智能方法及应用

ZHINENG FANGFA
JI YINGYONG

钟 珞 袁景凌 李 琳 钟 忺 著

科学出版社

智能方法及应用

钟　珞　袁景凌
李　琳　钟　忺　著

科学出版社
北　京

内 容 简 介

本书主要总结了目前比较常见的智能方法包括模糊计算、粗糙集与粒计算、群智能、神经网络、进化计算、人工免疫系统等，并从方法、模型和应用等三方面进行了阐述。重点讨论了智能挖掘分析方法、智能融合与优化方法以及智能方法在信息检索、推荐系统、观点挖掘、隧道监控、绿色计算等方面的典型应用。通过理论研究和具体实验分析，对各种常见智能方法及典型应用进行了剖析，并对未来智能技术进行了展望。

本书可供计算机相关专业的硕士、博士生及科研工作者参考，也适合对智能方法感兴趣的广大读者阅读。

图书在版编目(CIP)数据

智能方法及应用/钟珞等著. —北京：科学出版社，2015.3
ISBN 978-7-03-044600-9

Ⅰ.①智… Ⅱ.[illegible]钟 Ⅲ.①智能技术 Ⅳ.①TP18

中国版本图书馆 CIP 数据核字(2015)第 044962 号

责任编辑：张颖兵 杜 权/ 责任校对：肖 婷
责任印制：赵 博/ 封面设计：苏 波

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

双青印刷厂印刷
科学出版社发行 各地新华书店经销
*
开本：B5(720×1000)
2015 年 3 月第 一 版 印张：17 1/4
2015 年 3 月第一次印刷 字数：330

定价：80.00 元

(如有印装质量问题，我社负责调换)

序

随着智能方法在当今科学研究和工程实践中的不断应用发展，智能方法内容也越来越丰富。在各种应用中，经常会涉及一些经典算法或理论，如机器学习、数据挖掘、概率和统计等。它们在解决一些复杂的问题时大有用武之地，如专家系统、自然言语理解、模式识别、信息检索、智能推荐和博弈等。

智能方法及其应用领域的研究发展很快，涉及的学科领域越来越多，新的理论、方法与技术不断涌现，已形成多个研究方向。作为一门综合性学科，智能方法旨在研究如何利用计算机等现代信息技术模拟人类智能行为的算法，不仅是高等学校计算机科学与技术专业和软件工程专业等计算机类专业人员的研究热点，也得到其他信息类和管理类等学科领域专家学者的重点关注，并已经广泛地渗透到各学科领域和各行各业。

由于智能方法涉及面广，相关书籍的目标应定位为，帮助相关研究者掌握智能方法的基本理论、方法和技术，具有使用智能方法的基本能力。内容不应过于宽泛，而应突出重点，强调应用，深入浅出地介绍智能方法的理论与实现技术。本着这一指导思想，作者根据多年的研究和实践积累撰写了本书。本书共分 7 章，内容如下。

第 1 章为绪论，从智能计算方法和智能挖掘方法两个方面分别简要介绍相关经典和有影响力的算法。

第 2 章以智能计算方法为主，主要借鉴仿生学和拟物的思想，基于人们对生物体智能机理和某些自然规律的认识，采用数值计算的方法去模拟和实现人类的智能、生物智能、其他社会和自然规律。本章主要讨论模糊集计算、粗糙集计算、神经网络、进化计算、人工免疫计算和群智能计算。

第 3 章介绍与 Web 应用关系紧密的智能挖掘方法，是目前人工智能和数据库领域研究的热点话题。智能挖掘从大量数据中揭示出隐含的、先前未知的并有潜在价值的信息，是一种决策支持过程，本章介绍智能数据挖掘领域的主流算法，包括关联规则挖掘、序列模式挖掘、分类算法、聚类算法。最后讲解向量空间模型和语义网空间模型，它们是信息检索领域经典的数据表示模型。

