

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。



电子工业出版社

Publishing House of Electronics Industry
<http://www.phei.com.cn>

元昌安 主编
邓 松 李文敬 刘海涛 等编著

TP274
Y872

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

电子工业出版社

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。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

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 zlt@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

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010)88254396；(010)88258888

传 真：(010)88254397

E - mail: dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036

目 录

第 1 部分 数据挖掘应用基础	1
第 1 章 数据挖掘概述	2
1.1 数据挖掘的社会需求	2
1.2 数据挖掘的定义	3
1.2.1 数据挖掘的技术定义	3
1.2.2 数据挖掘的商业定义	4
1.3 数据挖掘系统分类	5
1.4 数据挖掘的应用领域	6
1.4.1 金融领域	7
1.4.2 数据挖掘在营销中的应用	7
1.4.3 电子政务领域	9
1.4.4 电信领域	9
1.4.5 工业生产领域	10
1.4.6 生物与医学	11
1.5 数据挖掘标准和规范	11
1.6 数据挖掘面临的挑战和局限性	12
1.7 数据挖掘的发展趋势	14
1.7.1 Web 挖掘	14
1.7.2 空间数据挖掘	14
1.7.3 生物信息或基因的数据挖掘	14
1.8 小结	15
第 2 章 数据挖掘可挖掘的知识类型	16
2.1 概念与类描述	16
2.2 关联模式	18
2.3 分类	19
2.4 聚类分析	20
2.5 预测	21
2.6 时间序列	22
2.7 偏差检测	23
2.8 小结	23
第 3 章 数据挖掘的体系结构与模型	24
3.1 数据挖掘的体系结构	24
3.2 数据挖掘的过程模型	25
3.2.1 Fayyad 模型	25
3.2.2 CRISP-DM 模型	27

3.3 小结	33
第 4 章 数据选择	34
4.1 数据挖掘的对象	34
4.1.1 数据库	34
4.1.2 数据仓库	35
4.1.3 文本	36
4.1.4 Web 信息	36
4.1.5 空间数据	37
4.2 选择建模数据	37
4.3 构造建模数据集	38
4.4 小结	39
第 5 章 数据预处理	40
5.1 数据预处理基本功能	40
5.1.1 数据清理	40
5.1.2 数据集成	42
5.1.3 数据变换	42
5.1.4 数据归约	42
5.2 数据预处理的方法	44
5.2.1 基于粗糙集理论的简约方法	44
5.2.2 复共线性数据的预处理方法	45
5.2.3 基于 Hash 函数取样的抽样技术数据预处理	48
5.2.4 基于遗传算法的预处理方法	50
5.2.5 基于神经网络数据预处理方法	51
5.2.6 Web 挖掘数据预处理方法	53
5.3 小结	54
第 2 部分 数据挖掘经典算法	55
第 6 章 回归分析	56
6.1 回归分析的基本原理	56
6.2 一元线性回归分析	58
6.2.1 一元线性回归模型	58
6.2.2 模型参数估计和估计平均误差	62
6.2.3 回归模型的校验	64
6.3 多元线性回归分析	68
6.3.1 多元线性回归模型	69
6.3.2 参数估计	69
6.3.3 多元回归方差分析和显著性检验	71
6.3.4 多元回归模型的残差分析	74
6.4 非线性回归分析	77
6.4.1 非线性模型	77
6.4.2 非线性模型的线性化	79
6.5 应用实例分析	82

6.5.1	算法描述	82
6.5.2	实例过程	83
6.5.3	结果分析	84
6.6	小结	85
第7章	贝叶斯分析	86
7.1	贝叶斯定理	86
7.1.1	贝叶斯理论基础	86
7.1.2	贝叶斯定理	87
7.2	贝叶斯分类	88
7.2.1	贝叶斯分类步骤	88
7.2.2	先验概率和后验概率	88
7.2.3	贝叶斯分类	88
7.2.4	贝叶斯分类器	89
7.3	朴素贝叶斯分类	90
7.3.1	贝叶斯假设和朴素贝叶斯	90
7.3.2	朴素贝叶斯计算	90
7.3.3	朴素贝叶斯分类举例	91
7.3.4	朴素贝叶斯分类的特点	93
7.3.5	朴素贝叶斯网络的扩展	94
7.4	EM 算法	94
7.5	贝叶斯信念网络	95
7.5.1	贝叶斯网络结构	95
7.5.2	贝叶斯网络应用	96
7.5.3	贝叶斯网络特点	98
7.6	应用实例分析	98
7.6.1	样本数据的选取依据和方法	98
7.6.2	使用贝叶斯构造网络算法训练数据	98
7.6.3	结果与评价	100
7.7	小结	100
第8章	聚类分析	101
8.1	聚类分析原理	101
8.1.1	聚类分析基础	101
8.1.2	聚类分析中的数据类型	103
8.1.3	区间标度变量	104
8.1.4	二元变量	105
8.1.5	分类型、序数型变量	106
8.1.6	向量对象	107
8.2	聚类分析常用算法分类	108
8.2.1	划分方法	109
8.2.2	层次方法	109
8.2.3	基于密度的方法	109
8.2.4	基于网格的方法	109

