

中国计算机学会学术著作丛书

知识发现

史忠植摇著

清华大学出版社

前摇摇言

随着计算机应用及数据库的日益普及,“丰富的数据与贫乏的知识”问题也日见突出,世界上的数据正以惊人的速度增长,堆积如山。不同领域的人们都期待着从这些数据中得到自己想要的答案,将信息变为知识,从数据矿山中找到蕴藏的知识金块。

知识发现正是这样一种从数据中挖掘知识的工具,它集数据收集、数据清洁、降维、规则归纳、模式识别、数据结果分析及评估、可视化输出等多种过程于一身,是统计学、计算机科学、模式识别、人工智能、机器学习及其他学科相结合的产物。它不仅被许多研究人员看作是数据库系统和机器学习方面一个重要的研究课题,而且被许多工商界人士看作是一个能带来巨大回报的重要领域。从数据库中发现出来的知识可以用在信息管理、查询响应、决策支持、过程控制等许多方面。从20世纪80年代中期的小范围研究到如今的蓬勃兴起,知识发现已经在企业界与科学界占据了一席之地。事实上,世界领先企业中的愿望都涉足知识发现的前瞻性研究或拥有一个或多个知识发现产品系统。它们帮助企业进行客户关系管理,减少不必要的投资,提高资金周转和回报,帮助人们迅速获取所需的知识和信息,提高工作效率,改进服务质量。

知识发现是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式而非平凡过程,它与数据仓库有着密切的联系。数据仓库是源于可操作数据的一个数据存储中心。数据仓库中的信息是面向主题的、稳定的且具有历史数据属性的,因此数据仓库用于存储大规模的数据集。知识发现与数据仓库、决策支持的结合预示着信息和知识管理的一个全新的变革。

本书全面而又系统地介绍了知识发现的方法和技术。全书共分15章。第1章是绪论,介绍知识发现的重要概念和任务。第2章讨论决策树,它是归纳学习方法中最实用的一种技术。关联规则挖掘是近几年应用最为广泛的方法,第3章将对重要的关联规则挖掘算法进行讨论。第4章讨论范例推理,它是一种有效的实用技术。第5章探讨模糊聚类法。第6章讨论粗糙集。第7章是贝叶斯网络,贝叶斯网络可以处理不完整和带有噪声的数据集,它用概率测度的权重来描述数据间的相关性。第8章探讨支持向量机,它在近几年知识发现研究中是极其活跃的研究课题。第9章讨论隐马尔科夫模型。第10章是神经网络,书中着重介绍几种实用的算法。第11章讨论进化和遗传算法。第12章介绍知识发现平台,接着,以宰藻知识发现、生物信息处理为例,介绍知识发现的应用。第13章关于宰藻知识发现。第14章介绍生物信息处理中基因组模式的发现。

本书是中国科学院计算技术研究所智能信息处理重点实验室有关机器学习和知识发现研究工作的总结。涉及的研究项目得到国家自然科学基金、国家高新技术计划、北京市自然科学基金、国家重点科技攻关项目的资助。参加该项研究工作的人员有叶

世伟副教授、何清副教授、李晓黎博士、叶施仁博士、宫秀军博士生、刘少辉博士生、郑毅、郑金华教授、张建博士、王军博士、张颖博士、李云峰博士、潘谦红博士、吴斌博士生、贾自艳博士生、游湘涛、任力安、李宝东等。本书得到清华大学出版社计算机专著出版基金的资助。

