数据挖掘与最优化技术及其应用

袁玉波　杨传胜　黄廷祝　徐成贤 著

科学出版社
www.sciencep.com

数据挖掘与最优化技术及其应用

袁玉波　杨传胜　黄廷祝　徐成贤　著

科学出版社

北京

内 容 简 介

本书介绍几类数据挖掘问题优化模型以及用于求解数据挖掘优化模型的优化算法,其中包括算法设计和数值实验.书中详细介绍了数据分类问题、数据聚类问题、回归问题、等基数的双目录分割问题、数据相关性问题的最优化数学模型以及关联规则挖掘算法和因果规则的近似表示理论.本书反映了数据挖掘的数学理论基础的最新研究成果.

本书可作为数据挖掘理论和算法及相关专业的研究生教材,也可作为相关专业科技工作者的参考书.

图书在版编目(CIP)数据

数据挖掘与最优化技术及其应用/袁玉波等著. —北京:科学出版社,2007

ISBN 978-7-03-019077-2

I.数… II.袁… III.数据采集-最佳化 IV.TP274

中国版本图书馆CIP数据核字(2007)第084756号

责任编辑:赵彦超/责任校对:邹慧卿

责任印制:赵德静/封面设计:王 浩

科学出版社出版

北京东黄城根北街16号

邮政编码:100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2007年7月第 一 版 开本:B5(720×1000)

2007年7月第一次印刷 印张:13 1/2

印数:1—3 000 字数:252 000

定价:36.00元

(如有印装质量问题,我社负责调换〈环伟〉)

序　言

21 世纪是信息化的世纪, 数据将充满人们的日常生活. 随着数据量的爆炸式增长, 激增的数据背后隐藏着越来越多重要信息, 人们希望能够对其进行更高层次的分析, 以便更好地利用这些数据. 目前的数据库系统可以高效地实现数据的录入、查询、统计等功能, 但无法发现数据中存在的关系和规则, 无法根据现有的数据预测未来的发展趋势, 缺乏挖掘数据背后隐藏的知识的手段, 从而导致了 "数据爆炸但知识贫乏" 的现象. 在 21 新世纪刚刚开始的时候, 我们回顾往昔, 人类在本世纪用什么来在历史的长河中留下发展痕迹呢? 就推动人类社会进步而言, 历史上能与网络技术相比拟的是什么技术呢? 有人甚至提出要把网络技术与火的发明相比拟. 火的发明区别了动物和人, 科学技术的种种重大发现扩展了自然人的体能、技能和智能, 而网络技术则大大提高了人的生活质量和人的素质, 使人成为 "全球人".

但是, 大量信息在给人们带来方便的同时也带来了一大堆问题: 一是信息过量, 难以消化; 二是信息的真假难以辨识; 三是信息安全难以保证; 四是信息形式不一致, 难以统一处理. 因此, 人们提出了一个新的口号: "要学会抛弃信息". 人们开始考虑: "如何才能不被信息淹没, 而是从中及时发现有用的知识, 提高信息利用率?" 在数据的海洋里, 掌握游泳技术的人必然是胜利者. 数据挖掘技术应运而生, 并显示出强大的生命力.

数据挖掘技术是当前机器学习、模式识别、计算机科学、智能计算技术、应用数学、统计学习方法以及智能机器人研究中重要的课题. 随着现代计算机技术、信息技术以及通信技术的迅猛发展, 如何从已有数据中分析、提炼和挖掘出隐含的、先前未知的、新奇的、对决策有潜在应用价值的知识, 已经越来越成为迫切需要解决的问题. 近二十年来, 关于此类问题的研究使得数据挖掘这个全新的学科和研究方向迅速崛起, 并在金融业、零售业、电信业以及基因序列组成研究等领域获得了广泛的应用.

目前已经有很多数据挖掘的书籍陆续出版, 但大多是从应用的角度出发, 对于从事数据挖掘的研究工作带有局限性, 读者往往读了很多本同类数据挖掘的著作, 也难以找到有用的创新点和知识. 本书选择 "数据挖掘与最优化技术及其应用" 为题, 通过分析几类数据挖掘问题, 建立相应问题的优化模型, 并研究用于求解数据挖掘的优化模型的具体优化算法, 进行算法设计和数值实验.

本书的主要内容集中于以下五个方面:

1. 挖掘问题最优化模型领域. 本书给出了数据分类问题、数据聚类问题、回

归问题以及数据相关性问题的最优化数学模型. 同时给出了新的光滑支持向量机模型、数据聚类优化模型和数据相关性的优化模型.