8.2.5 基于模型的方法	110
8.2.6 高维数据的聚类法	110
8.2.7 模糊聚类 FCM	110
8.3 划分聚类方法	110
8.3.1 典型的划分方法: k-means, k-medoids	111
8.3.2 算法实现	112
8.4 层次聚类方法	121
8.4.1 凝聚的和分裂的层次聚类	121
8.4.2 BIRCH: 利用层次方法的平衡迭代归约和聚类	123
8.4.3 ROCK: 分类属性的层次聚类算法	124
8.4.4 CURE: 使用代表点的聚类方法	125
8.4.5 Chameleon: 利用动态建模的层次聚类	126
8.5 基于密度的聚类方法	127
8.5.1 DBSCAN: 基于高密度连通区域的聚类	127
8.5.2 OPTICS: 通过点排序识别聚类结构	129
8.5.3 DENCLUE: 基于密度分布函数的聚类	129
8.6 基于网格的聚类方法	131
8.6.1 STING: 统计信息网格聚类	131
8.6.2 WaveCluster: 利用小波变换聚类	132
8.7 基于模型的聚类方法	132
8.7.1 统计学方法 COBWEB	132
8.7.2 神经网络方法 SOMs	133
8.8 高维数据的聚类方法	135
8.8.1 CLIQUE: 维增长子空间聚类方法	135
8.8.2 PROCLUS: 维归约子空间聚类方法	136
8.9 模糊聚类 FCM	136
8.9.1 模糊集基本知识	136
8.9.2 模糊 C 均值聚类	137
8.10 应用实例分析	138
8.11 小结	146
第 9 章 决策树算法	147
9.1 决策树算法原理	147
9.2 常用决策树算法	151
9.2.1 ID3 算法	151
9.2.2 C4.5 算法	153
9.2.3 CART 算法	157
9.2.4 PUBLIC 算法	159
9.2.5 SLIQ 算法	159
9.2.6 SPRINT 算法	160
9.3 决策树剪枝	161
9.3.1 预剪枝	162
9.3.2 后剪枝	162
9.4 由决策树提取分类规则	169

9.5	应用实例分析	170
9.5.1	类别属性信息熵的计算	170
9.5.2	非类别属性信息熵的计算	170
9.5.3	递归地创建决策树的树枝和叶子	170
9.6	小结	174
第 10 章	关联规则算法	176
10.1	关联规则基础	176
10.1.1	关联规则定义	176
10.1.2	关联规则分类	177
10.2	关联规则算法原理	178
10.2.1	关联规则挖掘算法的步骤	178
10.2.2	基本关联规则算法	178
10.2.3	复杂关联规则算法	181
10.3	分层搜索经典算法——Apriori 算法	181
10.3.1	频繁项目集的产生	182
10.3.2	产生关联规则	185
10.3.3	Apriori 算法性能分析	186
10.3.4	Apriori 算法改进	186
10.4	并行挖掘算法	187
10.4.1	并行算法思想	187
10.4.2	基于 Apriori 的并行算法	188
10.5	增量更新挖掘算法	190
10.5.1	增量挖掘	190
10.5.2	FUP 算法	191
10.6	多层关联规则挖掘	194
10.6.1	概念层次	194
10.6.2	多层关联规则挖掘方法	195
10.6.3	多层关联规则的冗余	197
10.7	约束性关联规则挖掘	197
10.7.1	数据挖掘中约束的作用	198
10.7.2	约束的类型	199
10.7.3	过滤事务数据库	200
10.7.4	算法 Separate	202
10.7.5	扩展的约束条件	203
10.7.6	时态约束关联规则挖掘	204
10.8	数量关联规则挖掘	205
10.8.1	数量关联规则挖掘问题	205
10.8.2	数量关联规则的分类	205
10.8.3	数量关联规则挖掘的步骤	206
10.8.4	数值属性离散化及算法	207
10.9	多维关联规则挖掘	208
10.9.1	多维关联规则挖掘原理	208
10.9.2	MAQA 算法	209

