管理、决策与信息系统丛书

DATA MINING IN FINANCE

金融数据挖掘

马超群 兰秋军 陈为民 著

科学出版社
www.sciencep.com

(F-0891.0101)

金融数据挖掘

DATA MINING IN FINANCE

ISBN 978-7-03-018651-5

9 787030 186515 >

定 价：35.00 元

高等教育出版中心·经管法出版分社

电 话：010-64002235

E-mail：mayue@mail.sciencep.com

管理、决策与信息系统丛书

金融数据挖掘

DATA MINING IN FINANCE

马超群　兰秋军　陈为民　著

国家自然科学基金资助项目(70371028)成果

科学出版社
北京

内 容 简 介

金融管理研究的一个显著特点是数据分析量大、不确定性因素多，面对当今时代的海量金融数据，基于传统统计技术建立的模型假设条件多，实际应用难以奏效。数据挖掘是20世纪90年代中期兴起的新技术，是发现数据中有用模式的过程，其目的在于使用所发现的模式帮助解释当前的行为或预测未来的结果，以人们容易理解的形式提供有用的决策信息。

本书对一些相对较成熟的挖掘技术的讨论，阐述其用途、解决思路、需注意的主要问题、步骤，以金融领域的具体案例介绍模型与方法的应用。全书包括金融数据预处理、分类技术、预测、聚类技术、神经网络与支持向量机、异常数据挖掘，并且介绍了这些领域的一些最新方法。

本书可作为信息管理与金融类专业本科生和研究生的教材，也可供从事数据挖掘技术与应用研究的科研人员、金融市场数据分析人员，以及数据挖掘应用软件的开发者参考。

图书在版编目(CIP)数据

金融数据挖掘/马超群，兰秋军，陈为民著. —北京：科学出版社，2007
(管理、决策与信息系统丛书)
ISBN 978-7-03-018651-5

Ⅰ. 金… Ⅱ. ①马…②兰…③陈… Ⅲ. 金融-数据-研究 Ⅳ. F830.41

中国版本图书馆CIP数据核字(2007)第024887号

责任编辑：马 跃 / 责任校对：桂伟利
责任印制：张克忠 / 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

铭浩彩色印装有限公司印刷
科学出版社发行 各地新华书店经销

*

2007年4月第 一 版 开本：B5(720×1000)
2007年4月第一次印刷 印张：18 1/2
印数：1—3 000 字数：343 000

定价：35.00元

(如有印装质量问题，我社负责调换〈环伟〉)

致　谢

几多辛苦之后，书稿即将开印，倍感欣慰，同时不得不表达我们心中油然而生的那份感谢。

这么多年来，我们在金融数据挖掘与风险管理领域开展了一系列的研究，先后获得了国家自然科学基金、国家社会科学基金、国家“863”计划、教育部新世纪优秀人才支持计划、教育部优秀青年教师支持计划、教育部高等学校博士学科点专项科研基金、留学回国人员科研启动基金等资助，我们衷心感谢以上相关单位的大力支持。

在研究过程中我们得到了国防科技大学汪浩教授、加拿大 York 大学吴建宏教授、朱怀平教授、甘国君博士、美国 Florida 大学 Uryasev 教授、Alabama 大学吴志坚教授等的指导与帮助，在此，向他们表示最诚挚的谢意！

我们还要特别感谢受邀来我们学术梯队作学术报告的国内外专家与教授，在与他们的交流中，让我们受益匪浅。

最后，感谢所有给予我们帮助与支持的相关研究工作者与科学出版社的同志。

作　者

2006 年 10 月

丛 书 序

管理理论、决策科学与信息系统技术在20世纪获得了巨大的发展。在20世纪80年代，为了推动这三大领域在中国的发展以及推动这些领域之间的学科交叉研究，中国科学院管理、决策与信息系统重点实验室在科学出版社的支持下编辑出版了这套“管理、决策与信息系统丛书”。这套丛书不求全而求新，以反映最新的研究成果为主。经过编委会的各位专家，特别是前任主编许国志院士的努力和作者们的辛勤劳动，这套丛书在社会上尤其是在科学界得到了广泛的关注和好评。

