

 大数据金融丛书 丛书主编 丁鹏

Quantitative
Matlab

量化投资

MATLAB数据挖掘技术与实践

卓金武 周英 编著

 中国工信出版集团

 电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

Q大数据金融丛书

Quantitative
Matlab

量化投资

MATLAB数据挖掘技术与实践

卓金武 周英 编著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

全书内容分为三篇。第一篇为基础篇, 主要介绍量化投资与数据挖掘的关系, 以及数据挖掘的概念、实现过程、主要内容、主要工具等内容。第二篇为技术篇, 系统介绍了数据挖掘的相关技术及这些技术在量化投资中的应用, 主要包括数据的准备、数据的探索、关联规则方法、数据回归方法、分类方法、聚类方法、预测方法、诊断方法、时间序列方法、智能优化方法等内容。第三篇为实践篇, 主要介绍数据挖掘技术在量化投资中的综合应用实例, 包括统计套利策略的挖掘与优化、配对交易策略的挖掘与实现、数据挖掘在股票程序化交易中的综合应用, 以及基于数据挖掘技术的量化交易系统的构建。

本书的读者对象为从事投资、数据挖掘、数据分析、数据管理工作的专业人士; 金融、经济、管理、统计等专业的教师和学生; 希望学习 MATLAB 的广大科研人员、学者和工程技术人员。

未经许可, 不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有, 侵权必究。

图书在版编目 (CIP) 数据

量化投资: MATLAB 数据挖掘技术与实践 / 卓金武, 周英编著. —北京: 电子工业出版社, 2017.1
(大数据金融丛书)

ISBN 978-7-121-30230-5

I. ①量... II. ①卓... ②周... III. ①Matlab 软件—应用—投资学 IV. ①F830.59-39

中国版本图书馆 CIP 数据核字 (2016) 第 259972 号

策划编辑: 李 冰

责任编辑: 李 冰

特约编辑: 田学清 赵树刚等

印 刷: 三河市双峰印刷装订有限公司

装 订: 三河市双峰印刷装订有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱

邮编: 100036

开 本: 787×980 1/16 印张: 27.5

字数: 717 千字 彩插: 4

版 次: 2017 年 1 月第 1 版

印 次: 2017 年 1 月第 1 次印刷

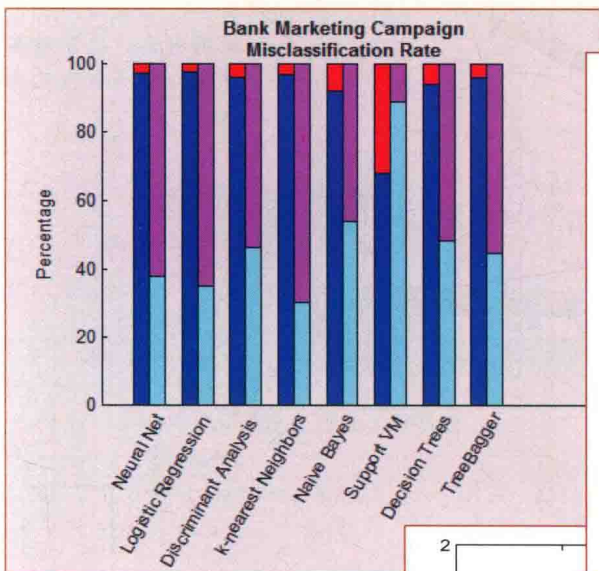
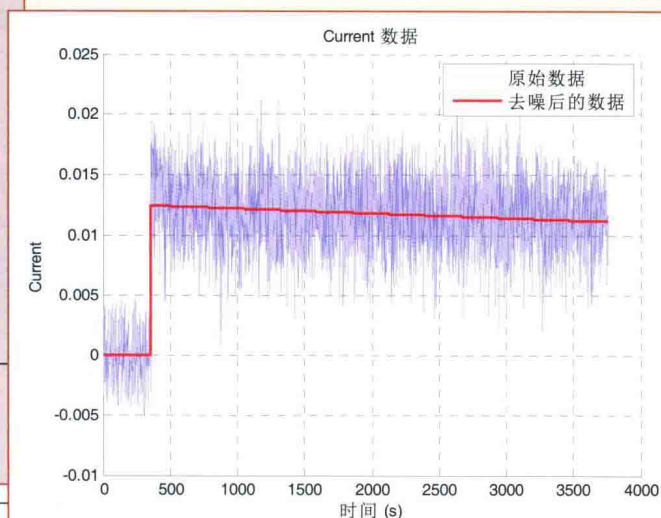
印 数: 3000 册 定价: 98.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888, 88258888。

质量投诉请发邮件至 zltts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

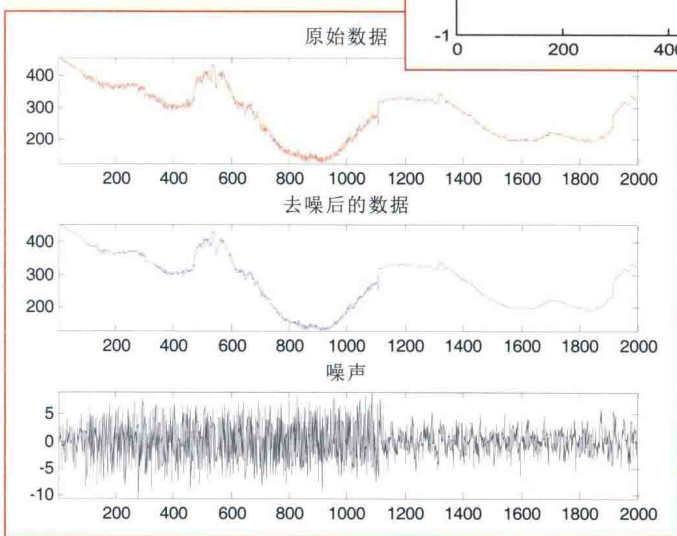
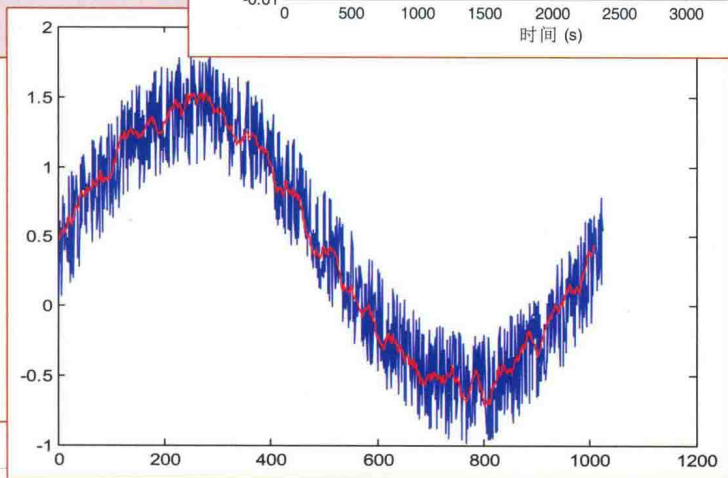
本书咨询联系方式: libing@phei.com.cn。

