知识发现与智能决策

张文宇　薛惠锋　薛　昱　苏锦旗　著

科学出版社

知识发现与智能决策

张文宇　薛惠锋　薛　昱　苏锦旗　著

科学出版社
北京

内 容 简 介

本书介绍知识发现、人工智能、数据仓库、联机分析处理和智能决策的基本概念与相关理论基础；分析知识发现与数据挖掘的对象与模式；综述数据预处理的作用和方法；深入探讨基于符号推理的数据挖掘方法、基于信息论思想的数据挖掘方法、基于进化思想的数据挖掘方法、基于集合论的数据挖掘方法和基于统计分析的数据挖掘方法，并将实例融入算法的具体应用；阐释智能决策支持系统，并对数据库与数据库管理系统、模型库与模型库管理系统、方法库与方法库管理系统、知识库与知识库管理系统以及人机对话管理系统进行详细说明，提出系统的逻辑框架和实现方案；最后给出知识发现与智能决策支持系统的应用案例。

本书可供从事知识发现与智能决策研究与开发的专业人士、技术管理人员以及从事知识发现与智能决策应用人员阅读，同时也可供高等院校计算机、管理和信息相关专业的教师与学生阅读。

图书在版编目(CIP)数据

知识发现与智能决策/张文宇等著. —北京：科学出版社，2014.11

ISBN 978-7-03-041320-8

Ⅰ.知… Ⅱ.张… Ⅲ.知识管理-智能决策-研究 Ⅳ.G302

中国版本图书馆CIP数据核字（2014）第143795号

责任编辑：李 敏 周 杰 / 责任校对：刘小梅

责任印制：赵德静 / 封面设计：李姗姗

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

中国科学院印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2015年1月第 一 版 开本：787×1092 1/16

2015年1月第一次印刷 印张：24 3/4 插页：2

字数：750 000

定价：100.00元

（如有印装质量问题，我社负责调换）

序

知识发现与智能决策的研究兴起于20世纪80年代，经过20多年的发展已经蔚为壮观。由薛惠锋教授提出总体思想和框架，并指导张文宇、薛昱和苏锦旗博士共同完成的《知识发现与智能决策》一书，是在作者前期出版的《信息港理论与实践》、《智能数据挖掘技术》、《数据挖掘与粗糙集方法》、《物联网智能技术》等著作基础上，站在智能信息科学研究的高度，以大量研究文献及作者承担完成的重点课题为背景和依托，有针对性地对智能决策理论框架和典型知识发现方法这两大主题进行的全面而深入的探究。

该书紧贴国内外学科动态，深入论述知识发现与智能决策的实质，合理分析新提出的学术观点。在展开具体研究时，能清晰把握知识发现与智能决策的理论、方法、技术和实际应用等方面的现状和趋势，在内容逻辑和形式体例安排上力求科学、合理、严密和完整。

全书共分四大部分。第一部分系统而全面地梳理知识发现与智能决策技术的研究背景、基础理论、主体思想及数据预处理过程，为后续的详细探讨知识发现做铺垫；第二部分从学科的宽度、专业的深度、哲学的高度详细论述了五大知识发现算法及典型应用；第三部分系统而全面研究了智能决策支持系统的框架结构，强调了以数据、信息、模型、方法、规则、知识、问题求解链为一体的科学决策流程；第四部分为知识发现与智能决策在电子商务领域、企业价值链、客户关系管理、企业财务管理、企业经营管理，产品质量管理等重要领域的实际应用。全书力求系统实用，章节安排层层推进，环环相扣，思路缜密，论证严谨。

近些年来有关知识发现和智能决策的书籍陆续问世，该书的特点在于能够把大家熟悉的研究课题做出新意，将一系列新思想、新方法引入进来，主要体现在两个方面：其一，该书以信息系统为基本研究对象，以知识发现为目的，系统而全面地阐述了五大数据挖掘算法。各种算法既有明确的应用目标，又有严格的数学模式，力求达到理论与实际、方法与应用的统一。其二，该书以系统科学和系统工程理论思想为指导，建立以知识发现为核心的智能决策框

