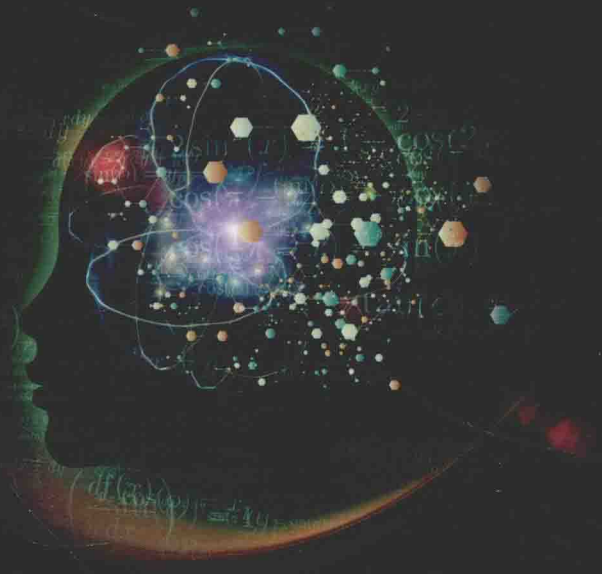

基于免疫计算的机器学习方法及应用

Machine Learning Methods and Applications Based on Immune Computing

徐雪松◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

湖南商学院学术著作出版基金资助出版

基于免疫计算的机器学习方法及应用

徐雪松◎著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书将免疫智能计算方法引入机器学习领域，致力于研究基于生物免疫原理的机器学习软计算方法。以免疫计算智能基本原理为线索，对其研究状况进行系统性的论述，从理论、算法构建及工程应用等方面对免疫机器学习进行介绍和分析。针对关联规则挖掘、数据分类、数据聚类、属性约简等机器学习及生物信息大数据挖掘等具体问题，提出一系列新方法，并展开理论及应用探讨。

本书可以为计算机科学、信息科学、人工智能和自动化技术等领域从事机器学习、数据挖掘及智能信息处理等相关专业技术人员提供参考，也适合信息管理、情报学、管理科学与工程、电子商务、计算机应用等专业的师生教学使用，还可供广大信息与知识工作者、有关管理和科技工作者学习参考。

图书在版编目（CIP）数据

基于免疫计算的机器学习方法及应用 / 徐雪松著. —北京：电子工业出版社，2017.8

ISBN 978-7-121-32363-8

Ⅰ. ①基… Ⅱ. ①徐… Ⅲ. ①机器学习－研究 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字（2017）第 182913 号

策划编辑：秦绪军　朱雨萌
责任编辑：赵　平
特约编辑：田学清　赵海军等
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　　印张：14.75　　字数：236 千字
版　　次：2017 年 8 月第 1 版
印　　次：2017 年 8 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件到 dbqq@phei.com.cn。

本书咨询联系方式：88254750。

前　言

近些年，随着信息技术的飞速发展，以博客、社交网络、基于位置（LBS）服务为代表的新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，在商务贸易和政府事务电子化、大规模工业生产过程中的智能监控和诊断、医疗领域的计算机诊断管理及科学计算等应用领域，产生了不断增长的海量数据源。数据正以前所未有的速度增长和累积，人类收集数据、存储数据的能力得到了极大提高，如何实现数据的智能化处理，从而充分利用数据中蕴含的知识与价值，已成为当前学术界与产业界的共识。在这样的大趋势下，人工智能、机器学习作为一种主流的智能数据处理技术，其作用日渐重要并受到了广泛关注。

机器学习是人工智能的核心研究领域之一。人工智能的根本在于智能——如何为机器赋予智能，而机器学习则是部署支持人工智能的计算方法。人工智能是科学，机器学习是让机器变得更加智能的算法。也就是说，机器学习成就了人工智能。基于人工智能所发展的仿生计算智能又为机器学习实践提供了强有力的工具。一般而言，经验对应于历史数据（如互联网数据、科学实验数据等），系统对应于数据模型（如决策树、支持向量机等），而性能则是模型对新数据的处理能力（如分类和预测性能等）。因此，机器学习的根本任务是信息和数据的智能分析与建模。