❖ 图 4-14 回归法去噪示意图

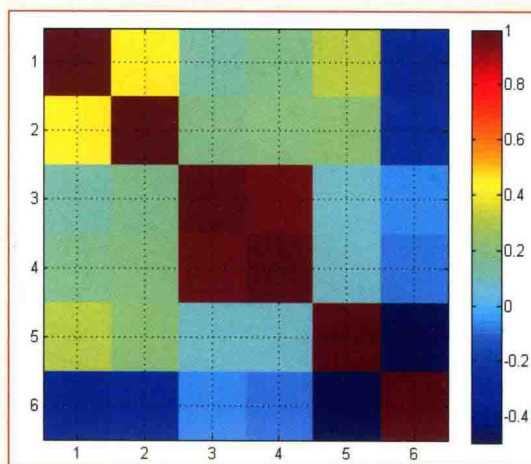


❖ 图 2-18 模型分类正确率柱状图

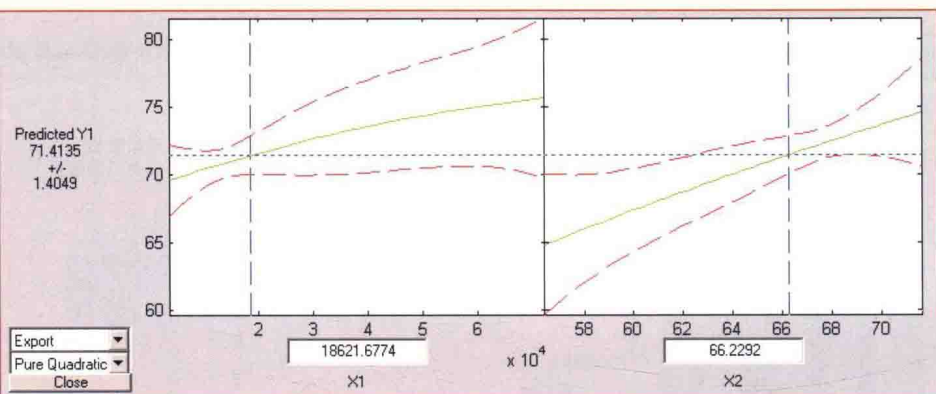
❖ 图 4-15 均值平滑法去噪示意图



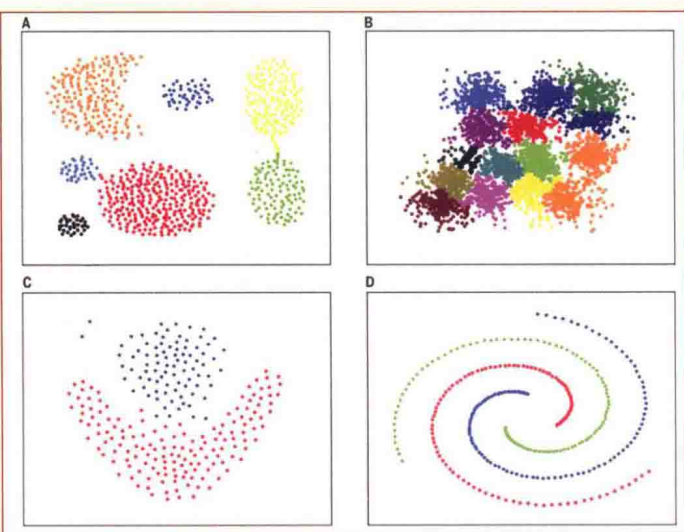
❖ 图 4-16 小波去噪效果示意图



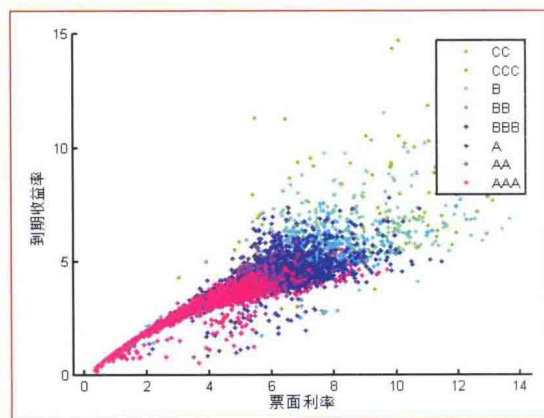
❖ 图 5-10 变量相关性强度图



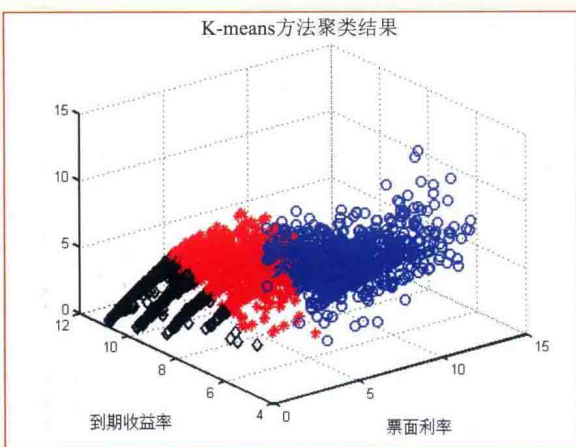
❖ 图 7-5 预期寿命与人均国内生产总值和体质得分的一个交互式画面



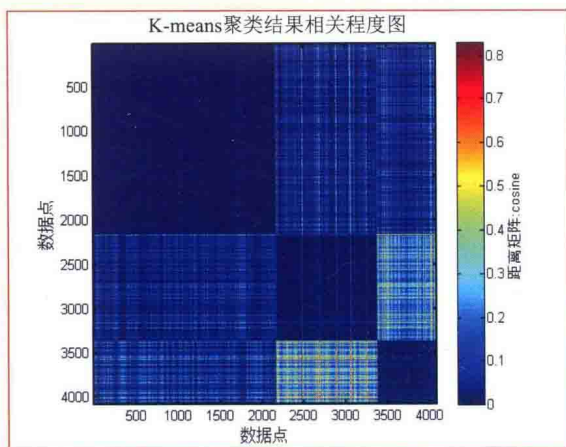
❖ 图 9-2 常见的类别特征



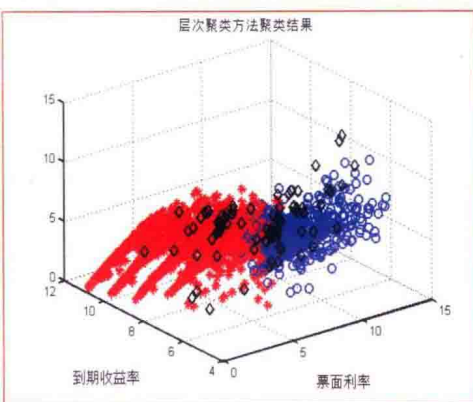
❖ 图 9-4 聚类效果图



