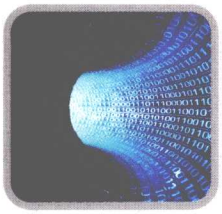
SHUJU WAJUE
SUANFA YU YINGYONG

◎ 黄添强 著



数据挖掘

算法与应用

数据挖掘算法与应用

黄添强 著

厦门大学出版社 XIAMEN UNIVERSITY PRESS
国家一级出版社
全国百佳图书出版单位

图书在版编目(CIP)数据

数据挖掘算法与应用/黄添强著. —厦门:厦门大学出版社,2011.11
ISBN 978-7-5615-4004-6

Ⅰ.①数… Ⅱ.①黄… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2011)第 164709 号

厦门大学出版社出版发行
(地址:厦门市软件园二期望海路 39 号 邮编:361008)
http://www.xmupress.com
xmup @ public.xm.fj.cn
厦门集大印刷厂印刷
2011 年 11 月第 1 版 2011 年 11 月第 1 次印刷
开本:787×1092 1/16 印张:14 插页:2
字数:358 千字 印数:1～2 000 册
定价:30.00 元
本书如有印装质量问题请直接寄承印厂调换

内容提要

数据挖掘是近年来计算机科学中最活跃的研究分支之一。本书分三部分介绍数据挖掘中的三个热点：空间数据挖掘、半监督学习与流形学习，并分别介绍了这三个研究热点的研究背景、研究现状、存在问题、最新算法与应用等。本书在论述这三个研究热点的研究现状并分析了存在的问题后，面向实际需要，提出最新的模型、算法与技术。这些模型与算法以作者研究成果为基础，具有一定的创新性与实际应用价值。这些成果可帮助广大研究工作者与工程技术人员拓展思路，并为数据挖掘的理论应用提供借鉴。

本书可供数据挖掘、机器学习及相关专业的研究人员、教师、研究生和工程人员参考。

前 言

数据挖掘技术是一门从数据中发现隐含的、先前不知道的、潜在有用的信息之技术，也是一门多学科交叉的技术，它涉及数据库、统计学、人工智能、机器学习、神经网络、模式识别、知识库系统、信息检索、高性能计算和可视化等多个领域。同时，它又是一门实用的学科，一出现即被许多领域与部门所应用。目前已经在自然科学、农业科学、医学科学、工程技术科学、经济学、社会学甚至文学、艺术学研究中得到了广泛的应用。

经过十几年的发展与积累，数据挖掘从多个学科吸收了营养，产生了许多高效又实用的算法；在许多应用领域开展了广泛的应用，并取得了大量的成果。空间数据挖掘、流形学习以及半监督学习是近几年学术界研究的热点。作者多年从事数据挖掘研究，尤其在空间数据挖掘、流形学习与半监督学习的研究中倾注了大量的心血。本书是在作者多年来对这三个研究领域进行深入的研究，并吸纳了国内外许多优秀成果的基础上撰写而成。同时，本书也是作者及其团队十几年来大部分研究工作的总结。

全书包括三部分，共 17 章。这三部分内容包括它们的研究背景、研究现状、存在问题、作者的研究成果及其典型应用。

第一部分包括 1～7 章，介绍空间数据挖掘技术。空间数据挖掘是指从空间数据中抽取隐含的知识、空间关系或非显式地存储在空间数据中的其他模式。它可以用来理解或重组空间数据、发现空间和非空间数据间的关系、构建空间知识库、优化查询等。第 1 章介绍空间数据挖掘技术的研究背景、意义、方法以及目前国内外研究现状。第 2 章介绍空间多维位置相关规则挖掘算法。第 3 章介绍基于多代表点的空间聚类算法。第 4、5、6 章介绍三种不同的空间离群点挖掘算法。第 7 章介绍移动对象轨迹的双重插值算法。

第二部分包括 8～13 章，介绍流形学习技术。流形学习是从高维采样数据中恢复低维流形结构，即找到高维空间中的低维流形，并求出相应的嵌入映射，以实现维数约简或者数据可视化。它可以发现物质的内在本质，揭示数据的内在结构，有效实现数据的非线性降维或可视化。第 8 章介绍流形学习背景、意义与研究现状。第 9 章介绍流形学习的基本概念，几个经典的流形学习算法以及它们存在的不足。第 10 章介绍基于局部相关维度的噪音流形学习算法。第 11 章介绍基于调和平均测地线核的流形学习算法。第 12 章介绍有监督的噪音流形学习算法及其在拉曼光谱数据上的应用。第 13 章介绍共享近邻的噪音流形学习算法及其在医院绩效考核中的应用。

