邵峰晶　于忠清　王金龙　孙仁诚　编著

数据挖掘原理与算法

（第二版）

科学出版社
www.sciencep.com

数据挖掘原理与算法
(第二版)

邵峰晶　于忠清　王金龙　孙仁诚　编著

科学出版社
北京

内 容 简 介

本书第一版是国内第一本对数据挖掘技术基础算法进行详细描述的实用性教材。第二版在第一版基础上进行了较多的修订和补充。在系统阐述数据挖掘与知识发现技术的产生、发展，以及相关概念、原理、基本方法的基础上，从实用的角度出发，对数据挖掘中的关联、分类、聚类、序列等算法和技术进行了剖析，对每种技术均提供了代表性算法。同时，结合作者近年来所做的研究，对数据挖掘的应用问题进行了分类论述。最后，对目前数据挖掘的最新进展、应用趋势等进行了总结。

本书可作为计算机、管理等专业高年级本科生与研究生课程的教材，也可作为数据挖掘领域的高级软件开发人员的参考书。

图书在版编目(CIP)数据

数据挖掘原理与算法/邵峰晶等编著. —2 版. —北京：科学出版社，2009

ISBN 978-7-03-025440-5

Ⅰ. 数… Ⅱ. 邵… Ⅲ. 数据采集 Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2009)第 153886 号

责任编辑：王志欣 任 静 / 责任校对：陈玉凤

责任印制：赵 博 / 封面设计：耕者设计工作室

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2003 年 8 月中国水利水电出版社第一版

2009 年 8 月第 二 版 开本：B5(720×1000)

2009 年 8 月第一次印刷 印张：26 3/4

印数：1—4 000 字数：522 000

定价：48.00 元

(如有印装质量问题，我社负责调换)

第二版前言

数据挖掘经过十几年的蓬勃发展，产生了丰硕的理论和应用成果。作为一门应用性较强的学科，数据挖掘技术已经渗透到国民经济的各个领域，引起学术界和产业界的极大关注，取得了广泛的应用，为各行各业的管理者提供了有价值的决策依据。这些都使我们迫切感觉到要对本书第一版进行大的修订，补充最新的理论和应用成果，以适应当前学科发展的需要。

本书自第一版出版以来，我们把其作为高年级学生和研究生学习数据挖掘课程的教材，取得了不少的教学经验，也发现了原书中的个别错误以及叙述不清楚的地方。第二版在原书的基础上，对原稿进行了改正并做了较大的更新，对内容进行了重新组织和整理，对数据挖掘中新出现的关键技术进行了介绍，详细描述了部分典型新算法；并根据作者近年来的研究成果增添了数据挖掘的应用章节，对数据挖掘的应用成果进行了论述；同时，对数据挖掘的最新进展进行了介绍和概括总结。这些有助于读者系统学习数据挖掘理论、技术和方法，通过应用实例的介绍能够给读者更加深刻的认识。

我们希望本书第二版的出版，不仅给学习数据挖掘课程的高年级学生和研究生提供一本内容比较全面的教材，而且也为开发数据挖掘相关系统的高级软件开发人员和从事该项技术的各个领域的科技工作者提供一本可读性较好的参考书，有助于进一步推动我国的数据挖掘研究与应用的深入开展。

本书的编写得到了中国工程院李德毅院士的关注和指导，作者在此表示衷心的感谢。青岛大学的隋毅、庞传军、纪俊、柯爽、杨坤等研究生也为本书的完成做了大量的工作，在此，一并表示衷心的感谢。

在数据挖掘蓬勃发展的今天，该项技术涉及了很多学科领域，由于我们的理论水平和实践经验都具有局限性，本书还存在不少不足之处，敬请读者在阅读本书时能够给我们提出宝贵建议，并对相关内容进行批评指正。

作　者

2009 年 5 月于青岛

第一版前言

数据挖掘技术是近几年国内外迅速发展起来的一门交叉学科，涉及数据库、统计学、人工智能与机器学习等多个领域。计算机的应用普及产生了大量的数据，数据挖掘就是利用上述学科的技术进行大数据量的处理。数据挖掘的应用领域非常宽广，从农业生产的预测到基因分类，从化学分子结构的识别到 NBA 教练临场更换队员，从信用卡欺诈到税务稽查，数据挖掘技术对未来社会的各个领域将起到越来越重要的作用。