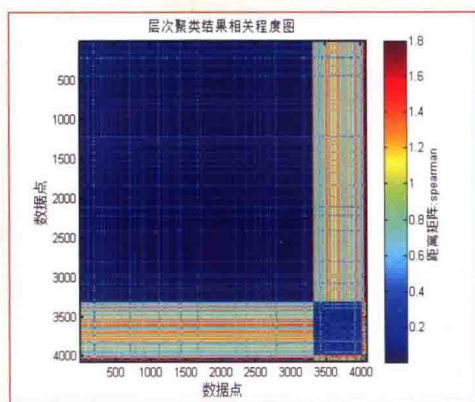
❖ 图 9-5 K-means 方法聚类效果图



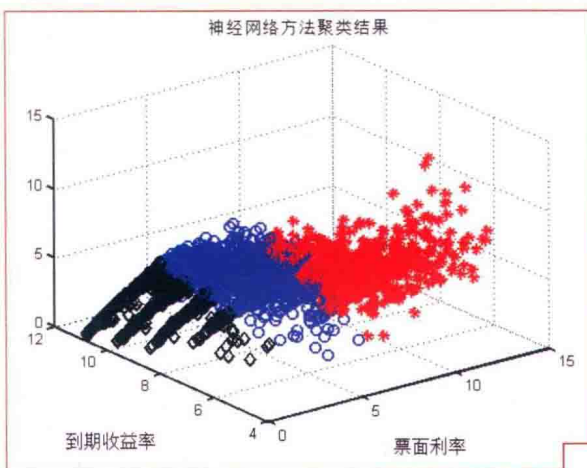
❖ 图 9-6 K-means 方法聚类结果簇间的相似度矩阵



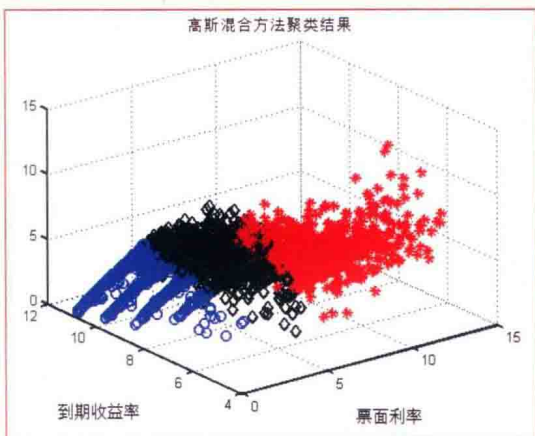
❖ 图 9-8 层次聚类方法产生的聚类效果图



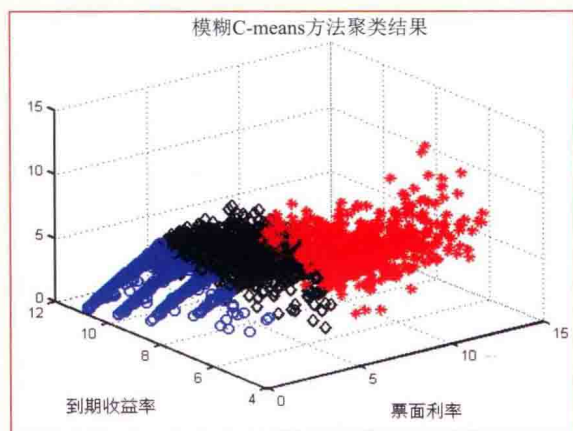
❖ 图 9-9 层次聚类方法产生的簇间相似程度图



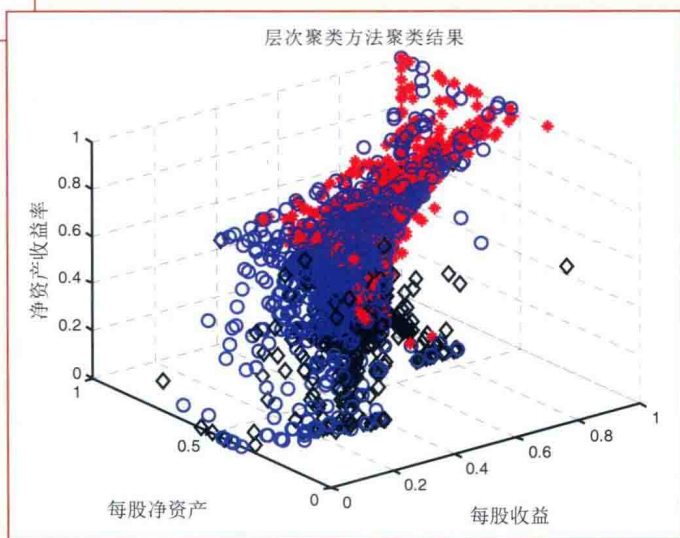
❖ 图 9-11 层次聚类方法产生的聚类效果图(一)



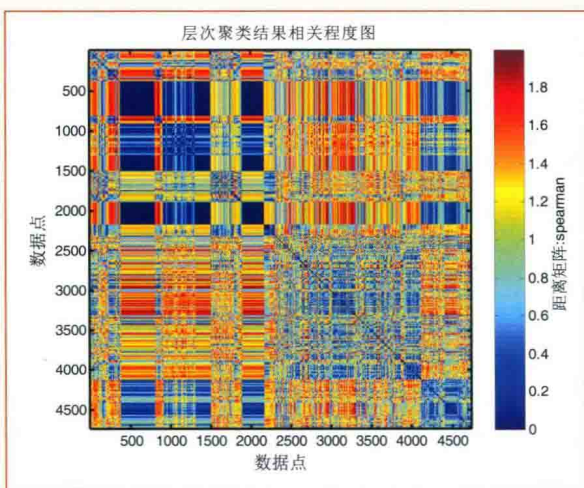
❖ 图 9-13 层次聚类方法产生的聚类效果图(三)



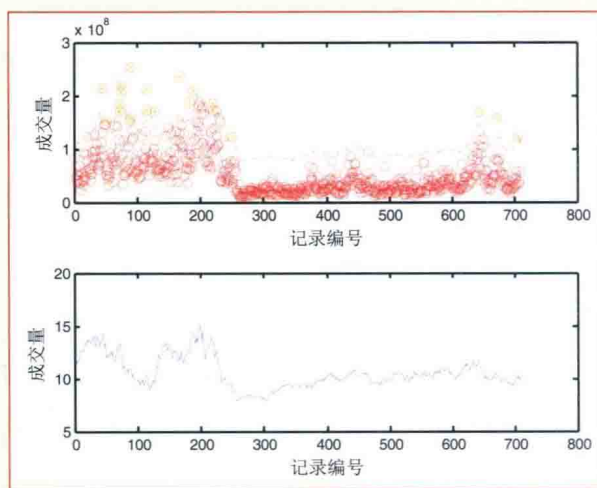
❖ 图 9-12 层次聚类方法产生的聚类效果图(二)