第三部分包括 14～17 章，介绍半监督学习技术。半监督学习是一门利用少量的标注样本和大量的未标注样本进行训练和学习的技术。半监督学习对于减少标注代价，提高学习机器性能具有非常重大的实际意义。第 14 章介绍半监督学习研究的背景、意义、内容与现状等。第 15 章通过分析协同训练算法在学习过程中引入噪声数据的负面影响，介

绍一种基于监督聚类的半监督分类算法。第16章通过分析和研究已有的经典算法在处理复杂数据方面的不足，介绍一种面向复杂数据聚类的半监督聚类算法。第17章针对已有的移动对象异常轨迹检测算法受应用背景和人为因素影响大，并且算法不稳定，可扩展性不好的问题，介绍一种基于半监督技术的异常轨迹检测算法。

作者要感谢参与本书整理、润色与校订工作的所有人员。首先，要感谢笔者的研究生余养强、李凯、曾文赋与卓飞豹！他们参与了部分内容的整理、修改或撰写工作。其中，余养强参与本书第三篇的部分内容整理与撰写；李凯参与本书第二篇的部分内容整理与撰写。其次，要感谢研究生陈智文、袁秀娟、吴铁浩、苏立超与李富贵等人！他们多次细心为本书进行校订。最后，要感谢福建省高校服务海西建设重点项目——基于数学的信息化技术研究(2008HX200941-4-5)、国家自然科学基金面上项目——基于智能技术的视频篡改取证研究(61070062)提供出版资助。

