国家新闻出版署“十五”重点图书

智能数据挖掘与知识发现

Intelligent Data Mining and Knowledge Discovery

陈　莉　刘　静　缑水平　刘　芳　焦李成　著

西安电子科技大学出版社
http://www.xduph.com

国家新闻出版署“十五”重点图书

智能数据挖掘与知识发现

焦李成　刘芳
缑水平　刘静　陈莉　著

西安电子科技大学出版社
2006

图书在版编目(CIP)数据

智能数据挖掘与知识发现/焦李成等著. —西安：西安电子科技大学出版社，2006.8
国家新闻出版署“十五”重点图书
ISBN 7-5606-1734-4
Ⅰ.智… Ⅱ.焦… Ⅲ.数据采集 Ⅳ.TP274
中国版本图书馆 CIP 数据核字(2006)第 084220 号

策　　划　陈宇光
责任编辑　夏大平　李惠萍
出版发行　西安电子科技大学出版社(西安市太白南路 2 号)
电　　话　(029)88242885　88201467　　邮　编　710071
http://www.xduph.com　　E-mail：xdupfxb@pub.xaonline.com
经　　销　新华书店
印刷单位　陕西光大印务有限责任公司
版　　次　2006 年 8 月第 1 版　2006 年 8 月第 1 次印刷
开　　本　787 毫米×960 毫米　1/16　印张 34.375
字　　数　687 千字
印　　数　1～4000 册
定　　价　52.00 元
ISBN 7-5606-1734-4/TP·0430
XDUP　2026001－1

内容简介

面对“人们被数据淹没，却饥渴于知识”的挑战，数据挖掘和知识发现技术应运而生，并得以蓬勃发展。数据挖掘涉及到人工智能、模式识别、机器学习、统计学等领域，因此，我们把体现当代科学技术发展特征的多学科间的知识交叉及最新成果反映到教材中来，同时本书从智能信息处理及数据挖掘两大主题出发，着重于介绍将智能信息处理中的最新技术如何应用于数据挖掘领域，如智能搜索、分类、聚类和智能决策等。

本书在介绍智能信息处理理论、方法、技术的基础上，全面系统地介绍了数据挖掘的概念、相关技术的原理及应用。全书共分9章。第一章主要从整体上介绍数据挖掘和知识发现的基本概念、研究现状及发展方向；第二章介绍了数据挖掘的理论基础；第三章详细论述了用于数据挖掘的计算智能方法的理论基础；第四章论述了神经网络和进化计算的分类方法；第五章全面论述了支撑矢量机与核分类方法；第六章详细论述了集成分类方法；第七章系统论述了数据挖掘中大规模数据聚类方法；第八章论述关联规则挖掘方法；第九章介绍数据挖掘实例及可视化。从第三章后的每一章都给出了所用方法的实验条件设置及实验结果。

本书可作为高校计算机、信号与信息处理、应用数学等专业的高年级本科生或研究生的教材，也可作为从事数据挖掘方面研究工作的科技工作者的参考资料。

前　言

人类正被数据淹没，却饥渴于知识。随着数据库技术的应用越来越普及，人们正逐步陷入“数据丰富、知识贫乏”的尴尬境地。知识信息的“爆炸”给人类带来莫大益处，但也带来不少弊端，造成知识信息的“污染”。面临浩瀚无际而被污染的数据，人们呼唤从数据汪洋中来一个去粗取精、去伪存真的技术。在这种形势下，数据挖掘(Data Mining，DM)应运而生。数据挖掘就是指从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的，人们事先不知道的，但又是潜在有用的，目标明确、针对性强、精炼准确的信息和知识的过程。

海量数据与知识贫乏导致了知识发现和数据挖掘的出现，当人们进入21世纪以后，可以预计知识发现(Knowledge Discovery，KD)与数据挖掘(KDDM)的研究又将形成一个新的高潮。Internet/Web的广泛使用，第三代通信技术的出现更加促进了这一研究的开展。由于DM涉及了人工智能、模式识别、机器学习、统计学等领域，因此，不同领域的学者利用各自不同的技术和方法对数据挖掘进行了卓有成效的研究。然而，如何将不同领域的理论、技术等进行融合将是下一阶段的研究中心，而计算机、通信和网络的三合为一为人们集成这些不同技术提供了硬件基础。可以预见，数据挖掘和知识发现的研究将会进一步深入进行并将有大量的产品问世，然而，对他们的研究将行进在充满挑战又极富发展潜力的漫长之路上。