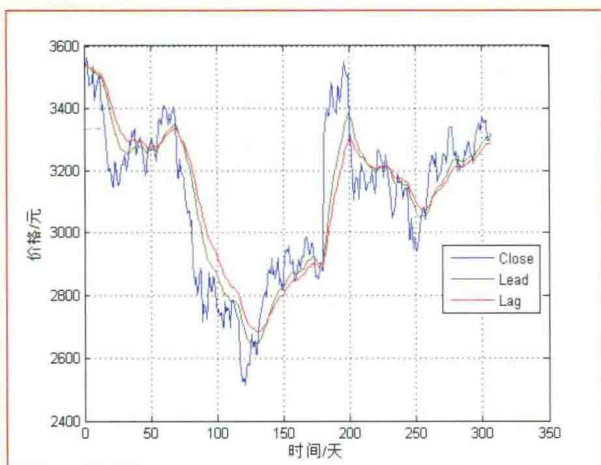
❖ 图 9-17 股票聚类结果效果图



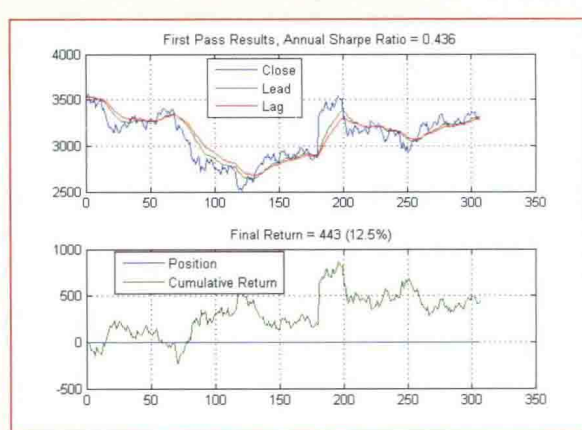
❖ 图 9-18 股票聚类结果簇间相似程度图



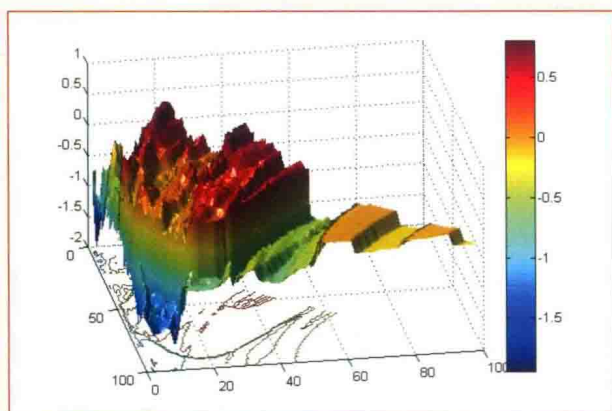
❖ 图 11-6 基于离群点诊断技术进行量化择时示意图



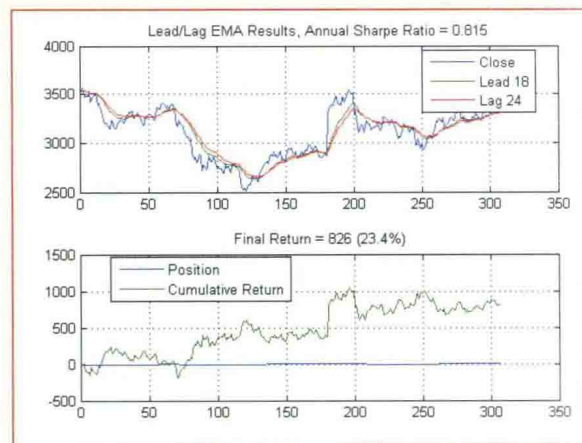
❖ 图 14-1 CSI300 收盘价的 20 日及 30 日移动均线



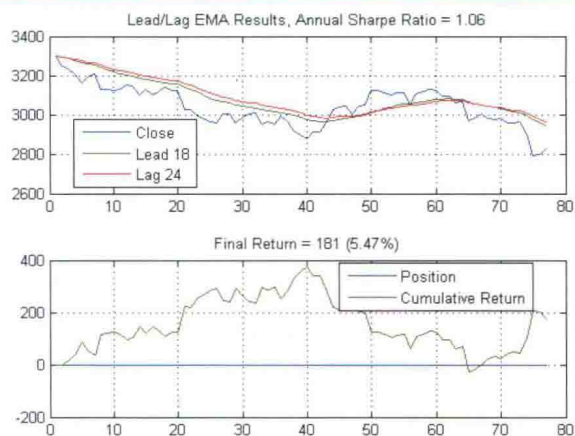
❖ 图 14-2 收益变化曲线



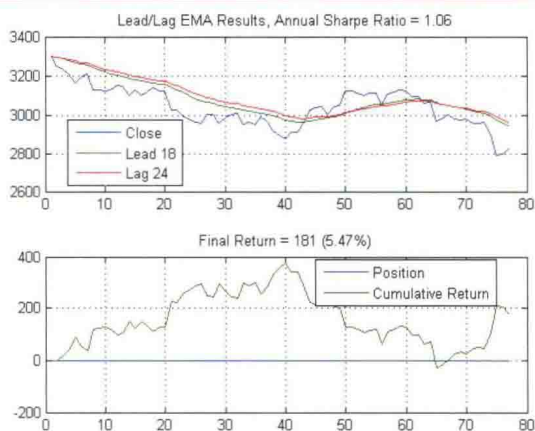
❖ 图 14-3 夏普率与参数 Lead 和 Lag 的空间分布图



❖ 图 14-4 最佳参数对应的投资收益图

