

智能数据挖掘技术

薛惠锋 张文宇 寇晓东 编著

西北工业大学出版社

【内容简介】 数据挖掘是人工智能、机器学习、数据库技术等多学科相结合的产物,是由计算机自动从已有数据中发现以前未知的、具有潜在应用价值的信息或模式的技术。数据挖掘是知识发现过程的重要内容。本书系统地介绍了数据挖掘技术的原理、方法和计算技术,可作为系统工程、控制工程及计算机类专业研究生的教材,也可供相关专业技术人员参考使用。

图书在版编目(CIP)数据

智能数据挖掘技术/薛惠锋,张文宇,寇晓东编著. 西安:西北工业大学出版社,2005.2

ISBN 7-5612-1891-5

I. 智… II. ①薛… ②张… ③寇… III. 数据采集—研究生—教材 IV. TP274

中国版本图书馆 CIP 数据核字(2005)第 009766 号

出版发行:西北工业大学出版社

通信地址:西安市友谊西路 127 号 邮编:710072

电 话:029-88493844 88471757

网 址:www.nwpup.com

印 刷 者:陕西兴平市印刷厂

开 本:850 mm×1 168 mm 1/32

印 张:8.625

字 数:208 千字

版 次:2005 年 2 月第 1 版 2005 年 2 月第 1 次印刷

定 价:21.00 元

前 言

当今世界正处在信息爆炸的时代,人类在如日中天的信息化浪潮中开启了 21 世纪的现代化之旅,信息技术以其在各个领域应用中的强大渗透性和高附加值成为时代发展的强劲动力。作为人类智慧的象征,智能技术是信息技术发展中最具诱惑力的研究前沿,知识发现(KDD)是其重要的研究课题之一。

如何在浩瀚的信息海洋中获取有用的信息是困扰人们有效利用信息资源的“瓶颈”问题。数据挖掘(Date Mining)就是要从已有的海量数据中发现以前未知的、具有潜在应用价值的信息和模式,解决高噪声数据与低音量知识的矛盾。它是人工智能、数据库技术、机器学习等多种技术的结合,是知识发现过程的关键步骤。从数据挖掘的基本概念和原理,到挖掘方法、算法及软件工具,许多学者、研究部门和 IT 公司都进行了广泛深入的研究,形成了较为完整的数据挖掘理论与方法体系,并出现了一些实用的数据挖掘工具。在石油勘探、金融、商业销售、产品制造、医疗、保险、化工等领域,这些研究成果得到广泛应用,取得了巨大的经济效益。

作为一个非常具有活力的研究领域,数据挖掘技术的知识在不断地翻新。作者虽涉足此领域研究多年,但也只是对其有了一个初步的把握。本书将作者了解的数据挖掘的有关技术及自己的研究成果融合成一个易于理解的体系呈现给大家。

书中第 1 章先介绍了人工智能的基本概念及研究领域,然后是知识发现和数据挖掘的概念与结构,以及相关的问题。第 2 章介绍了数据仓库的基本概念、结构和设计,以及工程规划方面的内

容。从第 3~7 章开始,详细介绍了数据挖掘的相关技术,即是:基于概率统计与神经网络的数据挖掘技术、基于信息论的数据挖掘技术、基于关联规则的数据挖掘技术、基于分类规则的数据挖掘技术、基于聚类规则的数据挖掘技术。第 8 章和第 9 章是数据挖掘应用方面的内容,分别介绍了基于 Web 的数据挖掘技术和基于数据挖掘的智能决策研究框架。