架体系，将智能决策科学流程、决策支持系统部件及知识发现方法有机结合起来，在统一的框架下将理论、方法、工具、实例融为一体。

该书作者长期从事知识发现与智能决策的教学和研究，他们学术思想活跃，工作刻苦认真，把多年的教学实践和科学研究成果体现在该著作中，为知识发现与智能决策的研究和教学作出了重要的贡献。我有幸被邀为该书作序，期待他们理论联系实际，有新的更多的研究成果面世。预祝他们在今后的科学研究和教学中取得新的成绩！

该书的出版无疑将引起广大读者的关注和重视，从而推动智能信息处理领域的深入研究和广泛应用。在此，希望各界朋友对知识发现和智能决策的研究给予关注，并望著者百尺竿头更进一步！

中国工程院院士 王礼恒

2014 年 10 月

前　言

20世纪80年代末90年代初，国内外广泛流传着一句耐人寻味的话语：我们沉浸在数据的海洋中，却渴望着知识的淡水。这句话生动地描绘了当时人们面对海量数据的迷茫与无奈。就在这时，世界商业巨头沃尔玛从其庞大的交易数据库中演绎了一个“啤酒和尿布的故事”，揭示了一条隐藏在海量数据中的、美国人的一种行为规律：年龄在25～35岁的年轻父亲下班后经常要到超市去给婴儿买尿布，而他们中有30%～40%的人顺手为自己买几瓶啤酒。受这条简单的客户行为模式的启发，沃尔玛调整了商品布局，并策划了促销价格，结果销售量大增。这一现象引起了科学界的注意，他们将“啤酒和尿布的故事”引申为“关联规则获取”，进而为了更好地适应新的信息环境和数字资源，“知识发现”作为一个明确的概念被提出来应对相关的问题和挑战。

高效地发现新颖、可用的知识是人类知识活动的主要目标，因此密切涉及信息和知识的各个领域，越来越多地应用知识发现。商业界发现了沃尔玛迅猛发展的秘诀，纷纷效仿。电信行业也沸腾了，各公司争先恐后地利用知识发现这一锐利武器解决它们面临的最紧迫的问题（如客户分群、客户流失原因及预测、业务套餐及响应、关联消费等）。工业界也行动了，它们从堆积如山的数据中，挖掘出指导生产和管理的决策规则。知识发现通常作为一个智能和决策辅助模块被嵌入到数据管理、信息系统、电子商务平台或其他的相关信息和数据应用中，帮助人们查找、获取信息和数据，并且根据不同的需求生产相应模式的知识，以便给出智能回答作为决策的支持。知识发现的技术也被很好地应用在各种效益效率分析、潜在关系预测和优化建模等关键领域，是人工智能领域的关键应用，它有效地提高了信息需求者获得知识的质量。

在知识发现的发展历程中，早期主要以算法研究为主，主要目标是解决海量数据和信息的处理与挖掘问题。知识发现的算法研究也通常被归为数据挖掘的研究，已经产生了大量的研究成果，形成了较为丰富的产品。但是，从知识发现的实质来讲，它是一个综合的知识活动和知识生产的过程，涉及规律、策

略和技术的集成，以及多学科和领域之间的相互渗透。孤立的算法和技术的研究难以形成有效的应用，必须结合到发现方法和应用的研究才能体现出更好的效果和更大的价值。因此，知识发现研究的重点也越来越转向基于领域和服务、面向智能决策的综合应用研究。

