宝典丛书

200万

数据挖掘原理与SPSS Clementine应用宝典

从数据挖掘原理与SPSS Clementine实践，层层深入技术内幕。

本书点面兼顾，目录分类细致而科学，方便快速查阅。

配套代码和精美PowerPoint 幻灯片课件。

元昌安 主 编

邓 松 李文敬 刘海涛 等编著

電子工業出版社
Publishing House of Electronics Industry
http://www.phei.com.cn

宝 典 丛 书

数据挖掘原理与SPSS Clementine应用宝典

元昌安 主编

邓 松 李文敬 刘海涛 等编著

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书从数据挖掘基础、数据挖掘经典算法、数据挖掘业务建模与模型评价、SPSS Clementine 数据挖掘实务这 4 方面对数据挖掘技术进行了全面介绍，既包含传统经典的数据挖掘方法，同时也包含了部分数据挖掘的最新研究成果；通过学习读者可以对数据挖掘理论有一定的认识，理解数据挖掘经典算法的实现，并且可以掌握数据挖掘建模以及 SPSS Clementine 数据挖掘实战。

本书共24章，分为4部分。第1部分　数据挖掘应用基础，包括第1～5章。通过本部分的学习可以了解掌握数据挖掘的基本概念及数据挖掘应用的基本原理。第2部分　数据挖掘经典算法，包括第6～15章，包括回归分析的基本原理以及各种回归分析的方法；贝叶斯网络的基本概念和一些常用的算法；聚类分析的原理和常用的聚类算法；决策树算法的原理和常用算法；关联规则的基本概念、原理以及常用算法；粗糙集基本概念，算法以及在数据挖掘中的应用；基本的神经网络模型的原理和算法；遗传算法的基本构成，算法及其在数据挖掘中的应用；支持向量机的基本原理和实现技术。第3部分　数据挖掘建模与模型，包括第16～17章。本部分是数据挖掘建模和模型评价的基础知识。第4部分　SPSS Clementine数据挖掘实务，包括第18～24章。本部分包括SPSS Clementine的使用入门和SPSS Clementine数据挖掘项目的实现和具体实施，最后讲解了SPSS Clementine的3个典型案例。

本书可作为高等院校计算机科学与技术专业、软件工程专业或信息类等相关专业的教材，也可作为有关数据挖掘方面的培训教材，以及所有拟从事数据挖掘领域工作研究的学生、学者、工程师的参考用书。

图书在版编目（CIP）数据

数据挖掘原理与 SPSS Clementine 应用宝典 / 元昌安主编. —北京：电子工业出版社，2009.8
（宝典丛书）
ISBN 978-7-121-08601-4

Ⅰ. 数… Ⅱ. 元… Ⅲ. ①数据采集－理论②数据采集－统计分析－应用软件，SPSS Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2009）第 046857 号

责任编辑：张月萍　　特约编辑：明足群
印　　刷：北京东光印刷厂
装　　订：三河市皇庄路通装订厂
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：43.25　　字数：1250 千字
印　　次：2009 年 8 月第 1 次印刷
定　　价：88.00 元（含光盘一张）

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到 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

随着计算机技术，特别是数据库技术的快速发展和广泛应用，各行各业积累的数据量越来越大。有数据表明，进入 20 世纪 90 年代，人类积累的数据量以每月高于 15%的速度增加，如果不借助强有力的挖掘工具，仅依靠人的能力来理解这些数据是不可能的。数据挖掘应运而生。

数据挖掘技术是一门交叉学科，涉及到数据库、统计学、人工智能和机器学习等多个领域。“数据挖掘”概念最早是由 Usama Fayaad 1995 年在加拿大蒙特利尔的第一届知识发现和数据挖掘国际会议上提出的，而从数据库中发现知识（KDD）一词首次出现却早在 1989 年举行的第十一届国际联合人工智能学术会议上。在国内对数据挖掘和知识发现的研究稍晚，1993 年国家自然科学基金首次支持国内学者对该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展数据挖掘和知识发现的基础理论及其应用研究。可以说数据挖掘技术及其应用研究正呈现蓬勃发展之势，近几年，数据挖掘方面的著作也如雨后春笋一样出现在读者面前，让读者应接不暇。随着数据挖掘研究逐步走向深入，人们越来越清楚地认识到，数据挖掘的研究主要有三个技术支柱，即数据库、人工智能和统计学。也就是说，数据挖掘技术中运用了大量的数据库、人工智能和统计学中的理论基础，这些理论基础实际上已经相对成熟，它们被综合地运用到数据挖掘技术中，一篇篇学术论文发表，一本本专著出版，使得数据挖掘技术快速发展。但由于这些理论的深奥以及一般技术人员实现上的困难，使得数据挖掘技术在实践中普及性地发展和应用成为一个重大难题。直至目前为止，有关数据挖掘的专著大都停留在理论方法论的介绍。而为了推动数据挖掘技术像数据库技术一样被广泛应用于实践中，需要数据挖掘的著作在介绍原理的同时，也要介绍一些详尽的算法和解决实际问题的建模技术，同时还需要对有关数据挖掘工具应用的详细介绍。本书正是在这种背景下应运而生的。

本书从数据挖掘基础、数据挖掘经典算法、数据挖掘业务建模与模型评价、SPSS Clementine 数据挖掘实务四个方面对数据挖掘技术进行了全面介绍，既包含传统经典的数据挖掘方法，同时也包含了部分数据挖掘的最新研究成果，让读者既对数据挖掘理论有一定的认识，同时在数据挖掘经典算法的实现、针对具体应用的建模以及数据挖掘工具的应用等方面达到实战的水平。