回顾管理理论的发展历史，我们不难发现一个趋势：系统的概念和方法越来越多地应用到管理的各个方面，并成为管理理论发展的第三阶段的重要特征。管理理论的第一阶段形成于20世纪初，以F. W. Taylor为代表，倡导科学的管理，为提高工厂劳动生产率而提出了标准化原理。管理理论的第二阶段，从20世纪20～30年代开始，以行为科学为特点，主要代表有A. H. Maslow，K. Lewin，R. Jannenbaum和D. McGregor等人。他们研究人的需要、动机、激励和定向发展；研究正式和非正式团体的形成、发展和成熟；研究个人在团体中的地位、作用、领导方式和领导行为等。管理理论的第三阶段出现在第二次世界大战后，这一阶段有各种学派，例如社会系统学派、决策理论学派、系统管理学派、管理科学学派和经验主义学派等。他们从不同角度强调系统的概念、理论和方法。这三个发展阶段并非截然分开，而是相互交叉的。

不论管理理论有多少学派，人们大致可以将它们分成三种模式：机械模式、生物模式和社会模式。生物模式认为：组织像一个生物，有头脑机构，有职能部门和分支机构。一个企业的目标可以分解，各部门完成其中的一部分。在这种模式下，目标管理得以发展。社会模式认为：各级组织都是一个交互的系统，它们有共同的目标、交互作用和信息联系，管理者是交互作用的中心。其特点是强调交互式管理(Interactive Management)和强调以系统方法来管理。这正是它不同于传统管理的地方。而传统管理大致可分为三类：回顾式(Reactive)管理、被动式(Inactive)管

理、预测式(Preactive)管理。回顾式管理是在自下而上地总结过去经验的基础上，去发现组织的弱点，找出克服其弱点的措施，并在条件允许下去逐个地解决问题。被动式管理的特点是危机管理，是“救火队”，领导疲于处理当前各种各样的问题。而预测式管理的决策基于对今后的经济、技术、顾客行为和环境等的预测。这三类管理可以混合成各种样式的管理方式，正像红、黄、蓝可以组成各种颜色一样。交互式管理强调系统的方法，认为某个企业出现的市场问题绝不仅仅是一个市场问题，而与R&D、生产、原材料供给和人事等有关，是一个系统的问题。回顾式管理的弱点是缺乏系统的观点。交互式管理强调要设计可见的未来，创造一条尽可能实现它的道路，这是“救火队”所不能做到的，但它又不把一切都寄托于预测。交互式管理还强调“全员参与”和“不断改进”。

决策理论学派以E. W. Simon等人为代表，是从社会系统学派中发展起来的。它认为决策贯穿于管理的全过程，管理就是决策。决策的优劣在很大程度上依赖于决策者的智慧、素养和经验。计算机技术的发展不仅使人们能够快速地解决决策中的复杂计算问题，而且可以有效地进行决策过程中的信息处理、分析等工作，从而达到提高决策质量的效果。今天正处在新的发展阶段的决策支持系统(DSS)和管理信息系统(MIS)正是集管理理论、系统理论和信息技术三大领域的交叉学科方向，它们为解决许多复杂决策问题提供了有力的工具。粗略地说，决策问题大致可分为三个层次：战略决策、结构决策和运行决策。战略决策是指与确定组织发展方向和远景有关的重大问题的决策。结构决策是指组织决策，运行决策是指日常管理决策。