为方便学生、教师、研究人员、专业人士以及企事业单位领导层和生产一线工作人员更好地理解知识与发现智能决策概念和技术，实现成功的知识发现智能应用，我们完成的这本书用大量的例子、简洁的语言描述关键的技术和算法，涵盖的领域包括知识发现与智能决策的相关概念理论、数据预处理、数据挖掘相关算法、智能决策支持系统。全书力求严谨求是，注意基本概念、基本知识、基本理论和相关术语正确理解与准确表达；从实践到理论，再从理论到实践，把抽象的理论与生动的案例有机地结合起来，使读者在理论与实践的交融中对知识发现和智能决策有全面和深入的理解与掌握；对知识发现与智能决策的理论、方法、技术和实际应用等各方面有清晰的现状了解和趋势把握，拓展读者的视野；在内容逻辑和形式体例上力求科学、合理、严密和完整，使之系统化和实用化。

本书由薛惠锋提出总体思想和框架，由张文宇、薛惠锋、薛昱、苏锦旗撰写，其中，第 1 ~3 章由西安邮电大学张文宇教授撰写，第 4 ~6 章由薛昱博士撰写，第 7 ~10 章由西安邮电大学苏锦旗博士撰写，第 11 章由中国航天社会系统科学与工程研究院薛惠锋教授撰写。中国航天系统科学与工程研究院侯俊杰、张文涛、张刚研究员在百忙中对本书的写作给予了悉心指导，全书的校对、修改和制图由西安邮电大学研究生马月、张宇飞、陈星、栾婧、孟旋、许明健、夏砚波、王秀秀、陶蓉、王磊等共同完成，在此表示感谢。

知识发现与智能决策是一门博大精深的学问，对其内容的理解，仁者见仁，智者见智。这些年来，知识发现与智能决策本身发生了很大的变化，尽管作者力求“与时俱进”，但也难免挂一漏万，书中的内容安排仅仅建立在作者的有限认识基础上，由于编写时间仓促，加之水平有限，书中难免有疏漏和不妥之处，恳请读者批评指正。

作　者

2014 年 8 月

目　　录

第1章 绪 论

1.1 知识发现

在许多领域中，随着数据的不断增多，一些大型数据库的规模已经远远超过人工所能分析的程度，因此数据库和知识发现（knowledge discovery in database，KDD）技术应运而生（李徽和李宛州，2001）。知识发现也是市场竞争的需要，它为决策者提供重要的、前所未有的信息或知识，从而产生不可估量的效益。

1.1.1 知识发现的历程

随着数据库系统的广泛开发和数据库技术的迅速发展，数据以前所未有的速度大量聚集在计算机中，但与之相配合的数据分析和知识提取技术在相当长一段时间里没有大的进展，使得存储的大量原始数据没有被充分利用，没有转化成为指导生产的“知识”，而是出现了“数据的海洋，知识的荒漠”这样一种奇怪的现象。于是，知识发现在这种背景下应运而生，并很快发展成为国际上数据库和信息决策领域最前沿的研究方向之一。

知识发现的研究经历了从机器学习到机器发现再到知识发现几个阶段，从20世纪80年代末，人们开始研究知识发现，1989年8月在美国底特律召开的第11届国际人工智能联合会议的专题讨论会上首次出现知识发现这个术语，法耶兹（Fayyad）首次给出了知识发现的定义“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”。随后在1991年、1993年和1994年都举行了知识发现专题讨论会，集中讨论海量数据分析算法、数据统计、知识表达、知识运用等问题。随着知识发现在学术界和工业界的影响越来越大，知识发现组委会于1995年把专题讨论会更名为国际会议，并改为大会代表自愿报名参加。1995年在加拿大蒙特利尔市召开了第一次知识发现国际学术会议，以后每年召开一次。

1.1.2 知识发现的内容

在知识发现'96国际会议上对知识发现做了如下定义：知识发现是识别出存在于数据库中有效的、新颖的、具有潜在效果的乃至最终可理解的模式的非平凡过程。知识发现是将数据变成信息、信息变为知识、知识形成策略、策略构成智能的活动，从而指导人类有效地分析问题和解决问题。知识发现过程从数据矿山中找到蕴藏的知识金块，将为知识创新和知识经济的发展作出积极的贡献。

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、医

疗卫生等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，可以是结构化、半结构化、非结构化的。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、方程或语义网络等。

