HZ BOOKS
华章教育

计 算 机 科 学 丛 书

Pearson

原书第2版

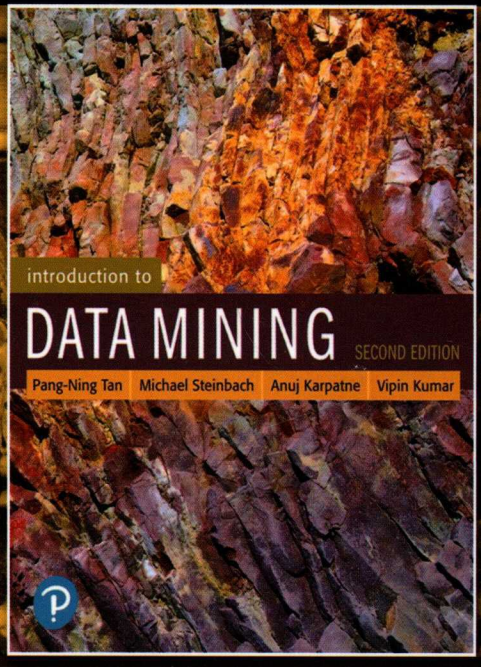
数据挖掘导论

[美] 陈封能（Pang-Ning Tan） 迈克尔·斯坦巴赫（Michael Steinbach） 阿努吉·卡帕坦（Anuj Karpatne） 维平·库玛尔（Vipin Kumar） 著

密歇根州立大学 明尼苏达大学

段磊 张天庆 等译

Introduction to Data Mining
Second Edition



从基础概念和算法的角度介绍数据挖掘所使用的主要原理与技术

机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

原书第2版

数据挖掘导论

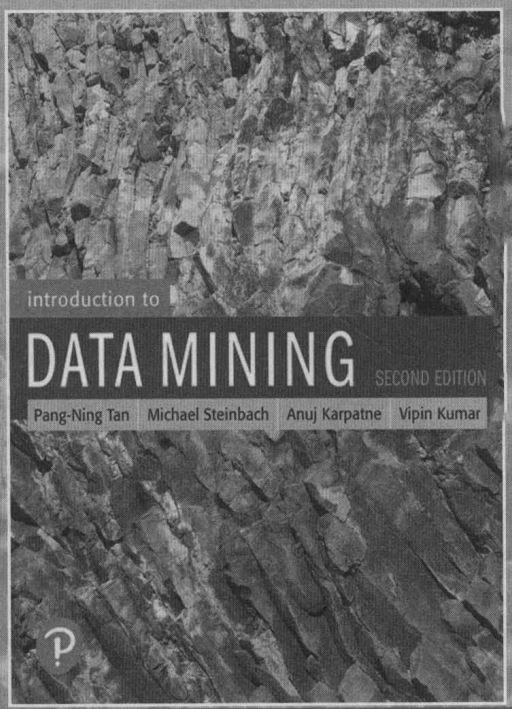
[美] 陈封能（Pang-Ning Tan） 迈克尔·斯坦巴赫（Michael Steinbach） 阿努吉·卡帕坦（Anuj Karpatne） 维平·库玛尔（Vipin Kumar） 著

密歇根州立大学 明尼苏达大学

段磊 张天庆 等译

Introduction to Data Mining

Second Edition



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘导论（原书第2版）/（美）陈封能（Pang-Ning Tan）等著；段磊等译．—北京：机械工业出版社，2019.7
（计算机科学丛书）
书名原文：Introduction to Data Mining, Second Edition

ISBN 978-7-111-63162-0

I. 数… II. ①陈… ②段… III. 数据采集－研究 IV. TP274

中国版本图书馆CIP数据核字（2019）第136744号

本书版权登记号：图字 01-2018-2522

本书所涵盖的主题包括：数据、分类、关联分析、聚类分析、异常检测和避免错误发现。通过介绍每个主题的基本概念和算法，为读者提供将数据挖掘应用于实际问题所需的必要背景。其中，分类、关联分析和聚类分析各自组织成两章的内容，一章讲述基本概念、代表性算法和评估技术，另一章深入讨论高级概念和算法。

