国外经典教材

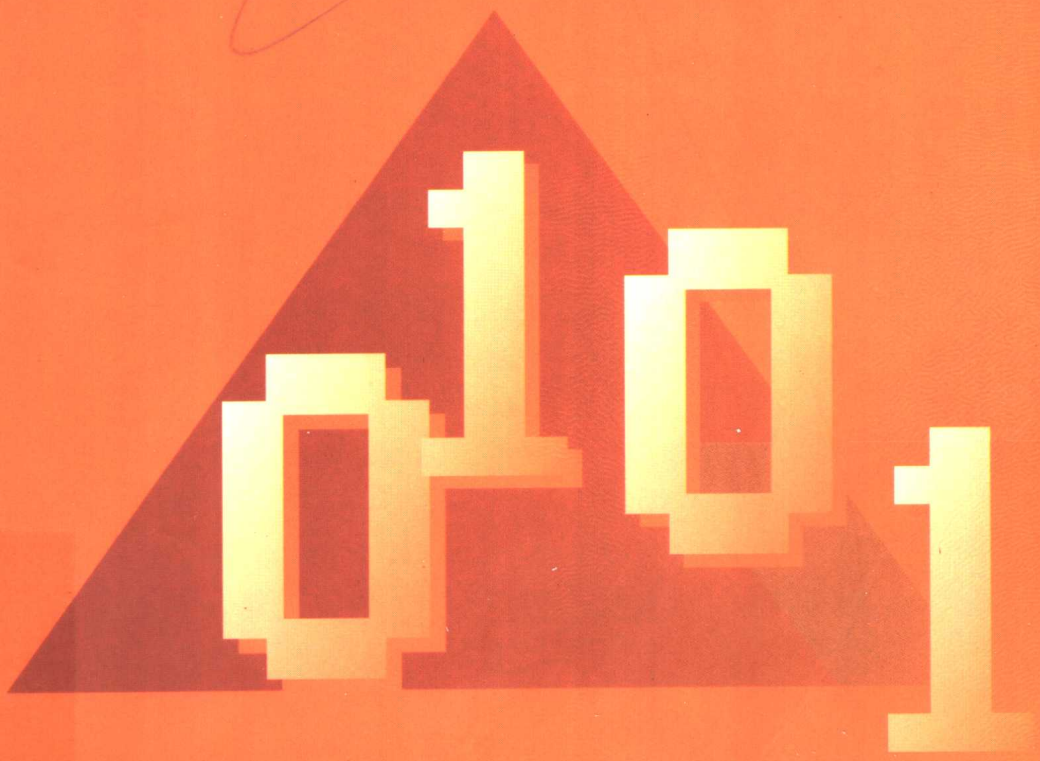
PEARSON Addison Wesley

数据挖掘教程

Data Mining

A Tutorial-Based Primer

(美) Richard J. Roiger 著
Michael W. Geatz
翁敬农 译



Pearson Education

清华大学出版社

国外经典教材

数 据 挖 掘 教 程

（美） Richard J. Roiger 著
Michael W. Geatz
翁敬农 译

清华大学出版社
北 京

内 容 简 介

本书为数据挖掘的基础教程，是作者多年来从事数据挖掘和专家系统课程教学经验的总结。它从商业角度介绍了数据挖掘的原理以及从数据中提取隐含模式的技术。本书首先帮助读者建立起数据挖掘的概念，进而通过 13 个数据挖掘示例帮助读者掌握数据挖掘的原理。本书的最后部分还介绍了结合专家系统和智能代理解决复杂问题的方法。

图书在版编目（CIP）数据

数据挖掘教程 /(美) 罗杰（Roiger , R. J.），（美）吉茨（Geatz , M. W.）著；翁敬农译. —北京：清华大学出版社，2003

(国外经典教材)

