中国科学院研究生院教材
Textbooks of Graduate University of Chinese Academy of Sciences

数据挖掘技术及应用

刘世平　编著

Technology and Application of Data Mining

高等教育出版社
Higher Education Press

中国科学院研究生院教材
Textbooks of Graduate University of Chinese Academy of Sciences

数据挖掘技术及应用

Shuju Wajue Jishu ji Yingyong

刘世平 编著

Technology and Application of Data Mining

高等教育出版社
Higher Education Press

内容提要

本书从应用的角度介绍数据挖掘的概念、原理、算法和技术，并提供丰富的真实案例。本书由4个部分组成，主要包括：数据挖掘和商业决策、数据挖掘技术、数据挖掘应用、专题分析。在应用部分，每一章中都包括一个特定的商业智能问题。每一章都以业务目标为起点，把业务问题逐步转化为技术问题。读者在阅读本书内容后，能够较好地掌握如何正确地将数据挖掘方法应用于实际的项目中，高质量地解决问题。

本书可作为高等院校“数据挖掘”课程的研究生教学用书，也可供本科高年级学生及工程技术人员参考。

图书在版编目(CIP)数据

数据挖掘技术及应用／刘世平编著. —北京：高等教育出版社，2010.1

ISBN 978-7-04-025779-3

Ⅰ.数… Ⅱ.刘… Ⅲ.数据采集 Ⅳ.TP274

中国版本图书馆CIP数据核字（2009）第201390号

策划编辑	孙惠丽	责任编辑	康兆华	封面设计	杨立新
责任绘图	吴文信	版式设计	张　岚	责任校对	杨雪莲
责任印制	韩　刚				

出版发行	高等教育出版社	购书热线	010-58581118
社　　址	北京市西城区德外大街4号	咨询电话	400-810-0598
邮政编码	100120	网　　址	http://www.hep.edu.cn
总　　机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印　　刷	高等教育出版社印刷厂	畅想教育	http://www.widedu.com
开　　本	787×1092　1/16	版　　次	2010年1月第1版
印　　张	23	印　　次	2010年1月第1次印刷
字　　数	450 000	定　　价	36.90元

物料号　25779-00

总 序

在中国科学院研究生院和高等教育出版社的共同努力下，凝聚着中国科学院新老科学家、研究生导师们多年心血和汗水的中国科学院研究生院教材面世了。这套教材的出版，将为丰富我院研究生教育资源，提高研究生教育质量，培养更多高素质的科技人才起到积极的推动作用。

作为科技国家队，中国科学院肩负着面向国家战略需求，面向世界科学前沿，为国家作出基础性、战略性和前瞻性的重大科技创新贡献和培养高级科技人才的使命。中国科学院研究生教育是我国高等教育的重要组成部分，在新的历史时期，中国科学院研究生教育不仅要为我院知识创新工程提供人力资源保障，还担负着落实科教兴国战略和人才强国战略，为创新型国家建设培养一大批高素质人才的重要使命。

集成中国科学院的教学资源、科技资源和智力资源，中国科学院研究生院坚持教育与科研紧密结合的“两段式”培养模式，在突出科学教育和创新能力培养的同时，重视全面素质教育，倡导文理交融、理工结合，培养的研究生具有宽厚扎实的基础知识、敏锐的科学探索意识、活跃的创新思维和唯实、求真、协力、创新的良好素质。

研究生教材建设是研究生教育中重要的基础性工作。由一批活跃在科学前沿，同时又具有丰富教学经验的科学家编写的中

总序

国科学院研究生院教材，适合在校研究生学习使用，也可作为高校教师和专业研究人员的参考书。这套研究生教材内容力求科学性、系统性、基础性和前沿性的统一，使学习者不仅能获得比较系统的科学基础知识，也能体会蕴于其中的科学精神、科学思想、科学方法，为进入科学研究的学术殿堂奠定良好的基础；不但是体现教学内容和教学方法的知识载体、开展教学的基本条件和手段，也是深化教学改革、提高教育质量、促进科学教育与人文教育结合的重要保证。