本书适用于数据挖掘专业高年级本科生和研究生教学，也可供相关技术人员参考。

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：张梦玲　　责任校对：李秋荣
印　刷：北京文昌阁彩色印刷有限责任公司　　版　次：2019年8月第1版第1次印刷
开　本：185mm×260mm　1/16　　印　张：30.75
书　号：ISBN 978-7-111-63162-0　　定　价：139.00元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　读者信箱：hzjsj@hzbook.com

文艺复兴以来，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的优势，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自 1998 年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与 Pearson、McGraw-Hill、Elsevier、MIT、John Wiley & Sons、Cengage 等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出 Andrew S. Tanenbaum、Bjarne Stroustrup、Brian W. Kernighan、Dennis Ritchie、Jim Gray、Afred V. Aho、John E. Hopcroft、Jeffrey D. Ullman、Abraham Silberschatz、William Stallings、Donald E. Knuth、John L. Hennessy、Larry L. Peterson 等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力相助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专门为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近 500 个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www. hzbook. com
电子邮件：hzjsj@hzbook. com
联系电话：(010)88379604
联系地址：北京市西城区百万庄南街 1 号
邮政编码：100037

华章科技图书出版中心

译者序

Introduction to Data Mining, Second Edition

大数据时代的万物互联极大地丰富了数据采集手段。人们所面对的数据无论是类型还是规模都达到了空前的高度。与此同时，数据的价值受到了各行各业的广泛关注，以数据科学为核心的科学研究第四范式深入人心。面向海量、多源、异构、复杂的数据，建立恰当的模型并设计高效的算法来挖掘数据中蕴含的未知知识成为当前计算机应用研究的重要任务。

获取数据所蕴含价值的需求催生了数据挖掘。数据挖掘技术自诞生以来，一直蓬勃发展，如今已然成为各类大数据服务、新一代人工智能应用的基础。数据挖掘技术的发展体现了从数据管理到知识管理的时代发展。

这本由数据挖掘领域著名专家 P. Tan、M. Steinbach、A. Karpatne 和 V. Kumar 编撰的教程《Introduction to Data Mining》是一部优秀著作，对于从事数据挖掘研究和应用的专业人士，是实现自我提升的最适合的专著之一。本书不仅内容全面，涵盖了数据、分类、关联分析、聚类、异常检测、避免错误发现等数据挖掘的重要主题，而且内容编排独具特色。对于每一个重要的主题，都分为两章介绍，一章讲述基本概念、代表性算法和评估技术，另一章则讨论高级概念和算法。因此，本书不仅适合数据挖掘入门者学习，也适合数据挖掘研究进阶者参考。值得一提的是，本书文辞精妙、语言生动，作者以引导、举例为叙述手段，重点讲述了如何用数据挖掘知识解决各种实际问题，着力让读者在学习基本数据挖掘概念的同时掌握应用数据挖掘解决问题的技巧，彰显了作者在此领域的深厚研究造诣和娴熟的教学手法。此外，全书各章都设有习题，以加深读者对关键知识点的理解。

我们受机械工业出版社华章公司的委托翻译此书，首先向原著作者 P. Tan、M. Steinbach、A. Karpatne 和 V. Kumar 致敬。在翻译过程中，我们不可避免地受到了第 1 版译著用词准确、文笔流畅的影响，借此机会，向第 1 版译者范明教授、范宏建老师表示衷心的感谢。同时感谢机械工业出版社华章公司的信任，给予我们为数据挖掘研究推广尽绵薄之力的机会。

本书由段磊、张天庆主译。四川大学研究生秦蕊琦、王婷婷、宋楷文、张晓慧、刘杰、张译丹、王新澳、崔丁山等也付出了极大的努力，在此对他们表示感谢。

我们在翻译过程中力求忠于原著，新的专业术语尽量符合原著语义，但由于水平和时间有限，译文难免有错误和不妥之处，恳请读者批评指正。

段　磊

2019 年 5 月于四川大学

自 12 年前的第 1 版以来，数据分析领域发生了很大的变化。采集数据和用数据做决策的速率不断提高，采集到的数据数量和种类也在不断增加。事实上，“大数据”这个术语已被用于指代那些可获得的海量、多样的数据集。此外，“数据科学”这个术语也被用于描述一个新兴领域，其中，数据挖掘、机器学习、统计学等诸多领域的工具和技术，被用于从数据(通常是大数据)中提取出可实际应用的见解。

