机器学习理论与算法

张燕平 张 铃 等 编著

科学出版社

机器学习理论与算法

张燕平　张　铃等　编著

科学出版社
北京

内 容 简 介

机器学习是人工智能研究领域中一个极具发展生命力的研究应用分支，已成为信息科学领域解决实际问题的重要方法。本书集中介绍了机器学习的一些典型方法、理论和应用领域，并首次系统地给出了构造性机器学习方法——覆盖算法。为了便于读者学习和研究书中所介绍的各类典型方法，在每章中还列出了相应的算法源代码。

本书通过研究大量丰富的文献资料和科研成果，对机器学习典型算法的过去做了应有回顾，对现状做出了必要剖析，对未来进行了充分展望。

本书可供高等院校计算机、自动化、电子工程等专业的高年级本科生、研究生、教师以及相关领域的研究人员与工程技术人员参考。

图书在版编目(CIP)数据

机器学习理论与算法 / 张燕平等编著. —北京：科学出版社，2012

ISBN 978-7-03-034318-5

Ⅰ. 机… Ⅱ. ①张… Ⅲ. ①机器学习②电子计算机-算法理论
Ⅳ. ①TP181②TP301.6

中国版本图书馆 CIP 数据核字(2012)第 095857 号

责任编辑：裴 育 / 责任校对：林青梅
责任印制：张 倩 / 封面设计：耕者设计工作室

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

骏 走 印 刷 厂 印刷
科学出版社发行 各地新华书店经销

*

2012 年 5 月第 一 版 开本：B5(720×1000)
2012 年 5 月第一次印刷 印张：18 3/4
字数：360 000

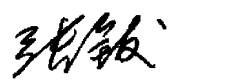
定价：60.00 元

(如有印装质量问题，我社负责调换)

序

机器学习是人工智能的重要分支，半个多世纪以来，无论在基础理论还是实际应用上均取得了重大进展，因此备受关注。由 T. M. Mitchell 提出，并得到广泛认可的机器学习的定义为“能够从经验中学习某类任务，并利用经验改善该任务性能的计算机程序”。目前的机器学习方法有基于规则和基于概率两大类。互联网与数字化技术的迅猛发展，为我们提供了海量的数据，因此数据驱动的概率方法近来得到飞速发展，并已设计出多种高速算法，被广泛地使用；相反，知识驱动的规则方法则逐渐被忽视。但是，目前概率统计方法在实际应用中遇到了瓶颈，因为要揭示数据中隐藏的复杂规律，需要大量的先验知识，如大量的标签数据，这就要求以概率为基础的机器学习能够融入大量的先验知识，但目前的概率方法难以做到这一点。机器学习的进一步发展，要求我们在关注基于数据的概率方法的同时，也要关注与了解其他基于知识的学习方法。知识表示及规则与概率方法的融合可能会成为今后新的研究热点。

该书作者张铃、张燕平及其团队长期从事人工智能、神经网络、遗传算法与粒计算等理论研究，建立了基于商空间的问题求解理论，提出了统计启发式搜索算法、构造性的学习算法以及多层规划方法等，这些与机器学习相关的研究成果都充分地反映在书中。目前已经出版的关于机器学习的书籍比较多，这些书籍均以介绍概率方法为主。该书的作者从更加宽广的角度看待机器学习，把机器学习分为搜索型、构造型与规划型三种策略，在介绍概率方法的同时，也介绍了其他的学习方法，内容丰富而全面。书中所介绍的构造性学习理论与覆盖算法、遗传算法等，其优点是容易引入启发式知识，这也许是今后概率与规则机器学习方法相融合的方向之一，正是广大读者所需要了解的，我以为这也是该书的特色所在。



中国科学院院士

2012 年 4 月于清华大学

前　言

机器学习致力于研究怎样用计算机模拟人类的学习行为和能力，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，即利用经验来改善系统自身的性能。它是人工智能的重要分支，也是计算机实现人类行为能力的重要方法。目前，机器学习的各种方法已应用到各个领域，也表现出了极大的吸引力。国际著名期刊 *Nature* 也对机器学习方法的研究成果进行过报道。国际上著名的机器学习会议（International Conference on Machine Learning，ICML）每年举行一次，报道最新的机器学习发展状况，进一步促进了机器学习领域的学术交流，从而使这门新兴的智能学习方法展现出勃勃生机。目前，机器学习已成为国际人工智能领域中备受关注的研究方向。