第 4 章主要介绍近年来产生发展的多种智能优化算法。围绕灰色理论和量子计算，讨论它们与经典智能方法的融合算法，主要是灰色神经网络、灰色粗糙集、量子神经网络与量子进化计算。本章重点关注这些算法的产生和发展、算法的基本思想和理论、基本构成、计算步骤以及实际例子和应用。

第 5 章为智能方法典型应用，给出智能方法在数据清理、知识约简及预测等方面的典型应用。

第 6 章为智能方法 Web 应用，介绍 Web 信息检索、Web 推荐系统及 Web 观点挖掘的基本方法，并给出应用研究成果。

第 7 章为智能方法研究拓展，给出全书的总结并讨论目前智能方法的热点研究领域。

本书内容详实、层次清晰、详略适当、重点突出、语言严谨、应用典型，便于科学研究和教学。可作为高等学校计算机等信息类和管理类相关专业的高年级本科生和研究生教材，也可作为有关科技人员的专业参考书籍。本书第 1、7 章由钟珞撰写，第 2、4、5 章由袁景凌、钟忺撰写，第 3、6 章由李琳、钟忺撰写。钟珞负责全书内容的审定并组织统稿。此外，感谢吕品、谢晶、杨光、董玲玲、陈旻骋等对本书编写和整理等进行的辅助工作。本书综合了相关专家学者的最新成果和本团队的研究工作，但由于本领域技术内容丰富、发展迅猛，书中难免存在不足或疏漏之处，希望广大读者批评指正。

钟　珞

2014 年 10 月

目　　录

第1章 绪 论

本章介绍智能计算和挖掘方法各领域的经典算法。首先讨论以神经网络、进化、遗传、免疫、生态等为主的智能计算方法的特点，然后介绍以数据处理为核心的十大经典智能挖掘方法。

1.1 智能计算方法

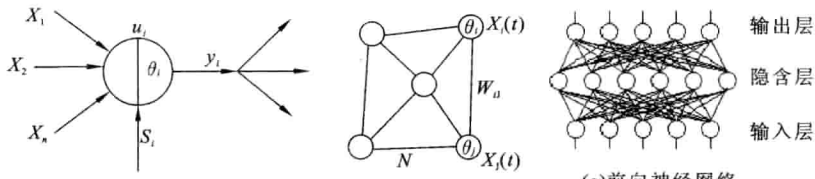
智能是个体有目的的行为、合理的思维，以及有效地适应环境的综合性能力。人工智能相对人的自然智能，用人工方法和技术，模仿、延伸和扩展人的智能(钟珞等，2009)。长期以来，人们从人脑思维的不同层次出发，对人工智能进行研究，形成符号主义、连接主义和行为主义。

传统的人工智能是符号主义，以 Newell 和 Simon 提出的物理符号系统假设为基础(吕品等，2012；史忠植，1998)。物理符号系统假设认为，物理符号系统是智能行为充分和必要条件。物理符号系统由一组符号实体组成，它们都是物理模式，可在符号结构的实体中作为组分出现。该系统可以进行建立、修改、复制、删除等操作，以生成其他符号结构。连接主义，或计算智能与分布式人工智能(Distributed Artificial Intelligence，DAI)是密不可分的。人们在研究人类智能行为中发现，大部分人类活动都涉及多个人构成的社会团体，大型复杂问题的求解需要多个专业人员或组织协作完成。“协作”是人类智能行为的主要表现形式之一，分布式人工智能正是为适应这种需要而兴起的。尤其是随着计算机网络、计算机通信和并发程序设计的发展，分布式人工智能逐渐成为人工智能领域的一个新的研究热点，作为人工智能的一个分支，DAI 主要研究在逻辑上或物理上分散的智能动作者如何协调其行为，即协调它们的知识、技能和规划，求解单目标或多目标问题，为设计和建立大型复杂的智能系统或计算机支持协同工作提供有效途径。分布式系统的本质决定了它是复杂的、非线性的、通过各子系统间的协同达到更高有序态的系统，因此，分布式人工智能的主要研究方法是连接主义的而不是符号主义的。