数据的增长为数据分析的各领域创造了大量的机会。其中，有着广泛应用的预测建模领域的发展最引人注目。例如，在神经网络(也称为深度学习)方面取得的最新进展，已经在许多具有挑战性的领域(如图像分类、语音识别以及文本分类和理解)表现出令人瞩目的成果。即使那些发展不是特别显著的领域(例如聚类、关联分析和异常检测等)也在不断前进。这个新版本就是对这些发展的响应。

概述 与第 1 版相同，本书第 2 版全面介绍了数据挖掘，方便学生、教师、研究人员和专业人士理解有关概念和技术。本书涵盖的主题包括：数据预处理、预测建模、关联分析、聚类分析、异常检测和避免错误发现。通过介绍每个主题的基本概念和算法，为读者提供将数据挖掘应用于实际问题所需的必要背景。与第 1 版一样，分类、关联分析和聚类分析都分两章讲述。前面一章(介绍章)讲述基本概念、代表性算法和评估技术，后面一章(高级章)深入讨论高级概念和算法。同第 1 版一样，这样做的目的是使读者透彻地理解数据挖掘的基础知识，同时论述更多重要的高级主题。由于这种安排，本书既可用作教材也可用作参考书。

为了帮助读者更好地理解书中讲述的概念，我们提供了大量的示例、图表和习题，并在网上公开了原有习题的答案。除了第 10 章的新习题，其余习题与第 1 版的基本一致。教师可以通过网络获取各章的新习题及其答案。对更高级的主题、重要的历史文献和当前趋势感兴趣的读者，可以在每一章结尾找到文献注释，本版对这部分内容做了较大的更新。此外，还提供了一个覆盖本书所有主题的索引。

第 2 版的新内容 内容上主要的更新是与分类相关的两章内容(第 3 章和第 4 章)。第 3 章仍使用决策树分类器进行讲解，但对适用于各种分类方法的主题讨论进行了大量的扩充，这些主题包括：过拟合、欠拟合、训练规模的影响、模型复杂度、模型选择以及模型评估中常见的缺陷等。第 4 章的每一节几乎都进行了重大更新，着重扩展了贝叶斯网络、支持向量机和人工神经网络的内容。对深度网络，我们单独增加了一节来介绍该领域当前的发展。我们还更新了 4.11 节“类不平衡问题”中有关评估方法的讨论。

关联分析内容的改进则更具体。我们对关联模式评估部分(第 5 章)以及序列和图形挖掘部分(第 6 章)进行了全面修订。对聚类分析的修订也很具体。在聚类分析的介绍章(第 7 章)增添了 K 均值初始化技术并更新了簇评估的讨论。聚类分析的高级章(第 8 章)新添了关于谱图聚类的内容。对异常检测部分也进行了大量的修订和扩展。我们保留并更新了现有方法，如统计学、基于最近邻/密度方法和基于聚类方法，同时介绍了基于重构的方法、单类分类和信息论方法。基于重构的方法通过深度学习范畴中的自编码网络进行阐述。关于数据的第 2 章也进行了更新，更新内容包括对互信息的讨论和基于核技术的讨论。

第10章讨论了如何避免错误发现并产生正确的结果，这一章的内容是全新的并且在当前关于数据挖掘的教科书中也是新颖的。该章讨论了关于避免虚假结果的统计概念(统计显著性、p值、错误发现率、置换检验等)，这些是对其他章中相关内容的补充，然后在介绍数据挖掘技术的内容中对这些概念进行了阐述。这一章还强调了对数据分析结果的有效性和可重复性的关注。新增的最后一章，是认识到这个主题的重要性后的产物，同时也是对“在分析数据时需要对相关领域有更深入的理解”这一观点的认可。

本版纸书删除了数据探索章节以及附录，但仍将其保留在网上。本版附录对大数据环境下的可伸缩性进行了简要讨论。

致教师 作为一本教材，本书广泛适用于高年级本科生和研究生教学。由于学习这门课程的学生背景不同，他们可能不具备广博的统计学和数据库知识，因此本书只要求最低限度的预备知识。数据库知识不是必需的，但我们假定读者有一定的统计学或数学背景，这些背景会让他们更容易学习某些内容。与以前一样，本书或者更确切地说是讨论主要数据挖掘主题的各章，都尽可能自成一体。因此，这些主题的讲授次序相当灵活。其中第2章、第3章、第5章、第7章和第9章是核心内容。对于第10章，建议至少给出粗略的介绍，以在学生解释他们的数据分析结果时引起一些注意。尽管应先介绍数据(第2章)，但可以按任意顺序来讲授基本分类(第3章)、关联分析(第5章)和聚类分析(第7章)。由于异常检测(第9章)与分类(第3章)和聚类分析(第7章)具备先后关系，所以后两章应先于第9章进行讲解。同时，可以根据时间安排和兴趣，从高级分类、关联分析和聚类分析章节(第4章、第6章、第8章)中选择多种主题进行讲解。我们还建议通过数据挖掘中的项目或实践练习来强化听课效果，虽然它们要花费一些时间，但这种实践作业可以大大提高课程的价值。

