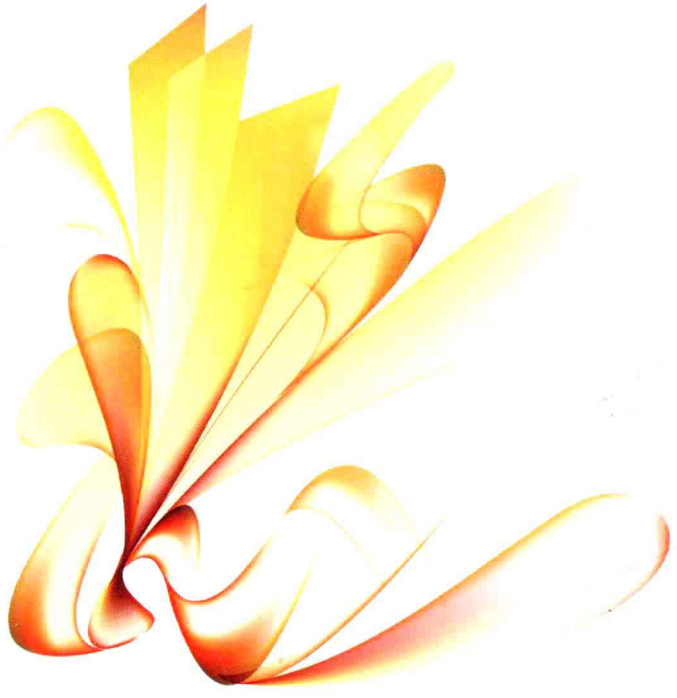
HZ BOOKS
华章科技

资深数据挖掘专家10余个行业、上百家大型企业、近10年数据挖掘应用与咨询经验结晶

详细讲解数据挖掘的核心技术、应用分类、建模方法、建模工具、二次开发技术、基于MATLAB和Hadoop的数据挖掘算法，深度与广度兼顾、实践与理论并举

涵盖金融、电信、电力、互联网、生产制造和公共服务等行业近20个完整案例，数据详实，实践性极强

大数据技术丛书



Data Mining: Practical Case Analysis

数据挖掘

实用案例分析

张良均 陈俊德 刘名军 陈荣◎著

CD-ROM

机械工业出版社
China Machine Press

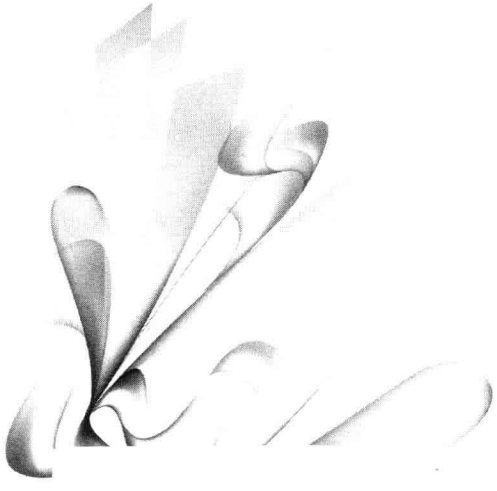
技术丛书

Data Mining: Practical Case Analysis

数据挖掘

实用案例分析

张良均 陈俊德 刘名军 陈荣◎著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据挖掘：实用案例分析/张良均等著. —北京：机械工业出版社，2013.6
（大数据技术丛书）

ISBN 978-7-111-42591-5

Ⅰ. 数… Ⅱ. 张… Ⅲ. 数据采集 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2013）第 109165 号

本书是数据挖掘实战领域颇具特色的一部作品，作者曾为 10 余个行业上百家大型企业提供数据挖掘服务，本书是其在数据挖掘领域探索近 10 年的经验总结之作。全书以实践和实用为宗旨，深度与广度兼顾，实践与理论并举。