“十年树木，百年树人”。我相信，经过若干年的努力，中国科学院研究生院一定能建设起多学科、多类型、多品种、多层次配套的研究生教材体系，为我国研究生教育百花园增添一支新的奇葩，为我国高级科技人才的培养作出新的贡献。

中国科学院常务副院长
中国科学院研究生院院长
中国科学院 院士

二〇〇六年二月二十八日

序　言

如今有许多企业和组织使用数据挖掘(data mining)技术从海量数据中提取信息,并利用这些信息做出关键的商业决策。对于主要关注于金融行业的数据挖掘和信息管理的商业智能(Business Intelligence)顾问来说,在过去的十几年里,当他们与来自不同部门的人员共事的时候,这些人经常会向他们提出与数据挖掘有关的各种各样的问题。在这些问题中,银行主管们经常提及的一个关键问题是:在现有的银行资源条件下,如何使用数据挖掘技术来做出更好的决策?这个问题的答案就在数据挖掘的定义中:数据挖掘是一个从海量数据中提取信息并利用这些信息做出关键的商业决策的过程。

为了有效地使用数据挖掘技术,相关技术人员除了要了解数据挖掘软件及其工作原理之外,还必须理解那些重要的业务问题以及确定能否使用数据挖掘的方法来解决这些问题。对业务问题和客户需求的较好理解将有助于在技术主管和商业用户之间形成很好的信息互动。与此同时,也必定会促使商业用户去了解那些用于解决业务问题的技术、算法和工具等。他们或许没有必要去了解如何做,但是至少可以知道什么是真正有用的。此外,无论从技术人员还是业务人员的角度来看,他们实际上能得到什么样的数据以及哪种类型的数据是必需的也是非常关键的因素。在如何应用数据挖掘的技术和方法来解决关键业务问题这件事上,在最终用户和技术人员之间存在着明显的差距。在现实生活中,最终用户通常知道哪些业务问题是必须得到解决的,哪些业务疑问是能够获得答复的。尽管他们相信数据挖掘是有帮助的,可是他们还没有完全意识到数据挖掘技术能被用来解决这些业务问题。另一方面,技术部门也许理解数据挖掘的工作原理或者知道如何使用数据挖掘工具,但是,在大部分时间里,技术主管们囿于业务问题之中并使用通常的行事风格来高效地解决问题。反过来,这又会导致如何使用算法和方法来解决业务难点和做出更好决策的问题。出现这种问题的部分原因在于最终用户和技术部门之间缺乏有效的沟通。但是,这种部门之间的交流也并不能完全解决根本性问题,原因在于懂技术的人员不熟悉业务,熟悉业务的人员不了解技术。

本书致力于在最终用户和技术人员之间架起一座桥梁。因此,本书涵盖了数据挖掘技术、数据挖掘在金融行业的应用以及数据挖掘工具等内容。本书由4个部分组成,它们是:数据挖掘和商业决策、数据挖掘技术、数据挖掘应用、专题分析。

第1部分包括一章的内容,其中简要地描述了关键的数据挖掘技术及其在金融机构中的潜在应用。根据每种技术的方法讨论了各类数据挖掘技术的优、缺点。

此外,还专门探讨了每种方法在金融行业的潜在应用。数据挖掘的过程和步骤也放在本章来讲述。总的来说,本章内容主要集中在数据挖掘的方法及其潜在的应用的描述上。因此,建议所有读者都要通读这部分。