支持材料 本书的读者可以在 http://www-users.cs.umn.edu/~kumar/dmbook/上获取相关材料：

- 课程幻灯片。
- 学生项目建议。
- 数据挖掘资源，如数据挖掘算法和数据集。
- 联机指南，使用实际的数据集和数据分析软件，为本书介绍的部分数据挖掘技术提供例子讲解。

其他支持材料(包括习题答案)只向采纳本书做教材的教师提供[⊖]。读者可通过邮箱 dmbook@cs.umn.edu 将意见和建议以及勘误发给作者。

致谢 许多人都为本书的出版做出了贡献。首先向家人表示感谢，这本书是献给他们的。正是有他们的耐心和支持，本书才能顺利完成。

感谢明尼苏达大学和密歇根州立大学数据挖掘小组的学生所做的贡献。Eui-Hong(Sam) Han 和 Mahesh Joshi 帮助我们准备了最初的数据挖掘课程。他们编制的某些习题和演示幻灯片已经收录在本书及教辅幻灯片中。小组中的其他学生也为本书的初稿提出建议或以各种方式做出贡献，他们是：Shyam Boriah、Haibin Cheng、Varun Chandola、Eric Eilertson、Levent Ertöz、Jing Gao、Rohit Gupta、Sridhar Iyer、Jung-Eun Lee、Benjamin Mayer、Aysel Ozgur、Uygar Oztekin、Gaurav Pandey、Kashif Riaz、Jerry Scripps、Gyorgy Simon、Hui Xiong、

⊖ 关于本书教辅资源，只有使用本书作为教材的教师才可以申请，需要的教师请联系机械工业出版社华章公司，电话 010-88378991，邮箱 wangguang@hzbook.com。——编辑注

Jieping Ye 和 Pusheng Zhang。还要感谢明尼苏达大学和密歇根州立大学选修数据挖掘课程的学生，他们使用了本书的初稿，并提供了极富价值的反馈。特别感谢 Bernardo Craemer、Arifin Ruslim、Jamshid Vayghan 和 Yu Wei 的有益建议。

Joydeep Ghosh(得克萨斯大学)和 Sanjay Ranka(佛罗里达大学)试用了本书的初稿。我们也直接从得克萨斯大学下列学生那里获得了许多有用的建议：Pankaj Adhikari、Rajiv Bhatia、Frederic Bosche、Arindam Chakraborty、Meghana Deodhar、Chris Everson、David Gardner、Saad Godil、Todd Hay、Clint Jones、Ajay Joshi、Joonsoo Lee、Yue Luo、Anuj Nanavati、Tyler Olsen、Sunyoung Park、Aashish Phansalkar、Geoff Prewett、Michael Ryoo、Daryl Shannon 和 Mei Yang。

Ronald Kostoff(ONR)阅读了聚类部分的初稿，并提出了许多建议。George Karypis 对创建索引提供了宝贵的帮助。Irene Moulitsas 提供了 LaTeX 支持，并审阅了一些附录。Musetta Steinbach 发现了图中的一些错误。

感谢明尼苏达大学和密歇根州立大学的同事，他们帮助创建了良好的数据挖掘研究环境。他们是：Arindam Banerjee、Dan Boley、Joyce Chai、Anil Jain、Ravi Janardan、Rong Jin、George Karypis、Claudia Neuhauser、Haesun Park、William F. Punch、György Simon、Shashi Shekhar 和 Jaideep Srivastava。还要向我们的数据挖掘项目的合作者表示谢意，他们是：Ramesh Agrawal、Maneesh Bhargava、Steve Cannon、Alok Choudhary、Imme Ebert-Uphoff、Auroop Ganguly、Piet C. de Groen、Fran Hill、Yongdae Kim、Steve Klooster、Kerry Long、Nihar Mahapatra、Rama Nemani、Nikunj Oza、Chris Potter、Lisiane Pruinelli、Nagiza Samatova、Jonathan Shapiro、Kevin Silverstein、Brian Van Ness、Bonnie Westra、Nevin Young 和 Zhi-Li Zhang。