本书在写作过程中,参考了众多同行专家的论著,西北工业大学资源与环境信息化工程研究所的荣群山博士对书稿的完成也做了大量工作,在此一并向他们表示感谢。

由于我们时间仓促,才疏学浅,加之数据挖掘是一项新技术,书中难免有错误和不当之处,恳请广大读者批评指正。

作者

2004 年 10 月

目 录

第 1 章 绪论	1
1.1 人工智能综述	1
1.1.1 人工智能的发展历史	1
1.1.2 人工智能的概念及研究领域	10
1.1.3 人工智能的发展前景	16
1.2 知识发现导论	18
1.2.1 知识发现(KDD)与数据挖掘的概念	18
1.2.2 KDD 过程及系统结构	19
1.2.3 KDD 研究的主要问题	22
1.2.4 KDD 应用及存在的问题	23
1.2.5 KDD 中不完备信息的问题	27
思考题	32
第 2 章 数据仓库的概念与结构	33
2.1 数据仓库的概念和特征	33
2.1.1 数据仓库的概念	33
2.1.2 数据仓库的特征	37
2.2 数据仓库工程规划	38
2.2.1 制定数据仓库工程规划的重要性	38
2.2.2 制定数据仓库工程规划的过程	39
2.2.3 数据仓库工程规划文档的内容	42

2.3	数据仓库系统的设计准则	48
2.4	数据仓库的结构和设计	50
2.4.1	数据仓库的数据模型	50
2.4.2	数据仓库的元数据管理	51
2.4.3	数据仓库的组件	57
2.4.4	数据仓库体系结构	58
	思考题	67
第3章	基于概率统计与神经网络的数据挖掘技术	68
3.1	基于概率统计的数据挖掘技术	68
3.2	基于神经网络的数据挖掘技术	71
	思考题	74
第4章	基于信息论的数据挖掘技术	75
4.1	信息论原理	75
4.1.1	基本思想	75
4.1.2	基本概念	76
4.1.3	信道模型及容量	81
4.2	基于互信息的 ID3 算法及改进算法	82
4.2.1	ID3 算法	82
4.2.2	改进算法	88
	思考题	95
第5章	基于关联规则的数据挖掘技术	97
5.1	基本概念及主要算法	97
5.1.1	基本概念	97
5.1.2	关联规则挖掘种类	98
5.1.3	关联规则挖掘算法综述	99

5.2 在线挖掘关联规则算法的改进	108
5.2.1 在线挖掘关联规则算法 Carma 简介	109
5.2.2 对 Phase I 的改进	111
5.2.3 对 Phase II 的改进	117
5.3 关联规则并行化挖掘算法	118
5.3.1 其他并行算法的回顾	118
5.3.2 IDD 并行算法	119
思考题	124
第 6 章 基于分类规则的数据挖掘技术	125
6.1 基于粗糙集合的分类方法	125
6.1.1 粗糙集合的历史与发展	125
6.1.2 粗糙集合的基本概念	131
6.1.3 粗糙微积分	139
6.1.4 基于粗集的数据过滤方法	146
6.1.5 RS 代数的公理化方法	152
6.1.6 可变精度粗集中的近似空间	158
6.1.7 知识表达逻辑	160
6.2 基于模糊集合的分类方法	169
6.2.1 模糊集合与凸模糊集	169
6.2.2 模糊关系及其基本性质	185
6.3 贝叶斯分类与推进方法分类	194
6.3.1 贝叶斯分类	194
6.3.2 推进方法分类	196
思考题	198
第 7 章 基于聚类规则的数据挖掘技术	199
7.1 聚类原理	199

7.1.1	属性聚类	200
7.1.2	概念聚类	202
7.2	聚类分析中的数据类型	205
7.2.1	区间标度变量	205
7.2.2	二元变量	206
7.2.3	混合类型的变量	207
7.3	相似性测度	208
7.3.1	样本点间的相似性测度	208
7.3.2	类与类之间的相似性测度	210
7.4	硬聚类	211
7.5	软聚类	213
7.6	模糊聚类	215
7.6.1	HCM 聚类方法	215
7.6.2	FCM 聚类方法	216
7.6.3	快速 FCM 聚类方法	217
7.7	空间对象聚类	221
	思考题	224
第 8 章	基于 Web 的数据挖掘技术	225
8.1	Web 挖掘概述	225
8.1.1	一些基本概念	227
8.1.2	Web 挖掘内容	230
8.1.3	Web 挖掘难点	233
8.2	Web 结构挖掘	235
8.2.1	Web 结构挖掘的意义	235
8.2.2	超链分析与页面分类	237
8.3	Web 内容挖掘	242
8.3.1	Web 信息获取	242

