国外信息技术优秀图书选译

数据挖掘方法与模型

Data Mining Methods and Models

Daniel T. Larose 著

刘燕权 胡赛全 冯新平 姜恺 译

高等教育出版社

国外信息技术优秀图书选译

SHUJU WAJUE FANGFA YU MOXING

数据挖掘方法与模型

Data Mining Methods and Models

Daniel T. Larose 著

刘燕权 胡赛全 冯新平 姜恺 译

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图字:01-2010-2231 号

图书在版编目(CIP)数据

数据挖掘方法与模型/(美)拉罗斯(Larose,D. T.)著;
刘燕权等译. —北京:高等教育出版社,2011.3
书名原文:Data Mining Methods and Models
ISBN 978-7-04-030968-3

Ⅰ. ①数… Ⅱ. ①拉… ②刘… Ⅲ. ①数据采集-
数学模型-研究 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2011)第 006259 号

策划编辑	刘 英	**责任编辑**	刘 英	**封面设计**	刘晓翔
版式设计	马敬茹	**责任校对**	刘 莉	**责任印制**	张泽业

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
			http://www.hep.com.cn
经 销	蓝色畅想图书发行有限公司	**网上订购**	http://www.landraco.com
印 刷	三河市华润印刷有限公司		http://www.landraco.com.cn
		畅想教育	http://www.widedu.com
开 本	787×1092 1/16	**版 次**	2011 年 3 月第 1 版
印 张	19.25	**印 次**	2011 年 3 月第 1 次印刷
字 数	380 000	**定 价**	49.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换。

物料号 30968-00

献　给

先我而去的亲人

父亲 *Ernest Larose*(*1920—1981*)，

母亲 *Irene Larose*(*1924—2005*)，

女儿 *Ellyriane Soleil Larose*(*1997—1997*)；

我可爱的孩子们

女儿 *Chantal Danielle Larose*(*1988*)，

女儿 *Ravel Renaissance Larose*(*1999*)，

儿子 *Tristan Spring Larose*(*1999*)。

译 者 序

随着信息技术，特别是数据库技术的快速发展和广泛应用，数据积累在各行各业越来越多，对数据分析的要求也越来越强烈。数据表明，20 世纪 90 年代以来，人类累计的数据信息量以每月高于 15% 的速度增加，如果不借助强有力的数据挖掘工具，仅依靠传统手工处理来理解这些数据是不可能的。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。缺乏挖掘数据所需的知识与手段，必将导致“数据爆炸但知识贫乏”的现象。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，而要发现数据中存在的关系和规则，根据现有的数据预测未来的发展趋势，则需新的技术和方法。

数据挖掘是近年来伴随着存储数据或瞬态数据的迅猛增长所激发的新技术和自动工具的需求，以智能方式将海量数据转换成有用的信息和知识，而发展出现的一门新兴交叉学科。数据挖掘涉及数据库、统计学、人工智能、机器学习等多个领域。计算机的应用产生了大量数据，数据挖掘就是利用上述学科的技术对海量数据进行处理，为决策者提供重要的有价值的信息知识。

在国内，越来越多的企业开始利用数据挖掘工具来分析数据，辅助决策层做出正确决策。然而，正是因为数据挖掘是新兴的交叉学科，一般的使用人员又没有接受过太多的统计知识方面的训练，使得人们对复杂的模型和算法望而却步，每个使用者在数据挖掘技术的使用上都是凭着感觉走，不知道哪种方法，这就急需数据挖掘方面的知识。目前国内现有教材一般都是大篇幅地阐述数据挖掘技术的原理和数学推导，很少有告诉读者怎么理解原理和使用方法的书籍，这对使用者来说难以解决实际问题。

《数据挖掘方法与模型》的引进弥补了同类书籍实用性不强的缺点，本书不局限于数据挖掘模型的理论阐述和数学推导，而是关注怎样使读者理解每一个模型并很好地在实践中运用各种模型。每一章均从实际生活中发生的实例入手，将每个模型的基本原理嵌套在案例中，告诉读者怎样理解模型的原理，怎样一步一步使用这些模型来解决实际问题，并对每个模型中有争议的问题和应注意的地方进行深入的讨论，让读者在实例中对这些复杂的数据挖掘技术不仅知其然还知其所以然，在面对大量数据时，学会采用恰当的数据挖掘技术，了解在该技术上要注意哪些问题。同时，在

每章后面都给出了关键点和复习题以及实际数据操作练习，方便读者更好地掌握这些模型方法的应用。

本书共7章，分为3部分，第1部分也就是第1章，是对降维方法的介绍，这是数据挖掘其他技术的一个先决条件，通过案例告诉读者怎样对大量数据先进行预处理，减少在使用具体技术进行数据处理时的麻烦和错误，提高数据分析结论的准确性。