第2部分描述了主要的数据挖掘技术。该部分提供关于数据挖掘技术的较为详细的描述。对算法比较感兴趣并想了解数学、统计学和神经网络技术的读者通过学习该部分内容能够更好地了解各类算法。在掌握这些基本的技术的基础上,数据挖掘人员能够研发新的算法和发现算法新的应用。这部分由3章构成。第2章讨论分类和聚类(clustering)技术,主要是神经网络聚类、统计学分类和传统的统计方法等。第3章是关于预测模型和分类技术的讨论。第3章讨论的预测方法有:逻辑回归、正态回归、决策树(decision tree)、神经网络(neural network)分类、径向基函数(radial basis function, RBF)、多元回归等。关于各种方法的优、缺点和潜在的应用也放在这部分来论述。最后,第4章讨论了链接分析,其中包括关联分析和序列模式分析以及相关的算法和函数。通过这3章帮助读者理解数据挖掘技术并更加有效地解释数据挖掘的结果,以便开发新的数据挖掘方法和发现数据挖掘技术的新应用。强烈建议数据挖掘人员通读这3章。此外,对于能力较强的数据挖掘用户也应该阅读这一部分。对于只想了解如何进行数据挖掘的商业用户而言,该部分是一个很好的参考资料,它能够帮助用户了解如何使用数据挖掘工具得到自己需要的结果。因此,理解这部分内容将大大改善用户和技术管理人员之间的沟通和交流。

第3部分是数据挖掘的应用部分。这部分由9章组成,主要讲述数据挖掘的应用。换句话说,就是集中讨论如何使用数据挖掘技术解决实际的商业问题。该部分的各章是相对独立的。每一章重在描述使用一种典型的算法解决某一类商业问题,并聚焦于与金融行业商业智能有关的特殊问题。各章都包括对商业问题和技术问题的论述,且从商业问题开始,然后再把它转换为技术问题。不过用来解决这些技术问题的方法和算法在此将只进行简要的讲述,详细的内容可参考第2部分的内容。对于商业问题的解决方案都相应地给出技术方面的结论和解释。最后从技术结论和商业解决方案的角度给出能够改进商业效益的可行的建议。特别地,该部分的组织形式非常适宜各章独立阅读,这种组织形式主要是从金融行业一些主要应用的角度考虑的。因此,读者可以根据商业问题分别找到相应的章节。换言之,它可以作为每章对特定问题做出指导的一个参考手册。本部分的目标就是使大家读完每章后能够把数据挖掘作为一个工具来解决实际的问题。通过这部分内容的阅读,相信读者应该具备解决金融领域大多数问题所必需的知识,并且这些问题是可以用数据挖掘工具来解决的。

第5章主要集中于细分,它可能是客户细分或市场细分。首先讨论直观的和分析的聚类方法。接下来讨论聚类的过程。换言之,就是通常在聚类分析中包含哪些步骤或过程,特别是分析性的聚类。然后讨论聚类在金融行业中的应用。在

本章的结尾，给出一个聚类分析的实例，它涵盖了业务问题、系统结构、聚类过程及结果的解释和展现。首先讨论聚类的原因在于聚类分析的广泛使用以及可以用它来进行数值的预测或建立分类模型，例如，客户赢利能力分析和目标建模。

在讨论了被广泛使用的聚类方法之后，接下来讨论银行最重要的数据挖掘应用：预筛选和目标建模。预筛选和目标建模都用于客户招徕，它可以划分为分类模型。一般来说，诸如逻辑回归、正态回归、神经网络和决策树都可用来建立分类模型。在第6章中将对与每种方法相关的算法进行简要的讨论，因为有关算法的细节已经在第3章中讨论过了。然后，讨论预筛选和目标建模的每一种方法的优点和不足。在理解了业务问题和技术方法以后，给出建立一个较好的目标建模的过程。为了解释如何使用这些方法和过程来建立一个模型，在本章结尾给出了一个建立目标模型的实例。

有了预筛选和目标模型以后，下一步自然而然就是进行市场营销活动了。在收到可能的客户的申请之后，需要有一个承销的过程，它将决定某项申请是被接受还是拒绝，对于那些可能的客户应该给出多大的信用额度，对于一个特定的客户银行将以什么样的利息率进行收费。第7章的承销模型能够解答这些问题。首先，讨论什么是承销模型以及承销模型的作用有哪些。接下来，探讨建立承销模型的过程以及算法和功能。最后，通过一个承销模型的实例来解释前面3个小节所描述的建模方法和过程。