书名原文：Data Mining A Tutorial-Based Primer

ISBN 7-302-07456-9

Ⅰ. 数… Ⅱ. ①罗… ②吉… ③翁… Ⅲ. 数据采集－教材 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2003）第 094373 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969

文稿编辑：李　强
封面设计：立日新设计公司
印 装 者：北京鑫海金澳胶印有限公司
发 行 者：新华书店总店北京发行所
开　　本：185×260　**印张**：24　**字数**：566 千字
版　　次：2003 年 11 月第 1 版　2003 年 11 月第 1 次印刷
书　　号：ISBN 7-302-07456-9/TP・5505
印　　数：1～4000
定　　价：45.00 元(附光盘 1 张)

译 者 序

当前，很多成功的企业正在应用数据挖掘来帮助它们更好地制度决策。利用功能强大的数据挖掘技术，可以把数据转化为有用的信息以帮助制定决策，从而在市场竞争中获得优势地位。数据挖掘是一个过程——是一个不断把商业经验和知识与数据相结合的过程。通过数据挖掘，可以更好地认识所面临的问题并发现新的市场机会，做出更加明智的决策。

数据挖掘的目标是找到能够帮助他们做出对其成功至关重要的决策的信息。例如，他们想知道这样一些情况："现有客户中哪些会对我们的新产品感兴趣？""这个贷款申请有合理的信用风险吗？"等等。数据挖掘中应用的方法包括传统的统计分析、分类、估计、预测和相关性分析或关联规则、聚集，也包括最新发展起来的一些诸如数据可视化、决策树和神经网络等一些较新的方法。越来越多的高等院校已经开设或正在准备开设数据挖掘方面的课程。

本书是新近推出的一本有关数据挖掘方面的好书。作者认为：数据挖掘模型的建立既是一门科学，也是一门艺术，是对"在实践中学习"的最佳诠释。书中自始至终都体现了作者这样的理念，采用全面的教程风格，提供了执行数据分析易于学习的一步一步的指南。通过列举各种数据挖掘技术建立模型简单详细的例子，揭去了数据挖掘的神秘面纱。本书提供了配套的数据挖掘数据集和一个用于数据挖掘的软件 iDA，该软件基于 Windows 风格和 Excel 配合使用。书中各章提供了关键术语的解释，提供了 3 种类型的习题：复习题、数据挖掘题和计算题。特别适于用作教材，这也是我们翻译此书的主要动机。

本书的作者 Richard J. Roiger 博士是美国明尼苏达大学计算机与信息科学系教授，多年从事数据挖掘与知识发现以及机器学习领域的教学、研究与应用工作。他为计算机相关专业的大学生开设了数据挖掘和专家系统课程，有着丰富的教学经验。本书的另一作者 Michael W. Geatz 是一位来自企业的专家，有着丰富的企业应用数据挖掘的经验。两位作者的合作，是学院经验和企业经验的结合，这样一种成书模式，应该引起我们关注，在积极推进我国教育教学改革的今天，笔者认为这样的模式是我们需要积极鼓励和倡导的。

感谢清华大学出版社对本书中译本出版的支持，感谢香港科技大学黄哲学博士、佛罗里达州州立大学终身教授，北京航空航天大学软件学院院长孙伟博士在百忙之中，审阅了部分译稿，并提出了很多好的建议。感谢北京应用文理学院信息技术系戴红讲师，她翻译了本书的部分章节。感谢北京航空航天大学计算机学院的苏淑文、黄坚、朱逸鹏、隋明祥等研究生对本书的译稿和录入所做的工作，全书由翁敬农副教授负责审核统稿。

本书内容涉及面广，许多术语目前尚无一致译法，虽几经斟酌，多方查找资料，仍难

免有词不达意之处，个别术语采用中英对照方式，抛砖引玉，大家一起探讨。我的联系方式，电话：010－82309816，邮箱：*wengjn@buaa.edu.cn*或 *jnweng@263.net*,欢迎读者来电来函，对书中不妥之处批评、指正。

