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。

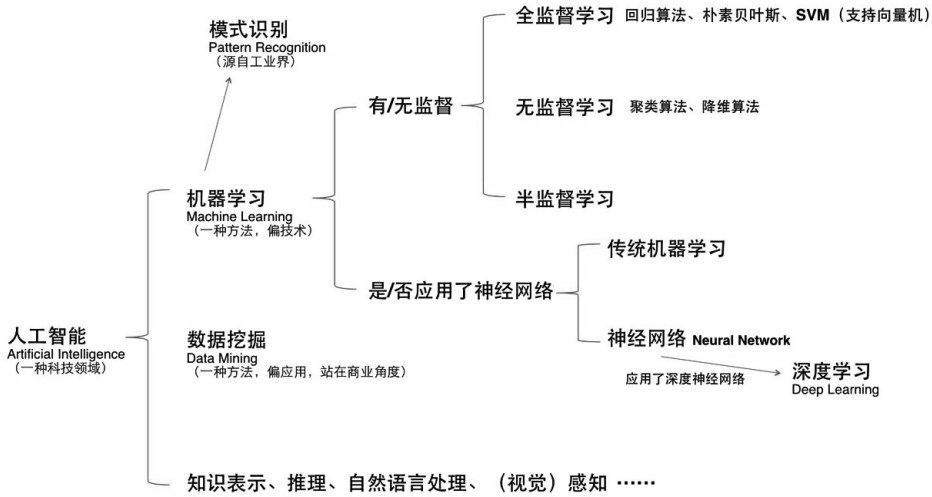

图 1-1　人工智能与机器学习等概念的相互关系

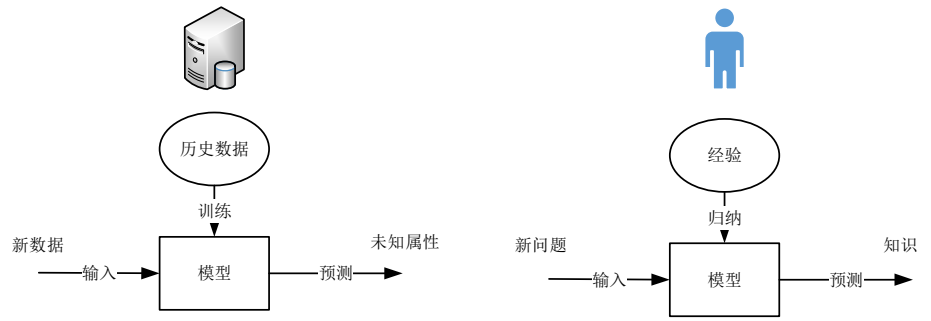

图 1-2　机器学习与人类思考模式的对比

几十年来，很多计算机科学和应用数学界的学者们总结出了不少教会计算机学习的办法，它们就是各式各样的机器学习算法。这些算法是数据科学家们手中的挖掘机，担负着将海量数据化腐朽为神奇的使命。因此，从数据科学的角度，我们可以对机器学习下这样一个定义：机器学习是指用某些算法指导计算机利用已知数据得出适当的模型，并利用此模型对新的情境给出判断的过程。由此看来，机器学习的思想并不复杂，它仅仅是对人类生活中学习过程的一个模拟。而在这整个过程中，最关键的是数据。因此，“机器学习”和“数据挖掘”通常被相提并论，并在许多场合被认为是可以相互替代的

术语。顾名思义，数据挖掘就是试图从海量数据中找出有用的知识。大体上看，数据挖掘可以视为机器学习和数据库的交叉，它主要利用机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据。

尤其是近几年互联网数据大爆炸，数据的丰富度和覆盖面远远超出人工可以观察和总结的范畴，而数据挖掘的算法能指引计算机在海量数据中挖掘出有用的价值，这也是无数学习者为之着迷的原因。一般意义上，大数据是指无法在一定时间内用常规机器和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合，呈现出大量化、多样化、快速化、低价化的特点。*Nature* 和 *Science* 等刊物也相继出版专刊来探讨大数据的研究。比如，2008 年 *Nature* 出版的专刊 *Big Data*，从互联网技术、网络经济学、超级计算、环境科学和生物医药等多个方面介绍了海量数据带来的挑战；2011 年，*Science* 推出关于数据处理的专刊 *Dealing with Data*，主要围绕科学研究中大数据的问题展开讨论，说明大数据对科学研究的重要性。计算社区联盟（Computing Community Consortium）在 2008 年发表了报告“*Big datacomputing: Creating revolutionary break throughs in commerce,science, and society*”，阐述了在数据驱动的研究背景下，解决大数据问题所需的技术，以及面临的一些挑战。美国一些知名的数据管理领域的专家学者则从专业的研究角度出发，联合发布 *Challenges and Opportunities with Big Data*，对大数据的影响、关键技术和应用领域等进行了详尽分析。进入 2012 年以来，大数据的关注度与日俱增。在 2012 年 1 月的达沃斯世界经济论坛上，特别针对大数据发布了报告“*Big data, bigimpact: New possibilities for international development*”。大数据本身是一个比较抽象的概念，单从字面来看，它表示数据规模的庞大。但是仅仅数量上的庞大显然无法看出大数据这一概念和以往的“海量数据”（Massive Data）、“超大规模数据”（Very Largedata）等概念之间有何区别，从这个意义上来说，我们需要对大数据时代的数据存在的若干形态进行新的诠释。

数据的形态之一：大数据。经典意义上的数据挖掘，通常是指对中等规模或海量数据进行分析。怎样才算海量数据，目前还没有明确的标准。而近

几年产生的大数据的提法，其概念在内涵和外延上都有了扩展。从本质上，可以认为大数据和海量数据是相似的。在实践中，不是记录数多的就称为大数据，通常大数据是数据量和数据维度均很大，且数据形式很广泛，如数字、文本、图像、声音等。同时大数据往往可能蕴含着丰富的规律和知识，所以在大数据之上应用数据挖掘就成为理所当然的活动了。

数据的形态之二：小数据。相对于大数据，在实践中还会存在不少特殊情况。例如，医学上有些疾病极为少见，只出现几百例甚至几十例就几乎是该病的总体了。这种数据可以称之为小数据。实际工程应用中需要对这些小数据进行深入分析和探索，以便挖掘出罕见疾病的特征，并给出临床诊断依据。面对这种数据类型，如果按照记录数依照传统数据挖掘观念、方法和技术，无法开展探索性的分析工作。数据挖掘的一个发展分支应该是从规模较小的、有限的数据中探索其中的规律和知识。

数据的形态之三：宽数据。有一种情况是小数据高维度、小样本大信息，可以称之为宽数据。如某些基因组信息数据量很少，通常只有几十例到几百例，但维度很高，通常有几百个到几千个。更极端的情况是个人大信息，即单个记录下的高维信息，如从宽带、移动支付、物联网、手机等媒介收集个人信息。在不远的将来会出现单独个体的高维数据，并需要解决此类数据挖掘的新理论和新算法。

数据的形态之四：深数据。还有一种数据涉及维度不是很宽，但是数据在某几个维度上跨度非常大，历史数据非常多，或者数据量的增长速度非常快，称之为深数据。如医学检查中心脑电图监测，每小时会产生几十万至几百万条数据。再如互联网服务商的服务器对互联网访问事件的日志记录，也是每小时会产生几十万至几百万条数据。这类数据，我们有时也称之为流数据。对这些深数据的挖掘也是非常具有挑战性的。一方面由于它的数据量非常大，另一方面也由于对这类数据进行挖掘的实时性要求较高。这些随着数据收集手段的进步而形成的各有特色的数据，正在逐步进入数据挖掘研究的视野。所以说数据挖掘应包括大数据挖掘、小数据挖掘、宽数据挖掘和深数

据挖掘。我们需要做的是处理好各类数据来获取知识，研究解决各类型数据的挖掘的新理论和新算法。这些数据的分析算法不完全与经典大数据挖掘相同，例如医学上的个性化精确治疗，就离不开涉及个人的宽数据和深数据。

数据挖掘是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。数据挖掘应该是一个广义的概念，甚至可以说不是一个传统意义上的定义，而是一类活动的集合。凡是有目的地探索数据中隐含的规律和知识的活动都可称为数据挖掘。在这里，我们重点强调的要素是：有目的、探索性地获取、数据中隐含的规律和知识。这就意味着不限任何方法和手段，无论是数学的还是非数学的，无论是复杂的还是简单的，只要能揭示数据中隐含的规律和知识都可以被称为数据挖掘。现在机器学习的范围已经逐步扩展到不确定性推理技术、人工智能、高性能计算，以及基于神经网络、模糊集理论、粗集理论、进化计算的软计算等研究领域，目前各种方向与技术之间相互融合，研究内容也纷繁复杂。从数据分析的角度讲，数据挖掘可分为关联挖掘、数据分类、数据聚类和预测等问题。根据数据类型对象的特点又可分为时序数据分析、空间数据挖掘、文本及 Web 挖掘和多媒体数据库等研究方向。从具体算法实施角度讲，数据挖掘又有兴趣度评价、并行算法、增量算法及算法的复杂性等研究问题。

数据挖掘受到了很多学科领域的影响，其中数据库、机器学习、统计学无疑影响最大。粗糙地说，数据库提供数据管理技术，机器学习和统计学提供数据分析技术。由于统计学界往往醉心于理论的优美而忽视实际的效用，因此，统计学界提供的很多技术通常都要在机器学习界进一步研究，变成有效的机器学习算法之后才能再进入数据挖掘领域。从这个意义上说，统计学主要是通过机器学习来对数据挖掘发挥影响，而机器学习和数据库则是数据挖掘的两大支撑技术。从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域。但能否认为数据挖掘只不过就是机器学习的简单应用呢？答案是否定的。一个重要的区别是，传统的机器学习研究并不把海量数据作

为处理对象，很多技术是为处理中小规模数据设计的，如果直接把这些技术用于海量数据，效果可能很差，甚至可能用不起来。

1.4 仿生计算智能与机器学习

今天，世界上各种科学技术互相交叉、渗透，许多研究课题已经不能单靠一个领域的理论和方法来解决，许多边缘学科正是多个领域交叉发展的结果，许多研究领域的理论和方法也越来越复杂，在信息及控制科学领域尤为突出。人们研究的问题越来越复杂，而传统方法解决问题的能力越来越有限，这就促使人们不断寻求新的方法和手段，比如人工智能的研究及迅猛发展。这些研究有助于人类更好地理解自然和宇宙。事实上，生命现象和生物的智能行为一直被人工智能研究者所关注，尤其是近 10 年来人工智能的成就与生物有着密切关系，不论是从结构上模拟的人工神经网络，还是从功能上模拟的模糊逻辑系统，抑或是着眼于生物进化微观机理和宏观行为的进化算法，都有仿生的痕迹。也正是模仿生物智能行为，并借鉴其智能机理，许多解决复杂问题的新方法不断涌现，丰富了人工智能的研究领域。经过近几十年的研究与实践，人工智能研究者开始认识到，要想仿效或逐步接近人类百万年进化才达到的大脑高级智能行为，无论是传统智能，还是单独的模糊逻辑系统，或是人工神经网络都无法完成。仿生计算智能不但是人工智能研究的基础，也是其发展思路的新思考，更是方法论转变的新成果。随着模糊逻辑、神经网络、进化计算及人工免疫系统等受进化论影响的新方法的不断完善，其仿生特点也日益突出。生物是自然智能的载体，因此生物学理所当然是人工智能研究灵感的重要来源。从信息处理的视角来看，生物体就是一部优秀的信息处理机，生命现象和生物智能行为引起了许多研究者的关注，不论是结构模拟的人工神经网络，或是功能模拟的模糊逻辑系统，还是着眼于生物进化宏观行为的遗传进化算法和借鉴生物免疫机理的人工免疫系统，都是模

拟生物智能行为，学习了其智能机理进而发展为人类可以使用的计算智能和信息处理技术。1994 年 6 月，IEEE 为促进多学科的渗透与结合，把人工神经网络、模糊技术和进化计算三个年会合并举行，在美国奥兰多召开了全球第一届计算智能大会（WCCI），出版了《计算智能、模仿生命》的论文集。此次会议是计算智能的第一次综合性大会，随后，计算智能成为大家关注的研究热点。目前国际上提出的计算智能（Computational Intelligence，CI）就是以人工神经网络为主导，与模糊逻辑系统、进化计算及信号与信息处理学科的综合集成；认为新一代的计算智能信息处理技术应是神经网络、模糊系统、进化计算、混沌动力学、分形理论、小波变换、人工生命等交叉学科的综合集成。尽管对计算智能的定义、内容，以及与其他智能学科分支的关系尚没有统一的看法，但计算智能的下列两个重要特征是人们比较认同的：

（1）计算智能与传统人工智能不同，主要依赖的是生产者提供的数字材料，而不是依赖于知识；它主要借助于数学计算方法的使用。这就是说，一方面，CI 的内容本身具有明显的数值计算信息处理特征；另一方面，CI 强调用“计算”的方法来研究和处理智能问题。需要强调的是，CI 中计算的概念在内涵上已经加以拓展和加深。一般地，在解空间进行搜索的过程都被称为计算。

（2）计算智能这个概念的提出显然不仅具有科学研究分类学的意义，其积极意义还在于促进基于计算的或基于计算和基于符号物理相结合的各种智能理论、模型、方法的综合集成，以便在计算智能这个主题下发展思想更先进、功能更强大、能够解决更复杂问题的大系统的智能科学成果。由此看来，当前计算智能发展的重要方向之一就是不断引进深入的数学理论和方法，以“计算”和“集成”作为学术指导思想，进行更高层次的综合集成研究。这种综合集成研究不仅不局限在模型及算法层次的综合集成的范畴，而且还进入了感知层次及认知层次的综合集成。

由于生物是自然智能的载体，因此生物学理所当然是人工智能研究灵感的重要来源。从信息处理的角度来看，生物个体本身就是一台优秀的信息处

理机，而其所具有的完美解决问题的能力让目前最好的计算机也相形见绌。人们已经从许多角度开创不同的学科来研究生物体系。其中一个重要领域就是生物信息处理系统，许多研究人员已经在工程领域应用生物系统的信息处理功能。如人工神经网络、模糊逻辑及进化计算就是模拟生物个体的某些特征而发展起来的智能算法，由于这些算法具有高度并行性，并且具有自组织、自适应、自学习等智能特征，通过“拟物”与“仿生”使问题得到解决，它们为解决某些复杂问题提供了新的启示。

一些传统的观点认为，机器学习算法的任务是寻找准确的知识或规则，换句话说，按照评价函数而言是最优的。随着机器学习领域研究的不断深入，人们逐渐认识到，采用机器学习算法的重点已不再是寻找准确无误的知识，而是发现一些新颖的、可被人理解并有意义的新知识，通过人的参与来做出更高一级的决策，这才是知识发现的最终目的。而这些发现就整个大型数据库而言，可能只是一些次优的规则。有鉴于此，基于进化算法的机器学习方法很快受到人们的重视。进化算法是一种迭代式搜索算法，它可以在很短的时间内找到许多问题的次优解，但为求全局最优解则需要付出很大的代价。为此，人们提出了一些将该算法与已有启发式算法相互结合的混合进化算法来提高搜索过程的整体性能，就进化算法本身的构成而言，它在个体生成时的两个主要算子（交叉和变异）都是在一定发生概率的条件下，随机、没有指导地迭代搜索，因此它们在为群体中的个体提供了进化机会的同时，也无可避免地产生了退化的可能。在某些情况下，这种退化现象还相当明显。另外，很多有待处理的数据挖掘问题都会有自身一些基本的背景知识和显而易见的特征信息，然而进化算法的交叉和变异算子却相对固定，在求解问题时，可变的灵活程度较小。这无疑对算法的通用性是有益的，但却忽视了问题的特征信息对求解问题的辅助作用，特别是在求解一些规模较大的数据问题时，这种忽视所带来的损失往往就比较明显了。利用问题自身的背景知识和特征信息来进行求解，正是很多人提出混合进化算法的初衷。就目前而言，能够用于解决机器学习问题的方法很多，从它们所运用的技术特点来看，这些方法主要有三种类型：基于信息论的启发式学习方法、神经网络的学习方

法和基于生物进化机理的进化学习方法，如遗传进化、遗传规划和 DNA 计算等。

20 世纪 40 年代，计算机的产生使机器学习的实现有了可能，并且自 50 年代中期到 60 年代中期成了机器学习的高峰时期。从 60 年代中期到 70 年代中期转入低潮，主要研究侧重于基于概念的学习和基于归纳的学习；在 70 年代中期到 80 年代中期又得到了迅速发展，特别是专家系统的成功应用，不同的学习策略和各种学习方法问世，示例归约学习成为研究主流。自动知识获取成为机器学习的应用研究目标，遗传算法应用于机器学习的思想已经被提出。最近 10 多年机器学习的研究和发展进入了一个崭新时期。1986 年，神经网络重新兴起，基于连接机制的学习开始向传统的符号学习挑战。神经网络将知识的表达蕴涵于网络连接中，处理隐层和反向传播算法的发展，显示出很强的学习能力，随着各种改进型算法不断被提出，显著地改善了机器学习系统的性能。

机器学习系统实际上是对人的学习机制的一种抽象和模拟，是一种理想的学习模型。基于符号学习的机器学习系统，如监督型系统、条件反射型学习系统、类比式学习系统、推理学习系统等，只具备一些较初级的学习能力。在物理、工程、技术应用、经济等实际应用领域中，常常存在一些复杂的物理系统，这些复杂系统往往需要由多个变量和多个参数的数学模型来描述，具有非线性、耦合性。同时，一些系统的参数或者结构并不是恒定不变的，而是具有一定的时变特性。基于传统知识处理方法的系统，在对认知领域有足够完备、清晰认识的基础上，可以很好地工作，但一旦所给信息缺损或模糊化，则其认知能力会急剧降低。这是因为传统的“硬”分析方法只能在给定的匹配模式下工作，对环境的适应能力较差，传统的“硬”知识处理方法不适合处理不确定知识。因此近年来，以自然计算为基础而发展起来的各种软计算方法为此提供了一种有效的解决方法。计算智能的本质与传统的“硬”分析方法不同，其目的在于适应现实世界遍布的不精确性。因此计算智能的指导原则是开拓对不精确性、不确定性和部分的容忍，以达到可处理性、鲁棒性、

低成本求解及与现实更好的紧密联系。在最终的分析中，智能计算并不追求问题的精确解，而允许存在不精确性和不确定性，所得到的是精确或不精确问题的近似解，这是人脑求解问题的体现。计算智能的作用模型是人的思维，利用不精确性、不确定性和部分方法论的一个聚合体，它们结合起来的效果比单独使用效果更好。用此方法得到的结果具有易处理性、鲁棒性及与现实相一致性，且这些结果常常好于只用传统的计算方法得到的结果。高精度对于实际应用有时是没有意义的，大部分情况下可以牺牲精度来换取速度，提高效率。计算智能不是单一方法，而是具有合作关系的多种方法的集成。这些方法主要包括模糊逻辑、神经网络、遗传算法和粗糙集理论等，它们是相互补充的而不是相互竞争的。

机器学习中的知识发现是从数据库中挖掘出隐含的知识，使这些知识变得可用。而在现实的数据库中，一方面大量积累的数据存在内在的不精确性，另一方面多属性数据又有其内在的复杂性。智能计算方法为处理数据挖掘中的不精确性和不确定性提供了有效的技术。从学科发展的角度来看，仿生计算智能的研究是各类自然科学（特别是生命科学）和计算机科学相交叉而产生的研究领域，它的发展完全顺应当前多交叉学科不断产生和发展的潮流。目前其在经典智能算法的理论及应用的基础上，已逐步发展出许多较有潜力的研究分支：DNA 计算、蚁群系统、遗传算法、人工免疫系统、神经网络计算、模糊计算等；开发了较多的新智能工具，如免疫算法、蚁群算法、变邻域搜索、进化算法、混合优化算法等。本书所介绍的以免疫计算智能为基础的机器学习方法，便是针对具体问题而采用的多种技术和方法的集成。

1.5 免疫计算与机器学习

生物体是一个复杂的大系统，其信息处理功能是由时间和空间尺寸相异的三个子系统完成的，即脑神经系统、免疫系统和内分泌系统。免疫系统是

生物，特别是脊椎动物（包括人类）所必备的防御机制，它由具有免疫功能的器官、组织、细胞、免疫效应分子和有关的基因等组成，可以保护机体抗御病原体、有害的异物及癌细胞等致病因子的侵害。免疫功能主要包括：免疫防御、免疫稳定和免疫监视。从工程应用和信息处理角度来看，生物免疫系统为人工智能提供的许多信息处理机制，正是充分认识到生物免疫系统中蕴涵丰富的信息处理机制，Farmer 等率先基于免疫网络学说给出了免疫系统的动态模型，并探讨了免疫系统与其他人工智能方法的联系，开始了人工免疫系统的研究。人工免疫系统（Artificial Imuune System，AIS）是模仿自然免疫系统功能的一种智能方法，它实现一种受生物免疫系统启发，通过学习外界物质的自然防御机理的学习技术，提供噪声忍耐、无教师学习、自组织、记忆等进化学习机理，结合了分类器、神经网络和机器推理等系统的一些优点，因此提供了新颖的解决问题的方法和途径。其研究成果涉及机器学习、数据处理、优化学习和故障诊断等许多领域，已经成为继神经网络、模糊逻辑和进化计算后人工智能的又一研究热点。

基于免疫的计算智能之所以获得广泛重视，在于具有以下优良的特性。

1）免疫系统的动态进化性

免疫系统在与病原体之间相互竞争的同时，也在不断进化。这一特性可用于动态优化算法的求解过程，随着环境的变化，得到的优化解也是动态进化的，而传统遗传算法一般是静态寻优，往往无法在线动态寻优，免疫系统的动态特性，以及与变化的环境不断适应的过程，为动态优化提供了一种新的思路。

2）免疫系统的多尺度进化特性

免疫系统实际上是一个在动态变化环境中通过自学习而不断完善的多时间尺度进化系统。免疫系统的鲁棒性是在慢速学习阶段，保证在外来攻击的大范围内获得满意的性能；而免疫系统的自适应性是在快速学习阶段实现的。

3）免疫系统的搜索方式

免疫系统对于特殊的入侵抗原选择最适宜的抗体时，也经历了一种搜索“最优抗体”的过程。免疫系统的搜索方式是将随机性和基于反馈的高度定向行为结合起来。免疫系统这种将随机性与定向性相结合来进行搜索的特性，使免疫系统在深度搜索和广度搜索之间取得了平衡，这对于优化搜索方式也是一种启发。

4）免疫系统抗体生存的周期性

免疫系统抗体群中记忆细胞可以存活几十年，将这种免疫抗体生命周期引入优化算法，可以用于动态环境中的种群个体的不同适应性的标度，这样种群个体就具有不同的生命周期，可以获得动态优化的效果。

根据以上免疫系统的生物特性而发展得到的免疫计算智能算法具有以下突出特点：免疫计算在记忆单元基础上运行，确保了快速收敛于全局最优解；而传统进化算法则是基于父代群体，不能保证概率收敛；免疫算法评价标准是计算亲和性，包括抗体-抗原的亲和度及抗体-抗体的亲和度。这一特性反映了真实免疫系统的多样性，而传统进化算法则是简单计算个体的适应度；免疫算法通过促进或抑制抗体的产生，体现了免疫反应的自我调节功能，保证了个体的多样性，而传统进化算法只是根据适应度选择父代个体，并没有对个体多样性进行调节；免疫算法沿用了交叉变异的思想，但新抗体产生还可以借助克隆选择、免疫记忆、疫苗接种等传统进化算法中没有的机理。因此，基于免疫原理发展起来的免疫算法能有效克服其他智能算法的进化早熟现象、群体多样性不足及搜索速度慢等问题。探讨免疫系统的运行机理，模拟其运行机制并提出免疫算法解决实际问题具有理论和现实意义，为研究新的优化算法及改进现有的优化算法提供了多种新思路。

目前基于免疫计算的机器学习研究已经有了大量算法，这些算法可分为两大类：基于免疫学原理的机器学习算法，以及与其他进化算法结合的混合算法。概括起来，在机器学习领域中，应用免疫计算相关原理的主要有以下几个方面：

（1）基于免疫克隆选择原理的机器学习算法。其利用免疫系统克隆选择的机制来实现优秀抗体的扩展和增生，并利用了免疫系统中的记忆机制保证算法能够最终收敛到全局最优解，在模式识别、字符识别、分类等问题中取得较好效果。

（2）基于免疫阴性选择原理的机器学习方法。免疫系统强大而高效的模式识别能力即从非自体分子区分自分子的能力，本质上就是分类的过程。可由此设计分类器，应用于入侵检测、模式识别、故障诊断、文本分类等问题。

（3）基于抗体浓度调节原理的免疫机器学习算法。其利用抗体多样性保持机制，提高了算法的群体多样性，能有效抑制早熟现象，使免疫算法具有较好的全局收敛性，可以适用于数据降维、知识约简等问题。

（4）基于免疫学原理与其他进化算法的结合。如免疫特征与遗传算法、神经网络结合的算法解决目标分类、聚类、回归分析与预测问题。

（5）模拟免疫系统的免疫网络原理的免疫网络算法。Farmer 最早证明了分类器和免疫独特型网络模型之间的相似性，从动力学系统角度研究免疫系统的一般模式识别性质，该研究确定了许多与 Holland 的分类器系统的类似性，用于解决聚类、分类、数据挖掘等问题。

（6）基于免疫抗体记忆及免疫响应的免疫优化算法。其引入了免疫记忆和抗体多样性等免疫特性，将随机搜索过程中的局部搜索和全局搜索采用不同的促进和抑制策略，有效保证了算法的收敛速度。

（7）基于免疫疫苗接种的免疫优化算法。该算法引入疫苗接种的概念，能充分利用问题的先验知识，改善算法的性能，在机器学习的有监督和半监督学习问题中获得了良好的效果。

正是由于免疫计算智能所具备的多种特性，使其在机器学习领域有着广泛的应用前景。本书的 3～7 章将重点介绍因以上相关原理而发展的机器学习方法及工程应用。

1.6 本书的内容及结构

本书内容涉及生态学、神经学、生物学、运筹学、计算机科学等多学科交叉，主要借助生物学免疫系统的机理，通过构造新的算子和算法，将免疫计算智能与机器学习中的数据约简、数据分类、关联规则挖掘、聚类分析及大数据分析等应用进行结合，将有关免疫计算智能的算法及技术运用于工程应用领域，发展出解决实际问题的新方法、新理论，以丰富机器学习与免疫计算智能的研究内容，拓展了机器学习软计算方法的设计与实现。

全书分为七章，具体内容如下。

第 1 章：绪论。本章在简单介绍人工智能和机器学习的概念、发展的基础上，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。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，介绍人工智能、大数据发展趋势及对机器学习的挑战。最后给出了本书的基本结构和各章节的主要内容，方便指导读者阅读。

第 2 章：机器学习主流技术与方法。本章将结合机器学习最新发展，简要阐述当前机器学习领域的主要技术和方法，主要包括对传统机器学习的数理统计方法及软计算方法，包括粗糙集、遗传算法、神经网络、支持向量机等，还对近十年来重要且主流的机器学习技术，包括度量学习、多核学习、多视图学习、集成学习、主动学习、强化学习和迁移学习等进行了简单介绍。探讨了多种机器学习模式，如监督学习、弱监督学习、半监督学习、流数据挖掘、社会网络分析等，以及如何与实际应用结合、解决现实问题的需要。

第 3 章：免疫计算的基础原理。介绍了生物免疫系统的结构和组成、免疫系统的工作机制，以及生物免疫学理论中的各种理论，对生物免疫系统形成一个全面的了解。研究和讨论了免疫机器学习的模型、原理及工作机制，为本书的后续研究提供了理论基础。

第 4 章：基于免疫聚类竞争的关联规则挖掘方法。针对数据挖掘中的关联规则挖掘广度及效率问题，利用免疫抗原与抗体对应于数据原记录和候选模式，在基于克隆选择原理免疫算法的基础上引入了聚类竞争机制，加速抗体亲和力的成熟，提高全局搜索能力。这种机制提高了抗体群的多样性，避免了抗体群被少数亲和力最高的抗体占满，从而提高关联规则获取的效能。通过实验可以发现基于免疫关联规则挖掘算法具有收敛速度快的特点，而且此算法同时具有相当好的全局及局部搜索能力，这样可以得到更多符合条件的关联规则。

第 5 章：基于小生境免疫粗糙集属性约简方法。围绕数据挖掘中的高维数据属性约简难题，在粗糙集核属性的基础上，融合小生境免疫优化提出一种决策属性约简方法。将粗糙集核属性参数作为抗体编码的先验信息，通过疫苗自适应提取算法对抗体群接种疫苗，提高抗体群的多样性及稳定性。为降低属性约简的计算复杂度，引入属性集合的分类近似标准作为免疫优化的亲和度，采用小生境免疫共享机制动态调整抗体群的亲和力，提高算法局部搜索能力。通过免疫记忆算子操作促使优良个体的保存，在保证收敛速度的同时具有较强的全局和局部寻优能力。通过滚动轴承故障诊断及 UCI 数据集的属性约简实验，显示本算法在属性约简精度和效率方面具有较好效果。

第 6 章：基于免疫阴性选择的数据分类器。根据免疫否定选择原理，设计了基于掩码分段匹配的否定选择分类器，克服连续 r 位匹配法的缺陷，给出了适用于免疫优化的分类规则编码及分类信息分的评价。通过免疫进化对其进行群体优化以约简数据规则集，避免了传统分类算法缺乏全局优化能力的缺点，提高了对样本的识别能力。将该分类器用于实现文本匹配选择分类，克服传统否定选择分类方法对大样本空间分类效果不好的缺点。实验结果表明本书方法提高了数据分类的准确性，在数据分类准确率及平均信息分上优于传统的分类方法。

第 7 章：免疫网络在生物信息学中的应用。根据免疫网络优秀的分类机理，研究了 aiNet 聚类模型和 AIRS 分类算法，提出了基于免疫进化网络的分

类器。该分类方法主要采用免疫记忆池的两次网络抑制操作来改善网络结构，使记忆细胞在“特异性”与“通用性”之间得到平衡，同时采用离散增量度量亲和力，克服生物信息数据挖掘应用中 DNA 序列的特征提取和亲和力度量对分类性能的影响，提高分类器泛化性能，更好地衡量序列之间的相似性。

最后为全书总结，结合大数据应用需求，介绍了未来机器学习技术的主要发展方向、对免疫计算的机器学习方法的研究，以及未来需重点突破的内容。

参考文献

[1] Nature.Big Data. [2012-10-02]. http://www.nature.com/news/specials/bigdata/index.html.

[2] Bryant R E, Katz R H, Lazowska E D. Big-Data computing: Creating revolutionary breakthroughs in commerce, science, and society. [2012-10-02]. http://www.cra.org/ccc /docs/init/Big_Data.pdf.

[3] Science. Special online collection: Dealing with data [EB/OL]. [2012-10-02] .http:// www.sciencemag.org/site/special/data/, 2011

[4] D Agrawal, P Bernstein, E Bertino, et al.Challenges and opportunities with big data-A community white paper developed by leading researchers across the United States. [2012-10-02] .http://cra.org/ccc/.

[5] H S Lin．数据挖掘技术与工程实践．洪松林，译．北京：机械工业出版社，2014．

[6] N JNilsson．人工智能．郑扣根，译．北京：机械工业出版社，2000．

[7] 史忠植．高级人工智能．北京：科学出版社，1998．

[8] G.E Hinton., Osindero, S. and Teh, Y., A fast learning algorithm for deep belief netsNeural Computation 18: 1527-1554, 2006.

[9] Min Gui, Anil Pahwa, Sanjoy Das. Analysis of Animal-Related Outages in Overhead Distribution Systems with Wavelet Decomposition and ImmuneSystems-Based Neural Networks. IEEE Transactions on Power Systems, 2009, 24(4): 1765-1771.

[10] Muhammad Rahmat Widyanto, Benyamin Kusumoputro, Hajime Nobuhara, etal.A Fuzzy-Similarity-Based Self-Organized Network Inspired by ImmuneAlgorithm for Three-Mixture-Fragrance Recognition. IEEE Transactions on Industrial Electronics, 2006, 53(1): 313-321.

[11] Cai KaiYuan, Lei Zhang. Fuzzy Reasoning as a Control Problem. IEEE Transactions on Fuzzy Systems, 2008, 16(3): 600-614.

[12] Licheng Jiao, Lei Wang. A Novel Genetic Algorithm Based on Immunity. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2000, 30(3): 552-561.

[13] Ashish Ahuja, Sanjoy Das, Anil Pahwa. An AIS-ACO Hybrid Approach for Multi-ObjectiveDistribution System Reconfiguration. IEEE Transactions on Power Systems, 2007, 22(3): 1101-1112.

[14] Marco Dorigo , Gianni Di Caro. Ant Algorithms for Discrete Optimization. Artificial Life, 1999, 5(3): 137-172.

[15] Hong-Wei Ge, Liang Sun, Yan-Chun Liang, etal. An Effective PSO and AIS-Based Hybrid Intelligent Algorithm for Job-Shop Scheduling. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2008, 38(2): 358-368.

[16] 焦李成，杜海峰，刘芳，等．免疫优化计算学习与识别．北京：科学出版社，2007．

[17] Dipankar Dasgupta, Advances in artificial immune systems. IEEE Computational Intelligence Magazine, 2006, 1(4): 40-49.

[18] De Castro L N, Von Zuben F J. Learning and optimization using the clonal selection principle. IEEE Trans on Evolutionary Computation, 2002, 6(1): 239-251.

[19] 戚玉涛，刘芳，焦李成．基于分布式人工免疫算法的数值优化．电子学报，2009，37（7）：1554-1561．

[20] 薛文涛，吴晓蓓，徐志良．用于多峰函数优化的免疫粒子群网络算法．系统工程与电子技术，2009，31（3）：705-709．

[21] 戚玉涛，刘芳，焦李成．基于信息素模因的免疫克隆选择函数优化．计算机研究与发展，2008，45（6）：991-997．

[22] 余航，焦李成，公茂果，等．基于正交试验设计的克隆选择函数优化．软件学报，2010，21（5）：950-967．

[23] 戚玉涛，焦李成，刘芳．基于并行人工免疫算法的大规模 TSP 问题求解．电子学报，2008，36（8）：1552-1558．

[24] Anna_SSwiecicka, Franciszek Seredynski, Albert Y. Zomaya. Multiprocessor Scheduling and Reschedulingwith Use of Cellular Automata and Artificial Immune System Support.IEEE Transactions on Parallel and Distributed Systems, 2006, 17(3): 253-262.

[25] Licheng Jiao, Yangyang Li, Maoguo Gong, etal. Quantum-Inspired Immune Clonal Algorithm for Global Optimization. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics, 2008, 38(5): 1234-1433.

[26] Felipe Campel, Frederico G. Guimarães, Hajime Igarashi. Multiobjective Optimization Using Compromise Programming and an Immune Algorithm. IEEE Transactions on Magnetics, 2008, 44(6): 982-985.

[27] Aldo Canova, Fabio Freschi, Michele Tartaglia. Multiobjective Optimization of Parallel Cable Layout. IEEE Transactions on Magnetics, 2007, 43(10): 3914-3920.

[28] Xiong Hao, Sun Cai-xin. Artificial Immune Network Classification Algorithm for Fault Diagnosis of Power Transformer. IEEE Transactions on Power Delivery, 2007, 22(2): 930-935.

[29] Slavisa Sarafijanovic, Jean-Yves Le Boudec. An Artificial Immune System Approach WithSecondary Response for Misbehavior Detection in Mobile adhoc Networks. IEEE Transactions on Neural Networks, 2005, 16(5): 1076-1087.

[30] Rogerio de Lemos, Jon Timmis, Modupe Ayara, etal. Immune-Inspired Adaptable Error Detectionfor Automated Teller Machines. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2007, 37(5): 873-886.

[31] Famer J D, Packard N H, Perelson A S. The Immune System, Adaptation, and Machine Learning. Physicia D, 1986, (2):187-204.

[32] Dasgupta D, Forrest S. Artificial immune systems in industrial applications. In：Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials. 1999, (1): 257-267.

[33] Dasgupta D, Attoh Okine N. Immunity based systems: A survey. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics. 1997, (1): 369-374.

[34] 王磊，潘近，焦李成．免疫算法．电子学报，2000，28（7）：74-78．

[35] 焦李成，杜海峰，刘芳，等．免疫优化计算学习与识别．北京：科学出版社，2006：11-114．

[36] De Castro L N, Von Zuben F J. The Clonal Selection Algorithm with Engineering Applications[C].In:Proceedings of GECCO'00 Workshop on Artificial Immune Systems and Their Applications, 2000, (1): 36-37.

[37] Dasgupta D.Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences. In:Proceedings of the IEEE SMC, 1997, (1): 873-878.

[38] Ishida Y. Fully Distributed Diagnosis by PDP Learning Algorithm: Towards Immune Network PDP Model. In:Proceedings of International Joint Conference on Neural Networks, San Diego, USA, 1990: 777-782.

[39] Y. Yu, Z.-H. Zhou. A new approach to estimating the expected first hitting time of evolutionary algorithms. Artificial Intelligence, 2008, 172(15): 1809-1832.

[40] Y. Yu, X. Yao, Z.-H. Zhou. On the approximation ability of evolutionary optimization with application to minimum set cover. Artificial Intelligence, 2012, 180-181: 20-33.

[41] C. Qian, Y. Yu, Z.-H. Zhou. On constrained boolean pareto optimization. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015: 389-395.

[42] C. Qian, Y. Yu, Z.-H. Zhou. An analysis on recombination in multi-objective evolutionary optimization. Artificial Intelligence, 2013, 204: 99-119.

[43] C. Qian, Y. Yu, Z.-H. Zhou. Pareto ensemble pruning. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, 2015: 2935-2941.

[44] C. Qian, Y. Yu, Z.-H. Zhou. Subset selection by Pareto optimization. In: Advances in Neural Information Processing Systems 28, Cambridge, MA: MIT Press, 2015, in press.

[45] 陈康，向勇，喻超．大数据时代机器学习的新趋势．电信科学，2013，28（12）：88-95．

[46] G. W. Zhang, Z. H. Zhan, K. J. Du, Y. Lin, W. N. Chen, J. J. Li, J. Zhang. Parallel particle swarm optimization using message passing interface. In: Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Singapore, 2014: 55-64.

[47] Z. H. Zhan, X. F. Liu, Y. J. Gong, J. Zhang, H. S. H. Chung, Y. Li. Cloud computing resource scheduling and a survey of its evolutionary approaches. ACM Computing Surveys, 2015, 47(4): 1-33.

[48] W. Z. Zhao, H. F. Ma, Q. He. Parallel k-means clustering based on Mapreduce. In: Lecture Notes in Computer Science 5931, Springer Berlin Heidelberg, 2009: 674-679.

[49] J. Zhang, Z. H. Zhan, Y. Lin, N. Chen, Y. J. Gong, J. H. Zhong. Evolutionary computation meets machine learning: A survey. IEEE Computational Intelligence Magazine, 2011, 6(4): 68-75.

[50] 何清，李宁，罗文娟，等．大数据下的机器学习算法综述．模式识别与人工智能，2014，27（4）：327-336．

[51] 何清，庄福振，曾立，等．PDMiner：基于云计算的并行分布式数据挖掘工具平台．中国科学（F 辑）信息科学，2014，44（7）：871-885．

[52] J. Chen, K. Li, J. Zhu, W. Chen. WarpLDA: A simple and efficient O(1) algorithm for latent Dirichlet allocation. arXiv:1510.08628, 2015.

[53] 李武军，周志华．大数据哈希学习：现状与趋势．科学通报，2015，60（5/6）：485-490．

[54] Y. Weiss, A. Torralba, R. Fergus. Spectral hashing. In: Advances in Neural Information Processing Systems 21, Cambridge, MA: MIT Press, 2008: 1753-1760.

[55] De Castro L N, Von Zuben F J. An Evolutionary Immune Network for Data Clustering. In: Proceedings of the IEEE SBRN, 2000: 84-89.

[56] 刘若辰，杜海峰，焦李成．免疫多克隆策略．计算机研究与发展，2004，41（4）：571-576．

[57] Alex A. Freitas, Jon Timmis. Revisiting the Foundations of Artificial Immune Systems for Data Mining.IEEE Trans on Evolutionary Computation, 2007, 11(4): 521-540.

[58] Tien Dung Do, Siu Cheung Hui, A. C. M. Fong, etal. Associative Classification WithArtificial Immune System. IEEE Transactions on Evolutionary Computation, 2009, 13(2): 217-238.

[59] Seralozs, en, Salih Gunes, Sadk Kara, etal. Use of Kernel Functions in Artificial ImmuneSystems for the Nonlinear Classification Problems. IEEE transactions on information technology in biomedicine, 2009, 13(4): 621-628.