一旦一个预期的客户成为真正的客户，需要对客户给予不断的、始终如一的关注和管理。有几个不同的模型可以用来对现有客户进行管理。一个最重要的模型是与任何类型的贷款都有关的不良行为和破产模型。消费者的破产有可能导致银行高达数十亿元的损失。在第8章中将讨论关于这类模型的几个问题，具体包括什么是不良行为和破产模型、建立不良行为和破产模型的方法和过程以及一个具体的实例。

除了不良贷款和贷款损失以外，欺诈是造成银行收益下降的另一个重要的方面。为了防止欺诈所造成的损失，最重要的一个步骤是识别潜在的欺诈。银行可能面临各种各样的潜在的欺诈。数据挖掘能够帮助金融机构有效地识别这些欺诈。在第9章中将介绍欺诈的类型以及如何建立一个有效的欺诈侦测模型，最后是一个关于欺诈侦测模型的实例。

在讨论了不良行为和破产模型之后，在第10章将要讨论的客户流失也是影响金融业纯收益的另一个关键因素。一般来说，保留一个现有的客户要比花费成本去招徕一个新的客户更加有利可图。而且，从一个新近招徕的客户身上实现赢利需要几年的时间。例如，大多数信用卡公司要三年左右才开始从他们的新卡成员身上赚到钱。因此，防止现有客户的流失是提高银行纯收入额的最为有效的方法之一。在这一章将讨论与流失有关的几个关键问题：为什么流失模型对银行来说很关键；如何建立流失模型；为什么赢利能力分析对于客户流失分析也是重要的；

如何设计有效的策略来预防与阻止客户流失;在本章最后将给出一个建立流失模型的实例。

管理现有的客户对银行来说非常具有挑战性。防止那些有利可图的客户离开银行和预测哪些客户可能会流失是非常重要的。遗憾的是,某些客户贷款成为坏账,并因此给银行造成损失。防止因这种情况给银行造成损失的一个方法是向那些不良贷款和破产账户收回损失。如何使用有限的资源来最大化收回贷款的数量,同时与那些欠贷未还的客户以及已经流失的客户保持良好的关系,这对银行的回收贷款的部门来说并不是一件容易的事。数据挖掘和商业智能能够在托收过程中起到非常重要的作用。在第11章,首先讨论贷款托收过程,然后将讨论贷款回收的主要挑战在于何时给欠贷的客户发送消息以及给这些客户发送什么样的消息。在此,数据挖掘能够起到关键的作用,它可以用来对客户行为和贷款可收回的程度进行评分。根据贷款的可回收性对客户进行由易到难的分级排列。但是,仅有评级还是不够的。必须事先准备一两种有效的方法,并且需要优化模型。特别地,在本章还将介绍水平营销的思想,并以此来设计一套收款策略,同时根据客户的背景资料和行为属性形成一系列的策略。首先,确定根据每个客户的背景资料和行为属性形成的最好策略是什么。换言之,最好的收款策略是根据每个客户的全部可获得的信息所形成的。如果第一种策略不大奏效,那就需要使用第二种策略,这将取决于第一种策略不能奏效的原因。有必要的话,还需要使用第3或第4种策略。这种方法的优点在于对每一个客户都使用最好的策略,以至于第一轮回收额将会达到最大。此外,即使前面的策略不能奏效,它们也将为后面的策略奠定基础,这就避免了任何随机的收款策略,而且能够给出一个对称的、一致的方法。在描述了水平营销的策略之后,接下来将讨论一个使用收回贷款方法的过程。

