平台为依托，基于平台提供的大数据集与高性能计算能力，引入自然语言处理技术，让机器能够理解地质报告，加强机器学习、深度学习、可视分析的应用，进行知识提取和模式识别，特别是有别于显性知识信息预测的隐性知识信息发现。

1.3 大数据挖掘的基本任务

大数据挖掘又称数据库知识发现(knowledge-discovery in databases, KDD),是指从大量的数据中自动搜索隐藏于其中的有着特殊关系性的信息的过程,是从大量数据中寻找其规律的技术。数据挖掘技术还经常用来增强信息检索系统的能力。它由以下三个阶段组成:①数据准备;②数据挖掘;③结果表达和解释。

目前,大数据挖掘方法在地球科学领域中的应用尚处于起步阶段。近二十年来,随着研究和勘探投入的增加,以及技术手段的进步,各类专题数据库中海量异构地质数据爆发式增长积累,例如,油气数据库、矿产数据库、水文数据库、土壤数据库、遥感数据库、地球物理数据库、地球化学数据库、地震数据库、岩石数据库等。一方面,这些数据库还在不断地扩充完善;另一方面,传统的分析方法已无法充分发掘出隐含在各类数据中的深层次关联信息,致使花费昂贵代价获取的地质数据价值无法充分实现。数据挖掘技术的发展和引入,使得海量、异构、空间相关的地质数据的深层次分析成为可能。通过对地学数据进行基于空间位置挖掘,以及地学数据间的关联分析、聚类、分类、回归等,可以在此基础上进行异常识别、矿产预测、地质背景判别分析等,从而得到更加深入的地质认识,或者进行矿产勘查等生产实践活动。

一个典型的应用是,近些年随着 GeoRock 和 PetDB 两个国际共享的岩石地球化学数据库的建立和完善,很多研究者从中获取不同类型、不同构造背景的岩石,在"全体数据"的基础上,对学界应用多年的玄武岩构造环境判别系列图解进行重新审视。由于新的判别方法所采用的数据多来自最近二十年的分析结果,一方面,数据测试质量相比 20 世纪 80 年代以前明显提高;另一方面,元素测试更加齐全,且可利用的数据量和分布范围早已呈数量级增长。这已成为大数据挖掘思维和技术应用的典型例子。

本书将以玄武岩为主线贯穿全文,分别对玄武岩数据进行数据预处理、逻辑回归、聚类分析等。其代码在每章最后一节中展示。具体如图 1-3 所示。

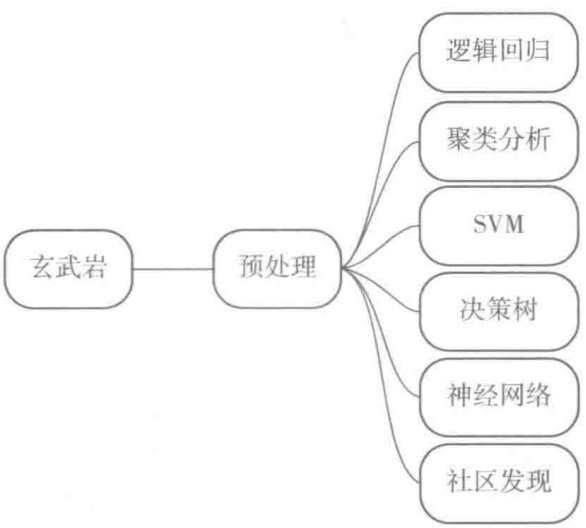

图 1-3 以 GeoRock 数据库中玄武岩为主线的大数据挖掘案例设计

在地学领域，有一类挖掘有特殊的意义：利用大数据技术对地质文本中地质实体的识别。地质文本是对一定区域范围内地质条件及地质事件的记录，其中，包括大量地质实体，且实体类型可以很多样。无论是对地质状况的描述、地质变化的说明，还是对地质灾害的统计，本质上都是对地质实体、相关附属信息及其之间关系的表达。地质实体是地质文本中的核心要素，其他属性和关系的描述都以地质实体为基础。伴随着传感器、测绘、定位等技术手段的不断发展，文本中对于地质实体的内容描述更加丰富、时空刻画更加精细、更新频率更加迅速。利用大数据技术对地质文本中的地质实体开展识别，就是地质文本的深度挖掘。地质实体识别能够有效辨别文本中的基本信息单位，帮助正确理解文本内容，同时基于提炼出的地质知识为广义文本数据挖掘中的信息抽取、信息检索、机器翻译、文摘生成等系列任务提供全面支持。

1.4 大数据挖掘建模过程

数据挖掘的步骤会随不同领域的应用而有所变化，每一种数据挖掘技术也会有各自的特性和使用步骤，针对不同问题和需求所制定的数据挖掘过程也会存在差异。此外，数据的完整程度、专业人员支持的程度等都会对建立数据挖掘过程有所影响。这些因素造成了数据挖掘在各不同领域中的运用、规划，以及流程的差异性。因此，对于数据挖掘过程的系统化、标准化就显得格外重要。