第 2 章

机器学习主流技术与方法

本章导读：

机器学习在过去十年经历了飞速发展，目前已经成为子领域众多、内涵丰富的学科领域。如何与实际应用结合、解决现实问题成为机器学习发展最主要的目标。在机器学习发展过程中，涌现出多种机器学习模式，例如监督学习、弱监督学习、半监督学习、代价敏感学习、流数据挖掘、社会网络分析等，这些学习模式都起源于实际应用中抽象出来的问题。本章将结合机器学习最新发展，简要阐述当前机器学习领域的主要技术和方法，主要包括传统机器学习的数理统计方法及软计算方法，如粗糙集、遗传算法、人工神经网络、支持向量机等，还会对近十年来重要且主流的机器学习技术，包括度量学习、多核学习、多视图学习、集成学习、主动学习、强化学习和迁移学习等进行简单介绍。

2.1 机器学习的发展

机器学习是人工智能研究中较为年轻的分支，它的发展过程大体上分为五个时期。

第一个时期：20 世纪 50 年代中叶到 60 年代中叶，属于启蒙时期。这个时期的研究目标是各类自组织系统和自适应系统，其主要研究方法是不断修改系统的控制参数和改进系统的执行能力，这种方法不涉及与具体任务有关的知识。这个时期的代表性工作主要有塞缪尔（Samuel）的下棋程序。1959 年美国的塞缪尔设计了一个下棋程序，这个程序具有学习能力，可以在不断的对弈中改善自己的棋艺。4 年后，这个程序战胜了设计者本人。又过了 3 年，这个程序战胜了美国一个保持 8 年不败的冠军。这个程序向人们展示了机器学习的能力，提出了许多令人深思的社会问题与哲学问题，但这种学习的结果远不能满足人们对机器学习系统的期望。

第二个时期：20 世纪 60 年代中叶到 70 年代中叶，被称为机器学习的平静时期。在这个时期，机器学习的发展步伐几乎处于停滞状态。本时期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。该时期的代表性工作有温斯顿（Winston）的结构学习系统和海斯罗思（HayesRoth）等的基本逻辑的归纳学习系统。这些研究虽然取得较大进展，但只能学习单一概念，而且未能投入实际应用。此外，神经网络学习机因理论缺陷未能达到预期效果而转入低潮。事实上，这个时期整个 AI 领域都遭遇了瓶颈。当时计算机有限的内存和处理速度不足以解决任何实际的 AI 问题。

第三个时期：从 20 世纪 70 年代中叶到 80 年代中叶，称为复兴时期。在此期间，人们从单概念学习扩展到多概念学习，探索不同的学习策略和方法。在本阶段已开始把学习系统与各种应用结合起来，并取得很大的成功，大大促进了机器学习的发展。1980 年，在美国卡内基梅隆大学召开了第一届机器

学习国际研讨会，标志着机器学习研究已在全世界兴起。经过一些挫折后，多层感知器由伟博斯在 1981 年的神经网络反向传播算法中具体提出。当然反向传播算法现在仍然是神经网络架构的关键因素。有了这些新思想，神经网络的研究又加快了。1985—1986 年，神经网络研究人员（鲁梅尔哈特、辛顿、威廉姆斯·赫、尼尔森）先后提出了 MLP 与 BP 训练相结合的理念。一个非常著名的机器学习算法由昆兰在 1986 年提出，称之为决策树算法，更准确地说是 ID3 算法，这是另一个主流机器学习的火花点。决策树是一个预测模型，代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径代表某个可能的属性值，每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有复数输出，可以建立独立的决策树以处理不同输出。在数据挖掘中，决策树是一种经常要用到的技术，可以用于分析数据，也可以用来作预测。

第四个时期：20 世纪 90 年代到 21 世纪初，称为现代机器学习技术成型期。1990 年，Schapire 最先构造出一种多项式级的算法，这就是最初的 Boosting 算法。一年后，Freund 提出了一种效率更高的 Boosting 算法。但是这两种算法存在共同的缺陷，那就是都要求事先知道弱学习算法学习正确的下限。1995 年，Freund 和 Schapire 改进了 Boosting 算法，提出 AdaBoost（Adaptive Boosting）算法，该算法的效率和 Freund 于 1991 年提出的 Boosting 算法几乎相同，但不需要任何关于弱学习器的先验知识，因而更容易应用到实际问题当中。同年，机器学习领域中一个最重要的突破——支持向量机（Support Vector Machine，SVM），由瓦普尼克和科尔特斯在大量理论和实证的条件下提出，从此将机器学习社区分为神经网络社区和支持向量机社区。支撑向量机、Boosting、最大熵方法（比如 Logistic Regression，LR）等模型的结构基本上可以看成带有一层隐层节点（如 SVM、Boosting），或没有隐层节点（如 LR）。这些模型无论是在理论分析上还是在应用中都获得了巨大的成功。另一个集成决策树模型由布雷曼博士在 2001 年提出，它由一个随机子集的实例组成，并且每个节点都是从一系列随机子集中选择。由于它的这个性质，被称为随机森林，随机森林也在理论和经验上证明了对过拟合的抵抗性。甚至连 AdaBoost 算法在数

据过拟合和离群实例中都表现出了弱点，而随机森林则是针对这些警告更稳健的模型。随机森林在许多不同的任务，像 DataCastle、Kaggle 等比赛中都表现出了成功的一面。这个时期的机器学习围绕三个主要研究方向进行：① 面向任务学习。在预定的一些任务中分析和开发学习系统，以便改善完成任务的水平，这是专家系统研究中提出的研究问题；② 认识模拟研究。主要研究人类学习过程及其计算机的行为模拟，这是从心理学角度研究的问题；③ 理论分析研究。从理论上探讨各种可能学习方法的空间和独立于应用领域之外的各种算法。这三个研究方向都有自己的研究目标，每一个方向的进展都会促进另一个方向的研究。这三个方向的研究都将促进各方面问题和学习基本概念的交叉结合，推动了整个机器学习的研究。

第五个时期：21 世纪初至今，这是现代机器学习技术的蓬勃发展期，其研究扩展到模式识别、计算机视觉、语音识别、自然语言处理等多个领域，如图 2-1 所示。

图 2-1　现代机器学习技术应用领域

在这个时期内，机器学习相关学习算法层出不穷，而且实用性越来越高，取得了许多突破性应用成果。

1997 年，“深蓝”在国际象棋上击败当时的世界冠军，成为了第一个击败当时世界国际象棋冠军的计算机系统。“深蓝”利用了 20 世纪 90 年代已经得到了发展的计算机能力来执行潜在走子方式的大规模搜索，并从中选择最好的走子步骤。

2011 年，在美国一个电视解密节目 Jeopardy！上，IBM 的 Watson 系统击败了两位人类冠军。

2012 年，ImageNet 分类赛与计算机视觉的进步：Alex Krizhevsky、Ilya Sutskever 和 Geoffrey Hinton 发表了一篇很具有影响力的论文。该论文所描述的模型赢得了 ImageNet 年度图像识别比赛，并极大地降低了图像识别的错误率。

2016 年，AlphaGo 击败世界围棋冠军棋手：Google DeepMind 创造的 AlphaGo 围棋系统以 4 : 1 的成绩击败了世界顶级围棋棋手李世石。

2017 年，学习玩扑克：卡内基梅隆大学的研究者开发了一个名叫 Libratus 的系统，其在历时 20 天的无限制德州扑克比赛上击败了四位人类顶级玩家。阿尔伯塔大学研究者开发的 Deepstack 系统也取得了类似的成功。

早期机器学习中的专家系统曾是人工智能研究工作者的骄傲。在研究一个专家系统时，知识工作者首先从领域专家那里获得知识，但知识获取是专家系统研究中公认的瓶颈问题。其次，在整理、表达从领域专家那里获得的知识时，知识表示又成为一大难题。最后，即使某个领域的知识通过一定方式获取并表达了，但这样做成的专家系统对常识和百科知识却出奇的贫乏。以上三大难题大大限制了专家系统的应用，使专家系统目前还停留在发动机故障诊断一类的水平上。人工智能学者，尤其是从事机器学习的科学工作者开始正视现实生活中大量的、不完全的、有噪声的、模糊的和随机的大数据样本，也开始走上了数据挖掘的道路。通过 50 多年的发展，目前机器学习已成为计算机科学中研究内涵极其丰富、新技术和新应用层出不穷的重要研究分支。国际上关于机器学习的主要学术会议包括每年定期举行的国际机器学习会议（ICML）、国际神经信息处理系统会议（NIPS）、欧洲机器学习会议

（ECML）及亚洲机器学习会议（ACML）等，主要学术期刊包括 *Machine Learning*、*Journal of Machine Learning Research*、*IEEE Transactions on Neural Networks and Learning Systems* 等。此外，人工智能领域的一些主要国际会议（如 IJCAI、AAAI 等）和国际期刊（如 *Artificial Intelligence*、*IEEE Transactions on Pattern Analysis and Machine Intelligence* 等）也经常发表与机器学习相关的最新研究成果。国内机器学习的重要学术活动包括每两年举行一次的中国机器学习会议（China Conference on Machine Learning，CCML），该会议目前由中国人工智能学会和中国计算机学会联合主办，中国人工智能学会机器学习专业委员会和中国计算机学会人工智能与模式识别专业委员会协办。此外，还有每年举行的中国机器学习及其应用研讨会（Chinese Workshop on Machine Learning and Applications，MLA）。

早期机器学习研究通常假设数据具有相对简单的特性，如数据来源单一、概念语义明确、数据规模适中、结构静态稳定等。当数据具有以上简单特性时，基于现有的机器学习理论与方法可以有效实现数据的智能化处理。然而，在大数据时代背景下，数据往往体现出多源异构、语义复杂、规模巨大、动态多变等特殊性质，为传统机器学习技术带来了新的挑战。为应对这一挑战，国内外科技企业巨头，如谷歌、微软、亚马逊、华为、百度等，纷纷成立以机器学习技术为核心的研究院，以充分挖掘大数据中蕴含的巨大商业与应用价值。可以预见，在未来相当长的一段时期内，机器学习领域的研究将以更广泛、更紧密的方式与工业界深度耦合，推动信息技术及产业的快速发展。

2.2 机器学习中的统计分析方法

数理统计是数学中最重要、最活跃的学科之一，然而它和机器学习技术结合得并不紧密，但一旦有了从数据查询到知识发现、从数据演绎到数据挖掘的要求，数理统计就获得了新的生命力。数理统计分析作为机器学习的三

个主要支柱之一，有许多寻找变量之间规律性的方法，而回归分析方法是其中最有效的方法之一。本节对作为数据挖掘机制之一的回归分析方法进行简单介绍。

机器学习利用了人工智能和统计分析的进步带来的许多好处，这两门学科都致力于模式发现和预测。一些新兴的技术同样在知识发现领域取得了很好的效果，如神经网络和决策树，在足够多的数据和计算能力下，它们几乎不用人的参与就能自动完成许多有价值的功能。机器学习把统计分析和人工智能的算法及技术封装起来，使人们不用了解这些技术的细节就能实现许多有价值的功能，从而使人们把更多的精力专注于所要解决的问题。机器学习与统计分析这两者之间的主要区别在于算法对大数据量的适应性，面对记录数达 10 万条以上的数据集，机器学习的算法必须仍然具有很好的适应性、鲁棒性。面对周期性数据集和流式数据集的更新时，机器学习需要考虑能对这些增量数据进行处理而不必从头计算一次。机器学习还需考虑如何处理数据集大于内存的问题及并行处理问题。而大多数统计分析技术都基于完善的数学理论和严谨的推理过程，预测的准确度还是令人满意的，但对使用者的数学基础有很高的要求。随着计算机性能的不断增强，便可以利用计算机强大的计算能力，通过相对简单和固定的方法完成复杂的推理过程。机器学习就其算法本身而言，很大一部分可以从数理统计中获得理论解释。但是作为一个整体的研究方向，应该从计算机的层面进行全局的考虑，即从系统的角度进行分析。毕竟机器学习是面向应用的，如果一个算法只能对几百条数据进行分析，那么它的用途将大打折扣。

在现实世界中，某个变量与其他一个或多个变量之间常存在着一定的关系。一般说来，变量之间的关系可分为两类：一类是确定性的关系，也就是通常所说的函数关系；另一类是非确定性关系，变量之间的这种非确定性关系称为相关关系。对于具有相关关系的变量，虽然不能找到它们之间的精确表达式，但是通过大量的观测数据，可以发现它们之间存在一定的统计规律性。设有两个变量 X 和 Y，其中 X 是可以精确测量或控制的非随机变量，而 Y

是随机变量，X 变化将使 Y 发生相应变化，但它们之间的变化关系是不确定的。若当 X 取得任一可能值 x 时，Y 相应地服从一定的概率分布，则称随机变量 Y 与变量 X 之间存在相关关系。设进行 n 次独立实验，测得实验数据如表 2-1 所示。

表 2-1　实验数据表

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

其中，x_i 及 y_i 分别是变量 X 与随机变量 Y 在第 i 次实验中的观测值 $(i=1,2,\cdots,n)$。取 $X=x$ 时随机变量 Y 数学期望 $E(Y)\big|X=x$ 时的估计值来表示这组观测值的最佳值，如式（2.1）。

$$\hat{y}=\hat{Y}\big|_{X=x}=E(Y)\big|_{X=x} \tag{2.1}$$

显然，当 x 变化时，$E(Y)\big|_{X=x}$ 是 x 的函数，如式（2.2）。

$$\mu(x)=E(Y)\big|_{X=x} \tag{2.2}$$

因此可以用一个确定的函数关系式如式（2.3），大致地描述 Y 与 X 之间的相关关系。

$$\hat{y}=\mu(x) \tag{2.3}$$

其中，函数 $\mu(x)$ 称为 Y 关于 X 的回归函数，式（2.3）称为 Y 关于 X 的回归方程。回归方程反映了 Y 的数学期望 $E(Y)$随 X 的变化而变化的规律。因此，从统计学角度，回归分析是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。回归分析按照涉及的变量的多少，分为一元回归分析和多元回归分析；在线性回归中，按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。如果在回归分析中只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，那么这种回归分析称为

一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且自变量之间存在线性相关，则称为多重线性回归分析。

回归分析中有多个自变量时，这里有一个原则问题，即这些自变量的重要性，究竟谁是最重要的、谁是比较重要的、谁是不重要的。然而，要找到合适的回归函数 $\mu(x)$ 是很困难的，通常总是限制 $\mu(x)$ 为某一类型的函数。函数 $\mu(x)$ 的类型可以由与被研究问题的本质有关的物理假设来确定。有些时候，我们很难精确地选择并确定函数 $\mu(x)$ 的类型，只能根据在实验结果中得到的散点图来确定。在确定了函数 $\mu(x)$ 的类型后，就可以设

$$\mu(x)=\mu(x;a_1,a_2,\cdots,a_k) \tag{2.4}$$

其中，$a_1,a_2,\cdots,a_k$ 为未知参数。于是问题就归结为：如何根据实验数据合理地选择参数 $a_1,a_2,\cdots,a_k$ 的估计值 $\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_k$，使方程式（2.5）在一定的意义下“最佳地”表现 Y 与 X 之间的相关关系。

$$\hat{y}=\mu(x;\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_k) \tag{2.5}$$

解决上述问题可以利用最小二乘法，即要求选取 $\mu(x;\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_k)$ 中的参数，使得观测值 y_i 与相应的函数值 $\mu(x;\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_k)$ $(i=1,2,\cdots,n)$ 的偏差平方和最小。最小二乘法（又称最小平方法）是一种数学优化技术。它通过最小化误差的平方和来寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小。最小二乘法还可用于曲线拟合，其他一些优化问题也可通过最小二乘法来表达。

下面从概率论的观点来说明最小二乘法的理论依据。设当变量 X 取任意实数 x 时，随机变量 Y 服从正态分布 $N(\mu(x),\sigma^2)$，即 Y 的概率密度为式（2.6）

$$f(y)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{1}{2\sigma^2}[y-\mu(x)]^2} \tag{2.6}$$

其中，$\mu(x)=\mu(x;a_1,a_2,\cdots,a_k)$，而 σ^2 是不依赖于 x 的常数。因为在 n 次独立实验中得到观测值 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，所以在利用最大似然估计法估计

未知参数 $a_1,a_2,\cdots,a_k$ 时，有似然函数 L

$$L(a_1,a_2,\cdots,a_k)=\prod_{i=1}^{n}f(y_i)=(\frac{1}{\sqrt{2\pi}\sigma})^n\mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-\mu(x_i)]^2} \tag{2.7}$$

要使似然函数 L 取得最大值，则应使式（2.7）指数中的平方和取最小值，如式（2.8）取最小值。

$$S=\sum_{i=1}^{n}\left[y_i-\mu(x_i;a_1,a_2,\cdots,a_k)\right]^2 \tag{2.8}$$

也就是为了使观测值 $(x_i,y_i)(i=1,2,\cdots,n)$ 出现的可能性最大，应当选择参数 $a_1,a_2,\cdots,a_k$ 使得观测值 y_i 与相应的函数值 $\mu(x;\hat{a}_1,\hat{a}_2,\cdots,\hat{a}_k)$ 的平方和最小。这就是最小二乘法的概率意义。

在式（2.8）中，分别求 S 对 $a_1,a_2,\cdots,a_k$ 的偏导数，并令它们等于零，得到式（2.9）

$$\begin{cases}\sum_{i=1}^{n}\left[y_i-\mu(x_i;a_1,a_2,\cdots,a_k)\right]\dfrac{\partial}{\partial a_1}\mu(x_i;a_1,a_2,\cdots,a_k)=0\\ \sum_{i=1}^{n}\left[y_i-\mu(x_i;a_1,a_2,\cdots,a_k)\right]\dfrac{\partial}{\partial a_2}\mu(x_i;a_1,a_2,\cdots,a_k)=0\\ \cdots\cdots\\ \sum_{i=1}^{n}\left[y_i-\mu(x_i;a_1,a_2,\cdots,a_k)\right]\dfrac{\partial}{\partial a_k}\mu(x_i;a_1,a_2,\cdots,a_k)=0\end{cases} \tag{2.9}$$

解方程组（2.9）求出参数 $a_1,a_2,\cdots,a_k$ 的估计值，代入式（2.5）即可得到回归方程。但是，一般来说，解方程组（2.9）是很困难的，仅当函数 $\mu(x_i;a_1,a_2,\cdots,a_k)$ 是未知参数 $a_1,a_2,\cdots,a_k$ 的线性函数时，可以比较容易地求得这些参数的估计值。

2.2.1 线性回归分析

线性回归是利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回

归系数的模型参数的线性组合。只有一个自变量的情况称为简单回归，多于一个自变量的情况叫作多元回归。回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。为了便于确定回归函数 $\mu(x)$ 中未知参数的值，首先讨论变量 Y 与 X 之间存在线性相关关系的情形。

设变量 Y 与 X 之间存在线性相关关系，则由实验数据得到的点 $(x_i, y_i)(i=1,2,\cdots,n)$ 将散布在某一直线周围，于是可以用线性方程（2.10）大致地描述变量 Y 与 X 之间的关系。

$$\hat{y} = a + bx \tag{2.10}$$

设随机变量

$$Y \sim N(a+bx, \sigma^2) \tag{2.11}$$

按最小二乘法确定未知参数 a 和 b 时，有偏差平方和

$$S = \sum_{i=1}^{n} (y_i - a - bx)^2 \tag{2.12}$$

为了使 S 取得最小值，分别求 S 对 a 和 b 的偏导数，并令它们等于零，得方程组

$$\begin{cases} \sum\limits_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum\limits_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{cases} \tag{2.13}$$

整理得

$$\begin{cases} na + \left(\sum\limits_{i=1}^{n} x_i \right) b = \sum\limits_{i=1}^{n} y_i \\ \left(\sum\limits_{i=1}^{n} x_i \right) a + \left(\sum\limits_{i=1}^{n} x_i^2 \right) b = \sum\limits_{i=1}^{n} x_i y_i \end{cases} \tag{2.14}$$

解方程组（2.14）得

$$\begin{cases}\hat{a}=\overline{y}-\hat{b}\overline{x}\\ \hat{b}=\dfrac{l_{xy}}{l_{xx}}\end{cases} \tag{2.15}$$

其中，$\overline{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i$，$\overline{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$，$l_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2=(n-1)s_x^2$，$s_x^2$ 是观测值 $x_1,x_2,\cdots,x_n$ 的样本方差。

$$l_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-n\overline{xy} \tag{2.16}$$

为了以后进一步分析的需要，再引进

$$l_{yy}=\sum_{i=1}^{n}(y_i-\overline{y})^2=(n-1)s_y^2 \tag{2.17}$$

其中，s_y^2 是观测值 $y_1,y_2,\cdots,y_n$ 的样本方差。

将由式（2.15）计算得到的 $\hat{a}$ 和 $\hat{b}$ 的值代入式（2.10），就得到所求的线性方程

$$\hat{y}=\hat{a}+\hat{b}x \tag{2.18}$$

这个方程称为 Y 关于 X 的线性回归方程，$\hat{b}$ 称为回归系数，对应的直线称为回归直线。对于变量 X 与 Y 的任何一组数据 $(x_i,y_i)(i=1,2,\cdots,n)$，只要 $x_1,x_2,\cdots,x_n$ 不全相等，则无论 Y 与 X 之间是否存在线性相关关系，都可以按上述计算方法求得一个线性方程。线性回归是回归分析中第一种经过严格研究并在实际应用中广泛使用的类型。这是因为线性依赖于其未知参数的模型比非线性依赖于其未知参数的模型更容易拟合，而且产生的估计的统计特性也更容易确定。

2.2.2 非线性回归分析

在实际问题中，有一类模型的回归参数不是线性的，也不能通过转换的

方法将其变为线性的参数，这类模型称为非线性回归模型。处理非线性回归的基本方法是，通过变量变换将非线性回归化为线性回归，然后用线性回归方法处理。假定根据理论或经验，已获得输出变量与输入变量之间的非线性表达式，但表达式的系数是未知的，要根据输入输出的 n 次观察结果来确定系数的值。按最小二乘法原理来求出系数值，所得到的模型为非线性回归模型。如果回归模型的因变量是自变量的一次以上函数形式，回归规律在图形上表现为形态各异的各种曲线，称为非线性回归。非线性函数的求解一般可分为将非线性变换成线性和不能变换成线性两大类。这里主要讨论可以变换为线性方程的非线性问题。这时，选择适当的回归曲线方程可能更符合实际情况。我们用下述两种方法来讨论非线性回归问题。根据专业知识或散点图，选择适当的曲线回归方程

$$\hat{y} = \mu(x;a,b) \tag{2.19}$$

其中，a 和 b 为未知参数。为了求参数 a 及 b 的估计值，往往可以通过变量置换，把非线性回归化为线性回归，然后用上述线性回归方法来确定这些参数的估计值。

为了便于读者选择适当的曲线类型，表 2-2 列举了某些常用的曲线方程，并给出了相应的化为线性方程的变量置换公式。

表 2-2　线性置换公式表

曲线方程	变换公式	变换后的线性方程
$\frac{1}{y} = a + \frac{a}{x}$	$u = \frac{1}{x}, v = \frac{1}{y}$	$v = a + bu$
$y = ax^b$	$u = \ln x, v = \ln y$	$v = a_1 + bu(a_1 = \ln a)$
$y = a + b\ln x$	$u = x, v = \ln y$	$v = a + bu$
$y = ae^{bx}$	$u = x, v = \ln y$	$v = a_1 + bu(a_1 = \ln a)$
$y = ae^{\frac{b}{x}}$	$u = \frac{1}{x}, v = \ln y$	$v = a_1 + bu(a_1 = \ln a)$

2.2.3 多元线性回归分析

随机变量 Y 与一个变量 X 的回归分析，通常称为一元回归分析。但在实际问题中，某个随机变量可能与多个变量有相关关系。研究随机变量 Y 与 $m(m\geq 2)$ 个变量 $x_1,x_2,\cdots,x_m$ 之间的相关关系，需要利用多元回归分析，这里我们仅简要地介绍多元线性回归。

设进行 n 次独立实验，得到的实验数据如表 2-3 所示。

表 2-3 实验数据

X_1	x_{11}	x_{12}	…	x_{1n}
X_2	x_{21}	x_{22}	…	x_{2n}
⋮	⋮	⋮	⋮	⋮
X_m	x_{m1}	x_{m2}	…	x_{mn}
Y	y_1	y_2	…	y_n

其中，$x_{1k},x_{2k},\cdots,x_{mk}$ 及 $y_k(k=1,2,\cdots,n)$ 分别表示变量 $X_1,X_2,\cdots,X_m$ 及 Y 在第 k 组的实例值。

若随机变量 Y 与变量 $X_1,X_2,\cdots,X_m$ 之间存在线性相关关系，则可用多元线性方程（2.20）来描述。

$$\hat{y}=a+b_1x_1+b_2x_2+\cdots+b_mx_m \tag{2.20}$$

设对于变量 $X_1,X_2,\cdots,X_m$ 的任意一组实数值 $(x_1,x_2,\cdots,x_m)$，随机变量

$$Y\sim N(a+\sum_{i=1}^{n}b_ix_i,\sigma^2) \tag{2.21}$$

与一元回归分析类似，我们利用最小二乘法确定其中的未知参数 $a,b_1,b_2,\cdots,b_m$ 的估计值。为了使偏差平方和（2.22）取得最小值，分别求 S 对 $a,b_1,b_2,\cdots,b_m$ 的偏导数，并让它们等于零。

$$S=\sum_{k=1}^{n}\left(y_k-a-b_1x_{1k}-b_2x_{2k}-\cdots-b_mx_{mk}\right)^2 \tag{2.22}$$

整理得方程组（2.23）

$$\begin{cases} na+\left(\sum_{k=1}^{n}x_{1k}\right)b_1+\cdots+\left(\sum_{k=1}^{n}x_{mk}\right)b_m=\sum_{k=1}^{n}y_k \\ \left(\sum_{k=1}^{n}x_{1k}\right)a+\left(\sum_{k=1}^{n}x_{1k}^2\right)b_1+\cdots+\left(\sum_{k=1}^{n}x_{1k}x_{mk}\right)b_m=\sum_{k=1}^{n}x_{1k}y_k \\ \cdots\cdots \\ \left(\sum_{k=1}^{n}x_{mk}\right)a+\left(\sum_{k=1}^{n}x_{mk}x_{1k}\right)b_1+\cdots+\left(\sum_{k=1}^{n}x_{mk}^2\right)b_m=\sum_{k=1}^{n}x_{mk}y_k \end{cases} \tag{2.23}$$

设 $\overline{x}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ik}$，其中，$i=1,2,\cdots,m$；$\overline{y}=\frac{1}{n}\sum_{k=1}^{n}y_k$。

$$l_{ij}=l_{ji}=\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)(x_{jk}-\overline{x}_j)=\sum_{k=1}^{n}x_{ik}x_{jk}-n\overline{x}_i\overline{x}_j \tag{2.24}$$

其中，$i=1,2,\cdots,m;j=1,2,\cdots m$；特别是当 $i=j$ 时，有

$$l_{ii}=\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)^2=(n-1)s_i^2 \tag{2.25}$$

其中，s_i^2 表示 X_i 观测值 $x_{i1},x_{i2},\cdots,x_{in}$ 的样本方差。

$$l_{iy}=\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)(y_k-\overline{y})=\sum_{k=1}^{n}x_{ik}y_k-n\overline{x}_i\overline{y} \tag{2.26}$$

其中，$i=1,2,\cdots,m$。利用消元法不难将方程组（2.23）化为下面的方程组

$$\begin{cases} a+\overline{x}_1b_1+\overline{x}_2b_2+\cdots+\overline{x}_mb_m=\overline{y} \\ l_{11}b_1+l_{12}b_2+\cdots+l_{1m}b_m=l_{1y} \\ l_{21}b_1+l_{22}b_2+\cdots+l_{2m}b_m=l_{2y} \\ \cdots\cdots \\ l_{m1}b_1+l_{m2}b_2+\cdots+l_{mm}b_m=l_{my} \end{cases} \tag{2.27}$$

于是，我们可以先从方程组（2.27）的后 m 个方程解得 $\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_m$；再代入第一个方程，即得式（2.28）

$$\hat{a} = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \cdots - b_m\overline{x}_m \tag{2.28}$$

最后，把解得的 $\hat{a},\hat{b}_1,\hat{b}_2,\cdots,\hat{b}_m$ 代入方程（2.28），就得到多元线性回归方程（2.29）

$$\overline{y} = \hat{a} + b_1\overline{x}_1 + b_2\overline{x}_2 + \cdots + b_m\overline{x}_m \tag{2.29}$$

2.3 机器学习中的现代技术方法

现实世界中的复杂系统往往需要由多个变量和多个参数的数学模型来描述，具有非线性、耦合性，同时，一些系统的参数或结构并不是恒定不变的，而是具有一定的时变特性。基于传统统计分析的知识处理方法，在对认知领域有足够完备、清晰认识的基础上，可以很好地工作。如果所给信息缺损或模糊化，则其认知能力会急剧降低。这是因为传统的统计分析方法只能在给定的匹配模式下工作，对环境的适应能力较差，不适合处理不确定知识。现实世界中的实际问题或系统往往具有高度非线性及复杂性的基本特征，这就迫切需要建立与之相适应的计算技术。机器学习中现代技术方法往往以智能软计算为基础，其目的在于适应现实世界遍布的不精确性。因此软计算的指导原则是开拓对不精确性、不确定性和部分的容忍，以达到可处理性、鲁棒性、低成本求解及与现实更好的紧密联系。在最终的实际问题求解分析中，软计算方法并不追求问题的精确解，而允许存在不精确性和不确定性，所得到的是精确或不精确问题的近似解，这是人脑求解问题的体现。软计算的作用模型是人的思维。软计算不是单一方法，而是具有合作关系的多种方法的集成。这些方法主要包括模糊逻辑、神经网络、遗传算法和粗糙集理论等，同时衍生出各种新颖而实用的机器学习方法。比如支持向量机、强化学习、相似学习、多核学习、集成学习、主动学习、迁移学习等，本节将对这些机器学习的现代技术和方法进行简要介绍。

2.3.1 粗糙集

粗糙集（Rough Set, RS）理论是 1982 年由 Z. Pawlak 提出的一种描述不完整性和不确定性的数学理论，它从新的角度对知识进行了定义，把知识看作关于论域的划分，从而认为知识是有粒度的，知识的粒度性是造成使用已有知识不能精确地表示某些概念的原因。这就产生了所谓的关于不精确的“边界”思想。粗糙集理论中的模糊性就是一种基于边界的概念，即一个模糊的概念具有模糊的不可被明确划分的边界。在没有掌握所有关于对象域的知识的情况下，人们只能用一对逼近来描述对象域上的集合。粗糙集理论认为知识就是将对象进行分类的能力。假定我们起初对全域里的元素（对象）具有必要的信息或知识，通过这些知识能够将其划分为不同的类别。若两个元素具有相同的信息，则它们就是不可区分的，即根据已有的信息不能够将其划分开，显然这是一种等价关系。不可区分关系是粗糙集理论最基本的概念，在此基础上引入了成员关系、上近似和下近似等概念来刻画不精确性与模糊性。

粗糙集理论与传统集合理论有相似之处，但是它们的出发点完全不同。传统集合理论认为：一个集合完全由它的元素所决定，一个元素要么属于这个集合，要么不属于这个集合。模糊集对此作了改进，它给成员赋予一个隶属度，使得模糊集能够处理一定的模糊和不确定数据，但是其模糊隶属度需要人为给定，这给它的应用带来了不便。传统集合理论和模糊集理论都是把成员关系作为原始概念来处理的，集合的并和交就建立在其元素的隶属度的 max 和 min 操作上，因此其隶属度必须事先给定（传统集合默认隶属度为 1 或 0）。在粗糙集理论中，成员关系不再是一个原始的概念，因此我们无须人为地给元素指定一个隶属度，从而避免了主观因素的影响。然而粗糙集理论和模糊集理论并不是互相竞争的理论，而是互补的。粗糙集理论基于知识的不可区分性，模糊集理论则侧重知识的模糊性。不可区分性和模糊性实际上是不完全知识的两个不同侧面。不可区分性是指知识的粒度，影响所讨论的域的定义。粗糙集假定我们起初对全域里的元素（对象）有必要的信息或知识，即我们通过这些知识能够将其划分为不同的类别。若我们对两个元素具

有相同的信息，则它们就是不可区分的，即根据已有的信息我们不能够将其划分开。模糊性是由自然语言的范畴经常是渐进的概念所导致，因此模糊性是指集合具有平滑的边界。借用图像处理中的一个例子说明，粗糙集理论是关于像素的大小，而模糊集理论是关于多个灰度级别的存在；模糊集理论依赖于表达成员关系的强度的有序关系，而粗糙集是基于等价关系，这些等价关系表示由不可区分对象类所形成的划分。因此，它们之间有一定的区别，是一种自然的补充。两种理论结合使用会带来好处，因为有时模糊性和粗糙性会相互影响。

机器学习过程中的知识在不同的范畴内有多种不同的含义。在粗糙集理论中，知识被认为是一种对对象进行分类的能力。

定义 2.1 设$U \neq \Phi$是感兴趣的对象组成的有限集合，称为论域。任何子集$X \subseteq U$称为U中的一个概念或范畴，空集也认为是一个概念。则U中的一簇概念就称为关于U的知识。

定义 2.2 设$U \neq \Phi$是论域，$C = \{X_1, X_2, X_3, \cdots, X_n\}$，使得$X_i \subseteq U$，$X_i \neq \Phi$，$X_i \cap X_i = \Phi$，且$\bigcup X_i = U$，则称$C$为$U$的一个划分，$X_i$称为划分$C$的一个等价类。$U$上的一簇划分称为关于$U$的一个知识库。$U$上的一个划分与其上的一个等价关系是等价的，每一个等价关系描述的是论域U上的某一个属性，即属性亦可看作一个等价关系。

一个知识系统S可以表示成$S = <U, A, V, f>$，其中，U是所有对象的集合，称为论域；$A = C \bigcup D$是属性的集合，子集C和D分别称为条件属性和结果属性；$V = U_{r \in R} V_r$是属性的集合，V_r表示属性$r \in R$值域，$f : U \times R \to V$是一个信息函数，它指定U中每一个对象x的属性值。粗糙集理论利用信息表来描述论域中的对象，是一张二维表，每一行描述一个对象，每一列描述对象的一个属性，信息表也称为决策表、属性-值表、数据表。信息表也可简记为$S = (U, A)$。

定义 2.3 对于知识系统$S = (U, A)$，$B \subseteq A$，定义B在U上的不可分辨关

系 Ind(B)为 $\text{Ind}(B)=\{(x,y)\in U\times U: f(x,a)=f(y,a),\forall a\in B\}$。如果$(x,y)\in \text{Ind}(B)$，则称 x 和 y B-不可分辨。

显然，不可分辨关系是一种等价关系，Ind(B)的所有等价类族，即由 B 决定的划分用 U/Ind(B)表示，或简记为 U/B，包含元素 x 的等价类用 $I_B(x)$ 表示。I_B 的等价类或划分 U/B 的块称为 B-基本集。令 $X\subseteq U$，R 为 U 上的一种等价关系。当 X 能表达成某些 R 基本类的并时，称 X 为 R 可定义的；否则称为 R 不可定义的。R 可定义集是论域 U 的子集，它可以在知识库中精确地定义，所以也可以称为 R 精确集；而 R 不可定义集不能在这个知识库中精确地定义，所以称为 R 非精确集或 R 粗糙集。

对于粗糙集可以近似地定义，粗糙集理论利用两个精确集——粗糙集的上近似集和下近似集来描述，即利用不可分辨关系导出的论域划分来描述论域的新子集。

定义 2.4 设 S 为信息表，U 为论域，X 为 U 的非空子集，$B\subseteq A$ 且 $B\neq\Phi$。集合 X 的 B-下近似集和 B-上近似集分别定义如式（2.30）、式（2.31）

$$\underline{B}(X)=\bigcup\{Y_i\in U/\text{Ind}(B): Y_i\subseteq X\} \tag{2.30}$$

$$\overline{B}(\text{X})=\bigcup\{Y_i\in U/\text{Ind}(B): Y_i\cap X\neq\Phi\} \tag{2.31}$$

$\underline{B}(X)$ 实际上是由那些根据已有知识判断肯定属于 X 的对象所组成的最大集合，有时也称为 X 的 B-正域，记为 $\text{POS}_B(X)$；而根据已有知识判断肯定不属于 X 的对象组成的集合称为 X 的 B-负域，记为 $\text{NEG}_B(X)$。$\overline{B}(X)$ 是由所有与 X 的交集不为空的等价类 $I_B(Y_i)$的并集，是那些可能属于 X 的对象组成的最大集合。

X 的 B-边界域 $\text{Bn}_B(X)$定义为：$\text{Bn}_B(X)=\overline{B}(X)-\underline{B}(X)$。$\text{Bn}_B(X)$是 X 的可疑域，不能肯定其中的元素是否属于 X。

如果 X 的 B-边界域为空，即 $\text{Bn}_B(X)=\Phi$，则集合 X 是关于 B 的精确集合，即 X 可表示为一定数量的 B-基本集的并集；如果 $\text{Bn}_B(X)\neq\Phi$，则集合 X 是关于 B 的非精确集合，并且利用 $\overline{B}(X)$ 和 $\underline{B}(X)$ 来近似。

具有 B-下近似和 B-上近似的集合 $X \subseteq U$ 称为一个粗糙集。

由于存在边界区域，即有些元素既不能在论域 U 的某个子集上被分类，也不能在它的补集上被分类，所以集合存在不精确性。集合的边界域越大，它的精确性就越低，粗糙集理论引入了近似精度的概念，用来度量集合定义的不精确程度。

定义 2.5 设 B 为属性集，则有属性集 B 定义 X（$X \neq \Phi$）的近似精度为式（2.32）

$$\alpha_B(X) = \frac{\underline{B}(X)}{\overline{B}(X)} \tag{2.32}$$

显然，$0 \leq \alpha_B(X) \leq 1$。如果 $\alpha_B(X)=1$，则 X 为关于 B 的精确集合；如果 $\alpha_B(X)<1$，则 X 为关于 B 的粗糙集。

粗糙集理论在知识表达系统的基础上定义了约简与核两个非常重要的概念，进而提供了分析多余属性的方法，对知识的处理是通过对决策表中的属性值的处理实现的。一般先删除重复的实例及多余的属性，对每个实例删除多余的属性值，然后求出最小约简，并根据最小约简求出逻辑规则。随着决策表的不断增大，知识约简的复杂性呈现指数增长，采用遗传算法寻求较优的约简是一种较好的方法。

（1）令 R 为一等价关系簇，并且 $r \in R$，如果 $\mathrm{Ind}(R) = \mathrm{Ind}(R - \{r\})$，则称 r 为 R 中可省略的，否则 r 为 R 中不可省略的。

当 $\forall r \in R$ 时，如果 r 不可省略，则簇 R 为独立的。在用属性集 R 来表达系统的知识时，R 为独立的，意味着属性集合中的每个属性都是必不可少的。通俗地说，R 是表达研究对象的属性集合，在近似表达中有一些特征作用不大，可以将这些属性删除而不影响我们对对象的表达，去掉冗余属性 r 后，剩下的属性集仍然保留其等价关系。当 R 是独立的时，如果存在属性子集 $P \in R$，则 P 也是独立的。

（2）对于属性子集 $P \in R$，若存在 $Q = P - r$，$Q \in P$，使得 $\mathrm{Ind}(Q)=\mathrm{Ind}(P)$，

且 Q 为最小子集，则 Q 称为 P 的约简，表示为 Red(P)。一个属性集合可以有多种化简。若 Q 为最小子集，就是说：不存在任何子集 $T \in Q$，有 $\mathrm{Ind}(T) = \mathrm{Ind}(P)$。

（3）P 中所有约简属性集中都包含不可省略关系的集合，即约简集 Red(P) 的交集，称为 P 的核，它是表达知识必不可少的重要属性集，表示为 Core(P)，如式（2.33）

$$\mathrm{Core}(P) = \bigcap \mathrm{Red}(P) \tag{2.33}$$

实际上，一般产生约简的方法是逐个向核中添加可省略的属性，并进行检查。由于可省略的属性关系集合的幂集的基数是多少，就有多少种添加的方式；所以最好的情况是所有不可省略的属性关系集合本身就是约简，此时的约简是唯一的。所以，计算所有约简与计算一个最佳约简（比如定义为关系最少）都是 NP 难题。

核属性是描述对象的条件属性不可缺少的属性。在条件属性中核以外的属性可以约简，核属性应取为无法约简的属性，核是所有属性的公共部分。

在应用中，一个分类对于另一个分类的关系非常重要。令 P 和 S 为 U 中的等价关系，S 的 P 正域记作 $\mathrm{Pos}_P(S)$，如式（2.34）

$$\mathrm{Pos}_P(S) = \bigcup P\_(S) \tag{2.34}$$

属性约简是数据挖掘研究的一个重要内容。通过对决策表中的条件属性进行简化，使得化简后的决策表具有与化简前的决策表相同的功能，但条件属性数目更少。由此可见，决策表的约简在工程应用中非常实用，同样的决策可以基于更少量的条件。化简后的决策表是一个不完全的决策表，它仅包含那些在决策时所必需的条件属性值，但具有了原始知识系统的所有知识。一般来说，一个属性子集可以有不止一种约简，也就是说，一个知识表达系统的决策表的约简不是唯一的。人们期望能找到具有最少属性的约简，即最小约简。但遗憾的是，Wong S.K.M 和 Ziarko. W 已经证明，找出决策表的最小约简是一个 NP 难题，导致 NP 难题的原因是属性组合的爆炸问题。

核属性的求取可以采用可辨识矩阵方法。可辨识矩阵（Discernibility Matrix）的概念是由波兰数学家 A. Skowron 提出的。在可辨识矩阵中浓缩进了信息表中的所有有关属性区分的信息，可以通过该矩阵非常方便地得到所约简信息表的属性核，从而大大提高了进行数据库知识发现的能力。

可辨识矩阵的定义如下。

定义 2.6 对于知识表达系统 $S=(U, A)$，$A=(C, D)$为属性集，C 是条件属性集合，D 为决策属性集合，$a(x)$为 x 在属性 a 上的取值，可辨识矩阵定义为式（2.35）

$$(C_{ij})=\begin{cases}\{a\in C, a(x_i)\neq a(x_j), D(x_i)\neq D(x_j)\} \\ 0, \qquad\qquad\qquad D(x_i)=D(x_j) \\ -1, \qquad a(x_i)=a(x_j), D(x_i)\neq D(x_j)\end{cases} \tag{2.35}$$

直观地解释，C_{ij} 就是能区分个体 x_i 和 x_j 的所有属性的集合。很显然，可辨识矩阵是一个对称矩阵。通过可辨识矩阵可以容易地求得决策表的核，决策表的核是唯一的，它可以作为最佳属性约简起点。在可辨识矩阵中属性组合数为 1 的属性的集合即为核，其余的有用属性可从属性不为 1 的矩阵元素中获得。

2.3.2 遗传算法

遗传算法（Genetic Algorithm，GA）最早是由美国 Michigan 大学的 Holland 教授于 20 世纪 70 年代提出来的。其核心思想是将生物进化过程中适者生存规则与群体内部染色体的随机信息交换机制相结合，是具有全局搜索能力的进化算法。其因具有简单通用、鲁棒性强、适于并行处理等特点，在机器学习、数据挖掘、组合优化等领域得到了广泛应用。

标准遗传算法（Standard Genetic Algorithm，SGA）的工作流程和结构形式是由 Goldberg 提出的，在实际应用过程中人们往往根据实际问题的需要，对 SGA 进行改变，使 GA 具备求解不同类型的优化问题的能力。

标准遗传算法的工作步骤如下，基本流程如图 2-2 所示。

（1）根据实际问题选择编码策略，把参数集合 X 和域转换为位串结构空间 S；

（2）定义适应值函数 $f(X)$，代表所求问题解；

（3）确定遗传操作算子策略，包括选择群体大小 n，选择、交叉、变异方法，以及确定交叉概率 p_c、变异概率 p_m 等遗传参数；

（4）随机初始化生成群体 P；

（5）计算群体中个体位串解码后的适应值 $f(X)$；

（6）按照遗传策略，运用选择、交叉和变异算子作用于群体，形成下一代群体；

（7）判断群体性能是否满足某一指标，或者已完成预定迭代次数，若满足则输出最优结果，不满足则返回工作步骤（6），或者修改遗传策略再返回工作步骤（6）。

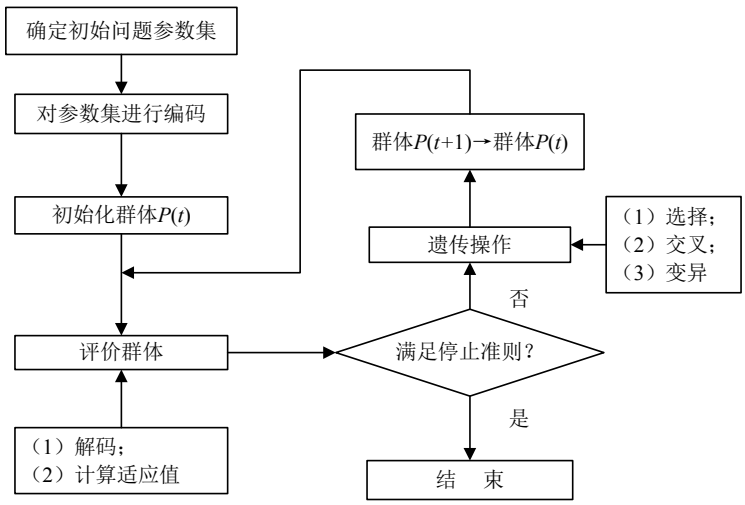

图 2-2　标准遗传算法的基本流程

由上可见遗传算法的实现由五大要素组成：参数编码、初始群体设定、适应函数的设计、遗传操作的设计和控制参数的设定。下面我们分别简述遗传算法的这四种要素。

1. 参数编码

如何将问题的解转换为编码表达的染色体是遗传算法的关键问题。在遗传算法中有二进制编码、字符编码、实数编码、整数编码、树编码等多种编码方式。采用长度为 L 的编码方式，形成规模为 n 的初始群体 P：

$P=\{a_1,a_2,\cdots,a_n\}$，$a_k=(a_{k1},a_{k2},\cdots,a_{kL})$，$a_{kl}\in$ 二进制数、整数、浮点数、字符等，$k=1,2,\cdots,n$；$l=1,2,\cdots,L$。

2. 初始群体设定

初始化群体通常使用的方法是按照采用的编码方式给种群赋予随机值，使每一个染色体代表问题空间中的一个解。

3. 适应函数的设计

适应值是群体中个体生存机会选择的唯一确定性指标。根据实际问题的含义，适应值可以是函数值、销售收入、利润等。为了能够直接将适应函数与群体中的个体优劣度量相联系，在遗传算法中适应值要求为非负，而且越大越好。而对于给定的优化问题 $\mathrm{opt}\,g(x)(x\in[u,v])$，目标函数有正有负，因此有必要通过变换将目标函数转换为适应函数，保证适应函数值为非负，且目标函数的优化方向对应适应值增大的方向。

对于最小化问题，建立如下适应函数 $f(x)$ 和目标函数 $g(x)$ 的映射关系：

$$f(x)=\begin{cases}c_{\max}-g(x) & \text{若}g(x)<c_{\max}\\ 0 & \text{其他}\end{cases} \tag{2.36}$$

其中，$c_{\max}$ 为 $g(x)$ 的最大值估计。

对于最大化问题，一般采用下述方法：

$$f(x)=\begin{cases} g(x)-c_{\min} & \text{若} g(x)>c_{\min} \\ 0 & \text{其他} \end{cases} \tag{2.37}$$

其中，$c_{\min}$ 为 $g(x)$ 的最小值估计。

4. 遗传操作的设计

1）选择算子

遗传算法的基本原理就是达尔文的自然选择原理，选择是遗传算法的推动力。选择是从当前的群体中选择适应值高的个体以形成交配池的过程。目前，主要有轮赌选择、Boltzmann 选择、排序选择、联赛选择等形式。下面简要介绍最常用的轮赌选择法。

