A PROSPECT FROM BIG DATA PERSPECTIVE

金融数据挖掘

基于大数据视角的展望

许伟　梁循　杨小平◎主编

知识产权出版社
全国百佳图书出版单位

A PROSPECT FROM BIG DATA PERSPECTIVE

金融数据挖掘

基于大数据视角的展望

许伟　梁循　杨小平◎主编

知识产权出版社
全国百佳图书出版单位

内容提要

全书结构分为五个篇章。第一篇介绍了数据挖掘方法。第二篇是银行数据挖掘篇，介绍了基于神经网络和支持向量机的信用评分方法。第三篇是证券数据挖掘篇，探讨了基于多种数据挖掘方法的股票价格预测、金融市场价格预测及股票自动交易系统。第四篇是保险及其他数据挖掘篇，研究了基于数据挖掘的保险欺诈监测、企业破产预测、财务报表欺诈监测等问题。第五篇从大数据的视角对金融数据挖掘进行了扩展和展望。

本书的读者可以是对数据挖掘算法感兴趣的计算机专业人士或是对金融信息挖掘感兴趣的领域专家，也可作为金融信息工程方向的工程硕士教材或参考书。

责任编辑：江宜玲　　**责任出版**：刘译文

图书在版编目（CIP）数据

金融数据挖掘　基于大数据视角的展望/许伟，梁循，杨小平主编. —北京：知识产权出版社，2013.2

ISBN 978-7-5130-1879-1

Ⅰ.①金… Ⅱ.①许… ②梁… ③杨… Ⅲ.①金融—数据收集—研究 Ⅳ.①F830.41

中国版本图书馆 CIP 数据核字（2013）第 024696 号

金融数据挖掘　基于大数据视角的展望

JINRONG SHUJU WAJUE　JIYU DASHUJU SHIJIAO DE ZHANWANG

许　伟　梁　循　杨小平　主编

出版发行：知识产权出版社

社　　址：北京市海淀区马甸南村 1 号

发行电话：010-82000860 转 8101/8102

责编电话：010-82000860 转 8339

印　　刷：知识产权出版社电子制印中心

开　　本：720mm×960mm 1/16

版　　次：2013 年 6 月第 1 版

字　　数：226 千字

ISBN 978-7-5130-1879-1

网　　址：http://www.ipph.cn

邮　　编：100088

传　　真：010-82000507/82000893

责编邮箱：jiangyiling@cnipr.com

经　　销：新华书店及相关销售网点

印　　张：14

印　　次：2013 年 6 月第 1 次印刷

定　　价：45.00 元

前　言

进入21世纪，中国金融业对信息化工作前所未有的重视，众多金融机构都建立起了自己的数据平台，形成了金融机构网络和垂直业务体系，实现了金融数据大集中。这些数据有三个特点。第一个特点是数据量大，一般达到PB级。第二个特点是类型多，如非结构化数据、半结构化数据、流数据等。第三个特点是价值密度低，有用的数据含量少。这都是大数据的典型特点。随着金融数据基础设施硬件建设的不断发展，如何处理每天产生的大数据，进行科学的分析处理，挖掘隐藏在数据内部各种有价值的关联，并及时提供决策支持，成为摆在金融业面前的新课题。

大数据时代的到来，使得世界不得不更多地使用智能数据挖掘技术。目前，金融数据挖掘引起了众多学者和业界人士的广泛关注。在这种背景下，本书力求把握金融数据挖掘的最新动向，开发金融数据挖掘的典型案例，从大数据的视角加以思考和探索，并为金融数据挖掘研究和应用的发展提供有益的支持。

编写本书的另一个出发点是，近年来我国金融业迅猛发展，金融信息化人才的需求量大大增加，相当多的毕业生进入了金融信息化行业。为满足实践的需要，很多大学的软件学院设立了金融信息系或金融信息专业，培养一批又一批的金融信息专业人才。笔者最近几年一直参与讲授一些院校的金融数据挖掘专业的工程硕士课程，感到缺少这样一本教材，所以组织师生编写了本书。