明尼苏达大学和密歇根州立大学的计算机科学与工程系为本书写作及研究提供了计算资源和支持环境。ARDA、ARL、ARO、DOE、NASA 和 NSF 等机构为本书作者提供了研究资助。特别是 Kamal Abdali、Mitra Basu、Dick Brackney、Jagdish Chandra、Joe Coughlan、Michael Coyle、Stephen Davis、Frederica Darema、Richard Hirsch、Chandrika Kamath、Tsengdar Lee、Raju Namburu、N. Radhakrishnan、James Sidoran、Sylvia Spengler、Bhavani Thuraisingham、Walt Tiernin、Maria Zemankova、Aidong Zhang 和 Xiaodong Zhang，他们有力地支持了我们的数据挖掘和高性能计算研究。

与培生出版集团的工作人员的合作令人愉快。具体来说，我们要感谢 Matt Goldstein、Kathy Smith、Carole Snyder 和 Joyce Wells。还要感谢 George Nichols 帮助绘图，Paul Anagnostopoulos 提供 LaTeX 支持。

感谢培生邀请的审稿人：Leman Akoglu(卡内基梅隆大学)、Chien-Chung Chan(阿克伦大学)、Zhengxin Chen(内布拉斯加大学奥马哈分校)、Chris Clifton(普度大学)、Joydeep Ghosh(得克萨斯大学奥斯汀分校)、Nazli Goharian(伊利诺伊理工学院)、J. Michael Hardin(阿拉巴马大学)、Jingrui He(亚利桑那州立大学)、James Hearne(西华盛顿大学)、Hillol Kargupta(马里兰大学巴尔的摩县分校和 Agnik 公司)、Eamonn Keogh(加利福尼亚大学河滨分校)、Bing Liu(伊利诺伊大学芝加哥分校)、Mariofanna Milanova(阿肯色大学小石城分校)、Srinivasan Parthasarathy(俄亥俄州立大学)、Zbigniew W. Ras(北卡罗来纳大学夏洛特分校)、Xintao Wu(北卡罗来纳大学夏洛特分校)和 Mohammed J. Zaki(伦斯勒理工学院)。

自本书第 1 版出版以来，我们收到了许多指出错别字和其他各种问题的读者和学生的意见。在此无法列举所有人的名字，但非常感谢他们的意见，相关问题已在第 2 版中予以修正。

第1章

Introduction to Data Mining, Second Edition

绪　论

数据采集和存储技术的迅速发展，加之数据生成与传播的便捷性，致使数据爆炸性增长，最终形成了当前的**大数据**时代。围绕这些数据集进行可行的深入分析，对几乎所有社会领域的决策都变得越来越重要：商业和工业、科学和工程、医药和生物技术以及政府和个人。然而，数据的数量(体积)、复杂性(多样性)以及收集和处理的速率(速度)对于人类来说都太大了，无法进行独立分析。因此，尽管大数据的规模性和多样性给数据分析带来了挑战，但仍然需要自动化工具从大数据中提取有用的信息。

数据挖掘将传统的数据分析方法与用于处理大量数据的复杂算法相结合。在本章中，我们将介绍数据挖掘的概况，并概述本书所涵盖的关键主题。首先介绍一些需要高级数据分析技术的应用。

商业和工业　借助POS(销售点)数据收集技术(条码扫描器、射频识别(RFID)和智能卡技术)，零售商可以在商店的收银台收集顾客购物的最新数据。零售商可以利用这些信息，加上电子商务网站的日志、客服中心的顾客服务记录等其他的重要商务数据，能够更好地理解顾客的需求，做出更明智的商业决策。

数据挖掘技术可以用来支持广泛的商务智能应用，如顾客分析、定向营销、工作流管理、商店分布、欺诈检测以及自动化购买和销售。最近一个应用是快速股票交易，在这个交易中，需要使用相关的金融交易数据在不到一秒的时间内做出买卖决定。数据挖掘还能帮助零售商回答一些重要的商业问题，如："谁是最有价值的顾客?""什么产品可以交叉销售或提升销售?""公司明年的营收前景如何?"这些问题促使着数据挖掘技术的发展，比如关联分析(见第5章和第6章)。