本书主要阐述机器学习方面最主要的一些算法并进行详细的介绍，同时吸纳了国内外许多具有代表性的最新研究成果，特别是覆盖算法的理论分析和应用。全书取材新颖、内容丰富，注重理论与实际的结合，主要介绍基于数据的机器学习，即对于一种未知的依赖关系，以观测为基础对其进行估计。现实世界中存在大量无法准确认识却可以进行观测的事物，因此这种机器学习在从现代科学、技术到社会、经济等各领域中都有着十分重要的应用。全书共 7 章：第 1 章绪论，主要介绍机器学习的思想、定义和研究现状及其面临的问题；第 2 章阐述统计学习理论与支持向量机算法；第 3 章重点阐述覆盖算法的理论基础和算法设计以及一些改进的算法，包括问题描述、领域覆盖、交叉覆盖、多侧面递进算法、核覆盖算法和概率模型的覆盖算法；第 4 章介绍当前的一个研究热点——集成学习和弱可学习理论，主要包括集成学习理论、分类器集成方法论、AdaBoost、选择性集成及其应用；第 5 章介绍数据流和增量学习的概念，详细叙述增量学习的理论；第 6 章阐述传统的遗传算法理论、现状及其应用，并给出了一些应用实例；第 7 章重点介绍决策树与贝叶斯网络的理论分析和一些常用的算法。

本书内容源于 973 计划项目（2007CB311003），国家自然科学基金项目（61073117、61175046），安徽省自然科学基金项目（11040606M145）的成果。本书能够出版，首先要感谢安徽大学计算机技术专业 2010 级博士班的张媛、吴蕾、张超、许荣斌、郑爱华等同学，尤其要感谢安徽大学计算机学院杜秀全老师，正是他

们共同的辛勤工作，使本书顺利出版。本书的出版获得安徽大学“211”工程“学术著作出版基金”和国家级特色专业(TS11483)项目资助，在此一并致谢。

由于作者的水平有限，书中难免存在不妥之处，恳请广大读者批评指正。

张燕平　张　铃

2012年1月

目　录

第1章 绪 论

人类所具有的最独特创造力在于可以通过已有经验与常识来学习并发现未知的事物，因此具备学习能力是人的一个极其重要的特征。随着科学技术的发展，人们开始探索如何制造智能机器来替代人的繁复的智力劳动，并且在某些方面已经取得了巨大成功。然而，机器不是人，它不具备人的思维、学习创造能力。如何使机器具备智能，使机器可以模拟人的大脑思维，可以像人一样地思考问题、学习新知识，就成为急需解决和发展的科学问题。机器学习就是这样的一门学科，它能够构建一些办法来有效地模拟人的大脑活动。目前，如何使机器具备拟人化的学习，进行更深层次的理解工作，还有很多问题有待探索和解决。Simon 认为，学习是一个系统对环境的适应性变化，它能够使得系统在下一次完成同样或类似的任务时更为有效。而 Michalski 认为，学习是构造或修改对于所经历事物的表示。

机器学习在人工智能的研究中占据着非常重要的地位，它逐渐成为人工智能研究的核心内容之一。现在针对机器学习的应用已遍及人工智能领域的各个分支，如专家系统、自动推理、自然语言理解、模式识别、计算机视觉、智能机器人、生物信息学等领域。在这些研究中，如何获取知识成为突出的瓶颈，人们试图采用机器学习的方法加以克服。

一般而言，机器学习的研究主要是从生理学、认知科学的角度出发，理解人类的学习过程，从而建立人类学习过程的计算模型或认知模型，并发展成各种学习理论和学习方法。在此基础上，研究通用的学习算法，进行理论上的分析，建立面向任务的具有特定应用的学习系统。

1.1 什么是机器学习

随着计算机技术的发展，人们已能够获取并存储海量数据。长期以来，研究者都在考虑如何利用这些数据，它们都表达什么样的知识。自然的，需要对这些数据进行“学习”。然而，究竟什么是学习，一直以来却众说纷纭。数学家、逻辑学家、心理学家和计算机学家都有着各自的看法，有些观点甚至差别较大。尽管如此，为了便于学科间讨论和评估学科的进展，首先需要给出一个明确统一的定义，即使这种定义是不完备的和不充分的。

