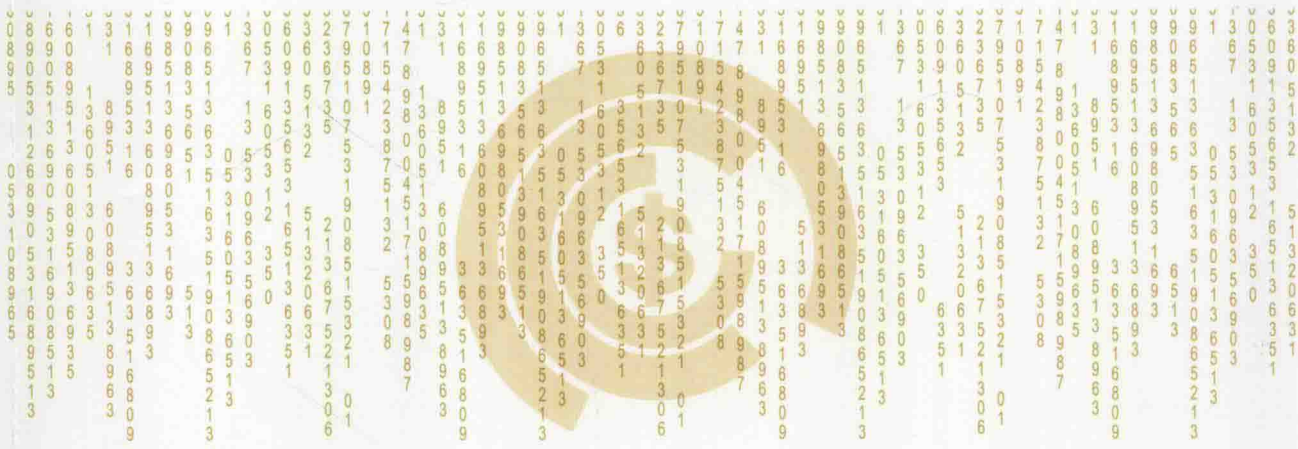
银行业信息科技风险管理高层指导委员会
银行业信息化丛书

金融数据挖掘与分析

郑志明　缪绍日　荆丽丽　等编著



Financial Data Mining and Analysis

机械工业出版社
CHINA MACHINE PRESS

银行业信息科技风险管理高层指导委员会
银行业信息化丛书

金融数据挖掘与分析

郑志明　缪绍日　荆丽丽　等编著

Financial Data Mining
and Analysis

机械工业出版社
CHINA MACHINE PRESS

本书针对金融行业数据量大、更新快的特点，着重介绍了数据挖掘与分析技术在金融行业尤其是银行业中的应用。本书的主要内容包括：数据挖掘概述、金融数据挖掘概述、基于大数据的金融数据挖掘概述、数据仓库技术、数据挖掘与分析技术、大数据挖掘与分析技术、数据挖掘技术在零售银行信用风险管理中的应用、数据挖掘技术在巴塞尔资本协议下的银行风险计量中的应用、数据挖掘技术在客户关系管理中的应用、数据挖掘技术在金融市场分析与预测中的应用、数据挖掘技术在互联网金融中的应用、基于大数据的金融科技战略与实施、数据安全与隐私保护，并针对当前的大数据浪潮，给出了金融数据挖掘与分析领域的应对策略。

本书主要供银行信息科技人员阅读，也可供从事数据挖掘与分析技术应用研究的科研人员和金融数据分析人员参考，还可作为信息管理与金融类专业教学参考书。

图书在版编目（CIP）数据

金融数据挖掘与分析/郑志明等编著. —北京：机械工业出版社，2015.10

（银行业信息化丛书）

ISBN 978-7-111-51805-1

Ⅰ.①金… Ⅱ.①郑… Ⅲ.①金融-数据处理-研究 Ⅳ.①F830.41

中国版本图书馆 CIP 数据核字（2015）第 243094 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