智能信息处理就是模拟人或自然界其他生物处理信息的行为，建立处理复杂系统信息的理论、算法和系统的方法和技术。其中，基于生物免疫机制发展而来的免疫计算智能信息处理技术是一门新兴的交叉学科。它与人工智能、人工生命科学、自动控制、运筹学、计算机科学、信息论、应用数学、仿生学、脑科学等有着密切的关系，是相关学科相互结合与渗透的产物。其主要面对的是不确定性系统和不确定性现象的信息处理问题，在机器学习、模式识别、复杂系统建模、分析和决策、系统控制、系统优化等领域具有广阔的应用前景。生物免疫系统是生命系统的主系统之一，免疫系统通过从不同种类的抗体结构中构造自己-非己非线性自适应网络，在处理动态变化环境中起着重要作用；同时它又具有高度自适应、分布、自组织等特性，蕴含着丰富的信息处理机理。免疫计算智能正是借鉴生物免疫系统信息处理机制而发展起来的智能信息处理技术。它具有噪声忍耐、无监督学习、模式识别、清晰的知识表达和学习记忆等进化学习机理，同时它吸取了传统进化计算、分类器、神经网络等的优点，从而提供了一种解决复杂机器学习问题的新选择。从工程上讲，它具有结合先验知识和免疫系统的适应能力；从信息科学讲，它具有强壮的鲁棒性和预处理能力。应当指出的是，基于免疫计算的机器学习和信息处理机制具有的多样性及其遗传机理，不仅可以用于全局进化的探索，改善已有进化算法中对局部探索不太有效的情况，而且在避免早熟及处理多准则和约束问题方面显示出良好的潜力。因而可能弥补神经网络等“黑箱”式学习模型难以表达学习知识的缺陷，有助于人们对问题的论证，同时将免疫信息处理与其他计算智能方法的集成可用于解决其他智能系统等难以解决的复杂问题。

因此，为读者提供人工智能领域的基于免疫计算的机器学习相关算法、技术和问题解决过程中的实践经验，是本书撰写的宗旨。本书以各类免疫机器学习方法和算法为核心，在概括了人工智能与机器学习、机器学习与免疫计算等概念的基础上，对现代机器学习技术和发展进行了简要介绍。重点介绍了免疫计算的生物学机制，以及各类免疫机器学习方法在数据分类、数据聚类、关联挖掘、数据降维、规则约简及生物大数据中的具体应用。

全书分为七章，内容包括：第 1 章绪论部分的人工智能、机器学习及免疫计算概念；第 2 章主流机器学习技术与方法；第 3 章免疫计算的基础原理；第 4 章免疫关联规则挖掘方法；第 5 章小生境免疫粗糙集属性约简方法；第 6 章免疫阴性选择数据分类器；第 7 章免疫网络在生物大数据中的应用。最后，还探讨了大数据背景下机器学习技术的发展方向，以及进一步研究的方向和面临的问题。

本书得到了国家留学基金项目、国家社科基金项目（14BJY066）、教育部人文社科青年项目（12YJCZH233）、湖南省自然科学基金项目（2016JJ2069）、国防科学技术大学博士后基金，以及广西跨境电商智能信息处理重点实验室培育基地等多方面的资助。同时，作者在科研和本书的撰写过程中得到了美国布兰迪斯大学 Professor Hong、美国麻省理工大学 Professor Yue 的支持和帮助，在此谨致以最诚挚的感谢。同时感谢国防科学技术大学张维明教授、广西财经学院王四春教授的指导和帮助。书中给出了主要算法实现机制和相应标准测试问题，便于读者使用和研究。另外，本书还参考和引用了一些论文和资料，在此也一并表示衷心的感谢。

感谢作者家人的大力支持和理解，将此书献给小女 Penny，在美国访学一年中，是你陪伴着我完成了本书。

最后感谢电子工业出版社的朱雨萌老师在本书出版过程中给予的大力帮助。

由于免疫计算及机器学习技术是一门新兴交叉学科，很多理论方法与应用技术问题还有待进一步深入探索和发展，加上作者学识所限，写作时间又十分仓促，因而书中难免存在不足之处，敬请专家和读者们批评指正。

