Clementine

数据挖掘方法及应用

薛　薇　陈欢歌　编著

PASW Modeler

统计分析教材

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

含光盘

Clementine 数据挖掘方法及应用

薛 薇　陈欢歌　编著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

数据挖掘是当前数据分析领域中最活跃最前沿的地带。本书以数据挖掘的实践过程为主线，通过生动的应用案例，从数据挖掘实施角度，系统介绍了经典的数据挖掘方法和利用 Clementine 实现数据挖掘的全部过程，讲解方法从易到难，说明问题从浅至深。本书力求以最通俗的方式阐述数据挖掘方法的核心思想与基本原理，同时配合 Clementine 软件操作的说明，希望读者能够直观了解方法本质，尽快掌握 Clementine 软件使用，并应用到数据挖掘实践中。为方便读者学习，书中所有数据和案例与所附光盘内容一致。

本书适合于从事数据分析各应用领域的读者，尤其适合于商业管理、财政经济、金融保险、社会研究、人文教育等行业的相关人员。同时，也能够作为高等院校计算机类、财经类、管理类专业本科生和研究生的数据挖掘教材。

图书在版编目（CIP）数据

Clementine 数据挖掘方法及应用 / 薛薇, 陈欢歌编著. —北京 : 电子工业出版社, 2010.9
ISBN 978-7-121-11778-7

Ⅰ. ①C… Ⅱ. ①薛… ②陈… Ⅲ. ①数据采集－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2010)第 175109 号

策划编辑：杨丽娟
责任编辑：杨丽娟
特约编辑：明足群
印　　刷：北京市顺义兴华印刷厂
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：787×1092　1/16　印张：19.75　字数：432 千字
印　　次：2010 年 9 月第 1 次印刷
印　　数：4000 册　　定价：38.00 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。

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

服务热线：（010）88258888。

前　言

数据挖掘是当前数据分析领域中最活跃最前沿的地带。

任何事物都有定性和定量两个方面，定量则产生数据。从数据分析入手是我们认识事物本质的基本手段。任何事物都是互相关联着的，从数据分析入手是我们把握事物之间联系的基本方法。任何事物都在永恒地变化发展着，从数据分析入手是我们探索事物发展规律的基本思路。所以我们进行数据分析，既是一种世界观，也是一种方法论。我们在研究着丰富多彩的客观世界的同时，也体现着分析者主观的智慧和自身的价值。

随着中国社会经济的蓬勃发展，在错综复杂的宏观、中观和微观的共同作用下，战略决策和战术选择都显得敏感而关键，越来越多的人们加入到数据分析的行列中来。这是一个非常富有挑战性的工作，不但有意思而且有意义。

IBM 公司于 2009 年 1 月公布了其“智慧地球”战略。该战略的主要思想是，将传感设备或智能仪表嵌入到建筑、电力、交通、管道等各种物体中，进行数据自动采集，之后基于互联网形成物物相联的物联网，然后通过超级计算机和云计算将数据整合，进行智能化分析和建模，从而实现社会与物理世界的融合。这是一个未来理想化的信息世界图景。

在这个智慧系统中，其核心是数据处理。为此，IBM 公司于 2009 年 7 月斥资 12 亿美元收购了著名的 SPSS 统计分析软件公司，将其应用广泛的 SPSS 统计分析软件和 Clementine 数据挖掘软件纳入麾下。同时对软件产品进行了整合，将 Clementine 更新命名为 PASW（Predictive Analytics SoftWare）Modeler，并快速推向市场。

目前，SPSS Clementine 软件已经连续若干年蝉联数据挖掘应用的王者，而业界对于 PASW Modeler 的认知则刚刚开始。所以本书继续沿用为广大读者所熟悉的 Clementine 这个名字。

Clementine 软件不但将计算机科学中许多机器学习的优秀算法带入到数据分析中来，同时也综合了一些行之有效的数据挖掘方法，成为内容最为全面、功能最为强大的数据挖掘产品。

Clementine 软件充分利用计算机系统的运算处理能力和图形展现能力，将方法、应用与工具有机地融合为一体，是解决数据挖掘问题的最理想工具。

Clementine 软件继续保持了 SPSS 产品的一贯风格：界面友好且容易使用。复杂的数学算法和冗余的输出结果被软件隐藏在程序系统内部。Clementine 软件始终把自己的

应用对象锁定在实际部门的业务分析人员，而不是一个数据分析专家。这种“傻瓜型”软件经常遭到一些精英学者的指摘，但是这恰恰成为 Clementine 成功开拓自己应用疆域的最有效利器。