总 策 划：张敬柱 黄养成

策划编辑：王华庆 责任编辑：王华庆 版式设计：霍永明

责任校对：张玉琴 封面设计：徐 超 责任印制：李 洋

保定市中画美凯印刷有限公司印刷

2016 年 1 月第 1 版第 1 次印刷

184mm×260mm · 17.75 印张 · 437 千字

0001—5500 册

标准书号：ISBN 978-7-111-51805-1

定价：69.80 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务

服务咨询热线：010-88361066

读者购书热线：010-68326294

010-88379203

网络服务

机 工 官 网：www.cmpbook.com

机 工 官 博：weibo.com/cmp1952

金 书 网：www.golden-book.com

教育服务网：www.cmpedu.com

“银行业信息化丛书”编委会

总序

信息化是推动经济社会变革的重要力量。坚持走中国特色的新型工业化、信息化、城镇化、农业现代化道路，是党中央立足全局、放眼未来、与时俱进的战略决策。2014年2月27日，中央网络安全和信息化领导小组的成立，更加体现了中央保障网络安全、推动信息化发展、维护国家利益的决心。银行业作为国家经济体系的重要行业之一，是信息化的重要推动主体、参与主体和受益主体。银行业持之以恒地贯彻落实国家信息化战略，不仅是推动加快我国信息化进程的必然要求，也是银行业改革发展、转型升级和更好地服务实体经济的内在需求。

近年来，我国银行业审时度势、积极作为，坚持基础建设与科技创新并重、提升服务与保障安全并举的科学发展导向，以推进信息化为契机，调整经营理念、优化经营机制、完善服务模式，在服务手段信息化、管理模式信息化、信息安全保障等方面取得积极进展，推动了银行业的核心竞争力、市场适应力和贴身服务能力的进一步提升。一是服务手段信息化发展迅速。电子银行、自助银行、智能支付终端等信息化服务渠道日渐普及，使得金融服务覆盖面更加广泛、服务方式更加便捷、服务产品更加丰富。二是管理模式信息化迈出实质性步伐。注重依托核心数据库、运用先进数据挖掘分析工具，推进银行经营决策逐步智能化，风险管理日趋精细化，产品创新逐渐体现个性化，银行业经营管理信息化水平不断提升。三是信息安全保障取得积极进展。银行业信息安全越来越受重视，相关科技基础设施建设步伐加快，多层次、立体化、全方位的信息安全保障体系正在逐步形成。

当然，我们也应该清醒地认识到，银行业信息化面临着复杂的内外部环境，核心技术受限、网络安全威胁、隐私保护和信息保密等挑战将长期存在，银行业自身认识不尽到位、技术储备不够充分、资源投入相对不足、过度依赖外包等问题仍较为突出，针对银行业特殊需求的信息化产品、工具和方法还比较单一，缺乏应对复杂需求的灵活创新能力。总的看来，银行业信息化还有很长的路要走，信息科技风险将成为当前和未来较长时期银行业的重要风险领域之一。

银行业信息化既不能因为成绩而骄傲自满，也不可因为差距而妄自菲薄，更不可因

为困难而畏首畏尾。各银行业金融机构要勇于直面困难、主动迎接挑战，坚决按照国家信息化总体战略部署，切实坚持“自主可控、持续发展、科技创新”的基本方向，紧紧抓住信息化发展机遇，推动信息服务和信息安全再上新台阶。一是借助信息化推动银行业金融机构治理能力现代化。积极引入先进的信息科技治理和管理理念，运用现代信息技术缓解治理中的信息不对称问题，推动流程银行建设，提高治理有效性。同时，理顺信息化建设的体制机制，加快信息化建设进程，为银行业转型发展提供有力保障。二是依托信息化推动金融服务智慧化。要充分利用互联网、移动计算蓬勃发展的大环境，积极应用大数据等新兴技术，创新思维模式，充分发挥金融数据和信息的价值，研发智能化、个性化、便捷化的产品和服务，灵活响应客户诉求，努力改善客户体验，尽力发掘潜在客户需求，增加产品和服务的吸引力，培育更为坚实的客户基础，形成新的业务和利润增长点。三是以自主创新增进安全可控能力。要坚持市场起决定作用的基本方针，探索形成以研发创新支持应用推广、以市场应用激发创新动力的良性正反馈机制。推动应用自主创新信息技术，建立自主创新信息技术落地银行业的配套机制，力争金融领域关键信息技术自主创新占比逐步提高，不断提升信息系统的开放性、灵活性和整体集约化水平。四是利用信息技术强化行业协作。要加强银行业信息化建设的统筹规划，促进信息化资源的集约共享，提升数据（灾备）中心布局的合理性，增强同业协同协作，共同应对外包集中度等风险。

