

目 录

第 1 章 数据仓库的概念与体系结构	(1)
1.1 数据仓库的兴起	(1)
1.1.1 数据管理技术的发展	(1)
1.1.2 数据仓库的萌芽	(3)
1.2 数据仓库的基本概念	(4)
1.2.1 元数据	(4)
1.2.2 数据粒度	(5)
1.2.3 数据模型	(5)
1.2.4 ETL	(6)
1.2.5 数据集市	(7)
1.3 数据仓库的特点与组成	(8)
1.3.1 数据仓库的特点	(8)
1.3.2 数据仓库的组成	(11)
1.4 数据仓库的体系结构	(15)
1.4.1 传统的数据仓库体系结构	(15)
1.4.2 传统数据仓库系统在大数据时代所面临的挑战	(16)
1.4.3 大数据时代的数据仓库	(20)
习 题	(23)
第 2 章 数据	(24)
2.1 数据的概念与内容	(24)
2.2 数据属性与数据集	(28)
2.3 数据预处理	(29)
2.3.1 数据预处理概述	(30)
2.3.2 数据清洗	(31)
2.3.3 数据集成	(35)
2.3.4 数据变换	(38)
2.3.5 数据归约	(39)
习 题	(47)
第 3 章 数据存储	(49)
3.1 数据仓库的数据模型	(49)

3.1.1	数据仓库的概念模型	(50)
3.1.2	数据仓库的逻辑模型	(52)
3.1.3	数据仓库的物理模型	(54)
3.2	元数据存储	(55)
3.2.1	元数据的概念	(55)
3.2.2	元数据的分类方法	(55)
3.2.3	元数据的管理	(57)
3.2.4	元数据的作用	(58)
3.3	数据集市	(59)
3.3.1	数据集市的概念	(59)
3.3.2	数据集市的类型	(60)
3.3.3	数据集市的建立	(60)
3.4	大数据存储技术	(61)
3.4.1	大数据的概念	(61)
3.4.2	传统数据库的局限	(62)
3.4.3	NoSQL 数据库	(63)
3.4.4	几种主流的 NoSQL 数据库	(64)
	习 题	(64)
第 4 章	OLAP 与数据立方体	(65)
4.1	OLAP 的概念	(65)
4.1.1	OLAP 的定义	(65)
4.1.2	OLAP 的准则	(66)
4.1.3	OLAP 的特征	(69)
4.2	多维分析的基本分析动作	(70)
4.2.1	切片	(70)
4.2.2	切块	(71)
4.2.3	钻取	(72)
4.2.4	旋转	(72)
4.3	OLAP 的数据模型	(73)
4.3.1	ROLAP 数据模型	(73)
4.3.2	MOLAP 数据模型	(75)
4.3.3	MOLAP 和 ROLAP 的数据组织与应用比较	(76)

4.3.4	HOLAP 数据模型	(77)
4.4	数据立方体的基本概念	(78)
4.4.1	数据立方体中的一些概念	(78)
4.4.2	数据立方体计算的一般策略	(79)
4.5	数据立方体的计算方法	(80)
4.5.1	多路数组策略计算完全立方体	(80)
4.5.2	从顶点方体向下计算冰山立方体	(80)
4.5.3	使用动态星树结构计算冰山立方体	(81)
4.5.4	快速高维 OLAP 预计算壳片段	(82)
	习题	(83)

第 5 章 数据挖掘基础 (84)

5.1	数据挖掘的兴起	(84)
5.1.1	数据挖掘的发展历程	(84)
5.1.2	数据挖掘的概述	(85)
5.1.3	大规模数据挖掘	(86)
5.2	数据挖掘的任务	(87)
5.2.1	关联规则	(87)
5.2.2	聚类分析	(88)
5.2.3	分类分析	(89)
5.2.4	回归分析	(90)
5.2.5	相关分析	(91)
5.2.6	异常检测	(92)
5.3	数据挖掘的流程	(92)
5.3.1	数据挖掘对象	(92)
5.3.2	数据挖掘分类	(93)
5.3.3	知识发现的过程	(94)
	习题	(96)

第 6 章 关联挖掘 (97)

6.1	关联规则的概念和分类	(97)
6.1.1	关联规则的概念	(97)
6.1.2	关联规则的分类	(99)
6.2	Apriori 算法	(100)

6.2.1	Apriori 算法概述	(100)
6.2.2	Apriori 算法的性质与步骤	(100)
6.2.3	Apriori 算法的实例	(101)
6.2.4	从频繁项集产生关联规则	(103)
6.3	FP-Growth 算法	(104)
6.3.1	FP-tree 的建立	(105)
6.3.2	FP-tree 上挖掘关联规则	(106)
6.4	挖掘算法的进阶算法	(107)
	习题	(110)
第7章	聚类分析	(112)
7.1	聚类分析概述	(112)
7.1.1	聚类分析的定义	(112)
7.1.2	聚类分析的分类	(113)
7.2	差异度的计算方法	(114)
7.2.1	聚类算法中的数据结构	(114)
7.2.2	区间标度变量的差异度计算	(115)
7.2.3	二元变量的差异度计算	(116)
7.2.4	标称型变量的差异度计算	(117)
7.2.5	序数型变量的差异度计算	(118)
7.2.6	比例标度型变量的差异度计算	(119)
7.2.7	混合类型变量的差异度计算	(119)
7.3	基于分割的聚类方法	(120)
7.3.1	分割聚类方法的描述	(120)
7.3.2	K-means 均值算法	(121)
7.3.3	PAM 算法	(122)
7.3.4	CLARA 算法和 CLARANS 算法	(125)
7.4	基于密度的聚类方法	(126)
7.4.1	基于密度的聚类方法描述	(126)
7.4.2	DBSCAN 算法	(127)
7.4.3	OPTICS 算法	(129)
7.5	谱聚类方法	(130)
7.5.1	谱聚类描述	(130)