第2部分包括第2章一元回归模型、第3章多元回归模型、第4章逻辑回归模型、第5章贝叶斯网络分析和第6章遗传算法。这5章都是经典的数据挖掘技术，书中没有对其进行大量的理论解释和推导，只通过实际案例引导读者怎样由已预处理的数据通过使用不同的挖掘技术从而得出所需结论，这对无论是数据挖掘入门者还是熟练者来说，都是不可多得的实际运用并熟练掌握这些技术的好教材。

第3部分为第7章，这是一个基于数据挖掘过程模型（CRISP－DM）上的多个案例研究，告诉读者数据挖掘的具体流程和每一流程上应该怎么做，并通过多个领域的案例来阐述这些流程和技术如何被运用，这也是对整本书的升华，使读者得到一个完整的数据挖掘体系方法。在这里，读者可以高屋建瓴地了解整个数据挖掘项目的全部过程和方法，准确把握完整的数据挖掘体系，为以后实际数据挖掘工作打下基础。

致谢

本书是我在北京大学软件与微电子学院作为客座教授执教数据挖掘课程时主持翻译的。北京大学软件与微电子学院2009OB922的同学参与了本书各章节的初译，胡赛全和冯新平研究生在译稿统稿中做了大量工作，为本书的完成做出了贡献；美国伊顿（Eaton）公司的姜恺先生对本书重要章节做了大量审校工作，保证了翻译的质量。本书的翻译顺利完成，得到刘英等编辑的帮助，得到高等教育出版社的大力支持，是他们敏捷、超前的思路才使读者能够有机会读到这样出色的一本书，谨此对他们表示衷心感谢。更要感谢的是原书的作者，是他使我们有机会系统地将数据挖掘的理论和实践揭示给中国读者。

参与本书各章节初译的北京大学软件与微电子学院的同学有：（排名顺序不分先后）柴楠、宫玲玲、黄温柔、张文静、董旭、厉鹏、孙子木、胡赛全、姜黎黎、潘硕、史俊平、刘佳、彭康、张萌、顾硕、马兰、赵志强、欧阳菲、孙亚红、李争艳、潘艳、苏毓仁、张治平、林莉、童菲、杨婷、饶侠、叶诚、张大川、夏露、陈昕、董洁、范国华、冯新平、龚涛、马茜、提云龙、何顺烽、廖丹、刘小敏、姚慧男、潘莉莉、瞿新芳、孙世超、符文君、花超、万浩华、薛飞、王敏、杨宇、郑江川及北京科技信息大学的张梦同学等。

由于时间仓促，此书如有疏漏和不妥之处，敬请读者指正。联系我们发E-mail至liuscsu@gmail.com。

刘燕权/Yanquan Liu
美国南康涅狄格州立大学

前　言

什么是数据挖掘?

> 数据挖掘是指从观察数据(经常是大量的)中分析探索到某些未知关系,并且用一种新的方式归纳数据,使得这些数据对于数据拥有者更加容易理解和有价值。
>
> ——David Hand, Heikki Mannila, Padhraic Smyth,《数据挖掘原理》, MIT 出版社, 剑桥,MA, 2001

根据在线技术杂志《ZDNET 新闻》(2001 年 2 月 8 日)的报道,数据挖掘被预言是"十年内最具有革命性的创新之一"。实际上,在《MIT 技术综述》中,数据挖掘被选定为能够改变世界的十大新兴技术之一。

由于数据挖掘代表了一个重要的领域,Wiley 将和我合作出版一系列关于数据挖掘的图书,第一阶段包括三本著作。此系列的第一本是 2005 年出版的《数据中的知识发现:数据挖掘简介》(*Discovering Knowledge in Data: An Introduction to Data Mining*),向读者介绍了数据挖掘这一迅速发展的领域。系列中的第二本即本书《数据挖掘方法与模型》(*Data Mining Methods and Models*),从建模的角度探讨数据挖掘的过程,这种复杂和有力的预测建模模型能够为广泛的商业和科研问题提供可行的数据分析结果。

为什么需要这本书?