为更好地推进落实银行业信息化战略，由银行业信息科技风险管理高层指导委员会指导推动，编著了“银行业信息化丛书”（简称“丛书”）。这套“丛书”致力于挖掘、研究、总结、提炼和传播国内外信息化最佳实践、宝贵经验和最新成果，内容涵盖银行业信息科技治理与管理、信息系统开发与应用创新、信息安全、基础设施与运行维护、信息科技监管等主要领域，可为银行业信息科技人才培养提供一些基础性、前瞻性、实用性的知识和信息。

展望未来，银行业信息化任务艰巨、时间紧迫。希望银行业在有关各方支持下，推动信息化工作更加积极主动、规范有效、科学前瞻，为我国银行业持续健康发展、提升服务水平提供坚实的支撑，为增强国家网络安全保障能力、提升信息化建设水平提供有力支持，为贯彻落实创新驱动发展战略、实现中华民族伟大复兴的中国梦做出积极贡献。

尚福林

前　言

随着信息技术尤其是计算机及互联网技术的飞速发展，金融行业每天都在产生着海量的数据。对这些数据进行统计、分析，挖掘出隐藏在数据内部有价值的信息，为金融行业的决策提供指导，已经成为具有挑战性的新课题。在大数据时代，金融行业尤其是银行业对数据挖掘与分析技术的需求已经迫在眉睫。

在这种背景下，本书从数据挖掘与分析技术的基础知识出发，紧紧把握金融数据挖掘与分析的最新动向，对数据挖掘与分析技术及其在金融行业中的应用进行了详细介绍，并对未来金融数据挖掘与分析的发展进行了展望。

本书分为4篇，共14章内容。

第1篇为基础篇，主要介绍了数据挖掘的背景、应用及大数据的基本思想，具体内容包括：第1章数据挖掘概述，主要介绍了数据挖掘技术的发展和应用领域；第2章金融数据挖掘概述，主要对数据挖掘技术在金融行业中的应用现状和必要性，以及金融数据挖掘的过程进行了介绍；第3章基于大数据的金融数据挖掘概述，介绍了大数据产生的背景及特点，并从大数据视角探讨了金融数据挖掘的新思维及系统架构。

第2篇为技术篇，主要介绍了与大数据相关的数据挖掘技术，具体内容包括：第4章数据仓库技术，主要介绍了数据仓库的基本概念及与数据仓库相关的数据预处理技术，并对基于数据挖掘技术的数据仓库系统框架设计进行了介绍；第5章数据挖掘与分析技术，对各种典型的数据挖掘与分析技术进行了较为全面的讲解；第6章大数据挖掘与分析技术，首先介绍了NoSQL数据库技术及海量数据分布式存储技术，接着对大图数据、序列数据等复杂数据的挖掘与分析进行了介绍，最后对新兴数据挖掘工具Spark、Mahout等进行了介绍。

第3篇为应用篇，主要介绍了数据挖掘技术在银行具体业务中的应用模式，具体内容包括：第7章数据挖掘技术在零售银行信用风险管理中的应用，首先介绍了数据挖掘技术在银行风险管理中的应用现状，然后以信用卡领域中的信用风险管理为例，介绍了申请风险评分模型、行为风险评分模型以及欺诈风险评分模型的开发和应用；第8章数据挖掘技术在巴塞尔资本协议下的银行风险计量中的应用，首先介绍了巴塞尔资本协议