8.3.2	Web 信息清理	245
8.3.3	Web 文本挖掘	250
	思考题	258
第 9 章	基于数据挖掘技术的智能决策研究框架	259
9.1	智能化交互式人机界面	259
9.2	问题求解器	261
9.3	方案设计决策支持	261
9.4	广义知识库管理系统	262
9.5	知识发现过程与数据挖掘管理器	264
	思考题	264
	参考文献	265

第 1 章 绪 论

半个多世纪以来,人工智能(AI)获得了很大发展并引起众多学科的日益重视,成为一门具有广泛应用的交叉学科和前沿学科。

1.1 人工智能综述

1.1.1 人工智能的发展历史

人工智能是计算机科学、控制论、信息论、神经生理学、语言学等多种学科互相渗透而发展起来的一门学科。人工智能的发展虽然已走过了 40 多年的历程,但是,人工智能至今尚无统一的定义。尽管学术界有各种各样的说法和定义,但就其本质而言,人工智能是研究、设计和应用智能机器或智能系统,来模拟人类智能活动的的能力,以延伸人类智能的科学。人类智能活动的的能力是指人类在认识世界和改造世界的活动中,经过脑力劳动表现出来的能力。一般地说,可概括如下:

(1)通过视觉、听觉、触觉等感官活动,接受并理解文字、图像、声音、语言等各种外界信息,这就是认识和理解外界环境的能力。

(2)通过人脑的生理与心理活动以及有关的信息处理过程,将感性知识抽象为理性知识,并能对事物运行的规律进行分析、判断和推理,这就是提出概念、建立方法、进行演绎和归纳推理、做出决策的能力。

(3)通过教育、训练和学习,日益丰富自身的知识和技能,这就

是学习的能力。

(4)对不断变化的外界环境(如干扰、刺激等外界作用)能灵活地做出正确的反应,这就是自适应能力。

不论从什么角度来研究人工智能,都是通过计算机等现代工具来实现的。计算机科学与技术的飞速发展和计算机应用的日益普及,为人工智能的研究和应用奠定了良好的物质基础。人工智能的发展使计算机更聪明、更有效,与人更接近。

1. 人工智能的起源

自古以来,人类对人工智能就有持久、狂热的追求,并凭借当时的认识水平和技术条件,设法用机器来代替人的部分脑力劳动,用机器来延伸和扩展人类的某种智能行为。例如,公元前 900 多年,我国就有歌舞机器人传说的记载。12 世纪末至 13 世纪初,西班牙的一位神学家和逻辑学家曾试图制造能解决各种问题的通用逻辑机。17 世纪,法国物理学家和数学家巴斯卡(B. Pascal)制成了世界第一台会演算的机械加法器并获得实际应用。随后,德国数学家和哲学家莱布尼兹(G. W. Leibniz)在加法器的基础上发展并制成了进行全部四则运算的计算机,他还提出了逻辑机的设计思想,即通过形式逻辑符号化,对思维进行推理计算。这种“万能符号”和“基于符号的推理计算”的思想是“智能机器”的萌芽,因而他被誉为数理逻辑的奠基人。进入 20 世纪后,人工智能相继出现了若干开创性的工作。1936 年,年仅 24 岁的英国数学家图灵(A. M. Turing)在他的一篇“理想计算机”的论文中,提出了著名的图灵机模型。1945 年,他进一步论述了电子数字计算机的设计思想。1950 年,他又在“计算机能思维吗?”一文中提出了机器能够思维的论述。1938 年,德国工程师苏斯(Zuse)研制成第一台累计数字计算机 Z-1。1946 年,在美国诞生了世界上第一台电子数字计算机 ENIAC。在同一时代,控制论和信息论的创立,生物学家设计的脑模型等,都为人工智能学科的诞生做出了理论和实验

