<<数据挖掘导论>>

图书基本信息

书名：<<数据挖掘导论>>

13位ISBN编号：9787115241009

10位ISBN编号：7115241007

出版时间：2010-12-10

出版时间：人民邮电出版社

作者：Pang-Ning Tan,Michael Steinbach,Vipin Kumar

页数：463

译者：范明,范宏建

前言

自从我和孟小峰等人翻译J. Han和M. Kamber的《数据挖掘：概念与技术》以来，我们高兴地看到数据挖掘的研究正在我国蓬勃开展。
许多学者和研究人员都对这个新兴的学科领域表现出了极大的兴趣，他们之中不仅有来自数据库领域的专家，而且不乏统计学、人工智能和模式识别、机器学习等领域的研究者。
国内的学者和研究者在数据挖掘方面的研究已经取得了一些令人鼓舞的成果，并且正在逐渐与国际学术界同步。
数据挖掘的产生和发展一直是分析和理解数据的实际需求推动的。
数据挖掘研究的进展也正是在于一直重视与其他领域研究者的合作。
数据挖掘从工业、农业、医疗卫生和商业的需求中获得动力，从统计学、机器学习等领域的长期研究与发展中汲取营养。
我们相信，只要有理解数据的需求，就有推动数据挖掘研究与应用发展的动力；只要依靠多学科的团队，就能应对新的数据分析任务带来的挑战。
P. Tan、M. Steinbach和V. Kumar编写的这本《数据挖掘导论》是继《数据挖掘：概念与技术》一书之后的另一本重要的数据挖掘著作。
三位作者都从事数据挖掘研究多年，其中Vipin Kumar教授是数据挖掘和高性能计算领域的国际知名学者。
本书原版在正式出版之前就已经被斯坦福大学、得克萨斯大学奥斯汀分校等众多名校采用。
J. Han教授也高度评价该书：“这是一本全新数据挖掘的教材，值得大力推荐。
它将成为我们的主要参考书。
”本书不需要读者具备数据库背景，只需要少量统计学或数学背景知识，而且取材涉及的学科和应用领域较多，实用性强，因此适合的读者面较广。
本书强调如何用数据挖掘知识解决各种实际问题，强调所挖掘的知识模式的评估。
例如，就像我们能够从天空中的白云想象出各种动物和物体一样，每个聚类算法能够从几乎所有的数据集中发现聚类。
如果数据集合中根本不存在自然的簇，所产生的聚类很难说具有实际意义。
全书共分10章。
范明负责第1~8章的翻译，范宏建负责第9章和第10章的翻译。
蒋宏杰、贾玉祥、许红涛和温箐笛也参加本书的最初翻译工作。
全书的译文由范明负责统一定稿。
在翻译的过程中，对发现的错误进行了更正，并得到原书作者的确认。
感谢P. Tan、M. Steinbach和V. Kumar为中文版撰写序言。
感谢人民邮电出版社图灵公司的编辑们，他们在第一时间内引进本书，并组织翻译，使得中文版能够如此之快地与读者见面。

<<数据挖掘导论>>

内容概要

本书全面介绍了数据挖掘的理论和方法，旨在为读者提供将数据挖掘应用于实际问题所必需的知识。
本书涵盖五个主题：数据、分类、关联分析、聚类和异常检测。
除异常检测外，每个主题都包含两章：前面一章讲述基本概念、代表性算法和评估技术，后面一章较深入地讨论高级概念和算法。
目的是使读者在透彻地理解数据挖掘基础的同时，还能了解更多重要的高级主题。
此外，书中还提供了大量示例、图表和习题。

本书适合作为相关专业高年级本科生和研究生数据挖掘课程的教材，同时也可作为数据挖掘研究和应用开发人员的参考书。

作者简介

陈封能(Pang-Ning Tan)现为密歇根州立大学计算机与工程系助理教授，主要教授数据挖掘、数据库系统等课程。
此前，他曾是明尼苏达大学美国陆军高性能计算研究中心副研究员（2002-2003）。

斯坦巴赫（Michael Steinbach）明尼苏达大学计算机与工程系研究员，在读博士。

库玛尔(Vipin Kumar)明尼苏达大学计算机科学与工程系主任，曾任美国陆军高性能计算研究中心主任。
他拥有马里兰大学博士学位，是数据挖掘和高性能计算方面的国际权威，IEEE会士。

范明,郑州大学信息工程学院教授，中国计算机学会数据库专业委员会委员、人工智能与模式识别专业委员会委员，长期从事计算机软件与理论教学和研究。
先后发表论史40余篇。

范宏建 澳大利亚墨尔本大学计算机科学博士。
先后在WWW、PAKDD、RSFDGrC、IEEE GrC和Australian AI等国际学术会议和IEEE Transactions on Knowledge and Data Engineering发表论文10余篇。
目前是澳大利亚AUSTRAC的高级分析师。

<<数据挖掘导论>>

书籍目录

章节摘录

插图：空间数据的重要例子是科学和工程数据集，其数据取自二维或三维网格上规则或不规则分布的点上的测量或模型输出。
例如，地球科学数据集记录在各种分辨率（如每度）下经纬度球面网格点（网格单元）上测量的温度和气压（见图2-4d）。
另一个例子，在瓦斯气流模拟中，可以针对模拟中的每个网格点记录流速和方向。
5.处理非记录数据大部分数据挖掘算法都是为记录数据或其变体（如事务数据和数据矩阵）设计的。
通过从数据对象中提取特征，并使用这些特征创建对应于每个对象的记录，针对记录数据的技术也可以用于非记录数据。
考虑前面介绍的化学结构数据。
给定一个常见的子结构集合，每个化合物都可以用一个具有二元属性的记录表示，这些二元属性指出化合物是否包含特定的子结构。
这样的表示实际上是事务数据集，其中事务是化合物，而项是子结构。
在某些情况下，容易用记录形式表示数据，但是这类表示并不能捕获数据中的所有信息。
考虑这样的时间空间数据，它由空间网格每一点上的时间序列组成。
通常，这种数据存放在数据矩阵中，其中每行代表一个位置，而每列代表一个特定的时间点。
然而，这种表示并不能明确地表示属性之间存在的时间联系以及对象之间存在的空间联系。
但并不是说这种表示不合适，而是说分析时必须考虑这些联系。
例如，在使用数据挖掘技术时，假定属性之间在统计上是相互独立的并不是一个好主意。

<<数据挖掘导论>>

编辑推荐

《数据挖掘导论(完整版)》是明尼苏达大学和密歇根州立大学数据挖掘课程的教材，由于独具特色，正式出版之前就已经被斯坦福大学、得克萨斯大学奥斯汀分校等众多名校采用。
《数据挖掘导论(完整版)》与许多其他同类图书不同，《数据挖掘导论(完整版)》将重点放在如何用数据挖掘知识解决各种实际问题。
只要求具备很少的预备知识——不需要数据库背景，只需要很少的统计学或数学背景知识。
《数据挖掘导论(完整版)》中包含大量的图表、综合示例和丰富的习题，并且使用示例、关键算法的简洁描述和习题，尽可能直接聚焦于数据挖掘的主要概念。
教辅内容极为丰富，包括课程幻灯片、学生课题建议、数据挖掘资源（如数据挖掘算法和数据集）、联机指南（使用实际的数据集和数据分析软件，《数据挖掘导论(完整版)》介绍的部分数据挖掘技术提供例子讲解）。

版权说明