的基本概况，接着对数据挖掘技术在新巴塞尔资本协议风险计量中的应用现状进行了介绍，最后介绍了基于新巴塞尔资本协议三大支柱的银行信用风险计量方法及主要过程；第 9 章数据挖掘技术在客户关系管理中的应用，首先介绍了在不同的生命周期阶段企业面临的不同业务要求，然后着重介绍了客户细分的方法与技术，接着介绍了客户价值的内涵及其评价体系的建立，最后介绍了基于数据挖掘技术的 CRM 设计方法；第 10 章数据挖掘技术在金融市场分析与预测中的应用，主要介绍了数据挖掘技术在金融产品价格模型、证券投资组合、交易平台等金融市场中的应用；第 11 章数据挖掘技术在互联网金融中的应用，对互联网金融中大数据技术的应用做了较为全面的介绍，并通过对基于大数据的征信管理、反欺诈检测和客户关系管理三个方面的具体介绍，使读者了解大数据技术具体的应用场景、应用方法和应用范围。

第 4 篇为展望篇，主要介绍了数据挖掘技术在金融行业中的应用前景和面临的问题，具体内容包括：第 12 章基于大数据的金融科技战略与实施，首先介绍了基于大数据的金融科技建设思路，接着介绍了数据挖掘技术下基于风险与收益平衡的差异化经营和基于客户需求的差异化经营，最后结合案例介绍了差异化思路与智能化工具的互动循环实践；第 13 章数据安全与隐私保护，主要介绍了云计算中的数据安全问题及大数据分析带来的隐私保护问题；第 14 章应对策略，针对当前的大数据浪潮，给出了金融数据挖掘与分析领域的应对策略。

本书由郑志明、缪绍日、荆丽丽、吴美玲、杨益明、黄熹微、林道新、汤瑛、李文锋、陈佳蔚编著。

在本书的编写过程中，参阅了大量的文献资料，在此向这些文献资料的作者表示衷心的感谢！

由于编者水平有限，再加上编写时间仓促，书中难免存在缺点和不足之处，恳请广大读者批评指正！

编著者

目　录

第2篇 技 术 篇

第3篇 应 用 篇

第 4 篇 展 望 篇

第1篇

基础篇

通过对数据进行分析来获取有用的知识是人们一直努力的方向。从统计分析到数据挖掘，再到大数据挖掘技术，数据挖掘与分析的理论和技术体系日益成熟。同时，数据挖掘与分析技术正逐渐应用于各个领域，特别是在金融领域，其信息化建设比其他行业成熟，积累的数据质量比较好，规模也比较大。随着金融企业服务模式的转型和提升，其对客户和业务理解与分析的需求不断提升，数据挖掘与分析技术在银行中的应用也日益深入。在互联网金融快速发展的今天，数据挖掘与分析技术已经成为金融企业信息化建设的重要支撑技术之一。随着大数据时代的到来，数据挖掘与分析的方法、技术和理念面临新的变化。在技术层面上，大量基于云计算平台的并行分布式计算平台和数据管理平台的使用，为分析和管理大规模的数据提供了新的技术支撑；在分析思想和模式层面上，大数据时代数据挖掘的思路存在很多的差别，多源异构数据的融合成为大数据时代数据挖掘与分析的特点。

第1章
数据挖掘概述

数据挖掘技术是数据处理技术发展到一定阶段的产物。经过多年的发展，数据挖掘技术体系日益成熟，在很多领域得以应用，甚至已经成为某些领域的核心技术。

总体而言，数据挖掘技术的不断发展，促进了人们对数据的理解和利用数据对未知事物进行预测的能力。

1. 对数据进行探索和理解的能力　主要是发现数据中包含的模式和规律，主要的技术包括三类：

（1）联机分析处理（On-Line Analytical Processing，OLAP）技术。OLAP 是一种对数据从不同的角度、不同层次进行观察和统计的技术。用户可以通过分块、下钻和上卷等操作，对数据实现汇总、合并和聚集计算，并通过自动报表等技术实现对信息和规律的展示。通过 OLAP 技术可以初步掌握数据的分布和特征。

（2）关联分析技术。关联分析技术是指分析数据项之间的关联性，即通过寻找数据集中数据项之间的共线性，进而获得数据之间的关联关系。例如，通过顾客同时购买啤酒和尿片的现象，了解这两种商品之间存在的关联关系。

（3）聚类技术。聚类是指以数据之间的相似性为基础，实现对数据项的划分，真正实现“物以类聚，人以群分”。例如，通过对银行的客户进行分析，将他们分成若干不同的类别，这样就可以针对不同类别的客户制定不同的销售策略。