数据库知识发现的研究非常活跃。在法耶兹的定义中，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 F（如关系数据库中的记录），模式是一个用语言 L 来表示的一个表达式 E，它可用来描述数据集 F 的某个子集 FE，E 作为一个模式要求它比对数据子集 FE 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一个过程，涉及数据准备、模式搜索、知识评价，以及反复地修改求精；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。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（decision support system，DSS）中可提高经济效益。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。

1.1.3 知识发现的过程

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，因此它还有很多不同的术语名称。除了知识发现外，主要还有如下若干种称法：数据挖掘（data mining）、知识抽取（information extraction）、信息发现（information discovery）、智能数据分析（intelligent data analysis）、探索式数据分析（exploratory data analysis）、信息收获（information harvesting）和数据考古（data archeology）等。其中最常用的术语是"知识发现"和"数据挖掘"。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（management inform ation system，MIS）领域，而知识发现则主要流行于人工智能和机器学习领域。

知识发现过程（图 1-1）可粗略地理解为三部曲：数据准备、数据挖掘和结果的解释评估。

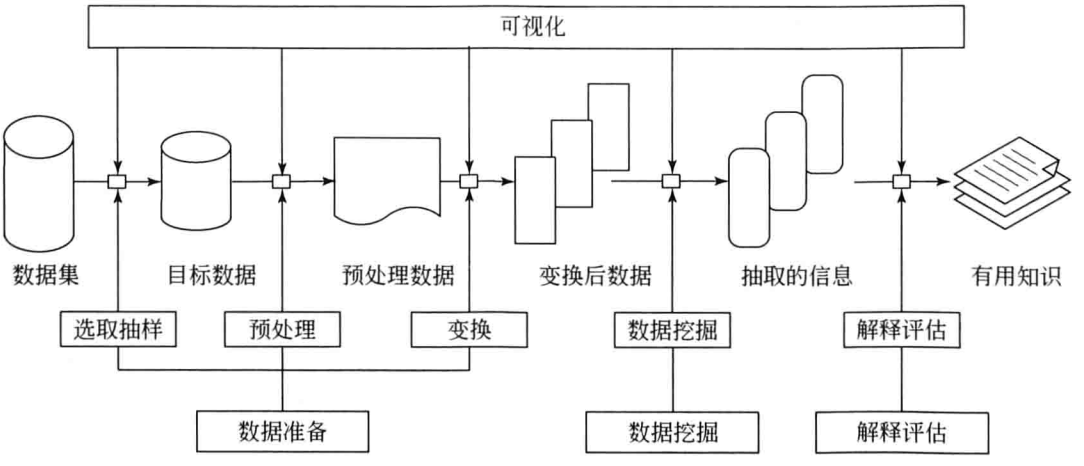

图 1-1 知识发现过程示意图

1.1.3.1 数据准备

数据准备又可分为三个子步骤：数据选取（data selection）、数据预处理（data preprocessing）和数据变换（data transformation）。数据选取的目的是确定发现任务的操作对象，即目标数据（target data），它是根据用户的需要从原始数据库中抽取的一组数据。原始数据库可以是异构的数据库和多源性数据文件。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，以便于概念性归纳）等。当数据挖掘的对象是数据仓库时，数据预处理已经在生成数据仓库时完成了，主要是通过在源数据中抽取数据，按数据仓库的逻辑数据模型的要求进行数据转换，再按物理数据模型的要求装载到数据仓库中去，即进行数据抽取、转换、加载（extract、transform、load，ETL）过程。数据变换的主要目的是消减数据维数或降维（dimension reduction），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。

1.1.3.2 数据挖掘

数据挖掘阶段首先要确定挖掘的任务或目的是什么，如数据总结、概念描述、分类、聚类、关联规则发现或序列模式发现和相关性分析等。确定了挖掘任务后，就要决定使用什么样的挖掘算法。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挖掘；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（descriptive）、容易理解的知识，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（predictive）知识。完成上述准备工作后，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，也是目前研究人员主要的努力方向，但要获得好的挖掘效果，必须对各种挖掘算法的要求或前提假设有充分的理解。