本书作者常年从事计算机数据分析的教学与科研工作，并长期跟踪研究 SPSS 公司数据分析产品，具有计算机应用与统计分析的双重学历背景。我们深知，对于数据挖掘这样一款综合方法性的软件工具来说，一个基层的读者应该从哪些方面入手，就能很快地掌握和使用 Clementine 开始数据挖掘工作，并从中受益。

本书默认读者具有以下特征：具有基础的计算机操作能力；不甚了解数据挖掘的原理和方法；有自己应用领域积累的数据，渴望使用数据挖掘方法解决实际问题。

所以，针对上述读者群，本书的特点是：

1．以数据挖掘过程为线索介绍 Clementine 软件

目前，具备基本的计算机操作能力已经不是读者的障碍，数据挖掘的过程与方法才是读者关心的主题和应用的难点。各领域众多的读者所面临的问题是：自己拥有的一批数据不知道如何使用 Clementine 进行组织，不知道如何利用 Clementine 对它们进行基本加工和整理；其次，不知道选择 Clementine 中的哪些方法对数据进行分析，不知道如何解释分析结果。

因此，本书以数据挖掘的实践过程为主线，从 Clementine 数据管理入手，说明问题从浅至深，讲解方法从易到难。这样，能使读者在较短时间内掌握 Clementine 的基本功能和一般方法，并可快速地运用于实际工作中。

2．数据挖掘方法、软件操作、案例分析的有机结合

目前，由于数据挖掘方法的中文资料相对不足，Clementine 相关书籍都比较侧重对其英文手册的翻译介绍，侧重于计算机操作过程的描述。而对数据挖掘方法则较多地罗列数学公式，对于输出结果也缺少恰当的解释。

本书作者配合实际案例，侧重数据挖掘方法核心思想和基本原理的阐述，使得读者可以直观理解方法，并正确掌握方法的应用范围，不至于滥用或者误用。同时介绍软件操作，使得读者能尽快熟悉 Clementine 软件，从而在理解方法与掌握操作的基础上对输出结果进行合理的解释。

3．数据挖掘方法讲解通俗，软件操作过程说明翔实

针对初学者的特点，本书力求以最通俗的方式对数据挖掘方法的核心思想与基本原理进行讲解，同时避免大量罗列数学公式、数学推导与数学证明，目的是使读者能够直观地了解方法的本质，并正确运用；介绍方法的同时也紧紧围绕 Clementine 的输出结果展开，以使读者理解分析结论的重要性，会合理地引用分析结果。另外，本书对 Clementine 的操作过程也给出了较为翔实的说明，但并非是对菜单功能清单的描述，而是将其穿插于分析案例的实现过程中。

本书适合于从事数据分析各应用领域的读者，尤其适合于商业管理、财政经济、金融保险、社会研究、人文教育等行业的相关人员。同时，也能够作为高等院校计算机类、财经类、管理类专业本科生和研究生的数据挖掘教材。

本书共分十章，由薛薇、陈欢歌执笔完成，全书最后由薛薇审核定稿。本书所附光盘配备全书的案例数据和数据流文件。

由于水平所限，书中难免出现错误，敬请读者批评指正。

编著者

目　录

第 1 章　数据挖掘和 Clementine 概述

数据挖掘，作为 20 世纪 90 年代中后期兴起的，具有鲜明跨学科色彩的应用和研究领域，因其注重减少数据分析方法对数据的限制性和约束性，注重与计算机技术结合以实现数据的可管理性以及分析的易操作性，已成为数据分析应用实践的新生代。同时，随着数据挖掘方法的不断成熟及其应用的日益普及化，数据挖掘软件的研发也取得了令人可喜的成果。目前，以 Clementine 为代表的数据挖掘软件，因其有效地将束之高阁的数据挖掘理论成果解放到数据分析实践中，已普遍应用于商业、社会、经济、教育、金融、医学等领域，并成为数据分析的主流工具，得到数据分析相关领域的极大关注。

1.1　数据挖掘的产生背景

数据挖掘的产生和兴起是在计算机数据库技术蓬勃发展，人工智能技术应用领域不断拓展，统计分析方法不断丰富过程中，为有效迎合数据分析的实际需求而逐步形成和发展起来的一门具有鲜明跨学科色彩的应用研究领域。