2. 支持向量机数据二分类技术. 本书主要给出一种用多项式光滑的新支持向量机模型, 并用 BFGS 算法和 Newton-Armijo 算法求解. 数值实验表明, 多项式光滑的支持向量机模型是十分有效的光滑模型. 对于数据分类这个研究方向, 感兴趣的读者还可以参看文献 [257~259], 在文献 [257] 中, Fadimeuney 和 Turkay 提出了一种基于整数规划的多分类模型, 对于研究支持向量机多分类模型非常有帮助; 在文献 [259] 中, Martin 采用了线性分割的思想研究数据分类. 本书主要研究支持向量机数据二分类技术, 使用函数光滑技术, 得出了一种多项式光滑的支持向量机模型, 并且使用 BFGS 和 Newton-Armijo 方法对光滑支持向量机模型进行了求解和比较, 说明了提出的模型是比较有效的. 关于支持向量机理论和应用研究, 这两年国际上涌现出很多优秀的成果, 有兴趣的读者可以参看文献 [260~317].

3. 聚类优化模型及其求解算法. 本书对数据聚类的 k 质心聚类算法提出了一种改进的 k 质心聚类算法. 在算法的初始化阶段用分段技术对初始中心进行处理, 数值实验表明, 新算法效率明显提高. 对于数据聚类这个研究方向, 由于原始问题和算法都比较单一, 所以相对来说受研究者的青睐程度较低. 主要关注 k 质心聚类算法, 我们给出了一种改变初始点选择的 k 质心聚类算法, 通过数值实验说明, 这是一种相对效率较高的算法. 但是, 近两年国际上也出现了很多更加优秀的 k 质心聚类算法和理论成果. 关于这个方面的研究, 感兴趣的读者可以参考文献 [318~329].

4. 等基数双目录分割问题. 本书阐述了等基数的双目录分割问题. 在书中将等基数的双目录分割问题转化为一个半定规划模型, 然后给出一个改进的随机算法, 通过算法理论分析, 得到其算法近似性能比为 0.6378, 高于 0.5. 由此, 在回答由 Jon Kleinberg 在 1998 年提出的公开问题方面取得了一定的进展. 分割问题实际上可以归为聚类方法的应用, 由于最优化领域对最大二分问题的理论和求解算法的进步, 这个方向不断取得进展. 对于其原始问题和应用研究读者可以参看文献 [330~332].

5. 关联规则挖掘. 本书研究了关联规则挖掘算法和因果规则的近似表示, 在书中提出一种新算法 —— 矩阵算法, 通过数值实验说明, 矩阵算法在减少产生项大集的运行时间上效果是显著的. 同时, 还用优化的思想对因果关系的近似表示进行了研究, 给出一种用一次多项式近似表示的方法, 近似表示的结果比较理想, 误差也较小. 对于关联规则挖掘方面, 由于其理论基础较薄, 研究趋于成熟, 所以近两年理论研究论文数量的增长相对趋于缓慢, 现在应用研究也是一个热点. 这一研究的意义在于推进了数学知识在关联规则挖掘中的应用. 关于这个方面的研究, 感兴趣的读者可以参考文献 [333~364].

这些研究工作只是数据挖掘研究的微少的部分, 数据挖掘领域还有很多内容

值得研究和应用. 从数据挖掘研究工作未来的发展看, 有两个方面是比较重要的: (1) 数据挖掘理论基础的研究, 包括数据处理优化算法理论、概率统计理论、微观经济观点以及数据库理论, 其中微观经济观点指的是把数据挖掘看成从数据中发现模式, 从而对企业决策过程 (如制定市场策略、产品开发计划等) 起指导作用的过程, 从这个观点上讲, 数据挖掘实际上是一个非线性的优化问题; (2) 数据挖掘系统产品开发, 包括数据挖掘语言的规范化、智能系统、挖掘结果的可视化以及实际应用系统的可伸缩性.

本书主要内容是一个新兴的研究领域, 本书期望使得最优化技术在数据挖掘中的应用更加广泛, 推进数据挖掘快速算法研究, 开创一个交叉研究领域. 另外, 粗糙集方法、遗传算法、模糊集方法、非线性方法、小波分析方法、神经网络方法等在数据挖掘中的应用前景非常广泛. 当然, 相对其他应用学科, 数学在数据挖掘中的应用尚处于初级阶段, 如果能将现有的数学领域的一些优秀成果应用到数据挖掘中来, 对数据挖掘的推动作用不可估量.

作者感谢科学出版社在出版本书过程中给予的帮助和支持. 由于作者的知识的局限性和写作技巧不高, 不足的地方请专家和读者谅解, 批评指正, 不吝赐教.