翁敬农

2003年8月　于北航

前　言

数据挖掘（data mining）是发现数据中有用模式的过程。数据挖掘的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。数据挖掘过程涉及下列几个研究方面：

- 数据收集与存储
- 数据选取与准备
- 模型建立与检验
- 解释与验证结果
- 模型应用

尽管在本书中我们对数据挖掘和知识发现的各个方面都进行了一定的说明，但一本书不能够浓缩数据挖掘的所有方面，本书的侧重点在于模型建立与检验（model building and testing），以及解释与验证结果（interpreting and validating result）。

为了帮助你更好地了解数据挖掘过程，我们提供了一个基于 Microsoft Excel 的数据挖掘工具，读者可以用它来实验性地建立和检验数据挖掘模型。智能数据分析器（Intelligent Data Analyzer ，iDA）是 Information Acumen 公司的产品，它通过提供可视化的学习环境、集成的工具集和数据挖掘处理支持系统，为商业或技术分析提供了支持。尽管我们推荐该软件，你也可以使用其它的软件包。书中的第 4 章、第 9 章以及第 5 章的第 10 节直接涉及到随书所属软件。

学习目标

撰写本书的初衷是帮助学生达到以下学习目的：

- 了解什么是数据挖掘以及如何用数据挖掘来解决实际问题。
- 识别某个数据挖掘解决方案对特定问题是否切实可行。
- 逐步学习知识发现的过程，写一篇关于数据挖掘对话结果的报告。
- 利用基本的统计和非统计技术评估数据挖掘对话的结果。
- 识别几种数据挖掘策略，了解每种策略的适用时机。
- 全面了解如何通过几种数据挖掘技术建立模型来解决问题。
- 概要了解数据仓库的结构以及如何利用数据仓库增加商业机会。
- 了解什么是联机分析处理（OLAP）以及如何应用它来分析数据。
- 了解专家系统是描述模拟人类行为的一般模型。

- 了解如何借助目标树设计基于规则的系统。
- 认识智能代理系统是协助我们处理日常任务的计算机程序。
- 了解结合专家系统的问题解决方法和数据挖掘策略能解决哪些类型的问题。
- 了解如何应用本书所附软件来解决实际问题。

读者对象

为主修或辅修商业或计算机科学的大学生开设的讲授数据挖掘课程的一学期中，我们为这本书准备了大部分的材料。 我们的课程还包括了基于规则的专家系统和智能代理的一个单元的内容。在编写本书时，我们主要针对以下 3 类读者群：

- 教育者：希望讲授数据挖掘和智能系统中的一个单元、一个专题或整个课程的教师。
- 学生：希望学习有关数据挖掘的知识并希望使用数据挖掘工具亲手实践的学生。
- 商业专家：需要了解如何应用数据挖掘和智能系统帮助解决其商业问题的专家。

各 章 特 点

我们所采取的方法，即模型建立既是一门艺术也是一门科学，是对“从实践中学习”的观点的最佳诠释。我们的观点为文中的多种特征所支持，下面列出了部分特征：

- 简单而详细的例子。我们通过列举不同数据挖掘技术如何建立模型的简单、详细的例子，揭去了数据挖掘的神秘面纱。因为它具有辅导教材的特点，本书适合作为数据挖掘和知识发现的自学指导书和大学教科书。
- 全面的教材风格。第 4 章、第 5 章、第 6 章、第 7 章、第 9 章和第 10 章中精选的小节提供了对执行数据分析的易于学习的一步一步的指南。
- 数据挖掘会话。数据挖掘会话允许学生使用提供的软件来完成数据挖掘过程的步骤。每个会话都突出显示以易于与正文区分开。
- 数据挖掘的数据集。来自商业、医学和科学研究中的不同数据集为数据挖掘做好了准备。
- 提示框。提示框用于介绍数据挖掘的数据集并强调重要的信息。
- 数据挖掘的 Web 站点。提供了链接到包含有趣的数据集的多个 Web 站点。
- 数据分析工具。举例说明了 Excel 中多种有用的数据分析工具。这些工具包括执行线性回归分析的 LINEST 函数，以及总结和分析数据的数据透视表。
- 关键术语定义。每章都介绍了关键术语，在每章最后提供了术语的定义列表。
- 章末习题。用于巩固各章中提到的技术和概念，这些习题分为 3 类——复习题、数据挖掘习题和计算性习题。适于通过实验完成的练习题加有标注。
    - 复习题 提出一些关于各章内概念和内容的基本问题。这些问题是用来检查