作　者

2017 年 3 月

于美国 波士顿

目 录

第1章

绪 论

本章导读：

随着信息技术及互联网应用的不断发展，人们在社会生活、科学研究等各个领域中的数据正以前所未有的速度产生并被广泛收集、存储。如何实现数据的智能化处理从而充分利用数据中蕴含的知识与价值，已成为当前学术界与产业界的共识。正是在这样的大趋势下，机器学习作为一种主流的智能数据处理技术，其作用日渐重要并受到了广泛关注。本章通过简单介绍人工智能和机器学习的概念、发展，分别阐述了人工智能、数据挖掘、仿生计算智能与机器学习的关系，并重点介绍了基于仿生计算原理的免疫计算在机器学习领域的基本概念、特性和发展。最后给出了本书的基本结构和各章节的主要内容，方便读者阅读。

1.1 引言

上天赐予了人类惊人的学习能力，从出生开始就不断学习和接收外界的反馈，掌握各种复杂的知识和技能，从而完成各项复杂的工作和任务，如自由行走、语言交流和图像识别等。人类不断地将这种第一学习体验加以修正、完善、发展和成熟，逐步形成人类的经验和智能。之后，人类利用这种学习概念来积累、拓展知识，开始对未知世界进行思考并预测结果。随着计算机技术和信息技术的快速发展，人类将这种学习概念和能力应用于与计算机相关的程序和任务中，不断赋予机器学习和具备智慧的能力。这些涉及上述计算过程中的技术，就是发展了 70 多年且目前正火热的“人工智能”。人工智能最初可以追溯至 1956 年，当时多名计算机科学家在美国达特茅斯举办的会议上共同提出了“人工智能”的概念。在随后的几十年中，人工智能一方面被认为是人类文明的发展方向，另一方面也被认为是难以企及的梦想。虽然计算机技术已经取得了长足的进步，但是到目前为止，还没有一台计算机能产生“自我”的意识。在人类的指导和大量数据的帮助下，计算机可以利用“机器学习”的技术表现得十分强大，但是离开了这两者，它就缺失了基本的辨识能力，直至 20 世纪 90 年代末，人工智能世界一个决定性时刻的到来。1997 年，国际象棋大师加里·卡斯帕罗夫对战 IBM 公司的“深蓝”计算机，“深蓝”计算机最终战胜国际象棋大师。本次胜利令外界对人工智能的看法发生彻底转变，并对其中重要的机器学习能力表现出极大的热情。在棋局对弈的过程中，象棋大师必须不断进行非常复杂的思考，考虑多种不同的走法及相应的策略。他们也可以自己进行学习，并创出新奇的走法。计算机利用机器学习技术，同样能够模仿这个过程，甚至将其应用到象棋这样的特别任务里，展露出人工智能巨大的潜力。得益于上述成功，人工智能不断发展，在过去几年，尤其是自 2015 年以后，人工智能实现了爆炸式发展。这在很大程度上是由于计算机的 CPU 和 GPU 的发展，使并行计算变得速度更快、成本更低、性

能更强大。与此同时，存储设备的容量变得越来越大。大数据的发展，使我们可以获得并充分学习和利用这些海量数据。无论是图片、文字、音频、视频，还是地图数据、实时交易信息等，都可用来实现机器学习的目的。2016 年，谷歌旗下的 DeepMind 公司使用深度学习算法，训练 AlphaGo 如何应对专业级棋手的走法，开始挑战非常复杂的围棋游戏。在对战其他围棋程序时的胜率达到 99.8%，并且在对战围棋专业选手李世石的比赛中取得 5 局 4 胜的好成绩。一时间，人工智能、机器学习等概念成为业界炙热的话题。这些国际一流企业所进行的应用实践充分证明了计算机可以像人类一样学习如何进行信息获取、数据处理、自主学习、建立模型和预测结果。随后，机器学习和人工智能技术将被应用于解决更为现实的问题。由著名的斯坦福大学的机器学习教授 Andrew Ng 和在大规模计算机系统方面的世界顶尖专家 Jeff Dean 共同主导的 Google Brain 项目，采用深度神经网络机器学习模型，在语音识别和图像识别等领域获得了巨大的成功。项目负责人之一 Andrew 称：“我们没有像通常做的那样自己框定边界，而是直接把海量数据投放到算法中，让数据自己说话，系统会自动从数据中学习”。另外一名负责人 Jeff 则说：“我们在训练的时候从来不会告诉机器：这是一只猫。其实是系统自己发明或者领悟了‘猫’的概念”。从看似很神奇却又真实的工程应用中我们可以了解到，机器学习是一门专门研究计算机怎样模拟或实现人类的学习行为，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的学科。在这些学习过程中，充分借鉴和应用了人工智能领域的多种理论、技术和方法。