1.1.3.3 结果解释和评估

数据挖掘阶段发现的模式，经过用户或机器的评估，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，这时则需要让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，甚至换一种挖掘算法（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，知识发现最终是面向人类用户的，因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂的另一种表示，如把分类决策树转换为“if…then…”规则等。

知识发现过程中要注意以下几点：

1）数据挖掘仅仅是整个过程中的一个步骤。数据挖掘质量的好坏受两个要素的影响：一是所采用的数据挖掘技术的有效性，二是用于挖掘的数据的质量和数量（数据量的大小）。如果选择了错误的数据或不适当的属性，或对数据进行了不适当的转换，则挖掘的

结果是不会令人满意的。

2）整个挖掘过程是一个不断反馈的过程。比如，用户在挖掘途中发现选择的数据不太好，或使用的挖掘技术产生不了期望的结果，这时，用户需要重复先前的过程，甚至重新开始。

3）可视化在数据挖掘的各个阶段都扮演着重要的作用。特别是在数据准备阶段，用户可能要使用散点图、直方图等统计可视化技术来显示有关数据，以对数据有一个初步的理解，从而为更好地选取数据打下基础。在挖掘阶段，用户则要使用与领域问题有关的可视化工具。在表示结果阶段，则可能要用到可视化技术。

1.1.4 知识发现的方法

知识发现的方法大致可分为如下几大类。

1.1.4.1 统计方法

统计方法是从事物的外在数量上的表现去推断该事物可能的规律性。科学规律性的东西一般总是隐藏得比较深，最初总是通过统计分析从其数量表现上看出一些线索，然后提出一定的假说或学说，再作深入的理论研究。当理论研究提出一定的结论时，往往还需要在实践中加以验证。就是说，观测一些自然现象或专门安排的实验所得资料，是否与理论相符、在多大的程度上相符、可能朝哪个方向偏离等问题，都需要用统计分析的方法处理。

近百年来，统计学得到了极大的发展。我们可用图 1-2 的框架粗略地刻画统计学发展的过程。

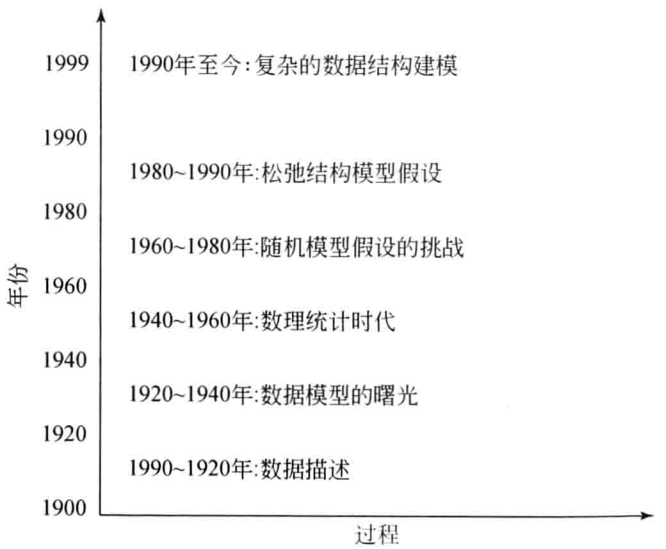

图 1-2 统计学发现过程图

其中，从 1960 ~ 1980 年，引导这一革命的是 20 世纪 60 年代的四项发现：①吉洪诺

夫（Tikhonov）、伊万诺夫(Ivanov)和菲利浦(Philips)发现的关于解决不适定问题的正则化原则。②帕仁(Parzen)、罗森布拉特(Rosenblatt)和陈瑟夫（Chentsov）发现的非参数统计学。③瓦普尼克（Vapnik)和车温尼克思（Chervonenkis）发现的在泛函数空间的大数定律及其与学习过程的关系。④柯尔莫哥洛夫（Kolmogorov）、索洛莫诺夫（Solomonoff）和沙坦（Chaitin）发现的算法复杂性及其与归纳推理的关系。