通过阅读本书, 读者朋友能够最终受益, 则是作者工作的最高荣誉.

祝每一位读者朋友学业顺利!

作　者

2006 年 10 月 18 日

于弗吉尼亚理工大学

目　录

第一章 引 言

数据挖掘技术是 20 世纪 80 年代后期兴起的一门交叉学科. 数据库、统计学、最优化技术、人工知识、模式识别、并行计算、机器学习、神经网络、数据可视化、信息检索、图像与信号处理和空间数据分析等在研究数据挖掘技术方面都有应用. 换句话说, 只要存在数据存储的地方, 数据挖掘就有存在的土壤. 尤其是信息泛滥的今天, 从过去的数据中获取有用的知识, 就显得更加重要. 由于数据挖掘是新兴学科, 涉及的理论背景也不是特别深, 研究可以从很多方面入手, 这种情况使数据挖掘成为一个热门的研究课题. 我国在该方向的研究工作相对滞后, 要使我国数据挖掘研究水平赶上国际水平, 在学术研究和软件开发两个方面, 同行的专家和学者们仍需要努力. 本书的目的是研究和探索最优化方法在数据挖掘的应用, 开辟优化方法与数据挖掘这一交叉研究课题, 拓宽数据挖掘理论研究的背景.

本章分五节, 主要介绍数据挖掘的一些基本概念和应用技术. §1.1 介绍数据挖掘的意义和研究数据挖掘的必要知识; §1.2 给出数据挖掘的基本概念、特点和数据库知识发现的主要步骤; §1.3 初步介绍数据挖掘的主要内容; §1.4 主要阐述数据挖掘的应用背景; §1.5 总体介绍本书的研究工作, 包括创新点和结构.

§1.1 数据挖掘的意义

随着现代信息技术、通信技术和计算机技术的高速发展, 数据库应用的范围、深度和规模不断扩大. 传统的信息系统大部分是查询驱动的, 数据库作为历史知识库对于一般的查询过程是有效的, 但当数据和数据库的规模急剧增长时, 传统的数据库管理系统的查询检索机制和统计分析方法已远远不能满足现实的需求, 它迫切要求能够自动、智能和快速地从数据库中挖掘出有用的信息和知识. 下面的两个例子可以说明信息资源呈爆炸性增长的趋势.

(1) 美国国家航空航天局 (National Aeronautics and Space Administration, NASA) 是负责太空计划的美国联邦政府机构, 于 1957 年创立. 除了太空计划, 美国国家航空航天局还进行长期的民用和军用航空宇宙研究, 被广泛认为是世界范围内太空机构的领头羊. 20 世纪 80 年代中期, 美国提出空间站的发展设想, 专门成立了对地观测系统 (EOS) 科学和飞行任务需求的研究工作小组, 从理论上提出了地球科学应用的基本任务, 确定了低轨道地球观测的基本需求, 从地球物理、气候过程、生物化学和水文等四大学科, 确定六个方面的观测内容. 应当说, 这个总结是

比较全面的, 它指导着一个时代相关技术的发展. 最初, 他们设想建立极轨空间站平台, 装载 EOS 有效载荷. 平台总重约 10t, 峰值功率 25kW, 数据容量 500Mpbs. 在这种平台上, 将装载五组仪器, 其中有 12 种新型对地观测敏感仪器. 具体罗列如下: 第一组, 地面成像和探测 (SISP), 包括中分辨率成像光谱仪、高分辨率成像光谱仪、高分辨率多频微波辐射计、光雷达大气探测和测高仪; 第二组, 主动式微波敏感仪 (SAM), 包括合成孔径雷达 (SAR)、雷达测高仪、雷达散射计; 第三组, 大气物理和化学监视器 (APACM), 包括多普勒光雷达、上层大气干涉仪、对流层成份监视器、上层大气成份监视仪、高能粒子监视器; 第四组, 实用化温度和湿度探测仪, 包括扫描辐射计、高分辨率红外探测仪; 第五组, 监视仪. 用于监视太阳、粒子和场、地球辐射收支等, 包括太阳紫外光谱辐照度监视仪、太阳常数监视仪、磁圈粒子和场探测仪、磁圈流体和场测量仪、地球辐射收支仪等. 每天 EOS 向地面接收站发回 $1 \sim 10 \times 10^{12}$(10^{12} 字节约合 954G) 以上字节的数据.