1.1.1　海量数据的分析需求催生数据挖掘

20 世纪 80 年代以来，随着计算机数据库技术和产品的日益成熟以及计算机应用的普及深化，各行业部门的数据采集能力得到了前所未有的提高，组织通过各自内部的业务处理系统、管理信息系统以及外部网络系统，获得并积累了浩如烟海的数据。以商业领域为例，美国著名的连锁超市 Wal-Mart 的数据库中已积累了 TB[①]级以上的顾客购买行为数据和其他销售数据。随着互联网和电子商务的普及，各类网上书店、网上银行、网上营业厅和网上商城等积累的 Web 点击流数据，存储容量也多高达 GB 级。另外，国家政府部门所积累的数据量也令人瞠目。例如，一次全国经济普查或人口普查所采集和处理数据量均在千万级以上。同时，各经济行业的企业内部也拥有大量的业务数据、财务数据和人事数据。

在严酷的市场竞争压力下，企业为更客观地把握自身和市场状况，提升内部管理和决策水平，管理者们面对如此丰富的海量数据，分析需求越来越强烈。他们希望利

① 1T=1000G。

用有效的数据分析工具，更多地挖掘出隐藏在数据中的、有价值的辅助于管理和决策的信息。

例如，制造业已从过去的粗放式生产经营模式过渡到精细化的生产管理。决策者需要了解客户偏好，设计最受市场欢迎的产品；需要制定合适的价格，确保企业的利润；需要了解市场需求，调整产销计划，优化库存结构；需要评估供应商质量，供应合同和订单违约率，提高产品合格率以及风险控制能力等。

再如，政府部门中的政策制定者们，为保证出台政策的科学性和全面性，也希望利用数据分析方法，对现有数据进行科学缜密的分析。

因此，正像著名的数据仓库专家 Ralph Kimball 在其著作中写的那样："我们花了二十多年的时间将数据放入数据库，如今是该将它们拿出来的时候了。"

然而，令人棘手的问题接踵而来。原来管理者们得不到想要的数据，是因为数据库中没有充足的数据，但现在他们似乎仍然无法快捷地得到想要的数据，其原因是数据库里的数据太多了。人们面对规模庞大、纷繁复杂的数据，漫无头绪无从下手，致使原本宝贵的数据资源成了使用者的负担。组织中的管理决策者无奈地感慨：基层业务人员尚且能够通过业务处理系统快速访问一定范围内的业务数据，而高层决策者却似乎缺少有效的工具，从数据库中获得利于决策制定的有价值的数据。于是，所谓的"信息爆炸"、"数据多但知识少"成为一种普遍的怪现象。

究其原因，一方面，对于基层业务人员来说，由于业务处理系统是依据一定的业务流程，符合一定的业务规范的，所以通过业务处理系统业务人员能够灵活自如地掌控"自己的"数据；而对于管理决策者，他们所需要的数据通常来自于各个业务处理系统，但由于业务处理系统是分散性的，加上管理、规划、设计、技术等诸多因素影响，各系统基本处于"封闭"状态，系统之间的数据交换需求极少，而且交换的渠道也不很畅通。尽管客观上各系统之间仍然存在数据重复录入、数据不一致性等问题，但由于基层业务处理具有"各自为政"的特点，因此对日常业务处理似乎并无大碍。然而由此形成的"信息（数据）孤岛"现象，对那些正在逐渐摒弃"凭经验"、"拍脑袋"决策方式的领导者们来说，却是一个大忌。他们深刻认识到，如果无法有效快捷地将各系统中的数据整合到一起，就无法及时得到全面准确的数据，更无法进行分析而做出正确决策。

另一方面，数据的定量分析是科学决策的前提。但实施定量分析需要深厚的专业知识，更需要有效的分析工具。但一般业务处理系统中的数据分析功能相对简单，通常只能制作各种数据汇总报表，无法实现对数据的深层次分析，因此不能很好地满足决策者的定量分析需求。

大规模海量数据的整合处理和深层次量化分析的实际需求，直接孕育了 20 世纪 90 年代初期的两项重大技术，这就是数据仓库技术和数据挖掘技术。数据仓库和数据挖掘的产生和发展，使得当今的计算机网络应用体系从业务管理层逐步跃升到决策支持层。