对于给定的规模为 n 的群体 $P=\{a_1,a_2,\cdots,a_n\}$，个体 $a_j\in P$ 的适应值为 $f(a_j)$，其选择概率为

$$p_s(a_j)=\frac{f(a_j)}{\sum_{i=1}^{n} f(a_j)},\quad j=1,2,\cdots,n \tag{2.38}$$

则个体 $a_j\in P$ 的累计选择概率为

$$p'_s(a_j)=\sum_{i=1}^{j} p_s(a_i),\quad j=1,2,\cdots,n \tag{2.39}$$

则在进行选择时产生一个[0,1]区间上服从均匀分布的随机数 rand（$\text{rand}\in[0,1]$），当 $p'_s(a_j)<\text{rand}\le p'_s(a_{j+1})$（$j\in\{1,2,\cdots,n-1\}$）时个体 a_{j+1} 被选中进入交配池，当 $\text{rand}\le p'_s(a_1)$ 时 a_1 被选中。

2）交叉算子

遗传算法中交叉操作模仿自然界有性繁殖的基因重组过程，其作用在于将原有的优良基因遗传给下一代个体，并生成包含更复杂基因结构的新个体。目前通常使用的交叉算子包括一点交叉、两点交叉、多点交叉、一致交叉等

形式。下面简要介绍一点交叉算子。

一点交叉算子是标准遗传算法中使用的最基础的一种交叉方式。对于从交配池中随机选择的两个染色体 $a_1 = a_{11}a_{12}\cdots a_{1l_1}a_{1l_2}\cdots a_{1(L-1)}a_{1L}$ 和 $a_2 = a_{21}a_{22}\cdots a_{2l_1}a_{2l_2}\cdots a_{2(L-1)}a_{2L}$，当 $\text{rand} < p_c$ 时（rand 为[0,1]区间上服从均匀分布的随机数，p_c 为交叉概率）随机选择一个交叉位置 $x \in \{1,2,\cdots,L-1\}$，假设 $l_1 \leq x \leq l_2$，对 a_1 和 a_2 中该位置的右侧部分的染色体位串进行交换，产生两个子串为 $a_1' = a_{11}a_{12}\cdots a_{1l_1}a_{2l_2}\cdots a_{2(L-1)}a_{2L}$ 和 $a_2' = a_{21}a_{22}\cdots a_{2l_1}a_{1l_2}\cdots a_{1(L-1)}a_{1L}$。这样就完成 a_1 和 a_2 这两个染色体之间的遗传信息交换，得到了两个新的染色体 a_1' 和 a_2'。在解决实际问题时，可针对特定的问题设计相应的交叉算子。

3）变异算子

变异操作模拟自然界生物体进化中染色体上某位基因发生的突变现象，从而改变染色体的结构和物理性状。目前通常使用的变异算子包括单点变异和多点变异等形式，其中常用的单点变异算子是指在群体中随机选择的一个染色体 $a_1 = a_{11}a_{12}\cdots a_{1l_1}a_{1l_2}\cdots a_{1(L-1)}a_{1L}$，当 $\text{rand} < p_m$ 时（rand 为[0,1]区间上服从均匀分布的随机数，p_m 为突变概率）随机选取一个变异位置 $x \in \{1,2,\cdots,L\}$，假设 $x = l_1$，根据特定的编码方式随机生成一个合法的基因 a_{1l_1}' 代替第 l_1 位基因 a_{1l_1}，得到一个新的染色体 a_1' 为 $a_1' = a_{11}a_{12}\cdots a_{1l_1}'a_{1l_2}\cdots a_{1(L-1)}a_{1L}$。在群体的进化过程中，交叉操作是基于重组和群体更替的主要手段，变异操作仅仅充当辅助作用。因此，交叉操作中的交叉概率 p_c 往往远大于变异操作中的变异概率 p_m。

2.3.3 神经网络

在传统机器学习所进行的计算都是建立在一种算法结构的基础上的，问题的求解都是将相应算法映射为传统计算机所能执行的机器指令序列来完成的。而神经网络的计算则是一种非编程的信息处理方式，在不确定的条件下，只要我们能准确地描述所要求的计算功能，并能给出体现该功能的大量例子，那么神经网络就可以通过这些例子来进行自我调节，直到达到所要求的计算

能力。有时甚至在没有例子可寻时，神经网络也可以根据一些输入信号通过自组织而达到某种计算能力，这种非编程的自适应信息处理方式称为神经计算。神经计算也是传统信息处理方式的一个有力补充。如果把神经网络看成由大量子系统组成的大系统，那么神经计算就是该系统状态的转换，其计算过程可以认为是状态的转换过程。神经计算主要有以下特点。

（1）大规模并行性、群集运算和容错能力。在大规模的神经网络系统中，有许多能同时进行运算的处理单元，信息处理在大量处理单元中并行而又有层次地进行，运算速度快。另外，神经网络系统并不是执行一串单独的指令，神经网络系统中的所有单元都是一起协同解决某一个问题，这是一种集团运算能力，所以信息的处理能力是由整个神经网络系统所决定的。在神经网络中，一个处理单元的失效并不能引起整个系统的失效，只不过是导致整个系统的性能降低，因此，神经网络有较好的容错能力。

（2）信息的分布式处理。与传统计算机不同，神经网络系统中信息的存储和处理是合二为一的，即信息的存储体现在神经元互连的分布上，信息在整个网络中作为一种连接的模式被存储起来，并以大规模并行分布式方式处理。

（3）学习和自组织能力。神经网络可以自动调节其结构来适应学习新的模式，这种变结构体系表现出了很强的对环境的适应性，以及对事物的学习能力。学习和适应体现在状态变化过程中神经网络系统内部结构和连接方式的改变，如 Hebb 学习规则，假设了两个处理单元若同时兴奋则引起它们之间连接强度的变化，这种变化最终会导致在外界输入作用下，网络系统内部有的信息通路增强，有的信息通路变弱甚至阻断，客观上造成网络系统内部结构和状态的变化。神经网络的学习能力，使它在一定程度上类似于大脑的学习功能，这种能力使之有广泛应用的可能性。

（4）多层神经网络系统有强大的解算能力和处理实际问题的能力。它可以处理一些环境信息十分复杂、知识背景不清楚、推理规则不明确的问题。在实际问题中，所提供的模式丰富多变，甚至相互矛盾，而制定决策又无法

可循，对于这些问题，神经网络系统通过学习，可以处理具体实例，给出满意的答案。

神经网络学习对于逼近实数值、离散值或向量值的目标函数提供了一种健壮性很强的方法，对于某些类型的问题，如学习解释复杂的现实世界中的传感器数据，人工神经网络是目前最有效的学习方法之一。神经网络的结构和性能是由神经元、网络结构和学习规则三方面共同决定的，下面简要介绍人工神经网络的主要概念。

1）人工神经元模型

人工神经元可以有多种模型，常用的人工神经元模型可用图 2-3 表示。

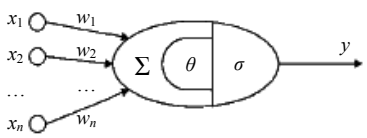

图 2-3　神经元模型

在图 2-3 中，$x_i(i=1,2,\cdots,n)$ 为神经元的输入信号；w_i 为相应的连接权系数，代表输入 x_i 的传递强度的比例；$\sum$表示输入信号加权求和；θ 表示神经元阈值；σ表示神经元的激励函数。该模型的数学表达式为：

$$s=\sum_{i=1}^{n} w_i x_i-\theta \tag{2.40}$$

$$y=\sigma(s) \tag{2.41}$$

神经网络的激励函数是一个重要的概念，激励函数不同就会形成不同的网络，具有不同的性能。常用的激励函数有以下几种。

（1）阈值函数：$\sigma(s)=\begin{cases}0 & s<0\\ 1 & s\geq 0\end{cases}$

（2）线性函数：$\sigma(s)=s$ -

（3）非线性函数：$\sigma(s)=1/(1+\mathrm{e}^{-s})$

2）神经网络的典型结构

单个神经元的功能非常有限，只有通过神经元的互连，构成神经网络，才具有处理复杂非线性映射的能力。神经元之间采取不同的连接方式可得到不同的神经网络。神经网络主要有以下几种连接方式：

（1）前向连接。网络中的神经元分层排列，每个神经元只与前一层神经元连接。最上一层为输出层，隐层可以是一层或多层，前向网络在神经网络中最常见，应用广泛。

（2）反馈连接。唯一与前向连接不同的是输出到输入之间有反馈回路。

（3）层内互连前向网络。与前向网络不同的是同一层神经元之间互相连接。

（4）互连网络。互连网络有局部互连和全互连两种，任何两个神经元都有连接是全互连，有些神经元之间没有连接是局部互连。

3）神经网络的学习方法

神经网络的学习方法就是网络连接权的调整方法。不同的网络结构有不同的学习方法，不同的学习方法有不同的功能。学习方法是多种多样的，最基本的学习规则有以下几种。

（1）Hebb 学习规则。它是最早提出的一种学习规则，可以表述为，若两个神经元同时兴奋，则它们之间的连接加强，如果 v_i 和 v_j 表示神经元 i 和 j 的输出值，w_{ij} 表示两个神经元之间的连接权值，则 Hebb 学习规则可以用式（2.42）表示：

$$\Delta w_{ij}=\alpha v_i v_j \tag{2.42}$$

其中，α 表示学习速率，Δw_{ij} 是权值 w_{ij} 的变化量，Hebb 学习规则是最基本的学习规则，可以说，其他学习规则都是基于 Hebb 学习规则的思想衍生而来的。

（2）δ 学习规则。这是一种有监督的学习规则，采用已知样本作为训练信息对网络进行学习，也称为误差校正规则。设 (X^k, Y^k) 为输入、输出样本对，$X^k=[x_1, x_2, \cdots, x_m]^{\mathrm{T}}$，$Y^k=[y_1, y_2, \cdots, y_n]^{\mathrm{T}}$。把 X^k 作为输入，在网络连接权值的作用下得到的实际输出为 $\overline{Y}^k=[\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_n]^{\mathrm{T}}$，则神经元 i 和 j 之间的权值 w_{ij} 的调整量为式（2.43）和式（2.44）

$$\Delta w_{ij}=\alpha\delta_j v_i \tag{2.43}$$

$$\delta_j=F(y_j-\overline{y}_j) \tag{2.44}$$

其中，α 为学习速率，$y_j-\overline{y}_j$ 为期望输出与实际输出之差。v_i 为第 i 个神经元的输出。函数 $F(\cdot)$ 根据具体问题而定，可以是线性的，也可以是非线性的。

δ 学习规则实际上是一种梯度方法，在许多网络中得到了应用，比如 BP 算法。

（3）相近学习规则。设 w_{ij} 为神经元 i 到 j 的连接权，v_i 为第 i 个神经元的输出，则权值调整量为：$\Delta w_{ij}=\alpha(v_i-w_{ij})$，当 $v_i=w_{ij}$ 时，$\Delta w_{ij}=0$。这种学习是使 w_{ij} 趋近于 v_i 值。下面以神经网络中应用较为广泛的 BP 神经网络为例介绍其模型结构，如图 2-4 所示。

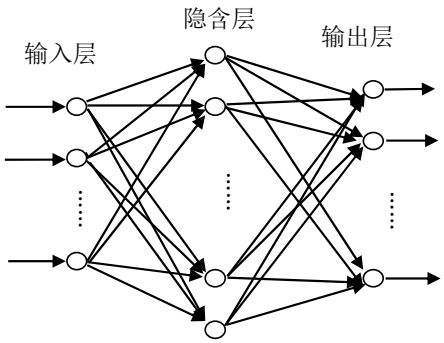

图 2-4　BP 神经网络模型

BP 神经网络又称为误差反向传播神经网络，在 BP 神经网络模型中，引入了中间隐含神经元层。故标准的 BP 模型由三个神经元层次组成，其最下层

称为输入层，中间层称为隐含层，最上层称为输出层。各层次的神经元之间形成全互连连接，各层次内的神经元之间没有连接，中间的隐含层可以有多层。为方便理解，下面以一个隐含层为例进行说明。

对于 BP 模型的输入层神经元，其输出与输入相同，即 $o_i = i_i$。中间隐含层和输出层的神经元的操作特性为：$net_{pj} = \sum w_{ij} o_{pij}$，$o_{pj} = f_j(net_{pj})$。其中，$p$ 表示当前的输入样本，w_{ij} 为从神经元 i 到神经元 j 的连接权值，o_{pij} 为神经元 i 对神经元 j 的当前输出，o_{pj} 为其输出。$f(\cdot)$ 为非线性可微非递减函数，一般取为 S 形函数，即 $f_j(x) = 1/(1+\mathrm{e}^{-x})$。

对多层网络进行训练时，首先要提供一组训练样本，其中的每个样本由输入样本和理想输出对 p 组成。当网络的所有实际输出与其理想输出一致时（实际应用时，只要满足一定的误差），表明训练结束。否则，通过修正权值，使网络的理想输出与实际输出一致。在网络不包含隐含神经元层的情况下，可以直接采用 Delta 学习算法，即对于输入/理想输出对 p，权值按式（2.45）进行调整：

$$\Delta_p w_{ij} = \eta(t_{pj} - o_{pj}) i_{pj} = \eta \delta_{pj} i_{pj} \tag{2.45}$$

其中，t_{pj} 为理想输出，o_{pj} 为实际输出，i_{pj} 为输入向量的第 i 个元素，$\delta_{pj} = t_{pj} - o_{pj}$ 为理想输出与实际输出之间的偏差，η 为学习率。

但是，在 BP 模型中引进了隐含层，因为隐含层的输出误差不能直接计算，故不能直接采用 Delta 学习算法来训练 BP 模型。但将 Delta 学习算法加以推广便可应用于 BP 模型。设网络输出误差为式（2.46）

$$E_p = \frac{1}{2} \sum_j (t_{pj} - o_{pj}) \tag{2.46}$$

设 $E = \sum_p E_p$ 为整个训练样本集中的所有样本产生的输出误差之和。要使 E 梯度下降，按式（2.47）调整权值：

$$\Delta_p w_{ij} = \eta \delta_{pj} o_{pi} = \eta(t_{pj} - o_{pj}) f'(net_{pj}) o_{pi} \tag{2.47}$$

该网络实质上是对任意非线性映射关系的一种逼近。由于采用的是全局逼近的方法，因而 BP 网络具有较好的泛化能力。BP 网络仅通过许多具有简单处理能力的神经元的复合作用使网络具有复杂的非线性映射能力，由于它在理论上的完整性和它能成功地应用于广泛的问题，所以具有重要的意义。概括起来，BP 网络的主要优点是：

（1）只要有足够多的隐层和隐节点，BP 网络可以逼近任意的非线性映射关系。

（2）BP 网络的学习算法属于全局逼近的方法，因而它具有较好的泛化能力。

（3）BP 网络的输入与输出之间的关联信息分布地存储于连接权中，由于连接权的个数很多，个别神经元的损坏只对输入与输出关系有较小的影响，因此，BP 网络显示了较好的容错性。

2.3.4 深度学习

自从 Hinton 教授 2006 年在著名期刊 *Science* 上发表《深度学习》一文以来，深度学习就受到了学术界和工业界研究人员的广泛关注。在数据和计算资源足够的情况下，深度学习在许多领域中体现出占据支配地位的性能表现，如语音识别、视觉对象识别、自然语言处理等领域。传统的方法是人们通过大量的工程技术和专业领域知识积累，进行人工设计特征提取机制，因此在处理未加工数据时表现出的能力有限。另外，大多数的分类等学习模型都是浅层神经网络结构，制约了对复杂分类问题的泛化能力。而深度学习作为一种特征学习方法，把原始数据通过一系列非线性变换得到更高层次、更加抽象的表达，这些都不是通过人工设计的，而是使用一种在线通用的学习过程不断从数据中学习获得的。深度学习主要通过建立类似于人脑的分层神经网络模型结构，对输入数据逐级提取从底层到高层的特征，从而能很好地建立从底层信号到高层语义的映射关系。相比传统的方法，具有多个处理层的深

度学习模型能够学习多层次抽象的数据表示，也受益于计算能力和数据量的增加，从而能够发现大数据中的复杂结构，在语音识别、图像分类等领域取得了较好结果。

2006 年，Hinton 提出的深度信念网络（Deep Belief Network，DBN）是第一批成功应用深度架构训练的非卷积模型之一。深度信念网络的引入开始了深度学习的复兴。在引入深度信念网络之前，深度模型被认为太难以优化，具有凸目标函数的核机器占据了研究前景。深度信念网络在 MNIST 数据集上的表现超过内核化支持向量机，以此证明深度架构是能够成功的。尽管现在与其他无监督或生成学习算法相比，深度信念网络大多已经失去了青睐并很少被使用，但它们在深度学习历史中的重要作用仍应该得到承认。深度信念网络是具有若干隐变量层的生成模型。隐变量通常是二值的，而可见单元可以是二值或实数，没有层内连接。通常，每层中的每个单元连接到每个相邻层中的每个单元，尽管构造更稀疏连接的 DBN 是可能的。顶部两层之间的连接是无向的，所有其他层之间的连接是有向的，箭头指向接近数据的层。

深度学习中最具有代表性的是深度卷积神经网络，这是一种前馈式神经网络，更易于训练，并且比全连接的神经网络泛化性能更优。卷积神经网络因其具有局部连接、权值共享、泛化和多网络层四个特征而非常适用于处理多维数组数据。自 20 世纪 90 年代以来，卷积神经网络被成功应用于检测、分割、识别、语音、图像的各个领域。比如最早是用时延神经网络进行语音识别及文档阅读，其由一个卷积神经网络和一个关于语言约束的概率模型组成，这个系统后来被应用在美国超过 10% 的支票阅读上；再如微软开发的基于卷积神经网络的字符识别系统及手写体识别系统；近年来，卷积神经网络的一个重大成功应用是人脸识别；Mobileye 和 NVIDIA 公司也正试图把基于卷积神经网络的模型应用于汽车的视觉辅助驾驶系统中。如今，卷积神经网络用于几乎全部的识别和检测任务。许多公司，如 NVIDIA、Mobileye、Intel、Qualcomm 及 Samsung 都积极开发卷积神经网络芯片，以便在智能手机、相机、机器人及自动驾驶汽车中实现实时视觉系统。

虽然深度学习在理论和应用上取得了一定的进展，但仍有一些问题亟待解决。第一，深度学习模型都是非凸函数，理论研究存在困难。第二，深度学习模型训练耗时，需要设计新的算法进行训练，或者采用并行计算平台来加快训练速度。如何克服深度学习的局限性来提高模型的性能是未来一段时间值得研究的问题。深度学习的动机源于脑科学，随着认知神经学的发展，科学家发现了许多与人脑动态学习相关的特性，如神经元自组织特性、神经元之间的信息交互特性、人类认知的进化特性等，而这些特性将为深度学习模型的构建提供更多的启发，促进深度学习的进一步发展。是否能够利用认知科学的一些新进展构造更好的深度学习模型，也是值得我们探讨的问题。

各大 IT 公司也非常关注深度学习的应用前景，纷纷成立相关的实验室。2012 年，华为成立诺亚方舟实验室，运用以深度学习为代表的人工智能技术对移动信息大数据进行挖掘，寻找有价值的规律。2013 年，百度成立深度学习研究院，研究如何运用深度学习技术对大数据进行智能处理，提高分类和预测等任务的准确性。国际 IT 巨头 Google、Facebook 等也成立了新的人工智能实验室，投入巨资对以深度学习为代表的人工智能技术进行研究。Hinton 等多位深度学习的知名教授也纷纷加入工业界，以深度学习为支撑技术的产业雏形正在逐步形成。

2.3.5 支持向量机

如前所述，统计学在解决机器学习的问题中起着基础性的作用。但是传统的统计学方法都是在立足于样本数目趋于无穷的前提下开展工作的。而在多数实际情况中，样本数目通常是有限的，这样很多方法都难以取得理想的效果。Vladimir N. Vapnik 等人从 20 世纪 60 年代起，就开始研究有限样本的机器学习问题，但是处在其他方法飞速发展的时期，这些研究没有得到充分重视。直到 20 世纪 90 年代，Vapnik 在统计学习理论的基础上，提出了支持向量机（Support Vector Machines，SVM）的概念。支持向量机在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并且具备完备的理论基础

和出色的学习能力，已成为机器学习领域的研究新热点，并在很多领域得到了成功的应用。

支持向量机是与相关的学习算法有关的监督学习模型，可以分析数据、识别模式，用于分类和回归分析。支持向量机方法是建立在统计学习理论的VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（对特定训练样本的学习精度）和学习能力（无错误地识别任意样本的能力）之间寻求最佳折中，以获得最好的推广能力。给定一组训练样本，每个标记属于两类，一个 SVM 训练算法建立一个模型，分配新的实例为一类或其他类，使其成为非概率二元线性分类。支持向量机中的一大亮点是在传统的最优化问题中提出了对偶理论，主要有最大最小对偶及拉格朗日对偶。除了进行线性分类，支持向量机可以使用核技巧，它们的输入隐含映射成高维特征空间中有效地进行非线性分类。低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间。但这个办法带来的困难就是计算复杂度的增加，而核函数正好巧妙地解决了这个问题。也就是说，只要选用适当的核函数，就可以得到高维空间的分类函数。在 SVM 理论中，采用不同的核函数将导致不同的 SVM 算法。

支持向量机的理论最初来自数据分类问题的处理，它巧妙地解决了最终所获得的分割平面位于两个类别的中心的问题。其机理可简单描述为：寻找一个满足分类要求的最优分类超平面，使得在保证分类精度的同时，能够使超平面两侧的空白区域最大。

1. 支持向量机分类

假设有训练样本集 $\{(\boldsymbol{x}_i, y_i), i=1,2,\cdots,l\}$，期望输出 $y_i \in \{+1,-1\}$，其中+1和−1 分别代表两类的类别标识。则若 $\boldsymbol{x}_i \in \boldsymbol{R}^N$ 属于第一类，对应的输出标记为正（ $y_i=+1$ ）；若 $\boldsymbol{x}_i \in \boldsymbol{R}^N$ 属于第二类，对应的输出标记为负（ $y_i=-1$ ）。学习的目标就是构造一个决策函数，将测试数据尽可能正确地分类。

1）线性支持向量机

当训练样本线性可分时，用星号和圆圈分别表示两类训练样本，如图 2-5

所示（二维情形）。分类线 H 把两类样本没有错误地分开，H_1 和 H_2 分别为通过各类样本中离分类线最近的点且平行于分类线的直线，则 H_1 和 H_2 之间的距离即为两类的分类间隔（Margin）。所谓最优分类线不但能将两类无错误地分开，而且使两类的分类间隔最大。前者保证经验风险最小，后者实际上是为了使置信范围最小，从而实现了实际风险最小的结构风险最小化原则。推广到高维空间，最优分类线就成为最优超平面（Optimal Hyperplane）。

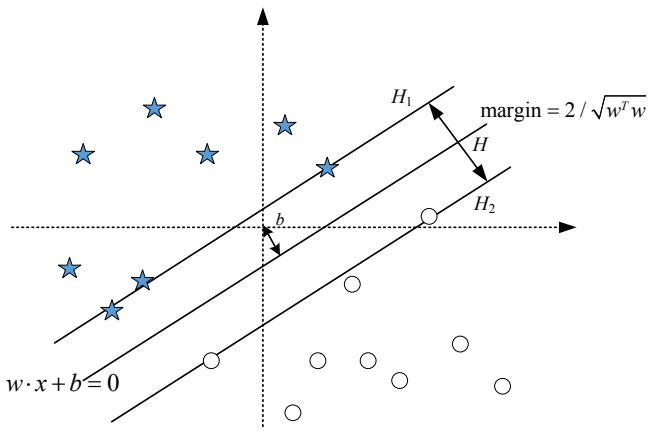

图 2-5　最优超平面示意图

假设存在分类超平面 $w\cdot x+b=0$。为使分类面对所有样本正确分类且具备分类间隔，满足式（2.48）

$$\left.\begin{array}{ll} x_i\cdot w+b\geq +1 & for \quad y_i=+1 \\ x_i\cdot w+b\leq -1 & for \quad y_i=-1 \end{array}\right\} \Leftrightarrow y_i\cdot(x_i\cdot w+b)-1\geq 0 \qquad (2.48)$$

可以计算出分类间隔为式（2.49）

$$\min_{\{x_i|y_i=+1\}}\frac{w\cdot x_i+b}{\|w\|}-\min_{\{x_i|y_i=-1\}}\frac{w\cdot x_i+b}{\|w\|}=\frac{2}{\|w\|} \qquad (2.49)$$

现在的目标就是在满足约束式（2.49）的条件下最大化分类间隔 $2/\|w\|$，即要求最小化 $\|w\|$。则求解最优分类超平面问题就可表示成约束优化问题，即在条件（2.49）的约束下，最小化函数（2.50）。

$$\Phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w\cdot w) \tag{2.50}$$

求解此约束最优化问题，引入 Lagrange 函数

$$L=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{l}\alpha_i \tag{2.51}$$

其中，$\alpha_i>0$ 为 Lagrange 系数。现在的问题就是关于 w 和 b 求解 L 的最小值。把式（2.51）分别对 w 和 b 求偏微分并令其等于 0，就可以把上述问题转化为一个较简单的对偶问题：求 L 的最大值，其约束条件为 L 关于 w 和 b 的梯度均为 0 及 $\alpha_i\geq 0$，即在约束条件（2.52）下对 α_i 求解函数（2.53）的最大值。

$$\sum_{i=1}^{l}y_i\alpha_i=0 \quad \alpha_i\geq 0,\ i=1,\cdots,l \tag{2.52}$$

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \tag{2.53}$$

若 α_i^* 为最优解，则 $w^*=\sum_{i=1}^{l}\alpha_i^* y_i x_i$，即最优超平面的权系数向量是样本向量的线性组合。这是一个不等式约束下的二次函数极值问题（Quadratic Programming，QP）。根据 Karush-Kuhn-Tucker（KKT）条件，这个优化问题的解必须满足：

$$\alpha_i\{y_i[(w\cdot x_i)+b]-1\}=0, \quad i=1,\cdots,l \tag{2.54}$$

因此，多数样本对应的 α_i 值将为 0，把 $\alpha_i\neq 0$ 对应于使式（2.54）中等号成立的样本称为支持向量（Support Vectors，SVs）。对学习过程而言，支持向量是训练集中的关键元素，它们离决策边界最近。如果去掉其他所有训练样本（或移动位置，但不穿越 H_1 或 H_2）再重新进行训练，得到的分类面是相同的。求解上述问题后得到的最优分类函数为式（2.55）

$$f(x)=\mathrm{sgn}\left\{\sum_{i=1}^{l}y_i\alpha_i^*(x_i\cdot x)+b^*\right\} \tag{2.55}$$

由于非支持向量对应的 α_i 均为 0，因此式（2.55）的求和实际上只对支持

向量进行。b^* 是分类的阈值，可以由任意一个支持向量用式（2.55）求得或通过两类中任意一对支持向量取中值求得。当样本集为线性不可分时，需引入非负松弛变量 ξ_i，$i=1,2,\cdots,l$，分类超平面的最优问题为式（2.56）

$$\begin{aligned}\min_{w,b,\xi_i}\quad & \frac{1}{2}w^T w+C\sum_{i=1}^{l}\xi_i \\ s.t.\quad & y_i(w^T\cdot x_i+b)\geq 1-\xi_i \\ & \xi_i\geq 0,\quad i=1,2,\cdots,l\end{aligned} \tag{2.56}$$

由式（2.56）有，当分类出现错误时，ξ_i 大于 0，因此 $\sum_{i=1}^{l}\xi_i$ 是训练集中错分样本数的上界。这就需要在目标函数中为分类误差分配一个额外的代价函数，即引入错误惩罚分量。其中，C 为惩罚参数，它控制对错分样本的惩罚程度，C 越大表示对错误分类的惩罚越大。采用拉格朗日乘子法求解这个具有线性约束的二次规划问题，即式（2.57）

$$L=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i[y_i(w\cdot x_i+b)-1+\xi_i]-\sum_{i=1}^{l}\beta_i\xi_i \tag{2.57}$$

其中，α_i,β_i 为拉格朗日乘子 $0\leq\alpha_i,0\leq\beta_i$，由此得到式（2.58）～式（2.60）

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{l}\alpha_i y_i x_i=0 \tag{2.58}$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{l}\alpha_i y_i=0 \tag{2.59}$$

$$\frac{\partial L}{\partial \xi_i}=C-\alpha_i-\beta_i \tag{2.60}$$

将式（2.58）～式（2.60）代入式（2.57），得到对偶最优化问题：

$$\begin{aligned}\max_{\alpha}\quad & \sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j x_i\cdot x_j \\ s.t.\quad & 0\leq\alpha_i\leq C,\quad i=1,\cdots,l \\ & y^T\alpha=0\end{aligned} \tag{2.61}$$

最优化求解得到的 α_i 可能是：① $\alpha_i=0$，② $0<\alpha_i<C$，③ $\alpha_i=C$；后两

者所对应的样本 $\boldsymbol{x}_i$ 为支持向量。只有支持向量对最优超平面、决策函数有贡献，故支持向量由此得名，对应的学习方法称为支持向量机。在支持向量中，②所对应的样本 $\boldsymbol{x}_i$ 称为标准支持向量（Normal Support Vector），③所对应的样本 $\boldsymbol{x}_i$ 称为边界支持向量（Boundary Support Vector）。根据 KKT 条件，拉格朗日乘子与约束的积在最优点为 0，即

$$\begin{cases}\alpha_i[y_i(w\cdot x_i+b)-1+\xi_i]=0\\ \beta_i\xi_i=0\end{cases} \tag{2.62}$$

对于标准支持向量（$0<\alpha_i<C$），由式（2.60）得到 $\beta_i>0$，则由式（2.62）得到 $\xi_i=0$，因此，对于任一标准支持向量，满足

$$y_i(w\cdot x_i+b)=1 \tag{2.63}$$

从而计算参数 b 为

$$b=y_i-w\cdot x_i=y_i-\sum_{x_j\in J}\alpha_j y_j x_j\cdot x_i,\ x_i\in \mathrm{JN} \tag{2.64}$$

为了计算可靠，对所有标准支持向量分别计算 b 的值，然后求平均，即

$$b=\frac{1}{N_{\mathrm{NSV}}}\sum_{X_i\in \mathrm{JN}}(y_i-\sum_{x_j\in \mathrm{J}}\alpha_j y_j(x_j,x_i)) \tag{2.65}$$

其中，N_{NSV} 为标准支持向量数，JN 为标准支持向量的集合，J 为支持向量的集合。

2）非线性支持向量机

在输入空间中构造最优分类面的方法类似于经典的感知器方法。该方法仅当样本集为线性可分时才能使经验风险等于零。由于许多问题都不是线性可分的，譬如异或问题（XOR），因此利用这种方法求得的解常常由于经验风险过大而失去意义。解决这个问题的一个方法是利用多层感知器，其实质是将近似函数集由简单线性指示函数扩展成由许多线性指示函数叠加成的一个更为复杂的近似函数集，再用 S 形函数来近似指示函数中的单位阶跃函数（或符号函数），从而得到使经验风险最小化的一种容易操作的算法。但是，这种

方法存在容易陷入局部极小点，网络结构设计依赖于先验知识及泛化能力较差等问题。另外一种方法是将输入向量映射到一个高维的特征空间，并在该特征空间中构造最优分类面，即支持向量机方法，它能够避免在多层前向网络中无法克服的一些缺点。理论证明，当选用合适的映射函数时，大多数输入空间线性不可分的问题在特征空间可以转化为线性可分问题来解决。但是，在低维输入空间向高维特征空间映射过程中，由于空间维数急速增长，就使得在大多数情况下难以在特征空间直接计算最佳分类平面。支持向量机通过定义核函数（Kernel Function），巧妙地将这一问题转化到输入空间进行计算。具体做法是通过某种非线性映射，将输入向量 $\boldsymbol{x}$ 映射到一个高维的特征空间，在这个高维的特征空间中构造最优分类超平面，如图 2-6 所示。

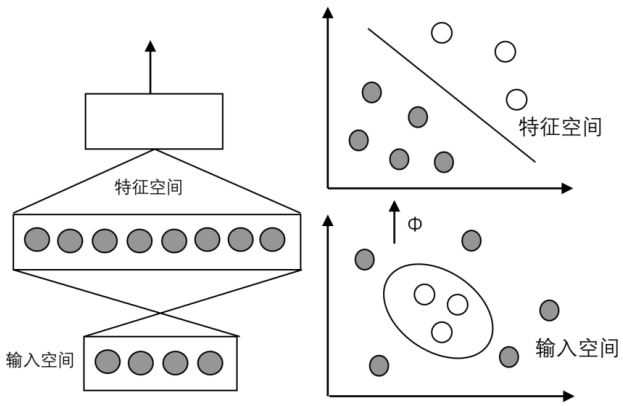

图 2-6　输入空间与高维特征空间之间的映射关系

在非线性情形下，最优分类超平面为 $w\cdot\boldsymbol{\Phi}(x)+b=0$；决策函数为 $f(x)=\mathrm{sgn}[w\cdot\boldsymbol{\Phi}(x)+b]$。则最优分类超平面问题描述为式（2.66）

$$\begin{aligned}\min_{w,b,\xi_i}\quad & \frac{1}{2}w^Tw+C\sum_{i=1}^{l}\xi_i \\ s.t.\quad & y_i(w^T\cdot\boldsymbol{\Phi}(x_i)+b)\geq 1-\xi_i \\ & \xi_i\geq 0,\quad i=1,2,\cdots,l\end{aligned}\tag{2.66}$$

用同样的方法可以得到对偶最优化问题为式（2.67）

$$\max_{\alpha} \quad \left\{ \begin{aligned} L &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) \\ &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \end{aligned} \right\}$$

$$s.t. \quad 0 \le \alpha_i \le C, \quad i = 1, \cdots, l \qquad (2.67)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$

其中，$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ 称为核函数。参数 b 可由式（2.68）计算。

$$b = \frac{1}{N_{\mathrm{NSV}}} \sum_{x_i \in \mathrm{JN}} (y_i - \sum_{x_j \in \mathrm{J}} \alpha_j y_j K(x_j \cdot x_i)) \qquad (2.68)$$

其中，N_{NSV} 为标准支持向量数，JN 为标准支持向量的集合，J 为支持向量的集合，其决策函数为式（2.69）

$$f(x) = \mathrm{sgn}\left\{ \sum_{i=1}^{l} y_i \alpha_i K(x_i \cdot x) + b \right\} \qquad (2.69)$$

用支持向量机求得的决策函数在形式上类似于一个神经网络，其输出是若干中间层节点的线性组合，而每一个中间层节点对应于输入样本与一个支持向量的内积，因此也被称为支持向量机网络，如图 2-7 所示。

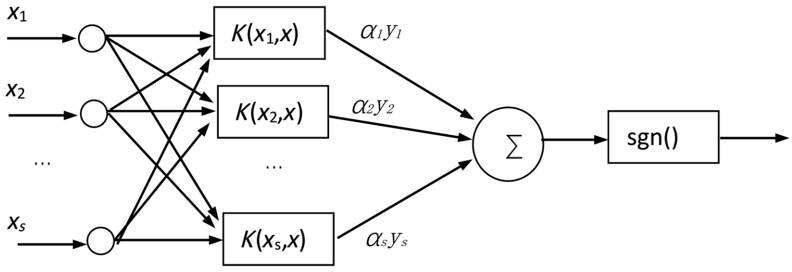

图 2-7　支持向量机网络示意图

2. 核函数

支持向量机可以采用不同的核函数来实现不同类型的学习机。目前常用的核函数主要有多项式核函数、径向基函数、多层感知器和动态核函数等。

1）多项式核函数

多项式核函数： $K(x,x_i)=[(x,x_i)+1]^d$ ，所得到的是 d 阶多项式分类器：

$$f(x)=\mathrm{sgn}(\sum_{\mathrm{SVs}} y_i a_i[x_i\cdot x+1]^d+b) \tag{2.70}$$

2）径向基函数

经典的径向基函数使用下面的判定规则：

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{l} a_i k_\gamma(|x-x_i|)+b) \tag{2.71}$$

其中， $k_\gamma(|x-x_i|)$ 取决于两个向量之间的距离 $|x-x_i|$ 。对于任何 γ 值，函数 $k_\gamma(|x-x_i|)$ 是一个非负的单调函数，当训练样本数趋向无穷大时它趋向零，最通用的判定规则是采用高斯函数：

$$k_\gamma(|x-x_i|)=\exp\{-\frac{|x-x_i|^2}{\sigma^2}\} \tag{2.72}$$

这里每个基函数的中心点对应一个支持向量，中心点本身和输出权值都是由 SVM 学习算法自动确定的。

3）多层感知器

支持向量机采用 Sigmoid 函数作为内积，这时就实现了包含一个隐层的感知器。隐层节点数目由算法自动确定。满足 Mercer 条件的 Sigmoid 核函数为式（2.73）

$$K(x_i,x_j)=\tanh(\gamma x_i^T x_j-\theta) \tag{2.73}$$

4）动态核函数

1999 年，Amari 和 Wu 通过对核函数的黎曼几何分析，提出了利用实验数据逐步修正原有的核函数，使之更好适应实际问题，设特征映射 $U=\Phi(x)$ ，则

$$\begin{aligned} \mathrm{d}U &= \sum_i \frac{\partial}{\partial x_i}\boldsymbol{\Phi}(x)dx_i \\ \left\|\mathrm{d}U\right\|^2 &= \sum_{i,j} g_{ij}(x)dx_i dx_j \end{aligned} \tag{2.74}$$

这里 $g_{ij}(x)=(\frac{\partial}{\partial x_i}\boldsymbol{\Phi}(x))\cdot(\frac{\partial}{\partial x_i}\boldsymbol{\Phi}(x))$ 称非负定阵 $(g_{ij}(x))$ 为 R^n 上的黎曼张量，$d_s^2=\sum_{i,j} g_{ij}(x)dx_i dx_j$ 为 R^n 上的黎曼距离。赋予黎曼距离 R^n 成为黎曼空间，如式（2.75）

$$d\upsilon=\sqrt{g(x)}dx_1\cdots dx_n \tag{2.75}$$

其中，$g(x)=\det(g_{ij}(x))$。直观地说，$g(x)$反映了特征空间中点 $\boldsymbol{\Phi}(x)$ 附近局部区域被放大的程度。因此，也称 $g(x)$为放大因子。

因为 $k(x,z)=(\boldsymbol{\Phi}(x)\cdot\boldsymbol{\Phi}(z))$ 可以验证 $g_{ij}(x)=\frac{\partial}{\partial x_i\partial z_j}k(x,z)\big|_{z=x}$，特别对高斯函数 $k(x,z)=\exp\{\frac{\left|x-z\right|^2}{2\sigma^2}\}, g_{ij}(x)=\frac{1}{\sigma^2}\delta_{ij}$。为了有效地将两类不同的模式区分开，希望尽量拉大它们之间的距离，即尽量放大分离面附近的局部区域。可以用修正核函数的办法达到此目的。设 $c(x)$是正的可微实函数，$k(x, z)$是高斯核，则式（2.76）也是核函数，且式（2.77）成立。

$$\tilde{k}(x,z)=c(x)k(x,z)c(z) \tag{2.76}$$

$$\tilde{g}_{ij}(x)=c_i(x)c_j(x)+c^2(x)g_{ij} \tag{2.77}$$

这里，$c_i(x)=\frac{\partial}{\partial x_i}c(x)$。Amari 和 Wu 设 $c(x)$有如下形式

$$c(x)=\sum_{x_i\in \mathrm{SV}} h_i \mathrm{e}^{\frac{\left\|x-x_i\right\|^2}{2\tau^2}} \tag{2.78}$$

这里，$\tau>0$ 是参数，h_i 是权系数。经计算可知

$$\tau\approx\frac{\sigma}{\sqrt{n}} \tag{2.79}$$

这样，新的训练过程由两步组成：首先用某个核 k（高斯核）进行训练，然后按照式（2.76）、式（2.78）和式（2.79）得到修正的核 $\tilde{k}$；再用 $\tilde{k}$ 进行训练。这种改进的训练方法不仅可以明显地降低错误率，而且还可减少支持向量的个数，从而提高计算效率。

2.3.6 强化学习

机器学习任务可以划分为监督学习、无监督学习和弱监督学习。监督学习面临的数据样本有完整的标记，即每一项观察都有与之对应的决策，机器从这样的样本中可以直接学习到从观察到决策的映射。无监督学习面临的数据样本完全没有标记，机器需要从数据中发现内部的结构信息。弱监督学习的目的与监督学习一致，然而其获得的样本并没有完整的标记。因标记缺失的形式和处理方式的不同，又可以分为半监督学习、主动学习、多标记学习和强化学习。在半监督学习中，只有少量的样本具有标记；在主动学习中，机器可以询问真实的标记，但需要考虑询问的代价；在多标记学习中，一个样本对应一组标记，因此需要处理巨大的标记组合空间问题；在强化学习中，机器需要探索环境来获得样本，并且学习的目的是长期的奖赏，因此样本的标记是延迟的。

强化学习研究学习器在与环境的交互过程中，如何学习到一种行为策略，并最大化得到的累积奖赏。与前面提到的其他学习问题的不同在于，强化学习处在一个对学习器的行为进行执行和评判的环境中：环境将执行学习器的输出，发生变化，并且反馈给学习器一个奖赏值；同时学习器的目标并不在于最大化立即获得的奖赏，而是最大化长期累积的奖赏。与监督学习相比，强化学习生成的模型与监督学习生成的模型并无本质区别，都是以对象的描述为输入、以决策值为输出的，但两者的学习过程有很大不同，如图 2-8 所示。

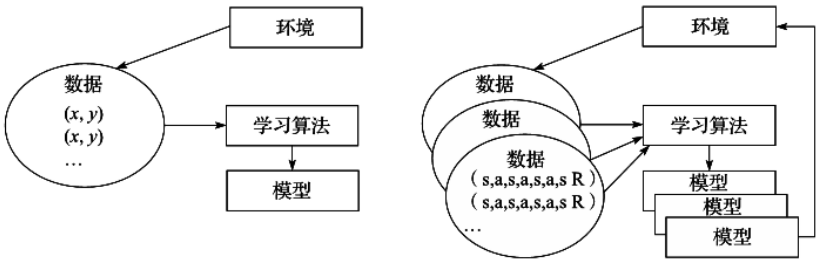

图 2-8　监督学习与强化学习对比图

强化学习的目标是最大化累积奖赏，这一点与马可夫决策过程（MDP）的目标一致，因此强化学习也常常用 MDP 来建模。一个 MDP 定义为四元组 <*S*, *A*, *T*, *R*>，其中，*S* 表示环境状态的集合。*A* 为动作集合，即学习器的输出值域。*T* 为转移函数，定义了环境跟随动作而发生的转移。*R* 为奖赏函数，定义了动作获得的奖赏。MDP 寻找最优动作策略以最大化累计奖赏。当 MDP 的四元组全部给出且 *S* 和 *A* 为有限集合时，求解最优策略的问题即转变为求解每一个状态上最优动作这一优化问题，而该优化问题通常可以通过动态规划来求解：在最终时刻，只需要考虑立即获得的奖赏，即可得知每个状态最优动作获得的奖赏。因为 MDP 四元组全部已知，实际上并不需要与环境交互，也没有“学习”的意思，动态规划就可以保证求解最优策略。

强化学习要面临的难题通常是 MDP 四元组并非全部已知，即“无模型”。最常见的情况是转移函数 *T* 未知及奖赏函数 *R* 未知，这时就需要通过在环境中执行动作、观察环境状态的改变和环境给出的奖赏值来求出 *T* 和 *R*。我们可以把强化学习方法分为基于值函数估计的方法和直接最大化累计奖赏的直接策略搜索方法。基于值函数估计的方法试图在与环境交互的过程中估计出每一种状态上每一个动作对应的累积奖赏，从而得出最佳策略。这一类方法的代表有时序查分学习方法 SARSA 和 Q-Learning。基于值函数估计的方法由于其目标并不是直接求得策略，而是通过值函数的学习来得到策略，即最终的策略是选择值函数大的动作。因此在较复杂的任务上会出现“策略退化”的现象，即虽然值函数估计较准确，但得到的策略却不好。直接最大化累计奖赏的直接策略搜索方法则不依赖于对状态上累积奖赏的估计，而直接优化策略