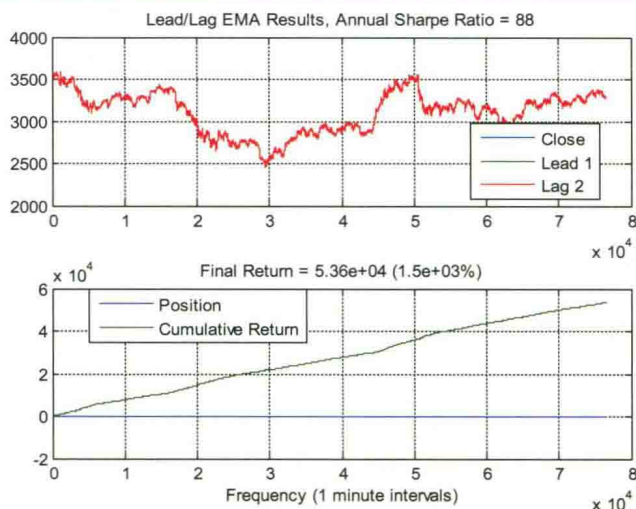


❖ 图 14-5 用新数据对策略进行验证得到的收益曲线



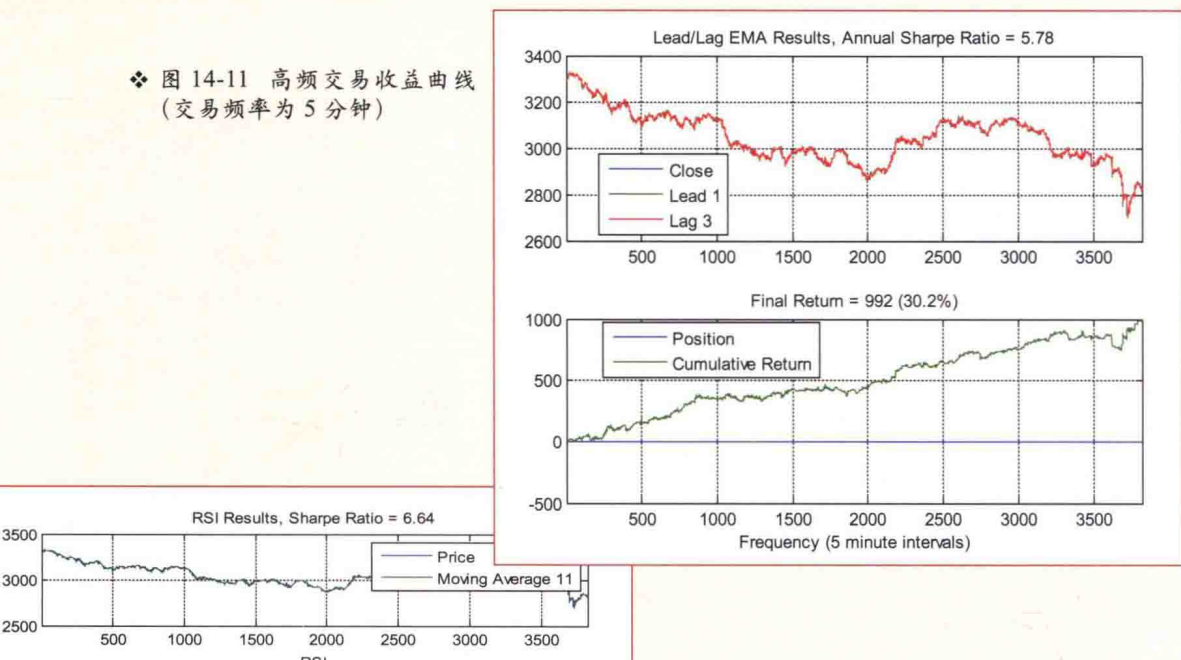
❖ 图 14-6 考虑交易费时夏普率与参数 Lead 和 Lag 空间分布图

❖ 图 14-7 用新数据对策略进行验证得到的收益曲线 (考虑交易费)

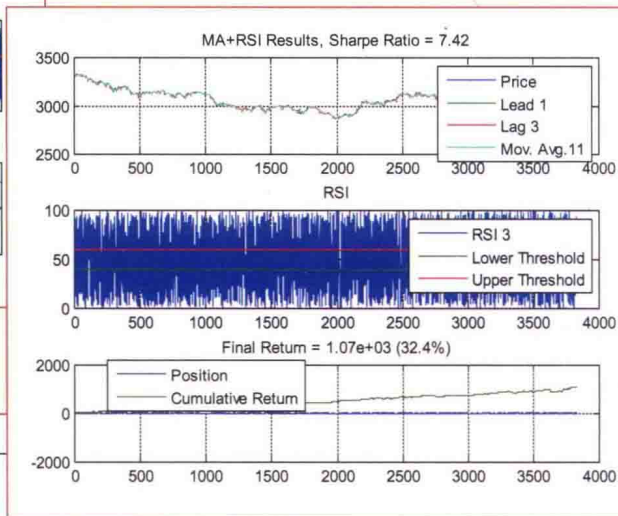
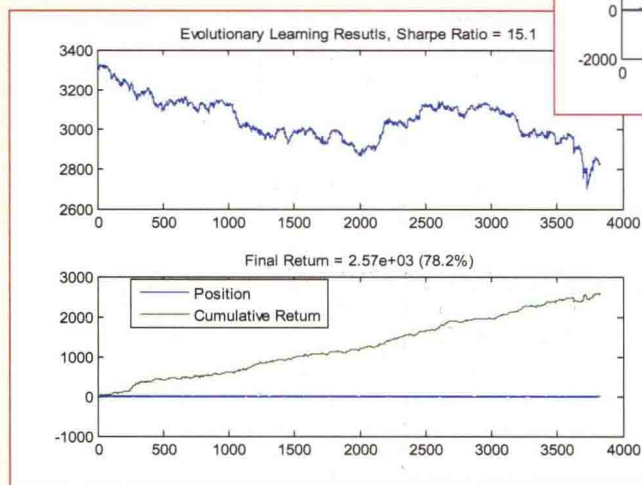


❖ 图 14-9 高频交易收益曲线 (交易频率为 1 分钟)

❖ 图 14-11 高频交易收益曲线
(交易频率为 5 分钟)



❖ 图 14-12 RSI 策略收益曲线



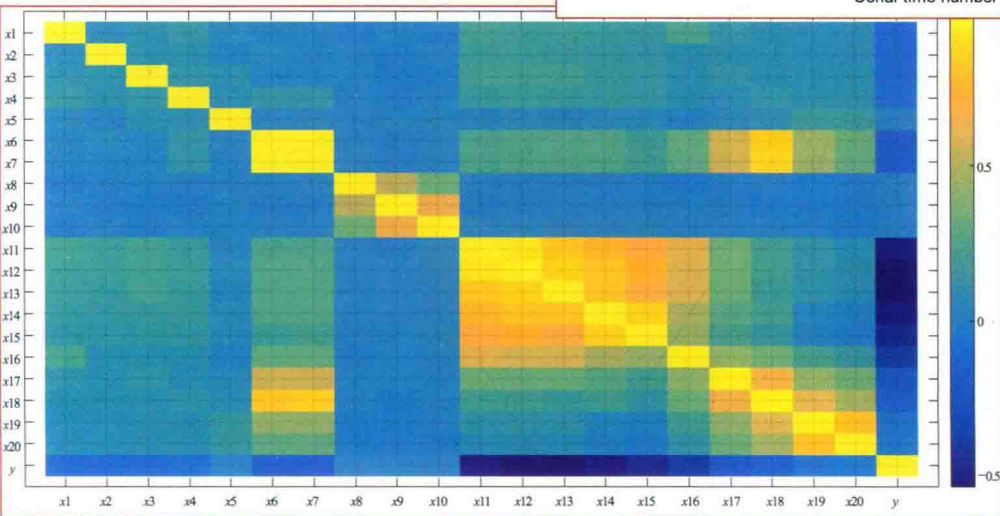
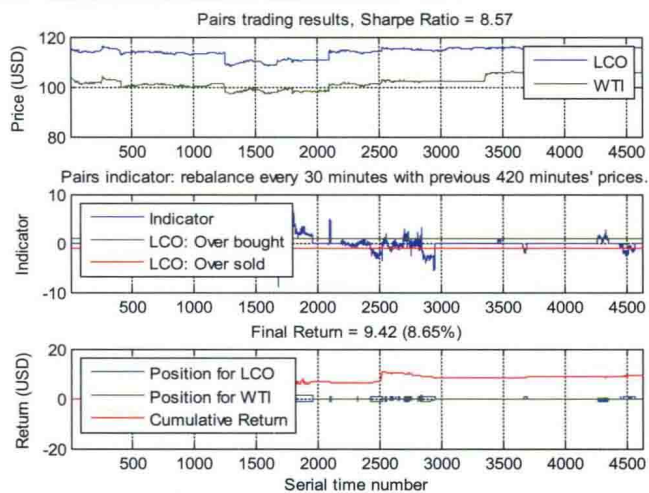
❖ 图 14-13 多信号策略收益曲线