10.9.3 确定多属性划分的聚类算法 CP	210
10.9.4 合并数量属性的相邻值	212
10.10 负关联规则挖掘算法	213
10.10.1 直接 Apriori 算法	213
10.10.2 “近似”负关联规则算法	214
10.11 加权关联规则挖掘算法	215
10.11.1 加权关联规则模型	215
10.11.2 加权关联规则发现算法——MINWAL(O)算法	215
10.12 应用实例分析	218
10.12.1 数据准备	218
10.12.2 挖掘关联规则	219
10.12.3 挖掘结果分析	220
10.13 小结	220
第 11 章 粗糙集理论	221
11.1 粗糙集基本概念	221
11.1.1 知识和知识库	221
11.1.2 不可分辨关系	222
11.1.3 上、下近似集	222
11.2 知识表达	223
11.2.1 知识表达系统	223
11.2.2 决策表	224
11.2.3 属性约简、核集的求取	225
11.2.4 属性值约简	225
11.2.5 决策规则	226
11.2.6 基于可辨识矩阵属性约简算法	226
11.2.7 信息熵的属性约简	227
11.3 粗糙集在数据预处理中的应用	228
11.3.1 属性约简的两种方法	228
11.3.2 粗糙集在神经网络中的应用——粗神经网络算法	231
11.4 小结	233
第 12 章 神经网络	234
12.1 神经网络基本原理	234
12.1.1 人工神经元模型	234
12.1.2 人工神经网络模型	235
12.1.3 神经网络的参数	236
12.1.4 神经网络的学习方法	237
12.2 BP 神经网络	239
12.2.1 BP 神经网络模型	239
12.2.2 BP 神经网络的 Java 实现	240
12.2.3 BP 神经网络的改进	247
12.3 径向基函数神经网络	251
12.3.1 RBF 神经网络结构	251

12.3.2	RBF 训练	252
12.3.3	RBF 神经网络算法分析	255
12.3.4	RBF 网络的应用	257
12.4	Hopfield 神经网络	258
12.4.1	Hopfield 神经网络概述	258
12.4.2	离散 Hopfield 神经网络	259
12.4.3	连续 Hopfield 神经网络	259
12.5	自组织神经网络	260
12.5.1	SOFM 网络模型	260
12.5.2	SOFM 网络聚类的基本算法	261
12.5.3	SOFM 算法分析	261
12.6	神经网络的应用	262
12.6.1	BP 神经网络在模式识别中的应用	262
12.6.2	基于 Hopfield 神经网络在优化问题中的应用	264
12.7	神经网络在数据挖掘中的应用	265
12.7.1	基于神经网络方法的数据挖掘过程	266
12.7.2	评价数据挖掘模型实现算法的指标	266
12.8	小结	267
第 13 章	遗传算法	268
13.1	遗传算法概述	268
13.1.1	遗传算法的基本理论	268
13.1.2	遗传算法的基本操作	271
13.1.3	遗传算法的编码方式	272
13.1.4	遗传算法的类型	273
13.2	基本遗传算法	273
13.2.1	基本遗传算法的流程	273
13.2.2	基本遗传算法的 Java 实现	277
13.3	改进遗传算法	287
13.3.1	分层遗传算法	287
13.3.2	自适应遗传算法	289
13.3.3	小生境遗传算法	290
13.3.4	并行遗传算法	292
13.3.5	混合遗传算法	294
13.4	基于遗传算法的数据挖掘	297
13.4.1	遗传算法的一般结构	297
13.4.2	遗传算法的组成要素	298
13.4.3	基于遗传算法的关联规则挖掘	299
13.4.4	基于遗传算法的聚类算法	300
13.4.5	基于遗传算法的分类算法	303
13.4.6	基于模糊遗传算法的建模	305
13.5	基因表达式编程	307
13.5.1	基因表达式编程国内外研究现状	307
13.5.2	基因表达式编程算法描述	307