主要内容

本书全面而细致地讲解了数据挖掘的原理、算法，以及 SPSS Clementine 数据挖掘工具应用实务。全书分为 4 部分，共 24 章。具体的篇章内容如下。

第 1 部分　数据挖掘应用基础，包括第 1~5 章。本部分是数据挖掘应用的基础部分，初学者通过本部分的学习可以了解与掌握数据挖掘的基本概念及数据挖掘应用的基本原理。内容包括数据挖掘的定义，数据挖掘的发展历史和数据挖掘技术在不同领域的应用；数据挖掘能够发现的知识模式以及相应的关键技术；数据挖掘的体系结构，以及现实中要完成一个数据挖掘项目任务时常用的

数据挖掘过程模型；数据挖掘的对象以及如何选择和构造建模数据集；数据预处理在数据挖掘过程中的重要意义，数据预处理的 4 个基本功能和数据预处理的几种方法。

第 2 部分 数据挖掘经典算法，包括第 6～15 章。本部分是数据挖掘的核心部分，是学习数据挖掘知识必须要熟练掌握和理解的内容。内容包括回归分析的基本原理以及各种回归分析的方法；贝叶斯网络的基本概念和一些常用的算法；聚类技术，重点讲解聚类分析的原理和常用的聚类算法；决策树算法的原理和常用算法，决策树的剪枝和由决策树提取分类规则的过程；关联规则的基本概念、原理以及常用算法；粗糙集基本概念，算法以及在数据挖掘中的应用；基本的神经网络模型的原理和算法；遗传算法的基本构成，算法及其在数据挖掘中的应用；支持向量机的基本原理，算法和实现技术，及其在数据挖掘中的具体应用；复杂对象的数据挖掘。

第 3 部分 数据挖掘建模与模型评价，包括第 16～17 章。本部分是数据挖掘建模和模型评价的基础知识，是学习数据挖掘必须熟悉和掌握的内容。本部分首先对数据挖掘建模进行概述。对数据挖掘建模的基本概念进行讲解。对于数据挖掘的入门者而言，掌握数据挖掘建模的相关概念也是很重要的。接下来讲解数据挖掘建模的基础知识。数据挖掘建模的基本概念和相关理论是建模的根本。此外对数据挖掘建模的基本原理进行了讲解，这样能够更好地把握数据挖掘建模。最后讲解数据挖掘模型评价的相关准则，对如何比较和评价数据挖掘模型有一个系统的研究，并能提供一些准则。

第 4 部分 SPSS Clementine 数据挖掘实务，包括第 18～24 章。本部分是应用数据挖掘工具 SPSS Clementine 进行数据挖掘的基础知识，是学习 SPSS Clementine 必须熟悉和掌握的内容，同时也是数据挖掘理论与实践的结合和运用。本部分首先对数据挖掘工具 SPSS Clementine 进行讲解。主要对 SPSS Clementine 使用入门进行讲解。接着讲解 SPSS Clementine 的数据管理，数据的图形化展示。接下来讲解 SPSS Clementine 数据挖掘建模和结果的输出。然后讲解 SPSS Clementine 数据挖掘项目的实施，包括数据挖掘项目实施步骤、数据挖掘项目周期、建立项目和报告、处理缺失值以及导入和导出 PMML 模型。最后讲解 SPSS Clementine 的 3 个典型案例。

本书特色

- 尽可能结合应用的实例，使理论和实际相结合，达到学以致用的效果。
- 从数据挖掘原理与 SPSS Clementine 实践，层层深入技术内幕。
- 本书点面兼顾，目录分类细致而科学，方便不同类型读者的快速查阅。
- 书中在介绍相关知识时，配备了大量的插图，使读者更容易阅读。
- 配套代码光盘，免去烦琐输入代码的工作，提高学习效率。此外，本书还配置了幻灯片课件，方便读者自学，也方便教学人员的备课。

读者对象

本书可作为高等院校计算机科学与技术专业、软件工程专业或信息类等相关专业的教材，也可作为有关数据挖掘技术方面的培训教材，以及所有拟从事数据挖掘领域研究的学生、学者和工程师的参考用书。

本书约定

本书的 SPSS Clementine 系统介绍以中文版 10.1 为操作界面，这是目前国内市场上最新的中文版，读者如果使用 Clementine 其他版本，其界面可能会有稍许差异。

致谢与分工

本书由元昌安主编，邓松、李文敬、刘海涛等编著。其中第 1 部分由丁超、覃晓、李文敬编写；第 2 部分由邓松、钟智、苏毅娟、彭昱忠、饶元、王艳、李文敬编写；第 3 部分由石亚冰、刘海涛编写；第 4 部分由廖剑平、李桂来、刘海涛编写；附录由蔡宏果完成。元昌安对全书进行了统稿。姚新军负责前期策划与后期质量控制。全书由成都易为科技有限责任公司审校，参与其他工作的同志还有：黄中林、王斌、张强林、王晓、万雷、李佳、王呼佳、吴艳、张赛桥、陶林、赵会春、余松、赵腾伦、虞志勇、李晓宁等。

由于时间有限，加之水平有限，书中不足之处在所难免，恳请读者批评指正。

反侵权盗版声明

目　录