读者是否理解了各章所传达的的主要观点。

- 数据挖掘习题　要求读者使用一个或多个数据挖掘工具来执行数据挖掘会话。
- 计算性习题　该类习题有一定的数学风格，因为它们要求读者执行一个或多个计算。许多计算题适于高年级的学生作为挑战内容。

各章内容

各章顺序以及本书各部分的划分是基于多年的数据挖掘和专家系统课程的教学经验。第Ⅰ部分介绍了理解数据挖掘处理过程的基本内容。该描述是比较通俗的且容易理解。这一部分介绍了基本的数据挖掘概念、策略和技术，学生能从中学到能用数据挖掘来解决的问题类型并精通随书软件的应用。该部分还描述了几种数据挖掘成功应用的现实生活中的例子。

了解了基本概念后，第Ⅱ部分通过介绍数据库中的知识发现（KDD）处理模型来形式化数据挖掘的问题求解。KDD 处理模型是数据挖掘科学方法的应用。它强调了“数据预处理是成功的数据挖掘的基础”这一事实。重点放在讨论数据仓库的作用和数据挖掘的评估技术上。

第Ⅲ部分详细描述了几种高级数据挖掘方法。介绍了当前大家感兴趣的主题，如神经网络学习、时间序列分析、对数回归以及基于 Web 的数据挖掘，并提供了 iDA 神经网络软件的使用指南。

尽管数据挖掘是一种适合许多应用的解决方法，但有时这种方法并不可行。幸运的是，当数据挖掘不可行时，其它用于创建有用决策制定模型的选择可能是合适的。第Ⅳ部分析了基于规则的系统和智能代理系统作为建立协助决策制定过程的模型的替代方法。重点放在将这些技术与数据挖掘相结合以解决复杂问题。

下面简单描述了各章的内容。

第Ⅰ部分　数据挖掘基础

- 第 1 章　概述数据挖掘过程各个方面。特别强调的是帮助学生确定数据挖掘什么时候适合作为解决问题的策略。
- 第 2 章　列出几种通用数据挖掘策略和技术的大纲。描述了用于评估数据挖掘会话结果的基本方法。
- 第 3 章　详细描述了一种决策树算法，用于创建关联规则的 apriori 算法，用于无指导聚类的 K-平均值算法，以及两种遗传学习技术。所提供的工具可用来帮助确定使用哪种数据挖掘技术来解决特定的问题。

- 第 4 章 数据挖掘工具 iDA 软件包的指南。描述了执行有指导学习和无指导聚类的一般方法。

第 II 部分 知识发现工具

- 第 5 章 介绍了 KDD 处理模型作为一种解决数据挖掘问题的形式化方法。这个模型的一个简化版被用来解决两类数据挖掘问题。
- 第 6 章 对数据仓库设计和 OLAP 进行了介绍，包括使用 Excel 数据透视表进行数据分析的使用指南。
- 第 7 章 描述了评估数据挖掘会话结果的形式化统计和非统计方法。并提供了使用 Excel 来计算属性相关性及显示散列图的指导。