另外, 加盟 "行星地球" 国际计划的卫星以及航天器每天的数据量更加庞大. 随着世界人口增加, 加速消耗地球的资源, 工业化竞争带来全球环境恶化. 过度地使用土地、森林面积减少, 使沙漠化加速漫延. 臭氧层破坏、CO_2 增加, 大气污染日趋严重. 地球的环境已经开始影响世界的农业、能源和人类健康. 人类必须面对这个涉及到自身生存的严重问题, 保护地球环境. 1989 年, 美国等 24 国提出 "行星地球计划"(Mission to Planet Earth, MTPE). 这项空间计划的目标是: 跟踪地球环境的变化过程; 记录自然过程同植物、动物和人类生活等的相互作用过程; 记录大气、海洋和陆地三者之间的相互作用过程. 该计划的基本任务是收集那些在地球环境方面对于国际组织选择正确方法和国家决定正确决策起作用的信息. 该计划将耗资数百亿美元, 耗时 15 年, 建立由一大批卫星或空间飞船组成的对地观测系统, 坚持长期观察和测量, 积累具有论断能力的数据. 20 世纪 90 年代初期世界 "冷战" 结束的政治形势, 更加有利于这一计划的实施. "行星地球计划" 的主要内容是建立三种空间平台, 根据总目标确定每种平台的卫星数量和星上有效载荷任务. 这样, 从 20 世纪 90 年代后期开始, 将发射二十多颗各种卫星, 它们共同构成整个对地球观测的完整系统. 极轨平台是指大型太阳同步极轨卫星. 目前已经确定并部分完成的三个卫星系列, 即美国的对地观测系统 (EOS-AM, -PM, -Chem-1)、欧空局的环境卫星 -1 (Envisat-1) 以及日本的高级对地观测卫星 (ADEOS). 它们均属于大型遥感平台, 重量在 3~8t 范围, 平台上安放多种遥感仪. 卫星由各国自行组织研制和发射, 有效载荷内容协商确定, 数据共享, 共同建立集成的空间对地球体系的观测系统. 地球同步轨道平台, 包括 5~6 颗静止轨道卫星, 分别由美、日、欧等国负责发射, 它们提供整个地球范围的连续的环境数据. 在波段上, 还将增加微波遥感仪, 提供地球表面温度和湿度的数据. 小型 "地球探针" 是一批针对性较强的小卫星. 例如, 测量臭气总量的光谱仪 (TOMS), 将由小火箭作为独立小卫星发射. 它们专门收集特殊的信

息, 作为 EOS 主体的补充. 由众多遥感仪器组成的庞大对地观测系统, 所产生的数据数量也是巨大的. 该计划将专门建立数据和信息系统 (例如 EOSDIS). 除了按常规方法建立数据标准格式的数据库之外, 还将研究巨量数据的管理方法、所有权及预订数据等方法. 特别, 受空间遥感商业化发展的影响, EOS 数据和信息也有可能走商业化的发展道路.

(2) 到 2006 年为止, 专业收集研究世界各个行业数据的 Fortune 500 公司在其电子数据库中, 已经拥有了 4×10^{16} 千字节 (约合 38147000G) 的数据, 这就需要有 4×10^{16} 字节的大存储器.

在这个信息时代里, 不同的群体都面临如何有效使用大量数据的问题. 例如, 如何从有关数据中获益? 怎样用历史数据为产生数据的过程建模? 怎样对某一过程的行为进行预测或预报? 对数据库中的数据作何种解释、如何分类等等. 据估计, 目前仅有 5% ~ 10% 的商业数据库被加以分析利用 [1]. 数据挖掘技术就是为迎合这种要求而产生并迅速发展起来的, 它为现代信息处理提供了一种新的方法和研究领域.

理解一个大的数据集合并进行抽象是十分困难的. 数据的增加一般沿两个层面: 领域的数目 (也称属性和维数) 和案例数. 人类的分析和抽象能力不适宜于高维和海量数据, 处理高维数据的一个标准方法是把数据投影到一个维数较低的子空间, 然后再在这个简化的空间内进行分析或建模. 随着维数的不断增长, 因为降维所可能选择的组合个数呈爆炸性增长; 另外, 向低维空间的投影, 往往可能把一个本来相对容易识别的问题转变成一个难于识别的问题, 而数据挖掘的某些算法 (如支持向量机算法) 利用反转技术有目的地逐渐增加维数, 使得诸如分类等问题变得相对容易.

然而, 即使我们接受减少维数是必须的这样一个事实, 探索和分析仍然需要由人进行指导, 那么仍然有一个投影选择的问题要解决. 对所有可能的投影方案进行实验以选取正确的样本子集是做不到的. 数据形象化的一个手段是应用数据挖掘算法对数据进行适当的简化. 例如, 一个聚类算法可以把嵌在高维空间中的有显著特征的数据子集选出, 再选定几个维数把它同其他类别的数据进行区分, 由此, 可以建立一个非常有效的具体模式, 使得分析人员可以发现那些隐含在高维空间中难于发现的模式 (知识) 或模型.