在进行数据挖掘技术的分析之前，还有许多准备工作要完成。数据挖掘完整的步骤如下：

(1)理解数据和数据的来源(understanding)。

(2)获取相关知识与技术(acquisition)。

(3)整合与检查数据(integration and checking)。

(4)去除错误或不一致的数据(data cleaning)。

(5)建立模型和假设(model and hypothesis development)。

(6)实际数据挖掘工作(data mining)。

(7)测试和验证挖掘结果(testing and verification)。

(8)解释和应用(interpretation and use)。

由上述步骤可看出，数据挖掘牵涉了大量的准备工作与规划工作，包括数据的净化、数据格式转换、变量整合，以及数据表的链接。

下面以 GeoRock 数据库中玄武岩构造环境判别分析为例，介绍大数据挖掘的建模过程。流程如图 1-4 所示。

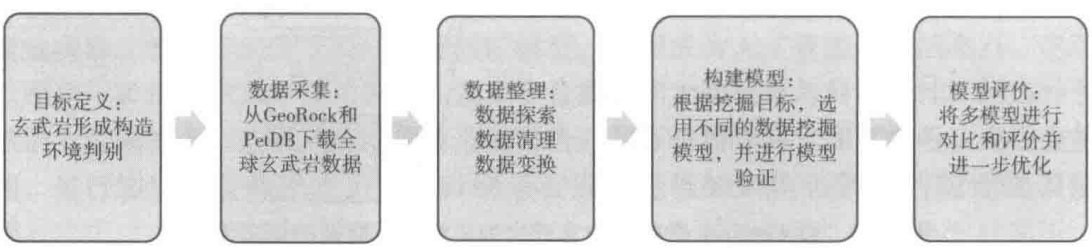

图 1-4 玄武岩挖掘建模过程

1.4.1 定义挖掘目标

要充分发挥数据挖掘的价值，必须先对目标有清晰的定义。针对玄武岩构造环境判别分析的数据挖掘，定义如下的挖掘目标：

(1)通过对全球玄武岩的岩石地球化学数据进行建模，实现利用全体元素数据对玄武岩进行聚类。

(2)以玄武岩形成的构造环境作为预测对象，利用玄武岩岩石地球化学数据实现对任一组玄武岩数据依其构造环境类型进行分类预测。

1.4.2 数据取样

在明确数据挖掘的目标后，需要从大数据系统中抽取出一个与挖掘目标相关的样本数据子集。抽取数据的标准是相关性、可靠性与有效性。通过数据样本的精选，不仅能减少数据的处理量，节省系统资源，而且能使要寻找的规律性被突显出来。

衡量取样数据质量的标准如下：

(1)资料完整无缺，各类指标项齐全。

(2)数据准确无误，反映的都是正常(无异常)状态下的水平。

对获取的数据，可再从中做抽样操作。常见的抽样方式如下：

(1)简单随机抽样。在采用简单随机抽样方式时，数据集中的每一组观测值都有相同的被抽样的概率。如按10%的比例对一个数据集进行随机抽样，则每一组观测值都有10%的机会被取到。

(2)等距抽样。如按5%的比例对一个有100组观测值的数据集进行等距抽样，则间距为20，等距抽样方式是取第20、40、60、80、100五组观测值。

(3)分层抽样。在这种抽样操作时，首先将样本总体分成若干层次(或者说分成若干个子集)。在每个层次中的观测值都具有相同的被选用的概率，但对不同的层次可设定不同的概率。这样的抽样结果通常具有更好的代表性，进而使模型具有更好的拟合精度。

(4)从起始顺序抽样。这种抽样方式是从输入数据集的起始处开始抽样。抽样的数量可以给定一个百分比，或者直接给定选取观测值的组数。

(5)分类抽样。在前述几种抽样方式中，并不考虑抽取样本的具体取值。分类抽样则依据某种属性的取值来选择数据子集。分类抽样的选取方式与上文所述的方式相同，只是抽样以类为单位。

1.4.3 数据探索与预处理

数据探索和预处理的目的是为了保证样本数据的质量，从而为保证模型质量打下基础。

数据探索就是针对如下问题的探索过程：①样本数据集是否达到原来设想的要求？②有没有什么明显的规律和趋势？③有没有出现从未设想过的数据状态？④属性之间有什么相关性？⑤它们可区分成怎样一些类别？数据探索主要包括：异常值分析、缺失值分析、相关分析、周期性分析等。对抽取的样本数据进行探索、审核和必要的加工处理，是保证最终挖掘模型的质量所必需的工作。可以说，挖掘模型的质量不会超过抽取

试读结束：需要全本请在线购买： www.ertongbook.com