本书共 12 章，分三个部分。第一部分是基础篇（第 1～4 章），主要对数据挖掘的基本概念、应用分类、建模方法及常用的建模工具进行了介绍，并对本书所用到的数据挖掘建模平台 TipDM 进行了说明。第二部分是实战篇（第 5～10 章），以案例的形式对数据挖掘技术在金融、电信、电力、互联网、生产制造以及公共服务等行业的应用场景进行了讨论；首先介绍案例背景，然后阐述分析方法与过程，最后完成模型构建；在介绍建模过程的同时穿插操作训练，把相关的知识点嵌入相应的操作过程中；此外，第 10 章精心设计了 6 个实验项目，读者可以通过本章介绍的方法动手实践，以巩固数据挖掘知识，在分析建模过程的同时，进一步增强动手能力。第三部分是高级篇（第 11～12 章），主要介绍基于第三方接口的数据挖掘二次开发技术，重点对常用的 WEKA 和 MATLAB 数据挖掘算法接口进行了探讨；最后对基于 Hadoop 框架的海量数据挖掘进行了说明，以满足读者更高层次的需求。

随书光盘中提供了本书的相关资料和案例资源，以及 6 个动手实验所使用的完整数据，方便读者动手实践书中所讲解的案例。

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：白 宇
北京市荣盛彩色印刷有限公司印刷
2013 年 7 月第 1 版第 1 次印刷
186mm×240mm · 26.25 印张
标准书号：ISBN 978-7-111-42591-5
ISBN 978-7-89433-973-7（光盘）
定 价：79.00 元（附光盘）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzjsj@hzbook.com

前　言

为什么要写这本书

现在，什么程序员最稀缺？什么技术最火？回答：数据挖掘！

数据挖掘是从大量数据（包括文本）中挖掘出隐含的、先前未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。数据挖掘有助于企业发现业务的趋势，揭示已知的事实，预测未知的结果，因此“数据挖掘”已成为企业保持竞争力的必要方法。

近年来企业所处理的数据每五年就会呈现倍数增长。大部分的企业并没有数据不足的问题，过度的数据重复与不一致才是大问题，这使得企业在使用、有效管理以及将这些数据用于决策过程方面都遭遇到了问题。因此未来几年，随着大数据迅速发展，数据挖掘将是极为重要的成长领域，其应用会越来越广泛，可以说，只要企业掌握有分析价值的数据源，皆可进行高价值的数据挖掘分析。目前数据挖掘主要应用在电信、零售、农业、互联网、金融、电力、生物、化工和医疗等行业。典型的应用如：客户细分、流失预警、价值评估、交叉销售、欺诈发现、精准营销、故障诊断等。

总的来说，跟国外相比，我国由于信息化程度不太高、企业内部信息不完整，零

售业、银行、保险、证券等对数据挖掘的应用并不太理想。但随着市场竞争的加剧，各行业应用数据挖掘技术的意愿越来越强烈，可以预计，未来几年各行业的数据分析应用一定会从传统的统计分析发展到大规模数据挖掘应用。

本书作者从实践出发，结合大量数据挖掘工程案例，总结出数据挖掘建模过程应完成的任务主要包括：数据探索、数据预处理、分类与回归、聚类分析、时序预测、关联规则挖掘、偏差检测等。因此，本书的编排以解决某个应用的挖掘目标为前提，先介绍案例背景，再阐述分析方法与过程，最后完成模型构建，在介绍建模过程的同时穿插操作训练，把相关的知识点嵌入相应的操作过程中。

本书光盘中附数据挖掘建模案例数据文件、数据挖掘算法工具包源程序及使用说明。

读者对象

❑ 需求分析及系统设计人员。

这类人员可以在理解数据挖掘原理及建模过程的基础上，结合数据挖掘案例完成精确营销、客户分群、交叉销售、客户流失分析、客户信用记分、欺诈发现等数据挖掘应用的需求分析和设计。

❑ 数据挖掘开发人员。

这类人员可以在理解数据挖掘应用需求和设计方案的基础上，结合本书提供的基于第三方接口快速完成数据挖掘应用的编程实现。

❑ 开设有数据挖掘课程的高校教师和学生。

目前国内不少高校将数据挖掘引入本科教学中，在数学、自动化、电子信息、金融等专业开设了数据挖掘技术相关的课程。目前这一课程的教学仍主要限于理论介绍，因为过于抽象，学生理解起来往往比较困难，教学效果不甚理想。本书提供的基于实战案例和建模实践的教学，能够使师生充分发挥互动性和创造性，理论联系实际，从而获得最佳的教学效果。