❖ 图 14-18 优化后的多交易策略收益曲线

❖ 图 15-1 两期货日间数据趋势图

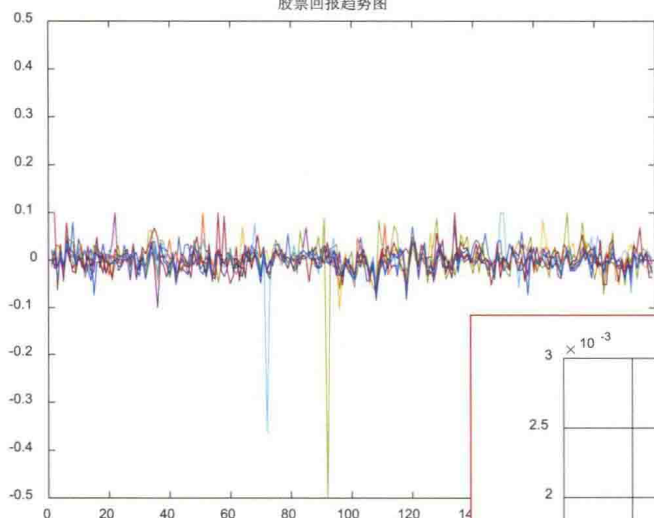


❖ 图 15-2 配对交易收益图

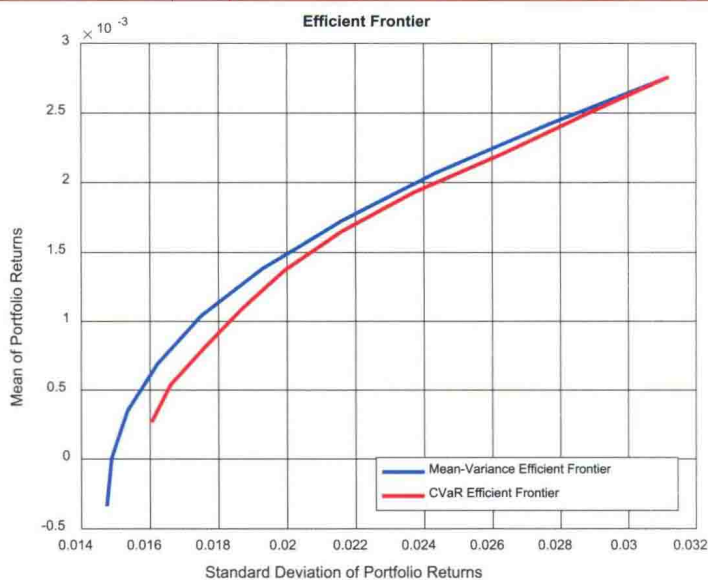


❖ 图 16-2 变量相关系数图

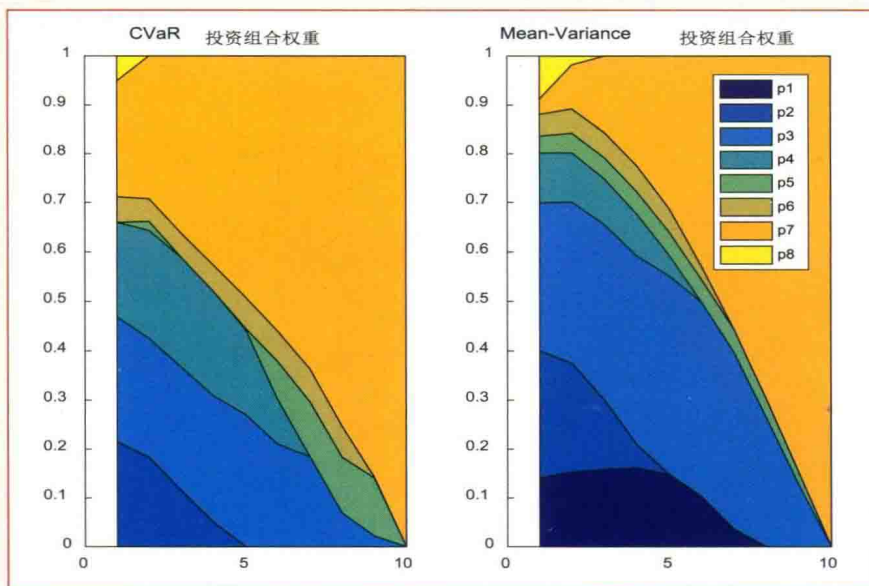
股票回报趋势图



❖ 图 16-4 要投资对象的收益曲线



❖ 图 16-5 组合投资的有效前沿曲线



❖ 图 16-6 投资权重分配图

除了你的才华，其他一切都不重要！

近年来，互联网和人工智能技术的飞速发展，推动传统金融大踏步前进，尤其是量化投资、互联网金融、移动计算等领域，用一日千里来形容亦不为过。2015 年年初，李克强总理在政府工作报告中提出制定“互联网+”行动计划，推动移动互联网、云计算、大数据等与各行业的融合发展。2015 年 9 月，国务院又印发了《促进大数据发展行动纲要》，提出“推动产业创新发展，培育数据应用新业态，积极推动大数据与其他行业的融合，大力培育互联网金融、数据服务、数据处理分析等新业态”。可见，大数据金融将会成为未来十年闪亮的领域之一。2012 年年初，中国量化投资学会联合中国工信出版集团电子工业出版社，共同策划出版了“量化投资与对冲基金丛书”，深受业内好评。在此基础上，我们再次重磅出击，整合业内顶尖人才，推出“大数据金融丛书”，引领时代前沿，助力行业发展。

本书特点

和卓金武认识是在一次行业会议上，我听到他演讲有关数据挖掘的课题，内容很详实精彩，会后与他交流后，发现他在量化投资领域做了很多年的研究，于是就建议他写一本有关基于数据挖掘的量化投资的教材，他欣然同意。一年后该书出版，受到业内很多好评。这次的改版更是对原先内容的一个大的升级和优化。