工具的巨大贡献。

1956 年的一次历史性聚会被认为是人工智能学科诞生的标志。1956 年夏季,在美国达特莫斯(Dartmouth)大学,由当时的年轻数学助教、现任斯坦福大学教授的麦卡锡(J. McCarthy)联合他的三位朋友——哈佛大学年轻数学和神经学家、现任麻省理工学院教授的明斯基(M. L. Minsky),IBM 公司信息研究中心负责人洛切斯特(N. Lochester)和贝尔实验室信息部数学研究员香农(C. E. Shannon)——共同发起,邀请 IBM 公司的莫尔(T. Moore)和塞缪尔(A. L. Samuel)、麻省理工学院的塞尔夫利奇(O. Selfridge)和索罗莫夫(R. Solomonoff)、兰德(RAND)公司和卡内基工科大学的纽厄尔(A. Newell)和西蒙(H. A. Simon)等 10 名年轻学者,举办了为期两个月的学术讨论会,讨论机器智能问题。经麦卡锡提议,在会上正式决定使用“人工智能”(Artificial Intelligence)这一术语,从而开创了人工智能作为一门独立学科的研究方向。麦卡锡因而被称为人工智能之父。从此在美国开始形成了以人工智能为研究目标的几个研究组,如纽厄尔和西蒙的 Carnegie-RAND 协作组,明斯基和麦卡锡的 MIT 研究组,塞缪尔的 IBM 工程研究组等。

1956 年,人工智能的研究取得了两项重大突破。第一项是纽厄尔、肖(J. Shaw)和西蒙的研究组编制了一个逻辑理论程序 LT(The Logic Theory Machine),模拟人们用数理逻辑证明定理的思想,采用分解、代入、替换等规则,证明了怀特海(A. N. Whitehead)和罗素(B. A. W. Russell)合著的《数学原理》第二章中的 38 条定理。1963 年,修订的程序在大机器上终于完成了该章中全部 52 条定理的证明。一般认为,这是用计算机模拟人的高级思维活动的一个重大成果,是人工智能研究的真正开端。第二项是 IBM 工程研究组的塞缪尔研制的西洋跳棋程序。这个程序可以像一个优秀棋手那样,向前看几步来下棋。尤其是它具有自学习、自组

织、自适应的能力,能在下棋过程中积累经验,不断提高棋艺。它能学习棋谱,在学习了 175 000 多个棋局后,可以根据棋局猜测棋谱所有推荐的走步,准确度达 48%,这是机器模拟人类学习过程的一次极有意义的探索。1959 年,这个程序战胜了设计者本人;1962 年,它又击败了美国某个州的跳棋冠军。

1957 年,纽厄尔、肖和西蒙通过心理学实验,发现了人在问题求解时思维过程的一般规律,大致可分为三个阶段:

- (1)先思考出大致的解题计划;
- (2)根据记忆中的公理、定理和推理规则组织解题过程;
- (3)进行方法和目的分析,不断修正解题计划。

基于这一规律,他们于 1960 年合作编制成功一种不依赖于具体领域的通用问题求解程序 GPS(General Problem Solver),能求解 11 种不同类型的问题。

1959 年,麻省理工学院研究组的麦卡锡发表了表处理语言 LISP。由于 LISP 语言可以方便地处理符号,很快成为人工智能程序设计的主要语言。LISP 语言武装了一代人工智能科学家,时至今日,仍然是研究人工智能的重要工具。

一连串的研究成果使醉心于人工智能远景的学者们做出了过于乐观的预言。1958 年,纽厄尔和西蒙曾充满自信地认为:在 10 年内,计算机将成为世界的象棋冠军;计算机将要发现和证明重要的数学定理;计算机将能谱写具有优秀作曲家水平的乐曲;大多数心理学理论将在计算机上形成。有人甚至断言:20 世纪 80 年代将是全面实现人工智能的年代,到了 2000 年机器的智能可以超过人的智能。

但是,事情的发展远非如此理想。塞缪尔的下棋程序在获得州的冠军之后,再也没有当上全国冠军。自然语言的机器翻译是人工智能研究最早并取得实验性成果的研究方向之一。人们以为只要用一部双向互译字典和某些语法知识即可很快地解决自然语