2. 利用对数据的理解实现对未知事物预测的能力　其主要数据挖掘技术是分类技术。分类技术是指通过对现有数据集进行分析，建立预测模型，并利用预测模型实现对未知事物或事件的分析。常用的分类预测模型包括决策树、神经网络、贝叶斯网络等。

随着数据挖掘支撑技术日益成熟及其逐步发挥的价值作用，数据挖掘技术正逐渐成为企业信息化建设的主要方向，并在各行业中深入应用。在企业信息化建设过程中，不管是国外还是国内，都逐步从支持企业核心业务管理与实施，转向通过对企业各个系统的数据整合，建立面向企业信息分析的数据仓库平台，以避免企业陷入“数据丰富但信息贫乏”的境况。通过对企业信息的分析，为企业的决策人员提供进一步决策的依据，避免其仅依据自己的直观感觉来进行决策。数据挖掘技术不仅广泛地应用于企业的

决策中，而且在其他领域也得以广泛应用，特别是在互联网和电子商务领域，通过对海量客户信息的挖掘与分析获取客户的特点，进而实现对客户的精准营销。在生物信息学领域更是由于存在对海量生物遗传数据进行分析能力，才使得这一学科得以形成和发展。正是由于数据挖掘技术的发展才逐步填补了数据和信息之间的鸿沟，将数据“坟墓”转换成知识“金块”。

1.1 数据挖掘技术的发展

20 世纪 90 年代，随着各个领域信息技术应用的推广，各行业建设了各种类型的信息系统，这些信息系统经过多年的运行，积累了大量的数据。为了发挥这些数据的价值，人们开始探索从海量的数据中获取有用的信息和知识的技术。数据挖掘技术就是在这一背景下产生的。当前，数据挖掘技术已经应用于各种领域，包括金融、智能交通、智慧城市、互联网、生物信息等。

数据挖掘是信息技术不断进化的结果。对数据进行分析和理解是数据处理技术的重要内容，大量的数学和计算机领域的专家一直致力于这方面的研究和开发工作。总体上讲，数据挖掘技术的产生有三个来源，如图 1-1 所示。

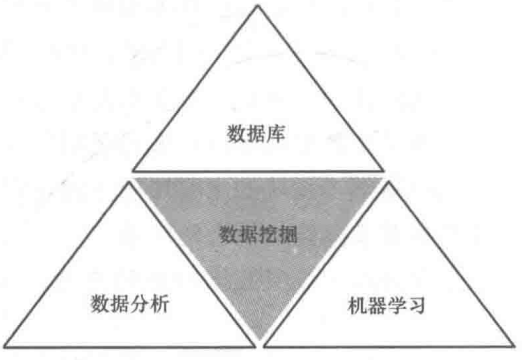


图 1-1 数据挖掘技术的产生来源

1. 数据库技术 通过对数据的查询和统计可以了解数据的基本分布情况，以及固有的模式和特征，是实现数据分析的一种初步而有效的方式，而数据库技术可以实现对大规模数据的有效管理和高效访问。数据库领域经过多年的发展，产生了一系列产品，在对大规模数据的分析方面主要可以提供两个方面的支撑：一是数据的组织和管理，即数据库技术针对各种类型的数据形成了一套有效的数据管理体系，包括数据库与数据仓库模式、元数据体系、数据质量体系等，这些技术可以较好地对各种不同类型、来源、结构的数据以一种比较固定和统一的模式进行管理；二是对海量数据的高效访问，即在数据库技术中不仅可以实现以约束为主的面向数据元组的查询，而且可以实现各种基于聚集函数的统计查询。在数据库管理系统中通过查询优化、事务处理、索引等机制，可以实现对数据的高效访问，以及对数据的初步理解和分析。数据库技术的发展历程如下：

（1）数据库技术起源于 20 世纪 60 年代，其原始动力是将原来以文件形式存储的数据以数据库的形式组织起来，以保证数据的一致性，并支持对数据的访问。自 20 世纪 70 年代以来，数据库管理系统的研究和开发日益系统化，从数据模型角度来看，历经层次和网状模型，最终发展到关系数据库模型的阶段。目前，关系模型已经成为主流