我国的数据挖掘技术一方面是科研机构停留在学术研究上，另一方面是利用国外公司的软件产品解决具体问题。为了提高学术水平，科研人员只得进行高水平但很难实用的算法研究；为了提高经济效益，销售国外软件公司的产品最稳健。但是，数据挖掘技术在解决实际问题的过程中需要的是成熟技术加针对具体问题的修正，因此，国内迫切需要对国外十余年的数据挖掘具体技术进行剖析，在掌握核心技术的前提下才能真正赶超。本书的背景是在我们三年前开始开发数据仓库产品及对数据挖掘技术进行了将近两年的跟踪的基础上，根据大量参考文献及内部技术报告，结合研究生的教学工作完成的。目前，我们已完成了开放式的数据挖掘平台及部分算法的实现。

本书的使用对象是在校高年级的本科生、研究生及各个领域的高级软件开发人员，书中介绍了大量的数据挖掘算法，各个算法具有很强的实用性。本书是国内第一本对数据挖掘技术基础算法进行详细描述的实用性书籍。

本书共分 9 章。第 1 章对数据挖掘从各个角度进行了剖析，从社会需求开始对数据挖掘的概念、数据挖掘的数据来源、数据挖掘的分类、体系结构、运行过程、数据挖掘与其他领域之间的关系、评价标准及未来的发展方向进行了全面的介绍。

第 2 章对数据挖掘的孪生兄弟——数据仓库技术进行了简单的介绍，由于数据挖掘技术的一个重要发展方向就是嵌入到数据仓库中，即数据挖掘所使用的大数据集直接来自于数据仓库。在简单地回顾了数据仓库技术之后，给出了一种多维数据的模型，这是实施联机分析处理(OLAP)的一种关键技术，同时简单介绍了我们自行开发的 OLAP 展示工具的体系结构，并介绍了数据仓库在银行的应用案例。

第 3 章讲述的是数据挖掘的数据预处理所涉及的概念及算法。干净而合乎要求的数据是数据挖掘成功应用的基础，对数据进行整理是一项枯燥而艰苦的工作。本章在介绍了数据挖掘的数据准备工作之后，给出了一种常用的数字属性的离散化及属性选择算法。数据挖掘虽然可以解决大数据集的问题，但在分布完全相同的前提下，算法处理十万条记录与百万条记录的时间代价是完全不同的。数

据采样技术同样有多种方法，每种方法适合解决的问题是不同的。本章最后一部分介绍了数据抽象问题，即如何将大量的数据进行概念提升。

第 4 章对关联分析给出了详细的算法。无论是在国内还是在国外，关联分析是数据挖掘发展的先行者，并且几乎与其他学科没有交叉。Apriori 算法是关联分析的基础，多值属性的关联分析所关心的问题是如何将连续数值的关联分析转化为布尔值，多层关联分析与约束性的关联分析都是解决实用问题的算法，本章最后给出了增量的关联分析解决算法。

第 5 章讲述了数据分类，给出了分类的各种基本算法，包括国外数据挖掘最早的 ID3 算法及 C4.5 算法。对来自统计学的 CART 算法给出了详细的描述，同时对如何解决大数据集问题的 SLIQ 算法及并行问题的 SPRINT 分类器也给出了详细的说明。

第 6 章讲述了多维访问与数据可视化。它虽然不是数据挖掘的直接内容，但聚类的多种算法都用到了多维数据访问的技术。而空间数据挖掘的基础则是多维访问。数据可视化技术中对数据的观察进行了阐述。

第 7 章给出了聚类的多种实用算法及基础算法。聚类算法采用了多种技术，用途非常广泛，本章给出了大量的详细的算法。分层的聚类来自于统计学，虽然不能解决大数据量问题，但作为基础还是进行了详细的说明。分区算法介绍了 PAM、CLARA 及 CLARANS 算法，其中对 CLARANS 算法进行扩充，可以用于空间数据挖掘。*k*-means 算法是最常见也是最实用的算法，特别介绍了处理离散数据的聚类算法 *k*-modes。OPTICS 是一种复杂的算法，用途也最广泛。BIRCH 的特色是只需访问一次数据库，对该算法给出了详细的描述。最后，对用途广泛的孤立点问题给出了最先进的算法。