第III部分 高级数据挖掘技术

- 第 8 章 给出了两个流行的神经网络模型。神经网络训练的详细解释是提供给更倾向于技术型的读者。
- 第 9 章 提供了采用 iDA 神经网络创建工具解决数据挖掘问题的应用指南。讲述使用有指导学习来评估无指导神经网络聚类结果的方法。
- 第 10 章 详述了几种统计技术，包括线性和对数回归、贝叶斯分类器以及 3 种无指导数据挖掘方法。提供了用 Excel 的 LINEST 函数执行线性回归的指导。
- 第 11 章 介绍了执行时间序列分析、基于 Web 的挖掘和文本数据挖掘技术。并讲述了作为提高模型性能方法的装袋（Bagging）和推进（Boosting）。

第IV部分 智能系统

- 第 12 章 介绍了人工智能和基于规则的系统。描述了使用目标树创建基于规则系统的一种通用方法。
- 第 13 章 介绍了基于规则系统中不确定性的来源。阐述了模糊逻辑和贝叶斯推理是推理不确定信息的方法。
- 第 14 章 介绍了智能代理这种计算机程序能帮助我们处理日常任务。讲述了一个结合智能代理、数据挖掘和专家系统解决困难问题的模型。

本书补充内容

本书附带有一个数据挖掘工具 iDA 软件包和一些用于数据挖掘的数据集。这是特别为课程讲授者设计的，以下对相关内容作简单的描述。

iDA 软件包

实践学习对于提高一名数据挖掘专家的技能是必需的。iDA 就是为给学生提供一个亲手体验数据挖掘过程而设计的软件。iDA 软件用在多个章节中以举例说明许多重要的数据挖掘概念。第 4 章、第 5 章、第 7 章、第 9 章、第 10 章、第 11 章和第 13 章这些章后的习题是为 iDA 软件设计的。

iDA 由 1 个预处理器、1 个报表生成器和 3 个数据挖掘工具组成。这 3 个数据挖掘工具是指用于有指导学习和无指导聚类的 ESX，一个用于创建有指导反向传播模型和无指导自组织映像的神经网络工具以及一个产生式规则生成器。作为 Excel 的插件，iDA 用户界面是 Microsoft Excel。我们选择 iDA 是因为它的灵活性和易用性。

iDA 数据集包

iDA 软件包括许多数据集。这些数据集来自 3 个一般应用领域——商业、医药卫生以及自然科学。所有的数据集都采用 Excel 格式且都可用。

可以从多个方面对数据集进行描述，包括数据实例的个数，属性的个数，缺失和噪声数据的数量，数据属性是否定义清楚，数据是分类类型、数值型还是两种数据类型的组合，在数据中是否存在定义明确的类，时间元素在数据中是否为固有的，输入属性是否可以和数据中已知类区分开，输入属性是否是相关的。由于这些因素影响了数据挖掘执行的方法，选择不同的 iDA 数据集以提供数据在这些方面的多样性。数据集还提供了许多常用功能：

- 提供给初学者实验数据以体验数据挖掘过程，不需要他们解决数据预处理问题。
- 给出了问题领域的范围以及适合数据挖掘解决方案的问题类型。
- 解释了数据挖掘的输出结果。
- 举例说明了知识发现的过程。
- 使读者认识到创建特定数据集的最佳模型，可能需要采用多种数据挖掘技术进行实验。

以下是部分 iDA 软件包数据集的简单描述，包括对每个数据集的一个或多个特征的简要说明。

1．商业应用

信用卡促销数据集。这个假想数据集包含了有关接受或拒绝各种促销的持卡者的信息。该数据集用于举例说明本文中所讨论的许多数据挖掘技术。

信用卡筛选数据集。这个文件包含了申请信用卡的个人信息。输出属性表明申请者是被接受还是被拒绝。为了保证数据保密性，输入属性已被转化为无意义的符号。

猎鹿人数据集。这个数据集包含了猎鹿人是否愿意为下次的打猎支付更多费用的信

息。数据中有许多不相关的输入属性。

股票指数数据集。该数据是纳斯达克和道琼斯工业指数每周平均收盘价的时间序列表示。

2. 医药和卫生

心脏病患者数据集。这个数据集包括两组人的医疗信息。第一组的成员都有过一次和多次心脏病发作病史，第二组成员没有心脏病病史。这个数据集包含了分类属性和数值属性的极佳的混合。