7.5.2 谱聚类算法描述	(131)
7.5.3 谱聚类实例	(132)
7.6 ICA 聚类分析	(133)
7.6.1 ICA 的起源和目的	(133)
7.6.2 ICA 模型和应用要求	(133)
7.6.3 ICA 应用场合	(135)
习题	(135)
第8章 分类	(137)
8.1 分类的基本知识	(137)
8.1.1 分类的概念	(137)
8.1.2 分类的评价标准	(138)
8.1.3 分类的主要方法	(138)
8.2 决策树分类	(139)
8.2.1 决策树算法概述	(139)
8.2.2 决策树的生成	(141)
8.2.3 决策树中规则的提取	(142)
8.2.4 ID3 算法	(143)
8.2.5 C4.5 算法	(145)
8.2.6 蒙特卡罗树搜索 (MCTS) 算法	(146)
8.3 SVM 预测	(147)
8.3.1 线性可分的 SVM	(147)
8.3.2 线性不可分的 SVM	(150)
8.3.3 SVM 的实现——手写数字图片的识别	(153)
8.4 KNN 算法	(154)
8.4.1 KNN 算法的描述	(155)
8.4.2 KNN 算法的实现	(156)
习题	(157)
第9章 神经网络	(159)
9.1 神经网络概述与定义	(159)
9.1.1 神经网络概述	(159)
9.1.2 神经网络的学习过程	(160)
9.2 限制玻尔兹曼机 (RBM)	(161)

9.2.1	RBM 的定义	(161)
9.2.2	RBM 的能量模型与学习方法	(162)
9.3	深度信念网络	(165)
9.3.1	DBN 反向传播算法介绍与改进	(165)
9.3.2	DNN 分类与代价函数选择	(170)
9.4	卷积神经网络 (CNN)	(173)
9.4.1	卷积神经网络定义与结构	(173)
9.4.2	CNN 两个特点与图形实例	(176)
9.5	循环神经网络 (RNN)	(179)
9.5.1	RNN 概述	(180)
9.5.2	RNN 训练	(181)
9.5.3	LSTMs 网络与函数展示图例	(182)
	习题	(186)
第 10 章	统计分析	(188)
10.1	回归分析	(188)
10.1.1	一元线性回归	(188)
10.1.2	多元线性回归	(191)
10.1.3	非线性回归	(193)
10.2	EM 算法	(194)
10.2.1	EM 算法的引入	(194)
10.2.2	EM 算法的推导	(196)
10.2.3	EM 算法的收敛性	(197)
10.3	Bayes 分类	(199)
10.3.1	Bayes 定理	(199)
10.3.2	简单 Bayes 分类	(200)
10.3.3	Bayes 信念网络	(201)
10.3.4	Bayes 网络的应用	(203)
	习题	(203)
第 11 章	非结构化数据挖掘	(204)
11.1	文本数据挖掘	(204)
11.1.1	文本数据挖掘的概念	(204)

11.1.2 文本数据挖掘技术	(208)
11.2 Web 数据挖掘	(214)
11.2.1 Web 数据挖掘的概念	(215)
11.2.2 Web 数据挖掘的分类	(216)
11.2.3 Web 数据挖掘的应用	(220)
11.3 多媒体数据挖掘	(221)
11.3.1 多媒体数据挖掘的概念	(222)
11.3.2 多媒体数据挖掘的分类	(223)
习题	(225)
第 12 章 知识图谱	(227)
12.1 知识图谱构建	(227)
12.1.1 知识图谱的概述	(227)
12.1.2 知识图谱的数据来源	(229)
12.1.3 多源异构数据的融合	(231)
12.1.4 知识图谱的表示	(232)
12.2 知识图谱技术	(233)
12.2.1 实体抽取	(234)
12.2.2 关系抽取	(235)
12.2.3 知识推理	(236)
12.3 知识图谱的典型应用	(238)
12.3.1 查询理解	(238)
12.3.2 自动问答	(240)
12.3.3 前景和挑战	(240)
习题	(241)
第 13 章 大数据挖掘算法	(242)
13.1 Hadoop 介绍	(242)
13.1.1 Hadoop 的基本概念	(242)
13.1.2 Hadoop 的基本组件	(244)
13.2 基于 MapReduce 数据挖掘算法	(247)
13.2.1 基于 MapReduce 的 K-means 并行算法	(248)
13.2.2 基于 MapReduce 的分类算法	(251)
13.2.3 基于 MapReduce 的序列模式挖掘算法	(253)

习题	(255)
参考文献	(256)