<<数据挖掘>>

图书基本信息

书名：<<数据挖掘>>

13位ISBN编号：9787302307143

10位ISBN编号：7302307148

出版时间：2013-1

出版时间：清华大学出版社

作者：坎塔尔季奇

译者：王晓海,吴志刚

<<数据挖掘>>

内容概要

随着数据集规模和复杂度的持续上升，分析员必须利用更高级的软件工具来执行间接的、自动的智能化数据分析。
《数据挖掘：概念、模型、方法和算法(第2版)》介绍了通过分析高维数据空间中的海量原始数据来提取用于决策的新信息的尖端技术和方法。

本书开篇阐述数据挖掘原理，此后在示例的引导下详细讲解起源于统计学、机器学习、神经网络、模糊逻辑和演化计算等学科的具有代表性的、最前沿的挖掘方法和算法。
本书还着重描述如何恰当地选择方法和数据分析软件并合理地调整参数。
每章末尾附有复习题。

本书主要用作计算机科学、计算机工程和计算机信息系统专业的研究生数据挖掘教材，高年级本科生或具备同等教育背景的读者也完全可以理解本书的所有主题。

◆ 介绍支持向量机(SVM)和Kohonen映射
◆ 讲解DBSCAN、BIRCH和分布式DBSCAN聚类算法
◆ 介绍贝叶斯网络，讨论图形中的Betweeness和Centrality参数测量算法
◆ 分析在建立决策树时使用的CART算法和基尼指数
◆ 介绍Bagging & Boosting集成学习方法，并详述AdaBoost算法
◆ 讨论Relief以及PageRank算法
◆ 讨论文本挖掘的潜在语义分析(LSA)，并分析如何测定文本文档之间的语义相似性
◆ 讲解时态、空间、Web、文本、并行和分布式数据挖掘等新主题
◆ 更详细地讲解数据挖掘技术商业、隐私、安全和法律方面的内容

作者简介

作者：（美）坎塔尔季奇（Mehmed Kantardzic） 译者：王晓海 吴志刚 王晓海，曾任总参某部应用研发中心副主任、信息服务中心主任，高级工程师，长期从事数据库应用系统的科研开发工作，负责主持多个大型数据库系统的开发和维护，荣获多项军队科技进步奖，享受军队优秀人才岗位津贴，出版多部论（译）著，在数据库挖掘、数据库应用开发、数据安全保护、数据恢复与数据去密等领域具有丰富的实践经验。
已出版的论著和译著：（（Oracle Streams 11g数据复制》，2012年，清华大学出版社\《SQL Server 2000管理、开发及应用实例详解》，2006年，人民邮电出版社、《空时编码技术》，2004年，机械工业出版社、《远程通信网络基础》，1996年，电子工业出版社。
吴志刚，工学博士，北京邮电大学副教授，长期从事网络与信息安全技术、数据库技术等领域的学术与科研工作，作为负责人主持过上述领域多项国家863计划、发改委产业化示范项目和国家级重大工程项目，获得技术专利2项，已在国内外学术期刊和国际会议上发表20余篇学术论文。

<<数据挖掘>>

书籍目录

章节摘录

版权页： 插图： 12.4分布式数据挖掘（DDM） 海量数据的涌现使得利用分布式系统对海量数据开展跨地理区域的分析的需求不断增长。
为海量数据驱动的知识发现，以及潜在的科学与商业理解带来了史无前例的发展机会。
在高性能分布式计算平台上（而不是集中式计算模型上）实现数据挖掘，其驱动力来自于技术和组织两个因素。
某些情况下，集中处理方式难以实现，因为需要长距离传输将大量的T级数据。
另外，集中方法违背了隐私规则，暴露了商业秘密，并带来其他一些社会问题。
这些问题的典型实例常见于医疗行业，其相关数据往往存在于多个组织商业机构中，例如制药公司、医院、政府实体（如美国食品和药物管理局）和非政府组织（如慈善和公共健康组织）。
每个组织都具有法律限制，例如隐私法规，有关专利信息的公司需求会给竞争对手带来巨大的商业利益。
因此既需要开发算法、工具、服务和基础结构用于实现分布式跨组织的数据挖掘，同时也需要考虑隐私保护问题。
这样一种朝着分布式、复杂环境发展的变化扩大了数据挖掘挑战的范围。
分布式数据所带来的新问题明显增加了数据挖掘过程的复杂性。
通过有线和无线网络，许多分布式计算环境，在计算和通信方面获得了进展。
这样的处理环境多数都涉及包含大量数据的分布式数据源、多个计算节点和分布式用户社区。
对这些分布式数据源进行监视和分析需要新的用于分布式应用的数据挖掘技术。
DDM领域处理这些问题——通过细致分析分布式源挖掘分布式数据源。
除数据分布外，网络的发展产生了大量复杂数据，包括自然语言文本、图像、时间序列、传感器数据、多关系及对象数据类型。
更复杂的是，包含分布式流数据的系统需要增量或在线挖掘工具，无论何时当底层数据发生变化时，需要完整地处理过程。
由于系统变化频繁，应用于如此复杂环境的数据挖掘技术必须适应巨大的动态变化，否则将会对系统的性能带来不良影响。
对所有这些特性提供支持的DDM系统需要有创新的解决方案。
Web架构（包含分层协议和服务）提供了合理的框架用于支持DDM。
新框架接受“融合通信和计算”的新趋势。
DDM接受数据可能自然地分布于不同的松耦合节点上的事实，这些分布的数据往往是通过网络连接起来的异构数据。
DDM提供用于通过分布式数据分析和使用最小数据通信建模发现新知识的技术。
同时，分布式系统交互需要以可靠、稳定、可扩展的方式实现。
最后，系统必须向用户隐藏技术方面的复杂性。
目前，能够通过e—services处理的商品不仅仅局限于类似电器、家具、机票等实体。
Intcmet及WWW的发展包含了软件、计算能力或有用的数据集这类资源。
这些新资源能够通过网络以服务的形式售卖或租赁给网络用户。
直观上看，数据挖掘适于作为一种e—service发布，因为该方法减少了高昂的用于支持该方法的基础架构的设置和维护开销。

<<数据挖掘>>

编辑推荐

《国外计算机科学经典教材:数据挖掘:概念、模型、方法和算法(第2版)》主要用作计算机科学、计算机工程和计算机信息系统专业的研究生数据挖掘教材，高年级本科生或具备同等教育背景的读者也完全可以理解《国外计算机科学经典教材:数据挖掘:概念、模型、方法和算法(第2版)》的所有主题。

版权说明