脊柱临床数据集。这个数据集包含了那些做过背部外科手术的患者的医疗信息。这些人中，有些人返回了工作岗位而有些人没有。这里并未给出每个属性含义的清晰定义。该数据集包含了数值数据和分类数据。

3. 自然科学

γ射线爆数据集。这个数据集包含了独立γ射线爆的记录信息。γ射线爆是来源于太阳系之外短暂的γ射线脉冲。这些脉冲是由 NASA 的康普顿γ射线天文台通过 BATSE（Burst And Transient Source Experiment）于 1991 年 4 月到 1993 年 3 月间观察到的。尽管天文学家同意γ射线爆存在分类，但是他们对存在一个特定的类结构持不同意见。

地球资源探测卫星图像数据集。这个数据集包括了表示一部分地球表面的数字化卫星图像的像素点。每个实例分类到 15 个类别中的一个，由于每个类都包括巨大的数据量，分类的准确度受到了特定模型参数设置的影响。

温度数据集。这个数据集提供了 56 个美国城市在一月份的平均最低华氏温度。同时提供了城市的纬度和经度值。所有的属性都是数值型的。

4. 杂项

泰坦尼克号数据集。这个数据集包含了 2 201 个实例，每个实例描述了一位泰坦尼克号上的乘客或船员的属性。输出属性表明该乘客或船员是否幸存。

为教师提供的补充材料

提供的下列补充材料有助于教师组织讲义和编写考题。

- PowerPoint 幻灯片。本书中的每个图和表都以 PowerPoint 表示。
- 测验题。每章提供多道测验题。
- 部分练习的答案。给出了大部分章后练习的答案。
- 课程规划者。课程规划者包含了关于讲稿格式方面的建议和讨论要点。规划者还

提供了在实验中使用有选择的章后练习的建议。

请注意这些补充材料仅仅适用于有使用资格的教师。请与你的 Addison-Wesley 销售代表联系或发 e-mail 给 *Computing@aw.com* 以索取这些材料。

推荐的课程大纲

有兴趣了解数据挖掘所带来的好处以及局限性这些最基本知识的读者，我们建议学习第 1 章、第 2 章、第 5 章和第 6 章，要想动手实验，还需学习第 4 章。

本书的第 I 部分、第 II 部分和第III部分为有关数据挖掘和知识发现的介绍性课程提供了素材。本书第IV部分还可以作为结合数据挖掘/专家系统的一门教程，其重点放在数据挖掘和知识发现。使用本书的人所需的预备知识很少，掌握电子数据表的基本操作、基本统计学和基础代数学的基础知识有助于更好地使用本书。

第 1 章为第 2~14 章提供了基本框架。第 2 章为第 3~11 章提供了必要的背景信息。如果你希望立即为学生提供一个亲手操作的学习体验，可以在学习第 1 章后，先切入第 4 章。学完第 2 章、第 3 章、第 4~7 章、第 10~12 章中的大部分内容后就可以任意顺序阅读。而第 9 章需要在第 8 章之后，第 13 和 14 章需在第 12 章之后阅读。

本书适合作为信息管理系统（MIS）和计算机科学专业的大学教材。它还可以为需要获得数据挖掘和知识发现初步知识的研究生提供指导性帮助。我们相信，在一个学期内能够讲授书中大部分的内容。以下是构造一门课程的一些可选方案。

1. 主修或辅修 MIS 专业的大学生的数据挖掘基础教程

包括第 1 章到第 6 章的详细内容，而第 3 章中的 3.3 节和 3.4 节可被忽略或简单带过。第 4 章需要足够的时间，使学生能轻松使用 iDA 软件工具。