从信息论的观点看，整个管理过程就是一个信息的接收、传输、处理、增功与利用的过程。计算机信息处理技术应用于管理走过了三个阶段：数据处理(EDP)、管理信息系统和决策支持系统。作为管理信息系统和决策支持系统的支持环境，相对独立于计算机软件的开发，需要研究和建立各类管理信息系统独特的支持软件系统和开发环境，例如分布式数据库管理系统和分布式知识库管理系统，面向用户、通用性较强和面向特殊用户的模型库、方法库管理系统，以及一些专门的用户接口语言。

展望未来，管理、决策与信息系统这个交叉学科的研究领域的发展有以下几个趋势：

1. 更加重视人的行为的研究，企业的管理将不仅强调竞争，而且应在竞争的前提下注重合作与协调；

2. 非线性建模与分析，将取得大的突破；

3. 互联网的飞跃发展，将为管理与决策分析提供新的研究问题以及支持平台。

这些趋势有两个重要特点：(1) 利用信息技术与数学中的最新成就去研究管

理与决策问题;(2)通过观察管理决策与信息系统发现其规律,形成数学与信息科学中具有挑战性的研究课题。

在这套丛书的编辑出版中,我们将不仅注重每本书的学术水平,而且也关注丛书的实用价值。因此,这套丛书有相当的适用面。丛书的作者们将竭尽全力把自己在有关领域中的最新研究成果和国际研究动态写得尽可能地通俗易懂,以便使更多的读者能运用有关的理论和方法去解决他们工作中遇到的实际问题。

本丛书可供从事管理与决策工作的领导干部和管理人员、大专院校师生以及工程技术人员学习或参考。

汪寿阳

序　言

金融——这一维系一国经济运行的纽带与联系各国经济的桥梁，在各国经济发展中占有举足轻重的地位。在金融市场全球化、电子化、虚拟化的发展背景下，当今世界进入到了一个信息化和数量化的时代，金融市场得到了飞速发展，资金转移速度不断加快，金融业每天产生的数据正在以惊人的速度增长，不过“数据丰富但知识贫乏”却已经成为了一种普遍现象。人们期待着能从这些汪洋大海般的金融数据（基金或者投资者帐户、股票价格、交易量、认购、申购、定期定额、赎回、预约赎回、基金转换、限制申购、巨额赎回、非交易过户、红利再投资、分红方式变更、转托管业务、资金清算，基金参数、其他相关业务数据）当中及时有效地挖掘出高附加值的信息资源或有用的知识为其经营管理决策服务。在这种需求推动下，数据挖掘的概念被提出来，并迅速得到学术界、IT 厂商以及各行业人士的普遍重视，与此相关的理论、技术与应用正得到蓬勃发展。

数据挖掘是一门交叉学科，它汇聚了数据库/数据仓库技术、机器学习、统计学等领域的技术与方法。同时它又是一门应用性很强的学科，各行业从其具体应用出发，有着不同的挖掘需求与方法。金融业是一个信息驱动的行业，信息技术已经广泛渗透到金融行业各个机构、各项业务、各个环节。信息技术的应用水平已成为衡量“新世界、新金融、新银行”的一个重要标准。在加入 WTO 后，我国金融企业不可避免地需要更广泛、更深入地参与国际竞争。而随着电子商务的日趋成熟与发达，营业网点的数量难以成为金融企业核心竞争力。网上银行、网上证券等基于互联网开展的金融业务将会得到迅猛发展并成为潮流。这种趋势推动着金融信息基础设施建设的迅速发展，也为金融企业累积更多、更完善的数据，同时也对数据挖掘技术的应用提出了更强劲、更紧迫的需求。目前中国金融市场上正好缺乏自主开发的先进的现代金融数据挖掘技术与方法。而国际金融服务机构已经盯住中国这个巨大的市场，中国金融行业普遍有建立适合中国国情的金融管理体系、开发研究新的管理技术的共识，许多有识之士也早就认识到决策智能化和管理信息化

是金融企业提升综合竞争力的必然选择，挖掘数据价值，推进知识管理，是金融企业创新业务模式的根本途径，金融企业对该技术的发展一直有着热切的期盼。因此，加强金融管理技术的研究与开发已是实践的迫切需要，时代的综合要求。