1.1.1 信息爆炸

计算机网络的发展使得人们对信息的采集、传播的速度和规模达到史无前例

的水平，实现了全球的信息共享与交互，它已经成为信息社会必不可少的基础设施。据统计表明，1986 年到 2007 年期间，全世界计算能力每年增长 58%，增长最快的信息处理能力是互联网和电话网络等双向通信领域，每年增长 28%，存储量每年增长 23%；而电视和无线电广播等单向信息发布渠道则要少得多，每年增长 6%。世界上通过特殊应用设备（如电子微控制器或图像处理器）处理信息的技术能力，大约每 14 个月就翻一番，而通用计算机（如个人电脑和移动电话）每 18 个月翻一番。全球人均通信能力每 2 年 10 个月就翻一番，而人均存储量大约每 3 年 4 个月增加两倍。现代通信和传播技术是由广播、电视、卫星通信和电子计算机通信等技术手段形成的复杂的网络，摆脱了传统的时间和空间障碍，将世界更进一步地融为一体。然而，这也带来了不少的副作用，如海量的信息有时让人无所适从，从如此多而复杂的海量信息中迅速而准确地获取自己最需要的信息，变得非常困难。这种现象被称为“信息爆炸”。

面对海量的信息，如何使信息资源得到有效的利用，提高信息的质量，已经成为一个世界性的亟待医治的网络顽症。因此，优化信息资源的开发、管理是使信息被有效利用的关键问题之一。例如，2009 年 5 月 18 日，美国易安信（EMC）公司对世界产生的数字化信息的比特数总量进行合计之后，发布了一个惊人的 22 位数字——4870 亿吉比特。如果将这些数字化的信息全部印成书籍并排列整齐，它们的长度将是从地球到冥王星距离的 10 倍。

针对海量数据相关问题，目前主要从技术和管理两个方面着手解决。从管理上，各国政府都颁布了相应的网络信息管理办法、条例等，但由于各国的国体、意识形态、习俗和道德观念的差异，很难有一个全球统一的标准。因此，对于全球的网络形成一个统一的控制“信息垃圾”的措施是不现实的。出于这样的认识，人们试图从技术上寻求办法。从 20 世纪 90 年代中期开始，各国日益将研究重点放在数据库技术、信息挖掘技术、信息标准化技术上，形成了信息获取技术的研究热潮，机器学习就是其中非常有效的方法之一。

1.1.2 学习的定义

学习是一种具有多侧面的现象。学习的过程可以看做是获取新的陈述性知识，通过教育或实践发展机械技能和认知能力，将新知识组织成为通用化和有效的表达形式，借助观察和实验发现新的事实和新的理论。

学习也可指学习者因经验而引起的行为、能力和心理倾向的比较持久的变化，它与成熟、疾病或药物等因素无关，而且不一定表现出外显行为。一般认为广义的学习具有三个要素：第一，主体身上必须产生某种变化，才能作出学习已经发生的推论；第二，这种变化是能相对持久保持的，因为有些因素（如适应、疲劳等）也可导致行为的暂时变化，另外学习之后的遗忘也是不可否认的；第三，主体的变化是由

他(它)与环境的相互作用而产生的,排除由成熟或先天反应倾向所导致的变化。

1.1.3 机器学习定义

机器学习(machine learning)是利用一些方法来使机器(如计算机)实现人的学习行为,以便获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。正因如此,机器学习的定义也出现很多不同的版本。例如,Langley 认为,"机器学习是一门人工智能的科学,该领域的主要研究对象是人工智能,特别是如何在经验学习中改善具体算法的性能"(Machine learning is a science of the artificial. The field's main objects of study are artifacts, specifically algorithms that improve their performance with experience)。Mitchell 的机器学习对信息论中的一些概念作了详细的解释,他认为,"机器学习是对能通过经验自动改进的计算机算法的研究"(Machine learning is the study of computer algorithms that improve automatically through experience)。Alpaydin 于 2004 年提出了他对机器学习的定义,"机器学习是用数据或以往的经验,来优化计算机程序的性能标准"(Machine learning is programming computers to optimize a performance criterion using example data or past experience)。依据本书写作目的,这里对"学习"设定一个宽广的定义,以使其包括任何计算机程序通过经验来提高某项任务处理性能的行为。

更准确地讲,"从有限观察概括特定问题世界模型的机器学习"与"从有限观察发现观测数据中暗含的各种关系的数据分析"的方法上,称其为基于数据集的机器学习。

例如,对识别学习问题,描述如下。

任务 T:识别或分类图像、文字或语音等;性能标准 P:分类的正确率;学习策略 L:已知分类的图像、文字或语音等的数据库的类别。

从有限观察概括特定问题:

任务 T:建立有限样本对应特定问题的模型;性能标准 P:预测或分类的正确率;学习策略 L:有限观测的样本集的正确分类法。

在这里,需要解决以下三个问题。

(1) 一致:假设事件 W 与给定样本集 Q 有相同的性质。例如,如果学习过程基于统计原理,则独立同分布(i. i. d)就是某种一致条件。

(2) 划分:将样本集放到 n 维空间,寻找一个定义在这个空间上的决策分界面(等价关系),使得问题决定的不同对象分在不相交的区域。

(3) 泛化:泛化能力是这个模型对世界为真程度的指标。从有限样本集合,计算一个模型,使得这个指标最大(最小)。

基于此,给出机器学习的定义:如果一个计算机程序针对某类任务 T,可以用

P 衡量的性能，根据某种学习策略 L 来不断完善，则称这个计算机程序从策略 L 中学习，这个学习是专门对任务 T 以目标 P 作为评价指标的。

1.2 机器学习的发展史

机器学习是人工智能研究比较年轻的一个分支，它的发展大体上可分为四个阶段。

第一阶段是从 20 世纪 50 年代中叶至 60 年代中叶，属于热烈时期，又称为通用学习系统的研究。这一时期基本上与人工智能学科的产生是同步的，其主要研究方法是不断修改系统的控制参数以改进系统的执行能力，不涉及与具体任务有关的知识，这种系统所应用的主要技术有神经元模型、决策论和控制论。

由于当时计算机技术水平有限，学者的研究主要停留在理论探索和构造以神经元模型为基础方面的工作，实验硬件系统只带有随机的或部分随机的初始结构。最具代表性的是被称为感知器的神经网络。系统的学习主要依赖神经元之间信号的传递。当时，麦克洛奇(McCulloch)和皮兹(Pitts)用离散决策元件模拟神经元的理论开展了应用符号逻辑来模拟神经元系统的工作；弗伦德勃(Friedberg)提出了进化过程的仿真。然而，这种神经元模型没有取得实质性的进展，最终在 60 年代末走入低谷。塞缪尔(Samuel)于 50 年代末设计了一种跳棋程序，引发另一种学习，即机械学习，并且取得了巨大成功。该程序随着使用次数的增加，它会积累性地记忆有价值的信息，很快能够达到跳棋大师级水平。

第二阶段是从 60 年代中叶至 70 年代中叶，称为冷静时期，即基于符号表示的概念学习系统研究。这个阶段的研究目标是模拟人类的概念学习过程，并采用逻辑结构或图结构作为机器内部描述。这时的人工智能研究重点也已转到符号系统和基于知识的方法研究。这一时期的工作主要有概念获取和各种模式识别系统的应用。其中，最有影响的开发工作当属温斯顿(Winston)的基于示例归纳的结构化概念学习系统、亨特(Hunt)和哈兰德(Hovland)的 CLS 以及巴查纳(Buchanan)等的 META-DENDRAL。

第三阶段是从 70 年代中叶至 80 年代中叶，称为复兴时期(基于知识的学习系统研究)。此阶段开始注重基于知识的学习系统的运用，研究者同时结合了聚类、类比推理和机器发现的工作。这一时期的工作主要有三个方面：①基于知识的方法；②开发各类学习方法；③结合生成和选择学习任务的能力。

第四阶段开始于 80 年代后期，联结学习和符号学习的深入研究使得机器学习领域的极大繁荣。主要可以分为以下几个方面：

(1) 机器学习已成为新的边缘学科并在高校形成一门课程。

(2) 结合各种学习方法,进行取长补短的多种形式的集成学习系统研究正在兴起。

(3) 机器学习与人工智能各种基础问题的统一性观点正在形成。

(4) 各种学习方法的应用范围不断扩大,一部分已形成商品。

(5) 数据挖掘和知识发现的研究已形成热潮,并在有关领域得到成功应用。

(6) 与机器学习有关的学术活动空前活跃。

尤其是近几年来,基于计算机网络的各种自适应、具有学习功能的软件系统的研制和开发都将机器学习的研究推向新的高度,网络环境已成为人工智能和机器学习的重要试验床。

1.3 机器学习的发展现状

机器学习从诞生开始至今已极其繁荣,这里有许多学者的贡献。然而,对机器学习进展产生重要影响的是以下三个发现:

(1) James 关于神经元是相互连接的发现。

(2) McCulloch 与 Pitts 关于神经元工作方式是"兴奋"和"抑制"的发现。

(3) Hebb 的学习律(神经元相互连接强度的变化)。

其中,McCulloch 与 Pitts 的发现对近代信息科学产生了巨大的影响。这项成果给出了近代机器学习的基本模型,加上指导改变连接神经元之间权值的 Hebb 学习律,成为目前大多数流行的机器学习算法的基础。这里列举出一些对机器学习发展比较重要的成果。