1.2 人工智能与机器学习

人工智能（Artificial Intelligence，AI）的思想萌芽可以追溯到 17 世纪的巴斯卡和莱布尼茨，他们较早萌生了“有智能的机器”的想法。19 世纪，英国数学家布尔和德·摩尔根提出了“思维定律”，这些可谓是人工智能的开端。

19 世纪 20 年代，英国科学家巴贝奇设计了第一架“计算机器”，它被认为是计算机硬件，也是人工智能硬件的前身。电子计算机的问世，使人工智能的研究真正成为可能。自图灵提出“弱人工智能”以后，更多的研究人员期望在此基础上机器能有自己的思维过程，从而形成“强人工智能”的想法。为了实现“强人工智能”，需要同时开展对脑神经科技、脑感知技术、智能机理和智能构造技术的研究，从而揭示人类智能的根本机理。在此基础上用智能机器去模拟、延伸和扩展人类智能，实现脑力劳动的自动化，而这正是人工智能研究的根本目标。

人工智能在 1956 年被正式提出前，其研究工作主要集中在探索智能及智能模拟的普适理论。这个一般术语用来描述一种由人类创造的技术，这种技术在解决问题时能够达到类似人类的智商程度。尼尔逊教授对人工智能下了这样一个定义：“人工智能是关于知识的学科——怎样表示知识以及怎样获得知识并使用知识的科学。”美国麻省理工学院的温斯顿教授认为：“人工智能就是研究如何使计算机去做过去只有人才能做的智能工作。”这些说法反映了人工智能学科的基本思想和基本内容，即人工智能是研究人类智能活动的规律，构造具有一定智能的人工系统。研究如何让计算机去完成以往需要人的智力才能胜任的工作，也就是研究如何应用计算机的软硬件来模拟人类某些智能行为的基本理论、方法和技术。人工智能为 21 世纪科技领域最前沿的技术之一，它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。其所使用的技术旨在根据数据和分析赋予计算机能够做出类似人类的判断。许多研究者深感单从符合主义、连接主义及行为主义来进行其研究有其局限性，甚至有些指导思想已被证明是错误的。人工智能应该从生物学而不单是物理学受到启示。在其他学科，尤其是生物技术的促进下，人工智能的研究随后进入了智能模拟的个性设计阶段。其主要特征是其研究不仅在方法上，而且在思想上呈现出多样性，发展了大量实用的方法，这一阶段是人工智能最具特色的发展阶段。人工智能研究是企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和

专家系统等。总的来说，人工智能研究的一个主要目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作。一般来说，人工智能分为计算智能、感知智能和认知智能三个阶段。

第一阶段为计算智能，即快速计算和记忆存储能力。十多年前，IBM 深蓝计算机战胜了国际象棋大师卡斯帕罗夫，当时震惊了世界。象棋机器人能够战胜人类，靠的就是超强的记忆能力的运算速度，能够预测到十几步以后的结果，这就属于计算智能。

第二阶段为感知智能，即视觉、听觉、触觉等感知能力。人和动物就是通过各种智能感知能力与自然界进行交互的。感知智能方面最形象的一个研究项目就是自动驾驶汽车，谷歌和百度都意欲在这个方面实现突破。机器不需要了解各种知识，只需要用各种传感器对周围的环境进行处理、自动控制就可以实现自动驾驶。

第三阶段为认知智能，也是目前各大科技巨头都在迫切寻找突破的领域，通俗来说就是“能理解、会思考”。人类有语言，才有概念，才有推理，所以概念、意识、观念等都是人类认知智能的表现，这也使人类能够明显区别于动物。人工智能将涉及心理学、哲学和语言学等学科，可以说几乎包括了自然科学和社会科学的所有学科。从思维观点看，人工智能不仅限于逻辑思维，更要考虑形象思维、灵感思维才能促进人工智能突破性的发展。认知智能是目前机器与人差距最大的领域：让机器学会推理和决策异常艰难，目前认知智能主要有传统和仿生两个主要流派。传统派认为，希望依靠知识工程或者通过知识图谱给大量信息加标签的方式进行量的堆积，用量变促进质变，实现真正的语义理解和认知智能，IBM 的沃森是这个流派的代表。仿生派希望参照人类大脑这个唯一的真正的智能，先研究人类大脑本身的运作机理，了解人脑神经元的结构，再通过人工神经网络进行规模、结构和机理上的模拟，通过仿生学思路实现人工智能的突破。