这四项发现也成为人们对学习过程研究的重要基础。下面我们列出与统计学有关的机器学习方法。

（1）传统方法

统计学在解决机器学习问题中起着基础性的作用。传统的统计学所研究的主要是渐近理论，即当样本趋向于无穷多时的统计性质。统计方法主要考虑测试预想的假设和数据模型拟合。它依赖于显式的基本概率模型。统计方法处理过程可分为三个阶段：①搜集数据：采样、实验设计。②分析数据：建模、知识发现、可视化。③进行推理：预测、分类。

常见的统计方法有回归分析（多元回归、自回归等）、判别分析（贝叶斯判别、费歇尔判别、非参数判别等）、聚类分析（系统聚类、动态聚类等）、探索性分析（主元分析法、相关分析法等）等。

（2）模糊集

模糊集是表示和处理不确定性数据的重要方法。模糊集不仅可以处理不完全数据、噪声或不精确数据，而且在开发数据的不确定性模型方面是有用的，可以提供比传统方法更灵巧、更平滑的性能。

（3）支持向量机

支持向量机（support vector machine，SVM）建立在计算学习理论的结构风险最小化原则之上，其主要思想针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率。而且SVM有一个重要的优点就是可以处理线性不可分的情况。

（4）粗糙集

粗糙集（rough set）理论由帕夫拉克（Z. Pawlak）在1982年提出。它是一种新的数学工具，用于处理含糊性和不确定性问题，在数据挖掘中发挥着重要的作用。粗糙集是由集合的下近似、上近似来定义的。下近似中的每一个成员都是该集合的确定成员，而不是上近似中的成员肯定不是该集合的成员。粗糙集的上近似是下近似和边界区的合并。边界区的成员可能是该集合的成员，但不是确定的成员。可以认为粗糙集是具有三值隶属函数的模糊集，即是、不是、也许。与模糊集一样，它是一种处理数据不确定性的数学工具，常与规则归纳、分类和聚类方法结合起来使用，很少单独使用。

1.1.4.2 机器学习

西蒙（Simon）对机器学习的定义是："如果一个系统能够通过执行某种过程而改进它

的性能，这就是学习”。这个说法的要点是：第一，学习是一个过程；第二，学习是相对一个系统而言的；第三，学习改变系统性能。过程、系统和改变性能是学习的三个要点。对上述说法，第一点是自然的。第二点中的系统则相当复杂，一般是指一台计算机，但是，也可以是计算系统，甚至包括人的人机计算系统；第三点则只强调“改进系统性能”，而未限制这种“改进”的方法。显然，西蒙对学习的这个说法是思辨的，但对计算机科学家来说，这是远远不够的。计算机学家更关心对不同系统实现机器学习的过程，以及改变性能的效果。图 1-3 给出一个简单的学习系统模型。

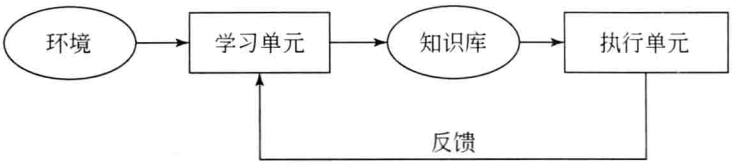

图 1-3　简单的机器学习系统模型

在 20 世纪 50 年代，机器学习采用了两种不同的研究方法。在控制理论中，使用多项式等为基函数，利用优化的方法建立模型，以刻画被控对象的行为，这个过程在控制理论中称为辨识或参数估计，甚至笼统地称为建模。而以罗森布拉特（Rosenblatt）的感知机为代表的研究，则是从心理学家麦卡洛克（McCulloch）和数理逻辑学家皮茨（Pitts）建立的神经网络和数学模型（MP 模型）出发，具体地说，就是将扩展为多个神经元的 MP 模型作为优化算法的数学基函数。但是从数学上讲，其区别仅仅是它们使用了不同的基函数，以及由此所带来的问题。这两种以优化为基础的方法至今还影响着机器学习的发展。