❑ 进行数据挖掘应用研究的科研人员。

许多科研院所为了更好地对科研工作进行管理，纷纷开发了适应自身特点的科研业务管理系统，并在使用过程中积累了大量的科研信息数据。但是，这些科研业务管理系统一般没有对这些数据进行深入分析的功能，对数据所隐藏的价值并没有充分挖掘利用。科研人员需要数据挖掘建模工具及有关方法论来深挖科研信息的价值，从而提高科研水平。

❑ 关注高级数据分析的人员。

业务报告和商业智能解决方案对于了解过去和现在的状况是非常有用的。但是，数据挖掘的预测分析解决方案还能使这类人员预见未来的发展状况，让他们的机构能够先发制人，而不是处于被动。因为数据挖掘的预测分析解决方案将复杂的统计方法和机器学习技术应用到数据之中，通过使用预测分析技术来揭示隐藏在交易系统或企业资源计划（ERP）、结构数据库和普通文件中的模式和趋势，从而为这类人员的决策提供科学依据。

如何阅读本书

本书共 12 章，分三个部分，通过对一个个真实案例深入浅出的剖析，使读者在不知不觉中能快速领悟看似深不可测的数据挖掘理论。读者在阅读过程中，应充分利用随书配套的案例建模数据，借助相关的数据挖掘建模工具，通过动手实践，帮助快速理解相关知识和理论。

第一部分是基础篇（第 1～4 章），主要对数据挖掘的基本概念、应用分类、建模方法及常用的建模工具进行了介绍；第 4 章对本书所用到的数据挖掘建模平台 TipDM 进行了说明。

第二部分是实战篇（第 5～10 章），其中第 5～9 章为案例部分，重点对数据挖掘技术在金融、电信、电力、互联网、生产制造以及公共服务等行业的应用场景进行了讨论。在过程组织上，按照先介绍案例背景、挖掘目标，再阐述分析方法与过程，最后完成模型构建的顺序进行，在介绍建模过程的同时穿插操作训练，把相关的知识点嵌入相应的操作过程中；第 10 章为实验部分，读者可以通过本章介绍的方法，动手实践以巩固数据挖掘知识及建模过程。

第三部分是高级篇（第 11～12 章），其中第 11 章对基于第三方接口的数据挖掘二次开发技术进行了说明，通过示例，介绍了如何基于 WEKA 和 MATLAB 等工具实现数据挖掘算法接口编程；第 12 章介绍了基于 Hadoop 框架开发的并行数据挖掘算法工具箱——TipCDM，并通过一个实际案例，介绍了基于云计算的海量数据挖掘的具体应用及实现过程。

勘误和支持

除封面署名外，参加本书编写工作的还有：张益铭、周积荣、曹晶、蒋世忠、张秋妮、曹焱峰、余春迪、阮鹏、余燕团、王军晓等。由于作者的水平有限，加之编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。为此，读者可通过作者微博（http://t.qq.com/waveletz）或 TipDM 官网（www.tipdm.com）反

馈有关问题。也可通过热线电话（40068-40020）或企业 QQ（40068-40020）进行在线咨询。

读者可以将书中的错误及遇到的任何问题反馈给我们，我们将尽量在线上为读者提供最满意的解答。随书光盘中提供了全部建模数据文件及源程序，也可以从智能中国网站（www.5iai.com）下载，我们会将相应的功能更新及时更正。如果您有更多的宝贵意见，也欢迎发送邮件至邮箱 5iai2008@gmail.com，期待能够得到你们的真挚反馈。