获得的累积奖赏。这一类方法的代表有使用策略梯度方法优化参数化策略的REINFORCE 方法，以及使用演化算法等全局优化算法来搜索策略的 NEAT+Q 方法等。由于强化学习框架的广泛适用性，被机器学习领域著名学者、国际机器学习学会创始主席 T. G. Dietterich 教授列为机器学习的四大研究方向之一。强化学习在实际问题上的广泛使用还面临诸多挑战，主要包括特征表示、搜索空间、泛化能力等方面的问题。

在经典强化学习的研究中，状态和动作空间均为有限集合，每一个状态和动作被分别处理。然而，一方面许多应用问题具有连续的状态和动作空间，如机械臂的控制。另一方面即使对于有限状态空间，状态之间也并非没有联系，如在棋盘上走棋有位置关系。因此如何将状态赋予合适的特质表示将极大地影响强化学习的性能，这一方面的工作包括使用更好的特征编码方式等。得益于深度学习技术的发展，特征可以更有效地从数据中学习，Google DeepMind 的研究者在 Nature 上发表了基于深度学习和 Q-Learning 的强化学习方法 Deep Q-Network，在 Atari 2600 游戏机上的多个游戏取得“人类玩家水平”的成绩。一方面可以看到特征的改进可以提高强化学习的性能，另一方面也观察到 Deep Q-Network 在考验反应能力的游戏上表现良好，而对于需要逻辑知识的游戏还远不及人类玩家。

由于强化学习关于累积奖赏的优化目标涉及多步决策，这使得策略的搜索空间巨大、累积奖赏目标极其复杂，优化非常困难。一方面需要研究更加有效的优化方法，例如使用 Cross-Entropy 等方法进行优化。另一方面通过引入模仿学习，可以极大缓解这一问题。在模仿学习中，存在能做到接近最优策略的“教师”，并且由“教师”进行示范，提供一批演示样本，这些样本可用于直接指导每一步的动作，因此可以借助监督学习帮助强化学习。同时模仿学习的另一作用是从演示样本中学习奖赏函数，称为逆强化学习，从而可以在应用问题中免去对奖赏函数的定义，例如 IJCAI Computers and Thought Award 得主斯坦福大学 Andrew Ng 教授使用逆强化学习进行运动轨迹规划。

2.3.7 度量学习

度量是计量的准则。脱离度量，收集的数据、分析的结果也就丧失了物理意义和现实指导。而距离的度量对众多机器学习方法的性能都起到了决定性作用：例如在分类方法中，K 近邻分类器使用了高斯核的核方法；在聚类方法中，K 均值聚类、谱聚类方法都与距离度量密切相关。

一般来说，对于任意样本 x, y, z 而言，距离度量函数需要满足自反（任意样本到自身的距离为 0）、对称（x 到 y 的距离等于 y 到 x 的距离）、非负（任意样本对之间的距离大于等于 0）及直递（三个样本之间的距离满足三角不等式）等性质。为了适应不同的具体应用场景，人们提出了诸如闵可夫斯基距离（欧几里得距离、曼哈顿距离、切比雪夫距离均为其特例）、马氏距离、海明距离等距离度量函数，并针对某些特定问题提出了一些衍生距离度量，例如，动态时间规整距离 DTW、推土机距离 EMD 等。

随着机器学习应用面的日益拓展，通过人工设计或定义特定的衍生距离度量函数未必吻合具体的问题，因此，通过学习获得问题相关的度量成为研究主题，美国卡内基梅隆大学机器学习系的邢波教授于 2003 年提出了距离度量学习。在随后的 10 余年里，各类距离度量学习方法不断被提出，并在诸如社交网络连接预测、强化学习的状态连接学习、信息检索与推荐、身份验证，甚至医疗效果评估等方面都获得了广泛应用。Weinberger 和 Saul 提出了一种利用邻域内三元关系进行度量学习的方法 LMNN。在 LMNN 中所有的约束关系都限于某个样本 $\boldsymbol{X}_i$ 的局部邻域，故此类方法也被称为局部距离度量学习方法。自 LMNN 提出后，局部距离度量学习方法得到众多研究者的青睐，多种扩展方案被分别提出，例如，能处理多任务的 mt-LMNN，可在不同集簇中学习多个度量的 mm-LMNN 等；在局部距离度量学习方面，Huang 等人提出了能够处理一定噪声和错误的鲁棒度量学习方法 RML；Chechik 等人借鉴 LMNN 的思想，直接对内积形式的相似度利用局部约束加以学习，并将相关算法运用于大规模图像检索，取得了很好的效果；利用与局部距离度量学习类似的

思想，研究者不仅针对马氏距离度量矩阵进行学习，甚至对前述的 EMD 距离进行了学习，例如χ^2-LMNN 就针对与直方图类特征对应的 EMD 距离进行学习；在局部信息和性质的利用方面，有些研究者甚至为每个样本都学习了合适的距离度量。

随着数据收集手段的提升，大数据时代已经开启。在大数据背景下，距离度量学习和降维之间的关系得到了研究者的关注。事实上，早在 2003 年 Goldberger 等人提出的 NCA 一文中就明确指出距离度量学习和降维之间的关系，Bellet 等人在 2005 年就明确指出：几乎每种线性距离度量学习方法都对应着一类降维策略。在意识到距离度量学习和降维的关系之后，研究者们提出了很多能够直接进行降维或者利用降维能力简化计算的距离度量学习方法。例如，Shi 等人提出在有限基上进行距离度量学习，仅需学习一组基的线性组合系数即可，从而消减了距离度量学习的计算量。值得注意的是，除了降维之外，距离度量学习研究者们也设计出了独到的高维数据处理方法，如 Qian 等人于 2014 年提出了一种基于随机投影的距离度量学习方法，通过随机投影降低数据维度，并通过对偶空间的基重构获得原空间的距离度量；Schultz、Joachims 和 Gao 等人都提出了学习一个对角距离度量矩阵代替学习完全的度量矩阵的替代方案等。此外，稀疏学习技术在距离度量学习研究中也获得了运用。

距离度量学习同样在计算机视觉、信息检索和生物信息学相关领域受到关注。在计算机视觉领域，距离度量学习除了被用于图像分类、物体识别、视觉追踪，还在一些计算视觉的本质问题上被利用，如图像表示方面等；信息检索的结果对距离和相似度的定义十分敏感，因此这方面的工作也相对丰富；对 DNA 和蛋白质分子的结构分析涉及诸如编辑距离和 DTW 方面的研究，度量学习在这些特殊距离度量处理方面也有对应的研究工作。

2.3.8 多核学习

核方法是机器学习中一类强有力的统计学习技术，被广泛应用于分类、

回归、聚类等诸多领域。核选择是核方法的关键内容，因其是提高核方法泛化性能的重要一环。多核学习（Multiple Kernel Learning，MKL）通过利用多个基本核的组合代替单核，将核选择问题转化为对组合系数的选择，有效地改进了核方法。其最早应用于生物信息学领域，例如在蛋白质功能预测与定位、蛋白质分子间的交互预测等问题中，由于来自异构源的数据具有不同的特性，可以通过多个基本核矩阵的线性组合实现异构数据源的融合，基于此训练分类器取得了很好的性能。

构造多核模型，最基本的方法就是考虑多个基本核函数的凸组合：

$$K(\boldsymbol{x},\bullet)=\sum_{i=1}^{M}\beta_i K_i(\boldsymbol{x},\bullet)\text{，}\ \beta_i\geq 0\text{，}\ \sum_{i=1}^{M}\beta_i=1$$

其中 $K_i(\boldsymbol{x},\bullet)$ 是基本核函数，M 是基本核的总个数，β_i 是组合系数，条件 $\beta_i\geq 0$ 可以确保由此产生的 Gram 矩阵是半正定的。因此，在 MKL 框架下，样本在特征空间中的表示问题转化为基本核与组合系数的选择问题。在这个由多个特征空间构建的组合空间中，利用了各个基本核的特征映射能力，通过将异构数据的不同特征分量利用对应的核函数进行映射，使数据在新的特征空间中得到更好的表达，能显著提高分类性能。MKL 的本质问题就是，如何得到这个组合的特征空间，即如何通过学习得到组合系数。近年来，研究者们提出了一系列 MKL 算法，主要侧重于算法的优化求解和性能提高两个方面。

近年来，多核学习已被成功应用于机器学习的许多领域，如多示例学习、半监督学习、增量学习等，并在生物特征识别、无人机、信息检索等领域得到了广泛应用。例如，在虹膜图像检测方面，研究者利用多核学习，融合了频谱能量分布、奇异倒谱直方图等多个特征，有效地提高了检测效果。在医学诊断方面，多核学习可克服采用单一核函数所导致的多个检测指标很难同时兼顾的问题，充分发挥了多个核函数不同的刻画能力，提高了检测算法的泛化能力和鲁棒性，较好地提升了诊断的准确度和敏感度，为医学诊断提供了更准确的信息。在无人机故障诊断方面，多核学习在单核的基础上进一步

融合了无人机平飞时俯仰角速率、爬升和下滑两种纵向飞行模态时速率陀螺发生冲击、偏差、卡死、乘性故障时俯仰角速率等多源数据信息，达到了更高的故障诊断准确性。在高光谱遥感图像分类方面，多核学习实现了空间特征和光谱特征的联合分类，分别从高空间分辨率的可见光图像和高光谱分辨率的高光谱图像中提取空间特征和光谱信息，构建多特征多核学习模型，有效地提高了空谱特征可利用性和高光谱遥感图像分类效果。

尽管多核学习取得了上述诸多优越性能，但其仍有一些问题亟待解决。首先，基本核的选择和组合方式缺乏理论依据。多核学习中的很多方法都是基于有限个基本核的线性组合加以讨论，基本核的选择也大都是启发式的。当面对一些复杂问题时，这些方法未必有效，有限个核函数融合的决策函数的性能也不可能达到处处最优。将多核学习由有限核向无限核扩展，以及考虑基本核的非线性组合方式，是一个重要的研究方向，现有的相关研究才刚刚起步。此外，目前的多核学习大多选择满足 Mercer 条件的正定核为基本核，但在实际应用中存在着大量的不定核，将不定核与多核学习相结合具有重要的理论与应用价值。其次，在“大数据”背景下，如何将多核学习扩展至大规模学习问题中需要进一步研究。对于大规模数据集，由于涉及多核矩阵的快速求解、高维多核扩展矩阵的各种分解等问题，通常的多核学习方法的学习效率会很低，如何提高其学习速度值得我们进行深入探讨。

2.3.9 集成学习

机器学习并不是为了替代传统的统计分析技术。相反，它是统计方法学的延伸和拓展。大多数的统计分析技术都基于完善的数学理论和严格的假定条件，而随着计算机能力的不断增强，我们有可能只利用计算机强大的计算能力，并通过相对简单和固定的方法达到传统统计方法无法达到的效果和目的。近年来，国内外有关机器学习的研究发展较快，由于集成学习（Ensemble Learning，EL）可以有效地提高模型的推广能力，因此从 20 世纪 90 年代开始，

对集成学习理论和算法的研究成为了机器学习的一个热点。早在 1997 年，国际机器学习界的权威 T.G. Dietterich 就将集成学习列为机器学习四大研究方向之首。四个大方向是指通过集成学习方法提高学习精度、扩大学习规模、强化学习和学习复杂的随机模型。而在今天，集成学习仍然是机器学习中最热门的研究领域之一，研究人员众多、成果层出不穷。现在已经有很多集成学习算法，比如 Bagging 算法、Boosting 算法、Arcing 算法、Random Forest 算法等。

与单一的学习模型相比，集成学习模型的优势在于能够把多个单一学习模型有机地结合起来，获得一个统一的集成学习模型，从而获得更准确、稳定和强壮的结果。集成学习的原理来源于 PAC 学习模型（Probably Approximately Correct learning）。Kearns 和 Valiant 最早探讨了弱学习算法与强学习算法的等价性问题，即提出了是否可以将弱学习算法提升成强学习算法的问题。如果两者等价，那么在学习概念时，只要找到一个比随机猜测略好的弱学习算法，就可以将其提升为强学习算法，而不必直接去找通常情况下很难获得的强学习算法。近年来，研究人员在集成学习方面，特别是分类集成方面进行了大量的探索和创新。大部分的集成学习模型都可以归为三大类：监督集成学习模型、半监督集成学习模型和非监督集成学习模型。监督集成学习模型又称为分类集成学习模型（Classifier Ensemble），包括一系列常见的分类技术。半监督集成学习模型包括多视图学习模型、共性最大化学习模型等。非监督集成学习模型又称为聚类集成（Cluster Ensemble）或一致性聚类（Consensus Clustering）学习模型。经过多年的研究，大量的聚类集成学习模型被提出来，如基于图论的聚类集成算法、基于多次谱聚类的聚类集成算法、混合模糊聚类集成算法等。然而集成学习模型的性能往往受到外在环境（如样本空间和属性空间）和内在环境（基本分类器的参数和基本分类器的权重）的影响。但是传统的集成学习模型没有考虑到这些因素的综合影响，也没有考虑到如何寻找最优的集成学习模型。而多角度自适应集成学习模型不但能够考虑到集成模型的内在环境，而且能够把握集成模型和外在环境之间的关

系。自适应集成模型之间会根据解决问题的需要进行一定的信息交互，不断地进行调整，直到达到最佳状态。多角度自适应集成学习模型将在传统集成学习模型的基础上，从多个不同角度加入自适应学习过程，从而获取最优的集成学习模型。

集成学习未来的发展趋势主要有两大块：集成学习模型的优化和集成学习模型的并行化。在大数据时代，数据来源各有不同，大数据的海量多元异构特性已经成为大数据智能处理的瓶颈。如何对多元数据进行融合和挖掘成为大数据智能处理亟须解决的问题。集成学习非常适合用于多元数据融合和挖掘，在集成学习里，集成器由一组单一的学习模型所构成，每一个学习模型都可以对应每一个来源的数据，并自动提取该数据源所蕴含的有价值的规律。因此，集成学习能够提供一个统一的框架用于分析异构性极强的多元数据，实现多元数据的融合、建模和挖掘，并从中寻找出有价值的数据语义，为政府的决策提供支持。然而，由于大数据的海量特性，使得集成学习模型的并行化处理技术变得日益重要。利用高性能服务器集群实现集成学习模型的并行化处理将成为集成学习未来的发展趋势之一。集成学习作为一种提升学习系统泛化性能的常用技术，在诸多领域有着广阔的应用前景。在美国NETFLIX 电影推荐比赛中，基于集成学习的推荐算法获得了第一名。在多次KDD 和 ICDM 的数据挖掘竞赛中，基于集成学习的算法都取得了较好的成绩。集成学习算法已成功应用于智能交通中的行人检测、车辆检测等，图像和视频处理中的动作检测、人物追踪、物体识别等，生物信息学蛋白质磷酸化位点预测、基因组功能预测、癌症预测等，数据挖掘中的脑电数据挖掘、数据流挖掘等。

2.3.10 主动学习

机器学习主要研究计算机如何利用经验数据提高自身性能，充分和高质量的数据是有效学习的基础和关键。在传统的有监督学习中，要求用于训练学习模型的数据均是已标记的。一般认为，已标记的数据越多，标记越精准，

基于这些数据训练得到的模型越高效。大数据时代为机器学习提供了丰富的原材料，使其发挥着越来越重要的作用，成为当前最热门的研究领域之一。然而，大数据提供机遇的同时也带来了严重的挑战，其中最典型的便是数据质量低下。在许多实际任务中，我们可以轻松获取大量数据，但这些数据大部分是未标注的。比如在图像分类任务中，绝大部分用户上传的照片缺乏准确的语义标签。因此如何从仅有少量标记的大数据中学习出有效模型是一个极具挑战的重要问题。

一个最直接的解决方案是先人工标注好所有数据再进行模型训练。面对海量数据时这种方案将耗费大量人力物力，显然这是不现实的。实际上，在某些现实任务中，即使标注少量数据也需要付出昂贵的代价。一个更合理的方案是挑选一部分数据进行标注。不同数据样本对于学习模型的贡献度是不一样的，如果我们能够选取一部分最有价值的数据进行标注，有可能仅基于少量数据就能获得同样高效的模型。要实现这一目标，关键在于如何选择出最有价值的数据样本，并去获取它们的标记信息。主动学习就是研究这一问题的一种机器学习框架。其核心任务是制定选择样本的标准，从而选择尽可能少的样本进行标注来训练出一个好的学习模型。

目前主要有三种主动学习场景：基于数据池的主动学习、基于数据流的主动学习及基于合成样本查询的主动学习。

（1）基于数据池的主动学习是最常见的一种场景，其假设所有未标记数据已经给定，形成一个数据池。主动学习算法迭代进行，每一次从未标记数据池中选择样本向专家查询标记，并将这些新标记的样本加入训练集，模型基于新的训练集进行更新，进而进入下一次迭代。

（2）基于数据流的主动学习假设样本以流的形式一个一个到达，因此在某时刻当一个样本到达的时候，算法必须决定是否查询该样本的标记。这种场景在一些实际应用中也比较常见，比如数据流源源不断产生，而又无法保存下来所有数据时，基于数据流的主动学习就更为适用。

（3）基于合成样本查询的主动学习并不是从已有样本中选择来查询标记信息，而是直接从特征空间里合成出新的样本进行查询。由于新合成的样本可能是特征空间里任意取值组合产生的，因此在某些应用问题中可能导致人类专家也无法标注这些合成样本。比如在图像分类任务中，任意像素取值合成的一幅图片可能并不能呈现出清晰的语义。

主动学习的关键任务在于设计出合理的查询策略，即按照一定的准则来选择被查询的样本。目前的方法可以大致分为三种策略：基于信息量的查询策略、基于代表性的查询策略及综合多种准则的查询策略。随着主动学习的广泛应用，一些实际任务中的新设置和新条件促进了主动学习技术的进一步延伸和发展。比如，在多标记学习任务中，一个样本可以同时具有多个标记，这时查询方式（以何种方式查询所选样本的监督信息）对主动学习性能非常关键。此外，在一些任务中，提供标记信息的不再是一个专家，而是一群可能提供错误信息的用户，这时如何从带有噪声的数据中获取正确的标记信息变得非常重要。在另外一些任务中，可能标注每个样本的代价不一样，这使得主动学习算法在选择样本的时候不仅要考虑样本可能带来的价值，还要考虑标注它可能花费的代价。这些新的主动学习设置和形式正引起越来越多的关注，使得其应用前景更为广阔。

随着大数据时代的来临，数据分析任务变得更加困难，同时也为主动学习的进一步发展和应用提供了巨大的机遇。首先，数据规模庞大但是质量低下，具有精确标记信息的数据尤其稀少。因此如何从海量数据中选择最有价值的部分数据进行人工标记成为了一个常见的重要步骤，这也恰恰是主动学习研究的内容。其次，数据分析任务的难度越来越高，许多学习任务仅仅依靠机器已经难以达到实用的效果。因此，人与机器在学习过程中进行交互成为了一种更有效、更现实的方案。在这样的背景下，主动学习可能会发展出更多新颖的设置，从传统查询样本标记衍生出更多的查询方式，从用户获取更丰富的监督信息。最后，随着数据来源的多样化趋势，主动学习在流数据、分布式学习、众包等场景下的研究和应用将会受到更多的关注。

2.3.11 迁移学习

在传统分类学习中，为了保证训练得到的分类模型具有准确性和高可靠性，都有两个基本的假设：① 用于学习的训练样本与新的测试样本满足独立同分布。② 必须有足够可用的训练样本才能学习得到一个好的分类模型。但是，在实际应用中我们发现这两个条件往往无法满足。首先，随着时间的推移，原来可利用的有标签样本数据可能变得不可用，与新来的测试样本的分布产生语义、分布上的缺口。其次，有标签样本数据往往很缺乏，而且很难获得。这就引起了机器学习中另外一个重要问题，如何利用少量的有标签训练样本或源领域数据建立一个可靠的模型，对具有不同数据分布的目标领域进行预测。

近年来，迁移学习已经引起了广泛的关注和研究。迁移学习是运用已存有的知识对不同但相关领域问题进行求解的一种新机器学习方法。它放宽了传统机器学习中的两个基本假设，目的是迁移已有的知识来解决目标领域中仅有少量有标签样本数据甚至没有的学习问题。迁移学习广泛存在于人类的活动中，两个不同的领域共享的因素越多，迁移学习就越容易，否则就越困难，甚至出现“负迁移”，产生副作用。比如，一个人学会了骑自行车，那他就很容易学会开摩托车；一个人熟悉了下五子棋，也可以轻松地将方法运用到下围棋中。但是有时候看起来很相似的事情，却有可能产生“负迁移”，比如，学会骑自行车的人学习骑三轮车反而会不适应，因为它们的重心不同。近几年来，已经有相当多的研究者投入到迁移学习领域中，每年在机器学习和数据挖掘的顶级会议中都有关于迁移学习的文章发表。

近十几年来，很多学者对迁移学习展开了广泛的研究，而且很多集中在算法研究上，即采用不同的技术对迁移学习算法展开研究。基于特征选择的迁移学习方法是识别出源领域与目标领域中共有的特征表示，然后利用这些特征进行知识迁移。首先选出所有领域（包括源领域和目标领域）共有的特征来训练一个通用的分类器；然后从目标领域无标签样本中选择特有特征来对通用分类器进行精化，从而得到适合于目标领域数据的分类器。基于特征

映射的迁移学习方法是把各个领域的数据从原始高维特征空间映射到低维特征空间，在该低维空间下，源领域数据与目标领域数据拥有相同的分布。这样就可以利用低维空间表示的有标签的源领域样本数据训练分类器，对目标测试数据进行预测。

迁移学习已在文本分类、文本聚类、情感分类、图像分类、协同过滤等方面进行了应用研究。迁移学习作为一个新兴的研究领域还很年轻，主要还是集中在算法研究方面，基础理论研究还很不成熟，因此值得我们进一步研究。迁移学习最早来源于教育心理学，下面借用美国心理学家贾德提出的“类化说”学习迁移理论来讨论下目前机器学习领域迁移学习研究存在的几个挑战性问题。

第一，贾德认为在先期学习 A 中获得的东西，之所以能迁移到后期学习 B 中，是在学习 A 时获得了一般原理，这种原理可以部分或全部运用于 A、B 之中。根据这一理论，两个学习活动之间存在的共同要素，是产生迁移的必要前提。这就是说，想从源领域中学习知识并运用到目标领域中，必须保证源领域与目标领域有共同的知识，那么如何度量这两个领域的相似性与共同性，是问题之一。第二，贾德的研究表明，知识的迁移是存在的，只要一个人对他的经验、知识进行了概括，那么从一种情境到另一种情境的迁移是可能的。知识概括化的水平越高，迁移的范围和可能性越大。把该原则运用到课堂上，同样的教材采用不同的教学方法，产生的迁移效果是不一样的，可能产生积极迁移也可能产生相反的作用。即同样的教材内容，由于教学方法不同，而使教学效果大为悬殊，迁移的效应也大不相同。所以针对不同的学习问题，研究有效的迁移学习算法也是另一个重要问题。第三，根据贾德的泛化理论，在讲授教材时重要的是鼓励学生对核心的基本概念进行抽象或概括。抽象与概括的学习方法是最重要的方法，在学习时对知识进行思维加工，区别本质的和非本质的属性，偶然的和必然的联系，舍弃那些偶然的、非本质的元素，牢牢把握那些必然的本质的元素。这种学习方法能使学生的认识从低级的感性阶段上升到高级的理性阶段，从而实现更广泛、更成功的正向

迁移。也就是说，在迁移学习的过程中，应该避免把非本质的、偶然的知识当成本质的（领域共享的）、必然的知识，实现正迁移。所以，如何实现正迁移避免负迁移，也是迁移学习一个重要的研究问题。针对以上讨论分析，未来的迁移学习研究可以在以下方面进行努力。

第一，针对领域相似性、共同性的度量，研究准确的度量方法；第二，在算法研究方面，对于不同的应用，迁移学习算法需求不一样。因此针对各种应用的迁移学习算法有待进一步研究；第三，关于迁移学习算法有效性的理论研究还很缺乏，研究可迁移学习条件，获取实现正迁移的本质属性，避免负迁移；第四，在大数据环境下，研究高效的迁移学习算法尤为重要。目前的研究主要还是集中在研究领域，数据量小且测试数据非常标准，应把研究的算法瞄准实际应用数据，以适应目前大数据挖掘的研究浪潮。尽管迁移学习研究还存在着各种各样的挑战，但是随着越来越多的研究人员投入到该项研究中，一定会促进迁移学习研究的蓬勃发展。

参考文献

[1] B. Settles, M. Craven. An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, 2008: 1069-1078.

[2] 周志华. 机器学习与数据挖掘. 中国计算机学会通讯，2007，3(12)：35-44.

[3] 中国机器学习白皮书．中国人工智能学会，2015．11．

[4] S.-J. Huang, R. Jin, Z.-H. Zhou. Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014. 36(10): 1936-1949.

[5] 周志华．机器学习．北京：清华大学出版社，2016．

[6] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, J. Ye. Batch mode active sampling based on marginal probability distribution matching. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012: 741-749.

[7] S.-J. Huang, S. Chen, Z.-H. Zhou. Multi-label active learning: Query type matters. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015: 946-952.

[8] P. Donmez, J. Carbonell, J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009: 259-268.

[9] D. Margineantu. Active cost-sensitive learning. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, UK, 2005: 1622-1623.

[10] R. S. Sutton, A. G. Barto. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.

[11] P. Abbeel, A. Coates, M. Quigley, A. Y. Ng. An application of reinforcement learning to aerobatic helicopter flight. In: Advances in Neural Information Processing Systems 19, Cambridge, MA: MIT Press, 2007: 1-8.

[12] Y. C. Wang, J. M. Usher. Application of reinforcement learning for agent-based production scheduling. Engineering Applications of Artificial Intelligence, 2005, 18(1): 73-82.

[13] J. J. Choi, D. Laibson, B. C. Madrian, A. Metrick. Reinforcement learning and savings behavior. The Journal of Finance, 2009, 64(6): 2515-2534.

[14] J. A. Boyan, M. L. Littman. Packet routing in dynamically changing networks:

A reinforcement learning approach. In: Advances in Neural Information Processing Systems 6, Burlington, MA: Morgan Kaufmann, 1994: 671-671.

[15] J. Frank, L. C. Seeberger, R. C. O'Reilly. By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. Science, 2004, 306(5703): 1940-1943.

[16] K. Samejima, Y. Ueda, K. Doya, M. Kimura. Representation of action-specific reward values in the striatum. Science, 2005, 310(5752): 1337-1340.

[17] T. G. Dietterich. Machine learning research: Four current directions. AI Magazine, 1997, 18(4): 97-136.

[18] C. H. Watkins. Learning from delayed rewards. Ph.D. Thesis, Kings College, University of Cambridge, 1989.

[19] P. L. Bartlett, J. Baxter. Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research, 2001, 15: 319-350.

[20] G. Rummery, M. Niranjan. On-line Q-learning using connectionist systems. Technical Report, University of Cambridge, 1994.

[21] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 1992, 8(3): 229-256.

[22] G. Konidaris, S. Osentoski, P. Thomas. Value function approximation in reinforcement learning using the Fourier basis. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, CA, 2011: 380-385.

[23] M. Bellemare, J. Veness, M. Bowling. Sketch-based linear value function approximation. In: Advances in Neural Information Processing Systems 25, Cambridge, MA: MIT Press, 2012: 2222-2230.

[24] X. Xu, D. Hu, X. Lu. Kernel-based least squares policy iteration for

reinforcement learning. IEEE Transactions on Neural Networks, 2007, 18(4): 973-992.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529-533.

[26] S. Mannor, R. Y. Rubinstein, Y. Gat. The cross entropy method for fast policy search. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013: 512-519.

[27] I. Szita, A. Lörincz. Learning tetris using the noisy cross-entropy method. Neural Computation, 2006, 18(12): 2936-2941.

[28] S. Schaal. Is imitation learning the route to humanoid robots. Trends in Cognitive Sciences. 1999, 3(6): 233-242.

[29] C. Atkeson, S. Schaal. Robot learning from demonstration. In: Proceedings of the 14th International Conference on Machine Learning, San Francisco, CA, 1997: 12-20.

[30] P. Abbeel, A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004: 1-8.

[31] B. Ziebart, A. Maas, J. Bagnell, A. Dey. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23th AAAI Conference on Artificial Intelligence, Chicago, IL, 2008: 1433-1438.

[32] A. Y. Ng, S. J. Russell. Algorithms for inverse reinforcement learning. In: Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, 2000: 663-670.

[33] P. Abbeel, D. Dolgo, A. Y. Ng, S. Thrun. Apprenticeship learning for motion planning with application to parking lot navigation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, 2008: 1083-1090.

[34] M. E. Taylor, P. Stone. Transfer learning for reinforcement learning domains: A survey. Journal of Machine Learning Research, 2009, 10: 1633-1685.

[35] M. E. Taylor, G. Kuhlmann, P. Stone. Autonomous transfer for reinforcement learning. In: Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems, Estoril, Portugal, 2008: 283-290.

[36] B. Da Silva, G. Konidaris, A. Barto. Learning parameterized skills. In: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, UK, 2012: 1679-1686.

[37] W. B. Knox, P. Stone. Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. Artificial Intelligence, 2015, 225: 24-50.

[38] J. Wang, X. Gao, Q. Wang, Y. Li. ProDis-ContSHC: Learning protein dissimilarity measures and hierarchical context coherently for protein-protein comparison in protein database retrieval. BMC Bioinformatics, 2012, 13(S-7): S2.

[39] 汪洪桥，孙富春，蔡艳宁，等．多核学习方法．自动化学报，2010，36（8）：1037-1050．

[40] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, W. S. Noble. A statistical framework for genomic data fusion. Bioinformatics, 2004, 20: 2626-2635.

[41] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st

International Conference on Machine Learning, Banff, Canada, 2004: 41-48.

[42] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, M. I. Jordan. Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research, 2004, 5: 27-72.

[43] S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf. Large scale multiple kernel learning. Journal of Machine Learning Research, 2006, 7: 1531-1565.

[44] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. More efficiency in multiple kernel learning. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007: 775-782.

[45] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. SimpleMKL. Journal of Machine Learning Research, 2008, 9: 2491-2521.

[46] Z. Xu, R. Jin, I. King, M. R. Lyu. An extended level method for efficient multiple kernel learning. In: Advances in Neural Information Processing Systems 22, Cambridge, MA: MIT Press, 2009: 1825-1832.

[47] Z. Xu, R. Jin, H. Yang, I. King, M. R. Lyu. Simple and efficient multiple kernel learning by group lasso. In: Proceedings of 27th International Conference on Machine Learning, Haifa, Israel, 2010: 1175-1182.

[48] S. V. N. Vishwanathan, Z. Sun, N. Ampornpunt. Multiple kernel learning and the SMO algorithm. In: Advances in Neural Information Processing Systems 23, Cambridge, MA: MIT Press, 2010: 2361-2369.

[49] R. Jin, T. Yang, M. Mahdavi. Sparse multiple kernel learning with geometric convergence rate. arXiv:1302.0315v1, 2013.

[50] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov. Efficient and accurate l_p-norm multiple kernel learning. In: Advances in Neural Information Processing Systems 22, Cambridge, MA: MIT Press, 2009: 997-1005.

[51] M. Varma, B. R. Babu. More generality in efficient multiple kernel learning. In: Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009: 1065-1072.

[52] A. Jain, S. V. N. Vishwanathan, M. Varma. SPG-GMKL: Generalized multiple kernel learning with a million kernels. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 2012: 750-758.

[53] C. Hinrichs, V. Singh, J. Peng, S. C. Johnson. Q-MKL: matrix-induced regularization in multi-kernel learning with applications to neuroimaging. In: Advances in Neural Information Processing Systems 25, Cambridge, MA: MIT Press, 2012: 1421-1429.

[54] C. Cortes, M. Mohri, A. Rostamizadeh. Learning non-linear comibinations of kernels. In: Advances in Neural Information Processing Systems 22, Cambridge, MA: MIT Press, 2009: 396-404.

[55] Q. Mao, I. W. Tsang, S. Gao, L. Wang. Generalized multiple kernel learning with data-dependent priors. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(6): 1134-1148.

[56] A. Nazarpour, P. Adibi. Two-stage multiple kernel learning for supervised dimensionality reduction. Pattern Recognition, 2015, 48(5): 1854-1862.

[57] C. Xu, D. Tao, C. Xu. A survey on multi-view learning. arXiv:1304.5434v1, 2013.

[58] A. Blum, T. Mitchell. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, 1998: 92-100.

[59] K. Nigam, R. Ghani. Analyzing the effectiveness and applicability of

co-training. In: Proceedings of the 9th International Conference on Information and Knowledge Management, McLean, VA, 2000: 86-93.

[60] V. Sindhwani, D. S. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In: Proceedings of the 25th International Conference on Machine Learning, Montreal, Canada, 2009: 976-983.

[61] Z.-H. Zhou, M. Li. Semi-supervised regression with co-training. In: Proceedings of the 19th International Joint Conferences on Artificial Intelligence, Edinburgh, UK, 2005: 908-916.

[62] S. Bickel, T. Scheffer. Multi-view clustering. In: Proceedings of the 4th IEEE International Conference on Data Mining, Brighton, UK, 2004: 19-26.

[63] S. Yu, K. Yu, V. Tresp, H. P. Kriegel. Multi-output regularized feature projection. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(12): 1600-1613.

[64] A. Sharma, A. Kumar, H. Daume, D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012: 2160-2167.

[65] Z.-H. Zhou, D.-C. Zhan, Q. Yang. Semi-supervised learning with very few labeled training samples. In: Proceedings of the 22nd National Conference on Artificial Intelligence, Vancouver, Canada, 2007: 675-680.

[66] J. He, R. Lawrence. A graph-based framework for multi-task multi-view learning. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, 2011: 25-32.

[67] J. Zhang, J. Huan. Inductive multi-task learning with multiple view data. In: Proceedings of the 18th ACM SIGKDD International Conference on

Knowledge Discovery and Data Mining, Beijing, China, 2012: 543-551.

[68] X. Jin, F. Zhuang, S. Wang, Q. He, Z. Shi. Shared structure learning for multiple tasks with multiple views. In: Lecture Notes in Artificial Intelligence 8189, Berlin: Springer, 2013: 353-368.

[69] M. Hodosh, P. Young, J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, 47(1): 853-899.

[70] L. Ma, Z. Lu, L. Shang, H. Li. Multimodal convolutional neural networks for matching image and sentences. arXiv: 1504.06063v1, 2015.

[71] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. The WEKA data mining software: An update. SIGKDD Explorations, 2009, 11(1): 10-18.

[72] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcaa, L. Sanchez, F. Herrera. KEEL data-mining software tool: dataset repository, integration of algorithms and experimental analysis framework. Journal of Multiple-Valued Logic and Soft Computing, 2011, 17(2-3): 255-287.

[73] M. Kearns, L.G. Valiant. Crytographic limitation on learning boolean formulae and finite automata. In: Proceedings of the 21st Annual ACM Symposium on Theory of Computing, Seattle, Washington, 1989: 433-444.

[74] L. Breiman. Bagging predictors. Machine Learning, 1996, 24(2): 123-140.

[75] Y. Freund, R. E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55(1): 119-139.

[76] L. Breiman. Random forests. Machine Learning, 2011, 45(1): 5-32.

[77] T. K. Ho. The random subspace method for constructing decision forests. IEEE Transactions Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[78] J. J. Rodriguez, L. I. Kuncheva, C. J. Alonso. Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630.

第 3 章

免疫计算的基础原理

本章导读：

仿生学模仿生物界的各种自然规律为科学的发展带来巨大的灵感，从飞翔的小鸟到飞机，从江河湖泊中畅游的鱼类到船舶潜艇，从穿山甲到防弹背心，从人脑到电脑……随着免疫学研究的深入，免疫系统呈现了一系列优良特性：分布式学习与记忆能力、模式识别能力、自适应性与鲁棒性等正引起人们越来越浓的研究兴趣，也取得了不少可喜的研究成果。随着生物学与医学免疫学的进步，生物免疫系统的一些机理不断被发现，研究结果表明，免疫系统是一个具有模式识别、学习、记忆、错误耐受等能力的分布式自适应复杂系统，其功能可与大脑媲美，这些特点对于解决许多复杂工程问题具有很好的启发作用。随着对免疫系统研究的深入，各种人工免疫算法不断被提出，并在各领域得到成功应用。

免疫学的概念、机理是开发人工免疫系统的生物理论依据，免疫系统具有许多复杂的、对实际工程问题很有启发的功能。本章对免疫学的基本概念、免疫细胞的主要免疫功能进行概述性介绍和分析，并对免疫系统运行机制进行理论阐述和概括，分析各机理的作用，为以后基于免疫学习与优化机理的免疫计算智能优化策略奠定生物免疫学基础。同时，对现有的人工免疫基本概念及系统模型及进行了说明。

3.1 免疫计算生物学基础

3.1.1 免疫学基本概念

在生物学领域中，免疫学是一门相对年轻的学科，然而，人类对自然免疫的认识可以追溯到 300 多年以前。早在 16 世纪，我国医学家就创造性地发明了“人痘”来预防天花。1796 年，英国医生 Edward Jenner 发明了“牛痘”，取代了人痘苗，是公认的现代免疫学发展的开端。法国免疫学家 Pasteur 发明了病毒细菌疫苗，奠定了经典免疫疫苗的基础。经过 300 多年的发展，免疫学已经从微生物学的一个分支发展成为了一门独立的学科，并派生出若干分支，如细胞免疫学、分子免疫学、神经与内分泌免疫学、生殖免疫学和行为免疫学等。表 3-1 总结了免疫学历史上比较重要的思想、理论和研究成果。

表 3-1 自然免疫学主要理论发展史

主要目标	时　　间	代表人物	思想、理论和研究成果
经验免疫时期	16 世纪起	中国民间	“人痘”的发明和应用
	1796—1870 年	Jenner	“牛痘”的发明和应用
		Koch	病理学

续表

主要目标	时　间	代表人物	思想、理论和研究成果
经验免疫时期	1870—1890 年	Pasteur	疫苗接种
		Besedovsky	神经－内分泌－免疫网络学说
		Metchinikoff	噬菌作用
科学免疫时期	1890—1910 年	Von Behring & Kitasato	发现抗体
		Ehrlich	发现细胞受体
	1910—1930 年	Bordet	免疫特性
		Landsteiner	半抗原
现代免疫时期	1930—1950 年	Breinl & Haurowitz	抗体合成
		Linus Pauling	抗原模型
	1950—1980 年	Burnet	克隆选择
		Niels Jerne	免疫网络与协作理论
分子水平研究	1980—1990 年	Susumu Tonegawa	受体的结构与多样性

免疫指机体对感染具有抵抗能力而使其不患疫病。免疫学是研究免疫系统的结构和功能，研究免疫系统识别并消除有害生物及其成分的应答过程及机制，理解其对机体有益的防卫功能和有害的病理作用及其机制的医学科学。

3.1.2　生物免疫系统的结构及组成

免疫系统是生物所具有且必备的防御机制。免疫系统不依靠任何中心控制，具有分布式任务处理能力，具有在局部采取行动的智能，也通过起交流作用的化学信息构成网络，进而形成全局观念。免疫系统多种多样，具有独

特性。其中人的免疫系统最为复杂，它是由免疫活性分子、免疫细胞、免疫组织和器官组成的复杂系统。免疫系统具有识别机制，能够从人体自体细胞、自体分子和外因感染的组织中，检测并消除病毒等病原体本身，以及因感染而引起的机能不良、功能紊乱、官能障碍等症状，并且能够“记忆”每一种感染源，这样当同样的感染源再次入侵机体的时候，免疫系统就能够更快地识别和应答，进行更有效的处理。

1. 生物免疫系统的结构

免疫系统的结构本质上是多层次的，由分布在几个层次的防御系统组成，图 3-1 为生物免疫系统的体系结构示意图。

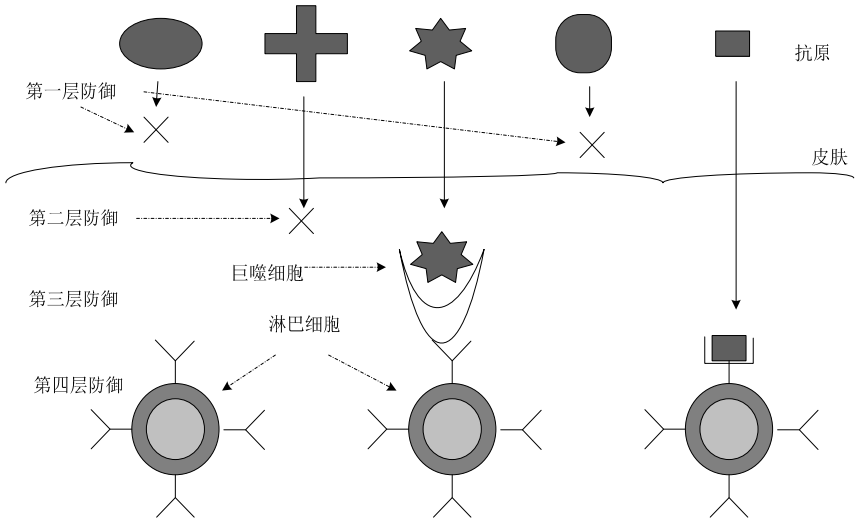

图 3-1　生物免疫系统的体系结构

（1）物理屏障：由皮肤和黏膜及局部细胞分泌的抑菌、杀菌的化学物质都属于物理屏障的范畴。

（2）生理屏障：唾液、汗液、眼泪之类的体液含有丰富的生物酶，能够有效分化和破坏微生物。

（3）免疫系统：可分为固有免疫系统和自适应免疫系统，由淋巴细胞、

噬菌细胞和细胞因子等组成，能有效地识别入侵的微生物，并采取相应的措施将其清除。

2. 生物免疫系统的组成

生物免疫系统由免疫细胞、组织和器官组成，其中最重要的是淋巴系统和补体系统。

1）淋巴系统

组成免疫系统的组织和器官分布于人体各处，以完成各种免疫防卫功能，它们就是人们熟知的淋巴细胞和淋巴组织。淋巴器官按照功能分为：中枢淋巴器官，由骨髓和胸腺组成，执行生成免疫细胞的功能；外周淋巴器官，由淋巴结、脾、扁桃体组成，成熟的免疫细胞在这些部位执行应答功能。淋巴系统的组成如图 3-2 所示。

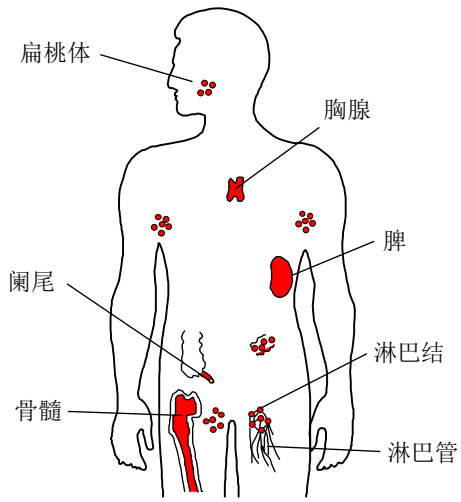

图 3-2　淋巴系统

2）免疫细胞

免疫细胞由造血干细胞、淋巴细胞、单核吞噬细胞系统组成，主要是在骨髓和胸腺中发育成熟，它们在血液和淋巴液中循环，图 3-3 所示为免疫系统

产生的细胞和分泌物的结构划分。免疫系统中某些免疫细胞负责固有免疫系统中的一般防御，而其他免疫细胞则在自适应免疫系统中担负起了对付高度特异化病原体的重任。

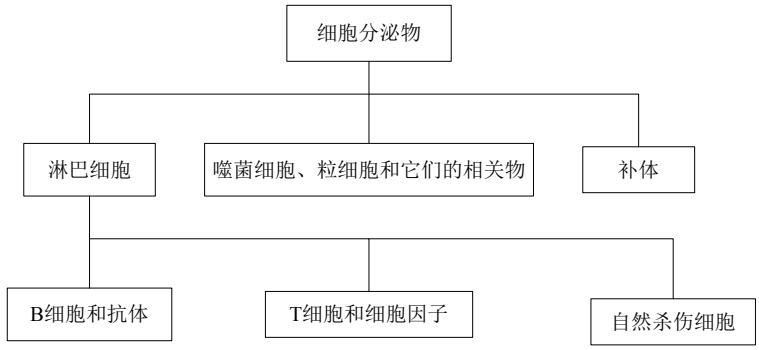

图 3-3　免疫系统的细胞和分泌物的结构划分

B 细胞是免疫系统的本质部分，是 B 淋巴细胞的简称。B 细胞是体内唯一能够产生抗体的细胞。它的表面含有可以识别特异性抗原的多种抗体分子，其多样性来自千百万种不同的 B 细胞克隆。B 细胞能与 T 细胞相互作用，并在后者的辅助下激活。