史忠植
二〇〇五年 八月

目 录

前言	I
第 1 章 绪论	1
1.1 知识	1
1.2 知识发现	1
1.3 知识发现的任务	1
1.3.1 数据总结	1
1.3.2 概念描述	1
1.3.3 分类	1
1.3.4 聚类	1
1.3.5 相关性分析	1
1.3.6 偏差分析	1
1.3.7 建模	1
1.4 知识发现的方法	1
1.4.1 统计方法	1
1.4.2 机器学习	1
1.4.3 神经计算	1
1.4.4 可视化	1
1.5 知识发现的对象	1
1.5.1 数据库	1
1.5.2 文本	1
1.5.3 半结构化信息	1
1.5.4 空间数据	1
1.5.5 图像和视频数据	1
1.6 知识发现与创新	1
第 2 章 决策树	2
2.1 归纳学习	2
2.2 决策树学习	2
2.3 混杂学习算法	2
2.4 关联学习算法	2
2.4.1 信息论简介	2
2.4.2 信息论在决策树学习中的意义及应用	2
2.4.3 关联算法	2
2.4.4 关联算法应用举例	2

第 5 章 模糊聚类	5
5.1 模糊聚类概述	5
5.1.1 模糊聚类结果的表示	5
5.1.2 模糊聚类的一般模型	5
5.2 传递闭包法	6
5.2.1 模糊相似系数的标定	6
5.2.2 传递闭包法	6
5.2.3 动态直接聚类法	6
5.2.4 最大树法	6
5.3 云团子聚类法	6
5.3.1 问题背景	6
5.3.2 云团等价标准型	6
5.3.3 置换等价类与平移等价类的记数公式	6
5.3.4 云团的结构	6
5.3.5 模糊最优等价阵的存在性	6
5.3.6 最优模糊等价阵的算法步骤	6
5.3.7 基于云团子模糊聚类的语音识别	6
5.4 系统聚类法	6
5.5 悦胸值聚类法	6
5.6 聚类有效性	6
5.7 聚类方法的比较	6
第 6 章 粗糙集	6
6.1 粗糙集概述	6
6.1.1 知识的分类观点	6
6.1.2 新型的隶属关系	6
6.1.3 概念的边界观点	6
6.2 知识的约简	6
6.2.1 一般约简	6
6.2.2 相对约简	6
6.2.3 知识的依赖性	6
6.3 决策逻辑	6
6.3.1 决策表的公式化定义	6
6.3.2 决策逻辑语言	6
6.3.3 决策逻辑语言的语义	6
6.3.4 决策逻辑的推演	6
6.3.5 规范表达形式	6
6.3.6 决策规则和决策算法	6
6.3.7 决策规则中的一致性和不分明性	6

远缘瑶决策表的约简	员源
远缘瑶瑶属性的依赖性	员缘
远缘瑶瑶一致决策表的约简	员缘
远缘瑶瑶非一致决策表的约简	员园
远缘瑶粗糙集的扩展模型	员猿
远缘瑶瑶可变精度粗糙集模型	员源
远缘瑶瑶相似模型	员缘
远缘瑶瑶基于粗糙集的非单调逻辑	员缘
远缘瑶瑶与其他数学工具的结合	员园
远缘瑶粗糙集的实验系统	员园
远缘瑶粗糙集的展望	员愿
第 苑章瑶贝叶斯网络	员怨
苑缘瑶概述	员怨
苑缘瑶瑶贝叶斯网络的发展历史	员怨
苑缘瑶瑶贝叶斯方法的基本观点	员园
苑缘瑶瑶贝叶斯网络在数据挖掘中的应用	员园
苑缘瑶贝叶斯概率基础	员园
苑缘瑶瑶概率论基础	员园
苑缘瑶瑶贝叶斯概率	员源
苑缘瑶贝叶斯学习理论	员愿
苑缘瑶瑶几种常用的先验分布选取方法	员苑
苑缘瑶瑶计算学习机制	员怨
苑缘瑶瑶贝叶斯问题求解	员员
苑缘瑶简单贝叶斯学习模型	员猿
苑缘瑶瑶简单贝叶斯学习模型	员猿
苑缘瑶瑶简单贝叶斯模型的提升	员缘
苑缘瑶瑶提升简单贝叶斯分类的计算复杂性	员苑
苑缘瑶贝叶斯网络的建造	员苑
苑缘瑶瑶贝叶斯网络的结构及建立方法	员苑
苑缘瑶瑶学习贝叶斯网络的概率分布	员愿
苑缘瑶瑶学习贝叶斯网络的网络结构	员园
苑缘瑶贝叶斯潜在语义模型	员猿
苑缘瑶半监督文本挖掘算法	员园
苑缘瑶瑶网页聚类	员园
苑缘瑶瑶对含有潜在类别主题词的文档的类别标注	员苑
苑缘瑶瑶基于简单贝叶斯模型学习标注和未标注样本	员愿
第 愿章瑶支持向量机	圆猿
愿缘瑶统计学习问题	圆猿