另外, 一个反映数据挖掘是十分必要的因素是数据集合增长的速度远远超过传统的手工分析技术所能处理的程度. 如果想及时地利用由数据提供的信息, 用传统的分析技术方法, 达到这个目的是不可能的, 实际上, 这意味着大多数数据没有得到利用, 信息资源被浪费, 这样一种情况在一个谁能较好地利用数据资源谁就能获得明显效益的充满竞争的环境中是多么让人心痛. 而大多数团体都会面临这样的问题, 无论是商业企业、科学研究还是政府机关, 这就导致一个严肃的问题, 即迫切需要重新考虑数据收集和数据分析的策略.

在市场竞争非常激烈的今天, 企业或个人都经常面临着复杂的决策问题, 不仅需要快速作出决策, 而且需要分析与解决决策问题中多种不确定性所带来的困难. 好的决策可以使企业或个人的发展获得巨大的成功, 坏的决策一定会使企业或个人招致失败. 所以, 应用数据挖掘技术开展决策分析的理论与方法研究, 不仅对中国管理科学的发展具有重大的理论意义, 而且对解决许多复杂的实际决策问题也有特别重要的现实意义. 结合对中国实际问题的研究, 提出新的不确定性决策理论与方法是决策方法研究的主要课题.

下面是国际上决策分析学界关注的一些主要科学问题.

(1) 不确定性决策问题的建模与分析技术研究. 面对复杂的决策问题, 试图完全用数学模型进行精确刻画似乎是不现实的, 即使对某些问题可行, 但求解与分析也是非常困难的. 因此, 从 20 世纪 90 年代初开始, 人们就借助新发展的信息技术来处理或支持处理复杂的决策问题. 例如, 用数据挖掘 (data mining) 技术帮助解决商场货物摆放决策问题, 著名的“啤酒与尿布”案例便是一个成功的例子. 又如, 用多智能体 (multi-agents) 技术实现复杂问题决策支持已成为决策支持系统研究的一条主线. 除了要重视这些重要趋势外, 应在解决中国的复杂决策问题中提出自己的不确定性决策的建模与分析技术.

(2) 随机决策模型的求解算法. 不少实际决策问题可以用随机优化模型来近似刻画与分析, 例如, 流域水资源分配与管理问题可用一个随机规划模型来分析与求解; 国家或地区的环境规划决策问题可由一个多层多目标随机优化模型来刻画; 多阶段金融投资决策问题可用动态随机规划来建模与分析. 无论是作为问题的完整模型还是作为问题建模的一部分子模型, 它们都具备随机、高维和非线性特征, 数学上求解是相当困难的. 所以, 这是不确定性决策方法能否真正解决复杂决策问题的关键之一, 也是国际上决策科学界特别关注的一个研究方向. 在这个方面, 演化算法的研究应受到足够的重视.

(3) 风险度量与管理. 在不确定性决策中, “风险”是一个特别重要的概念. 如何对风险进行度量和管理是决策分析研究的核心内容之一. 传统上, 常用收益的方差来刻画风险, 近年来国际上对市场风险的度量提出了一个新概念: value at risk(简记为 VaR), 用来度量由基本市场因子的不利运动所导致的下方风险位势 (downside risk potential). VaR 的提出受到了金融界和学术界的重视, 但如何从实际数据计算 VaR? 如何针对一般的不确定性数据过程提高 VaR 估计的精度? 此外, 对灾害这一类管理决策问题中的风险又该如何度量与管理? 如何预测风险? 这些都是迫切需要研究的重要理论问题. 中国未来 10 年中, 在航空航天、金融、灾害和其他社会经济活动中都将面临一系列与风险管理相关的决策问题.

(4) 新的效用理论. 效用最优化是决策分析研究的基础. 在不确定性决策中, von Neumann-Morgenstern 的期望效用准则起着非常重要的作用, 但期望效用准则

本身也存在着相当的局限性. 国外已提出了条件期望效用和非期望效用等理论, 但仍需要提出新的适用于解决实际问题的效用理论. 这是决策分析的一个最基本的理论问题, 也是国际学术界特别关注的一个重要研究方向.