本书介绍了金融数据挖掘的一些典型应用。本书与作者以前出版的《网络金融》《数据挖掘算法与应用》《互联网金融信息系统的设计与实现》《电子商务理论与实践》《网络金融信息挖掘导论》《网络金融系统设计与实现案例集》《互联网金融信息智能挖掘基础》和《支持向量机算法及其金融应用》八本书之间的关系见图0－1。本书是这些书籍在金融数据挖掘方面的延续和补充。

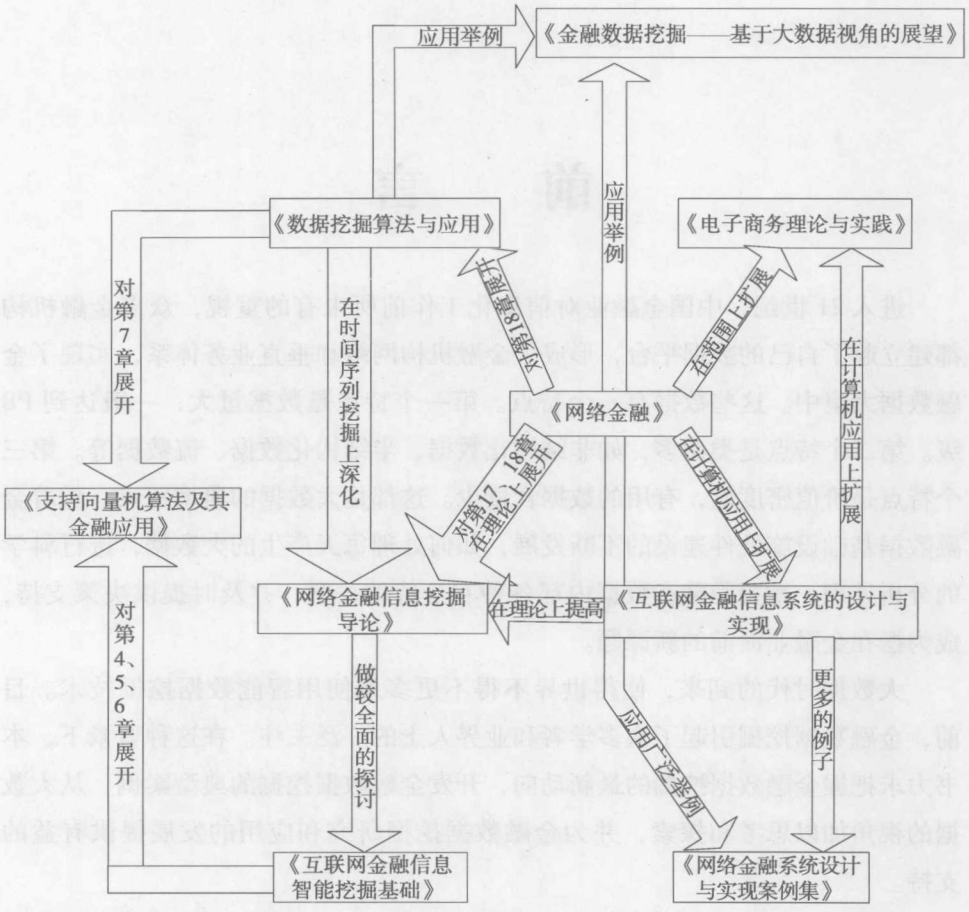

图 0-1　本书与另外八本书关系图

本书分为五篇。第一篇介绍了数据挖掘的基本方法。第二篇是银行数据挖掘篇，介绍了基于神经网络、支持向量机的信用评分方法及银行信贷的评价方法。第三篇是证券数据挖掘篇，讨论了基于粗糙集的股票价格预测方法、基于网络信息的金融市场价格预测研究及基于数据挖掘的股票自动交易系统。第四篇是保险及其他数据挖掘篇，选择性地探讨了基于数据挖掘的保险欺诈监测方法、基于 Logistic 回归的企业破产预测研究、基于数据挖掘的财务报表欺诈监测方法研究及基于时间序列模型的原油期货价格预测研究等问题。第五篇从大数据视角对金融数据挖掘进行扩展并进行了展望。