本书可以作为计算机专业研究生、高年级本科的教材或供工程技术人员参考。

由于作者的研究水平与实践经验有限，加上出书过程时间仓促，书中的疏漏、错误与不妥之处在所难免，欢迎各位专家、学者及广大读者批评指正！

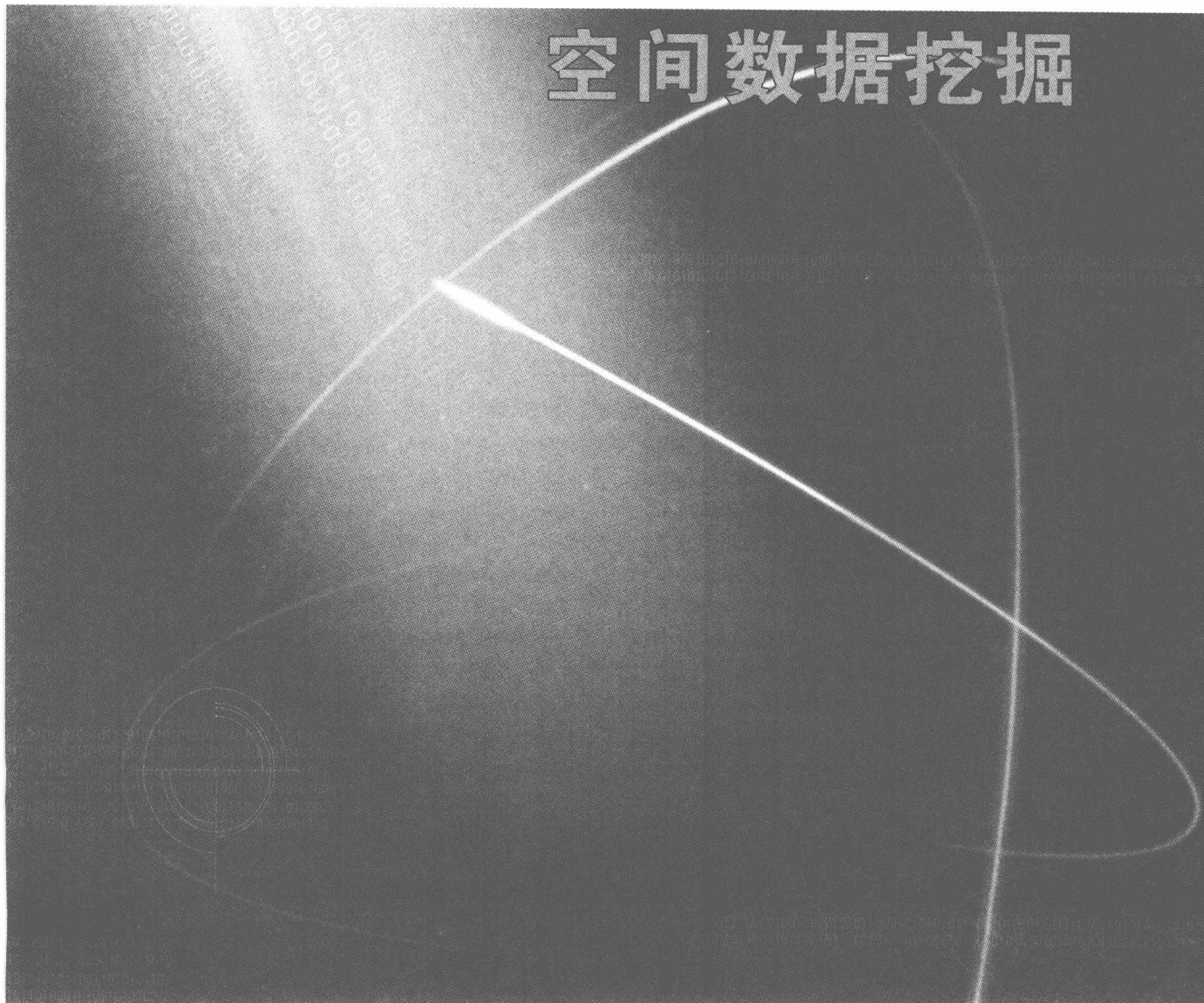
作者

2011年10月

目 录

第一篇

空间数据挖掘



第 1 章　空间数据挖掘研究绪论

一、空间数据挖掘研究背景及意义

空间数据是具有空间或位置分量的数据，它可以被看作定位于一定物理空间的数据对象[1]. 空间属性的表示可以通过特定的位置属性来实现，如地址、经纬度，或是隐含在基于位置的数据库划分中. 对空间数据的查询可以通过空间运算符的查询来进行，如“包含”、“相交”、“东”、“南”、“附近”等. 由于空间数据本身固有的特点，空间数据常常使用特定的数据结构来存储，如四叉树[2]、k-d 树[3]、R^*-树[4]等.

空间数据挖掘(spatial data mining, SDM)是指从空间数据中抽取隐含的知识、空间关系或非显式地存储在空间数据中的其他模式[5]. 它可以用来理解或重组空间数据，发现空间和非空间数据间的关系，构建空间知识库，优化查询等.

空间数据挖掘技术的产生与发展来自两大推动力. 第一，由于数据挖掘研究领域的不断拓展，由最初的关系数据和事务数据挖掘，发展到对空间数据的挖掘. 空间信息正在逐步成为各种信息系统的主体和基础. 空间数据是一类重要、特殊的数据，有着比一般关系数据库和事务数据库更加丰富和复杂的语义信息，包含着更丰富的知识. 因此，尽管数据挖掘最初产生于关系数据库和事务数据库，但由于空间数据的特殊性，从空间数据库中发掘知识很快引起了数据挖掘研究者的关注. 许多数据挖掘方面的研究工作也从关系型和事务型数据库扩展到空间数据库. 第二，大量的数据通过传感器和其他的数据采集设备源源不断地收集. 在地学领域，随着卫星和遥感技术的广泛应用，空间和非空间数据收集和存储日益丰富. 在医疗领域，各种医疗成像大量产生. 在交通领域，各种传感器监测数据不断传入数据收集中心. 随着数字城市实施，各种空间数据大量产生. 海量的空间数据在某种意义上已经超过了人们的处理能力，传统的空间分析难以胜任从这些海量的数据中提取和发现空间知识. 数据挖掘与知识发现的出现很好地满足了空间数据处理的需要，推动了数据挖掘的技术在空间数据中的应用，促使空间数据挖掘产生与发展.

空间数据挖掘具有广阔的发展前景. 大量的空间数据从遥感、地理信息系统、多媒体系统、医学影像等多种应用的传感器或其他数据采集设备中收集出来，这些数据中隐含了大量的知识. 例如，NASA 的对地观测系统(earth observing system, EOS)每天都产生 1 TB的数据[6]. 卫星图像是以地球表面为参考框架，但它不是唯一的参考框架. 一块芯片可以是参考框架，整个人体在医疗图像中也可作为参考框架，甚至超市交易中包含邮政编

码的一条记录都可以认为是空间信息.有人认为80%的数据与空间信息与有关.这些收集到的数据远远超过了人脑的分析能力,而现代社会要求人们能够获取包含空间信息的知识.空间数据挖掘为此提供了途径.这种研究有助于发现有价值的与未知的知识,在交通、生态、公共安全、公众健康、气候、基于位置的服务等领域有广泛的用途.

二、空间数据挖掘与经典数据挖掘的区别

由于空间数据与传统数据存在区别,空间数据与传统数据存储方式不同,空间数据挖掘所查找的模式与经典数据挖掘所查找的模式不同,决定了空间数据挖掘与经典数据挖掘的差异.