有了前面所描述的许多必需的部分以后,一个最重要的步骤是进行赢利能力分析。赢利能力分析包含两个方面的内容。一方面是分析现有客户的赢利能力,另一方面是预测未来的赢利能力。预测未来的赢利能力需要客户的历史数据,例如,利润级别、人口统计学属性和行为变量。赢利能力分析的结果可以用来对期望的客户进行优先次序的排列,即根据客户潜在的赢利能力排序。有许多方法可以用来完成赢利能力分析。一个是使用数据挖掘的方法来帮助预测未来的赢利能力,并根据客户的背景资料对客户进行细分。赢利能力分析的结果也能够用来分配市场营销活动的预算,即为了更加有效地招徕所期望的客户而进行的营销活动预算。赢利能力分析的结果也能够用于客户关系管理之中,例如,渠道设计和制定客户保持策略等。第12章将主要描述客户的赢利能力分析。

除了招徕一个新客户到银行外,能够增强银行与客户的关系以及提高赢利能力的一个重要方法是交叉销售。这是第13章将要讨论的内容,也是目前银行少有的几个热点问题之一。对银行来说,这一主题之所以较热的原因有许多。首先,它能增强或深化银行与其客户之间的关系,因而能降低客户流失率。其次,把另一种

产品销售给现有的客户的成本将比为了特定的产品和服务去招徕新的客户的成本低得多。最后,向客户销售新的产品和服务将需要很低的成本。因此,这又将提高银行的赢利能力。交叉销售能够提高组织的赢利能力和降低客户流失率。建立交叉销售模型的技术和方法也将在本章进行概括性的介绍。如何把这些方法用于解决实际的商业问题将在后面的章节中详细讨论。在本章的结尾用一个实例说明如何建立交叉销售的模型及如何使用这一模型。

第4部分是专题分析,共包括7章。第14章和第15章分别介绍了分销网络决策和银行金融产品的价格决策。在分销网络决策中,将描述如何为每种产品和服务选择正确的渠道。随后,重点讨论如何进行渠道的选择,并集中说明如何使用空间分析的技术,以及如何把空间分析和数据挖掘结合起来进行渠道的选择和定位。在第15章将讨论银行的价格决策,即如何使用数据挖掘的方法进行价格决策,从而制定出比较合理的价格。

除了数值数据以外,今天人们收集的许多数据以文本的形式存在。如何把这些文本数据转化为信息,可以使用文本挖掘(text mining)的方法来实现。因为许多文本中包含了关于客户行为方面的有用的信息,因此,进行文本挖掘显得非常重要。例如,根据信用卡交易数据能够知道客户在哪里消费(商场名称)、购买了哪些产品(运动、时装和玩具)。这种富于洞察力的信息对于客户招徕和客户保持来说非常有用。第16章将讨论如何把文本转化为信息以及如何使用文本中的信息制定商业决策。在本章的结尾将给出一个实例来说明文本挖掘的过程。

众所周知,目前客户关系管理(customer relationship management, CRM)是一个非常热点的问题。对于不同的人员、机构和商业环境来说,它具有不同的含义。但是,关键是从不同的角度来观察客户并得到有关客户的整体描述。客户看起来是什么人、他们需要什么以及他们的行为是什么等,在对这些问题有一个比较全面的理解之后,金融组织就应该尽力在正确的时间通过正确的渠道把正确的产品交付给正确的客户。主要是注意这里的4个“正确”。正确地做的关键是信息可以从如下几方面得到,但并不局限于这些方法,它们是客户背景资料分析、客户行为分析、客户赢利能力分析和客户风险分析等。可以根据客户细分、预筛选模型、目标模型、流失模型、交叉销售模型、收益率模型以及其他模型等进行分析。问题是如何把它们集成在一起。第17章将重点讨论如何把客户和市场信息集成在一起,以实现在正确的时间通过正确的渠道把正确的产品提供给正确的客户,因此,本章看起来更像总结性的一章,通过它把大多数章节集成起来。第18章给出了一个关于上市公司财务预警的具体的应用实例,以供读者在解决实际面临的业务问题时作参考。