数据挖掘是人工智能领域的一个重要分支，在学术界发展了大概 30 年，最近在业内改头换面，以“大数据”的名词大放异彩。这个技术在过去十年各行各业都得到了广泛的应用，如语音识别、人脸识别等。在金融投资的领域，数据挖掘主要用于各种交易模式的识别、策略优化等，最近几年也得到了越来越多的研究人员的支持。这本《量化投资——MATLAB 数据挖掘技术与实践》，以详实的内容、深入的研讨，将数据挖掘技术用于量化投资的方方面面，可以对读者起着非常重要的借鉴作用。

本书的基础篇阐述了与数据挖掘有关的基础理论，包括数据挖掘原理、数据挖掘过程及数据挖掘的主要工具，特别是通过几个案例进行快速的 MATLAB 入门，让读者了解 MATLAB 中与数据挖掘有关的一些函数和工具箱的作用。

技术篇阐述了有关数据挖掘的各种技术，包括数据的准备、数据的探索、关联规则方法、数据回归、分类方法、聚类方法、预测方法、诊断方法、时间序列方法和智能优化方法等。其中，关联规则、分类方法、聚类方法是数据挖掘的最主要的内容。关联规则解决

不同因素之间的因果关系问题，试图从大量的数据中，找到看似不相关的因素之间背后有可能隐藏的逻辑关系。分类方法则对历史知识进行分类处理，试图找到对未来的预测。例如，可以将股票分为“涨”和“跌”两大类别，学习分类模型后，可用于对未来股票的走势进行预测。聚类是一种无监督的学习，也就是说，在没有历史样本的情况下，找到背后的大致规律。这特别适用于波动不规律的品种的投资策略。马尔科夫链是一种预测方法，可以用于大盘走势的预测。我们日常看到的 K 线都属于时间序列。时间序列的研究方法可以用于 K 线走势的预测，包括 ARMA 模型、ARIMA 模型、ARCH 模型、GARCH 模型等。

实践篇则介绍了多个基于数据挖掘的策略，包括统计套利策略、配对交易策略，以及有关数据挖掘系统的一些阐述等。统计套利用到的模式和分类技术；配对交易用到的协整关系和配对交易函数，可以用基于数据挖掘的技术来实现。书稿的最后两章分别阐述了基于 Quantrader 平台的量化投资和基于数据挖掘技术的量化交易系统，为投资者提供实际交易的辅助。

卓金武的这本书在理论方面具有很深的探讨，在实践方面又有多个案例，是一本不可多得的优秀教材，特此推荐。

美好前景

中国经济经过几十年的高速发展，各行各业基本上已经定型，能够让年轻人成长的空间越来越小。未来十年，大数据金融领域是少有的几个有着百倍、甚至千倍成长空间的行业，在传统的以人为主的分析逐步被数据和模型替代的过程中，从事数据处理、模型分析、交易实现、资产配置的核心人才（我们称之为宽客），将有广阔的舞台可以充分展示自己的才华。在这个领域中，将不再关心你的背景和资历，无论学历高低，无论有无经验，只要你勤奋、努力，脚踏实地地研究数据、研究模型、研究市场，实现财务自由并非是遥不可及的梦想。对于宽客来说，除了你的才华，其他一切都不重要！

丁鹏 博士

中国量化投资学会 理事长

《量化投资——策略与技术》作者

“大数据金融丛书”主编

2016.10 上海

前言

量化投资交易策略的业绩稳定，市场规模和份额不断扩大，得到越来越多投资者的认可。中国金融业飞速发展，尤其是 2010 年股指期货的推出，使得量化投资和对冲基金逐步进入国内投资者的视野。2012 年丁鹏博士所著的《量化投资——策略与技术》出版后，更是推动了量化投资技术在国内的普及。目前，量化投资、对冲基金已经成为中国资本市场热门的话题之一，各投资机构纷纷开始着手打造各自的量化投资精英团队。

量化投资是将投资理念及策略通过具体指标、参数的设计，融入到具体的模型中，用模型对市场进行不带任何情绪的跟踪，简单而言，就是用数量化的方法对股票、期货等投资对象进行估值，选取适合的对象进行投资。目前，量化投资的书籍主要集中在模型和策略及工具的使用上，但关于如何产生这些量化模型、量化策略的书籍的确非常少。金融市场瞬息万变，为了更好地进行量化投资，我们要不断去验证既有模型的有效性，同时要想在金融行业保持自己的竞争力，又必须不断开发新的模型，而验证模型、开发模型，所谓量化投资的主要内容，都需要数据的支撑。另一方面，金融领域是数据资源保存最好、最为丰富的行业，在金融领域已经积累了大量的数据，同时每天还在产生大量的交易数据、价格数据等数据信息。这些数据资源正好为量化投资提供了很好的数据基础，那么问题的关键就是如何利用金融业丰富的数据资源更好地进行量化投资。

数据挖掘技术是从数据中挖掘有用知识的一门系统性的技术，刚好解决了数据利用的问题，所以，数据挖掘与量化投资便很自然地结合在一起。但数据挖掘在国内也是一个新领域，所以，还没有关于量化投资与数据挖掘相结合的相关书籍。另外，目前关于数据挖掘的几本书基本都是译著，由于语言和文化的差异，国内读者读起来相对吃力。在这样的背景下，能有一本书介绍如何利用数据挖掘技术进行量化投资还是很好的。

巧合的是，笔者在 MathWorks 平时的工作职责之一是支持金融客户，相当比例是关于量化投资的，二是支持其他商业客户的数据挖掘，所以对这两个领域都有一定的了解。在一次研讨会上，丁鹏博士与笔者讨论了数据挖掘在量化投资中应用的话题，感觉这是个非常好的课题，建议笔者写一本这样的书。笔者对这个课题也非常感兴趣，于是就开始了这本书的创作。

《量化投资——数据挖掘技术与实践（MATLAB 版）》于 2015 年 6 月出版，一年以内已四次印刷，不少热心的读者发来 E-mail 与作者讨论书中的内容，并提出建议和不足。为了回馈读者，也为了让本书质量更好，于是在 2016 年年初就启动了新版的编写工作。新版主



要变动的地方有两个方面：一个是对上一版中发现的不足进行了修订，另一个是增加了 MATLAB 快速入门和基于 Quantrader 平台的量化投资两个章节。另外，有关程序化交易内容的章节也进行了大幅修改，主要是让程序直接与 Wind 数据对接。

本书内容

全书内容分为三篇。

第一篇为基础篇，主要介绍一些基本概念和知识，包括量化投资与数据挖掘的关系，以及数据挖掘的概念、实现过程、主要内容、主要工具等内容，并在第三章介绍了 MATLAB 快速入门。