随着互联网不断改变我们日常生活中互动和做决定的方式，能够生成大量的在线体验数据，例如网页浏览、信息传递，以及在社交网站上发布信息，这为使用Web数据的商务应用提供了机会。例如，在电子商务领域，用户的在线浏览或购物偏好数据可以用来推荐个性化的产品。数据挖掘技术也在支持其他基于互联网的服务方面扮演着重要的角色，如过滤垃圾信息、回答搜索查询，以及建议社交圈的更新和联系。互联网上大量的文本、图像和视频使得数据挖掘方法有了许多进展，如深度学习(这将在第4章进行讨论)。这些进展推动了诸多应用领域的进步，如目标识别、自然语言翻译与自动驾驶。

另一个经历大数据快速转型的应用领域是移动传感器和移动设备的使用，如智能手机和可穿戴计算设备。借助更好的传感器技术，可以利用嵌入在相互连接的日常设备上的低成本传感器(称为物联网(IOT))来收集物理世界的各种信息。在数字系统中，物理传感器的深度集成正开始产生大量与环境相关的多样化和分布式的数据，可用于设计方便、安全、节能的家庭系统，以及规划智能城市。

医学、科学与工程　医学、科学与工程界的研究者正在快速收集大量数据，这些数据对获得有价值的新发现至关重要。例如，为了更深入地理解地球的气候系统，NASA已经部署了一系列的地球轨道卫星，不停地收集地表、海洋和大气的全球观测数据。然而，由于这些数据的规模和时空特性，传统的方法常常不适合分析这些数据集。数据挖掘所开发

的技术可以帮助地球科学家回答如下问题："干旱和飓风等生态系统扰动的频度和强度与全球变暖之间有何联系?""海洋表面温度对地表降水量和温度有何影响?""如何准确地预测一个地区的生长季节的开始和结束?"

再举一个例子，分子生物学研究者希望利用当前收集的大量基因组数据，更好地理解基因的结构和功能。过去，传统方法只允许科学家在一个实验中每次研究少量基因，微阵列技术的最新突破已经能让科学家在多种情况下比较数以千计的基因特性。这种比较有助于确定每个基因的作用，或许可以查出导致特定疾病的基因。然而，由于数据的噪声和高维性，需要新的数据分析方法。除了分析基因序列数据外，数据挖掘还能用来处理生物学的其他难题，如蛋白质结构预测、多序列校准、生物化学路径建模和系统发育学。

另一个例子是利用数据挖掘技术来分析越来越多的电子健康记录(EHR)数据。不久之前，对患者的研究需要手动检查每一个患者的身体记录，并提取与所研究的特定问题相关的、具体的信息。EHR 允许更快和更广泛地探索这些数据。然而，只有患者在看医生或住院期间才能对他们进行观察，并且在任何特定访问期间只能测量关于患者健康的少量细节，因此存在重大挑战。

目前，EHR 分析侧重于简单类型的数据，如患者的血压或某项疾病的诊断代码。然而，很多类型更复杂的医学数据也被收集起来，例如心电图(ECG)和磁共振成像(MRI)或功能性磁共振成像(fMRI)的神经元图像。尽管分析这些数据十分具有挑战性，但其中包含了患者的重要信息。将这些数据与传统的 EHR 和基因组数据集成分析是实现精准医学所需的功能之一，旨在提供更加个性化的患者护理。

3

1.1　什么是数据挖掘

数据挖掘是在大型数据库中自动地发现有用信息的过程。数据挖掘技术用来探查大型数据库，发现先前未知的有用模式。数据挖掘还可以预测未来的观测结果，比如顾客在网上或实体店的消费金额。

并非所有的信息发现任务都被视为数据挖掘。例如查询任务：在数据库中查找个别记录，或查找含特定关键字的网页。这是因为这些任务可以通过与数据库管理系统或信息检索系统的简单交互来完成。而这些系统主要依赖传统的计算机科学技术，包括先进高效的索引结构和查询处理算法，有效地组织和检索大型数据存储库的信息。尽管如此，数据挖掘技术可以基于搜索结果与输入查询的相关性来提高搜索结果的质量，因此被用于提高这些系统的性能。