愿爱瑶经验风险	愿爱缘
愿爱瑶灾况维	愿爱缘
愿爱瑶学习过程的一致性	愿爱原
愿爱瑶学习一致性的经典定义	愿爱原
愿爱瑶学习理论的重要定理	愿爱原
愿爱瑶灾况熵	愿爱缘
愿爱瑶结构风险最小归纳原理	愿爱远
愿爱瑶支持向量机	愿爱愿
愿爱瑶线性可分	愿爱愿
愿爱瑶线性不可分	愿爱识
愿爱瑶核函数	愿爱员
愿爱瑶多项式核函数	愿爱员
愿爱瑶径向基函数	愿爱员
愿爱瑶多层感知机	愿爱员
愿爱瑶动态核函数	愿爱圆
愿爱瑶基于分类超曲面的海量数据分类方法	愿爱远
愿爱瑶分圆曲线定理	愿爱远
愿爱瑶杂圆直接方法基本思想	愿爱原
愿爱瑶实现算法	愿爱缘
愿爱瑶实验结果分析	愿爱缘
第 愿爱瑶隐马尔科夫模型	愿爱怨
愿爱瑶马尔科夫过程	愿爱怨
愿爱瑶隐马尔科夫模型	愿爱园
愿爱瑶似然概率和前反向算法	愿爱员
愿爱瑶前向算法	愿爱圆
愿爱瑶反向算法	愿爱圆
愿爱瑶灾况算法	愿爱远
愿爱瑶计算期望	愿爱远
愿爱瑶学习算法	愿爱原
愿爱瑶耘算法	愿爱原
愿爱瑶梯度下降	愿爱缘
愿爱瑶灾况学习	愿爱远
愿爱瑶基于状态驻留时间的分段概率模型	愿爱远
愿爱瑶杂圆模型的构成	愿爱苑
第 愿爱瑶神经网络	愿爱园
愿爱瑶概述	愿爱园
愿爱瑶基本的神经网络模型	愿爱园
愿爱瑶神经网络的学习方法	愿爱园

第 4 章 瑶 人工神经元及感知机模型	4.0
瑶 基本神经元	4.1
瑶 感知机模型	4.2
第 5 章 瑶 前向神经网络	5.0
瑶 前向神经网络模型	5.1
瑶 多层前向神经网络的误差反向传播(FF)算法	5.2
瑶 FF 算法的若干改进	5.3
第 6 章 瑶 径向基函数神经网络	6.0
瑶 插值问题	6.1
瑶 正规化问题	6.2
瑶 FF 网络学习方法	6.3
第 7 章 瑶 反馈神经网络	7.0
瑶 离散 FF 网络	7.1
瑶 连续 FF 网络	7.2
瑶 FF 网络应用	7.3
瑶 双向联想记忆模型	7.4
第 8 章 瑶 随机神经网络	8.0
瑶 模拟退火算法	8.1
瑶 玻尔兹曼机	8.2
第 9 章 瑶 自组织特征映射神经网络	9.0
瑶 网络的拓扑结构	9.1
瑶 网络自组织算法	9.2
瑶 有教师学习	9.3
第 10 章 瑶 进化和遗传算法	10.0
瑶 概述	10.1
瑶 基本遗传算法	10.2
瑶 基本遗传算法的构成要素	10.2.1
瑶 基本遗传算法的一般框架	10.2.2
瑶 遗传算法的数学理论	10.3
瑶 模式定理	10.3.1
瑶 积木块假设	10.3.2
瑶 遗传算法欺骗问题	10.3.3
瑶 隐并行性	10.3.4
瑶 遗传算法的基本实现技术	10.4
瑶 编码方法	10.4.1
瑶 适应度函数	10.4.2
瑶 选择算子	10.4.3
瑶 交叉算子	10.4.4