20 世纪 50 年代以后一段时间，符号智能体系取得了巨大的成功，但 80 年代中期以来，这种经典人工智能的发展由辉煌转入相对停滞状态，而计算智能在神经网络的带动下异军突起(石纯一等，1993)。与生命科学、系统科学密切联系是计算智能的突出特点，正是由于这个特点，不仅计算机科学家，而且众多其他学科的学者也加入计算智能的研究中来，极大地促进了它的发展。

1.1.1 神经网络

神经网络是连接主义的经典代表。神经网络是由大量神经元广泛互连而成的复杂网络系统，诞生于 1943 年(吕品等，2012)。单一神经元可以有许多输入、输出。神经元之间的相互作用通过连接的权值体现。神经元输出是其输入的函数。常用的函数类型有线性函数、S 型函数和阈值型函数。虽然单个神经元的结构和功能极其简单和有限，但大量神经元构成的网络系统的行为是极其丰富的。单个神经元、Hopfield 网络模型和前向神经网络的结构如图 1-1 所示。



(a)单个神经元　(b)Hopfield网络　(c)前向神经网络的结构与工作方式

图 1-1　神经网络

神经网络的基本特点如下。

(1) 大规模并行处理：神经网络能同时处理与决策有关的信宿，例如，虽然单个神经元的动作速度不快，但网络的总体处理速度极快。

(2) 容错性：由于神经网络包含的信息是分布存储的，即使网络某些单元和连接有缺陷，它仍然可以通过联想得到全部或大部分信息。

(3) 自适应和自组织性：神经网络系统可以通过学习不断适应环境，增加知识的容量。

1.1.2 遗传算法与演化计算

1. 遗传算法

遗传算法(Genetic Algorithm)是模拟达尔文的遗传选择和自然淘汰的生物进化过程的计算模型，是由 Holland 于 1975 年首先提出的(Holland，1975)。其主要特点是群体搜索策略和群体之间的信息交换。与解析法、穷举法、随机法等传统搜索方法相比，遗传算法具有不需搜索空间的知识、并行爬峰、编码方法适应性广等特点。遗传算法是所谓“演化计算”的一种。遗传算法的基本流程和要素如图 1-2 所示。

根据模式定理(Schemata Theorem)，遗传算法中串的运算实际上是模式的运算，遗传的进化实际上是模式的进化。低阶、短定义距及平均适应度高于群体适应度的模式(积木块)在子代中将以指数级增长，积木块在遗传算子作用下，相互结

合，能生成高阶、长距、高平均适应度的模式，可最终生成全局最优解。

遗传算法具有隐并行性，在对 n 个串个体进行运算时隐含处理了 $O(n^3)$ 个模式。实际上，在自然进化过程的任何时刻，总是同时有大量的物种在彼此独立地向前进化。在同一物种内部，也同时存在着大量的个体在通过自然选择、交配和基因突变而进化。显然，自然界的进化过程本身就是一个并行过程。遗传算法源于自然进化，自然也就继承了自然进化过程所固有的并行性。

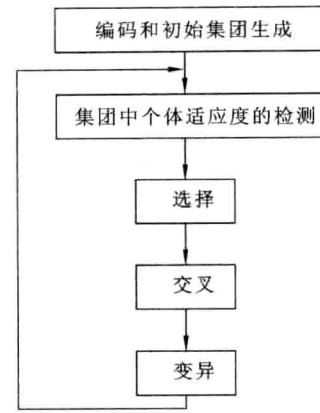


图 1-2 遗传算法

标准遗传算法是生物遗传过程的一个非常简化的模拟。事实上，由于遗传以及更广泛的进化与生态的关系是密不可分的，在遗传算法中引入生态因素是非常重要的。这方面经典的有小生境(Niche) 技术(史忠植，1998；Futuyma，1986)。