第 8 章介绍了序列模式及时间序列。序列模式给出了最早也是最实用的算法，时间序列只是介绍了概貌，没有给出具体的算法，因为时间序列本身就是一门交叉学科。

第 9 章介绍了我们开发的开放式的数据挖掘平台，限于篇幅只是给出了体系结构，对数据挖掘平台中所用的 OLE DB For Data Mining 及可预测模型描述语言 PMML 也进行了简单的介绍。

书中的第 1 章、第 2 章和第 7 章由邵峰晶教授编写，其余章节由于忠清研究员编写。在本书的编写过程中得到了南京大学徐洁磐教授、北京大学的邵维忠教授及青岛市副市长马论业教授的多次指导，在此表示感谢。青岛海尔青大海威软件公司的刘志强、林永及贾胜中三位工程师在海威数据仓库与数据挖掘软件及资料方面给予了大力支持，李洁小姐在文字及图形的整理方面做了大量的工作，在此一并表示谢意。

由于时间仓促，书中的错误与不足之处在所难免，敬请读者批评指正。

作　者

2003 年 6 月

目　录

第1章 导 论

1.1 数据挖掘的社会需求

一切新事物的产生都是由需求驱动。让计算机能够自动、智能地分析数据库中的大量数据以获取信息是推动挖掘型工具产生并发展的强大动力。从生产成本的角度看，公司的人工费用在不断提升，产品与服务的价格持续下降，激烈的市场竞争迫使决策者想办法降低成本及扩大产品与服务的销售量来提高公司的竞争力。从计算机应用角度看，无论硬件与网络性能的提高，还是软件技术与功能的提高，都要求软件从单纯的管理功能向综合的分析功能转变。从数据管理角度看，历史数据是一笔宝贵的财富，而且这些数据正以几何级数或指数方式增长。从软件技术发展方向看，海量数据的智能分析对原来各个领域的技术都带来了极大的挑战，需要采用综合性的技术来迎接这些挑战。

随着数据库技术的飞速发展以及人们获取数据手段的多样化，人类所拥有的数据急剧增加，随着大容量、高速度、低价格的存储设备相继问世，当今数据库的容量越来越大，已经达到 TB[①]，甚至 PB[②] 的水平，但能够对这些数据进行有效分析处理的工具却很少。数据库系统往往只对已有数据进行存取和简单操作，人们很难通过这些操作获取数据隐含的深层语义，而这些描述数据整体特征和发展趋势的信息在决策制定过程中具有更加重要的价值和意义，它们可以指导政府、企业决策以获取更大的社会效益和经济效益。

为了能够得到这些隐藏在海量数据背后的具有决策价值的知识，数据挖掘应运而生，该技术就是要从“数据矿山”中挖掘蕴藏的“知识金块”，为人们提供有价值的、隐藏的商业和科学“情报”，从而应用于商务管理、市场分析、生产控制和科学研究与探索等诸多方面。例如，股票经纪人需要从日积月累的大量股票行情信息的历史数据中发现市场规律，以预测未来趋势；大型超市管理员希望从过去几年的销售记录中获取顾客的消费行为习惯，以便及时变换营销策略；地质学家想通过对地球资源卫星发回的大量数据和照片的分析来发现有开采价值的矿物资源，等等。著名的“啤酒和尿布”的例子能够说明数据挖掘的作用：在世界杯期间，

① Terabyte，1TB=1000GB

② Petabyte，1PB=1000TB

美国加州某个超级市场通过数据挖掘，从顾客每天购买记录数据库中发现，下班后前来购买婴儿尿布的顾客多数是男性，而且他们往往也同时购买啤酒。于是该超市的经理重新布置了货架，把啤酒类商品布置在婴儿尿布货架附近，并在二者之间放上土豆之类的佐酒小食品，同时把男士们需要的日常生活用品也就近布置。这有效刺激了顾客的随机购买欲望，促进了上述几种商品的销量大增。

通过上面的例子可以看出，数据挖掘能够为决策者提供重要的、极有价值的信息或知识，从而产生不可估量的效益。经过十几年的技术发展，国外在数据挖掘技术方面积累了丰富的经验，取得了大量成果，而且越来越多的大中型企业开始利用数据挖掘技术分析公司数据，以辅助决策。

在国内，数据挖掘已从单纯的研究走向产品开发及技术应用，随着国内市场经济的不断完善，国内对数据挖掘的市场需求正在高速增长，众多软件开发商涉足该领域，随着市场的成熟与用户应用水平的提高，将会出现大量国产软件产品。