元遗传算法变异算子	猿缘
元遗传算法约束条件的处理方法	猿缘
元遗传算法的高级实现技术	猿缘
元遗传算法反转操作	猿缘
元遗传算法变长度染色体遗传算法	猿远
元遗传算法小生境遗传算法	猿远
元遗传算法混合遗传算法	猿苑
元遗传算法改进遗传算法	猿园
元遗传算法并行遗传算法	猿员
元遗传算法遗传算法应用	猿圆
元遗传算法优化神经网络连接权值	猿圆
元遗传算法用遗传算法优化神经网络连接结构	猿猿
第 4 章 知识发现平台 酝酝酝	猿缘
元概述	猿缘
元数据仓库	猿苑
元数据仓库含义	猿苑
元数据	猿愿
元数据管理	猿怨
元数据仓库和数据挖掘技术的结合	猿怨
元酝酝酝的体系结构	猿园
元数据挖掘模型	猿园
元系统功能	猿员
元体系结构	猿圆
元元数据管理	猿猿
元酝酝酝元数据的内容	猿猿
元酝酝酝元数据库	猿源
元酝酝酝元数据对象模型	猿源
元数据仓库管理器	猿苑
元酝酝酝数据仓库的基本结构	猿愿
元主题	猿怨
元数据抽取和集成	猿园
元数据抽取和集成的元数据	猿猿
元数据仓库建模及 酝酝 的实现	猿源
元算法库管理	猿愿
元数据挖掘算法的元数据	猿愿
元可扩展性的实现	猿怨
元数据挖掘算法的接口规范	猿园
元数据挖掘任务规划	猿圆

第 1 章 生物信息知识发现	猿源
1.1 概述	猿源
1.2 基因的基本结构	猿远
1.3 生物信息数据库与查询	猿苑
1.3.1 基因和基因组数据库	猿苑
1.3.2 蛋白质数据库	猿怨
1.3.3 功能数据库	猿园
1.4 序列比对	猿员
1.4.1 序列两两比对	猿员
1.4.2 多序列比对	猿猿
1.5 核酸与蛋白质结构和功能的预测分析	猿源
1.5.1 核酸序列的预测方法	猿源
1.5.2 针对蛋白质的预测方法	猿缘
1.6 基因组序列信息分析	猿苑
1.7 功能基因组相关信息分析	猿园
1.7.1 大规模基因表达谱分析	猿园
1.7.2 基因组水平蛋白质功能综合预测	猿员
1.8 数据库资源和公共数据库	猿园
参考文献	猿苑
索引	猿愿

第 员章 摇摇论

摇摇知摇摇识

摇摇人类从工业社会向知识社会演进时,政治经济中心正从“生产”转向“发现、发明和创新”。知识正在成为创新的核心,知识创新成为知识经济发展的最主要的动力源泉。知识经济对物质文明发展能够发挥巨大的推动作用,依靠无形资产的投入来实现可持续发展的,推动经济全球化发展。

在信息科学中,信息是根据表示数据所用的约定,赋予数据的意义。数据是事物、概念或指令的一种形式化的表示形式,以适合于用人工或自然方式进行通信、解释或处理。信息是数据所表达的客观事实。数据是信息的载体,与具体的介质和编码方法有关。20世纪40年代,香农对信息的数学本质进行过研究,提出了著名的香农信息论。他用熵的概念来研究信息的容量,采用比特作为度量信息的单位。