B 细胞有三个主要的功能：产生抗体、提呈抗原和分泌细胞因子参与免疫调节。B 细胞在免疫应答和清除病原体的过程中起主要的作用，在清除病原体的过程中受到刺激，分泌抗体结合抗原，但其发挥免疫作用要受 T 辅助细胞的帮助。

3）抗原与抗体

所谓抗原指凡是能够诱导免疫系统发生免疫应答，并能与其产生的抗体或效应细胞在体内或体外发生特异性反应的物质，图 3-4 为抗体与抗原示意图，抗原表面被抗体识别的部分被称为抗原决定基。

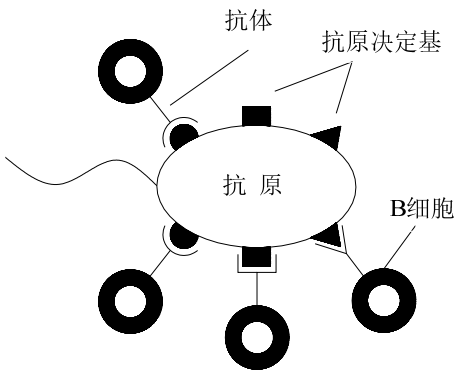

图 3-4 抗体与抗原

抗体是 B 细胞识别抗原后克隆扩增分化为浆细胞所产生的一种蛋白质分子，也称为免疫球蛋白分子。抗体结合由外部入侵的感染微组织或有毒物等抗原，然后依靠自己或者借助免疫系统其他元素（如 T 细胞等）帮助破坏这些抗原，消除对人体的威胁。抗体由抗体决定基和独特型组成，抗体决定基是抗体上识别抗原决定基的部分，而独特型是抗体上可供自身免疫细胞识别的抗原决定基。抗体是具有两种截然不同的功能区的分子：一部分是保持相对静态的区域，称为稳定区，简称 C 区；另一部分则是负责与不同的多种感染抗原结合的分子变化区，称为可变区，简称 V 区，图 3-5 为抗体结构示意图。正是可变区为免疫系统提供了大部分鲁棒性和自适应能力。

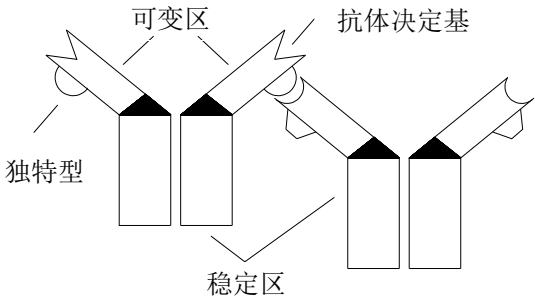

图 3-5 抗体结构图

3.1.3 免疫系统功能及机制

免疫是机体的一种特异性生理反应，通过识别和排除抗原等异物维持机体内环境的稳定，其功能是通过众多免疫细胞和免疫分子之间的相互作用实现的。机体的免疫功能是在淋巴细胞、单核细胞和其他有关细胞及其产物的相互作用下完成的。免疫系统是机体执行免疫功能的机构，是产生免疫应答的物质基础。免疫系统的主要功能有：

（1）免疫防御，即机体防御病原微生物的感染；

（2）免疫（自身）稳定，即机体通过免疫功能经常消除那些损伤和衰老的细胞以维持机体的生理平衡；

（3）免疫监视，即机体通过免疫功能防止或消除体内细胞在新陈代谢过程中发生突变的和异常的细胞。

图 3-6 所示为经典免疫反应中各因素间的相互关系。

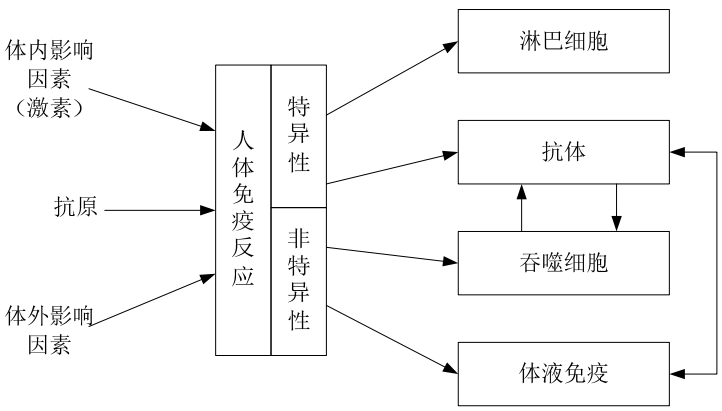

图 3-6 免疫反应中各因素间的相互关系

免疫反应是针对有害于机体的异物的防御机制。当异物（抗原物质）侵入机体时，巨噬细胞先将其摄取，并将该抗原信息传递至相当于电子计算机的免疫记忆装置，在这里分拆处理抗原信息，如断定为“非己”，则免疫活性细胞——T 淋巴细胞和 B 淋巴细胞活化增殖，对该抗原发动攻击。免疫系统的

免疫应答机制及其许多重要的功能，如免疫识别、免疫学习、免疫记忆、免疫宽容和免疫自适应调节等是免疫学研究的重要内容。下面将分别对其进行简要描述，以阐述免疫原理对人工免疫系统研究的重要借鉴作用。

1. 免疫应答与记忆

免疫应答指免疫细胞对抗原分子的识别、活化、分化和产生免疫效应的全过程。免疫应答一般是由抗原引发的、由多种免疫细胞参加的一系列反应。首先是抗原递呈细胞把经过加工处理的抗原递呈给相应的 T 细胞（抗原特异性 T 细胞克隆），活化以后的细胞通过细胞膜上的分子或分泌的细胞介素进一步活化其他免疫细胞，产生相应的免疫效应，促进对抗原和靶细胞的吞噬，最终清除抗原，维持机体内部的平衡和稳定。

图 3-7 为免疫应答示意图。免疫系统有两种免疫应答类型：一种是遇到病原体后，首先并迅速起防卫作用的称为固有免疫应答；另一种是适应性免疫应答。在固有免疫应答中，执行固有免疫功能的有皮肤和黏膜的物理阻挡作用及局部细胞分泌的抑菌、杀菌物质的化学作用等。固有免疫一般发生在感染的早期。

适应性免疫应答是以 T 细胞和 B 细胞为执行载体的。它们在免疫过程中帮助识别和破坏清除一些特定物质（抗原）。抗原就是任何能被 T 细胞和 B 细胞识别并刺激 T 细胞及 B 细胞进行特异性免疫应答的物质。B 细胞是抗体的唯一产生者，但是当 B 细胞识别到抗原的时候，只有在与抗原的匹配程度超过某一阈值的时候，才有可能分化成浆细胞，否则没有达到阈值的 B 细胞就会凋亡。即使超过了阈值，B 细胞也并不能马上分化成浆细胞，而需要在 T 细胞也同时检测到 B 细胞提呈的抗原缩氨酸之后，由 T 细胞分泌刺激物对 B 细胞进行刺激，B 细胞才会分化成浆细胞制造大量的抗体。没有 T 细胞的协同刺激，B 细胞就不能分化。

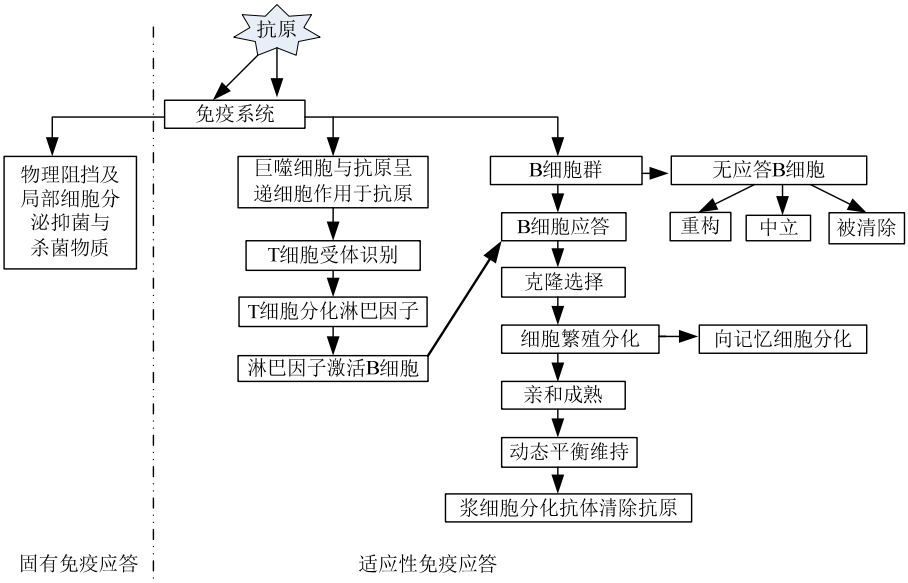

图 3-7　免疫应答

B 细胞和 T 细胞表面都存在大量的抗原识别受体，这些受体通过对于特定的抗原决定基高度特异的来识别抗原。T 细胞表面的抗原识别受体称为 T 细胞受体，B 细胞表面的抗原识别受体称为 B 细胞识别受体，简称抗体。B 细胞抗体与完整的抗原分子表面上的抗原决定基相互作用，当 B 细胞的抗体能够完全或部分识别抗原表面的抗原决定基的时候，B 细胞就通过克隆扩增，进入到高频变异和受体编辑阶段，实现对抗原决定基的高度特异识别。T 细胞只与细胞表面分子进行相互作用。T 细胞分泌能够杀死或者促进其他细胞（如 B 细胞）生长的化学物质，在免疫调节中起重要作用，通过识别细胞表面的异常分子，T 细胞必须判断是直接溶解该细胞，还是寻求与其他细胞进行合作来清除该细胞。

T 细胞和 B 细胞与抗原结合后开始活化，但活化后并不立刻表现出防卫功能，而是经过免疫应答过程，四五天后才生成效应细胞，对已被识别的病原体施加杀伤进行破坏并清除。适应性免疫应答是继固有免疫应答之后发挥效应的，在最终清除病原体、促进疾病治愈及防止再感染中起主导作用。

适应性免疫应答又分为两种类型：初次免疫应答和二次免疫应答。

1）初次免疫应答

初次免疫应答发生在免疫系统遭遇某种病原体第一次入侵时，此时免疫系统对感染产生大量抗体，帮助清除体内抗体。图 3-8 为免疫应答过程示意图，*x* 轴表示时间，*y* 轴表示抗体浓度。

在图中可以看出，在免疫系统第一次遭遇病原体时的初次应答中，抗体浓度开始升高，但几天以后抗体的浓度又开始下降，直到再次遇到同种抗原。初次应答是对以前从未见过的病原体的应答过程，因此应答的过程很慢，需要较长的时间来清除病原体。

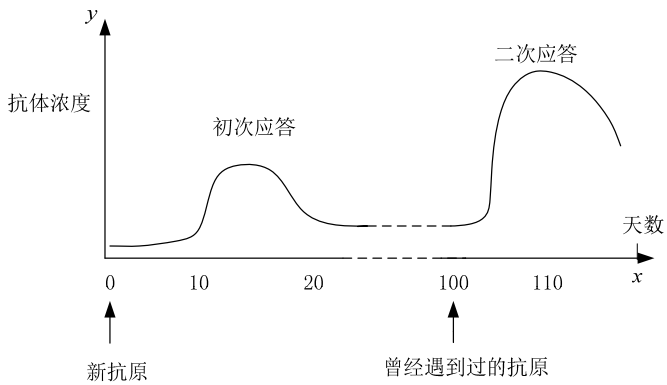

图 3-8　免疫应答过程

2）二次免疫应答

初次免疫应答时，免疫系统首次遭遇抗原并将其清除出体外，但免疫系统中仍保留一定数量的 B 细胞作为免疫记忆细胞，这使得免疫系统能够在再次遭遇异物后仍能快速识别并清除抗原，这个过程称为二次免疫应答。二次免疫应答是在初次免疫应答分化遗留下来的记忆细胞的基础上实现的，如图 3-8 所示，当初次免疫应答遭遇过的相同抗原或类似抗原出现时，免疫记忆细胞迅速从休眠状态活化，并通过克隆增生过程产生大量的 B 细胞和抗体，

中和入侵的抗原，因此应答的反应速度非常快，省去了学习的时间，这就是免疫系统实现自适应应答的一种重要机制，即免疫记忆。在二次免疫应答中，对引起初始免疫反应及造成免疫系统 B 细胞和抗体数量迅速增加的抗原是特异的，必须是曾经遭遇过的相同或类似的抗原。

2. 阴性选择与自体耐受

淋巴细胞的阴性选择指的是：淋巴细胞识别了自体细胞上的抗原，结果导致该淋巴细胞死亡或无反应力，被免疫系统删除。B 细胞抗体上存在独特型（抗原决定基），同样 T 细胞上也含有抗原决定基。因此无论是 B 细胞还是 T 细胞都有可能被淋巴细胞所识别，并被当作非自体细胞，成为免疫系统的攻击目标，这对免疫系统来说是致命的，因此免疫系统提供了阴性选择机制来消除或抑制那些识别自体特异抗原的淋巴细胞。免疫耐受指免疫活性细胞接触抗原性物质时所表现的一种特异性的无应答状态。它是免疫应答的另一种重要类型，也是机体免疫调节的内容之一，其表现与正向免疫应答相反，也与各种非特异性的免疫抑制不同，后者无抗原特异性，对各种抗原均呈现应答或低应答。免疫耐受现象是由于抗原诱导的专一性淋巴细胞功能缺失或死亡，而导致的机体对该抗原反应功能丧失或无应答的现象。抗原侵入机体后可能导致淋巴细胞的活化，也可能产生免疫耐受，这是淋巴细胞对抗原的识别和应答的两种可能结果。诱导免疫耐受的抗原称为耐受原，而诱导产生正常免疫反应的抗原称为免疫原。

3. 克隆选择与扩增

克隆选择原理的基本思想是只有那些能够识别抗原的细胞才能进行扩增，只有那些细胞才能被免疫系统选择并保留下来，而那些不能识别抗原的细胞则不被选择，也不进行扩增。

1959 年，Burnet 提出的克隆选择学说认为，免疫细胞是随机形成的多样性的细胞克隆，每一克隆的细胞表达同一特异性的受体，当受到抗原的刺激时，细胞表面受体特异识别并结合抗原，导致细胞进行克隆扩增，产生的大

量后代细胞合成大量相同的特异性抗体。

克隆扩增指少数与抗原结合亲和力较高的 B 细胞通过分裂产生大量相同的 B 细胞。当 B 细胞克隆扩增时，它经历一个自我复制超变异的随机过程，免疫系统此时产生大量的抗体从体内清除感染的抗原，并为抵制下一次某个时候类似但不同的感染做好准备。在某些情况下，克隆是通过细胞变异形式完成的，这一机制使免疫系统具有自适应性，也就是通过调整特殊的变异机制产生抗体分子基因密码变异。在克隆扩增过程中，也会产生一定数量的自由抗体，图 3-9 为克隆选择扩增原理示意图。

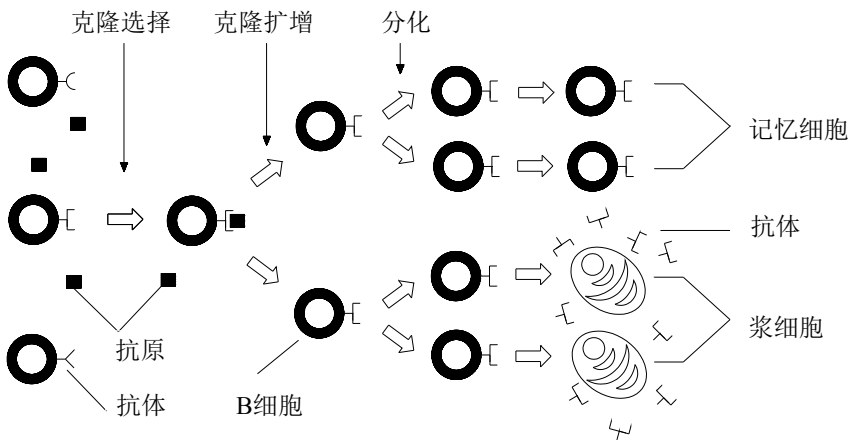

图 3-9　克隆选择扩增原理

4. 免疫系统的多样性及调节

免疫调节指免疫应答过程中免疫系统内部各细胞之间、免疫细胞与免疫分子之间、免疫系统与其他系统如神经、内分泌、遗传系统之间的相互作用，从而构成了一个相互协助又相互制约的网络结构，使免疫应答维持合适的强度以保证内环境的稳定。

免疫系统的多样性指免疫系统能够产生多种多样的抗体来应对自然界几乎无限多的抗原。免疫系统通过淋巴细胞上的抗体决定基结合病原体上的抗

原决定基来识别并清除入侵的抗原，因此要保证结合的效率和有效清除抗原、治愈感染，免疫系统就必须有充分的多样性淋巴细胞受体，以确保至少有一些淋巴细胞能够结合任何给定的抗原决定基。

免疫系统的多样性主要靠体细胞高频变异、受体编辑和随机生成新抗体来实现。

1）体细胞高频变异

B 细胞与抗原结合后被激活，激活后的 B 细胞就进入了克隆扩增阶段，在克隆扩增期间 B 细胞将会以极高的频率发生变异，该过程称为体细胞高频变异。体细胞高频变异是克隆扩增期间产生的重要变异形式，对受体多样性的产生起重要作用。体细胞高频变异的实质是抗体可变区的 DNA 基因片段重新排列，从而改变了可变区的结构，形成了一种新的抗体。

B 细胞在克隆选择与扩增过程中所进行的体细胞高频变异过程，使变异后的子代 B 细胞增加了具有不同于父代受体结构的抗体决定基，因此就会有不同的抗原决定基亲和力。新的 B 细胞具有与在淋巴结内捕获的抗原决定基发生结合的机会。如果不结合将很快凋亡；如果结合成功，则离开淋巴结，分化为浆细胞和记忆 B 细胞。

变异的迅速积累对于免疫应答的快速成熟是必需的，但是多数变化会导致更弱。如果一个细胞刚刚采用一种有用的变异，并以同一速率在下一次免疫应答期间继续变异，则衰弱变化的积累可能引起变异优点的损失。免疫系统为了避免这种情况发生，在体细胞高频变异爆发之后，进行克隆选择和扩增，给亲和力提高了的细胞以呼吸空间。同时，选择机制也可以依靠亲和力来调节高频变异过程，使具备低亲和力受体的细胞进一步变异，而具备高亲和力受体的细胞则可以不激活高频变异。

2）受体编辑

如图 3-10 所示，B 淋巴细胞被抗体激活进入了克隆扩增时，B 细胞抗原受体将会发生基因重组，这个过程被称为受体编辑。受体编辑发生时，现有 B

细胞上抗体基因片段将会与遗传基因库中的 DNA 基因片段进行重组，形成新的特异识别抗体，这样产生的子代 B 细胞就可能比父代 B 细胞具有与特异抗原更高的亲和力。受体编辑是免疫系统保持高度多样性的又一重要机制。

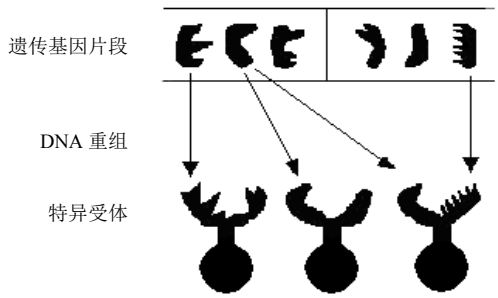

图 3-10　受体编辑的示意图

当 B 细胞抗体经过变异和编辑后对于外部入侵抗原的亲和力降低时，抗体将无法与抗原相结合，这样 B 细胞将会死亡，这就是免疫系统的克隆删除功能。研究成果表明，在抗原结合部位的形态空间中，受体编辑具有在亲和力域内避免局部优化的能力，而体细胞高频点变异对搜索局部区域有良好的作用。图 3-11 中 x 轴表示所有可能的抗体抗原结合形式，y 轴表示抗体与抗原之间的亲和力。

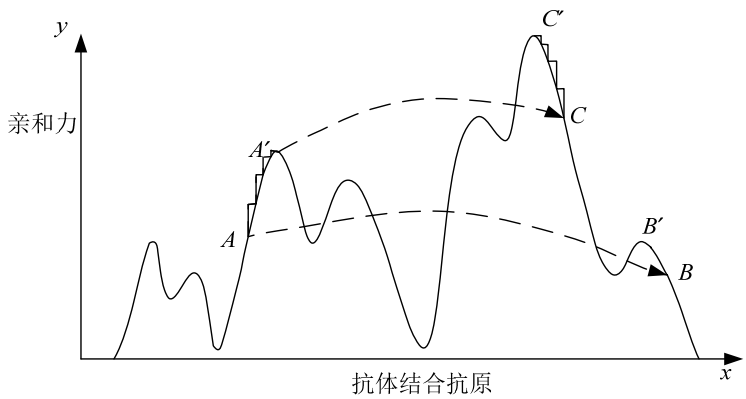

图 3-11　抗原结合部位的形态空间

3）随机生成新抗体

虽然受体编辑和高频变异能够使免疫系统保持一定的多样性，但一般的生物免疫系统只同时存在 105 种不同的蛋白质，而自然界却有 107 种不同的外部蛋白质或模式（抗原决定基）需要识别，按照数目分析，免疫系统的多样性显然不够充分去结合每一个可能的抗原决定基。在这种情况下可能会引起严重的问题，生物体如何才能识别这些外部病原体呢？免疫系统通过动态性解决这个问题。为了保持免疫系统的高度多样性，骨髓每天都要随机产生大量新的抗体，而大量免疫系统原有的没有与抗原结合的抗体将会凋亡，新产生的这些抗体进入到免疫系统虽然可能因不能结合抗原而最终导致死亡，但是却能够增加并保持免疫系统的多样性，以应对那些可能从未碰到过的病原体。

5. 免疫独特型网络

免疫网络理论是由美国学者、诺贝尔奖获得者 Jerne 于 1974 年提出的，该理论指出免疫系统是由在没有抗原的情况下也能相互识别的细胞分子调整的网络构成。免疫独特型网络指免疫系统是识别独特型集合的抗体决定基组成的巨大而复杂的网络。网络中的每一个抗体在识别抗原的同时也被其他抗体识别。淋巴细胞能够对正或负的识别信号做出反应。一个正反应信号将导致细胞扩增、细胞活化和抗体分泌，而负反应信号则导致耐受和抑制。

免疫独特型网络示意图如图 3-12 所示，当抗原 Ag 侵入时，抗体 1 特异识别该抗原，两者之间具有很高的亲和力，因此抗体 1 所在的淋巴细胞受到抗原的刺激开始扩增、分化并分泌大量与抗体 1 相同的抗体。此时抗体 2 恰恰能够特异识别抗体 1 上的独特型，抗体 1 又刺激抗体 2 的增加，相当于抗体 2 特异识别抗体 1 的抗原决定基（独特型），依此类推。反过来，抗体 2 的增加将会抑制抗体 1 的产生和数量，而抗体 1 又会抑制和清除体外的入侵者抗原 Ag。因此整个免疫系统就是在这种相互促进、相互抑制的过程中保持一种动态平衡，即保持免疫系统的稳定。

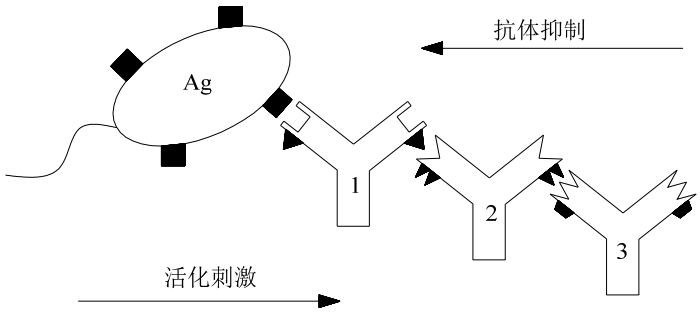

图 3-12 免疫独特型网络示意图

淋巴细胞通过抗原-抗体相互作用而动态地联系。不仅抗原，由淋巴细胞产生的抗体也起抵制其他淋巴细胞的作用，这样便可以勾画出一幅免疫系统内在的抗原内影像，如图 3-13 所示。

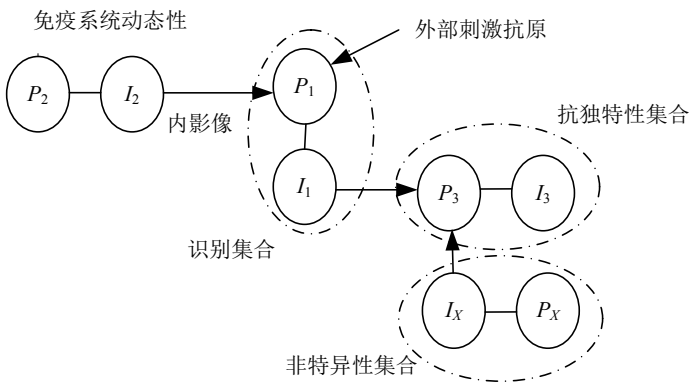

图 3-13 免疫独特型网络与抗原内影像示意图

当免疫系统内充满某一种抗原时，根据不同程度的特异性，其抗原决定基被一组称为 P_1 的不同抗体决定基识别。这些抗体决定基连同一定独特型出现在抗体和受体分子上，这样抗原决定基的集合 P_1 与独特型 I_1 的集合相关。用 P_1I_1 表示识别抗体分子和潜在的与抗原反应的淋巴细胞的整个集合。在免疫网络内部，集合 P_1 的每一个抗体决定基识别一组独特型，整个集合 P_1 识别一个更大的独特型集合。独特型集合 I_2 称为抗原决定基（或抗原）的内部影像，因为它被识别抗原的同一集合 P_1 所识别。集合 I_2 与出现在分子和集合 P_2I_2

的细胞受体上的抗体决定基集合 P_2 相关。再者，集合 P_1I_1 的每一个独特型被一组抗体决定基识别，这样整个集合 I_1 被一个更大的与抗独特型集合 I_3 一起出现的抗体和抗独特型集合 P_3I_3 的淋巴细胞上的抗体决定基 P_3 识别。按照这种机制规律，得到一个更大的识别（也被识别）抗体集合，这样可以形成免疫记忆，即抗体保留了抗原的特征内影像。

除了识别集合 P_1I_1，还有一个免疫球蛋白并行集合 P_xI_x 和细胞受体，细胞受体表示分子中与结合部位有关的集合 I_x 的独特型，该结合部位与外源抗原决定基不符。图中箭头表示当一个独特型被细胞受体上的抗体决定基识别时的刺激作用，以及当抗体决定基识别细胞受体上的独特型时的抑制作用。免疫独特型网络理论是人工免疫系统的生物学理论基础，许多免疫计算智能、人工免疫算法与模型设计都是基于该理论的基本思想而进行的。

6. 免疫系统反馈

免疫系统反馈机制主要实现两个任务：一是迅速对外部物质的出现进行应答；二是稳定免疫系统。免疫系统显示两种类型的应答：一种是体液应答；另一种是细胞应答。在体液应答中，抗体由 B 细胞产生，并去中和抗原；在细胞应答中，杀伤 T 细胞捕捉微生物细胞或者被病毒感染的细胞，然后杀死它们，如图 3-14 所示。由于 T 细胞在两种应答中都起重要作用，所以该机制称为 T 细胞调节。

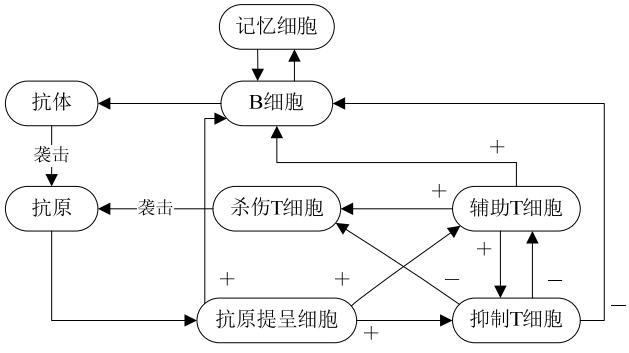

图 3-14　免疫系统反馈

在外部物质被抗原提呈细胞消化后，抗原提呈细胞把抗原的有关信息传递给辅助 T 细胞并激活辅助 T 细胞，然后辅助 T 细胞刺激 B 细胞、杀伤 T 细胞和抑制 T 细胞。激活 B 细胞和杀伤 T 细胞的调节机制是免疫系统最重要的反馈机制。

辅助 T 细胞和外部物质激活抑制 T 细胞，而抑制 T 细胞禁止其他细胞的活动，如辅助 T 细胞、B 细胞和杀伤 T 细胞等，最后使免疫系统的反应平静下来，这是免疫系统中的另一种反馈机制。禁止机制和主反馈机制之间的合作使免疫反馈机制对外部物质迅速反应，并使免疫系统迅速稳定。

免疫系统具有分布式、自适应性、鲁棒性等许多重要的特性。免疫系统的分布式特性首先取决于病原体的分布式特征，即病原体是分散在机体内部各处的，对遍布全身的病原体的检测和清除由遍布全身的效应细胞完成；其次免疫系统的分布式特性有利于加强系统的鲁棒性，从而使免疫系统不会因为局部组织损伤而使整体功能受到很大影响，也不会因为在应答过程中单个免疫协同发生错误而导致灾难性后果。分散于机体各部分的淋巴细胞采用学习的方式实现对特定抗原的识别，完成识别的抗体以正常细胞变异概率的 10 倍进行变异，使得其亲和度提高的概率大大增加，并通过分化为效应细胞和记忆细胞分别实现对抗原的有效清除和记忆信息保留，这个过程实际上是一个适应性的应答过程。由于免疫应答机制是通过局部细胞的交互起作用而不存在集中控制，所以系统的分布式进一步强化了其自适应特性。

3.2 人工免疫基本原理

人工免疫系统理论及应用是近年来刚兴起的新研究领域，起源于 20 世纪 80 年代末 90 年代初，国际上的几位专家在人工免疫系统方面作了开创性工作，如 Farmer、Dasgupta、Jerne、Forrest、De Castro、Forrest、Fukuda 等，国内是在 20 世纪 90 年代末才相继有文献出现的。人工免疫系统本质上是依据免疫

系统的机理、特征、原理开发的并能解决工程问题的计算或信息系统。关于AIS 的定义尚未达成共识，在此引用文献的定义，即所谓 AIS（从工程和科学角度），就是研究借鉴、利用生物免疫系统（主要是人类的免疫系统）的各种原理和机制而发展的各类信息处理技术、计算技术及其在工程和科学中应用而产生的各种智能系统的统称。从此意义上讲，AIS 不仅涉及计算系统，而且涉及信息处理的其他技术。开发免疫系统的任务不是激发人们对免疫系统的内在功能的运行规律进行深入探讨，而是激发人们提取免疫系统中对解决工程问题有益的特征和免疫机制，借助免疫学原理对各种运行机制进行有机组合，建立能有效解决实际问题的智能系统。

20 世纪 70 年代，美国诺贝尔奖获得者，生物学家、医学家、免疫学家 Jerne 提出了免疫系统的网络假说，开创了独特型网络理论，给出了免疫网络的数学框架，Perelson 对此进行了进一步阐述。1986 年，Farmer 提出了基于免疫网络的假说，构造了免疫系统的动态模型，并与 Holland 的分类器系统进行比较，提出了一些有价值的学习算法的构造思想。他们的研究工作为建立有效的基于免疫原理的计算系统和智能系统的发展开辟了道路。Forrest 等人基于免疫系统阴性选择理论提出了一种用于计算机安全系统的人工免疫系统模型。De Castro 等人在克隆选择原理的基础上提出了一种基于克隆选择原理的人工免疫系统模型及人工免疫网络模型，都为使免疫系统成为有效的解决工程问题的灵感源泉作出了巨大的贡献。随后的研究者不断从生物免疫系统中抽取隐喻机制，越来越多的研究者与研究机构渐渐认识到了人工免疫系统在计算机信息安全、机器学习、最优化、模式识别、故障诊断、图像处理、自动控制与数据挖掘等领域潜在的应用前景，开始关注人工免疫系统理论的发展与应用，并投入到人工免疫系统这一新兴的热门研究领域中，这极大地推动了人工免疫系统的发展、人工免疫算法模型设计、算法实现及其在工程领域中的应用。

3.2.1 人工免疫系统基本概念

人工免疫系统正是在研究借鉴、利用生物免疫系统信息处理机制的基础上发展而来的各类信息处理技术、计算技术及其在工程学科中应用而产生的各种智能系统的统称。人工免疫系统是一个跨越多个学科的研究领域，是与生物免疫系统相对应的工程概念。因此在人工免疫系统中完全套用生物学定义，照搬生物学过程，是不可能也不必要的。为了更好地描述人工免疫系统算法，以下将简要阐述几个常用的免疫学术语及其在人工免疫系统中的含义。

1. 抗原

在人工免疫系统中，一般指问题及其约束，与进化算法中的适应度函数类似，是人工免疫系统算法的始动因子及重要的度量标准。

2. 抗体

在人工免疫系统中一般指问题的候选解，与进化算法中的个体相似，抗体的集合称为抗体群。在实践中，一般抗体是以编码的形式出现的，常用的编码形式有二进制和十进制。例如，一个抗体结构为八位二进制数，则位串“01110100”即代表该抗体。

3. 抗体–抗原亲和力

抗体对抗原结合力的大小常用亲和力表示，以反映抗体的单个结合部位和单价抗原（或表位）的结合力。在人工免疫系统中，用这一概念来表示抗体不同位置（编码）对抗原（或目标函数）的影响。

4. 抗体–抗原亲和力

反映整个分子抗体与抗原之间总的结合力。在人工免疫系统中，一般指候选解所对应的目标函数值或候选解对问题的适应性度量。

5. 抗体-抗体亲和度

反映抗体与抗体间的结合能力。在人工免疫系统中一般指候选解间的距离，对应于二进制编码一般采用海明距离，而实数编码一般采用范数，而且多为欧几里得距离。

6. 疫苗

是指根据进化环境或待求问题的先验知识，所得到的对最佳个体基因的估计。

7. 免疫记忆单元

在人工免疫系统中，记忆单元是由特定抗体组成的抗体群，用于保持种群多样性，以及求解过程中的最优解。

8. 克隆

生物的增生过程。在人工免疫系统中的克隆算子是基于克隆选择学说的，充分结合了选择、扩展、变异和交叉的综合算子。

3.2.2 人工免疫系统基本原理及机制

1. 免疫阴性选择的系统模型

Perelson 和 Forrest 介绍了一种基于阴性选择和耐受的人工免疫系统（Artificial Immune System，ARTIS）模型，该模型包括多样性、分布式计算、错误耐受、动态学习和适应性与自我检测。ARTIS 是分布式自适应系统的一般框架，该系统基于免疫系统检测抗原的原理，用于计算机安全系统。该模型较为完整地给出了从免疫系统到人工免疫系统的模型映射，提供了完整的人工免疫系统问题求解范式。

2. 免疫克隆选择原理的系统模型

De Castro 等人在克隆选择原理的基础上建立了人工免疫系统模型，称为克隆选择算法（Clonal Selection Algorithm，CSA），并将该模型应用于模式识别与优化领域当中，取得了较为满意的成果。该模型主要考虑以下几个方面：

（1）使功能性的记忆细胞从指令系统分离；

（2）受刺激最强的个体选择和克隆；

（3）未受刺激的细胞死亡；

（4）亲和力成熟和较高亲和力克隆的重新选择；

（5）多样化的产生和保持；

（6）与细胞亲和力成比例的高频变异。

克隆选择学说的中心思想是，抗体是天然产物，以受体的形式存在于细胞表面，抗原可与之选择性地反应。抗原与相应抗体受体的反应可导致细胞克隆性增殖，该群体具有相同的抗体特异性，其中某些细胞克隆分化为抗体生成细胞，另一些形成免疫记忆细胞来参加之后的二次免疫反应。此外，该学说认为，免疫耐受是由于自身抗原或胚胎成熟过程中引入的抗原所致的“克隆流产”。

克隆选择是生物体免疫系统自适应抗原刺激的动态过程。这一过程中所体现出的学习、记忆、抗体多样性等生物特性，正是人工免疫系统所借鉴的。

3. 免疫网络模型

1）免疫独特型网络

免疫网络学说认为，免疫系统中的各个细胞克隆不是处于一种孤立的状态，而是通过自我识别、相互刺激和相互制约构成了一个动态平衡的网络结构；构成相互刺激和相互制约的物质基础即为抗原与抗体结合时所表现出的独特型和抗独特型。与 Burnet 的克隆选择学说着重强调免疫系统对抗原识别

不同，Jerne 的独特型网络调节学说建立在抗体自身识别的基础上，认为免疫系统淋巴细胞上分布的特异性抗原受体可变区（V 区）组成内网络，通过免疫细胞相互识别 V 区上的抗原决定簇来实现免疫系统的功能；对外来抗原的应答，是建立在识别自身抗原基础上的反应。

独特型网络调节学说免疫系统是一个由细胞、分子和器官组成的复杂系统，主要用于限制异物对肌体的侵害，并由此产生抗体，引发免疫应答。独特型网络调节学说主要与抗体有关。抗体上能够被其他抗体识别的部分，叫作独特型抗原决定簇；能对它识别并引起反应的抗体叫抗独特型抗体，该抗体又可以被其他特型抗体所识别。独特型网络调节学说把免疫系统设想为网络，抗体克隆之间的活动通过其独特型彼此沟通、互相联系、互相制约。网络学说立足于抗体分子的双重性，既可与特定抗原结合发挥抗体作用，又可借助自身的独特型抗原决定簇引发免疫反应。由此独特型网络强调免疫系统中各个细胞克隆不是处于一种孤立状态，而是通过自我识别、相互刺激和相互制约构成一个动态平衡的网络结构。网络模型有线性模式、环型模式等。

2）aiNet 网络模型

De Castro 提出了一种基于克隆选择原理的免疫网络模型，名为 aiNet 网络。建立该模型的目的是为了研究未标识数据集合的聚类和过滤问题。该网络具有减少冗余、描述数据结构、包括聚类形状等特征。系统内可能的交互作用可以表示为连接图形式。该网络模型描述为：

aiNet 是一个边界加权的图，图中的节点（相当于免疫系统中的细胞）无须全部连接，两个节点之间的连接称为边界，每一个边界都被赋予了一个权重或连接强度。aiNet 网络具有进化性，因为该模型通过克隆选择和体细胞变异来控制网络的动态性和可塑性。它定义了一个连接强度矩阵来度量网络细胞之间的亲和力。aiNet 模型不区分 B 细胞和抗体，采用抗体和抗原之间的空间距离表征它们之间的亲和力，而抗体与抗体之间的亲和力则取决于它们的相似性。

3）多值免疫网络模型

多值免疫网络是一种基于免疫应答原理的网络模型，由 Zhang Tang 于 1997 年提出。在该模型中，首先，抗原被 B 细胞吞噬并出现在 B 细胞的表面，即抗原提呈；抗原提呈细胞被辅助 T 细胞发现，分泌白细胞介素（IL+）激活免疫应答；IL+成为 B 细胞的第二信号，刺激 B 细胞分解为浆细胞，然后合成并分泌抗体；如果抗体增加，抑制 T 细胞受刺激后分泌白细胞介素（IL−），抑制免疫应答。该模型抗原为输入，抗体为输出，输出不由 B 细胞决定，而是由 B 细胞与 T 细胞之间的相互作用决定；抑制 T 细胞在免疫系统中也起到重要作用；辅助 T 细胞与 B 细胞的连接权值作为记忆模式，抗体作为输入模式与记忆模式的误差。该网络模拟免疫系统 B 细胞和 T 细胞相互作用的机制，与生物免疫系统极其类似，这种模型不但具有良好的记忆能力，而且还具有较强的噪声抑制能力。

4）免疫联想记忆模型

在免疫系统消灭抗原后，若下次同样的抗原侵入机体后能够迅速识别该抗原，因而说明免疫系统具有记忆功能。免疫记忆是通过生成永久记忆细胞来实现对抗原记忆的，是一种分布式具有鲁棒性的联想记忆。Abbattista 基于免疫网络的学习和自适应原理提出了免疫联想记忆模型，并将其应用于模式识别。该模型用 n 维空间中的某些特定点来记忆模式，分为学习和回忆两个阶段。学习阶段可以找到代表输入模式的空间中某些特定点，回忆阶段可以在学习得到的模式中找到与输入模式相匹配的模式。

5）PDP 免疫网络

由于免疫系统是一个分布式系统，PDP 免疫网络模型是通过将并行分布式（PDP）理论与免疫学理论结合而提出的一种免疫网络模型。在网络中，输入单元为识别抗原的淋巴细胞，输出单元为分泌识别特异抗原的抗体的浆细胞，而隐层单元为可产生独特型抗体的淋巴细胞，各单元之间的亲和力为连接强度，而推测抗体亲和力的变化是通过学习规则来实现的。由输入单元、输出单元、隐层与各单元的活化规则及相互之间的连接权矩阵构成免疫 PDP

网络。PDP 免疫网络由于固有的分布特性，其信息处理速度非常快且具有很强的错误耐受能力。

6）多 Agent 免疫模型

Epstein，Axtell 于 1996 年提出了一种基于多 Agent 的模型。在该模型中，基于海明形态空间，用定长的二进制字符串表示免疫系统，每一个 Agent 都有一个独特的免疫系统，能够适应一种疾病。如果任何免疫系统的字符串与疾病匹配，认为对该疾病免疫。免疫应答是免疫系统试图修改它的局部结构与所遭遇疾病的匹配过程。该模型使用两种类型规则来控制系统：Agent 免疫应答规则与疾病传送规则。多 Agent 模型特别适合于计算机网络安全与病毒防御的应用。

7）有限资源人工免疫系统（RLAIS）

有限资源人工免疫系统是 Timmis 在 Hunt 与 Cooke 等人提出的人工免疫系统数据分析方法的基础上提出的一个著名网络模型。在免疫系统内，产生有限数目的 B 细胞，因而推断免疫系统内的资源是有限的，这将导致 B 细胞之间的竞争。该模型不再表示个体 B 细胞，引入人工识别球（Artificial Recoginazation Ball，ARB）的概念，每个 ARB 表示同样 B 细胞的集合；ARB 必须为了基于它们的刺激水平而竞争，刺激水平越高，ARB 就可以拥有越多的 B 细胞；如果 ARB 失去所有 B 细胞，则被从网络中清除。RLAIS 允许新模式学习，不影响已经学习到的模式，RLAIS 具有强大的模式识别能力。

3.3 免疫计算学习及优化方法

基于免疫计算智能的学习及优化方法的基本思想是将抗原对应于目标函数和约束条件，抗体对应于搜索空间的解，用抗原与抗体之间的亲和力来对解进行评价和选择。当某种抗体的数量大于某个阈值时，产生抗体的细胞将

发生分化，分成为抑制性细胞和记忆细胞。抑制性细胞抑制这种抗体的进一步增加，记忆细胞将此抗体对应的解记为局部最优解。基于免疫原理的优化算法是研究全局快速优化方法的新尝试。相关算法研究起源于 20 世纪 90 年代初，在这一阶段出现了许多基于免疫原理的智能算法，例如免疫遗传算法、免疫规划算法、克隆选择算法、模式跟踪算法等。从免疫学原理出发的一般免疫算法框架尚未建立，目前 De Castro 提出的克隆选择算法具有代表性，克隆选择算法的出现标志着从免疫学角度开发智能算法的出现，尽管此算法存在着对特定的优化问题效果不理想的现象，但它的出现标志着从免疫系统自身机理出发开发智能优化算法解决优化问题的开始。

从克隆选择原理出发所获得的克隆选择算法能很好地解释抗体应答抗原的部分作用机制。但这种算法未体现抗体之间的作用关系，因而导致算法搜索过程中出现群体多样性不足的现象。然而免疫独特型网络原理刻画了免疫系统中抗体与抗体、抗体与抗原之间的作用关系，这种关系表现为调节机制，其中抗体之间的调节表现为抗体之间的抑制和促进，即亲和力高且浓度低或期望繁殖率高的抗体受到鼓励或促进，反之则受到抑制。抗体的抑制和促进可根据抗体的浓度及亲和力设计随机选择算子进行刻画。

克隆选择原理及免疫独特型网络原理的结合更能合理地反映免疫系统与抗原的作用机制。深入挖掘复杂免疫系统的内在机制，提取有益于构建免疫算法的免疫特征，充分考虑抗体或克隆的选择具有的随机性，同时开发既能合理体现免疫系统原理及机理，又能有效解决工程问题的智能优化算法，是目前人工免疫系统理论及应用领域的重要研究课题。现在已有的基于免疫的方法大致有：基于免疫的智能优化算法、免疫网络算法、模式跟踪算法、阴性选择算法及免疫 Agent 算法等。

本书引入了其对于免疫计算智能算法与人工神经网络及进化算法的综合比较，从单元组成与结构、学习、知识表达与存储、性能等方面进一步阐述其原理及其他计算智能方法的异同，如表 3-2 所示。

表 3-2　进化（遗传）算法、人工神经网络与免疫算法的比较

特　点	免疫算法	进化（遗传）算法	人工神经网络
基本单元	抗体特征字符串	染色体字符串	神经元
结构	分布式或网络结构，结构松散	分布式，结构松散	神经元网络，结构固定
学习	改变抗体浓度及亲和度及抗体记忆实现学习	通过适应度及群体进化实现学习	通过改变神经元间的连接权值实现学习
知识存储	抗体及网络	染色体	神经元连接权值
动态性能	学习/记忆/进化	进化	学习/记忆
动态过程	删除/补充新抗体	删除/补充新染色体	建立/调整网络权值
鲁棒性	种群/网络的个体	种群的个体	网络神经元个体
控制能力	免疫响应，条件及反馈原理	进化运算原理	权值调整
非线性	结合活化函数	神经元激活函数	—

由表可见，免疫计算智能算法依托于免疫原理，具有容噪、泛化和记忆能力，并且通过竞争实现并行分布处理能力，既有一般进化算法的特点又有其特有的优势。但免疫系统自身机理所隐含的丰富思想与工程问题相结合的研究还处于初级阶段。从应用角度来看，在免疫计算智能中，已有的大部分基于免疫机理的算法是针对具体问题而设计的，且免疫算子的设计不具有通用性，因而限制了其应用范围。由于免疫系统是一种极为复杂的自适应系统，其动态机制的模拟及各种机制的组合是一种复杂过程，因此建立免疫算法的一般框架极为困难，导致至今有各种类型的基于免疫的算法。另外，目前免疫计算智能的研究正处于大量开发智能方法的阶段，并且属于应用型的研究方向，这些方法的应用越来越广泛，许多基于免疫的方法已在实际问题中获得广泛而有效的应用。

参考文献

[1] 莫宏伟．人工免疫系统原理与应用．哈尔滨：哈尔滨工业大学出版社，2003：1-47．

[2] 陈慰峰．医学免疫学．3 版．北京：人民卫生出版社，2001：1-18．

[3] 朱锡华．生命的卫士：免疫系统．北京：科学技术文献出版社，1999：12-18．

[4] Forrest S, Perelson A S, Allen L. Self-Nonself Discrimination in a Computer. In: Proceedings of the IEEE Symposium on Research in Security and Privacy, 1994: 202-212.