数据库中的数据挖掘与知识发现

数据挖掘是**数据库中知识发现**(Knowledge Discovery in Database，KDD)不可缺少的一部分，而 KDD 是将未加工的数据转换为有用信息的整个过程，如图 1.1 所示。该过程包括一系列转换步骤，从数据预处理到数据挖掘结果的后处理。

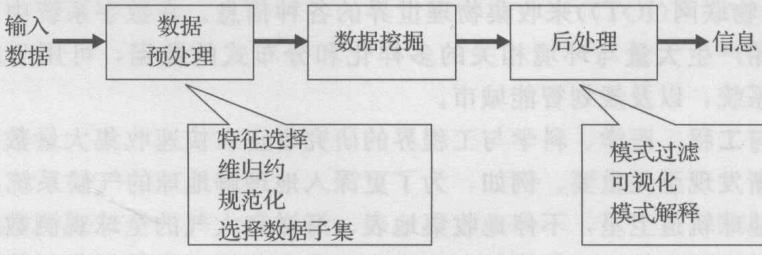


图 1.1　数据库中知识发现(KDD)过程

4

输入数据可以以各种形式存储(平面文件、电子表格或关系表),并且可以存储在集中式数据库中,或分布在多个数据站点上。**预处理**(preprocessing)的目的是将原始输入数据转换为适当的格式,以便进行后续分析。数据预处理涉及的步骤包括融合来自多个数据源的数据,清洗数据以消除噪声和重复的观测值,选择与当前数据挖掘任务相关的记录和特征。由于收集和存储数据的方式多种多样,数据预处理可能是整个知识发现过程中最费力、最耗时的步骤。

"结束循环"(closing the loop)通常指将数据挖掘结果集成到决策支持系统的过程。例如,在商业应用中,数据挖掘的结果所揭示的规律可以与商业活动管理工具结合,从而开展或测试有效的商品促销活动。这样的结合需要**后处理**(postprocessing)步骤,确保只将那些有效的和有用的结果集成到决策支持系统中。后处理的一个例子是可视化,它使得数据分析者可以从各种不同的视角探查数据和数据挖掘结果。在后处理阶段,还能使用统计度量或假设检验,删除虚假的数据挖掘结果(见第 10 章)。

1.2 数据挖掘要解决的问题

前面提到,面临大数据应用带来的挑战时,传统的数据分析技术经常遇到实际困难。下面是一些具体的问题,它们引发了人们对数据挖掘的研究。

可伸缩 由于数据产生和采集技术的进步,数太字节(TB)、数拍字节(PB)甚至数艾字节(EB)的数据集越来越普遍。如果数据挖掘算法要处理这些海量数据集,则算法必须是可伸缩的。许多数据挖掘算法采用特殊的搜索策略来处理指数级的搜索问题。为实现可伸缩可能还需要实现新的数据结构,才能以有效的方式访问每个记录。例如,当要处理的数据不能放进内存时,可能需要核外算法。使用抽样技术或开发并行和分布式算法也可以提高可伸缩程度。附录 F 给出了伸缩数据挖掘算法的技术总体概述。

高维性 现在,常常会遇到具有成百上千属性的数据集,而不是几十年前常见的只具有少量属性的数据集。在生物信息学领域,微阵列技术的进步已经产生了涉及数千特征的基因表达数据。具有时间分量或空间分量的数据集也通常具有很高的维度。例如,考虑包含不同地区的温度测量结果的数据集,如果在一个相当长的时间周期内反复地测量,则维数(特征数)的增长正比于测量的次数。为低维数据开发的传统数据分析技术通常不能很好地处理这类高维数据,如维灾难问题(见第 2 章)。此外,对于某些数据分析算法,随着维数(特征数)的增加,计算复杂度会迅速增加。

异构数据和复杂数据 通常,传统的数据分析方法只处理包含相同类型属性的数据集,或者是连续的,或者是分类的。随着数据挖掘在商务、科学、医学和其他领域的作用越来越大,越来越需要能够处理异构属性的技术。近年来,出现了更复杂的数据对象。这种非传统类型的数据如:含有文本、超链接、图像、音频和视频的 Web 和社交媒体数据,具有序列和三维结构的 DNA 数据,由地球表面不同位置、不同时间的测量值(温度、压力等)构成的气候数据。为挖掘这种复杂对象而开发的技术应当考虑数据中的联系,如时间和空间的自相关性、图的连通性、半结构化文本和 XML 文档中元素之间的父子关系。

数据的所有权与分布 有时,需要分析的数据不会只存储在一个站点,或归属于一个机构,而是地理上分布在属于多个机构的数据源中。这就需要开发分布式数据挖掘技术。分布式数据挖掘算法面临的主要挑战包括:(1)如何降低执行分布式计算所需的通信量?(2)如何有效地统一从多个数据源获得的数据挖掘结果?(3)如何解决数据安全和隐私问题?