尽管KDD和DM的研究可以继承大量在计算机科学和模式识别理论中已有的理论和技术，但是它仍面临着大量的问题。

1. 巨大的数据量以及高维数据问题

目前，数据集合中拥有数百万条记录的数据库已大量存在，对这样的数据库进行优化分析会产生数据的组合爆炸，因此要考虑将最优解转换为可接受

解，并要使用降维、去噪等处理方法。

2. 数据缺失问题

由于数据库不是为知识发现定做的，因而就会有一些重要数据或重要属性缺失的问题出现。

3. 变化的数据和知识问题

变化的数据有可能使原有的模式不正确。因此要考虑模式的更新功能，并且要能够利用原来的知识发现新模式，减少分析量。

4. 模式的易读性

要使知识发现结果易于非计算机专业人员理解，即需要进一步加强人机对话能力。

5. 与其他系统集成问题

数据库的发展趋势是不仅要存储数字化数据，而且要存储许多非标准化数据，与这类模型的数据接口问题日趋严重，只有实现这些不同模型数据的有效结合，才具有实用价值。

针对海量数据挖掘存在的维数灾难及小样本问题，传统机器学习理论的“小样本”统计假设受到了极大的限制和挑战。信息稀疏问题摆在了人们的面前：属性巨大，样本稀少，而一些属性的数值对问题世界没有意义，大量数据存在于边界上。而对实际问题，目标函数各异，表示形式多样，多个需求叠加的“平均”，无人喝彩。面对海量数据的不确定性、“小世界”无尺度性、非线性、非平稳、非高斯、非局域、非标号、非结构以及半结构化、域知识特征等如何能够实现有效的智能数据挖掘与知识发现成为一个有重大应用潜力的课题，相应的方法学也面临新的挑战。

近年来，人们在向自然进化学习，向人脑学习，向遗传免疫学习的基础上，发展了多种自然计算和智能信息处理的方法。从 1996 年开始，我们在国家“863”计划、国家自然科学基金重点项目、教育跨世纪人才基金、国家“973”子项等课题资助下，展开了相关课题的研究，并取得了一定的成果，这些成果主要包括：自适应子波网络、多尺度几何网络、免疫克隆计算、量子克隆计算、量

子克隆进化计算、方向多分辨脊波网络、协同神经网络分类、海量数据的组织协同进化分类、基于免疫克隆算法的特征选择、子波支撑矢量机、模糊多类支撑矢量机、子波核匹配追踪学习机、集成分类器、基于免疫克隆算法的选择性SVMs集成、核匹配追踪分类集成。免疫进化聚类及混合属性特征大数据聚类、克隆进化聚类及混合属性特征大数据聚类、克隆网络结构聚类、基于有限资源的模糊网络结构聚类、量子聚类和核聚类、谱聚类、基于组织进化和组织多层次进化的关联规则挖掘与Web日志挖掘、基于克隆算法的多维数据及孤立点挖掘、基于Petri网的可视化模型的实现等。

可以说，本书是我们在该领域工作的小结，是智能信息处理研究所10余年来工作的集体结晶。特别感谢保铮院士多年来的悉心培养和教导；感谢中国科技大学陈国良院士和IEEE进化计算杂志主编、英国Birmingham大学的姚新教授的指导和帮助；感谢国家自然科学基金委信息科学部的大力支持；感谢王磊、李映、张莉、周伟达、郑春红、杨淑媛、钟伟才、李洁、刘若辰、王爽、薄列峰、公茂果、张向荣、马秀丽、李阳阳、王玲、李青、侯冲、尚荣华、马文萍、卢滨、王蓉芳、王静、胡锐、胡颖等同学所付出的辛勤劳动。本书的部分内容借鉴了国内外其他专家和作者的最新研究成果，同时该书也得到了西安电子科技大学出版社的关心和支持，在此深表感谢！