同时，两者在技术和产品的上互相补充和互相促进，逐渐形成了融合发展的可喜局面，为最终形成具有一定通用意义的决策支持系统奠定了良好的基础。

1.1.2　应用对理论的挑战催生数据挖掘

应用需求对理论研究的牵引力是巨大的，没有应用背景的理论研究是没有价值的。在海量数据管理和分析应用呼声不断的同时，相关理论研究和应用实践的脚步也未曾停止。数据库与数据仓库、人工智能与机器学习、统计学理论应用的发展是数据挖掘诞生的坚实基础。

一、数据库和数据仓库

计算机应用从其刚刚诞生时的以数值计算为主跨越到当今的以数据管理为主，数据库的理论实践起到了巨大的推波助澜的作用。从最初的文件系统研究，到后来的层次模型、网状模型，直至 1969 年 E.F.Codd 提出关系数据模型，可以说数据库理论开创了数据管理的新时代。数据库以其卓越的数据存储能力和数据管理能力，得到了极为广泛的应用。随着数据库中数据的不断积累以及人们对海量数据分析需求的强烈，数据库的理论实践开始思考这样的问题：是否存在更有效的存储模式实现高维海量数据的存储管理？数据库仅仅是用来存储数据的吗？难道数据库对数据的管理只仅仅停留在简单的查询和汇总上吗？

应用呼唤理论的发展和理论的再实践。通过数据库研究者们的不懈努力，在数据库基础上逐渐发展完善起来的数据仓库技术，已经成为一种有效的面向分析主题的数据整合、数据清洗和数据存储管理集成工具。同时，在机器学习和统计学等领域研究成果的基础上，数据仓库正在不断吸纳经典的数据分析方法并将其融合到商业产品中。

例如，许多知名的数据库厂商，如 Microsoft 公司的 Sql Server 产品提供了多种典型的数据挖掘算法；Oracle 公司的 Orcale 产品包含了包括关联规则和贝叶斯算法在内的众多数据挖掘算法；而 IBM 公司更是通过斥资 12 亿美元收购了业界极为著名的，麾下拥有 SPSS 统计分析软件和 Clementine 数据挖掘产品的 SPSS 软件公司，扩展了 IBM 的“信息随需应变”软件组合和商业分析能力。IBM 表示，收购 SPSS 将增强公司“信息议程战略”（Information Agenda initiative）的业务实力，帮助客户公司更有效地将信息转化为战略资产。

与此同时，研究者们也在为实现数据仓库中数据和分析模型的无缝交互，以及不同数据库仓库产品间的数据挖掘分析方法共享而不懈努力着。例如，1999 年，Microsoft 公司提出了 OLE（Object Linking and Embedding）DB（DataBase）for DM（Data Mining）规范，研发了模型建立、模型训练和模型预测的数据挖掘语言。其核心思想是利用 SQL 和 OLE DB 将数据库中的关系概念映射到数据挖掘中；包括 IBM、Microsoft、Oracle、SAS、SPSS 等大公司在内的数据挖掘协会，提出了预测模型标记语言 PMML（Predictive

Model Markup Language），它标准化了常见数据挖掘算法的模型内容，并以 XML 格式存储，使不同软件之间的模型交换和共享成为可能。以 Microsoft 公司的 Sql Server 产品和 IBM 公司的 Clementine 产品为例，当用户在计算机中安装了 Sql Server，如果在 Clementine 中建立和执行数据挖掘流，则 Clementine 会将挖掘流提交给数据库，并利用数据库系统所提供的各种数据管理优化机制，直接读取数据库中的数据而不必下载到 Clementine 中，且模型结果可存储于数据库中。

二、人工智能和机器学习

人工智能和机器学习的理论研究一开始就具有浓厚的应用色彩。针对如何利用计算机模拟人脑的部分思维，如何利用计算机进行实际问题的求解等，人工智能和机器学习的理论研究主要集中在基于谓词演算的机器定理证明技术和二阶演绎系统等方面，可以说成果丰硕。然而，其理论实践过程中出现了许多问题。

例如，作为人工智能和机器学习研究成果之一的专家系统，在某种意义上能够代替专家给病人看病，能够帮助人们识别矿藏，但却很难解决那些看似简单但却极为复杂的问题。如专家系统建立中的知识获取过程，出现了诸如人脑是如何思维的，计算机技术人员应以怎样的方式与领域专家交流才能全面获取其专业知识，如何克服知识交流过程中的随意性、跳跃性等一系列问题；再如，专家系统的知识表示过程中，出现了因计算机的知识表示通常是“机械”化的“如果……那么……”方式，而专家的领域知识丰富多彩，并不是所有知识都能够概括成“如果……那么……”的模式等问题；再如，专家系统中获取和存储的知识绝大部分是领域的专业知识，常识性知识很少。但没有常识的专家系统有时会比傻子还傻。人工智能学家 Feigenbaum 曾估计，一般人拥有的常识存入计算机大约有 100 万条事实和抽象经验。将如此庞大的事实和抽象经验整理表示并存储在计算机中，难度是极大的。