一门新技术的应用与推广，需要在理论和方法上事先做大量创造性、前瞻性的探索工作。本专著是在国家自然科学基金资助下完成的研究成果，是作者多年来对该领域研究探索的一个总结。它也是国内第一本关于金融数据挖掘的专著。该书提出的一些金融数据挖掘模型、建模思想和算法都具有重大的创新性，将对研究开发金融数据挖掘的相关系统具有重要的指导作用，进而对缩小我国在该领域与国际水平的差距，推动我国金融管理向科学化、定量化、信息化、可视化管理与远程监控方向发展，提高金融管理的效率，增强金融风险的识别与控制能力具有重要的理论与现实意义。

本书一方面既系统地总结了针对金融领域的一些成熟的数据挖掘理论与技术，如神经网络、决策树、聚类、Apriori 算法等；也对数据挖掘发展中的一些前沿技术，如支持向量机等做了精辟的阐述；更重要的是反映了作者近几年来的创新研究成果，如偏好的相似性度量问题、基于共同机制的关联模式挖掘算法、局部时间序列挖掘方法 TSEOPM 等。该书在写作方式上，先介绍数据挖掘理论、技术模型和方法，然后辅以金融领域的具体应用案例，这样既能使读者在理论与方法上对挖掘模型有系统和深刻地理解，同时也能使读者从具体案例的研读中对模型的价值获得感性认识。这无论是对数据挖掘技术的学习者、金融数据分析人员，还是数据挖掘应用软件的开发人员都是大有裨益的。

石 勇

2007 年 1 月

前　言

金融管理研究的一个显著特点是数据分析量大、不确定性因素多，基于传统统计技术建立的模型假设条件多，实际应用难以奏效；当今，金融正处在一个信息化和数量化的时代，每天都有不计其数的数据在产生。在这种情况下，如何刻画中国金融市场的本质特性与特征量以回答金融市场不断演化的复杂性机理；如何对金融行业进行风险识别、评估与度量；如何从“海量”的金融实际数据（基金或者投资者账户、股票价格、交易量、认购、申购、定期定额、赎回、预约赎回、基金转换、限制申购、巨额赎回、非交易过户、红利再投资、分红方式变更、转托管业务、资金清算，基金参数、其他相关业务数据）中挖掘出高附加值的信息资源为金融机构及其监管部门提供科学化的管理决策支持技术已成为我国金融管理研究领域所迫切需要解决的问题。

本书提出的具有突破性意义的建模思想及其算法、为解决数据挖掘过程中的问题而提出的一些具有创新价值的方法、研究与开发出的金融数据挖掘高端技术，所有这些对推动高端金融科技发展，缩小中国在此领域与国际水平的差距，推进中国金融管理向科学化规范化方向进一步发展，对提高金融监管效率，增强风险识别、预警、防范与控制能力，对中国金融管理实现定量化、信息化、可视化管理与远程监控，对丰富金融数据挖掘技术都具有重要的理论与现实意义。

数据挖掘是20世纪90年代中期兴起的新技术，是发现数据中有用模式的过程，其目的在于使用所发现的模式帮助解释当前的行为或预测未来的结果，以人们容易理解的形式提供有用的决策信息。数据挖掘技术的应用领域非常广泛，各行各业只要涉及到大量数据的分析处理，数据挖掘就有用武之地。而在金融领域，许多金融业务活动，如客户分析、投资决策、风险管理、价格预测都需要对大量历史数据进行分析，分析数据是许多金融从业人员的“基础”与“日常性”工作。特别是由于我国金融信息化的快速发展，金融领域积累了非常庞大的数据，如何将这些数据化成有用的决策信息，是令很多金融分析人士十分感兴趣的问题。希望本书能给