如果你的学生没有学过基础统计学课程，第 7 章可忽略或简单带过。如果跳过第 7 章，那么 2.5 节（评估性能）需要多花些时间。学过至少一门商业统计课程的学生可以学习第 7 章的内容。

需要学习第 8 章，但 8.5 节是可选的。在第 9 章上需要花相当多的时间，它向学生介绍如何使用 iDA 神经网络软件工具。

第 10 章是可选的。有一些统计知识背景的学生将会发现，线性回归和对数回归以及贝叶斯分类器是很有趣的。对于第 11 章，所有学生都需要了解时间序列分析以及基于 Web 的数据挖掘和文本数据挖掘。11.4 节是可选的。如果时间允许，花一天或两天的时间讨论基于规则的系统（第 12 章）。

2. 强调数据挖掘智能系统的本科生 MIS 课程

MIS 专业本科生的数据挖掘课程学习计划：涵盖第 12~14 章的所有内容。忽略前面提到的可选章节以节省时间。如果时间允许，通过为学生准备提供动手机会的实验，该实验使用简单的基于规则的专家系统建立工具，以补充说明第 12~14 章的内容。

3. 计算机科学专业本科生主修或辅修的数据挖掘基础课程

包括第 1 章到第 5 章的详细内容。用一天或两天的时间学习第 6 章，它为学生提供了数据仓库设计的基础知识。包括第 7~11 章的大部分内容。如果时间较紧，你可以限制用在 10.4 节、10.5 节和 11.4 节上的时间。花些额外的时间在第 12 章上。

对于一门更深的课程，决策树属性选取（附录 C）和性能评估统计（附录 D）的内容可以作为正常课程的一部分。如果希望学生能使用一个或多个公共领域的数据挖掘工具，这些工具可从 *www.kdnuggets.com* 下载。

4. 强调数据挖掘智能系统的本科生计算机科学课程

以下是计算机科学专业本科生的数据挖掘教程计划。需要学习第 12~14 章的内容。如果时间不够，你可能希望仅仅学习第 10 章和第 11 章中特别感兴趣的章节，一种计划就是包括 10.1 节、10.2 节以及 10.4 节中的一个小节。

如果时间允许，可以通过为学生提供动手实验机会，该实验使用基于规则的专家系统建立工具，从而补充第 12~14 章的内容。你可能希望学生能使用一个或多个公共领域的数据挖掘工具，这些工具可从 *www.kdnuggets.com* 上下载。

5. 数据挖掘简明课程

对于有兴趣快速掌握数据挖掘初步知识的大学生和研究生来说，可以将时间用在第 1 章、第 2 章、第 4 章、第 5 章上。他们还可以从第 8 章（从 8.1 节到 8.4 节）和第 9 章中获得神经网络的实践知识。

致　谢

在很多人的帮助下，才使得本书的出版成为现实。感谢 David Haglin 和 Jon Haakila，他们对数据挖掘性能评估提出了重要的建议。感谢为本书附带的众多数据集做过预处理的人们。非常感谢 Yifan Tang 和 Suzy 对各章所做的评论。感谢所有参与出版工作的商业和计算机科学专业的大学生。

还要感谢我们的产品协调员 Keith Henry ，特别感谢 Addison-Wesley 的所有工作人员，他们为保证本书的高质量出版付出了辛勤的劳动。深深地感谢我们的编辑 Maite Suarez-Rivas。最后，感谢本书以下的评审者，他们的创造性建议在本书的初稿修订中起到了特殊的帮助作用。

Ananth Grama	Purdue University
John Keane	Department of Computation, UMIST-UK
Selwyn Piramuthu	Decision and Information Sciences, University of Florida
Mary Ann Robbert	Bentley College
Lynne Stokes	Southern Methodist University
Stuart A.Varden	Pace University

目　录

第Ⅰ部分 数据挖掘基础

第 II 部分 知识发现工具

第III部分 高级数据挖掘技术