言之间的互译问题,实际上,由机器翻译出来的文字有时会出现十分荒谬的错误。

自从人工智能形成一个学科之后,许多学者遵循的指导思想是:研究和总结人类思维的普遍规律,并用计算机来模拟人类的思维活动。他们认为,实现这种计算机智能模拟的关键是建立一种通用的符号逻辑运算体系。但是,由于人类的认知和思维过程是一种非常复杂的行为,至今仍未能被完全解释,也由于现实世界的复杂性和问题的多样性,老一辈人工智能科学家为之奋斗的通用逻辑推理体系至今也没有创造出来。他们早期的代表作——通用问题求解程序 GPS——的通用性受到严格的限制,只能对具有相当小的状态集和良定义的形式规则的问题有效。人工智能的早期研究只能停留在实验室里,作为研究的实验系统或演示系统,不能解决实际问题。科学家们开始对人工智能探索人类思维普遍规律的研究战略思想进行反思。

2. 人工智能的发展

20 世纪 60 年代中期以后,人工智能由追求万能、通用的一般研究转入特定的具体研究,通用的解题策略与特定领域的专业知识及实际经验结合,产生了以专家系统(ES, Expert System)为代表的基于知识的各种人工智能系统,使人工智能走向社会,走向实际应用研究。斯坦福大学当时的年轻教授费根鲍姆(E. A. Feigenbaum)重新举起了英国 16 世纪哲学家和自然科学家培根(Francis, Bacon)的旗帜:“知识就是力量”,于 1965 年开创了基于知识的专家系统这一人工智能研究的新领域。与通用问题求解程序 GPS 的系统不同,专家系统并不试图发现很强有力的和很通用的问题求解方法,而是把研究范围缩小在一个特定的相对狭小的专业领域中。人类专家之所以成为专家,是因为他拥有解决自己专业领域问题的大量专门知识,包括各种有用的经验。

在费根鲍姆的主持下,第一个专家系统课题 DENDRAL 化学

分子结构分析系统于 1965 年在斯坦福大学开始研究,1968 年研制成功。该系统能根据质谱仪数据推断未知有机化合物的分子结构。它是一个启发式系统,把化学专家关于分子结构质谱测定法的知识结合到控制搜索的规则中,从而能迅速消去不可能为真的分子结构,避免了搜索空间以指数级膨胀。通过产生全部可能为真的分子结构,它甚至可以找出那些人类专家可能漏掉的结构。DENDRAL 及附属的 CONGEN 系统商品化后,每天为上百个国际用户提供化学结构的解释。这一研究成果使人们看到,在某个专门领域里,以知识为基础的计算机系统完全可能相当于这个领域里的人类专家的作用。

MACSYMA 系统是麻省理工学院于 1968 年开始研制的大型符号数学专家系统。该系统从应用数学家那里获得了几百条关于一个表达式与另一等价表达式之间转换的规则,擅长于易引起组合爆炸的符号表达式的化简,能执行微分、积分、解方程、泰勒级数展开、矩阵运算、向量代数等 600 多种不同的数学符号运算。1971 年研制成功后,由于它具有很强的与应用分析相结合的符号运算能力,很多数学和物理学研究人员及各类工程师争相使用 MACSYMA 系统,遍及美国各地的很多用户每天都通过 ARPA 网与它联机工作达数小时。

在 DENDRAL 和 MACSYMA 的影响下,在化学、数学、医学、生物工程、地质探矿、石油勘探、气象预报、地震分析、过程控制、计算机配置、集成电路测试、电子线路分析、情报处理、法律咨询和军事决策等各方面出现了一大批专家系统。著名的 MYCIN 系统就是斯坦福大学人工智能研究所于 1973 年开始研制的一个诊断和治疗细菌感染性血液病的专家咨询系统。该系统可以看成是 DENDRAL 系统的直接后继者,并且具有更广泛的影响。MYCIN 拥有的知识包括约 450 条“前提-结论”型的关于细菌性血液感染的诊疗规则,系统根据提供的数据和主动向医生询问获得