(5) 网络环境下的决策理论与方法. Internet 对企业的生产组织方式和决策模式已经产生了前所未有的影响, 研究网络环境下的决策理论与方法就显得特别重要. 由于电子商务的快速发展, 基于 Internet 的招投标决策与拍卖已成为国际决策分析学术界激烈竞争的一个热点, 并被 INFORMS 和 ACM 列为 21 世纪初最有挑战性的决策科学问题. 许多与决策分析理论和方法相关的问题在网络环境下都需要重新开展研究.

以上是在管理科学中不确定性决策理论与方法的五个主要研究课题, 也是研究最优化方法在数据挖掘中应用的五个主要应用研究课题. 国家自然科学基金委员会也列出 130 万元的重点项目研究此课题. 详细信息可以在如下网址找到:

http://www.nsfc.gov.cn/nsfc/cen/xmzn/2005xmzn/02zd/02/05.htm.

内容如下:

最优化与数据挖掘 (G0110)

以用最优化及其相关方法探寻并发展高水平的、具有实际应用价值的数据挖掘技术, 旨在高效率、高精度地从海量的数据中发现潜在、新颖及有用的知识. 主要研究内容应涉及基于最优化及其相关方法的分类、聚类、关联、预测、模式等数据挖掘技术的探讨, 从建模、特征、算法、有效性、实用性等方面寻求众多最优化及其相关方法与各种数据挖掘技术的结合; 研究应考虑问题及求解中的非结构化、非线性、近似性、不确定性等特点, 并应结合实际管理决策的应用问题展开研究.

拟资助经费: 130 万元

主管科学部: 管理科学部 (管理科学一处管理科学与工程学科负责受理)

相关交叉科学部: 信息科学部

从上面信息可见数据挖掘技术的理论和算法的研究已经是时代的要求. 它是信息时代的产物, 是处理人工不可能完成的企业或个人的要求时而产生的应用工具.

下一节将给出数据库知识发现和数据挖掘的一般性定义.

§1.2 数据库知识发现

通常数据挖掘被视作以提取有用信息为目的的“数据簇聚”或“数据产生”过程. 数据为信息处理者提取新的和有用规则服务, 并能够根据已有的信息对实际未发生行为的结果作出预测. 数据挖掘是从大量数据中挖掘出隐含的、先前未知的、对决策有潜在价值的知识和规则. 这些规则蕴含了数据库中一组对象之间的特定联系, 揭示出一些有用的信息, 为经营决策、市场策划、经营预测、工业控制提供依

据. 通过数据挖掘, 有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来, 并从不同角度显示, 从而使大型数据库作为一个丰富可靠的资源为知识归纳服务.

数据库知识发现 (knowledge discovery in database, KDD) 是指识别出存在于数据库中有效的、新颖的、具有潜在价值的、最终可理解的、模式的、非平凡知识的过程. 数据挖掘的结果知识可能是概念、模式、规则、规律、可视化或者特殊数据等. 它是一个众多学科相互交融形成的、有广阔应用前景的新兴领域, 其中包括数据库、统计学、最优化技术、粗糙集、人工知识、模式识别、并行计算、机器学习、神经网络、数据可视化、信息检索、图像与信号处理和空间数据分析等. 知识发现的整个过程包括在指定的数据库中用数据挖掘算法提取模型, 以及围绕数据挖掘进行的预处理和结果表达等一系列计算步骤. 数据挖掘算法是整个过程的核心, 通常占整个过程 15% ~ 25% 的工作量. 数据挖掘是知识发现的一个关键步骤, 包括特定的数据挖掘算法, 具有可接受的计算效率, 生成特殊的模式; 知识发现强调知识是数据发现的最终产品, 利用相应的数据挖掘算法, 按指定方式和阈值提取有价值的知识, 包括数据挖掘前对数据的预处理、抽样及转换和数据挖掘后对知识的评价解释过程.

一般的, 数据库知识发现包括以下主要迭代步骤:

(1) 问题定义. 首先, 知识发现的计划目标必须确定, 这些目标必须是可行有用的. 也就是说, 一旦知识发现的目标达到, 就有公司或者个人对结果知识加以利用. 另外, 知识发现的应用数据也必须在这一阶段确定.

(2) 数据预处理. 这一阶段包括数据采集、数据清理、数据筛选和数据转换.

数据采集是从互连网或其他途径得到必要的数据, 还包括重新进行数据表示, 统一符号差异, 从不同的数据桌面连接形成整齐的数据源.

数据清理是检测和解决数据冲突、数据异常、噪音数据、错误数据、含糊数据等; 另外, 还有应用变换和组合产生新的数据域, 比如说应用变化率或者滚动求和等. 这一步可以说占整个知识发现工作工作量的 70%, 甚至还要更多.