予他们一定的帮助。

关于数据挖掘技术的书籍已经不少，但大多数是从理论与技术角度进行介绍，较少以实例的形式对其具体应用进行描述。数据挖掘过程包括数据收集、数据选取与准备、挖掘模型的建立与检验、挖掘结果的解释与验证以及结果的应用。每一个步骤都非常重要，而数据挖掘本身是一个多学科交叉领域，涉及机器学习、数据库、统计学等众多学科知识，其具体的研究内容非常广泛。本书基于作者对金融数据挖掘多年的研究经验，其写作目的是通过对挖掘技术的讨论，阐述其用途、解决思路、步骤，再辅以金融领域的具体案例介绍模型与方法的应用，从而使从事金融行业相关读者，既能在理论与技术上对其有较深刻的理解，又能通过对具体案例操作的研读获得感性认识。

全书分为 9 章，第 1 章绪论，从整体的角度介绍数据挖掘产生的背景、技术、步骤以及一些关键问题，讨论金融数据的基本特点与挖掘业务需求；第 2 章金融数据预处理，主要讨论金融数据来源、数据可能面临的错误与去噪处理方法；提出的小波去噪方法克服了频谱分离、卡尔曼滤波、维纳滤波、匹配滤波等常见的去噪方法所不能解决的问题，对具有非平稳、非线性、缺乏先验信息等特点的金融时间序列数据的去噪效果比较好；第 3 章关联规则挖掘技术，在介绍多层关联规则挖掘与多维关联规则挖掘的基础上给出了银行卡关联规则挖掘的案例分析，提出了一种基于共同机制的时间序列关联模式挖掘算法，设计出了一种高效率获取不同序列间关联模式的算法，该挖掘方法快速、简单，能探测出微观局部隐含存在的关联性；第 4 章分类技术，首先介绍了分类的基本概念与方法及评价，探讨了判别式分类、决策树分类、贝叶斯分类和粗糙集分类方法，以及分类技术在信用卡管理中的应用；第 5 章预测技术，介绍了传统的线性回归与非线性回归方法，灰色预测与组合预测技术，以及混合预测模型在股票价格预测中的应用；第 6 章神经网络与支持向量机，主要介绍了前向型神经网络、Hopfield 网络、自组织特征映射神经网络、统计学习理论、支持向量机及其在金融预测中的应用；第 7 章聚类分析，主要介绍了聚类的相关概念，研究了数据类型及相似性度量、分割聚类算法、层次聚类法、基于密度与模型的聚类方法；第 8 章时间序列数据挖掘，首先给出了时间序列相似性度量的一般方法，从心理偏好角度研究了时间序列相似性度量，提出了一种新的时间序列相似度量的主观偏好模型及其偏好系数的“锚点”估计方法。该方法从水平偏移、幅度伸缩和波动一致性等角度来度量相似程度，并将投资者决策偏好融入相似度模型，从而在数据挖掘过程中，可充分利用挖掘者的经验背景知识，提高了模式的有趣性和有效性；创造性提出了用金融时间序列局部模式来挖掘金融市场局部规律特征的建模思想；提出了一种新的具有预测性的时间序列事件征兆模式挖掘方法 TSEOPM(Time Series Event Omen Pattern Mining)，该方法主要基于“模式点聚集性”原理，通过聚类、候选模式生成以及候选模式判别技术获取序列中隐含

的预测性序列模式。它与以往的序列模式挖掘重在发现“所有”频繁的模式不同，主要是将目标聚集于能预测挖掘者感兴趣的事件的序列模式。因而其搜索空间大为减少，而且模式是有趣的，避免了常见的所谓“坏模型”问题对检验结果的影响；第9章异常数据挖掘，给出了异常挖掘的一般方法，以及异常挖掘在金融市场中的应用。

本书可作为信息管理与金融类专业高年级本科生和研究生的教材，也可供从事数据挖掘技术与应用研究的科研人员、金融市场数据分析人员，以及数据挖掘应用软件的开发者参考。

作 者

2006年12月

目　录