本书面向金融领域的实际问题，从数据角度来看银行、证券、保险等行业

进行挖掘的可行性和重要性，对信用评分、股市预测、保险分析等进行了深入研究，并取得了较好的应用效果。随着金融领域数据的剧增，一定会经历从金融数据研究到金融大数据研究的转变。虽然本章涉及一些大数据的处理方法和基本技术，但要走的路还很漫长。希望通过本书抛砖引玉，引起学者和业界人士对金融大数据分析与挖掘的重视，同时也希望本书能作为有效的分析工具对金融企业的发展提供决策支持。

本书对金融数据挖掘进行了深入研究，在银行、证券、保险及相关方向取得了一些研究进展。在银行数据挖掘中，针对信用评分、银行贷款等决策问题，利用神经网络、支持向量机等智能分析工具进行了深入研究和实证分析，取得了较好的应用效果。在证券数据挖掘中，对股票市场预测和股票自动交易系统进行了深入研究，分别提出了利用粗糙集方法提炼股票市场的预测规则，引入网络大数据增加股票市场的预测精度，基于小波分析和神经网络构造股票自动交易系统，为优化投资组合、提高资金使用效率提供了有效的工具。在保险及其他数据挖掘领域，分别对保险欺诈、企业破产、财务报表欺诈及原油市场预测进行了研究，提出了面向不平衡数据的保险欺诈监测方法，基于 Logistic 模型的企业破产预测模型、基于集成学习的财务报表欺诈监测模型及基于 ARIMA 的原油价格预测模型，取得了较好的应用效果。虽然目前大数据的研究与应用在金融业还处于初级阶段，但是价值已经显现出来。未来，大数据可能成为最大的金融交易商品。我们深信，未来金融大数据将会成为金融业进行重要活动的基础设施。

本书的特色在于面向金融应用，不仅介绍数据挖掘算法本身，更注重如何将数据挖掘方法应用到金融实际问题中。本书的实践证明，数据挖掘方法在金融领域大有所为。本书从应用实际出发，取得了较好的应用效果。值得一提的是，为了增强金融市场预测的效果，本书力求利用大数据分析技术研究网络信息等对金融市场的影响。大数据分析与挖掘技术在金融领域将大有所为，书中对大数据及金融大数据挖掘进行了一些探讨和思考。

由于目前专门针对专业硕士的教材不多，本书也仅仅是在这方面的一个探索。考虑到专业硕士应该以应用特别是领域前沿的应用为导向，在本书结构的构思上，我们以实例分析为主线展开。本书力图把数据挖掘算法和应用“揉”在一起介绍，力求活学活用。

本书的出版，得到中国人民大学科学研究基金（中央高校基本科研业务费专项资金资助）项目（10XNI029）、国家自然科学基金资助项目（70871001、71001103、71271211）、北京市自然科学基金项目（9122013、4132067）和北京市优秀人才培养资助。

作者的一些同事和学生，也参加了本书的编写。他们是王炎、马跃峰、张黛玲、林娜娜、纪阳、李启东、胡敏章、罗易、朱浩然、周晨曦、杜玮、王佳佳、程成、金鑫、王婴、桂斌，周杰等。