正是这样，人工智能和机器学习的应用重心开始从博弈、自然语言理解、专家系统等领域向更具应用意义的数据分析方面转移。机器学习方法，如决策树、神经网络、推理规则等，能够模拟人类的学习方式，向数据案例和经验学习，并通过学习实现对新事物所具模式的识别和判断，而这种方式恰恰为数据分析提供了极为绝妙的研究思路。

三、统计学

统计学发展至今已有几百年的历史，它为数据收集、整理、展现和分析过程提供了完整的理论框架和实践依据。然而在信息技术迅猛发展，数据量高速膨胀、数据类型日益丰富、数据管理和分析需求不断提升的当今，统计学的理论研究和应用实践也面临着诸多挑战。这主要体现在理论研究和应用实践两个方面。

● 理论研究方面

例如，在数据采集能力极为有限的过去，人们只能通过研究少量样本来推断总体特征。此时，作为统计学传统方法的推论统计具有极高的应用价值。但在数据采集能力极

强的今天，有时摆在人们面前的不再是“小样本”而是海量的高维总体，此时推断不再有意义，且原本较小的参数差异在大样本条件下都表现出了“显著”；再如，经典统计分析方法往往是模型驱动式的。以统计学中应用极为广泛的线性回归分析方法为例，它即是首先确定模型，然后利用数据建立、验证模型，最后应用模型。这样的研究模式是建立在对模型的“先知先见”基础上的。但在数据庞大结构复杂的今天，这种“先知先见”几乎是不可能的，数据驱动式的分析思路似乎更为现实。因此，基于模型假设进行总体推断和检验的传统分析方法已显露出很大的局限性。

为克服统计分析方法应用过程中的诸多问题，20 世纪 60 年代，稳健统计开始盛行。它通过敏感性分析、异常值诊断等手段，开创性地解决了当数据与理论分布假设有偏差的分析问题。20 世纪 70 年代中期，John Tukey 提出的探索性数据分析（EDA）方法，开始打破统计方法中分布假设的古典框架，注重从数据的特征出发研究和发现数据中有价值的信息。在之后至今的几十年发展历程中，统计方法在与数据相结合的道路上硕果累累，许多新的统计技术应运而生。在摆脱古典框架约束方面，通过马尔可夫链蒙特卡罗（Markov Chain Monte Carlo，MCMC）模拟以及 Bayes 统计等方法，着力解决复杂模型识别和分析问题。利用 Jack-knife（刀切法）、Cross-Validate（交叉验证）、Bootstrap 等方法解决模型评价和选择问题。此外，在分析结果展示方面，除传统的数学语言表示之外，统计也力图更多地借助现代计算机技术，实现高维数据分布特征以及分析结果的图形化展示，数据的可视化技术已成为统计和计算机界共同的热门话题。

● 应用实践方面

例如，数据整理是统计分析必不少的重要环节。在数据量相对较少的过去，数据整理可以通过手工或借助简单工具实现。但随着数据量的快速膨胀，这个问题不再仅仅是个量变而成为一种质变。从工作量看，数据整理的工作量已经占到整体统计分析工作量的 70%到 80%或更高；从工作方式看，手工或借助电子表格软件整理数据的方式已显得无能为力。

表面看上述问题源于数据整理手段和工具效率不高，但本质上却源于数据的存储组织模式。因为，数据整理的高效率是建立在良好的数据组织模式基础上的，只有好的数据组织模式才可能支撑高效率的数据整理。因此，过去在统计应用视野之外的数据存储和组织问题，今天不得不成为统计应用实践的焦点，统计应用与计算机数据库技术相结合已是大势所趋。

再如，整体解决方案已成为统计应用实践的大趋势。过去，人们的统计应用实践往往呈现出“片段性”的特点，原本完整的统计应用呈“割裂”状。以企事业统计为例，统计应用实践应包括建立指标体系，采集数据，存储和管理数据，分析数据和制定决策等多个相互影响和制约的环节。但如果将其割裂开，必然会出现各自为政、各行其是的局面。于是，一些统计人员脑子中“我只负责指标框架设计不考虑具体实施”、“你给我数据，我