[5] 丁永生，任立红．人工免疫系统：理论与应用．模式识别与人工智能，2000，13（1）：52-59．

[6] Janis Kuby. Immunology. W H:Freeman and Company, 1994: 123.

[7] Jerne N K. Towards a Network Theory of the Immune System. Annual Immunology, 1974, (125): 373-389.

[8] Perelson A S, Oster G. F. Theoretical Studies of Clonal Selection: Minimal Antibody Repertoire Size and Reliability of Self-Nonself Discrimination. Journal of theory Immune, 1979, 81: 645-670.

[9] Takahashi K, Yanada T. Application of an Immune Feedback Mechanism to Control Systems. JSME International Journal, Series C, 1998,41(2): 184-191.

[10] 李涛．计算机免疫学．电子工业出版社，2004．

[11] 黄席樾，张著洪，何传江，等．现代智能算法理论及应用．科学出版社，2005．

[12] 焦李成，杜海峰，刘芳，等. 免疫优化计算学习与识别. 科学出版社. 2006.

[13] 肖人彬，曹鹏林，刘勇. 工程免疫计算. 科学出版社. 2007.

[14] De Castro L N, Von Zuben F J. Learning and optimization using the clonal selection principle. IEEE Trans on Evolutionary Computation, 2002, 6(1):239-251.

[15] 戚玉涛，刘芳，焦李成. 基于分布式人工免疫算法的数值优化. 电子学报，2009，37（7）：1554-1561.

[16] 薛文涛，吴晓蓓，徐志良. 用于多峰函数优化的免疫粒子群网络算法. 系统工程与电子技术，2009，31（3）：705-709.

[17] 戚玉涛，刘芳，焦李成. 基于信息素模因的免疫克隆选择函数优化. 计算机研究与发展，2008，45（6）：991-997.

[18] 余航，焦李成，公茂果，等. 基于正交试验设计的克隆选择函数优化. 软件学报，2010，21（5）：950-967.

[19] 戚玉涛，焦李成，刘芳. 基于并行人工免疫算法的大规模 TSP 问题求解. 电子学报，2008，36（8）：1552-1558.

[20] Anna_SSwiecicka, Franciszek Seredynski, Albert Y. Zomaya. Multiprocessor Scheduling and Reschedulingwith Use of Cellular Automata and Artificial Immune System Support.IEEE Transactions on Parallel and Distributed Systems, 2006, 17(3): 253-262.

[21] Licheng Jiao, Yangyang Li, Maoguo Gong, etal. Quantum-Inspired Immune Clonal Algorithm for Global Optimization. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics, 2008, 38(5): 1234-1433.

[22] Felipe Campel, Frederico G. Guimarães, Hajime Igarashi. Multiobjective Optimization Using Compromise Programming and an Immune Algorithm. IEEE Transactions on Magnetics, 2008, 44(6): 982-985.

[23] Aldo Canova, Fabio Freschi, Michele Tartaglia. Multiobjective Optimization of Parallel Cable Layout. IEEE Transactions on Magnetics, 2007, 43(10): 3914-3920.

[24] Xiong Hao, Sun Cai-xin. Artificial Immune Network Classification Algorithm for Fault Diagnosis of Power Transformer. IEEE Transactions on Power Delivery, 2007, 22(2): 930-935.

[25] Slavisa Sarafijanovic, Jean-Yves Le Boudec. An Artificial Immune System Approach WithSecondary Response for Misbehavior Detection in Mobile adhoc Networks. IEEE Transactions on Neural Networks, 2005, 16(5): 1076-1087.

[26] Rogerio de Lemos, Jon Timmis, Modupe Ayara, etal. Immune-Inspired Adaptable Error Detectionfor Automated Teller Machines. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2007, 37(5): 873-886.

[27] Famer J D, Packard N H, Perelson A S. The Immune System, Adaptation, and Machine Learning. Physicia D, 1986, (2): 187-204.

[28]Dasgupta D, Forrest S. Artificial immune systems in industrial applications. In: Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials. 1999, (1): 257-267.

[29] Dasgupta D, Attoh Okine N. Immunity based systems: A survey. In: Proceedings of IEEE International Conference on Systems, Man, and Cybernetics. 1997, (1): 369-374.

[30] De Castro L N, Von Zuben F J. The Clonal Selection Algorithm with Engineering Applications[C].In:Proceedings of GECCO'00 Workshop on Artificial Immune Systems and Their Applications, 2000, (1): 36-37.

第 4 章

基于免疫聚类竞争的关联规则挖掘方法

本章导读：

在机器学习的研究领域，关联规则挖掘是数据挖掘的重要任务之一，Agrawal 和 Srikant 提出了著名的 Aprion 算法后，很多学者研究和提出的算法从各个方面对关联规则挖掘进行了改善，但这些方法也不同程度地存在收敛速度慢、正确关联规则的提取率不高等问题。而且，挖掘关联规则需要找出所有满足支持度要求的频繁模式，这将面临极大的搜索空间。人工免疫系统基于生物免疫系统抗体多样性的遗传机理，提供了一种多点、随机的智能搜索技术，系统引入了免疫记忆机制，具有卓越的搜索能力。另外，免疫抗体和抗原由氨基酸的不同排列组成，而数据记录或模式也由不同的属性值排列组合而成，因此很容易在关联规则挖掘与免疫系统之间建立对应关系。本章将在第 3 章的基础上，将克隆选择

运用于关联规则的提取，从而实现免疫算法与数据挖掘的结合。针对数据挖掘中的关联规则挖掘广度及效率问题，利用免疫抗原与抗体对应于数据原记录和候选模式，在基于克隆选择原理免疫算法的基础上引入了聚类竞争机制，加速抗体亲和力的成熟，提高全局搜索能力。这种机制提高了抗体群的多样性，避免了抗体群被少数亲和力较高的抗体占满，导致那些有可能经过亲和力成熟过程成为全局最优解的抗体受到遏制，从而提高关联规则获取的效能。通过实验可以发现基于免疫关联规则挖掘算法具有收敛速度快的特点，而且此算法同时具有相当好的全局及局部搜索能力，这样可以得到更多符合条件的关联规则。

4.1 基本概念及问题描述

在数据库的知识发现中，关联规则就是描述在一个事务中物品之间同时出现的规律的知识模式。更确切地说，关联规则通过量化的数字描述物品甲的出现对物品乙的出现有多大的影响。现实中，这样的例子很多，例如超市利用前端收款机收集存储了大量的售货数据，这些数据是一条条的购买事务记录，每条记录存储了事务处理时间，以及顾客购买的物品、物品的数量及金额等。这些数据中常常隐含如下形式的关联规则：在购买笔记本的顾客当中，有 75%的人同时购买了笔。这些关联规则很有价值，管理人员可以根据这些关联规则更好地进行规划，如把笔记本和笔摆放在一起能够促进销售。有些数据不像售货数据那样很容易就能看出一个事务是许多物品的集合，但稍微转换一下思考角度，仍然可以像售货数据一样处理。比如人寿保险，一份保单就是一个事务。保险公司在接受保险前，往往需要记录投保人详尽的信息，有时还要到医院做身体检查。保单上记录有投保人的年龄、性别、健康状况、工作单位、工作地址、工资水平等，这些投保人的个人信息就可以被看作事务中的物品。通过分析这些数据，可以得到类似以下这样的关联规

则：在年龄在 40 岁以上、工作在 A 区的投保人当中，有 45%的人曾经向保险公司索赔过。在这条规则中，“年龄在 40 岁以上”是物品甲，“工作在 A 区”是物品乙，“向保险公司索赔过”则是物品丙。可以看出，A 区可能污染比较严重、环境比较差，导致工作在该区的人健康状况不好，索赔率也相对比较高。

为了不失一般性，设 $I=\{i_1,i_2,\cdots,i_n\}$ 是项的集合，与任务相关的数据 D 是一组事务集，其中每个事务 T 也是项集，显然满足 $T\subseteq I$ 。设 A 是一个项集，称事务 T 包含 A，当且仅当 $A\subseteq T$，关联规则是 $A\Rightarrow B$ 形式的一种蕴涵，其中 $A\subset I$， $B\subset I$ 且 $A\bigcap B=\Phi$，则有如下定义：

定义 4.1 称关联规则 $A\Rightarrow B$ 在 D 中具有大小为 s 的支持度。如果 D 中事务包含 $A\bigcup B$ 的百分比为 s，即 $\text{support}(A\Rightarrow B)=P(A\bigcup B)$ ；

定义 4.2 称规则 $A\Rightarrow B$ 在 D 中具有大小为 c 的置信度，如果 D 中包含 A 的事务同时也包含 B 的百分比是 c，即 $\text{confidence}(A\Rightarrow B)=P(A\mid B)$ 。

置信度是对关联规则的准确度的衡量，支持度是对关联规则重要性的衡量。支持度说明了这条规则在所有事务中有多大的代表性，显然，支持度越大，关联规则越重要。有些关联规则置信度虽然很高，但支持度却很低，说明该关联规则实用的机会很小，因此也不重要。

设 D 中有 e%的事务支持物品集 B，则 e%称为关联规则 $A\rightarrow B$ 的期望可信度。期望可信度描述了在没有任何条件影响时，物品集 B 在所有事务中出现的概率。如果某天共有 100 个顾客到商场购买物品，其中有 20 个顾客购买了笔记本，则上述关联规则的期望可信度就是 20%。作用度是置信度与期望可信度的比值，描述物品集 A 的出现对物品集 B 的出现有多大的影响。因为物品集 B 在所有事务中出现的概率是期望可信度；而物品集 B 在有物品集 A 出现的事务中出现的概率是可信度，通过可信度与期望可信度的比值反映了在加入“物品集 A 出现”的这个条件后，物品集 B 的出现概率发生了多大的变化。

期望可信度描述了在没有物品集 A 的作用下，物品集 B 本身的支持度。作用度描述了物品集 A 对物品集 B 的影响力的大小。作用度越大，说明物品集 B 受物品集 A 的影响越大。一般情况下，有用的关联规则的作用度都应该大于 1，只有关联规则的可信度大于期望可信度，才说明 A 的出现对 B 的出现有促进作用，也说明了它们之间某种程度的相关性，如果作用度不大于 1，则此关联规则也就没有意义了。

在关联规则的四个属性中，支持度和置信度能够比较直接地形容关联规则的性质。从关联规则的定义可以看出，任意给出事务中的两个物品集都存在关联规则，只不过属性值有所不同。如果不考虑关联规则的支持度和置信度，那么在事务数据库中可以发现无穷多的关联规则。事实上，人们一般只对满足一定的支持度和可信度的关联规则感兴趣。因此，为了挖掘出有意义的关联规则，需要给定两个阀值：最小支持度（minsurp）和最小置信度（minconf）。同时满足这两个阀值的规则称为强规则。如果项集满足最小支持度，即项集出现的频数大于或等于最小支持度与 D 中事务总数的乘积，则称它为频繁集。

在关联规则的挖掘中对数据、业务规则的理解并选取合适的编码尤为重要，数据挖掘工具能够发现满足条件的关联规则，但它不能判定关联规则的实际意义。在发现的关联规则中，可能有两个主观上认为没有多大关系的物品，它们的关联规则支持度和可信度却很高，这就需要根据业务知识、经验，从各个角度判断这是一个偶然现象还是有其内在的合理性。反之，可能有主观上认为关系密切的物品，结果却显示它们之间相关性不强。只有很好地理解关联规则，才能去其糟粕、取其精华，充分发挥关联规则的价值。

因此关联规则的挖掘通常分两步：第一步是找出所有的频繁集；第二步是由频繁集生成强关联规则。第一步尤为关键，决定着整个算法的性能，因为潜在频繁集的数量与项的总数呈指数关系，这将需要遍历一个极大的搜索空间。因此免疫克隆选择实现了一种多点和随机的搜索策略，为关联规则挖掘问题提供了一种新颖的解决方法。本节基于免疫克隆选择机制，建立免疫

聚类关联规则挖掘算法（Immune Cluster Association Rule Mining，ICARM），使得抗体群中亲和力高、满足支持度条件的抗体能够进入克隆选择扩增的过程，在选择亲和力较高的抗体的同时，抑制与它类似或相同的抗体进入克隆扩增机制，从而提高获取关联规则的速度及准确度。

4.2 数据表达及初始化

由于原始记录和候选模式都可以视为由基因构成的染色体，因此将数据库中的记录作为抗原，候选模式作为识别抗体。通过抗体与抗原的比较，可以得到它们的相似程度和包含关系。向每个抗体呈递所有的抗原，与较多抗原匹配程度高的抗体将获得更大的存活和克隆变异的机会。免疫学习的过程既是个体亲和度提高的过程，也是频繁模式生成并通过免疫记忆机制得以保存的过程。而且，强关联规则由记忆细胞所代表的频繁模式生成。免疫克隆算法依靠编码来实现与问题本身无关的搜索，可以把相关属性用相应的整数值代替，从而产生种群数目为n的初始抗体。

在数据挖掘中待处理的每条记录被视为一个抗原。搜索过程中生成的候选模式则可用抗体来代表，抗体除包含候选模式信息外，还包含三项重要信息：B 细胞数、刺激阈值和支持度。算法执行过程中通过聚类竞争产生的较优个体将以免疫记忆的形式得到保存，记忆细胞包含两项：抗体编码和支持度。

由于算法须频繁地对个体施加匹配、变异等操作，因此需要用一种直接的数字化形式来表示记录或模式。这里采用十进制编码，将每个具体的属性取值用整型的属性值编号代替（之前所有连续型变量都已被离散化）。设$W=\{W_1,W_2,\cdots,W_n\}$是属性的有限集合，$V=\{V_1,V_2,\cdots,V_n\}$（V_i是属性W_i的值域）是属性的值域集，那么属性W_i的编码取值为 0～V_i，取值为 0 时表示该属性与其他属性无关联。这样，记录或模式将被表示成属性编码的序列。定

义初始抗体群 $\boldsymbol{Ab}$，规模为 N，每个抗体的编码长度为 L，可定义所有抗体组成的形态空间为 $\boldsymbol{S}$，那么 $\boldsymbol{Ab} \in \boldsymbol{S}^{N \times L}$。

4.3 免疫关联规则挖掘

针对 4.1 节描述的关联规则挖掘问题，通过本节设计的算法主要从以下几个方面寻求问题的解决：

（1）采用优良种群保持策略，通过聚类将抗体依相似浓度划分为多个子种群空间，构建各个小局域中优秀抗体能够快速克隆扩增。聚类族中对最优个体的保持，一个类族代表搜索域中一个小局域，通过对其区间内抗体实施克隆选择操作，增强每个小搜索局域中的优秀个体独立获得克隆扩增及亲和力成熟的机会，从而提高抗体群分布的多样性。

（2）采用基于实数编码的混合变异算子，保持高斯变异概率及柯西变异概率的平衡，维持抗体变异的全局及局部特性，避免具有不同高度的峰值点被吞并。

（3）通过抗体补充，尽可能地在进化的整个过程中维持群体的多样性，使得算法最终能定位于尽可能多的峰值点。

4.3.1 抗体聚类与竞争克隆

在抗体选择与克隆过程中实施聚类竞争的克隆选择机制，使得抗体群中亲和力高且浓度低的抗体能够进入克隆选择扩增的过程，在选择亲和力较高的抗体的同时，抑制与它类似或相同的抗体进入克隆扩增，竞争机制设计的依据是个体群（抗体群或外在抗原群）中个体相互激励和抑制的机理。个体的竞争力刻画了个体在群体中的优越程度，而浓度则刻画了群体中相似个体所占的比重。该算法的特点是聚类中个体活跃度反映了个体的浓度和其竞争

力的关系，在鼓励高竞争度个体的同时，抑制高浓度的个体，进而增强了群体的自我调节能力。

定义初始抗体群 $\boldsymbol{Ab}$，规模为 N，每个抗体的编码长度为 L，可定义所有抗体组成的形态空间为 $\boldsymbol{S}$，那么 $\boldsymbol{Ab}\in \boldsymbol{S}^{N\times L}$。编码方式可选择实数编码、二进制编码、序号编码、字符编码。在函数优化问题当中抗原对应优化的目标函数，抗体对应优化函数的可行解。

具体的聚类算法流程如下。

步骤 1：设定总的聚类中心标准数为 M；

步骤 2：根据亲和力计算。在抗体群 Ab 中选择亲和力最高的抗体作为第一个聚类中心，令 $C=1$；

步骤 3：计算余下的 $N-C$ 个抗体（$N\neq M$）到 C 个聚类中心的距离和，选取 C 个聚类中心的距离和最大的抗体（浓度值最小）作为下一个新的聚类中心，并令 $C=C+1$；

步骤 4：判断 C 是否等于 M，如果不等于回到步骤 3，否则进入步骤 5；

步骤 5：计算余下的 $N-M$ 个抗体（非聚类中心）到 M 个聚类中心之间的距离和，按照到聚类中心距离最小的原则，将抗体归入到与之距离最近的聚类中心所代表的聚类；

步骤 6：将抗体按照所在的聚类进行编号，并计算出每个聚类所包含的抗体数量 D_i（$i=1,2,\cdots,M$）。

聚类完成后，得到规模为 N 的聚类抗体群 $\boldsymbol{Ab}_t$，这 N 个抗体分别属于 M 个聚类，即：

$$\boldsymbol{Ab}_t=\{\boldsymbol{Ab}_{t1},\boldsymbol{Ab}_{t2},\cdots,\boldsymbol{Ab}_{tM}\} \tag{4.1}$$

其中，$\boldsymbol{Ab}_{ti}$（$i=1,2,\cdots,M$）表示第 i 个聚类抗体群，其包含的抗体数量为 D_i。

在每个聚类中引入竞争克隆，首先选出聚类中亲和力最高的抗体，

再取出代表聚类中心的抗体，组成规模为 T（$T=2M$）的优秀抗体群 Ab_e，即：

$$\boldsymbol{A}b_e=\{\boldsymbol{A}b'_{e1},\boldsymbol{A}b'_{e2},\cdots,\boldsymbol{A}b'_{eM},\boldsymbol{A}b''_{e1},\boldsymbol{A}b''_{e2},\cdots,\boldsymbol{A}b''_{eM}\} \tag{4.2}$$

其中，$\boldsymbol{A}b_e\in S^{T\times L}$；$\boldsymbol{A}b'_{ei}$ 为第 i 个聚类中亲和力最高的抗体；$\boldsymbol{A}b''_{ei}$ 为第 i 个聚类的中心；$i=1,2,\cdots,M$。

将经过聚类竞争的这 T 个优秀抗体按照式（4.3）的克隆规模函数进行克隆扩增。

$$N_i=\text{round}\left(N\times\text{aff}(Ab_i)\Big/\sum_{j=1}^{T}\text{aff}(Ab_j)\right) \tag{4.3}$$

其中，N_i 为第 i 个抗体的克隆规模，round(·) 为取整函数，且 $i=1,2,\cdots,T$。$N_c=\sum_{i=1}^{T}N_i$。其中，N_c 为克隆后的抗体数量，可知 $N_c\approx N$。故亲和力越高的抗体，克隆规模越大，克隆复制出的相同抗体就越多。经过克隆后，原来的 T 个优秀抗体就分别扩张成为了 T 个规模分别为 N_i（$i=1,2,\cdots,T$）的小抗体群。则优秀抗体群 $\boldsymbol{A}b_e$ 变为了规模为 N_c 的克隆抗体群 $\boldsymbol{A}b_c$，即：$\boldsymbol{A}b_c=\{\boldsymbol{A}b_{c1},\boldsymbol{A}b_{c2},\cdots,\boldsymbol{A}b_{cT}\}$，其中 $\boldsymbol{A}b_{ci}$（$i=1,2,\cdots,T$）表示 $\boldsymbol{A}b_e$ 中第 i 个抗体 Ab_{ei}（$i=1,2,\cdots,T$）经过克隆复制后形成的小抗体群，其包含的抗体数量为 N_i。

但是固定的聚类半径不利于搜索更多的极值点；过大的聚类半径容易导致次极值点被吞并；过小的聚类半径可能放慢收敛的速度。为了克服上述问题，本书拟对聚类半径采用自适应调节机制，该方法的基本思想是：算法在迭代搜索过程中，记住每一次搜索过程的所有聚类中心集合，如果某一聚类中心经过多次搜索性能没有改进，则此时可减小聚类半径，否则需要增加聚类半径。

由上述可知，相似的抗体经过聚类后将会归属于同一个聚类，那么在这些相似的抗体中，只有亲和力最高的抗体和代表聚类中心的抗体才能够被选中并进入克隆扩增及亲和力成熟过程，这样就防止了遗传算法等进化算法中

少数适应值很高的相似个体经选择操作后充满整个种群，而使算法陷入局部最优解的早熟收敛现象发生。一个聚类代表搜索域中的一个小局域，采用聚类竞争选择机制使得每个聚类（每个小搜索局域）中的优秀个体可以获得克隆扩增，实现亲和力成熟的机会，这样将大大提高抗体群分布的多样性，使免疫克隆算法在深度搜索和广度搜索之间取得平衡，有效提高免疫克隆算法开发与探索的能力。

4.3.2 抗体编码及初始化

为了能够将抗体成功地进行适应度评价并确定聚类族中的优良种群，需要计算抗体之间的距离，即通过浓度对克隆扩增实施竞争。

假设每个基因位上采用的字符集大小为 X（采用二进制编码，其字符集就是 $\{0,1\}$，此时 $X=2$），那么整个抗体群 $\boldsymbol{Ab}$（规模为 N，编码长度为 L）中所有抗体的第 j 位基因的信息熵

$$H_j(N)=\sum_{i=1}^{N}-p_{ij}\log_X p_{ij} \tag{4.4}$$

其中，$j\in\{1,2,\ldots,L\}$，p_{ij} 是字符集中某个字符在第 i 个抗体的第 j 个基因上出现的概率。如果在位置 j 上抗体群中所有抗体的字符都相同，那么 $H_j(N)=0$。多样性的平均信息熵

$$H(N)=\frac{1}{L}\sum_{j=1}^{L}H_j(N) \tag{4.5}$$

根据熵的定义可知两个抗体越相似，则它们的平均信息熵越小；两者越不相似，两者的平均信息熵越大。因此可采用两个抗体之间的平均信息熵来定义两个抗体 u 和 v 之间的距离 dis_{uv}。

$$\mathrm{dis}_{uv}=\frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{2}-p_{ij}\log_X p_{ij} \tag{4.6}$$

基于二进制编码的抗体浓度计算中普遍采用抗体群平均信息的概念计算

抗体亲和度和浓度。当抗体群的所有抗体在同一基因座上的等位基因各不相同时，抗体群的平均信息熵最大，抗体的亲和（相似）度最小；同一基因座上的等位基因全部相同时，抗体群的平均信息熵最小，抗体的亲和（相似）度最大。然而，在二进制编码的优化计算中，这种利用信息熵的概念计算抗体（解）亲和（相似）度和抗体浓度的方法存在一定的问题，即存在遗传进化算法中难以避免的“海明悬崖”。例如，对于由抗体 $\boldsymbol{A}b_v=\{011111111\}$ 和抗体 $\boldsymbol{A}b_u=\{10000000\}$ 构成的抗体群，其平均信息熵 $H(2)=0.693147$，亲和度 $\mathrm{aff}(\boldsymbol{A}b)=0.591$，计算结果表明 u 和 v 是两个很不相同的抗体，但是，我们知道在求解优化问题时它们是两个很接近的解，从这个意义上说，抗体 u 和 v 应该是相似的。

为避免这个问题，我们采用实数编码的抗体距离定义。实数编码是连续参数优化间的自然描述，不存在编码和解码过程。在实数编码下，我们将一个实参数向量构成一个抗体，一个实数对应成一个基因，一个实值对应成一个等位基因。使用实数编码有以下优点。

（1）基因的实数编码消除了因编码精度不够，使得搜索空间中具有较优适应值的可能解未能够表示出来的隐患。当对连续参数进行二进制编码时，存在一个潜在的危险，即可能没有足够的精度来将那些具有最好适应值的参数值充分表示出来。在定义区域比较大的情形下，提高二进制编码精度往往需要很长的编码长度。比如，在区域 $[-100,100]$ 上要想取到每一维上 10^{-6} 的精度，就要求相应的二进制位串长至少为 28×30=540，这给保存和运算带来很大的负担。而直接采用实数编码时，表示同样的区域只需精度在 10^{-6} 以上的 30 个实数组成的位串即可。

（2）作用在实数编码基因上的免疫抗体具备了利用连续变量函数具有的渐变性的能力。渐变性表示变量小的变化所引起的对应函数值的变化也是小的。我们所考虑的问题多数为连续函数，都具备这种渐变性，这一特性被证明对某些优化问题是至关重要的。

（3）采用实数编码可以消除“海明悬崖”。如果采用实数编码，根据形态

空间理论，文中采用欧几里德空间距离计算公式，假设抗体 u 的坐标由 $(au_1, au_2, \cdots, au_i)$ 给定，抗体 v 的坐标由 $(av_1, av_2, \cdots, av_i)$ 给定，那么它们之间的距离用式（4.6）表示，且定义了一个抑制阈值 σ_s，当两个抗体之间的欧几里德距离小于这个抑制阈值时，将采取相应的措施实现相似抗体间的抑制。

$$\mathrm{dis}_{uv} = \sqrt{\sum_{i=1}^{L}\left(au_i - av_i\right)^2} \tag{4.7}$$

初始种群中的抗体是随机得到的，从数据集中随机地抽取一条记录，并随机地将若干基因位置为 0，这样就得到一个抗体聚类；若在抗体种群中不存在相同个体，则将其添加到种群中；重复这个操作若干次，直到群的规模满足要求为止。按照到聚类中心距离最小的原则，将抗体归入到与之距离最近的聚类中心所代表的聚类，并将抗体按照所在的聚类进行编号，并计算出每个聚类所包含的抗体数量 D_i（i=1,2,⋯,M），聚类方式完成后，得到规模为 N 的聚类抗体群 $\boldsymbol{A}b_t$，这 N 个抗体分别属于 M 个聚类，即：

$$\boldsymbol{A}b_t = \{\boldsymbol{A}b_{t1}, \boldsymbol{A}b_{t2}, \cdots, \boldsymbol{A}b_{tM}\} \tag{4.8}$$

其中，$\boldsymbol{A}b_{ti}$（$i = 1,2,\cdots,M$）表示第 i 个聚类抗体群，其包含的抗体数量为 D_i。在每个聚类中引入竞争机制，首先选出聚类中亲和力最高的抗体，再取出代表聚类中心的抗体，组成规模为 T（$T = 2M$）的优秀抗体群 $\boldsymbol{A}b_e$，即：

$$\boldsymbol{A}b_e = \{\boldsymbol{A}b'_{e1}, \boldsymbol{A}b'_{e2}, \cdots, \boldsymbol{A}b'_{eM}, \boldsymbol{A}b''_{e1}, \boldsymbol{A}b''_{e2}, \cdots, \boldsymbol{A}b''_{eM}\} \tag{4.9}$$

其中，$\boldsymbol{A}b_e \in \boldsymbol{S}^{T\times L}$，$\boldsymbol{A}b'_{ei}$ 为第 i 个聚类中亲和力最高的抗体，Ab''_{ei} 为第 i 个聚类的中心，$i = 1,2,\cdots,M$，相似的抗体经过聚类后将归属于同一个聚类，那么在这些相似的抗体中，只有亲和力最高的抗体和代表聚类中心的抗体才能够被选中并进入克隆扩增及亲和力成熟过程，一个聚类代表搜索域中一个小局域，采用聚类竞争选择机制使每个聚类中的优秀个体可以获得克隆扩增实现亲和力成熟的机会，这样将大大提高抗体群分布的多样性，使免疫克隆算法在深度搜索和广度搜索之间取得平衡，有效提高关联规则开发与探索的能力。

4.3.3 抗体亲和力定义

在算法执行过程中，支持度大于支持度阈值的优秀个体都将被作为记忆细胞保存下来。这样，记忆细胞所代表的模式都是满足最小支持度要求的模式，可以很容易提取出同时满足最小置信度要求的关联规则。因为反映关联规则性能的支持度和置信度都是统计量，所以此算法以全体训练样本为抗原。可以选用支持度作为筛选条件，以置信度作为抗原-抗体亲和度函数，即 $\mathrm{aff}(\boldsymbol{A}b_i)=C$ ， C 为置信度。

4.3.4 抗体操作

1. 交叉算子

针对抗体的构造和编码方式，采用 OX 交叉算子。OX 交叉算子的实现步骤如下，其说明如图 4-1 所示。

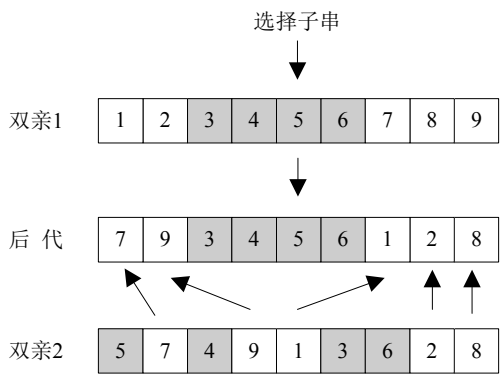

图 4-1 OX 交叉算子的说明

步骤 1：从第一双亲中随机选择一个子串。

步骤 2：将子串复制到第一个空子串的相应位置，产生一个原始后代。

步骤 3：删去第二双亲中子串已有的自然数，得到原始后代需要的其他自然数的顺序。

步骤 4：按照这个自然数顺序，从左到右将这些自然数定位到后代的空缺位置上。

2. 变异算子

因为采用了自然数编码，所以变异方式选择相应的互换变异算子。互换变异指随机选择一个抗体中的两个位置，并将这两个位置上的自然数进行交换。针对单个抗体，产生的基因交换次数 K 是随机的，交换的位置也是随机产生的。例如，图 4-2 中的染色体，假设随机产生的基因交换次数 $K=1$，随机产生的两个位置为第 3 位和第 6 位，则互换变异算子的说明如图 4-2 所示。

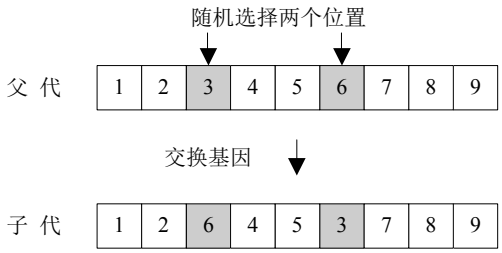

图 4-2 互换变异算子的说明

此外，变异算子还采用了逆转算子，并将逆转算子进行了改进。当逆转算子能够使染色体的适应度增大时，就进行逆转操作，如此反复，直到抗体的适应度不再增加为止，这实质上就是一种局部爬山算法，这样就使爬山算法的局部搜索能力与免疫算法的全局搜索能力得到了有机的结合。逆转算子就是随机地选择染色体中的两个位置，并将这两个位置之间的子串首尾倒置。如图 4-3 中所示的抗体，假设随机产生的两个位置为第 3 位和第 6 位，则逆转算子的说明如图 4-3 所示。

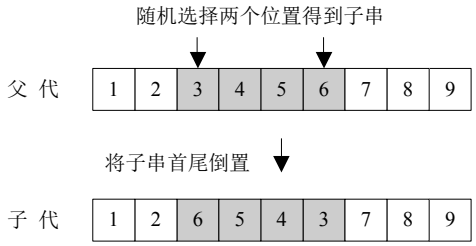

图 4-3 逆转算子的说明

抗体变异是产生高亲和力抗体及多样化抗体的主要环节，是抗体空间到自身的随机映射，$S^N \to S^N$。免疫抗体操作 T_g^C 在单一抗体周围产生一个变异解的群体，利用局部搜索增加提高抗体和抗原亲和度的可能性，这是一般进化算法所不具备的机理。变异保证了新产生的抗体具有充足的多样性，这是免疫系统自适应、自学习特性的重要体现。通过变异过程，使得所建立的算法模型具有了自适应、自学习特性，能更好地适应外界环境的变化。系统通过重复进行上述工作来增强自己识别抗原的能力。

由于抗体针对候选模式编码，执行过程中需频繁地对个体施加匹配、变异等操作。抗体变异算子指单个抗体按照一定的突变概率 P_m，随机选取抗体中的一个或多个点，并将这些点上的基因抽出，再随机插入原抗体的某个位置，形成新的抗体。

4.4 免疫关联规则挖掘方法及分析

算法初始随机产生每一个属性值，以概率（$1-p_i$）选取其他属性值，该值为从 1 到此属性值个数间随机选择的一个整数。当某一个属性对应的选取概率 $p_i=0$ 时，此属性一定存在于所挖掘出的关联规则之中；若 $p_i \neq 0$，则其对应的属性不一定存在于所挖掘出的关联规则之中。所以要挖掘出包含特定属性的关联规则时，应使此属性的选取概率 p_i=0，其余属性的选取概率 p_i 一般取 0.2～0.5。

其算法流程如下。

步骤 1：初始抗体，这些抗体中各个属性的顺序相同且应保证每个抗体满足支持度阈值条件。由此形成最初的抗体种群初始化抗体群 $\boldsymbol{A}b$，随机产生 N 个抗体。

步骤 2：按此前所述方法进行抗体聚类竞争，得到 M 个聚类抗体群，然后选出每个聚类中亲和力最高的抗体和代表每个聚类中心的抗体，得到由 T（$T=2M$）个优秀抗体组成的抗体群 $\boldsymbol{A}b_e$。

步骤 3：将这 T 个抗体按照式（4.3）进行克隆扩增，得到规模为 N_c 的抗体群 $\boldsymbol{A}b_c$。

步骤 4：对 $\boldsymbol{A}b_c$ 中的抗体按照重组变异概率进行重组变异操作后，进行克隆删除操作。以概率 p_m^c 从 $\boldsymbol{A}b_c$ 抽取抗体，对一个或多个属性进行实值变异，删除此种群中不满足支持度条件的抗体，得到规模为 N_c 的抗体群 $\boldsymbol{A}b_m$。

步骤 5：对 $\boldsymbol{A}b_m$ 中抗体免疫选择操作，当抗体与抗体之间的距离小于抑制阈值 σ_s 时，清除亲和力较低不满足支持度条件的抗体，得到抗体群 $\boldsymbol{A}b_d$。

步骤 6：随机产生规模为 N_r 的抗体群 $\boldsymbol{A}b_r$，选出亲和力最高的 N_s（$N_s=N_c-N_d$）个抗体加入到抗体群 $\boldsymbol{A}b_d$ 中，若得到的抗体同时满足最小支持度和最小置信度条件，输出此抗体，并把此抗体还原成原始属性值保留在种群中，留作下一代计算的初始抗体种群。

步骤 7：判断是否满足终止条件，不满足则转至步骤 2 继续执行，满足则结束计算。

算法从一个初始群体出发，不断重复执行免疫克隆聚类、竞争选择、扩增、交叉和变异等过程，使群体进化趋近优化目标。免疫克隆选择算法形同进化算法，属于有限齐次马尔可夫链，并证明了其收敛性。同样基于聚类竞争的免疫克隆选择优化算法的整个聚类竞争及免疫操作过程状态变化均在有限空间中进行，种群序列 $\{A(n),n\geq 0\}$ 是有限的。由于 $A(k+1)=T(A(k))=$

$T_c^C(A(k))T_s^C(A(k))T_g^C(A(k))T_e^C(A(k))T_d^C(A(k))\Theta_i(A(k))$ 均与 n 无关，仅相邻状态 $A(k+1)$ 与 $A(k)$ 有关，可知 $\{A(n),n\geq 1\}$ 是有限齐次马尔可夫链。

在上述算法中初始种群的规模为 n，初始种群中的全部近似解看成状态空间 S 个体，$s_i \in S$ 表示 S 中的第 i 个状态，V_k^i 表示随机变量 V 在第 k 代时所处的状态 s_i。另 $f(x)$ 是 X 上的适应度函数，表示 $s' = \{x \in X \mid f(x) = \max f(x)\}$，则可定义算法的收敛性有：

$$\lim_{k\to\infty} \sum_{s_i \cap s'} p\{A_k^i\} = 1 \tag{4-10}$$

该定义表明，当算法迭代到足够多的次数后，群体中包含全局最佳个体的概率接近 1，称之算法收敛。

设随机过程 $\{A(k)\}$ 的转移概率为 $p_{ij}(k)$，且 $p_{ij}(k) = p\{A_{k+1}^j / A_k^i\} \geq 0$；记 $p\{A_k^i\}$ 为 p_i^k，$p_k = \sum_{i\in I} p_i(k)$，由马尔可夫链的性质可知

$$p_{k+1} = \sum_{s_i \in s} \mid \sum_{j\in I} p_i(k)p_{ij}(k) = \sum_{i\in I}\sum_{j\in I} p_i(k)p_{ij}(k) + \sum_{i\notin I}\sum_{j\notin I} p_i(k)p_{ij}(k)$$

$$\sum_{i\notin I}\sum_{j\notin I} p_i(k)p_{ij}(k) + \sum_{i\notin I}\sum_{j\in I} p_i(k)p_{ij}(k) = \sum_{i\notin I} p_i(k) = p_k$$

$$\therefore \sum_{i\notin I}\sum_{j\notin I} p_i(k)p_{ij}(k) = p_k - \sum_{i\notin I}\sum_{j\in I} p_i(k)p_{ij}(k)$$

$$\therefore 0 \leq p_{k+1} \leq p_k - \sum_{i\notin I}\sum_{j\in I} p_i(k)p_{ij}(k) \leq p_k \leq 1$$

又 $\lim_{k\to\infty} p_k = 0$，则

$$1 \geq \lim_{k\to\infty} \sum_{s_i \cap s'} p_i(k) \geq \lim_{k\to\infty} \sum_{i\in I} p_i(k) = 1 - \lim_{k\to\infty} p_k = 1$$

可证明式（4-10）概率 1 收敛。

由于算法中将抗体通过聚类分为有限个聚类族，并在各个小局域内实现优秀抗体克隆，相对于对整个抗体群进行抗体克隆选择及变异而言降低了运算量。为了不失一般性，令算法的迭代代数为 N，L 为种群克隆的规模。算法复杂度主要是由克隆选择、交叉及变异的复杂度决定 $O(N \cdot L^3)$。在聚类过程

中将种群划分为 M 个类族，操作复杂度为 $O(Q)$，理想情况下，每个族中抗体种群为 D，因此算法时间复杂度为 $O(N \cdot Q \cdot D^3)$，鉴于 $D \ll L$，故算法运算速度较标准克隆选择算法有所提高。

4.5 仿真实验及应用

4.5.1 UCI 数据集仿真实验

为了验证本章给出的关联规则挖掘算法规则提取效率，我们采用 UCI 机器学习数据库最常用的 Iris 数据集。Iris 数据集包含 150 个对象，均匀分布在 3 个目标类中；有 4 个连续的条件属性，其中两个与分类的相关性较强。为适用于算法，先将其离散化并且无缺失值。为便于比较，我们分别采用 C5.0 和 ICARM 来产生规则集规则，如表 4-1 和表 4-2 所示。

表 4-1 基于 C5.0 算法产生的 Iris 规则集

Rule1（覆盖 35 个例子） Petal-Length≤1.9 → class Iris: Setosa
Rule2（覆盖 32 个例子） Petal-Length≥1.9 ∧ Petal-Length≤5 ∧ Petal-Width≤1.6 → class Iris:Versicolor
Rule3（覆盖 29 个例子） Petal-Width > 1.6 → class Iris: Virginica
Rule4（覆盖 28 个例子） Petal-Length>5 → class Iris-Virginica
Default class: Iris: Setosa

表 4-2 基于 ICARM 算法产生的 Iris 规则集

Rule1（覆盖 35 个例子）
Petal-Length≤1.9 → class Iris: Setosa
Rule2（覆盖 32 个例子）
Petal-Length≥1.9 ∧ Petal-Length≤4.89 ∧ Petal-Width≤1.58 → class Iris:Versicolor
Rule3（覆盖 29 个例子）
Petal-Width > 1.6 → class Iris: Virginica
Rule4 （覆盖 28 个例子）
Petal-Length>4.81 → class Iris-Virginica
Default class: Iris: Setosa

由表可知，算法 ICARM 产生与 C5.0 相同的规则形式，而区别在于对于规则 2 的属性 Petal-Length 右边界的值改变为 4.89，Petal-Width 右边界的值变为 1.58；对于规则 4 属性左边界的值变为 4.81。可见 ICARM 算法导致了精度的提高和规则的可理解性。

表 4-3 表明，采用算法优化后的规则精度优于 C5.0 算法，以利于查询和理解。

表 4-3 免疫聚类竞争算法规则提取精度

	C5.0 rule set		ICARM rule set	
	Training set	Testing set	Training set	Testing set
Setosa	25/25	25/25	25/25	25/25
Versicolor	24/25	23/25	24/25	24/25
Virginica	25/25	24/25	25/25	25/25

续表

	C5.0 rule set		ICARM rule set	
	Training set	Testing set	Training set	Testing set
Total Value	74/75 98.67%	72/75 96%	74/75 98.67%	73/75 98.67%

4.5.2 教学质量规则挖掘与分析

为了进一步说明 ICARM 的数据挖掘方法，结合实践过程考查如下应用实例。随机抽取某校教师教学质量评估表样本，并将年龄、职称和评定分数三项输入数据库，忽略其他信息，通过数据挖掘找出其中的关系。表 4-4 给出了部分教学评价信息视图，共有 1800 条记录。

表 4-4 教师教学质量评估表

编　号	年　龄	性　别	职　称	评价等级
13	36	男	副高	中等
50	43	男	副高	良好
76	31	女	中级	良好
99	37	男	副高	优秀
135	41	女	正高	优秀
245	51	男	副高	良好
257	29	男	初级	中等
……	……	……	……	……

在表 4-5 中，年龄是数量属性，将它转换成布尔类型是为了离散化，将年龄分为四个组分别是：C1（21~30）；C2（31~40）；C3（41~50）；C4（51~60）。职称和评价等级是类别属性，须进一步化为布尔类型。根据实际情况对职称、评定等级的范围作了限定，将职称分为：Bl（初级）；B2（中级）；B3（副高）；

B4（正高）；评价等级分为：A1（优秀）；A2（良好）；A3（中等）；A4（差）。

表 4-5 离散后的数据值

A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
1	0	0	0	0	0	1	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0	1	0
0	1	0	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	1	0	0	1	0
0	1	0	0	0	0	1	0	0	0	0	1
0	0	0	0	1	0	0	0	1	0	0	0

如果需要挖掘形如 $A_1 \bigcap A_2 \bigcap \cdots \Rightarrow$ Category 的关联规则，取变异概率为 0.2，交叉概率为 0.8，初始种群 n 选为 100，迭代次数为 200。挖掘评价等级为优秀（excellent）的教师状态信息，设置 minsurp=20；minconf=5；p_i=0 形成初步关联规则：

A: $\begin{aligned} &\text{age}(31 \bigcap 35) \Rightarrow \text{category(excellent)} \\ &[\text{surp} = 27.34\%; \text{conf} = 9] \end{aligned}$

B: $\begin{aligned} &\text{age}(36 \,||\, 39) \Rightarrow \text{category(excellent)} \\ &[\text{surp} = 46.4\%; \text{conf} = 13] \end{aligned}$

C: $\begin{aligned} &\text{age}(36 \,||\, 39) \bigcap \text{certified(senior)} \\ &\Rightarrow \text{category(excellent)} \\ &[\text{surp} = 52.1\%; \text{conf} = 24] \end{aligned}$

为比较其规则提取效能，我们选用标准 Aprion 算法及进化关联规则（EAM）算法进行比较（见表 4-6）。

表 4-6　ICARM 算法规则提取效能比较

算　法	提取的关联规则数	规则提取率	运算时间（s）
Aprion	18	100%	360
EAM	14.7	81.7%	45
ICARM	17.2	95.6%	49

规则提取率为相应算法挖掘出的规则数与总规则数的比例，由表 4-6 可以看出，传统的 Aprion 算法挖掘出的规则数最多，EAM 及 ICARM 算法挖掘出的规则不完全。但 Aprion 算法在挖掘关联规则时运算量随维数增加而加大，ICARM 在规则提取率上优于基于进化算法的关联挖掘，在算法运算量上大大小于标准 Aprion 算法。无论是平均置信度还是最佳置信度，都有明显的优势。

参考文献

[1] Agrawal R, I miclinski T, Swami A. Database mining: A performance perspective. IEEE Trans Knowledge and Data Enginnering, 1993, 5: 914-925.