信息经过加工和改造形成知识。知识是人类在实践的基础上产生又经过实践检验的对客观实际的可靠的反映。知识是人脑创新的成果,是人类智慧的结晶。智慧是人类文明的源泉,是推动历史发展的永衡动力,是生产力诸要素中的核心。知识一般可分为陈述性知识、过程性知识和控制性知识。陈述性知识提供概念和事实,描述系统状态、环境和条件,使人们知道是什么。例如,在一个知识检索系统中,陈述性知识包括陈述具体事实的数据库内容。过程性知识提供有关状态的变化、问题求解过程的操作、演算和动作的知识。智能信息检索系统利用过程性知识处理陈述性知识。用控制策略表示问题的知识常称为控制性知识。控制性知识,即元知识,包含有关各种处理过程、策略和结构的知识,常用来协调整整个问题求解的过程。

知识具有下列特性:

(员) 知识的客观性。虽然知识是人脑对信息加工的成果,但这些成果是客观的,人类对自然、社会、思维规律的认识是客观的,这些规律的运行是不以人的意志为转移的。

(2) 知识的相对性。人类对自然、社会、思维规律的认识必须有一个过程。在一段时间内认为正确的东西,经过变革,可能发生变化。1955年第5届国际人工智能大会上,阿佩尔的计算机和思维奖授予阿佩尔的阿佩尔,他提出基于行为的人工智能,认为智能不要知识表示,智能不要推理^[1]。1956年第6届国际人工智能大会上,阿佩尔的计算机和思维奖授予斯坦福大学的阿佩尔,以表彰她在概率推理、机器学习方面的贡献^[2]。

(3) 知识的进化性。人类在认识客观世界和主观世界的过程中,不断对真理的长河加入新的内容,知识不断更新,例如对物质结构的认识,对基因的认识等。

(4) 知识的依附性。知识有载体,载体分层次。离开载体的知识是没有的。随着载体的消失,知识也跟着消失。

(5) 知识的可重用性。在使用过程中知识可以反复重用。当然,要根据具体情况作

具体分析,灵活应用知识。

(远)知识的共享性。基础研究一般由政府进行投资,所得到的科学知识具有共享性;但最新的技术知识受到知识产权法保护,使用者只有支付一定的费用,才能获得这种知识的使用权。知识产权的保护对发展技术和知识经济是非常重要的国策。

知识发现

知识发现是从数据集中抽取和精化新的模式。知识发现的范围非常广泛,可以是经济、工业、农业、军事、社会、商业、科学的数据或卫星观测得到的数据。数据的形态有数字、符号、图形、图像、声音等。数据组织方式也各不相同,可以是有结构、半结构或非结构的。知识发现的结果可以表示成各种形式,包括规则、法则、科学规律、方程或概念网等。

目前,关系型数据库应用广泛,并且具有统一的组织结构,一体化的查询语言,关系之间及属性之间具有平等性等优点。因此,数据库知识发现(从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程)的研究非常活跃。该术语于1989年出现,定义为“知识发现是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程”。在上面的定义中,涉及几个需要进一步解释的概念:“数据集”、“模式”、“过程”、“有效性”、“新颖性”、“潜在有用性”和“最终可理解性”。数据集是一组事实(如关系数据库中的记录)。模式是一个用语言来表示的一个表达式,它可用来描述数据集的某个子集。作为一个模式要求它比对数据子集的枚举要简单(所用的描述信息量要少)。过程在知识发现中通常指多阶段的处理,涉及数据准备、模式搜索、知识评价以及反复的修改求精;该过程要求是非平凡的,意思是要有一定程度的智能性、自动性(仅仅给出所有数据的总和不能算作是一个发现过程)。有效性是指发现的模式对于新的数据仍保持有一定的可信度。新颖性要求发现的模式应该是新的。潜在有用性是指发现的知识将来有实际效用,如用于决策支持系统里可提高经济效益。最终可理解性要求发现的模式能被用户理解,目前它主要是体现在简洁性上。有效性、新颖性、潜在有用性和最终可理解性综合在一起称为兴趣性。