致谢

本书的案例主要来自作者承担的国家及省部级项目和与合作单位的研究应用实践，如独立承担的科技部中小企业创新基金项目——基于云计算和 SOA 架构的海量数据挖掘平台；与广东省电科院合作的智能用电海量数据挖掘项目；与广州翰思软件有限公司合作的基于数据挖掘和 GIS 技术的房地产自动评估系统；与广州因孚网络科技有限公司合作的基于云计算的海量数据挖掘平台的研发及应用示范；与西南交通大学合作的数据挖掘技术在混合厌氧消化系统优势营养互补机制研究；与南京中医药大学合作的数据挖掘技术在乳腺癌证素变化规律及截断疗法研究；与华南师范大学合作的企业信息预测开发平台；与广东工业大学合作的应用统计实践教学基地建设项目；与广东石油化工学院合作的云计算环境下 Web 结构挖掘研究及应用等。

本书编写过程中，得到了广大企事业单位科研人员的大力支持！在此谨向广东电力科学研究院、广西电力科学研究院、广东电信规划设计院、珠江/黄海水产研究所、华南师范大学、广东工业大学、西南交通大学、南京中医药大学、华南理工大学、湖南师范大学、广州中医药大学、武汉理工大学、广东石油化工学院、中山大学、浙江大学、广州大学、河南理工大学、甘肃中医学院、番禺职业技术学院、大连海事大学、广州从兴电子开发有限公司、广州泰迪智能科技有限公司、广州太普软件科技有限公司、中科普开（北京）科技有限公司、EasyHadoop 社区等单位给予支持的专家及师生致以深深的谢意。

在本书的出版过程中，得到了参与中国数据挖掘建模竞赛（http://c.5iai.com）的众多师生及机械工业出版社华章公司杨福川老师、白宇编辑等无私的帮助与支持，在此一并表示感谢。

张良均

2013 年 4 月于广州

目　录

第二部分 实 战 篇

第三部分 高 级 篇

第一部分

基 础 篇

本部分内容

- ❑ 初识数据挖掘
- ❑ 数据挖掘的应用分类
- ❑ 数据挖掘建模
- ❑ 顶尖数据挖掘平台 TipDM

第1章　初识数据挖掘

随着计算机技术、网络技术、通信技术和 Internet 技术的发展，以及各行各业业务操作流程的自动化，企业内积累了大量业务数据，这些数据动辄以 TB 计算。这些数据和由此产生的信息是企业的财富，它如实地记录着企业运作的状况。面对大量的数据，迫使人们不断寻找新的工具，来对企业的运营规律进行探索，为商业决策提供有价值的信息，使企业获得利润。能满足企业这一迫切需求的有力工具就是**数据挖掘**。对于企业而言，数据挖掘有助于发现业务的趋势，揭示已知的事实，预测未知的结果。从这个意义上讲，知识是力量，数据挖掘是财富。

1.1　什么是数据挖掘

数据挖掘（Data Mining，DM）：就是从大量数据（包括文本）中挖掘出隐含的、未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程；是利用各种分析工具在海量数据中发现模型和数据之间关系的过程。这些模型和关系可以被企业用来分析风险、进行预测。

数据挖掘的目的就是从数据中“淘金”，就是从数据中获取智能的过程。

Gartner Group 提出：“数据挖掘是通过仔细分析大量数据来揭示有意义的新的关系、模式和趋势的过程。它使用模式认知技术、统计技术和数学技术。”

The META Group 的 Aaron Zornes 表示：“数据挖掘是一个从大型数据库中提取以前不知道的可操作性信息的知识挖掘过程。”

总之，由于企业内产生了大量的业务数据，这些数据和由此产生的信息是企业的财富，它如实记录了企业运作的状况。通过数据挖掘分析，能帮助企业发现业务的趋势，揭示已知的事实，预测未知的结果。数据挖掘已成为企业保持竞争力的必要方法。

1.2　数据挖掘在企业商务智能应用中的定位

报告和商业智能解决方案对于了解过去和现在的状况是非常有用的。但是，预测分析解决方案还能使用户预见未来的发展状况，使其能够先发制人，而不是处于被动。数据分析和数据挖掘系统的目的是带给我们更多的决策支持信息，并不是取