《数据挖掘方法与模型》是《数据中的知识发现》一书的延续,为读者提供了:

- 用于揭示隐藏于信息中的"金子"的模型和技术;
- 数据挖掘算法的实际应用;
- 在大规模数据集中实现数据挖掘的经验。

白盒方法:理解基本算法和模型结构

盲目的黑盒数据挖掘方法会产生代价昂贵的错误,避免这种错误的最好方法就是用白盒方法代替。白盒方法强调的是对算法的理解和软件潜在的统计模型结构。

《数据挖掘方法与模型》通过以下方式应用白盒方法:

- 引导读者概览不同的数据挖掘算法
- 提供算法在大规模实际数据集上运作的实例
- 测试读者对算法和概念的理解程度
- 给读者提供一个在大规模数据集上做真实数据挖掘的机会

算法概览

《数据挖掘方法与模型》将带领读者浏览各种不同算法的步骤和微妙之处,使用样本数据集运算,让读者清楚地感受算法的内在原理。比如,在第 2 章中,观察的是一个新的输入数据如何改变模型结果;同样的,在第 6 章中,使用选择、交叉、变异的操作数,以便逐步寻求到算法的最优解。

算法和模型对大规模数据集的应用

《数据挖掘方法与模型》提供了各种算法在大规模数据集中的实际应用。比如,在第 1 章中,应用主成分分析法解析加利福尼亚州的人口普查真实数据;在第 3 章中,使用真实数据详细解析了营养分级和谷物含量的关系。这些数据集在本系列书的网站 www. dataminingconsultant. com 均可找到。

章节习题:测试读者对算法概念的理解程度

《数据挖掘方法与模型》包括了 110 个章节习题,使读者知道对内容的掌握程度,同时也可以与数字和数据打交道,乐趣悠然。概念辨析题的主旨是帮助读者澄清一些有挑战性的概念;数据集应用,挑战读者在小规模数据集上应用数据挖掘算法的能力,逐步找到已经计算完善的解决方案。比如,在第 5 章中,要求读者找到本章数据集和网络中提到的最大后验概率分类。

动手分析:通过挖掘数据来学习数据挖掘

第 1 章到第 6 章都向读者提出了需要实战分析数据的问题,读者可以利用这些机会,将自己新学习到的数据挖掘专业知识应用于解决实际问题。《数据挖掘方法与模型》给读者提供了一个可以实践的框架,比如,在第 4 章中,读者可以通过处理真实的信用认可分类数据,用本书中学到的方法构建自己的逻辑回归模型,这些逻辑回归模型对原模型提供支持,包括原有解释和指示变量。

案例研究:方法集合

《数据挖掘方法与模型》一书详细分析了一个案例研究:直接邮购营销回馈建模。在这里,读者可看到他或她将所学聚合以获取其可行和获利的方案。该案例研究包括超过50页的图标分析、探索性数据分析、预测模型、用户资讯分析等,以及根据用户不同需求所提供的相应解决方案。使用用户定制的成本/收益表,而不是用一般的方法如整体误差率评估这些模型,这样可反映分类误差的真实花费。因此,分析人员可以通过比较用于评估已记录个体客户利润的模型,来预测通过这些模型评估大量客户可获取的金钱额度。

数据挖掘作为过程

《数据挖掘方法与模型》继续了数据挖掘覆盖面作为一个过程。具体标准过程使用的是CRISP - DM框架(Cross - Industry Standard Process for Data Mining),意为跨行业数据挖掘标准流程。CRISP - DM要求把数据挖掘视为一个不可分割的整体过程,从业务问题沟通,通过数据选择和管理,数据预处理,建立模型,模型评价,直到最后的模型部署。因此,这本书不仅可用于分析人员和管理人员,也可供数据管理专业人员、数据库分析人员和决策者参考。

软件

本书包括的软件有:

- Clementine 数据挖掘软件
- SPSS 统计软件
- Minitab 统计软件
- WEKA 开源数据挖掘软件

Clementine (http://www.spss.com/clementine/)是广泛应用于数据挖掘的软件之一,由SPSS发布,其基础软件也用于这本书。SPSS软件试用版可下载限期使用,其网站是www.spss.com。Minitab是一个易于使用的统计软件包,用户可从其网站www.minitab.com下载试用版。

WEKA:开源替代工具

WEKA (Waikato Environment for Knowledge Analysis)机器学习Workbench是根据GNU通用公共许可证的开放源码软件,包括一个完成许多数据挖掘任务的工具集。《数据挖掘方法与模型》提出了一些实际操作,分步教程实例使用WEKA 3.4,同时输入文件可从本书的配套网站www.dataminingconsultant.com下载。通过对以下类型的分析向读者展示如何使用WEKA:逻辑回归(第4章)、朴素贝叶斯分类器(第5章)、贝叶斯网络分类(第5章)和遗传算法(第6章)。如需更多有关WEKA的信

息,见 http://www.cs.waikato.ac.nz/~ml/。WEKA 的实例和训练由 James Steck 提供,作者深表谢意。James Steck (james_steck@comcast.net)是作者在 2004—2005 年间的研究生助理。他是康涅狄格州立大学在 2005 年(成绩 4.0)第一批完成硕士数据挖掘课程的学生之一,并获得第一届数据挖掘研究生学术奖。James 与他的妻子和儿子生活在华盛顿的雷尼奥。