在 20 世纪 60 年代末，明斯基（Minsky）对感知机的批评使这类研究陷于停顿。但是，他运用数学方法对当时的人工神经网络（artificial neurel networks，ANN）模型进行了精辟的分析，指出了人工神经网络求解问题的能力局限性和问题，这个思想对 80 年代兴起的人工神经网络的研究是有意义的。从 50 年代末到 80 年代的 20 余年间，在人工智能领域中，机器学习的研究完全脱离了这种基于统计的传统优化理论为基础的研究方法，而提出一种以符号运算为基础的机器学习。这种以符号运算为基础的机器学习方法，可以从塞缪尔（Samuel）的下棋系统中发现其原型。

人工智能的研究者根据认知心理学的原理研究各种机器学习的方法，以符号运算为基础的机器学习代替了以统计为基础的机器学习，成为人工智能研究的主流。在这个时期，就学习机制来说，主要是归纳机器学习。其中代表性的学习算法有 AQ11 和 ID3。同时，使用不同学习机制的研究层出不穷。20 世纪 80 年代中期，基于解释的学习（explanation-based learning）和类比学习也引起人们极大的兴趣，特别是与类比学习原理相近的基于案例的学习（case-based learning）解决实际问题的能力强。这些研究丰富了机器学习的研究。

1984 年瓦伦特（Valiant）提出了可学习理论，并将可学习性与计算复杂性联系在一起。1986 年布鲁默（Blumer）等证明了 VC（Vapnik-Chervonenkis dimension）维度与 Valiant 的“大概逼近正确”（probably approximately correct，PAC）（Patterson and Aebersold，2003）的可学习理论之间的联系。“大概逼近正确”本身提出的理论课题也派生出被称为“计算

学习理论（computational learning theory，COLT）”的学派，并且这方面的国际会议定期召开。1995 年瓦普尼克（Vapnik）在统计学习理论研究的基础上，指出了经验风险最小的问题，提出结构风险最小化。在这一理论框架指导下，产生了支持向量机学习方法，这是一种构造性的学习方法。

达尔文进化论是一种稳健的搜索和优化机制，对计算机科学，特别是对人工智能的发展产生了很大的影响。大多数生物体是通过自然选择和有性生殖进行进化。自然选择决定了群体中哪些个体能够生存和繁殖，有性生殖保证了后代基因中的混合和重组。自然选择的原则是适应者生存，不适应者被淘汰。自然进化的这些特征早在 20 世纪 60 年代就引起了美国（密歇根大学）霍兰德（Holland）的极大兴趣。霍兰德注意到学习不仅可以通过单个生物体的适应实现，而且可以通过一个种群的许多代的进化适应发生。受达尔文进化论思想的影响，他逐渐认识到在机器学习中，为获得一个好的学习算法，仅靠单个策略的建立和改进是不够的，还要依赖于一个包含许多候选策略的群体的繁殖。考虑到他们的研究想法起源于遗传进化，霍兰德就将这个研究领域取名为遗传算法（genetic algorithm）。一直到 1975 年霍兰德出版了那本颇有影响的专著《自然系统和人工系统的适应》（*Adaptation in Natural and Artificial Systems*），遗传算法才逐渐为人所知。

20 世纪 80 年代，基于试错方法、动态规划和瞬时误差方法形成了强化学习（reinforcement learning）。1984 年萨顿（Sutton）提出了一种基于马尔可夫（Markov）过程的强化学习。1996 年，卡尔布林（Kaelbling）在总结强化学习的研究时指出，实现这种学习的手段就是自适应机制。1998 年，麻省理工学院出版社出版了萨顿和巴尔托（Barto）的著作《强化学习：导论》（*Reinforcement Learning*：*An Introduction*），将这些研究统称为适应性计算。根据西蒙的说明，这也是一种学习，但是，在机制上，这类机器学习理论不同于人工智能意义上的机器学习，其主要区别是：这类机器学习强调对变化环境的适应，这意味着，它们需要建立一种基于反馈机制的学习理论。下面列出了几种目前较为常用的机器学习方法。