的数据,运用相应的规则进行推理,最终给出对患者进行诊断和治疗方面的咨询性建议。经专家小组对医学专家、实习医生及 MYCIN 系统的行为进行正式测试评价,认为 MYCIN 的行为超过了临床医生助手的作用,尤其在诊断和治疗菌血症和脑膜炎方面有相当高的准确率,在 MYCIN 系统框架基础上建立的肺功能专家系统 PUFF 曾在旧金山太平洋医疗中心使用过相当长的一段时间。

由拉特格尔斯(Rutgers)大学于 20 世纪 70 年代中期开发的 CASNET 是一个诊断和治疗青光眼的专家咨询系统。CASNET 的知识不使用 MYCIN 的静态规则的表示方式,而是按因果关系把疾病表示为与病理生理状态相连接的网络。系统共有 150 个状态、350 个测试、50 个分类表。系统利用因果网络推断疾病类型,预测治疗效果。此外,系统还包括广泛的医学参考资料,能对疾病做出相当完善的解释。经评价,该系统接近专家水平,曾被美国和日本一些学术团体用于疾病研究。

CADUCEUS(原名 INTERNIST)系统也是 20 世纪 70 年代研制的医疗咨询专家系统,由匹兹堡大学的计算机专家和内科专家合作研制,用于内科疾病诊断。它的知识表示方式与 CASNET 类似,它的知识库是当时专家系统中最大的知识库之一,据称拥有比人类任何个人都多的内科知识。它用疾病树表达疾病分类知识,知识库中包含 500 多种内科疾病及 10 万条疾病与症状的相关关系。它还采用了一些复杂的策略来区别和组合多种疾病。因此,它能正确地对人类专家感到棘手的复杂病案做出诊断。CADUCEUS 用于医学教学和某些由美国国家健康研究机构赞助的医疗服务点。

HEARSAY I 和 II 是卡内基-梅隆大学于 20 世纪 70 年代先后研制的两个语音理解系统。HEARSAY II 通过称为“黑板”的全局数据库来组合专家的知识。它的语音理解能力可以和一个

10 岁的孩子接近,但离专家水平甚远。这是因为语音理解是一个更为困难的研究方向。

PROSPECTOR 是一个著名的地质勘探专家系统,由斯坦福大学人工智能研究所于 1976 年开始研制。它能帮助地质学家解释地质矿藏数据,提供硬岩石矿物勘探方面的咨询,包括勘探评价、区域资源估值、钻井井位选择等。该系统的知识来自地质学家的矿床模型,用似然推理网络表示知识,网络的节点为有关探测证据和重要地质假设的各种断言,有向弧表示所连接的节点之间的推理规则,用贝叶斯(Bayes)概率推理处理不确定的数据和知识。该系统拥有 12 种矿藏知识库,共有 1 100 多条规则,另有 400 种岩石和地质术语。它的能力足以与地质学家相比,并已在实际应用中取得巨大的经济效益,例如,它发现了一个钼矿,据说该系统在这个钼矿勘探与开采中的使用价值可能超过 1 亿美元。

XCON(原名 R1)是由卡内基-梅隆大学从 1978 年开始研制的一个 VAX 计算机配置专家系统,它能根据 DEC 公司的 VAX 机的用户对每一部件的需求订单,以及关于部件组合方面的上千条规则型约束知识,形成一个功能上可以接受的 VAX 系统配置。经评价小组的正式性能测试,认为该系统具有足够的专家水平。后来,系统又进行了多次改进和发展,最后成为一个成熟的系统,在 DEC 公司担负了日常的 VAX 计算机系统配置任务,每年为 DEC 公司可节省 1 500 万美元的设计费用。

专家系统的一系列研究成果展示了人工智能应用的广阔前景,社会各界对人工智能的兴趣与应用需求与日俱增。卫生界因某些医学专家系统的成功而受到极大鼓舞;军界期望它能开辟军事决策和指挥自动化的新局面;工商企业界出自对新技术的厚望,对人工智能的热情迅速高涨起来,过去对人工智能研究的态度十分保守的 IBM 公司也很快改变了态度。20 世纪 70 年代后期,随着专家系统技术的逐渐成熟和应用领域的不断开拓,人们从各种