配套网站:www.dataminingconsultant.com

读者可在配套网站 www.dataminingconsultant.com 中找到本书的辅助材料和作者其他数据挖掘书籍资料。读者也可以找到书中所需要的数据集。很多数据集亦可下载,读者可使用本书中提供的方法与模式练习。作为一套完整的数据挖掘资源,此网站不仅包括勘误表,也包括数据集的超链接、数据挖掘组和研究论文。

本网站最主要的优势是为采用本书作为教科书的教师提供一个完善的终端,可获取如下资源:

- 所有练习题的答案,包括操作分析
- 每一章的 PPT 讲稿
- 样本数据挖掘过程的项目,作者在自己的课程上所需的资源
- 真实世界的数据集,可用于课程项目
- 章测验多项选择题
- 每章网络资源

《数据挖掘方法与模型》作为教科书

《数据挖掘方法与模型》适合作为数据挖掘介绍课程的教材,本书的特点是:

- 详述了数据挖掘过程
- 采用白盒的方法,强调了理解算法结构:
  - 算法概览
  - 大型数据集上应用算法
  - 每章练习
  - 操作分析
- 逻辑性表述,遵从 CRISP-DM 标准过程和数据挖掘任务
- 详细的案例研究,汇集了《数据挖掘方法与模型》和《数据中发掘知识:数据挖掘引言》两书中获取的经验
- 配套网站提供了各种相关资源

《数据挖掘方法与模型》适用于高年级本科或研究生课程。在几个章节中假定学生已掌握微积分知识,但无微积分知识的读者也可理解内容的主旨。提前学习一门统计导论课程更好,但并不是必需的。计算机编程或数据库知识亦非必需。

致谢

我要感谢所有在威利(Wiley)为本书工作的人员，特别是指导和帮助我的编辑Val Moliere。衷心感谢为本书提供WEKA资料的James Steck。

我还要由衷感谢我在美国中康涅狄格州立大学做数据挖掘项目科学研究的同事:Chun Jin, Daniel S. Miller, Roger Bilisoly 和 Darius Dziuda 博士;数学系主任Timothy Craine 博士;康涅狄格大学统计系主任 Dipak K. Dey 博士;西菲尔德学院数学系主任 John Judge 博士。没有你们，这本书依然是一个梦想。

感谢我的母亲 Irene R. Larose,她今年去世了,以及使这一切成为可能的父亲Ernest L. Larose。感谢我的女儿 Chantal 带给我的可爱艺术品和无限的快乐。感谢我的双胞胎孩子 Tristan 和 Ravel，分享这台电脑和他们独特的意境。我愿将我永恒感激之情特别呈送给我亲爱的妻子 Debra J. Larose，为她的耐心、爱心和与我“永恒的金带”。

在生活中携手共创，
我们将迈入
梦想的门槛……
——忧郁蓝调合唱团

Daniel T. Larose 博士
美国中康涅狄格州立大学数据挖掘项目主任
www. math. ccsu. edu/larose

目　录

第1章 降维方法

数据挖掘中降低维度的必要性
主成分分析法
因子分析
用户自定义复合

1.1 数据挖掘中降低维度的必要性

通常用于数据挖掘的数据库可能有上百万条记录和数千个变量。所有变量都是独立而没有任何关联的现象是不常见的。如《数据中发掘知识:数据挖掘引言》[1]中所提及的那样,数据分析人员需要防范多重共线性,即预测变量之间相互关联的情形。多重共线性会导致解空间的不稳定,从而可能导致结果的不连贯。如在多元回归中,即使单个变量的回归结果均不显著,预测变量的多重共线性集可能导致回归整体相对显著。即使上述的不稳定性得以避免,包含具有高度相关性变量的模型往往会强调其某一特定成分,该成分实质上被重复计算。

贝尔曼[2]指出,样本量需要符合一个多元函数,该函数跟随变量数呈现指数关系递增。换句话说,高维空间本身具有稀疏性。正如这个经验法则(empirical rule)告诉我们的,在一维空间的正态分布中,有68%的值介于正负标准差之间,而在10维多元正态分布中,只有0.02%的数据属于类似的高维空间。

在考察预测变量和回应变量之间的关系时,过多地使用预测变量会不必要地复杂化分析过程。这违反了简约原则,即应将预测变量的数目保持在可控的范围内。另一方面,过多的变量会妨碍查找规律的建立,因为新的数据对所有变量作出的反应很可能和建模中采用的数据反应不同。

此外,仅在变量层面上分析可能会忽略变量之间的潜在联系。例如,几个预测变量可能落入仅反映数据某一方面特征的一个组(一个因素或一个组成部分(components))内。例如,储蓄账户余额、支票账户余额、房屋资产、所持股票的价值