数据选择是从给定数据库中选择出与某一项分析任务有关的数据. 换句话说, 就是从一个数据集合中选出关于变量或者数据样本的子集, 挖掘就在这个子集上进行.

数据转换是把数据变换或统一成适合挖掘的形式, 比如说通过汇总或其他操作.

(3) 数据挖掘. 是知识发现的核心步骤, 它甚至成为知识发现的代名词. 在这里, 一些灵活的方法将用于提取模式. 有意义的模式是一系列研究的形式表达式或者是表达式的集合, 包括分类规则、分类树、回归、聚类、系列模型、相关性、规律和可视化.

(4) 知识集成. 包括模式评价、模型保持和应用以及知识表示.

知识发现的过程是迭代的. 比如说经过清理和准备的数据, 或许会发现从一个给定的数据源得到的数据是不可用的, 或许事先没有考虑到或没有识别的数据源在对另外数据的考虑下需要并入挖掘数据源. 所以需要迭代地进行对数据的清理工作.

数据挖掘是数据库知识发现中与算法有关的一步, 现有的数据挖掘技术分为 5 类, 即相关性建模、聚类 (也称为分割)、数据概括、预测模型以及发现变化和偏差. 从国内外目前的研究进展来看, 各学科的研究自成一派, 没有突破各个领域的技术界限, 没有融合各领域的不同方法, 尤其是未将并行优化的诸方法集成用于数据库中的数据挖掘, 从而提高实时性, 并解决随机的、动态的、不完全的及混沌数据的数据挖掘, 即所谓智能数据挖掘. 而且以往多数技术都是在驻留于内存的数据之上进行挖掘, 没有把这些技术与数据库技术相集成. 近年来, 有些技术已开始定位于大型数据库上的挖掘, 即基于磁盘存贮进行挖掘. 从而出现了关系数据库的数据挖掘、面向对象数据库的数据挖掘等.

数据挖掘是近年来一个十分活跃的研究领域. 从数据库中发现知识一词首先出现在 1989 年举行的第十一届国际联合人工智能学术会议上. 到目前为止, 由美国人工智能协会主办的 KDD 国际研讨会已召开了 9 次, 由美国电气电子工程师学会 (IEEE) 主办的 KDD 国际会议 2003 年 11 月在美国弗罗里达州召开第 4 次会议, 2004 年 11 月在乌克兰召开第 5 次会议. 规模由原来的专题讨论会发展到国际学术大会, 仅以 2002 年为例, 就有 20 多个国际会议列有 KDD 专题. 以数据挖掘为主题的其他国际会议主要有:

ADMA(The International Conference on Advances Data Mining and Applications), 2005 年召开第 1 届国际会议;

PAKDD(The Pacific-Asia Conference on Knowledge Discovery and Data Mining), 2005 年召开第 9 届国际会议;

PKDD(The European Conference on Principles and Practice of Knowledge Discovery in Databases), 2005 年召开第 9 届国际会议;

ICDE(The International Conference on Data Engineering), 2005 年召开第 21 届国际会议;

DASFAA (The International Conference on Database Systems for Advanced Applications), 2006 年召开第 11 届国际会议;

ICDM(The Fifth IEEE International Conference on Data Mining), 2005 年召开第 5 届国际会议;

MLDM(The International Conference on Machine Learning and Data Mining), 2005 年召开第 5 届国际会议.

这两年国内也有相当多的数据挖掘和知识发现方面的研究成果. 许多学术会

议上都设有专题进行学术交流. 目前, KDD 的研究重点集中于发现方法的研究和实际的系统应用. 在系统方面, 国际上有影响的典型数据挖掘系统有 SAS 公司的 EnterpriseMiner, IBM 公司的 InterligentMiner, SG 公司的 Setminer 等.

下面给出一些更加一般的定义.

数据库知识发现　从数据中发现真实、新奇, 有潜在应用价值, 同时最终可以被解释的模式 (知识) 的全过程. 它包括从数据库中对数据进行选取或取样、清理或预处理、转换或必要的简化, 由数据挖掘产生模式, 直至对得出的模式进行解释和评估等过程.

数据挖掘　数据挖掘是 KDD 全过程中同算法有关的一步, 借助于算法, 在可接受的计算范围内从数据枚举模式或模型 (结构).