1.2 什么是数据挖掘

数据挖掘(data mining，DM)，也称数据库中知识发现(knowledge discovery in database，KDD)，就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的非平凡过程(Piatetsky-Shapiro 1996)。当前，还存在很多和数据挖掘相近的术语，如数据分析、知识抽取、模式分析、数据考古(data archeology)、数据采捞、商业智能以及决策支持等。

数据挖掘起源于多种学科，包括数据库、人工智能、数理统计、机器学习、可视化、并行计算等，其中最重要的是数据库、机器学习和统计学三个领域，这些不同的历史影响使得不同领域的学者对数据挖掘功能持有不同观点(Chen 1996，Ramakrishnan 1999，Zhou 2003)。

数据挖掘的诞生可追溯到20世纪80年代，1989年8月在美国底特律召开的第11届国际人工智能联合会议(International Joint Conference on Artificial Intelligence，JICAI)，举行了数据库中知识发现的专题讨论(KDD Workshop)。接着，美国人工智能学会在1981年、1983年和1994年相继举行了KDD Workshop。在这些讨论会的基础上，美国计算机学会成立了知识发现与数据挖掘专业委员会SIGKDD，并于1995年在加拿大蒙特利尔召开了第一届知识发现与数据挖掘国际学术会议(The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)。迄今为止，该会议已经召开了14届。第15届SIGKDD国际会议将第一次离开北美，于2009年在法国巴黎召开。作为数据库研究、开发和应用最活跃的分支之一，数据挖掘兴起至今，已成为许多国际学术会议关注的焦点，如SIGKDD、

SIGMOD、VLDB、ICDE、ICDM、SDM、PKDD、PAKDD等。

数据挖掘技术主要包括关联规则(association rule)发现、分类(classification)、聚类(clustering)分析、泛化(generalization)和预测(prediction)等。迄今为止,人们已发现了很多数据挖掘方法,如频繁项集和关联规则、决策树(decision tree)方法、贝叶斯(Bayesian)方法、人工神经网络(artificial neural network,ANN),等等。各式各样的数据挖掘理论被提出与采用,如模糊集合(fuzzy sets)、粗糙集(rough sets)理论、数理统计、机器学习(machine learning)、人工神经网络、决策树、模式识别(pattern recognition)、高性能计算等。数据挖掘发展至今,已经开发了很多数据挖掘平台,如美国SAS公司的SAS Enterprise Miner、IBM公司的Intelligent Miner、IBM公司Almaden研究中心开发的QUEST系统、加拿大Simon Fraser大学的DB Miner、新西兰Waikato大学开发的Weka、中国科学院计算技术研究所的MS Miner等。近几年,数据挖掘研究的重点已从算法研究向具体应用过渡,从实验室原型走向商品化阶段。

数据挖掘可以处理多种多样的数据,如关系数据库中的结构化数据,文本、图形、图像等非结构化数据,甚至是Web信息、生物信息等非结构化异构型数据。通过数据挖掘可以帮助决策者寻找数据间潜在的关联,发现被忽略的要素,所发现的知识可被用于信息管理、查询优化、决策支持、过程控制等,从而对趋势预测和决策行为提供帮助。

数据挖掘技术从一开始就是面向应用的,作为商业智能(business intelligence,BI)实现的最深层次,在其体系中占据重要位置。数据挖掘不仅面向特定数据库的简单检索查询调用,而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理,以指导实际问题求解,通过发现事件间相互关联关系来预测未来活动,从而帮助企业减少不必要的投资,同时提高资金回报。将人们对数据的应用,从低层次的末端查询操作,提高到为各级经营决策者提供决策支持。从数据挖掘定义上看,它与传统数据分析最主要的区别在于是否存在明确的前提下挖掘信息并发现知识。数据挖掘所得到的信息应具有先前未知、有效和易于使用三个特征,其与传统数据分析技术的对比如表1.1所示。

表1.1 传统数据分析技术与数据挖掘技术的对比

	传统的数据分析技术	数据挖掘技术
工具特点	回顾型、验证型	预测型、发现型
分析重点	已经发生了什么	预测未来知识、解释发生原因
分析目的	从最近销售单中发现客户	锁定未来可能客户,以减少销售成本
数据大小	数据维、维中属性数、维中数据少	数据维、维中属性数、维中数据庞大
技术现状	成熟	统计分析工具成熟,其他正在发展