本书的最后两章,即第19章讨论了与数据挖掘技术密切相关的数据可视化技术;第20章将集中于软件包的简单介绍。现在,有许多机构提供数据挖掘服务和数据挖掘软件。第20章将从功能、易用性、灵活性、可扩展性、兼容性以及其他的

角度对数据挖掘软件进行简要论述。

本书是作者根据在北美洲和亚太地区的许多金融机构做数据挖掘顾问时总结的经验撰写而成的。本书的初衷是给金融组织中的各个部门的商务主管们写的，目的是让他们通过有效地使用数据来快速而又准确地制定商业决策。在市场和风险管理部门供职的业务经理们可以把本书作为一本参考手册，以帮助他们使用数据挖掘工具来解决特定的业务问题。本书的目的也在于帮助风险分析人员和市场营销分析人员使用数据挖掘的方法来解决他们在日常的业务运作过程中所面临的问题。技术分析人员也可以把本书作为一本参考书，用以指导他们的数据挖掘过程。此外，它也适合于那些想要了解他们应该收集什么数据以及如何来组织这些数据的人员。本书还为那些业务分析人员或者是经理以及技术分析人员或者是IT经理提供了大量的实例，以便他们能够遵循这些特定的步骤来解决所面临的问题。而且，无论业务经理还是IT经理都能够使用本书来指导他们之间的沟通并用这些技术方法来共同解决业务问题。此外，本书的目的还在于帮助技术人员和业务人员对数据挖掘的结果进行恰当的解释，以便决策人员能够根据这些结果来针对组织所面临的问题制定关键的商业决策。除了组织中的业务和技术人员以外，本书也适合那些打算学习数据挖掘或选修“数据挖掘”课程的学生作为教学参考书使用。尽管本书是根据作者在金融机构的工作经验写成的，但是其他机构的相关工作人员也能够使用，特别是那些来自不同组织的人员都可以使用它，例如，政府机关、电信部门、高科技公司、零售商店以及那些在商业决策过程中有兴趣使用数据挖掘技术的人员等。原因在于数据挖掘的方法对于不同的行业是相通的，而数据挖掘的过程却因不同的行业而异。因此，本书中所描述的数据挖掘方法和过程(根据具体问题进行选取)完全可以并且能够应用到其他行业中去。

致　谢

当本书呈现在读者面前时，需要感谢的人实在有很多。在本书的写作过程中，得到了诸多师长亲朋的热情指点与帮助，谨在此向所有帮助、支持过本书写作的朋友、同事致以由衷的谢意，感谢大家这么多年来对我的帮助和关怀。如果没有你们的帮助，这本书的出版将是不可能的。

首先要感谢陪伴我走过13年数据挖掘职业生涯的商业伙伴，是他们的项目给了我很多实践的机会，也正是那些前沿性的项目让我能够把数据挖掘理论应用到解决金融行业的实际问题中来。这些尖端应用带来的挑战给了我研究数据挖掘理论的动力和解决实际问题的能力。感谢那些曾经和我一起奋斗到深夜的朋友们，他们的睿智、大度和专业能力无疑增强了我的工作热情和乐趣，让我至今仍然热爱着数据挖掘这个领域，并志在将其作为我终生的事业。让我难忘的合作伙伴包括：北美洲的 Bank of America、Wells Fargo、Bank One、First USA、Discover Card、Equifax、Experian、Novel Scotia Bank 等；亚洲的 Development Bank of Singapore、Overseas United Bank、United Overseas Bank、Thai Farmers Bank 等。这些金融机构中至今还有很多朋友一直与我保持着联系和沟通，我要衷心地说出来：谢谢你们，我的朋友！