2. 演化计算

演化计算(Evolutionary Computation)，又称进化计算，是遗传算法的超集，其特点是群体搜索策略和群体中个体之间的信息交换。目前研究的进化算法主要有遗传算法、进化规划(Evolutionary Programming，EP)和进化策略(Evolutionary Strategies，ES)(Futumyma，1986)。尽管它们之间很相似，但历史上这三种算法是彼此独立发展起来的。

遗传算法由美国 Holland 创建，后由 DeJong 等进行了改进；进化规划最早由美国的 Gfogel、Owens 和 Walsh 提出；进化策略由德国的 Rechenberg 和 Schwefel 建立。三种算法既有许多相似之处，同时也有很大的不同：①进化规划和进化策略都把变异作为主要的搜索算子，而在标准遗传算法中，变异只处于次要地位；②交叉在标准遗传算法中起着重要作用，而在进化规划中完全省去，在进化策略中与自适应结合在一起使用非常重要；③标准遗传算法和进化规划都强调随机选择机制的重要性，而从进化策略的角度看，选择是完全确定的，没有合理的根据表明随机选择原则的重要性；④进化规划和进化策略确定地把某些个体排除在选择复制之外，而标准遗传算法一般对每个个体都指定一个非零选择概率(Robert，2006)。

此外，在所谓的“智能进化”中，除了考虑遗传因素外，还考虑到学习，这就是进化的强化学习(ERL)(Futumyma，1986)。在 ERL 中，评价网络的结构和结合强度均由遗传决定，而行动网络的结合强度则有可能通过学习决定。而且这种学习信号的源泉，也仅限于评价网络的报酬信号。这是通过遗传决定的评价网络的报酬信号对行动网络结合强度实行最优化的一种强化学习方法。

1.1.3　免疫信息处理

人体免疫系统是一个高度进化、复杂的生理机制，免疫系统通过高度复杂的网络结构来识别和排除抗原性异物，维护体内环境的稳定。免疫过程中所具有的识别能力、学习和免疫记忆功能，以及自适应调节机制等特性具有重要的工程应用价值。1994 年以来，免疫信息处理成为国际上新的研究热点。从工程应用的角度，可抽取出免疫反应的概念化模型，如图 1-3 所示。图中免疫识别能够识别自我(Self)和非我(Non-Self)，对模式识别和计算机病毒防治等都具有重要的的意义。

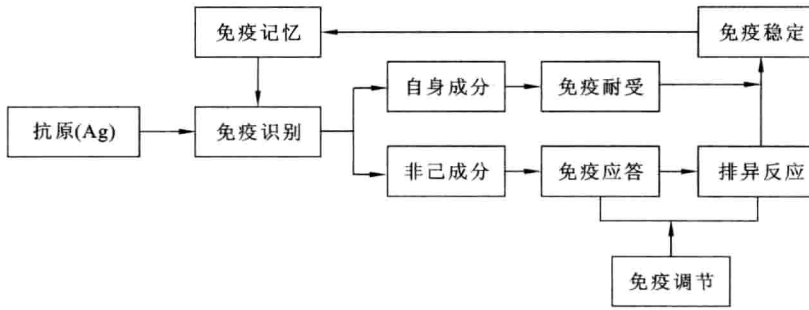


图 1-3　免疫反应模型

在免疫调节中，一个比较成功的学说是 Jerne 网络模型。这种高度联结的网络具有非线性动力系统所具有的稳定平衡点(该点对应于免疫记忆)，当抗原侵入时，扰动网络的平衡，通过内部节点间的作用关系而形成新的平衡点。该网络学说能较好地解释免疫记忆、免疫学习及免疫耐受等重要特点。基于该网络学说的思想和模型已在组合优化问题、自适应控制等方面取得了较好的效果。

免疫系统与脑神经系统在系统行为上具有很多相似性，如识别、学习和记忆性能等，但是它们却有着不同的信息处理机制。免疫细胞间呈现着一种相互刺激和抑制的对称作用关系，这不同于脑时间元之间的作用关系。免疫系统广泛分布于全身，它们通过在时间和空间上分布式的网络结构来实现各种免疫功能，并且这种网络结构和作用关系是随着环境的不同而不断变化的。而经典人工神经网络是一种固定连接的网络模型。