13.5.3 基因表达式编程的主要遗传操作	308
13.6 小结	310
第 14 章 支持向量机	311
14.1 支持向量机基础	311
14.1.1 机器学习的基本问题	311
14.1.2 经验风险最小化问题	312
14.1.3 VC 维与学习一致性理论	313
14.1.4 结构化风险最小化	315
14.2 支持向量机的基本原理	317
14.2.1 线性支持向量机	317
14.2.2 广义线性支持向量机	320
14.2.3 非线性支持向量机	322
14.3 支持向量机的实现技术	326
14.3.1 chunking 块算法	326
14.3.2 Decomposing 算法	328
14.3.3 SMO 算法	330
14.3.4 SMO 算法源代码	331
14.3.5 SMO 算法的特点和优势	341
14.4 支持向量回归机	341
14.4.1 不敏感损失函数	342
14.4.2 支持向量回归机 (SVR) 模型	343
14.5 支持向量机的改进算法	345
14.5.1 V-SVM 算法	345
14.5.2 One-class SVM 算法	346
14.5.3 RSVM 算法	347
14.5.4 LS-SVM 算法	347
14.5.5 WSVM 算法	348
14.5.6 模糊支持向量机算法 (FSVM)	348
14.5.7 多类值支持向量机算法	349
14.6 支持向量机在数据挖掘中的应用	352
14.6.1 支持向量机在医疗诊断中的应用	353
14.6.2 支持向量机时间序列预测模型	354
14.7 小结	355
第 15 章 复杂对象数据挖掘	356
15.1 空间数据库挖掘	356
15.1.1 空间数据概述	356
15.1.2 空间数据挖掘中的基础计算模型	358
15.1.3 空间数据挖掘基础	363
15.1.4 几种空间数据挖掘算法	365
15.2 多媒体数据挖掘	368
15.2.1 多媒体数据挖掘概述	369
15.2.2 多媒体数据挖掘方法	371

15.3	文本挖掘	373
15.3.1	文本挖掘概述	374
15.3.2	文本的预处理	375
15.3.3	文本挖掘方法	377
15.4	挖掘互联网	380
15.4.1	挖掘 Web 页面布局结构	381
15.4.2	挖掘 Web 链接结构识别权威 Web 页面	382
15.4.3	挖掘 Web 上的多媒体数据	383
15.4.4	Web 文档的自动分类	384
15.4.5	Web 使用挖掘	384
15.5	挖掘数据流	386
15.5.1	流数据处理方法和流数据系统	386
15.5.2	流 OLAP 和流数据立方体	388
15.5.3	数据流中的频繁模式挖掘	389
15.5.4	动态数据流的分类	390
15.5.5	聚类演变数据流	391
15.6	时间序列数据挖掘	393
15.6.1	趋势分析	393
15.6.2	时间序列分析中的相似性搜索	395
15.7	挖掘事务数据库中的序列模式	396
15.7.1	序列模式挖掘	396
15.7.2	挖掘序列模式的可伸缩方法	398
15.7.3	基于约束的序列模式挖掘	399
15.7.4	时间相关序列数据的周期性分析	400
15.8	挖掘生物学数据中的序列模式	401
15.8.1	生物学序列比对	402
15.8.2	生物学序列分析	403
15.9	小结	409
第 3 部分 数据挖掘建模与模型		411
第 16 章 数据挖掘建模		412
16.1	数据挖掘建模概述	412
16.1.1	原型与模型	412
16.1.2	模式与模型	413
16.1.3	知识层次理论	413
16.1.4	模型与数据	416
16.1.5	知识结构与框架	416
16.1.6	认识决策	417
16.2	数据挖掘建模基础	419
16.2.1	数据挖掘建模	420
16.2.2	建模与挖掘的结合	423
16.2.3	模型分类	427
16.2.4	建模行为	430

16.3 数据挖掘建模原理	432
16.3.1 建模要求	432
16.3.2 建模原则	432
16.3.3 简化模型	433
16.3.4 建模流程	434
16.3.5 建模素质	439
16.4 小结	440
第 17 章 数据挖掘模型评价	442
17.1 基于损失函数的标准	442
17.1.1 混淆矩阵	442
17.1.2 准确率及误差的度量	443
17.1.3 两个评价模型成本的可视化工具	445
17.1.4 评估分类器的准确率	447
17.2 基于统计检验的准则	449
17.2.1 统计模型之间的距离	449
17.2.2 统计模型的离差	451
17.3 基于计分函数的标准	453
17.4 贝叶斯标准	454
17.5 计算标准	455
17.5.1 交叉验证标准	455
17.5.2 自展标准	456
17.5.3 遗传算法	460
17.6 小结	461
第 4 部分 SPSS Clementine 数据挖掘实务	463
第 18 章 SPSS Clementine 基础	464
18.1 认识 SPSS Clementine	464
18.1.1 SPSS Clementine 运行方式	465
18.1.2 Clementine 的组成构件	466
18.1.3 SPSS Clementine 选项设置	470
18.2 SPSS Clementine 应用领域	474
18.3 SPSS Clementine 数据挖掘入门	475
18.3.1 SPSS Clementine 中鼠标以及快捷键的使用	475
18.3.2 SPSS Clementine 中构建数据流	476
18.3.3 数据流中节点的设置	476
18.3.4 对数据流的设置和操作	481
18.4 小结	484
第 19 章 SPSS Clementine 数据管理	485
19.1 各种格式数据的导入	485
19.1.1 从开放数据库中导入数据	486
19.1.2 从无格式文本文件中读取数据	489
19.1.3 从固定字段的文本文件中读取数据	490