感谢作者家人的大力支持和理解。

由于作者水平有限，书中不妥之处在所难免，恳请读者批评指正。

著　者

2006年5月

目　录

第一章 绪 论

近年来，由于计算机性能提高、成本下降及数据管理技术的成功运用，使各部门内部的信息化程度越来越高，同时也造成了大量数据的积累。“数据丰富，知识贫乏”，决策者很难从海量的数据中提取出有价值的知识的现状，促使人们产生了对数据分析工具的强烈需求。利用数据分析工具所获取的信息和知识，可以广泛地用于商务管理、生产控制、市场分析、工程设计和科学研究与探索等诸多方面。

事实上，数据挖掘是高级数据分析工具，是信息技术自然演化的结果。20 世纪 60 年代，计算机的应用领域从科学计算转向事务处理，人们利用磁带、磁盘等计算机的相关技术，进行数据的收集、回顾和静态传输，这便是早期的数据库技术；1970 年 E. F. Cold 创建了关系数据库模型，实现了从层次和网状数据库系统到关系数据库系统的开发，形成了在数据收集基础上的数据索引和数据组织技术、查询处理及查询优化、事务处理和联机事务处理(OLTP)，集成了关系数据库管理系统(DBMS)、结构查询语言(SQL)和开放数据库互联(ODBC)技术，使数据可在记录层上进行动态传输，从而使数据库应用得到了迅速推广。80 年代起，人们广泛运用关系模型，研究和开发新的、功能强大的数据库系统。如基于高级数据模型的扩充关系模型、面向对象模型、对象—关系模型和演绎模型，基于应用的空间模型、时间模型、多媒体模型、主动模型、知识库模型等。80 年代后期以来，随着数据仓库和决策支持技术的出现，利用在线分析处理(OLAP)、多维数据库(MDDB)和数据仓库(DW)技术，可保证在数据收集的基础上进行动态的多层次的数据传输。90 年代，一种新的数据库系统——Web 数据库系统出现。进入 2000 年以来，新一代的综合信息系统应运而生。随着查询和事务处理的各种大型数据库系统的应用，数据分析和理解已成为新的目标，人们期待着从数据中发现有价值的信息和知识，发现事物发展的趋势，期待自动分析工具的出现[1,2]。所以，利用大型数据库技术、并行计算机、高级算法进行主动的信息传输已成为数据库系统和机器学习中的一个关键性的研究课题[3,4]。

近年来，Internet 及相关技术又使计算机、网络、通信三合为一。网络经济、注意力经济等新的概念的出现，以其巨大的社会效益和极富挑战与机遇的内涵，成为信息科学最引

人注目的研究课题。然而，网络在快捷、方便地带来大量信息的同时，也带来了一大堆的问题：诸如信息过量，难以消化；信息真假难以辨识；信息安全难以保证；信息形式不一致，难以统一处理，等等。如何理解已有的历史数据并用以预测未来的行为，如何从这些海量数据中发现信息，变被动的数据为主动的知识，如何快速、准确地获得有价值的网络信息和网络服务，为用户提供重要的、未知的信息或知识，指导政府决策、企业决策，获取更大的经济效益和社会效益，这些都迫使人们去寻找新的、更为有效的数据分析手段，对各种"数据矿藏"进行有效的挖掘以发挥其应用潜能。20 世纪 80 年代后期至今，高级数据分析——数据挖掘(Data Mining，简称 DM)与数据库中的知识发现(Knowledge Discovery in Database，简称 KDD)正是在这样的应用需求背景下产生并迅速发展起来的、开发信息资源的一套科学方法、算法及软件工具和环境。数据挖掘和知识发现(Knowledge Discovery，简称 KD)是集统计学、人工智能、模式识别、并行计算、机器学习、数据库等技术为一体的一个交叉性的研究领域[5-7]。

1.1 数据挖掘概述

简单地说，数据挖掘是提取或"挖掘"知识。目前，数据挖掘可以从统计学、数据库和机器学习等三个角度进行定义。"挖掘"一词最早出现于统计学中。从统计学的角度，数据挖掘是指分析所观察的数据集以发现可信的数据间的未知关系并提供给数据拥有者可理解的、新颖的和有用的归纳数据[8]。从数据库的观点来看，数据挖掘是指从存储在数据库、数据仓库或其它信息仓库中的大量数据中发现有趣的知识的过程[9]。从机器学习的角度，数据挖掘定义为从数据中抽取隐含的、明显未知的和潜在有用的信息[10]。