免疫与遗传系统之间也是相互区别和联系的，由于抗原、抗体的特性是通过基因编码体现的，体内多样性抗体的产生也是基于免疫细胞分裂时进行的基因交叉和变异而实现的，这种基于遗传交叉的多样性操作，以及基于变异和选择等自适应群体层次的操作，对构成免疫识别和记忆具有重要的作用。但免疫系统与遗传系统有着本质的区别：遗传算法是一种单一功能个体的进化，对于每个个体，只能适

应某个问题或环境，一旦环境变化，进化将前功尽弃；而免疫系统是将环境（非己）和自己相互作用直接考虑的。和遗传算法相比，免疫系统还有如下特征。

（1）不是独立地对每个个体进行评价和选择，而是以共同作用为前提，考虑共生关系与系统化。

（2）自我与非我的识别是一种特殊的模式识别，因此有特殊的多样性和选择操作。

（3）自适应性体现在包括结构层次在内的各种层次中。

（4）更能在线适应变动的环境。

（5）免疫与生态系统具有某些联系。在博弈生态系统中，博弈可产生抗体[6]。

1.1.4 生态计算

生态计算或计算生态学（the Ecology Computation），是神经网络、遗传算法、演化计算、免疫信息处理等在更高层次的概括——尽管其发展与上述各分支是相互独立的（曹先彬等，1999；Robert，2006）。

从历史的角度，生态计算具有几个相对独立的来源。而这些相对独立的来源却得到了类似的结论，这无疑从一个侧面反映了生态计算作为计算智能的集大成者，其产生是在智能计算发展到一定阶段后的必然结果。其发展可描述如下：

（1）开放信息系统→计算生态学；

（2）博弈论→博弈生态系统→生态动力学；

（3）反馈与控制理论→生态学模型；

（4）非平衡态统计物理学→非平衡相变→反应扩散方程和序参量方程→进化方程。

生态系统中的自组织可从几个方面考察。从博弈论的角度考察，生态系统中的各种策略（如K策略、R策略），根据其适应度、淘汰、变异等，缓慢地进化。那些能够适应环境的策略类型或者相互依存的策略集团自发地形成具有新的秩序的组织。从协同的角度，计算的过程可以看成一种相变，是系统处在一种非线性结构下产生新的更有序的空间结构和时间结构的过程。

生态计算的关键不是设计一种新的方法直接解决实际问题，它更多是面向智能计算的一般原理，说明智能计算中的某些基本原则的。在这个意义上，也可以把它称为“广义生态学”。下面还要详细地讨论这个问题。

1.1.5 各领域的内在联系

智能计算的各领域之间不是相互独立的，而是有着深刻的内在联系。连接主

义的思维方式与传统的符号主义不同，智能计算的各领域用不同方式实现了连接主义的计算，即研究简单个体如何在简单交互规则指导下，构成具有复杂智能行为的高层系统。由此带来各种算法的统一特点，如社会性、并行性、单元的智能性、开放性等。

各领域服从统一的模型，即“开放式计算系统”模型。智能计算是多个简单个体通过生态行为，或者说是社会行为，自组织地形成智能的过程。系统是由异构的、分布的、动态的、大规模的、自主的成分构成的计算系统，一个复杂的计算任务由大量的计算单元非同时的计算行为完成；执行这些任务的单元的全部特性对其他单元甚至系统本身也是未知的；大量的单元的行为决定是基于它们对系统的不完全知识和延迟的甚至是矛盾的信息做出的。

遗传算法是进化计算的特例，而进化实际上是一种特殊的生态行为（曹先彬等，1999）。进化包含遗传的因素，但进化不能只从遗传的角度出发来考察问题，同样演化计算也不能仅以遗传变异机制为限（如经典演化计算所做的）。进化是非生物环境和生态系统共同作用的结果，而且环境通常通过生态系统对物种的进化起作用；物种在进化中相互作用，形成所谓的“协同进化”（Coevolution）。因此，如果把以遗传算法为代表的演化计算作为“微观演化”或“基因进化计算”，那么，演化计算的进一步研究将导致“生态进化计算”的产生。