[2] Agrawal R, Srikant R. Fast algorithm for mining association rules. Proceeding 1994 International conference Very Large Data Bases. Santiago: Chile, 1994: 487-499.

[3] Han Euihong, George K, Kumar V. Scalable parallel data mining for association rules. Proceeding of the ACM SIGMOD97. New York: ACM Press, 1997: 277-288.

[4] B Hetzler, W M Harris, S Harvre, and P Whitney, Visualizing the full spectrum of document relationships. In: Proceedings of the Fifth International

Society for Knowledge Organization Conference, 1998: 168-175.

[5] 刘芳，孙杨军．基于多克隆选择的多维关联规则挖掘算法．复旦学报．2004，43（5）：742-744．

[6] 梁美莲，梁家荣，郭晨．基于人工免疫系统的关联规则挖掘算法．计算机应用，2004，24（8）：50-53．

[7] De Castro L.N, Fetnando J, V.Zuben.Learnng and Optimization Using the Clonal Selection Principle. In:IEEE Trans on Evolutionary Computation, 2002, 6(3): 239-251.

[8] De Castro L N, Von Zuben F J. The Clonal Selection Algorithm with Engineering Applications. In:Proceedings of GECCO'00 Workshop on Artificial Immune Systems and Their Applications, 2000, (1): 36-37.

[9] 焦李成，杜海峰，刘芳，等．免疫优化计算．北京：科学出版社，2006：57-71．

[10] Han J，Kamber M．数据挖掘概念与技术．范明，孟小峰，译．北京：机械工业出版社，2001．

[11] J. Zhou, G. Ding, Y. Guo. Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th ACM Conference on Research and Development in Information Retrieval, Gold Coast, Australia, 2014: 415-424.

[12] G. Ding, Y. Guo, J. Zhou. Collective matrix factorization hashing for multimodal data. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014: 2083-2090.

[13] F. Shen, C. Shen, Q. Shi, A.V.D. Hengel, Z. Tang. Inductive hashing on manifolds. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013: 1562-1569.

[14] L. Li, W. Chu, J. Langford, R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, 2010: 661-670.

[15] W. Li, X. Wang, R. Zhang, Y. Cui, J. Mao, R. Jin. Exploitation and exploration in a performance based contextual advertising system. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D. C., 2010: 26-37.

[16] L. Zhang, R. Jin, C. Chen, J. Bu, X. He. Efficient online learning for large-scale sparse kernel logistic regression. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012: 1219-1225.

[17] A. Daniely, A. Gonen, S. Shalev-Shwartz. Strongly adaptive online learning. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.

[18] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 1958, 65: 386-407.

[19] N. Cesa-Bianchi, G. Lugosi. Prediction, Learning, and Games. Cambridge, UK: Cambridge University Press, 2006.

[20] V. Dani, T. P. Hayes, S. M. Kakade. Stochastic linear optimization under bandit feedback. In: Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland, 2008: 355-366.

[21] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, A. Rakhlin. Stochastic convex optimization with bandit feedback. SIAM Journal on Optimization, 2013, 23(1): 213-240.

[22] A. D. Flaxman, A. T. Kalai, H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In: Proceedings of the

16th Annual ACM-SIAM Symposium on Discrete Algorithms, Vancouver, Canada, 2005: 385-394.

[23] W. Smart, M. Zhang. Applying online gradient descent search to genetic programming for object recognition. In: Proceedings of Australasian Workshop on Data Mining and Web Intelligence, Dunedin, New Zealand, 2004: 133-138.

[24] P. Li, M. Wang, J. Cheng, C. Xu, H. Lu. Spectral hashing with semantically consistent graph for image indexing. IEEE Transactions on Multimedia, 2013, 15(1): 141-152.

[25] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, Y. Zhuang. Sparse multi-modal hashing. IEEE Transactions on Multimedia, 2014, 16(2): 427-439.

[26] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Berlin: Springer, 2009.

[27] S. Shalev-Shwartz, Y. Singer, N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007: 807-814.

[28] S. Shalev-Shwartz. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 2011, 4(2): 107-194.

[29] S. Bubeck, N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning, 2012, 5(1): 1-122.

[30] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the 20th International Conference on Machine Learning, Washington D. C., 2003: 928-936.

[31] E. Hazan, A. Agarwal, S. Kale. Logarithmic regret algorithms for online

convex optimization. Machine Learning, 2007, 69(2-3): 169-192.

[32] H. B. Mcmahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, 2011: 525-533.

[33] L. Zhang, J. Yi, R. Jin, M. Lin, X. He. Online kernel learning with a near optimal sparsity bound. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013: 621-629.

[34] H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 1952, 58(5): 527-535.

[35] L. Zhang, T. Yang, R. Jin, Z.-H. Zhou. Online bandit learning for a special class of non-convex losses. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, 2015,

[36] B. Awerbuch, R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In: Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, 2004: 45-53.

[37] R. Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. Advances in Applied Probability, 1995, 27(4): 1054-1078.

[38] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 2002, 3: 397-422.

[39] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 2003, 32(1): 48-77.

[40] J. Abernethy, E. Hazan, A. Rakhlin. Competing in the dark: An efficient

algorithm for bandit linear optimization. In: Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland, 2008: 263-274.

[41] P. Auer, N. Cesa-Bianchi, P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 2002, 47(2-3): 235-256.

第 5 章

基于小生境免疫粗糙集属性约简方法

本章导读：

国内外智能故障诊断中通常存在着知识获取瓶颈问题，目前多采用机器学习方法来解决。当系统模型和参数信息不准确或难于建模时，传统的状态估计、参数估计方法往往会失效，同时由于数据多元异类且属性繁多，以及现场环境的影响，某些特征属性的获取困难或计算过于复杂，而包含这些特征的属性约简集合不一定是最优选择，因此多个属性约简集合的快速获取具有重要现实意义。同时，高维数据及多维属性约简是机器学习中的一项主要内容。本章提出了一种基于粗糙集及免疫优化的混合机器学习方法。首先将粗糙集核属性中的特征信息引入免疫疫苗的编码，并对初始抗体群接种疫苗，在免疫优化过程中采用小生境及免疫记忆共享机制，促使优良种群的进化，实现类精英保持策略。将属

性集合的分类近似标准作为适应度目标进行优化，通过免疫记忆加速抗体的成熟，在加强全局及局部搜索能力的同时有效保持了该算法快速收敛的特性。通过工业轴承故障诊断及汽车数据仿真及对比实验，证明本书方法能够快速地对数据规则集进行约简，并且能够取得最简约简和约简完备性较好的平衡，为属性约简 NP 难题的解决提供了新的思路，也为工业生产数据挖掘提供更好的解释能力。

5.1 问题描述

目前，工业生产过程的数据使用存在诸多瓶颈：① 存在大量噪声数据、含糊不确定的知识描述，并存在大量定性、定量、半定性半定量的数据形式，关键工艺节点上还可能采集有视频、音频等异构数据，造成信息的多元化。② 由于工业现场环境的影响，很多重要的工业参数和信息难以获取或丢失，再加上仪表的记录误差、数据和分析中的人为因素等，存在数据缺失、不完备等问题。③ 分布式生产过程的数据繁杂、属性多、耦合性强、可读性差，海量信息导致科学计算规模增大，给系统分析、建模和优化带来极大困难。由于分布式生产过程中数据和参数规则的准确性直接关系到专家系统及求解效果，近年来许多学者致力于工业环境下的混合数据表述和属性约简方法研究。有学者针对实际检测数据产生的模糊规则库缺乏良好的完备性问题，提出了从数值数据中提取模糊规则的算法，验证了其逼近能力和对不确定数据干扰的鲁棒性。Ningler 研究了自适应变精度粗糙集方法，为采用粗糙集方法处理不完备信息系统提供了新的技术手段。有学者从属性（集）的可辨识性和不可辨识性出发，提出了面向大规模异类信息的并行知识约简算法。可见，基于粗集、模糊集等软计算技术的数据挖掘算法针对不确定、不完备、多源异类信息的表述及海量数据的规则约简有着显著优势，可用较低的计算成本获得满意的处理结果。Wong S.K.M 和 Ziarko. W 根据属性多元化及属性组合

的爆炸现象，指明了找出数据集的最小约简是一个 NP 难题。从信息论角度将进化算法应用在工业数据的最优属性的约简中是一个可行方法。但为求全局最优解则需要付出很大的代价，存在早熟和退化现象。在很多数据挖掘问题中都会有一些特定的背景知识和特征信息，但进化算法的交叉和变异算子却相对固定，忽视了问题的特征信息对求解问题时的辅助作用，使算法的灵活度较小，特别是在求解一些规模较大的数据问题时就比较明显了。为此，人们提出了混合进化算法来提高搜索过程的整体性能。

目前，在工业数据融合及故障诊断系统中，当系统模型和参数信息不准确或难于建模时，传统的状态估计、参数估计方法往往会失效。本章针对工业生产数据的属性规则多元化、不确定性问题，将粗糙集核属性信息进行免疫疫苗编码，作为特征信息接种到抗体群。而且，在免疫优化过程中引入小生境及免疫记忆共享机制，促使原始抗体群及记忆库中的抗体优良模式的保存，有效地提高抗体的逃逸能力，提高抗体群的多样性及稳定性，实现从不确定规则属性约简集合中较为快速而准确地寻求最优选择和次优选择，其优点是可以根据实际情况来进行优化选择，从而提高粗糙集理论的实际应用能力。

5.2 基本概念及理论

粗糙集理论是 1982 年由 Z. Pawlak 提出的一种描述不完整性和不确定性的数学理论，它从新的角度对知识进行了定义，把知识看作关于论域的划分，从而认为知识是有粒度的，知识的粒度性是造成使用已有知识不能精确地表示某些概念的原因。这就产生了所谓的不精确的“边界”思想。粗糙集理论认为知识就是将对象进行分类的能力。假定我们起初对全域里的元素（对象）具有必要的信息或知识，通过这些知识能够将其划分为不同的类别。若我们对两个元素具有相同的信息，则它们就是不可区分的，即根据已有的信息不

能够将其划分开，显然这是一种等价关系。不可区分关系是粗糙集理论最基本的概念，在此基础上引入了成员关系、上近似和下近似等概念来刻画不精确性与模糊性。

本书采用分类近似质量来衡量属性的分类能力。

定义 5.1 设 $X \subseteq U$ 为论域上的一个子集。条件属性集为 C，决策属性集为 D。X 的下近似 $B_{-}(X)$，$|U|$ 表示集合中元素的个数，$Y_i \subseteq X$，Y_i 为决策属性的类别，则信息系统的分类近似质量为

$$f_B(U,D)=\sum_{i=1}^{n}\left|B_{-}(Y_i)\right|/|U| \tag{5.1}$$

式（5.1）中分类近似质量表示应用条件属性集合能确切进行分类的比例。由于符合分类条件的属性集合非常多，考虑到约简子集的个数随着属性个数的增加呈指数增长，部分子集对约简贡献不大且影响进化的效率，因此对目标进行优化时只寻找那些分类能力出众且属性个数较少的子集。设定优化的亲和力函数为

$$F(B)=f_B(U,D)+[\frac{N-m}{N}] \tag{5.2}$$

其中，N 为所有属性的个数，m 为所选择抗体中所含属性的个数。这样优化求解条件属性集中所有属性值可以在对数据库的一次扫描中同时求出，计算量得以简化。

5.3 属性信息编码及小生境免疫优化

5.3.1 疫苗提取及初始抗体种群

在实际的免疫优化的计算中，初始群体若是接近问题解，将缩短求解时间，提高算法效率。因此在实际操作过程中，首先，需对所求解的问题进行

具体分析，从中提取出最基本的特征信息（疫苗）；其次，对此特征信息进行处理，以将其转化为求解问题的一种方案；最后，将此方案以适当形式转化为免疫算子，以实施具体操作。

疫苗的正确选取会对群体的进化产生积极的推动作用，对算法的运行效率具有重要意义。

免疫疫苗是在原有的进化算法中引入免疫概念和方法，是为了从理论上探讨在处理疑难问题时利用局部信息寻找全局最优解的可行性和有效性，具体而言，它通过局部特征信息以一定的强度干预全局并行的搜索进程，抑制或避免求解过程中的一些重复和无效的工作，以克服原进化策略算法中交叉和变异算子操作的盲目性。算法在执行时，可以有针对性地抑制群体进化过程中出现的一些退化现象，从而使群体适应度相对稳定地提高。另外，免疫进化算法较适于求解一些难度随规模的扩大而迅速增大的问题或 NP 问题。针对这一类问题选取疫苗时既可以根据问题的特征信息来制作免疫疫苗，也可以在具体分析的基础上考虑降低原问题的规模，增设一些局部条件来简化问题。这种简化后的问题求解规律就可作为选取疫苗的一种途径。不过在实际的选取过程中应考虑到：一方面，原问题局域化处理越彻底，局部条件下的求解规律就越明显，这时虽然易于获取疫苗，但寻找所有这种疫苗的计算量会显著增加；另一方面，每一个疫苗都是利用某一局部信息来探求全局最优解的，即估计该解在某一分量上的模式，所以没有必要使每个疫苗做到精确无误。因此，一般可以根据对原问题局域化处理的具体情况，选用目前通用的一些迭代优化算法来提取疫苗。

但是对于某一具体的待求问题，特征信息往往不止一个，这时可随机地选取一种或按照一定的逻辑关系进行组合后再予以考虑。此外，疫苗实质是对最优个体在某一分量上值的估计，疫苗的正确与否有待于其后的选择机制作进一步的判断，即疫苗只会影响算法的搜索效率而不涉及算法的收敛性。免疫算子是由疫苗提取、接种疫苗和免疫选择三个步骤完成的，接种疫苗是为了提高个体的适应度，免疫选择则是为了防止群体的退化。具体叙述如下。

设个体 $x = x_1 x_2 \cdots x_l$。前面已提及，疫苗的提取可以通过利用所求问题的一些特征信息或对问题的先验知识来进行，先验知识可以是最优个体某些分量 x_i 的大概取值范围，也可以是一些分量之间一定的制约关系。

根据核属性定义可知，任何决策表的相对核都具有唯一性。所以以属性核为启发式信息，将其作为疫苗注入抗体编码，对初始群体的选取进行了优化。在进化过程中从各代种群中选出优良个体并提取免疫疫苗，再对后代种群个体接种疫苗。

疫苗不是一个成熟或完整的个体，它仅具备最佳个体局部基因位上的可能特征。疫苗的正确选择会对群体的进化产生推动作用。而且，在某些情况下，由于对所求解的问题一时很难形成较为成熟的先验知识，从而无法正确地提取疫苗或为了提取正确的疫苗而花费大量的工作，使疫苗失去意义。为了提高算法的通用性与应用的便利性，我们可以采用自适应疫苗抽取算法。主要过程如下：通过对 $k-1$ 代保留下来的最优个体群和当前代的最优个体群进行分析，抽取该最优个体 x_1 和 x_2 基因位的共同特点和有效信息用作疫苗。

接种疫苗，是指按照先验知识强制性修改 x 的某些基因位上的基因，使所得个体以较大概率具有更高的适应度值。在该操作中应满足下述条件：

若个体 y 已是最优个体，即 y 的每一个基因位都与最优个体相同，则个体 y 以概率 1 转移为 y_0；若 y 是最差个体，即 y 的每一个基因位都与最优个体不相同，则个体 y 以概率 0 转移为 y_0。

设第 k 代的群体为 $A = \{A_1, A_2, \cdots, A_k\}$，对群体 A 进行接种疫苗，指按照一定比例 $p(0 \leq p \leq 1)$ 随机抽取 $N_P = N \bullet p$ 个个体而进行的一种操作。接种疫苗是利用疫苗确定位上的等位基替代个体相应位上等位基因的操作。接种疫苗加速了优良模式的繁殖，修复了被交叉、变异破坏的优良模式。通过种群与疫苗库相互作用、协同进化，从而极大地提高了其收敛速度。免疫选择对接种了疫苗的个体进行检测，若其适应度不如父代，说明在交叉、变异、接种过程中出现了严重的退化，这时该个体将被父代中所对应的个体所替代，如

果子代适应度优于父代，则子代将代替父代进入下一代种群，免疫选择对算法的收效性将起到决定性作用，它的作用在于加强接种算子的积极作用，消除其负面影响，具有较强的鲁棒性。

5.3.2 抗体编码及接种疫苗

假设有 12 个条件属性 $\{z_1, z_2, \cdots, z_{12}\}$ 的决策表的核属性为 $\{z_3, z_6, z_8\}$，那么我们在初始群体选择时对其编码为“**1**1*1****”，并将其携带的先验信息引入抗体编码。令 $A = a_1, a_2, \cdots, a_l$ 是二进制的抗体编码，在抗体 A 中，a_i 为抗体基因并分为 n 段，每段长为 l_i。接种疫苗过程中采用自适应疫苗抽取算法，通过对 $k-1$ 代保留下来的最优个体群 $\boldsymbol{A}b(k-1)$ 和当前代的最优个体群 $\boldsymbol{A}b_{(k)}$ 进行分析，分别提取两代中最优个体基因位上共同的有效信息用作疫苗。按照先验知识强制性修改抗体群部分基因，使所得个体具有更高的亲和度。算法实现如图 5-1 所示。

Begin
While(condition=true)
统计父代群体，确定 *k*-1 代及 *k* 代最佳个体
$Ab_{\text{best}}^{k} = \text{statistics}(Ab_i^k \mid i = 1, \cdots, n)$；
$Ab_{\text{best}}^{k-1} = \text{statistics}(Ab_i^{k-1} \mid i = 1, \cdots, n)$
分解最佳个体，提取免疫基因信息：
$H = \{h_j = Ab_{\text{best},j}^{k} = Ab_{\text{best},j}^{k-1} \mid j = 1, 2, \cdots, l)$；
Gauss 变异：$Ab_i^k = \text{Mutation}(Ab_i^k)$；
for *i*=1 to *n*
if{ *p* }=True
J=random(*l*);
接种疫苗：$Ab_{H,i}^{k} = \text{Vaccine}(Ab_i^k, Ab_i^{k-1}, h_J)$；
免疫检测：if $Ab_{H,i}^{k} < Ab_i^{k-1}$, then $Ab_i^k = Ab_i^{k-1}$;
else $Ab_i^k = Ab_{H,i}^{k}$；
end；
end；
end；

图 5-1 疫苗自适应提取及接种算法

其中，$\boldsymbol{A}b_{H,i}^{k}$ 为对第 k 代第 i 个抗体 $\boldsymbol{A}b_{i}^{k}$ 接种疫苗后所得抗体；p 为接种疫苗的概率；$\boldsymbol{A}b_{\text{best},j}^{k}$ 表示第 k 代优秀抗体的基因位信息；$\text{Vaccine}(\boldsymbol{A}b_{i}^{k},\boldsymbol{A}b_{i}^{k-1},h_j)$ 表示按模式 h_j 修改个体 $\boldsymbol{A}b_{i}^{k}$ 上基因的接种疫苗操作；J 表示随机生成 1～l 的一个任意正整数；n 和 l 分别为群体规模和个体基因长度。接种疫苗加速了优良模式的繁殖，修复了交叉、变异破坏的优良模式。通过种群与具有先验信息的疫苗相互作用、协同进化，保证了种群的多样性及收敛速度。

5.4 小生境免疫共享机制及免疫算子操作

在生物学上，小生境指特定环境中的组织功能或角色，而且把有共同特性的组织称为物种。自然界的小生境为新物种的形成提供了可能性，是生物界保持近乎无限多样性的根本原因之一。为了使进化算法能够有效地处理优化问题，许多学者将小生境技术引入进化算法。1970 年，Cavicchio 率先在进化算法中引入了基于预选择机制的小生境技术。1987 年，Goldberg 和 Richardson 提出了一种基于共享机制的小生境技术，其基本思想是：解空间中峰周围的子空间中的个体相对独立的生长繁衍。由于小生境技术中将每一代个体划分为若干类，每个类中选出若干适应度较大的个体作为一个类的优秀代表组成一个种群，再在同一种群中或不同种群中实施交叉、变异等操作形成新种群，同时采用预选择机制、排挤机制或分享机制完成选择操作，可以更好地保持解的多样性，同时具有很高的全局寻优能力和收敛速度。Deb 和 Goldberg 提出了两种形式的共享策略，其结果优于排挤模型，并提出一种基于适应度值共享模型的小生境技术。同时，提出共享半径的概念：如果两个个体距离小于一个预定的共享半径，那么就认为它们在同一个小生境中。对于大多数问题，每个小生境的形状和尺寸是不一样的，一个共享半径不适用于所有问题。国内，也有许多人员对小生境技术问题进行了大量研究，提出了基于隔离机制的小生境技术，该方法不仅能够有效地保证群体中解的多样性，而且具有很强的引导进化能力。

本章将这种机理运用于免疫算法，并结合免疫记忆功能，提出了基于免疫记忆共享的优化算法，通过小生境技术将每个子类种群中的优良个体保存于免疫记忆池中，再通过共享选择策略，利用抗体群及免疫记忆抗体群中个体间的相似度的共享函数来调整各个个体的适应度，并分别进行进化，当新一迭代过程开始时，随机从免疫记忆池中选入一定比例的优良抗体注入抗体群进行补充操作，因此在群体进化过程中，算法能够依据调整后的新适应度来进行优良个体的选择操作，以维护群体的多样性及信息共享。

将抗原视为问题（P），抗体 $\boldsymbol{Ab}$ 为此问题的候选解，抗原对抗体 $\boldsymbol{Ab}$ 的亲和力为 $\mathrm{aff}(\boldsymbol{Ab})$，群体规模指定为 N。免疫算子包括克隆选择、亲和突变、抗体补充。另外，利用共享适应度小生境实现方法的思想，构建具体获取多种记忆细胞的共享机制算法，其目的是增强群体多样性及保存优良个体，提高算法搜索性能。

1. 抗体间共享函数

在这种机制中，我们首先定义共享函数，共享函数是将问题空间的多个优良解在空间中区分开来，每个优良解周围接受一定比例数目的个体，同时也是关于两个个体之间关系密切程度的函数，当个体之间的关系比较密切的时候，共享函数值较大（接近于 1），反之则较小（接近于 0）。在免疫算法中，亲和力函数等值于适应值函数。

定义 5.2 对于群体 $P=\{\boldsymbol{Ab}_1,\boldsymbol{Ab}_2,\cdots,\boldsymbol{Ab}_n\}$，令 $H(d_{ij})$ 表示个体 $\boldsymbol{Ab}_i$ 和 $\boldsymbol{Ab}_j$ 间的共享函数，其表达式为

$$H(d_{ij})=\begin{cases}1-\left(\dfrac{d_{ij}}{r_s}\right)^a, & d_{ij}<r_s\\ 0, & d_{ij}\geq r_s\end{cases} \tag{5.3}$$

其中，r_s 为小生境半径，d_{ij} 为个体间距离测度，d_{ij} 采用欧式距离。a 为共享函数调整参数，一般 a 取 2。

定义 5.3 对于群体 $P=\{\boldsymbol{Ab}_1,\boldsymbol{Ab}_2,\cdots,\boldsymbol{Ab}_n\}$，个体 $\boldsymbol{Ab}_i$ 在群体中的适应度为

$$S_i = \sum_{j=1}^{n} H(d_{ij}), i = 1,2,\cdots,n \tag{5.4}$$

定义 5.4 基于个体的共享适应度值的调整方法为：

$$f^*(\boldsymbol{A}b_i) = \frac{f(\boldsymbol{A}b_i)}{S_i}, i = 1,2,\cdots,n \tag{5.5}$$

可见，共享模型是一种特殊的非线性适应值标度变换，其依据是群体中个体间的相似性。该机制限制了群体内特殊抗体的无限制增长，共享半径 r_s 的取值影响算法的搜索性能。

2. 免疫记忆算子

设抗体群 $\boldsymbol{A}(b) = \{\boldsymbol{A}b_1, \boldsymbol{A}b_2, \cdots, \boldsymbol{A}b_n\}$，其亲和力函数表示为 $F(\boldsymbol{A}b)$，由欧几里德距离计算抗体间距离测度。根据 $d_{uv} < r$，确定局部小生境优良种群 A_S，并初始化记忆种群 $\boldsymbol{A}_M$。

$$A_M = \{Ab_{m1}, Ab_{m2}, \cdots, Ab_{mm}\} \tag{5.6}$$

如果 A_S 中包含种群 A 中亲和力最高的抗体，则保留其中一个最高亲和力的抗体，并按 $F^*(\boldsymbol{A}b_i) = \dfrac{F(\boldsymbol{A}b_i)}{S_i}, i = 1,2,\cdots,n$ 更新其余抗体亲和度。按亲和力大小进行各抗体排序，选择前 M 个抗体更新记忆抗体集 $\boldsymbol{A}_M$。

3. 免疫检测算子

免疫检测算子对接种了疫苗的个体进行检测，若其亲和度不如父代，说明在交叉、变异、接种过程中出现了严重的退化，这时该个体将被父代中所对应的个体所替代。如果子代亲和度优于父代，则子代将代替父代进入下一代种群，免疫选择对算法的收效性将起到决定性作用，它的作用在于加强接种算子的积极作用，消除负面影响，具有较强的鲁棒性。利用更新后的亲和度及处罚函数对子种群中浓度高、亲和度低的抗体进行处罚。按亲和力大小进行各抗体排序，选择前 M 个抗体更新记忆抗体集 P_M。

这样，通过免疫记忆及共享适应度机制调整抗体群的亲和度，抑制浓度

高的抗体，选择亲和力成熟的抗体保持或更新抗体群。促使原始抗体群及记忆库中的抗体优良模式的保存。算法在执行过程中，同时对多个子种群进行进化，当某个子种群发现最优解时，将该最优解选入小生境集合的免疫记忆池中，该子种群重新初始化并再次进行进化。在各个子种群中，当个体与任意一个小生境核的距离小于小生境半径 r_s 时，将其适应度缩减为一个较小值，各个子种群按新的亲和力进行进化。

本书通过小生境技术将种群中的优良个体保存于免疫记忆池中，再通过共享选择策略调整各个个体的适应度。当新一迭代过程开始时，随机从免疫记忆池中选入一定比例的优良抗体注入抗体群进行补充操作，并根据新的适应度对优良个体进行选择，以维护群体的多样性及实现信息共享。

5.5 算法执行过程

设本书算法 Immune Memory Sharing Reduction Algorithm（IMSRA）的输入参数（Input）为决策表 $S=(U,R,V,f)$，$R=C\bigcup D$，属性集合为 R，条件属性和决策属性分别为 C 和 D，算法输出（Output）为决策表中的一个约简属性 R。算法执行过程如图 5-2 所示，当适应度不再提高则停止进化，选出适应度最高的 N_s 个个体，得到优化的属性约简 R。

基于小生境免疫进化算法中的变异算子是其主要操作算子，它保证了算法的全局搜索能力，通过对群体内现有信息进行重组来发现与环境更为适应的个体而在进化算法中起着核心作用。它的作用有二：一方面，变异可以保持搜索的全局性，给群体带来新的继承信息，并可维持群体的多样性，防止出现早熟现象，此时，变异概率应取尽可能大的值；另一方面，变异使得进化算法具有局部搜索能力，而此时变异概率应取尽可能小的值，以保持群体的稳定性和收敛性，这是一对不可调和的矛盾。进化算法正是通过交叉和变异这一对算子的相互配合而又相互竞争的操作使其兼顾全局和局部的均衡搜

索能力。但如何有效地配合使用交叉和变异操作仍是目前进化算法的一个重要研究方向和难点。显然，选择算子借用了适者生存这一原则，即个体适应度值越高被选择的机会就越大，由于随机误差的存在，实际运算中选择的结果将会与理论计算的期望值有一定的偏差。为此在进行点的选择时要尽可能使选出的点位于不同的区域（每个区域的点视为同种类）。

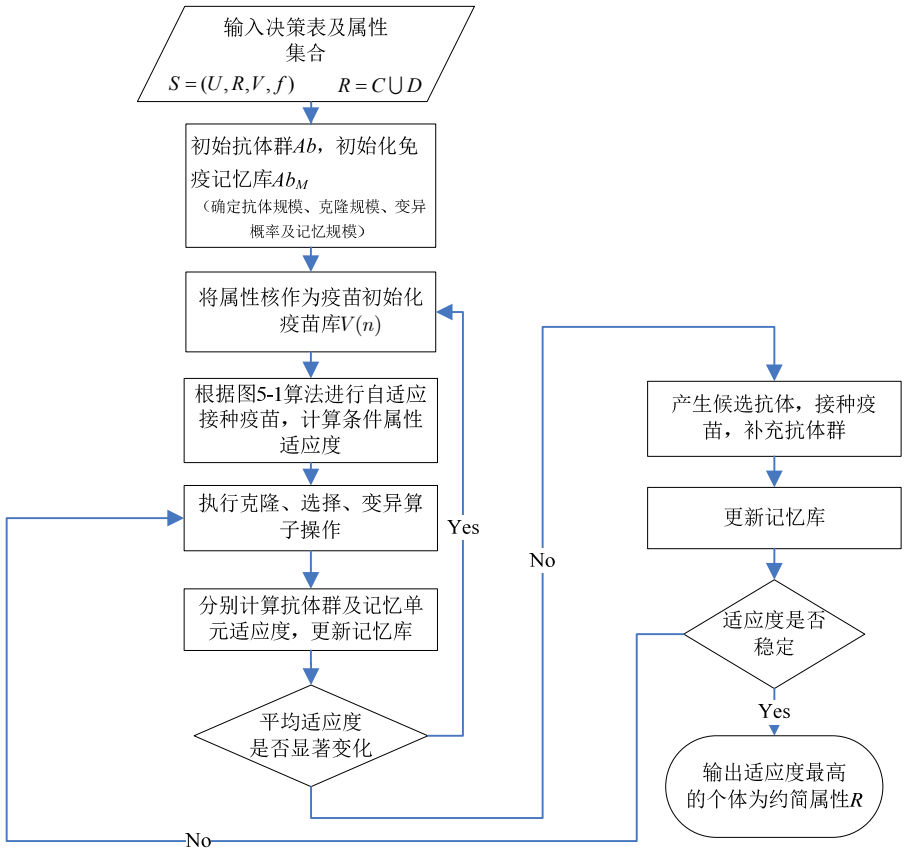

图 5-2　程序流程图

算法协同进化在两个群体上，当某个子种群发现最优解时，将该最优解选入小生境集合并记忆保存。当个体与任意一个小生境核的距离小于小生境半径 r 时，将其适应度缩减为一个较小值，各个子种群按新的亲和力进行进化。

令抗体群规模为 N，记忆库规模为 N_0，克隆抗体规模为 N_c，最大迭代数为 $g_{\max}$，则初始抗体群及记忆单元的时间复杂度分别为 $O(N)$ 和 $O(rN+N_0)$，克隆重组操作时间复杂度为 $O(N_c)$，抗体群及记忆库更新操作时间复杂度为 $O(NN_c^2)$ 及 $O(NN_0^2)$，对记忆单元学习操作时间复杂度最差为 $O(N_0^2)$。因此本书算法总时间复杂度最差为

$$
\begin{aligned}
&O(N)+O(rN+N_0)+g_{\max}[(O(r(N+N_0)^2+O(N_n)+O(NN_n^2)+O(NN_0^2)+O(N_n^2)] \\
&=O(N+rN+N_0)+g_{\max}(r(N+N+N_0)^2+N_n+NN_n^2+NN_0^2+N_n^2) \\
&=O(g_{\max}(r(N+N_0)^2+NN_n^2+NN_0^2+N_n^2)) \\
&=O(g_{\max}(r(N+N_0)^2+(N+1)N_n^2+NN_0^2))
\end{aligned}
$$

5.6 实验仿真及应用

5.6.1 实验一

在工业故障诊断应用中某些特征属性获取困难或计算过于复杂，而包含这些特征的属性约简集合不一定是最优选择。由粗糙集理论可知，属性约简中如果没有结合专业知识，那么优化的单一属性约简集合就有可能与现场的情况存在差别。因此多个属性约简集合的获取是非常有必要的，其优点是可以根据实际情况来进行优化选择，从而提高粗糙集理论的实际应用能力。实验一和实验二分别对本书方法在属性约简完备性及约简效率方面进行了比较分析。分析机械滚动轴承的振动、声学征兆数据集来进行故障诊断，其故障决策信息的获取是一个重要环节。设部分数据集论域 $U=\{1,2,\cdots,16\}$，条件属性集 $C=\{S_1,S_2,\cdots,S_{12}\}$，决策属性 $D=\{0\,|\,1\}$，它包括低频/高频故障及正常两种状态。每组样本利用频谱特征提取了 12 个条件属性和 1 个决策属性，其中条件属性 $S_1,S_2,\cdots,S_{12}$ 分别表示测点振动数据在频谱区间（ 0～0.2、0. 2～0.4、0.4～0.6、0.6～0.8、0.8～1.0、1、2、3、4、5 倍频和大于 5 倍频）中的最大幅值。由于表征频率分量上幅值大小的条件属性连续变化，因此采用了模糊 C

均值聚类方法进行离散化处理，将每个属性离散成高、中、低三种状态，如表 5-1 所示。

算法在 CPU 主频 2.0GHz，内存 2GB 的系统运行环境中，参数初始值选择如表 5-2 所示。运行 30 次取平均值的实验结果如表 5-3 所示。

表 5-1　滚动轴承技术信息离散数据表

U	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	D
1	高	高	低	低	中	低	中	高	中	低	低	高	0
2	高	中	中	中	高	低	中	中	中	高	中	低	0
3	低	高	高	中	中	中	高	低	中	高	中	低	0
4	高	高	高	低	低	高	低	低	高	高	中	高	0
5	高	中	高	低	中	低	高	高	高	低	高	高	1
6	高	高	中	中	中	高	低	低	中	中	高	低	0
7	高	高	中	低	高	低	高	高	低	中	中	低	0
8	中	中	高	高	高	高	高	高	低	高	低	中	0
9	高	高	中	低	高	高	中	低	低	中	高	高	0
10	高	低	中	低	低	高	低	高	低	高	低	低	0
11	高	中	低	高	低	高	高	低	中	中	高	中	0
12	高	低	中	中	低	中	高	低	高	高	高	中	0
13	高	高	低	中	低	中	中	中	中	中	中	高	0
14	高	高	中	高	高	中	高	低	低	高	高	中	1
15	高	高	中	高	中	中	中	中	中	高	高	低	1
16	高	高	中	中	高	低	高	高	高	中	高	中	1

表 5-2　算法参数

参数名	参数值	参数名	参数值
共享函数调整 a	2	共享半径 r	0.8~2.0
克隆规模 n_c	20	补充抗体群 N_r	10
变异概率 p_m	0.1	迭代次数 g	100
抗体规模 N	100	接种概率 p	0.3
记忆规模 n_m	0.5		

表 5-3　属性约简计算结果

迭代次数	最优个体	个体适应度
1	110000111000	0.6321
3	100001110100	0.6521
5	000001111100	0.7344
9	110011000000	0.8542
13	110010000000	0.8741

当属性约简过程中搜索的属性数目庞大时，寻找最小约简是一个十分费时的过程。由表 5-3 可以看到本书算法在第一代中即寻找到具有 5 个条件属性的约简个体，经过进一步迭代，系统在第 13 代即寻找到最小约简$\{S_1\ S_2\ S_5\}$，种群快速收敛并保持稳定，具有较快的全局收敛速度。将本书方法与 Rosetta 软件（http://www.lcb.uu.se/tools/rosetta/）进行属性约简完备性比较，两种方法计算属性最小约简结果一致（见表 5-4）。但本书方法约简更为全面，具有较好的完备性，其所得条件属性更为直观、简单。图 5-3 为种群变化曲线及个体适应度变化。

表 5-4 属性约简完备性比较

本书方法属性项	Rosetta 属性项
$S_1\ S_2\ S_7\ S_8\ S_9$	$S_1\ S_2\ S_5\ S_6\ S_7\ S_8\ S_9\ S_{10}\ S_{11}\ S_{12}$
$S_1\ S_6\ S_7\ S_8\ S_{10}$	$S_1\ S_2\ S_5\ S_6\ S_7\ S_8\ S_9\ S_{10}$
$S_2\ S_6\ S_7\ S_8\ S_9$	$S_2\ S_6\ S_7\ S_8\ S_9$
$S_1\ S_2\ S_5\ S_6$	$S_6\ S_7\ S_8\ S_9\ S_{10}$
$S_1\ S_2\ S_5$	$S_1\ S_2\ S_5$

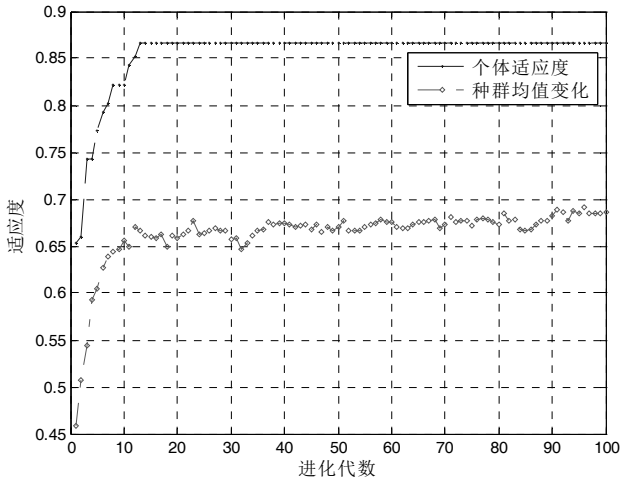

图 5-3 种群变化曲线及个体适应变化

5.6.2 实验二

采用本章文献[30]一个关于汽车的决策表来验证本书提出方法的有效性。其中，论域 $U=\{1,2,\cdots,21\}$，条件属性集为 $C=\{$类型，汽缸，涡轮式，燃料，排气量，压缩率，功率，换挡，质量$\}$，决策属性 $D=\{$里程$\}$。本书对各条件属性按序用 $x_1,x_2,\cdots,x_9$ 来表示。利用本书提出的约简算法对此数据库进行约

简，选择实验参数同实验一。最后种群中的个体趋于统一，适应度不再变化，得到的最小约简集为$\{x_1,x_5,x_9\}$，对应的最小相对约简是{类型，排气量，质量}。图 5-4 为属性约简过程中个体最佳适应度变化情况。从图中可以看出，种群收敛迅速，并最终趋于稳定。图示中显示在第 7 代找到了相对最小约简$\{x_1,x_4,x_5,x_9\}$，并于 15 代左右适应度不再发生变化，得到的最小约简集为$\{x_1,x_5,x_9\}$。

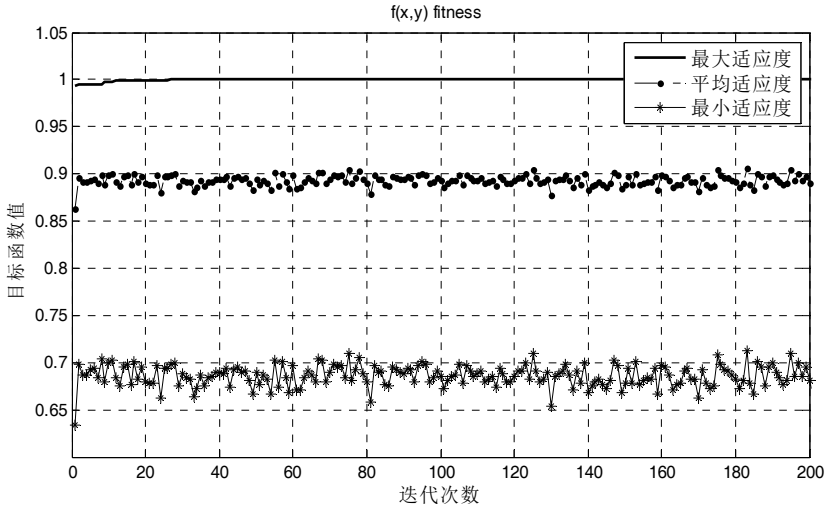

图 5-4　个体适应度变化曲线

为了进一步检验算法的性能，在相同运算条件下运行 50 次，取平均值得到的实验结果如表 5-5 所示，并与本章文献[5]，文献[10]，文献[30]进行典型值比较。

表 5-5　算法实验比较

方　　法	文献[5]	文献[10]	文献[30]	本书方法
100110001	30 代	3 代	9 代	3 代
100010001	—	—	—	15 代
平均代数	30	3		2

从实验对比结果可以看出，四种方法均能寻找到最小相对约简{100110001}，而采用本书方法 50 次运算中最快找到最优个体{100110001}的代数为 3，最慢代数为 5 代，平均 3 代运算即可寻找到约简。文献[5]的方法在近 30 代之后求得最优，而文献[10]的可行域概念的遗传约简方法平均 3 代可求得最小约简，在求解速度上有了显著提高。但文献都没能有效寻找到约简值{100010001}，约简属性的完备性不及本书方法。实验表明，本书提出的算法用于求解工业数据中的相对属性约简是有效的，同时本书中的算法与已有的基于遗传算法和启发式信息为代表的约简算法相比，其算法的编码和适应度函数简单，且利用了免疫疫苗信息作为启发性信息，并采用小生境局部搜索，使得算法快速而有效。RSIA 与 SGA 收敛性比较如图 5-5 所示。

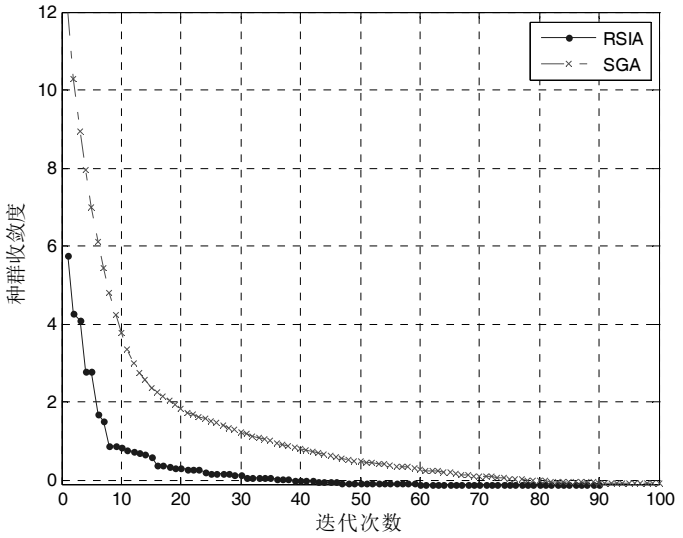

图 5-5　RSIA 与 SGA 收敛性比较

5.6.3　实验三

为验证本书方法对复杂决策表约简的有效性，选择 UCI 数据库中 6 个决策表为实验数据（数据来源：http://kdd.ics.uci.edu/summary.data/），其基本信

息如表 5-6 所示。为统一运行环境，实验选择 Petium4 2.8GHz，内存 512MB 的微机系统，编程工具采用 VC++6.0，算法参数选择如表 5-2 所示，并与文献[39]所介绍的 B、C 两种算法进行典型值比较，运行 10 次的平均结果如表 5-6 所示。

表 5-6　决策表基本信息及约简比较

决策表	U	C	Core	最小约简数目	文献[39]算法 B			文献[39]算法 C			本书算法		
					约简结果	是否含冗余属性	执行时间/ms	约简结果	是否含冗余属性	执行时间/ms	约简结果	是否含冗余属性	执行时间/ms
Patient	90	8	8	8	8	否	1.58	8	否	0.73	8	否	1.22
Monks-1	432	6	3	3	4	是	2.97	3	否	1.32	3	否	1.93
Vote	435	16	7	9	10	是	24.05	9	否	18.12	9	否	16.34
Tic-Tac-Toe	958	9	0	8	8	否	17.90	8	否	23.57	8	否	15.02
Led17	2000	22	14	18	22	是	452.13	18	否	226.86	18	否	187.6
Poker	25010	10	5	7	7	否	902.27	7	否	760.57	7	否	522.45

从表 5-6 中的数据可以看到，本书算法和文献[39]的算法 C 执行效率普遍高于文献[39]算法 B，并且随着数据集的增大，这种优势也越明显。在对本书算法和算法 C 单独比较时可以发现：当约简集与核属性集属性数目比值较小时，本书算法性能基本与算法 C 一致，其原因主要是本书算法在抗体群和记忆抗体群协同优化过程中，接种疫苗及更新记忆库时所耗费的计算量在优化对象规模较小时所占比重偏高。例如，数据集 Patient 和 Monks-1 的约简集和核属性集是相同的，算法 C 在求出核属性集以后就等于约简算法完成了，但本书算法却增加了规则集的寻优过程，而且增加了算法开销。

在论域和属性集较大的情况下，本书算法的优越性要高于算法 B、C。主

要是在处理规模较大的属性约简过程中，本书通过核属性参数对优化过程进行先验信息的引导，算法的编码和适应度函数简单，并通过免疫记忆种群实现精英个体的保留，在提高全局搜索能力的同时提高了抗体群的多样性，也在加强局部及全局搜索能力的同时保持了该算法的快速收敛特性。特别的是，在数据集 Tic-Tac-Toe 约简过程中，算法 C 花费的时间却大于算法 B，这是因为该数据集中核属性数目为 0，约简集中属性均为非核属性，算法 C 的最后阶段需要对约简集中所有属性进行反向消除检查，这必然要花费一些时间，由于 Core(P)为 0，本书算法执行过程中不再产生疫苗库进行接种疫苗，因此执行效率有提高。分析表中属性约简结果可以看出，算法 C 和本书算法的约简结果是完备的，并且获得的属性约简是最小约简，而算法 B 却不一定能保证约简的完备性。

参考文献

[1] Ryszard S. Michalski Ivan Bratko Miroslav Kubat. Machine Learning and Data Mining: Methods and Applications. 2008.