由于知识发现是一门受到来自各种不同领域的研究者关注的交叉性学科,因此导致了很不同的术语名称。除了知识发现外,主要还有如下若干种称法:“数据挖掘”(从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程)、“知识抽取”(从数据集中抽取知识)、“信息发现”(从数据集中发现信息)、“智能数据分析”(从数据集中分析信息)、“探索式数据分析”(从数据集中探索信息)、“信息收获”(从数据集中收获信息)和“数据考古”(从数据集中考古)等等。其中,最常用的术语是“知识发现”和“数据挖掘”。相对来讲,数据挖掘主要流行于统计界(最早出现于统计文献中)、数据分析、数据库和管理信息系统界;而知识发现则主要流行于人工智能和机器学习界。

知识发现过程可粗略地理解为三部曲:数据准备、数据开采以及结果的解释评估(见图1.1)。

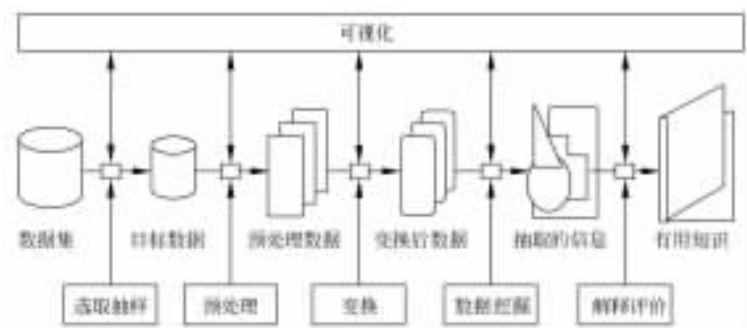


图 员 猿 知识发现过程示意图

猿 数据准备

数据准备又可分为三个子步骤：数据选取（数据源选择）、数据预处理（数据清洗、数据集成）和数据变换（数据预处理选择）。数据选取的目的是确定发现任务的操作对象，即目标数据（数据子集），它是根据用户的需要从原始数据库中抽取的一组数据。数据预处理一般可能包括消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换（如把连续值数据转换为离散型的数据，以便于符号归纳，或是把离散型的转换为连续值型的，以便于神经网络归纳）等。当数据开采的对象是数据仓库时，一般来说，数据预处理已经在生成数据仓库时完成了。数据变换的主要目的是消减数据维数或降维（数据选择），即从初始特征中找出真正有用的特征以减少数据开采时要考虑的特征或变量个数。

猿 数据挖掘阶段

数据挖掘阶段首先要确定开采的任务或目的是什么，如数据总结、分类、聚类、关联规则发现或序列模式发现等。确定了开采任务后，就要决定使用什么样的开采算法。同样的任务可以用不同的算法来实现，选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来开采；二是用户或实际运行系统的要求，有的用户可能希望获取描述型的（数据描述）、容易理解的知识（采用规则表示的开采方法显然要好于神经网络之类的方法），而有的用户或系统的目的是获取预测准确度尽可能高的预测型（数据预测）知识。

完成了上述准备工作后，就可以实施数据挖掘操作了。具体的数据挖掘方法将在后面章节中作较为详细的论述。需要指出的是，尽管数据挖掘算法是知识发现的核心，也是目前研究人员的主要努力方向，但要获得好的采掘效果，必须对各种采掘算法的要求或前提假设充分的理解。

猿 结果解释和评价

数据挖掘阶段发现出来的模式，经过用户或机器的评价，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要整个发现过程退回到发现阶段之前，如重新选取数据、采用新的数据变换方法、设定新的数据挖掘参数值，甚至换一种采掘算法（如当发现任务是分类时，有多种分类方法，不同的方法对不同的数据有不同的效果）。另外，由于最终是面向人类用户的，因此可能要对发现的模式进行可