（1）规则归纳

规则归纳反映数据项中某些属性或数据集中某些数据项之间的统计相关性。AQ 算法是有名的规则归纳算法。关联规则的一般形式为 X1 ∧ … ∧ Xn ：Y［C，S］，表示由 X1 ∧ … ∧ Xn 可以预测 Y，其可信度为 C，支持度为 S。近年来提出了许多关联规则算法。

（2）决策树

决策树的每一个非终结节点表示所考虑的数据项的测试或决策。一个确定分支的选择取决于测试的结果。为了对数据集分类，从根节点开始，根据判定自顶向下，趋向终结节点或叶节点，当到达终结节点时，则决策树生成。决策树也可以解释为特定形式的规则集，以规则的层次组织为特征。

（3）案例推理

案例推理是直接使用过去的经验或解法来求解给定的问题。案例常常是一种已经遇到

过并且具有解法的具体问题。当给定一个特定问题，案例推理就检索范例库，寻找相似的范例。如果存在相似的案例，它们的解法就可以用来求解新的问题。该新问题被增加到案例库，以便将来参考。

(4) 贝叶斯信念网络

贝叶斯信念网络（恩门，2000）是概率分布的图表示。贝叶斯信念网络是一种直接的、非循环的图，节点表示属性变量，边表示属性变量之间的概率依赖关系。与每个节点相关的是条件概率分布，它描述该节点与它的父节点之间的关系。

(5) 科学发现

科学发现是在实验环境下发现科学定律。在著名的 BACON 系统（Sam，2003）中，核心算法基本上由两种操作构成：第一种操作称为双变量拟合，判定一对变量之间的关系；第二种操作是合并多对关系到一个方程中。

(6) 遗传算法

遗传算法是按照自然进化原理提出的一种优化策略。在求解过程中，通过最好解的选择和彼此组合，可以期望解的集合越来越好。在数据挖掘中，遗传算法用来形容变量间的依赖关系假设。

1.1.4.3 神经计算

神经网络是指一类新的计算模型，它是模仿人脑神经网络的结构和某些工作机制而建立的一种计算模型。这种计算模型的特点是，利用大量的简单计算单元（即神经元）连成网络，来实现大规模并行计算。神经网络的工作机理是通过学习，改变神经元之间的连接强度。

1943 年表卡洛克（McCulloch）和匹兹（Pitts）公布了他们对神经元模型的研究结果。在更为广泛的科学发展进程中，这个研究的历史意义是首次发现了人类神经元的工作方式，并给出了这种工作方式的数学描述。这项研究在科学史上的意义是非同寻常的，它第一次揭示了人类神经系统的工作方式。它对近代信息技术发展的影响也是巨大的，计算机科学与控制理论均受到这项研究的启发。由于匹兹的努力，这个研究结论并未仅仅停留在生物学的结果上，他为神经元的工作方式建立了数学模型，正是这个数学模型深刻地影响了机器学习的研究。

在 20 世纪 50 年代，以罗森布拉特（Rosenblatt）的感知机为代表的研究则是从 MP 模型出发，具体地说，就是将扩展为多个神经元的 MP 模型作为优化算法的数学基函数。20 世纪 70 年代末以来，人工智能在模拟人的某些认知活动上取得很大的进展，专家系统、智能计算机受到重视，人们突然感到了传统的人工智能系统与人的自然智能相比存在一些明显的不足。人工智能与人的自然智能相比在感知能力上的差距很大，人能够毫不费力地识别各种复杂的事物，能从记忆的大量信息中迅速找到需要的信息，人具有自适应、自学习等创新知识的能力，这些都是现有计算机无法比拟的。因此人们又重新将目标转向神经