非传统分析　传统的统计方法基于一种假设-检验模式，即提出一种假设，设计实验来收集数据，然后针对假设分析数据。但是，这一过程劳力费神。当前的数据分析任务常常需要产生和评估数千种假设，因此需要自动地产生和评估假设，这促使人们开发了一些数据挖掘技术。此外，数据挖掘所分析的数据集通常不是精心设计的实验的结果，并且它们通常代表数据的时机性样本(opportunistic sample)，而不是随机样本(random sample)。

1.3　数据挖掘的起源

如图 1.1 所示，虽然数据挖掘最开始被认为是 KDD 框架中的一个中间过程，但是多年来它作为计算机科学的一个学术领域，关注着 KDD 的所有方面，包括数据预处理、数据挖掘和后处理。它的起源可以追溯到 20 世纪 80 年代末，当时组织了一系列围绕数据库中知识发现的主题研讨会，汇集了来自不同学科的研究人员，讨论关于应用计算技术从大型数据库中提取可利用的知识的挑战和机遇。这些由来自学术界和工业界的研究人员和实践者参加的研讨会很快成为非常受欢迎的会议。会议的成功举办，以及企业和行业在招聘具有数据挖掘背景的新员工时所表现出的兴趣，推动了这一领域的巨大发展。

该领域最初建立在研究人员早先使用的方法和算法之上。特别是，数据挖掘研究人员借鉴了如下领域的思想方法：(1)来自统计学的抽样、估计和假设检验；(2)来自人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地采纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理以及可视化和信息检索，并将其延伸至解决大数据挖掘的挑战。

一些其他领域也起到重要的支撑作用。特别是，需要数据库系统提供高效的存储、索引和查询处理。源于高性能(并行)计算的技术在处理海量数据集方面常常是非常重要的。分布式技术还可以帮助处理海量数据，并且当数据不能集中到一起处理时显得尤为重要。图 1.2 显示了数据挖掘与其他领域之间的联系。

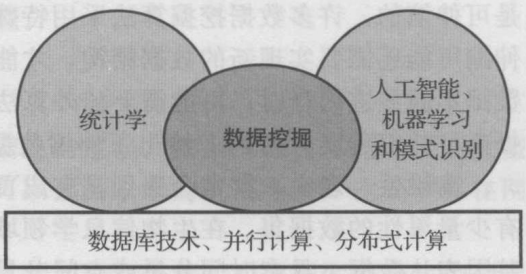


图 1.2　数据挖掘汇集了许多学科的知识

数据科学和数据驱动发现

数据科学(data science)是一个研究及应用工具和技术从数据中获取有用的见解的跨学科领域。虽然它被认为是一个具有独特身份的新兴领域，但其中的工具和技术通常来自数据分析的许多不同领域，如数据挖掘、统计学、人工智能、机器学习、模式识别、数据库技术以及分布式和并行计算(见图 1.2)。

数据科学作为一个新兴领域出现是一种共识。现有的数据分析领域，通常没有为新兴应用中出现的数据分析任务提供一整套分析工具。相反，处理这些任务通常需要广泛的计算、数学和统计能力。为了说明分析此类数据面临的挑战，请设想以下示例。社交媒体和网络为社会科学家提供了大量观察和定量测量人类行为的新机会。为了进行这样的研究，社会科学家会与具备网页挖掘、自然语言处理(NLP)、网络分析、数据挖掘和统计等技能的分析师合作。与传统的基于调查的社会科学研究相比，这种分析需要更为广泛的技术和工具，并且涉及的数据量更大。因此，数据科学必然是一个建立在许多领域持续合作基础上的高度跨学科的领域。

数据科学的数据驱动方法强调从数据中直接发现模式和关系，特别是在大量数据中，通常不需要广泛的领域知识。这种方法中一个值得注意的成功例子是神经网络的进步，即