第二篇为技术篇，是本书的主体，系统介绍了数据挖掘的相关技术及这些技术在量化投资中的应用实例。这部分又分为如下三个层次：

(1) 数据挖掘前期的一些技术，包括数据的准备（收集数据、数据质量分析、数据预处理等）和数据的探索（衍生变量、数据可视化、样本选择、数据降维等）。

(2) 数据挖掘的核心六大类方法，包括关联规则、回归、分类、聚类、预测和诊断。对于每类方法，则详细介绍了其包含的典型算法，包括基本思想、应用场景、算法步骤、MATLAB 实现程序和应用案例。同时，对每类方法还介绍了一个在量化投资中的应用案例，以强化这些方法在量化投资中的实用性。

(3) 数据挖掘中特殊的实用技术，包含两章内容，一是关于时序数据挖掘的时间序列技术，二是关于优化的智能优化方法。这个层次也是数据技术体系中不可或缺的技术。时序数据是数据挖掘中的一类特殊数据，并且金融数据往往都具有时序性，所以针对该类特殊的数据类型，又介绍了时间序列方法。另外，数据挖掘离不开优化，量化投资也离不开优化，所以又以一章智能优化方法来介绍两个比较常用的优化方法，遗传算法和模拟退火算法。

第三篇为实践篇，主要介绍数据挖掘技术在量化投资中的综合应用实例，包括统计套利策略的挖掘与优化、配对交易策略的挖掘与实现、基于 Wind 数据的程序化交易，基于 Quantrader 平台的量化投资，最后一章——基于数据挖掘技术的量化交易系统，则给出了集成主流数据挖掘技术的量化投资系统的框架，读者可以利用该框架，依据书中介绍的数据挖掘技术，结合自己的情况，开发出属于自己的量化交易系统，从而轻松实现从理论到实践的跨越，更好地利用数据挖掘技术在量化投资的领域乘风破浪，不断创造佳绩。

本书特色

综观全书，可发现本书的特点鲜明，主要表现在：

(1) 方法务实，学以致用。本书介绍的方法都是数据挖掘中的主流方法，都经过实践的检验，具有较强的实践性。对于每种方法，本书基本都给出了完整、详细的源代码，这对读者来说，具有非常大的参考价值，很多程序可供读者直接套用并加以学习，并可以直接转化为自己的量化投资实战工具。

(2) 知识系统，易于理解。本书的知识体系应该是当前数据挖掘书籍中最全、最完善的，不仅包含详细的数据挖掘流程、数据准备方法、数据探索方法，还包含六大类数据挖掘主体方法、时序数据挖掘方法、智能优化方法。正因为有完整的知识体系，读者读起来才有很好的完整感，从而更利于理解数据挖掘的知识体系，这对于读者学习本书内容非常有帮助。

(3) 结构合理，易于学习。在讲解方法时，由浅入深，循序渐进，让初学者知道入门的切入点，让专业人员又有值得借鉴的“干货”。基础篇、技术篇和实践篇的结构部署也让本书独树一帜，让读者在学习数据挖掘和量化投资的过程中有一个循序渐进的过程，使读者在短时间内成为一位数据挖掘高手，同时成为一位量化投资高手。

(4) 案例实用，易于借鉴。绝大多数实例都是量化投资领域的实例。所以，综观全书，本书都在有意引导读者思考如何让数据挖掘在量化投资中产生更实际的价值。

(5) 主线明晰，脉络分明。本书涉及的知识面宽广，以数据挖掘和量化投资为中心，辐射银行、债券、营销、零售等领域和学科。为了与书稿主线保持一致，这些所涵盖的领域虽然只是略微带过，但是从侧面折射出数据挖掘技术真的广泛服务于社会各个领域。在现代社会，某学科单打独斗的时代已经过去了，本书在无形之中已经树立了一个意识：各学科的知识之间是相通的，运用知识的最高境界是各学科知识的大融合。

(6) 理论与实践相得益彰。对于本书的每个方法，除理论的讲解，都配有一个典型的应用案例，读者可以通过案例加深对理论的理解，同时理论也让案例的应用更有信服力。技术的介绍都是以实现实例为目的，同时提供大量技术实现的源程序，方便读者学习，注重实践和应用，秉承笔者务实、切近读者的写作风格。

(7) 内容独特，趣味横生。很多方法和内容是同类书籍所没有的，这无疑增强了本书的新颖性和趣味性。

(8) 文字简介、明了，易于阅读。在本书编写过程中，在保证描述精准的前提下，摒弃那些刻板、索然无味的文字，让文字充满活力，更易于阅读。

读者对象

- 从事投资工作的专业人士，包括证券、基金、私募、信托、银行、保险等领域的从业者。

- 从事数据挖掘、数据分析、数据管理工作的专业人士。
- 金融、经济、管理、统计等专业的教师和学生。
- 从事量化投资或数据挖掘方向研究的科研工作者。
- 希望学习 MATLAB 的工程师或科研工作者。因为本书的代码都是用 MATLAB 编写的，所以，对于希望学习 MATLAB 的读者来说，本书也是一本很好的参考书。

致专业人士

对于从事量化投资的专业人士来说，书中的数据挖掘技术是值得借鉴的技术，至少会有助于挖掘或启发策略。书中的实例都具有一定的实战背景，含有一些从数据挖掘层面的策略，大家可以尝试将这些技术和策略融入自己的思想和策略中，以让自己的策略更强大。另外，对于书中介绍的各方法的理论，如果你有很好的数学或计算机背景，且有时间或感兴趣，可以认真看，否则，理论部分可以直接跳过。但是对于每种方法的思想和应用场景，读者一定要领悟，这样当遇到合适的场景后，读者就可以马上想到用哪种方法，然后直接借鉴书中的代码就可以轻松地将这些方法应用到自己的量化投资实践中。

对于从事数据挖掘的专业人士来说，大家可以关注整个数据挖掘知识体系和数据挖掘的流程，因为本书的数据挖掘知识体系是当前数据挖掘数据中体系比较全面和完善的。另外，数据挖掘流程也介绍得很详细，具有很强的操作性。此外，书中的算法案例和综合应用案例，也算是本书的特色，值得借鉴。

致教师

本书系统地介绍了数据挖掘设计的理论、方法和案例，可以作为金融、经济、管理、统计等学科的本科专业教材或研究生教材。相比一般的数据挖掘教材，量化投资更容易激发学生的学习兴趣，兴趣是最好的老师，这对开展教学是非常有利的。

书中的内容虽然系统，但也相对独立，教师可以根据课程的学时安排和专业方向的侧重，选择合适的内容进行课堂教学，其他内容则可以作为参考章节。授课部分，一般包含第一篇的 3 章和第二篇的前 8 章，总共 11 章内容，如果课时较多，则可以增加其他章节，包括后面案例的学习内容。

在进行课程备课的过程中，如果您需要书中的一些电子资料作为课件或授课支撑材料，可以直接给笔者发邮件（70263215@qq.com）说明您需要的材料和用途，笔者会根据具体情况，为您提供力所能及的帮助。