同样，在国内的很多客户也给我提供了大力的帮助和支持。在数据挖掘方面，给我留下非常深刻印象的包括上海证券交易所、新浪网、兴业银行、国家开发银行、中国人民银行和国家广播电视总局。让我感到最为欣慰的是在结束这些项目之后，几乎在我合作过的每个客户那里，都结交了许多非常好的真心朋友。当然，在做项目的过程中，他们的支持、理解与友情更是让我深受感动。正是这些合作伙伴让我认识到了自己工作的价值，让我体会到了友情的重要和可贵，让我感受到了知识的意义和力量。在此特别要感谢的有朱丛玖、周勤业博士、白硕博士、赵小平博士、郑刚博士、董国群、黄宏斌、石晓成博士、陈中苏博士、鲍康虎、吴宏英、吴尚荣、谢为国、侯维栋、白新民、郑梅迪、王魏、董燕、梁名高、皮六一、周国庆、林永峰博士等很多朋友。

其次，我要感谢的是在工作期间给予我关怀和支持的领导和同事，包括 Ed Murphy、Erik Kraglas、Michael Rotheman、Harry Mathis、John Rollins、Thomas Lee、Jim Kraemer、Yuhui Yao 和 Jenny Yap，是他们把我引领到了数据挖掘领域，同时给了我很多的指点和帮助。他们中的很多人同时也都是和我一起战斗在客户第一线的“战友”。其中，有两位人物对我的影响尤为深刻。记得当年为了解决美国最大的征信管理公司 Equifax 的一个难题时，当时已经年近花甲的 Harry Mathis 博士常常

和我一起加班到凌晨三、四点。此外，为了新加坡发展银行（Development Bank of Singapore）的一个项目，我当时的领导，时年已经52岁的Ed Murphy和我一起工作到凌晨四点，然后让我“早点”回家休息，以便能做好第二天早上的成果汇报。在陪我出来的路上，他去接了一把自来水往自己的脸上一擦，然后说：“我清醒啦，可以完成汇报材料的最后一段。”当五点多他离开办公室时，已经没有出租车，没有地铁，他是六点多在地铁站的椅子上被人叫醒的，也许当时是被当做流浪人叫醒的；但是当八点半我们开始向客户汇报工作时，他依然精神抖擞，思路敏捷，神采飞扬地向客户报告工作的进展情况和阶段性成果。现在无论Mathis博士还是Murphy先生都已经退休，离开了他们曾经为之付出青春的岗位，做出了他们对于这个行业应有的贡献。但是他们的这种敬业精神却从未消逝，一如既往地激励着我去做我喜欢的事业和完成这本书的写作：做有意义和有价值的事情。

2002年回国之后，很多朋友的关怀和支持更是让我深受感动。在此特别要感谢的是中国科学院研究生院的王颖和潘辛平教授，是他们让我有了和学生交流的机会，让我在教学中不断地总结和提升。他们的支持、肯定和帮助是这本书逐渐成形的原动力。

除了领导的支持之外，更要特别感谢的还有在吉贝克公司工作的同事，他们的支持和帮助也给了我无穷的动力，包括田凤、李华明、张志旺、申爱华、陈彬和陈苗军等。他们参与的部分项目都已经成为本书中的案例的原型，其中有些人员同时参与了书稿的整理，特别是张志旺和陈彬。对于他们所付出的努力和贡献，在此郑重地说声：谢谢你们，我的好同事！

还需要感谢的是高等教育出版社的相关编辑，他们的专业能力和敬业精神令人钦佩。感谢他们对于书稿超时的耐心和等待，他们的细致工作使本书更臻完善。

最后，最不能忘记的是我的家人。感谢父母多年的养育之恩和无私的爱，感谢妻子陈怀清的支持以及孩子刘恺伦和刘恺儒的理解。虽然孩子年幼，很需要我的关怀和陪伴，但是他们幼小的心灵似乎已经理解了“爸爸忙”的含义。在此我要深深地说声：谢谢你们，我的家人！

目　录

第 1 部分　数据挖掘和商业决策

第 2 部分　数据挖掘技术