由于作者水平和时间的限制，书中一定存在不少缺点和错误，恳请读者批评指正。

目　录

第一篇　金融数据挖掘概述

第二篇　银行数据挖掘

第三篇 证券数据挖掘

第四篇 保险及其他数据挖掘

第五篇 基于金融大数据视角的展望

第一篇

金融数据挖掘概述

第1章 绪 论

近年来，随着金融信息化的迅速发展，金融机构已经搭建起数据平台，逐步实现数据大集中，形成金融数据。与此同时，数据挖掘技术在过去几十年里得到了长足的发展，技术与方法日趋完善，应用到了各个领域。金融领域利用数据挖掘技术，不仅可以用数据“说话”，为金融决策提供更加有效的支持，而且可以为金融服务提供更准确的信息和知识，为消费者提供有针对性的个性化服务。金融数据挖掘得到了众多学者和业界人士的广泛关注，在这种背景下，本书力求把握数据挖掘的最新动向，开发金融数据挖掘的典型案例，为金融数据挖掘的不断发展提供有益的支持。

1.1 金融领域进行数据挖掘的必要性

金融领域涉及银行、证券、保险及其他相关内容，包括银行信贷、信用评分、市场分析、投资组合、保险定价、智能定损、金融欺诈等。金融领域的研究内容相当广泛，但不确定性是金融市场的本质，也是金融领域需要研究的核心内容（马超群等，2007）。为了捕捉金融市场的不确定性，更好地提高金融市场效率，需要使用数据建模方法对金融市场进行有效刻画。目前，数据建模方法已经应用于金融领域，用以把握金融市场的规律和趋势，达到了良好的应用效果。但由于传统的数据建模方法基于一些有严格要求的假设，当假设条件不满足时，难以对金融数据进行建模，因此难以把握金融市场规律。随着金融行业的不断发展壮大，银行、证券、保险及其他相关机构不断融合，信息化程度大大提高。而且随着云计算技术的不断发展，金融数据正在逐步实现大集中。在这种情况下，不苛求严格假设的数据挖掘技术与算法在金融数据的支撑下就有了用武之地，并且发挥出极大的优势，为刻画金融市场的规律和趋势提供了有效的分析工具。

与其他领域的数据相比较，金融数据具有多种特点。（1）金融数据具有多样性。作为社会经济系统的一部分，金融系统的数据不仅受到物理数据（客户数据、交易数据、经济数据等）的影响，而且受到网络信息、心理行为信息的强烈影响，甚至一些主观数据的变化也会导致金融市场的剧烈波动。（2）金融数据的关系复杂。金融市场是一个复杂系统，数据之间的关系有时很难用一个简单的数学公式或线性函数来表示，呈现出高度的复杂性和非线性性。（3）金融数据具有动态性。金融市场随着时间的推移会发生剧烈变化，但仍受前期市场的影响，呈现出动态特征。为了更好地研究金融市场，需要利用这些物理数据、网络信息及心理行为信息。这些信息是不断变化的，便形成了一个巨大的数据仓库。金融数据的高度复杂性，使得一般的数据建模方法在进行金融数据建模时失效，而数据挖掘方法具有灵活性、自适应性及非线性等特征，在处理金融数据时可以达到较好的应用效果。实际的应用效果也证实了这一点，因此数据挖掘方法应用在金融领域是可行的。

1.2 金融数据挖掘的应用领域

数据挖掘方法已经应用到金融业的各个领域，从金融业务的划分来看，金融数据挖掘在银行、证券、保险及其他领域已经有了一些典型的应用案例，下面进行简要介绍，后续章节将展开详细论述。

1.2.1 银行数据挖掘

1. 贷款审批

随着企业的发展及消费者生活水平的提高，对银行贷款的需求越来越大，银行在决定是否发放贷款给客户时面临着严峻挑战。银行该如何评定一个新客户的级别并发放贷款？银行又该如何处理老客户提出的贷款申请？应用数据挖掘方法为贷款审批提供决策支持，是银行数据挖掘的一个重要应用领域。

2. 信用评级

信用评级是由专业部门或机构按照一定的方法和程序在对客户进行全面了解、调研和分析的基础上，得出其信用可靠性、安全性评价的管理活动。随着中国市场经济体制的建立，信用评级在防范商业风险、资本市场公平公正及为