与传统信息处理方法相比, 数据挖掘技术有其自身的特点:

(1) 处理对象为大规模数据库, 数据规模十分巨大;

(2) 信息查询一般是由决策制定者 (用户) 提出的即时随机查询, 往往没有精确的查询要求, 需要靠数据挖掘技术寻找其可能感兴趣的东西;

(3) 在一些应用中, 某些行动并没有实际发生或很少发生, 因而它们对输出所造成的影响没有在数据库中体现出来, 需要利用数据挖掘技术从数据库中提取有用的规则, 为这些情况作出预测;

(4) 在一些应用中, 由于数据变化迅速, 可能很快过时, 因此要求数据挖掘技术能快速对数据变化作出反应以提供决策支持, 数据挖掘既要发现潜在的规则, 还要管理和维护规则, 而规则是动态的, 当前的规则只能反应当前状态的数据库特征, 随着新数据的不断加入, 规则需要随之更新;

(5) 数据挖掘中规则的发现主要基于大样本的统计规律, 发现的规律未必适用于所有的数据, 当达到某一阈值时便可认为有此规律.

随着关系数据库系统的广泛应用、数据仓库技术的发展, 数据预处理 (比如数据采集、数据清理、数据筛选和数据变换) 可以通过构造数据仓库和执行联机分析处理 (OLAP) 操作来完成. 实际上, 数据库知识发现退化为一个步骤: 数据挖掘. 当然一些必要的挖掘后处理工作还是必要的, 比如说, 对挖掘结果的解释、模型保持等. 这也就是在很多场合人们把数据库知识发现和数据挖掘看成是同义词的原因.

数据挖掘是一个应用工具. 它只是帮助商业人士更加深入透彻地去分析和应用数据, 无法告之使用者它具体的贡献是多少, 只是告之使用者该怎样去做, 至于为什么, 使用者无从得知. 也许这也是数据挖掘的魅力所在.

接下来在下一节中对数据挖掘的工作作简单的阐述.

§1.3 数据挖掘的主要内容

一般说来, 数据挖掘工作可以分为两类: 描述式的数据挖掘和预测式的数据挖掘. 前者是用一种简明扼要的方法描述数据集合和表示有意义的数据性质; 后者是参考可靠的数据集合构造一个或一组模型, 并且尝试描述或预测新的数据集合的行为 [2~5].

一个数据挖掘系统会完成下列数据挖掘工作中一种或者多种.

(1) 类的阐述 (描述). 简明扼要地对数据进行收集、概括, 并且将它们区分开. 概括收集实际上就是特征归类; 比较两个数据集合称为数据类比或者说数据区分. 类别描述包括数据分布的属性的全部, 比如说变量情况、变量的个数等. 例如, 类别的描述可以用于分析一个欧洲公司对亚洲的销售数据库, 对过去的销售业绩作一个综合回顾, 总结归纳出畅销商品情况等, 也可以分析亚洲和非洲的销售数据库的差异, 也就是数据类比.

(2) 数据分类. 数据仓库、数据库或者其他信息库中隐藏着许多可以为商业、科研等活动的决策提供所需要的知识. 分类与预测是两种数据分析形式, 它们可以用来抽取能够描述重要数据集合或预测未来数据趋势的模型. 分类方法用于预测数据对象的离散类别; 预测方法用于预测数据对象的连续取值. 数据仓库、数据库或者其他信息库中隐藏着许多可以为商业、科研等活动的决策提供所需要的知识. 分类与预测是两种数据分析形式, 它们可以用来抽取能够描述重要数据集合或预测未来数据趋势的模型. 分类方法 (classification) 用于预测数据对象的离散类别 (categorical label); 预测方法 (prediction) 用于预测数据对象的连续取值. 换句话说, 数据分类是基于数据的特征, 分析训练数据集合, 并且给每个类一个确定模型. 这里训练集合是诸如类别已经知道的目标集. 一个分类过程是能更好地去理解数据库的每一个类, 并且能够对未知数据进行分类的过程. 这一个过程产生一棵决策树、类的模型或者分类规则集合.

分类技术在很多领域都有应用, 例如, 可以通过客户分类构造一个分类模型来对银行贷款进行风险评估; 当前的市场营销中很重要的一个特点是强调客户细分. 客户类别分析的功能也在于此, 采用数据挖掘中的分类技术, 可以将客户分成不同的类别, 比如呼叫中心设计时可以分为: 呼叫频繁的客户、偶然大量呼叫的客户、稳定呼叫的客户、其他, 帮助呼叫中心寻找出这些不同种类客户之间的特征, 这样的分类模型可以让用户了解不同行为类别客户的分布特征; 其他分类应用如文献检索和搜索引擎中的自动文本分类技术; 安全领域有基于分类技术的入侵检测等. 比如说, 从过去的病人数据库中挖掘出病人的分类模型, 然后根据病人的症状将病人进行分类, 即确定是哪一种病.