(1)空间数据比经典数据复杂.空间数据包含扩展的对象,如点、线、面等.有着比一般关系数据库和事务数据库更加丰富和复杂的语义信息,包含着更丰富的知识.空间数据包括非空间属性与空间属性.非空间属性描述非空间特征,包括地名、人口、温度、气压等一切可用的数值属性;空间属性用来描述空间特征,包括经纬度、实体形状、空间方位关系等.

(2)空间数据存在于连续的数据空间,而经典数据通常是离散的.

(3)空间模式是基于局部的,而传统的模式是基于全局的.

(4)空间关系是不明确的,如叠加、相交等拓扑关系,东、西、南、北等方位关系.经典数据关系是明确的,如数学关系、逻辑关系等.

(5)空间数据存在空间自相关.空间对象的观察不是独立的,空间特征存在空间自相关(autocorrelation)[7],地理学家把它作为第一定律:每件事物都与其他事物相关,但邻近事物间的相关性比距离较远的事物的相关性要大得多[8].经典的数据分析通常假设数据采样是独立的,但在空间数据中是不成立的.

(6)空间数据是海量的.空间数据库中存储的数据具有多源、多维、时态性的特点.数据的多源是指数据来源多种多样,数据格式也不尽相同,可以是遥感、图形、声音、视频和文本数据等.数据是多维的.一个城市级的GIS系统,其数据量一般可达到GB的数据量级.例如沈阳市1∶500的基础地图就有2.4 GB[9].美国建造的"暗物质望远镜",每天的观测数据高达18 TB[10].

(7)空间数据存储与存取的复杂性.空间对象用空间数据类型和对象的空间关系表示;空间数据库有许多不同于关系数据库的特性.如空间数据库中的拓扑或距离信息,通常以复杂的空间多维索引结构组织,通过空间数据存取方法存取,常常需要空间推理、几何计算和空间知识表示技术.

这些特性决定了空间数据挖掘比传统的数据挖掘更加复杂,更具有挑战性.

三、空间数据挖掘技术的主要方法及特点

常用的空间数据挖掘技术包括空间数据概化、空间规则挖掘、空间分类、空间趋势预测、空间聚类、空间离群点查找等.

(一)空间数据概化

空间数据概化包含两个方面:一般化与特殊化.一般化是指由低层面的信息派生出高层面的信息的过程.空间数据概念的层次分为空间的与非空间的.当一般化空间数据时,非空间数据需要作适当的修改以反映与新的空间区域关联的非空间数据;同样,一般化非空间数据时,空间数据也需要作适当的修改.故可以把空间数据的一般化分为:空间数据主导的一般化与非空间数据主导的一般化[11].空间数据的特殊化是指基于空间关系的层次结构中渐进求精的过程.与空间相关的应用中存在大量的数据,在找到更精确的答案之前可以用近似的答案代替.最小边界矩形法就是用对象的近似外形来代替实际外形的一种方法.大多数空间索引技术,如四叉树、R-树等,都用到了渐进求精的技术.

(二)空间规则挖掘

特征化是对数据库或它的子集进行描述的过程[12].空间规则是一种特殊的特征化[13].空间数据挖掘中有三类空间规则:空间特征规则、空间判别规则、空间关联规则.空间数据挖掘中最经常需要挖掘的一类知识是空间关联规则.Koperski 与 Han[5]最早把关联规则挖掘的技术用到空间数据库中.空间关联规则是关于空间对象的规则,是形如 $P_1 \wedge \cdots \wedge P_m \rightarrow Q_1 \wedge \cdots \wedge Q_n(c\%)$的规则,这里 $P_1 \cdots P_m$, $Q_1 \cdots Q_n$ 至少有一个谓词是空间谓词,$c\%$为规则的置信度,表示有 $c\%$满足前项的对象产生规则的结果.空间关联规则与传统关联规则的不同点在于:要分析的数据不再是一组事务的集合,而是空间对象的集合.另一种空间规则是空间同位规则,同位规则试图将关联规则泛化为空间索引的点对象集.在这种规则的查找中,用邻域的概念代替事务的概念,最终得出点对象的属性关联.

(三)空间分类

分类在数据挖掘中是一项非常重要的任务,目前在商业上应用最多.分类的目的是学会一个分类函数或分类模型,该模型能把数据库中的数据项映射到给定类别中的某一个.空间分类是对空间对象进行划分.可以用空间属性、非空间属性或两者对空间对象进行划分.

Ester 等[14]提出基于 ID3[15]算法的空间分类算法.该方法利用近邻图分类.近邻图是由空间对象构造的一种图.每个对象对应图中的一个结点,依据这个对象的近邻来构造