数据库中的知识发现(KDD)是识别有效的、新颖的、具有潜在用处的、可理解的数据模式的过程。数据库(DB)与数据库中的知识发现(KDD)的关系从名称上就体现了明显的区别。数据库提供了基本数据模型下的存储和数据操作。而 KDD 的过程说明了知识发现常常意味着经验、重复、用户的交互及许多设计、决策和习惯。简单地讲，KDD 表示了从低层数据抽象高层知识的整个过程。通过数据库中的知识发现，人们可以从数据库的数据及相关集合中抽象出有用的知识、数据的规律性或高层的信息。对于 KDD，还有一些类似的术语，如从数据库中挖掘知识、知识提取、数据考古、数据捕捞、数据/模型分析等。

本质上，数据挖掘(DM)与数据库中的知识发现(KDD)也是不同的，但也有一些人把数据挖掘和 KDD 等同看待，其实 DM 仅仅是 KDD 的一个步骤。典型的 KDD 的过程如下：

(1) 数据清理：消除噪声或不一致的数据。

(2) 数据集成：将多种数据组合在一起。

(3) 数据选择：从数据库中检索与分析任务相关的数据。

(4) 数据变换：将数据变换或统一成适合挖掘的形式，如通过汇总或聚集操作。

(5) 数据挖掘：使用不同的智能方法提取数据模式或模型。

(6) 模式(型)评估：根据某种兴趣度度量，识别表示知识的真正有趣的模式。

(7) 知识表示：使用可视化和知识表示技术，向用户提供所挖掘的知识。

KDD 的工作流程如图 1 - 1 所示。

数据挖掘过程可以与用户或知识库交互，将有趣的模式提供给用户，或作为新的知识存放在知识库中。比较广义的观点是：数据挖掘是从存放在数据库、数据仓库或其他信息库中的大量的数据中挖掘有趣知识的过程。

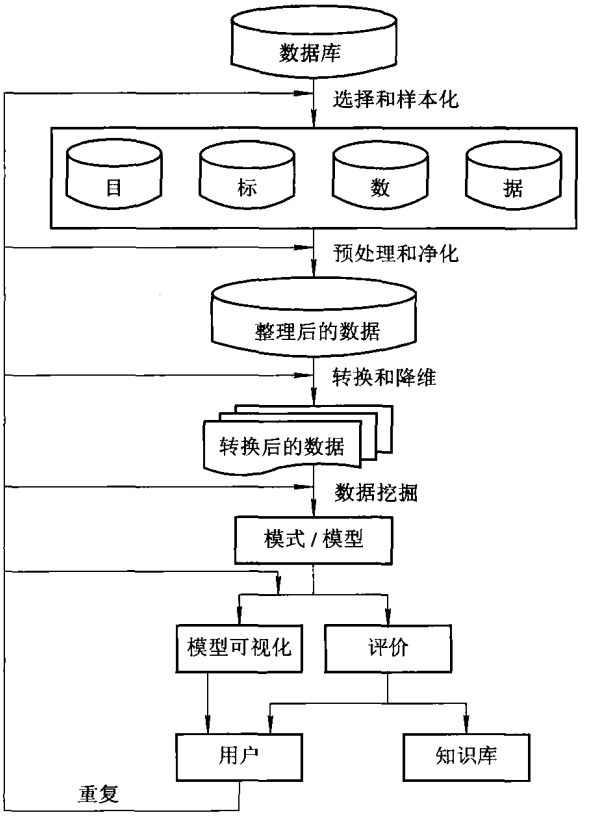

图 1 - 1 KDD 的步骤

按照这样的观点，典型的数据挖掘系统具有如下组成：

· 数据库、数据仓库或其他信息库：这是一个或一组数据库、数据仓库、电子表格或其他类型的信息库。可以在此数据集上进行数据清理和集成。

· 数据库或数据仓库服务器：根据用户的数据挖掘请求，数据库或数据仓库服务器负