商业银行的风险决策提供依据等方面发挥着重要作用。信用评级主要应用于用户信用评价、贷款审批决策等方面，数据挖掘方法在该领域有着广泛的应用前景。

3. 客户细分

随着银行业的迅速发展，争取优质客户成为银行间竞争的焦点。谁拥有优质客户，谁就赢得市场先机。为了更好地识别优质客户，针对不同客户进行差异化服务，成为银行客户关系管理的重要问题。因此，对银行客户细分变得越来越重要，客户细分研究也越来越多。利用数据挖掘方法，特别是聚类分析，可以清晰地发现不同类型客户的特征，挖掘不同类型客户的特点，为金融业的优质服务提供有效的决策支持。

1.2.2 证券数据挖掘

1. 市场预测

金融市场分析和预测不仅可以为企业投资带来可观的利润，而且可以规避市场风险，是目前最有吸引力的研究课题之一。在缺少风险回避机制的金融市场中，机构投资者往往倾向于短线操作，频繁的大额交易行为会引起股价的波动，而市场参与者在金融预测工具的帮助下，能够进行更加理性的投资决策。同时，决策过程对信息的需求反过来刺激信息披露的程度，不仅缓解了短期投资依赖内部消息的现象，也促进了有效市场的形成。但是，金融市场是一个复杂系统，呈现出复杂性、非线性等特征，传统方法由于苛刻假设不能很好地应用于金融市场预测，而数据挖掘技术的出现，为金融分析和预测提供了有效的分析工具。数据挖掘在金融市场预测中有着广泛的应用，并取得了大批的研究成果。

2. 投资组合

目前，投资组合理论广泛应用于金融市场，降低了金融市场的投资风险，提高了资金的使用效率。投资组合是基于历史收益—风险进行建模的，未考虑预期的不确定性。利用最新的数据挖掘技术不仅可以更好地刻画预期的不确定性，改进已有的投资组合模型，使之更加符合现实需求，同时可以为投资组合模型的求解提供更为精确的手段，从而为投资者提供更为精准的知识。

3. 自动交易

股票自动交易系统是一种避免受到投资者情绪影响的、借助计算机设计的理性交易系统。投资者在作出错误的购买决策时，往往容易陷入沉没成本误区，在面临损失时，人们往往是风险偏好的，总期待能收回投入成本，而不愿抛售已持有的股票，这往往导致更大的亏损。自动交易系统最大的作用就是止盈止损，如果购买决策错误，则应该把损失控制在一定范围以内，如果购买决策正确也不应该贪得无厌，判断股票价格不可能再涨的时候就应该卖掉。自动交易系统的核心就是自动交易算法，而数据挖掘技术为自动交易算法提供了技术支持。基于数据挖掘技术的自动交易算法已经被学术界开发使用，并在业界有着广泛的应用价值。

1.2.3 保险及其他数据挖掘

1. 交叉销售

保险交叉销售是一种借助 CRM（客户关系管理）发现顾客的多种需求，并通过满足其需求而销售多种相关服务或产品的新兴营销方式。通过使用数据挖掘方法，特别是关联规则，可以发现客户购买产品的关联与偏好，为客户关系管理提供有效的分析工具。

2. 欺诈监测

随着国内保险业的发展，保险欺诈问题日益突出，给保险公司和社会带来了极大危害。因此，研究保险欺诈监测与防范具有重要意义。利用数据挖掘方法，分析并确定欺诈行为的特征，从而对保险欺诈行为进行实时监测和预警，这对保险业的健康有序发展有着重要的应用价值。

3. 智能定损

保险公司收取保费后，在标的出险时需要及时进行理赔服务。由于标的具有差异性，为标的的精准赔付带来了困难。利用数据挖掘技术，可以挖掘已有标的的定损规律，从而对标的损失进行精确估计和预测，为保险智能定损提供了有效的分析工具。