1954 年,Barlow 与 Hebb 在研究视觉感知学习时,分别提出了不同假设:Barlow 倡导单细胞学说,假设从初级阶段而来的输入集中到具有专一性响应特点的单细胞,并使用这个神经单细胞来表象视觉客体;而 Hebb 主张视觉客体是由相互关联的神经细胞集合体来表象,并称其为 Ensemble。这两个假设对机器学习研究有重要的启示作用。1957 年,Rosenblatt 在 McCulloch 与 Pitts 模型的基础上首先提出了感知机算法[1]。该算法主要操作是:首先,借用最简单的 McCulloch 与 Pitts 模型作为神经细胞模型;然后,根据 Hebb 集群的考虑,将多个这样的神经细胞模型根据特定规则集群起来,形成神经网络,并将其转变为下述机器学习问题。1969 年,Minsky 与 Papert 出版了对机器学习研究具有深远影响的著作 *Perceptron*(感知机)[2]。在该著作中提出了扼杀了感知机的研究方向的 XOR 问题。1986 年,Rumelhart 等提出了 BP 算法,该算法解决了 XOR 问题,使得沉寂近二十年的感知机研究方向重新获得认可[3]。1960 年,Widrow 提出了 Madaline 模型[4],他的基本思想是先将线性模型(如感知机)考虑为神经细胞模型(而不是简单的 McCulloch 与 Pitts 模型),再基于 Hebb 神经元集合体假设,将这些局部模型集群

为对问题世界的表征，由此解决线性不可分问题。

自 1992 年开始，Vapnik 将有限样本统计理论介绍给全世界，并出版了统计机器学习理论的著作[5]。主要涉及机器学习中两个相互关联的问题，泛化问题与表示问题。前者包含两个方面的内容：其一，有限样本集合的统计理论；其二，概率近似正确的泛化描述。而后者则主要集中在核函数，由此，将算法设计建立在线性优化理论之上。

近年来，许多学者对机器学习算法作了深入的研究。例如，最近邻算法对不相关特征的研究[6]。朴素贝叶斯分类器容易受冗余特征的影响，因为朴素贝叶斯分类器假定各特征间相互独立[7]。决策树算法有时会对训练集过适应，产生过于复杂的树，移除不相关和冗余特征可以产生一棵较小、容易理解的树。由此，特征选择的作用主要有：去除不相关特征、冗余特征、甚至噪声特征，提取关键特征，提高学习算法的泛化性能和运行效率，得到更加简单和容易理解的学习模型[8,9]。自 90 年代以来，许多应用领域，如基因工程[10]、文本分类[11]、图像检索[12]等，大规模数据的处理问题不断涌现。大规模数据的特征选择也对现有的特征选择算法提出了严峻的挑战，特征选择引起机器学习领域学者广泛的研究兴趣。Kira 和 Rendell 于 1992 年提出了一种著名的特性选择算法——Relief[13]，该算法属于一种特征权重算法（feature weighting algorithms），根据各个特征和类别的相关性赋予特征不同的权重，权重小于某个阈值的特征将被移除。Almuallim 和 Dietterich 于 1992 年提出了利用信息论度量度来选择相关特征[14]。1994 年，Stanford 大学的 John、Kohavi 和 Pfleger 全面讨论了特征选择问题[15]。1997 年，新加坡国立大学的 Dash 和 Liu 对在此以前的特征选择方法进行了总结[8]，Dash 和 Liu 依据评判准则和搜索策略对特征选择算法进行了系统的分类。其中，特征选择的评判准则分为：距离测度（distance measures）、信息测度（information measures）、关联测度、一致性测度以及分类错误率测度；特征选择的搜索策略分为：完全搜索策略、启发式策略以及随机搜索策略。2002 年，Molina 等对各种不同的特征选择算法，做了全面的比较和评价[16]。近年来，集成学习成为机器学习领域中的研究热点之一，特征选择在集成学习中的运用主要涉及个体分类器的产生过程，特征选择一方面可以提高个体分类器的精度，另一方面可以增强个体间的差异性，从而提高集成学习的泛化能力[17]。Ho 提出了一种称为随机子空间法（random subspace）的集成学习算法[18]，利用随机产生的不同的特征子集，训练多个个体。Opitz 则提出了一种利用遗传算法搜索特征子集的集成学习算法，其中遗传算法的个体适应度函数综合考虑了个体分类器的准确率和个体间的差异性，Opitz 通过在 UCI 数据集上的实验证明了该算法优于 Bagging 和 Boosting 算法。

目前国际上相关机器学习的期刊和会议也有很多，见表 1.1，供读者参考。