机器学习（Machine Learning，ML）是一门专门研究计算机如何模拟或实现人类的学习行为以获取新的知识或技能，重新组织已有的知识结构使之不

断完善自身性能的学科。机器学习是人工智能的核心研究领域之一，其最初的研究动机是让计算机或系统具有人的自主学习能力以便实现智能。机器学习中一个最宝贵的方面是适应能力，通过适应性学习能提高预测的准确度。英特尔机器学习主管尼迪·查普尔在解释人工智能与机器学习关系时提到："人工智能的根本在于智能——如何为机器赋予智能。而机器学习则是部署支持人工智能的计算方法。人工智能是科学，机器学习是让机器变得更加智能的算法。也就是说，机器学习成就了人工智能"。机器学习是人工智能增长最快的部分，尤其是目前的深度学习在现实工业应用中的成功，带来了机器学习技术的蓬勃发展，拓展了人工智能的适用领域。深度学习是机器学习的分支，是机器学习的一种实现方式。机器学习系统将任务分解，让机器利用深度学习技术可以去完成这些任务。像无人驾驶汽车、精准医疗及更强大的预防医疗，甚至更好的电影推荐都将成为可能。借助于机器学习，人工智能将走过科幻小说阶段，C-3PO 机器人和终结者将会成为现实。目前基于深度学习的机器学习系统，在学习过程中，最关键的在于表面区域或深度。更复杂的问题可以由更多神经元和层块来解决。这个系统用于对系统进行训练，将已知的问题和答案应用于解决任何给定的问题，这就创造了一个反馈回路。训练结果是一个加权结果，这种加权会传递给下一个神经元来决定该神经元的输出——通过这种方式，它根据各种可能性建立起一个更为准确的结果。深度学习已经应用到更复杂的任务当中，在这些任务里规则更为不明确也更加复杂。大数据时代将提供一些更有利于推动使用机器学习的工具。我们可以看到机器学习应用于任何与模式识别相关的东西中，例如面部识别系统、语音助手和用于防止诈骗行为的分析。由于有这些更为复杂和更尖端的算法的帮助，尤其是大数据和分布式计算的有效支持，人工智能正进入一个新时代。

1.3 数据挖掘与机器学习

上一节提到近些年来，随着大数据及信息技术的飞速发展，博客、社交网络、基于位置服务为代表的新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，在商务贸易和政府事务电子化、大规模工业生产过程中的智能监控和诊断、医疗领域的计算机诊断管理和科学计算等应用领域，产生了不断增长的海量数据源。数据正以前所未有的速度不断地增长和累积，人类收集数据、存储数据的能力得到了极大提高，对这些数据进行分析以发掘数据中蕴含的有用信息，成为几乎所有领域的共同需求。正是在这样的大趋势下，机器学习和数据挖掘技术的作用日渐重要，并受到了广泛的关注。

机器学习是人工智能的核心研究领域之一，而数据挖掘又是机器学习的核心技术之一。人工智能与机器学习等概念的相互关系如图 1-1 所示。目前被广泛采用的机器学习的定义是“利用经验来改善计算机系统自身的性能”。事实上，由于“经验”在计算机系统中主要是以数据的形式存在的，因此机器学习需要设法对数据进行分析，这就使得数据挖掘逐渐成为智能数据分析技术的创新源之一，并且为此而受到越来越多的关注。我们常说机器学习就是研究计算机怎样模拟人类的学习行为，以获取新的知识或技能，并重新组织已有的知识结构使之不断改善自身。人类是通过从以往发生事情的经验中进行学习的。而对于计算机来说，它们学习的经验其实就是数据和信息。我们通过不断给计算机“喂吃”（输入）数据，计算机通过算法“消化”（训练）数据，并不断“成长”（输出）为一个模型，然后将这个模型运用到新的数据上作出新的预测或决策。机器学习与人类思考模式的对比如图 1-2 所示。