生态模型具有共性。对生态的理解不能停留在自然生态系统的层次上，在智能计算中，“生态”是描述计算系统中各主体间相互关系的概念，广义上，神经网络、进化系统、多主体系统、免疫系统等都是生态系统。由于系统的层次化，生态关系也有不同的层次，如人作为细胞的生态系统，而本身又是社会经济“生态系统”的一个“细胞”。一个统一的生态理论能解释这不同应用领域、不同层次的系统的共性问题，并对智能计算的终极目标——人工生命理论产生直接的积极影响。在这个意义上，不能不说生态的研究在智能计算中是处于一个承上启下的地位的。

综上所述，智能计算作为一个整体，具有明确的研究思路、理论背景、数学手段和应用前景，其各领域和智能计算一般理论均具有极大的理论意义和应用能力。

1.2 智能挖掘方法

众所周知，互联网已经成为一个非常流行并且功能强大的信息存储、传播、获取的平台以及知识发现的数据库。然而，由于信息量和用户量的大幅快速增长，网络用户经常遇到信息超载的问题。另外，互联网上大量有价值的数据或者信息知识可以通过先进的数据挖掘方法来发现。据认为，挖掘这方面的知识将大大有利于网站的设计和 Web 应用程序的开发，并促进其他相关的应用，如商业智能、电子商务、娱乐广播等。

智能挖掘方法已经吸引了大量从事数据管理、信息检索，尤其是知识发现和机器学习领域中的人工智能等方面的学者和研究人员，并且，由于互联网上的数据内容巨大增长和电子商务应用的迫切需要，近年来许多研究机构都有涉及这个话题[7,8]。国际权威的学术组织 the IEEE International Conference on Data Mining (ICDM)2006 年 12 月评选出了数据挖掘领域的十大经典算法：C4.5、k-Means、SVM、Apriori、EM、PageRank、AdaBoost、kNN、Naive Bayes 和 CART。不仅是选中的十大算法，其实参加评选的 18 种算法都可以称得上是经典算法，它们在数据挖掘领域都产生了极为深远的影响。下面对十大经典算法的核心思想进行一个简要的介绍。

1.2.1 决策树类模型

1. C4.5 决策树

机器学习中，决策树是一个预测模型。它代表的是对象属性值与对象值之间的一种映射关系。如图 1-4 所示，树中每个节点表示某个对象，每个分叉路径则代表某个可能的属性值，而每个叶节点则对应具有上述属性值的子对象。决策树仅有单一输出；若需要多个输出，可以建立独立的决策树以处理不同输出。决策树一般都是自上而下生成的。选择分割的方法有多种，但是目的都是一致的，即对目标类尝试进行最佳的分割。从根节点到叶子节点都有一条路径，这条路径就是一条“规则”。决策树可以是二叉的，也可以是多叉的。

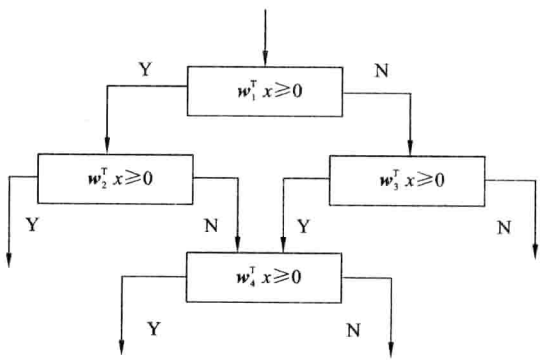


图 1-4 二叉决策树

2. ID3 算法

ID3 算法是 Quinlan 在 1975 提出的分类预测算法，当时还没有提出数据挖掘这个概念。该算法的核心是“信息熵”。信息熵就是一组数据包含的信息及概率的