[2] Agrawal R, I miclinski T, Swami A. Database mining: A performance perspective. IEEE Trans Knowledge and Data Enginnering, 1993, 5: 914-925.

[3] 胡清华，于达仁，谢宗霞．基于邻域粒化和粗糙逼近的数值属性约简．软件学报，2008，19（3）：640 - 649．

[4] 杨静，邱菀华. 基于离差的模糊多属性决策法及其应用．系统工程，2008，26（6）：107-110．

[5] 代建华，李元香．粗集中属性约简的一种启发式遗传算法．西安交通大学学报，2002，36（2）：1287-1290．

[6] 梁霖，徐光华．基于克隆选择的粗糙集属性约简方法．西安交通大学学报，2005，39（11）：1231-1235．

[7] 丁卫平，王建东，管致锦．基于量子蛙跳协同进化的粗糙属性快速约简．电子学报，2011，39（11）：2597-2603．

[8] 刘清．Rough 集及 Rough 推理．北京：科学出版社，2001．

[9] 马建敏，张文修，朱朝晖．基于信息量的序信息系统的属性约简．系统工程理论与实践．2010，30（9）：1679-1683．

[10] 李订芳，章文，李贵斌等．基于可行域的遗传约简算法．小型微型计算机系统．2006，27（2）：312-314．

[11] 苗夺谦，周杰，张楠．基于代数方程组的属性约简研究．电子学报，2010，38（5）：1021-2027．

[12] 王熙照，王婷婷，翟俊海．基于样例选取的属性约简算法．计算机研究与发展，2012，49（11）：2305-2310．

[13] 沈艳军，汪秉文．基于实数编码的克隆选择算法及其应用．华中科技大学学报，2004，32（2）：41-42．

[14] Goldberg D E, Richardson J. Genetic algorithm with sharing for multimodal function optimization. Proceedings of the Second International Conference on Genetic Algorithms, 1987: 41-49.

[15] Deb K, Goldberg D E . An investigation of niche and species formation in genetic function optimization. In: Proceedings of the Third International Conference on Genetic Algorithms, 1989: 42-50.

[16] Goldberg D E, Deb K,Horn J. Massive multi-modality, deception, and genetic algorithms. Proceedings of the Second International Conference on Parallel Problem Solving from Nature, Berlin, Springer, 1992(2): 37-46.

[17] 林焰，郝聚民，纪卓尚，等．隔离小生境进化算法研究．系统工程学报，2000，15（1）：86-91．

[18] Agrawal R, Srikant R. Fast algorithm for mining association rules. Proceeding 1994 International conference Very Large Data Bases (VLDB' 94). Santiago: Chile, 1994: 487-499.

[19] Han Euihong, George K, Kumar V. Scalable parallel data mining for association rules. Proceeding of the ACM SIGMOD97. New York: ACM Press, 1997: 277-288.

[20] B Hetzler, W M Harris, S Harvre, and P Whitney, Visualizing the full spectrum of document relationships[A]. In: Proceedings of the Fifth International Society for Knowledge Organization Conference, 1998: 168-175.

[21] 王永富，王殿辉，柴天佑．一个具有完备性和鲁棒性的模糊规则提取算法．自动化学报，2010，36（9）：1337-1343．

[22] Ningler M, Stockmanns G, Schneider G, et al. Adapted variable precision rough set approach for EEG analysis. Artificial Intelligence in Medicine, 2009, 47: 239-261.

[23] 钱进，苗夺谦，张泽华．云计算环境下知识约简算法．计算机学报，2011，36（12）：84-95．

[24] M. Kryszkiewicz. Rough set approach to incomplete information systems. Information Sciences, 1998, 112(1): 39-49.

[25] Ke Liangjun, Feng, Zuren, Ren.Zhigang. An effcient ant colony optimization approach to attribute reduction in rough set theory. Pattern Recognition Letters, 2008, 29: 1351-1357.

[26] 谢宏．程浩忠．牛东晓．基于信息熵的粗糙集连续属性离散化算法．计算机学报，2005，28（9）：1570-1574．

[27] 冯林，王国胤，李天瑞．连续值属性决策表中的知识获取方法．电子学报，2009，31（11）：2433-2438．

[28] 徐雪松，章兢，贺庆．基于疫苗提取及免疫优化的粗糙集属性约简．控制与决策，2008，23（5）：497-502．

[29] 马建敏，张文修，朱朝晖．基于信息量的序信息系统的属性约简．系统工程理论与实践，2010，30（9）：1679-1683．

[30] 孔芝，高利群，王立谦．自适应和声搜索算法及在粗糙集属性约简中的应用．控制与决策，2009，24（10）：1580-1584．

[31] Jiao Licheng, Li,Yangyang .Quantum inspired immune clonal algorithm for global numerical optimization. IEEETransactions on System, Man and Cybernetics, Part B, 2008, 38(5): 1234-1253.

[32] G. Ding, Y. Guo, J. Zhou. Collective matrix factorization hashing for multimodal data. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014: 2083-2090.

[33] F. Shen, C. Shen, Q. Shi, A.V.D. Hengel, Z. Tang. Inductive hashing on manifolds. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013: 1562-1569.

[34] P. Li, M. Wang, J. Cheng, C. Xu, H. Lu. Spectral hashing with semantically consistent graph for image indexing. IEEE Transactions on Multimedia, 2013, 15(1): 141-152.

[35] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, Y. Zhuang. Sparse multi-modal hashing. IEEE Transactions on Multimedia, 2014, 16(2): 427-439.

[36] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Berlin: Springer, 2009.

[37] S. Shalev-Shwartz, Y. Singer, N. Srebro. Pegasos: Primal estimated

sub-gradient solver for SVM. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007: 807-814.

[38] L. Li, W. Chu, J. Langford, R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, 2010: 661-670.

[39] 葛浩，李龙澍，杨传健．基于冲突域的高效属性约简算法．计算机学报，2012，35（2）：250-352．

[40] W. Li, X. Wang, R. Zhang, Y. Cui, J. Mao, R. Jin. Exploitation and exploration in a performance based contextual advertising system. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D. C., 2010: 26-37.

[41] L. Zhang, R. Jin, C. Chen, J. Bu, X. He. Efficient online learning for large-scale sparse kernel logistic regression. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012: 1219-1225.

[42] A. Daniely, A. Gonen, S. Shalev-Shwartz. Strongly adaptive online learning. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.

[43] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 1958, 65: 386-407.

[44] N. Cesa-Bianchi, G. Lugosi. Prediction, Learning, and Games. Cambridge, UK: Cambridge University Press, 2006.

[45] S. Shalev-Shwartz. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 2011, 4(2): 107-194.

[46] S. Bubeck, N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Foundations and Trends in Machine Learning, 2012, 5(1): 1-122.

[47] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In: Proceedings of the 20th International Conference on Machine Learning, Washington D. C., 2003: 928-936.

[48] E. Hazan, A. Agarwal, S. Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 2007, 69(2-3): 169-192.

[49] H. B. Mcmahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, 2011: 525-533.

[50] L. Zhang, J. Yi, R. Jin, M. Lin, X. He. Online kernel learning with a near optimal sparsity bound. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, 2013: 621-629.

[51] H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 1952, 58(5): 527-535.

[52] L. Zhang, T. Yang, R. Jin, Z.-H. Zhou. Online bandit learning for a special class of non-convex losses. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, 2015: 3158-3164.

第 6 章

基于免疫阴性选择的数据分类器

本章导读：

随着网络的迅速发展，网络信息资源已涵盖了社会生活的各个方面，社会进入到信息极为丰富的数字化时代，同时由于文本信息量的快速增长，网络信息过载问题日益突出，促使网络挖掘技术和网络信息检索技术迅速发展。文本分类是处理这些海量数据的一个重要方法，已经成为信息检索、知识挖掘和管理等领域的关键技术。通过自动分类可以将网络文本按照类别信息分别建立相应的数据库，提高中文搜索引擎的查全率和查准率，也可以建立自动的分类信息资源，为用户提供分类信息目录。本章针对文本分类的特点，依靠免疫机理的自体–非自体的检测，介绍了文本分类器的实现机制。针对大样本空间对于连续匹配所带来的计算效率降低问题，提出掩码分段匹配规则，给出了适用于免疫优化的文

本分类规则编码及分类信息评价标准，并通过免疫进化对其进行群体优化，生成更为简洁、便于理解的分类规则集。通过进化学习的思想，实现文本规则优化及压缩，便于分类器更为高效、准确地实现数据集的分类。通过实验数据的测试，并与连续 r 值匹配否定选择分类方法及传统的基于决策树算法、kNN 及贝叶斯方法进行了比较，结果表明该分类方法的可行性，适应于大样本空间文本分类，并在精度及平均信息分上优于传统的分类方法。通过统计显著性检验其方法的有效性，为文本分类研究提出了一种更优化、可靠的方法。

6.1 问题描述

文本分类（Text Categorization）指在给定分类体系下，根据文本内容自动确定文本类别的过程。文本分类的研究可以追溯到 20 世纪 60 年代，早期的文本分类主要基于知识工程（Knowledge Engineering），通过手工定义一些规则来对文本进行分类，这种方法费时、费力，且必须对某一领域有足够的了解才能写出合适的规则。到 20 世纪 90 年代，随着网上在线文本的大量涌现和机器学习的兴起，大规模的文本（包括网页）分类和检索重新引起研究者的兴趣。文本分类系统首先通过在预先分类好的文本集上训练，建立一个判别规则或分类器，从而对未知类别的新样本进行自动归类。大量的结果表明它的分类精度可以和专家手工分类的结果相媲美，并且它的学习不需要专家干预，能适用于任何领域的学习，这使得它成为目前文本分类的主流方法。

总而言之，机器学习理论对于文本分类的研究起了不可低估的作用，在这之前文本分类的研究一度处于低潮，但是文本分类的实际应用和它自身固有的特性给机器学习提出新的挑战，这使得文本分类的研究仍是信息处理领域一个开放的、重要的研究方向。文本分类系统的主要任务是在给定的分类体系下，根据已经掌握的每类若干样本的数据信息，总结出分类规律，建立

判别公式和判别规则，当遇到新样本点时，只需根据总结出的判别公式和判别规则，就能判别该样本点的所属类别。从数学的角度来看，文本分类是一个映射过程，将未标明类别的文本映射到分类体系下已有的类别中，这种映射可以是一对一的映射，也可以是一对多的映射，通常某些文本不但可以与一个类别相关联，也可以与多个类别相关联。网络文本分类是信息检索、知识挖掘和管理等领域的关键技术，文本分类的精确程度取决于特征提取的科学性和分类算法的科学性。近年来，人们结合人工智能的技术研究了各种特征抽取和分类算法，提出了许多模型。D3 算法和 C4.5 算法是较早提出的两个著名的数据分类算法，其基于决策树的操作具有简单实用的数据分类特性，但基于有指导学习的决策树分类方法往往容易导致过度学习或过度拟合问题。朴素贝叶斯分类器和 TAN 分类器以及进而演化提出的进化算法及其一些仿生计算分类方法，大大丰富了数据挖掘的分类方法。随着统计学习理论的成熟，神经网络、支持向量机、AdaBoost 等方法被应用到文本分类器的构建中，并取得了很好的效果。目前关于文本分类算法的研究很多，概括起来主要分为以下几类：① 基于统计的方法，如朴素贝叶斯，kNN、类中心向量、支持向量机、最大熵等方法；② 基于连接的方法，如人工神经网络；③ 基于规则的方法，如决策树等。而目前仿生智能方法在数据挖掘、情报分析、文本分类、时序数据分析等应用方面有进一步的发展。

6.2 基本概念及原理

免疫系统是哺乳动物抵御外来有害物质侵害的防御系统，它在生物体的生命活动中起着至关重要的调节作用。免疫系统能够产生识别“自己-异己”的抗体，利用抗体与抗原的特异性匹配进行特异性检测和识别。借鉴否定选择原理，目前在数据挖掘、故障诊断、文本数据挖掘、时序数据中的异常检测和入侵检测等应用方面有进一步的运用。

否定选择是 T 细胞在胸腺中产生、成熟过程中的一个重要阶段。未成熟 T 细胞在胸腺中与大量的“自己”细胞进行匹配操作，与“自己”细胞匹配的 T 细胞死亡，只有不与任何“自己”细胞匹配的未成熟 T 细胞才最终生长为成熟 T 细胞，此过程称为否定选择过程。在计算机免疫系统中，检测器生成过程采用否定选择过程的称为遵循否定选择原则。否定选择算法基本步骤如下。

（1）定义自己：定义长度为 L 的有限个字母串的类集 S。

（2）随机产生一个长度为 L 的字符串 n，依次与 S 进行匹配。一般采用部分匹配原则，而不是精确或完美匹配。

（3）不断地将 n 与集合 S 进行匹配，一旦发生匹配，则丢弃该字符串并回到步骤（2）。

（4）如果 n 中没有任何位与之匹配，则认为 n 成熟，输出到监测集 R 中。

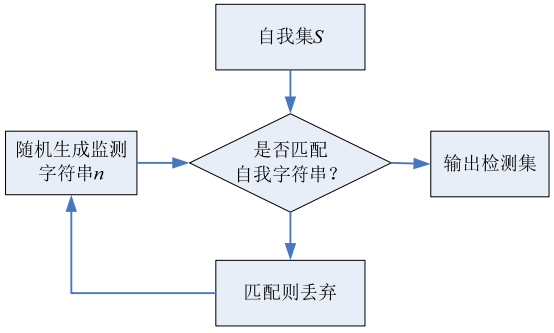

图 6-1　否定选择算法流程图

然而，在生物免疫系统中，抗体和抗原的结合更多地表现出不完全匹配特性，完全匹配只是其中的一个特例，因此部分匹配规则更受关注。通常有许多部分匹配规则，如海明规则、连续 r 位的匹配规则等。在海明规则中，通常设定阈值的大小，以确定两个字符串是否匹配。当两个字符串之间的海明距离大于等于阈值时，称这两个字符串是匹配的，反之则不匹配；在连续 r 位的匹配规则中，需要根据连续匹配的位数来确定两个字符串是否匹配。当连续匹配的位数大于等于 r 值时，两个字符串匹配，否则不匹配。在否定选择算

法中，初始检测器的生成是一个重要的步骤，目前主要有穷举法和基于连续位匹配规则的生成算法。但由于该算法的时间和空间复杂度与匹配阈值呈指数关系，因此它不适用于 l 和 r 较大的情况。

6.3 文本分类规则编码

从数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本集合映射到已有的类别中，该映射可以是一一映射，也可以是一对多的映射，因为通常一篇文本可以同多个类别相关联。用数学公式表示为：$\theta: S \rightarrow C$，其中，S 为待分类的文本集合，C 为分类主题中的类别集合。在文本分类中，训练方法和分类算法是分类系统的核心部分。目前，文本分类的许多研究致力于二元问题，但是有各种各样的文本信息数据源，比如网上新闻、电子邮件、数字字典，都是由不同的主题构成，因此构成了多种分类问题。其一般方法是将此工作分成不同的二元分类问题，对一个新文本进行划分，只需要应用二元分类器及它们的联合形成一个决策。其缺点是忽视了不同类之间的联系。本书对待分类规则编码采用 Holland 的密歇根编码方案，将一条规则对应为一个个体，而分类规则对应一个种群。假定特征属性类型为离散型，如果一个特征属性有 n 个可能的取值，则在二进制值中为其分配 n 位，每一位与特定的取值对应，取 0 表示析取式中没有该取值，取 1 表示析取式中有该取值。对于类别属性，则使用连续二进制进行表示。

6.3.1 个体编码

为了不失一般性，考虑数据集 D 包含两个特征属性的两类情况：$X\{x\text{kind}s1, x\text{kind}s2, x\text{kind}s3, x\text{kind}s4\}$ 和 $Y\{y\text{kind}s1, y\text{kind}s2, y\text{kind}s3\}$，以及四个类别属性“class_$a$”，“class_$b$”“class_$c$”和“class_$d$”，则规则（IF <$x$=$x$kind1 or $x$kind4> AND <y=$y$kind2 or $y$kind3> THEN class =class b）可以表

示为二进制串((1001011)，(01))，而((0110101)，(11))对应规则 (IF <x=$x$kind2 or $x$kind3> AND <y=$y$kind1 *or* $y$kind3> THEN class=class d)。这样编码后每一个分类规则可对应一个免疫抗原，整个群体的进化过程实质是规则前提的进化。

6.3.2 亲和力定义

为了提高分类匹配效果，亲和力函数为分类器精度及分类信息分的线性组合。

1. 分类精度

确保较小的分类错误率是衡量分类算法的一个重要标准，分类正确率判别计算式为

$$f_c = (N_n - N_e)/N_n \tag{6.1}$$

其中，N_n 为数据实例集，N_e 为错误的分类实例数目。

2. 分类信息分

在某些特定条件下，默认的分类器依然可以达到很高的分类精度。因此文中引入信息分概念，来消除类别优先可能性所带来的影响，从而可以充分考虑到分类器适当的似然响应和先验概率产生的影响。

定义 6.1 平均信息分

$$\mathrm{In}f = \frac{\sum_{i=1}^{M} \mathrm{In}f_i}{M} \tag{6.2}$$

其中，M 代表测试实例集，第 i 次测试实例的分类信息分定义为

$$\mathrm{In}f = \begin{cases} -\log_2 P(C_i) + \log_2 P'(C_i), & P'(C_i) \geq P(C_i) \\ -(\log_2(1-P(C_i)) + \log_2(1-P'(C_i))), & P'(C_i) < P(C_i) \end{cases} \tag{6.3}$$

其中，C_i 是第 i 次测试实例的类别，$P(C_i)$ 是其先验概率，$P'(C_i)$ 是分类器返

回的概率。如果正确类返回的概率比先验概率大，则信息分为正值，表明获得的信息是正确的。反之则取信息分为负，表示获取信息错误。分类的精度在某些特殊的实例中的表现会很不一致，而信息分则要稳定得多。我们将其进行线性组合，得到亲和力计算函数为

$$F = f_c + f_{\mathrm{Inf}} \tag{6.4}$$

6.3.3 免疫优化

对编码后的分类规则进行免疫进化的过程包括免疫选择、交叉变异。通过免疫选择对个体进行亲和力检测，清除亲和力较低及不满足支持度条件的抗体。同时进行免疫接种及抗体补充操作，以保持种群的多样性。通过交叉变异随机地把个体中的一个属性值变换为属于同一属性域中的另一个属性值，即一条规则中的某个属性或几个属性对应的边界值与另外一条规则的相应部位进行交叉。通过免疫群体进化算法对抗体群进行全局搜索及压缩，从而输出优化后更为简洁、便于理解的规则集。

6.4 掩码匹配的否定选择分类器

本节通过掩码匹配构造相应记忆库来建立检测分类器，使其具有学习能力。对将进入记忆库的有效检测器与已有的记忆有效检测器进行亲和度计算，如果亲和度大于一定的阈值，则进入记忆有效检测器集，否则剔除，从而生成分类器。

定义 6.2 阈值：采用八位二进制编码，用于编码检测器识别阈值（Recognition Distance Threshold，RDT）。通过实数阈值表示检测器和特征向量之间的匹配程度。

定义 6.3 模式：由 l 个符号组成的符号串，即长度为 l 的二进制串组成。

本书中用U表示所有模式的集合，N表示所有非我模式的集合，简称“非我集”；S表示所有自我模式集合，简称“自我集”。

定义 6.4 掩码：由长度为l的二进制串组成的符号串。

掩码具有与模式特征向量一样的位数。掩码基因覆盖到模式基因上时，在掩码基因位是 1 的位置，改变模式基因中的对应位为*通配符；掩码基因中为 0 的位置的模式基因对应位不变。

定义 6.5 匹配：在一定的匹配规则下，两个模式串a和b的相似程度超过匹配阈值，则称a和b匹配，记为$\mathrm{Match}(a,b)$。

定义 6.6 匹配强度：设定两个模式串a和b，采用二者对应的模式进行匹配，匹配强度必须比检测器的阈值更大，才认为发生匹配。匹配强度= 匹配位中非*个数/非*位总数。

每一个检测器用三元组表示，每一个抗体组含有四个基因段，如图 6-2 所示。

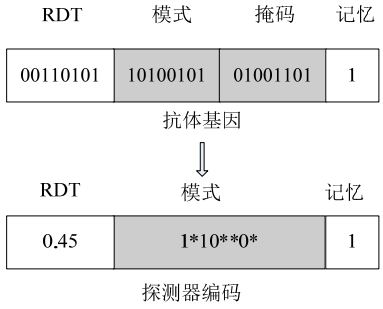

图 6-2　检测器编码生成

第一个染色体 RDT 基因段表示检测器和特征向量之间的匹配程度。第二个和第三个基因段分别为模式和掩码基因。第四个基因段为记忆段，存储检测器状态信息，当发生匹配后修改该记忆段中状态为 1，否则记 0，便于免疫变异后扩增过程中寻找最优解向量。这样检测器模式将染色体的四个基因段转化成三个基因段。类似免疫系统，最终得到的检测器有 0 和 1 两种形式，形式 0 为检测器，分类任何它所匹配的特征向量为非自体；形式 1 则分类任何它所匹配的特征向量为自体。

根据匹配强度计算得到二者的匹配结果为 1/2=0.5，大于阈值 0.45，则将其划分为自体类别，如图 6-3 所示。

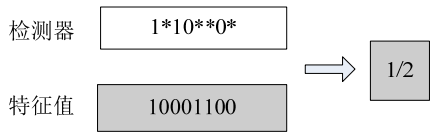

图 6-3　检测器匹配特征向量

6.5　免疫进化分类实现

通过免疫否定选择原理构建分类检测器，结合免疫克隆进化优化对分类规则组成的群体进行操作，在群体进化过程中采用最佳保留技术，保证算法能以概率 1 收敛到全局最优。其实现基本步骤如图 6-4 所示。

步骤 1：按照 6.3.2 节的定义随机在一定范围内产生检测器 T_l；

步骤 2：设定识别自我匹配的阈值和待识别分类的匹配阈值 RDT；

步骤 3：判断检测器和自我集的每一个编码个体是否匹配，如果不匹配转向步骤 1；

步骤 4：判断分类有效检测器集是否已满，如果不满转向步骤 1；如果已满转向步骤 5，并输出分类器 C_T；

步骤 5：判断已有的 C_T 中对分类规则记忆的有效集 M 是否大于等于 1。如果 M>1，计算将要加入记忆有效检测器集的检测器与已有的记忆有效检测器集中每一个检测器的欧氏距离，如果有一个以上小于一定的阈值，则抛弃该检测器转向步骤 3；如果大于一定的阈值，该检测器则加入到分类器的检测器；

步骤 6：根据分类规则编码初始抗体群，这些抗体中各个属性的顺序相同且应保证每个抗体满足阈值条件；

步骤 7：对抗体群进行免疫进化操作，清除亲和力较低不满足支持度条件的抗体，同时进行免疫交叉及抗体补充操作；

步骤 8：通过分类检测 C_T 对优化后的分类规则集进行匹配，高于设定阈值的归为与检测器类别相同的己类，否则划为异类；

步骤 9：循环计算是否满足收敛条件，如果满足则结束进化，输出最优（或较优）的分类规则集 R。

图 6-4　算法流程

这样可通过对一个有限的训练集 T 进行有指导学习而建立一个预测模型，用以描述预定的数据集及数据库 D，即对 D 中的数据进行分类。

6.6 仿真实验及应用

6.6.1 实验一

为了测试本书建立的否定选择分类器的有效性，构造的数据样本为二维二类云分布数据点，共包含 3000 个样本。每个点的类属值并不代表它的类别信息，且是不连续无序分布。数据集中叠加了隐含属性数据，如图 6-5 所示。图 6-6 是基于本书 MMNSC 分类算法对数据集规则划分结果的显示，其中分类器对数据集划分为四个边界区域，中心点显示了分类器的记忆单元。o 表示识别类别为文中描述的 1 类分类器，识别自体的特征值；*表示为 0 类分离器，识别类别为非自体的特征值。图 6-7 显示了免疫进化后分类器对数据集的划分，其中对非自体中隐含的自体属性进行了正确区分，充分表现了免疫系统的自我识别与区分能力，具有较好的噪声数据及模糊信息处理效果。对数据集进行了 20 次随机实验，其正确率在 98%以上，且在免疫进化过程中产生的记忆单元数较少，运行效率较高。

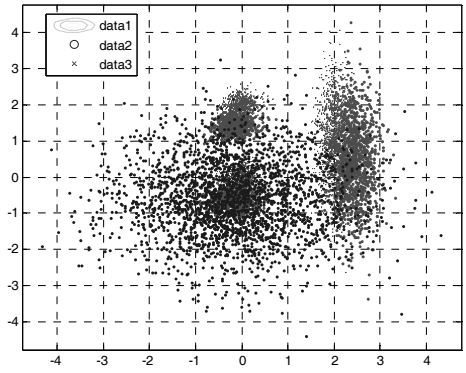

图 6-5 二维二类人造数据集及分布

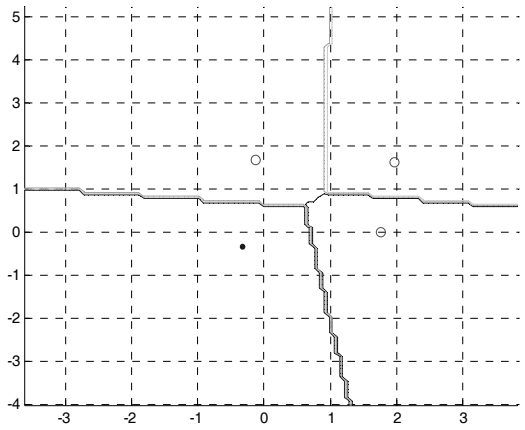

图 6-6　分类器边界区域及免疫记忆单元

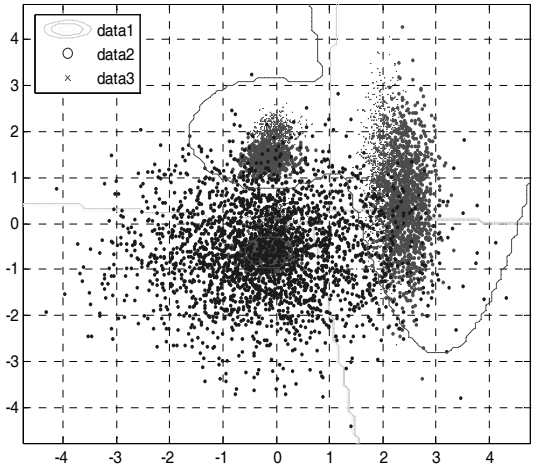

图 6-7　数据分类结果

6.6.2　实验二

本书选择由复旦大学提供的实验数据语料集（http//www.nlp.org.cn），语料集分为十个主题类：环境、计算机、交通、教育、经济、军事、体育、医药、艺术、政治。该语料集具有一定的综合性，具有多归属和类别层次化的

特点。通过该语料集来测试分类器，在一定程度上反映了分类器的实际分类性能。为便于比较，我们在同等条件下（相同硬件及软件运行环境），采用本书算法与连续 r 位的匹配规则的免疫否定选择算法，均选择循环代数为 200，初始抗体群规模 $N=50$，变异概率为 $P_m=0.2$。同时采用 Matlab 编写了传统贝叶斯及 kNN 分类算法，并根据多种群协同优化的分类规则提取流程，采用简单遗传算法和蚁群算法协同优化，分别对该实验数据进行了 20 次随机测试，统计结果如表 6-1 所示。

表 6-1　不同方法下的分类效果比较

类别	本书方法		连续 r 值匹配否定选择		贝叶斯		kNN		文献[3]方法	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
政治	89.56	46.89	83.01	47.11	83.54	49.89	81.06	36.89	90.36	46.89
经济	83.23	57.65	81.78	53.01	76.43	48.62	74.24	47.44	81.25	57.65
环境	87.32	59.56	83.09	55.16	81.52	61.76	84.32	62.56	88.72	59.56
体育	81.98	77.24	80.33	73.03	71.46	67.64	71.98	74.24	85.08	77.24
教育	86.31	78.32	83.23	80.02	71.31	73.52	83.97	63.12	84.00	78.32
交通	89.11	78.10	83.98	81.32	81.34	73.11	83.13	68.10	87.12	78.10
军事	86.28	72.45	82.21	70.43	72.58	77.55	76.68	70.49	88.68	72.45
医药	79.02	72.22	75.20	73.26	69.12	67.12	67.12	67.22	80.12	72.22
艺术	86.31	80.43	84.22	83.09	76.35	74.43	74,89	66.48	85.90	80.43
计算机	79.73	82.33	74.93	80.59	70.03	56.36	75.33	67.72	82.41	82.33

根据以上实验结果可以看出，本书方法在各样本集的分类精度均优于 r

值连续匹配免疫分类方法。在召回率上，连续 *r* 值匹配否定选择方法在教育、交通、医药及艺术四个类别集中的召回率高于本书方法。经分析可知，该四类数据样本集偏小，两分类方法分类效果差异不明显，但当政治、计算机等数据样本集大时，时间和空间复杂度与匹配阈值呈指数关系，本书方法在分类效率及效果上优于传统否定选择分类方法。本书方法与协同进化方法，对于政治、计算机等大样本集的分类效果明显，但由于协同进化方法引入信息熵等先验知识，在具体实验运行中所抽取规则的数量少，普遍精确度高，平均长度短。本书方法较贝叶斯及 kNN 方法的精度及召回率要高一些，这是由于本书算法采用免疫掩码匹配方式使错误分类的样本尽量少，文中引入信息分概念，来消除类别优先可能性所带来的影响，从而可以充分考虑到分类器适当的似然响应和先念概率产生的影响。

我们用统计显著性检验实验分类精度结果的显著性。在 SPSS13.0 软件中，采用了统计分析中的方差分析。设原假设 H0：贝叶斯、kNN、决策树、本书方法所得结论在统计上没有差异。备择假设 H1：贝叶斯、kNN、决策树、本书方法所得结论在统计上存在差异。在进行方差分析之前对数据进行方差齐性检验，可知数据满足方差分析的条件。表 6-2 为对不同算法分类精度采用 SPSS 输出的 ANOVA 方差分析表。

表 6-2 ANOVA 方差分析表

Precision	平方和	df	均方	F	显著性
组间	708.938	4	177.235	8.267	0.000
组内	964.746	45	21.439		
总数	1673.68	49			

由表 6-2 可知 F 值为 8.267，显著性为 0.000，在统计上是非常显著的，因此拒绝方差分析的原假设，认为本书方法、连续 *r* 值匹配否定选择、贝叶斯、kNN、文献[1]方法所得结论在统计上存在差异。为了更加明显地得到哪些方

法之间的差异是显著的，选择了方差分析中的 LSD（最小显著性差异）方法比较本书方法与其他方法之间的差异，从表 6-3 可以看出，在显著性水平为 0.1 时，本书方法与连续 r 值匹配否定选择、贝叶斯、kNN 之间的差异是显著的，仅与文献[3]方法的差异不显著。因为两者同样采用了进化优化策略，对大样本空间的分类效果显著，具有一定的优越性。

表 6-3　不同算法分类精度统计显著检验比较

方法（I）	比较方法（J）	均值差（I-J）	标准误差	显著性	90% 置信区间	
					下限	上限
本书方法	连续 r 值匹配否定选择	3.68700*	2.07069	.082	.2094	7.1646
	贝叶斯	9.51700*	2.07069	.000	6.0394	12.9946
	kNN	7.39722*	2.12743	.001	3.8244	10.9701
	文献[3]方法	.43318	2.02308	.831	-2.9644	3.8308

*. 均值差的显著性水平为 0.1。

为进一步验证本书分类方法的有效性，选择了四个斯洛文尼亚卢布尔雅拉中心医科大学的医疗诊断文本数据源对其进行测试，并将其与传统基于决策树（Assistant R）算法、朴素贝叶斯及 K 最近邻算法进行结果比较。这些诊断数据主题分别包括确定转移型病人主要肿瘤位置问题（PRIM）、预测肿瘤移植手术后 5 年内乳癌复发率问题（BREA）、确定淋巴疾病类型问题（LYMP）、风湿病诊断（RHEU）。

表 6-4　医疗数据集的基本特征

类别名称	类别个数	属性个数	属性平均值	实例个数	主要类别比例	熵
PRIM	22	17	2.2	339	25	3.89
BREA	2	10	2.7	288	80	0.73

续表

类别名称	类别个数	属性个数	属性平均值	实例个数	主要类别比例	熵
LYMP	4	18	3.3	148	55	1.28
RHEU	6	32	9.1	355	66	1.73

实验中使用的数据的特征总结在表 6-4 中（分别为 PRIM，BREA，LYMP，RHEU）。诊断的类别和熵显示了诊断的难度。属性个数可近似反映病人病情好坏的程度。主要类别比例近似等于大多数可能诊断中的优先诊断的可能性。实例个数表明数据样本集的多少。其中类别表明该数据样本在数据库中的编号。

实验过程中随机地从实例中选择 70%用来学习，30%用来测试，表 6-5 表明不同分类方法所体现的分类平均信息分。表 6-6 中所列的结果是 10 次运行的精度平均结果及召回率。在分类评价中平均准确度随着平均每个答案的信息分给出来。信息分能消除类别优先可能性所带来的影响的性能度量。

在分类规则进化过程中，选择最大进化代数 100，初始抗体群规模 $N=50$。整个抗体的编码长度 $L=44$，候选抗体群数量 $N_r=20$，交叉概率 $P_c=0.8$，变异概率 $P_m=0.3$。

表 6-5 医疗数据集分类系统的平均信息分

类别	贝叶斯	kNN	决策树（AssistantR）	本书方法
PRIM	1.61 ± .14	1.15 ± .11	1.07 ± .11	1.45 ± .14
BREA	0.06 ± .04	0.02 ± .02	0.05 ± .06	0.05 ± .03
LYMP	0.78 ± .08	0.53 ± .08	0.61 ± .09	0.73 ± .06
RHEU	0.53 ± .06	0.43 ± .05	0.41 ± .08	0.49 ± .06

表 6-6　医疗数据集分类精度及召回率比较

类别	贝叶斯		kNN		决策树（AssistantR）		本书方法	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
PRIM	49.2 ± 3.9	56.1	42.1 ± 5.0	49.1	38.9 ± 4.7	41.1	48.7 ± 3.1	64.2
BREA	77.3 ± 4.2	71.2	79.5 ± 2.7	81.7	78.5 ± 3.9	67.2	79.9 ± 4.0	76.3
LYMP	84.2 ± 2.7	73.3	82.6 ± 5.7	83.2	77.0 ± 5.9	67.1	86.4 ± 4.1	78.4
RHEU	67.1 ± 4.7	71.4	66.0 ± 3.6	65.3	63.8 ± 4.9	59.4	71.3 ± 3.8	69.3

为避免实验误差带来的影响。我们以 PRIM 分类为例，对四种算法 10 次随机测试结果采用统计显著性检验验证其分类效果。选择了方差分析中的 LSD（最小显著性差异）方法比较本书方法与其他方法之间的差异，从表 6-7 可以看出，在显著性水平为 0.01 时，本书方法与连续 r 值匹配否定选择、kNN 之间的差异是显著的，仅与贝叶斯的差异不显著。

表 6-7　不同算法分类精度统计显著检验比较

方法（I）	比较方法（J）	均值差（I-J）	标准误差	显著性	90%置信区间	
					下限	上限
本书方法	贝叶斯	.82000	.97405	.405	-.8245	2.4645
	kNN	7.63000*	.97405	.000	5.9855	9.2745
	决策树	10.68000*	.97405	.000	9.0355	12.3245

*. 均值差的显著性水平为 0.1。

在算法的比较中我们可以看出，在 PRIM 分类评价中，贝叶斯分类精度及信息分均高于其他分类方法，在其他三个类别的分类比较中，本书方法所得分类精度及信息分都较高。这样的实验结果验证了贝叶斯方法由于需要估计

概论，在样本量较少的情况下不宜发挥作用。由于本书方法不需要先验知识，通过对数据集的全局进行优化，输出规模更小、结构简单便于理解的规则集，通过否定选择分类器进行规则匹配，克服了其仅适合样本稀疏或小样本空间的问题，在分类效果上优于决策树（Assistant R）方法，同时在分类效率上高于 kNN、贝叶斯方法。

参考文献

[1] Perelson A S, Oster G. F. Theoretical Studies of Clonal Selection: Minimal Antibody Repertoire Size and Reliability of Self-Nonself Discrimination. Journal of theory Immune, 1979, 81: 645-670.

[2] Takahashi K, Yanada T. Application of an Immune Feedback Mechanism to Control Systems. JSME International Journal, Series C, 1998, 41(2): 184-191.

[3] 刘赫，刘大为，裴志利．基于多种群协同优化的文本分类规则抽取方法．自动化学报，2009，10（35）：1335-1340．

[4] 苏金树，张博锋，徐昕．基于机器学习的文本分类技术研究进展．软件学报，2006．9（17）：1848-1859．

[5] Lewis D D. Naive Bayes. The Independence Assumption in Information Retrieval. Proc of the 10 th European Conference on Machine Learning. Chemnitz, Germany, 1998: 4-15.

[6] Ramon M,Sebastiam P.Robust Baycs classifiers. Artificial Intelligence, 2001, 125(1-2): 209-226.

[7] 胡昌平，胡吉明．个性化服务中基于支持向量机的用户兴趣挖掘分析．情报学报，2009，28（4）：543-547．

[8] 吉汉强，李丽舒．数字资源分类方法的探讨．图书馆论坛，2011，31（1）：101-104．

[9] 李辉，史忠植，许卓群．运用文本领域的常识改善基于支持向量机的文本分类器性能．中文信息学报，2002，16（3）：2-13．

[10] 夏火松，彭柳艳，余梦麟．自动情感文本分类研究综述．情报学报，2011．30（5）：530-540．

[11] M, Kryszkiewicz. Rough set approach to incomplete information systems. Information Sciences, 1998, 112(1): 39-49.

[12] 莫宏伟，谭娜，金鸿章，等．免疫阴性选择分类器在信息恢复中的应用．计算机学报．2005．28（8）：1314-1319．

[13] 唐华，曾碧卿．基于遗传算法和信息熵的文本分类规则抽取方法研究．中山大学学报（自然科学版）．2007，47（5）：18-23．

[14] Ryszard S. Michalski Ivan Bratko Miroslav Kubat. Machine Learning and Data Mining: Methods and Applications. 2008.

[15] Agrawal R, I miclinski T, Swami A. Database mining: A performance perspective. IEEE Trans Knowledge and Data Enginnering, 1993, 5: 914-925.

[16] Agrawal R, Srikant R. Fast algorithm for mining association rules[A]. Proceeding 1994 International conference Very Large Data Bases (VLDB' 94). Santiago: Chile, 1994: 487-499.

[17] Han Euihong, George K, Kumar V. Scalable parallel data mining for association rules. Proceeding of the ACM SIGMOD97. New York: ACM Press, 1997: 277-288.

[18] B Hetzler, W M Harris, S Harvre, and P Whitney, Visualizing the full spectrum of document relationships. In: Proceedings of the Fifth International Society for Knowledge Organization Conference, 1998: 168-175.

[19] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 2002, 3: 397-422.

[20] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 2003, 32(1): 48-77.

[21] J. Abernethy, E. Hazan, A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In: Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland, 2008: 263-274.

[22] P. Auer, N. Cesa-Bianchi, P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 2002, 47(2-3): 235-256.

[23] V. Dani, T. P. Hayes, S. M. Kakade. Stochastic linear optimization under bandit feedback. In: Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland, 2008: 355-366.

[24] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, A. Rakhlin. Stochastic convex optimization with bandit feedback. SIAM Journal on Optimization, 2013, 23(1): 213-240.

[25] A. D. Flaxman, A. T. Kalai, H. B. McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In: Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms, Vancouver, Canada, 2005: 385-394.

[26] W. Smart, M. Zhang. Applying online gradient descent search to genetic programming for object recognition. In: Proceedings of Australasian Workshop on Data Mining and Web Intelligence, Dunedin, New Zealand, 2004: 133-138.

[27] B. Awerbuch, R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In: Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, 2004: 45-53.

第 7 章

免疫网络在生物信息学中的应用

本章导读：

医学的进步及生物技术的飞速发展使人类逐步加深了对自身遗传信息的了解，尤其是随着人类基因组计划（Human Genome Project，HGP）的实施，人们逐步掌握了破解遗传密码的基本元素。现代的高通量测序技术，使人们可以非常容易地获得现存物种的全基因组核酸序列或蛋白质组的氨基酸序列。基因序列正以每天超过 50 万个碱基对的速度加入到数据库中，其提交的序列速度呈指数级增长，大约每 14 个月就会增长一倍。这些数据的特点是海量、高维度、数据变量复杂、分析处理复杂。所以，在当前基因组信息爆炸的时代，如何处理和分析这些海量的生物信息数据是生物信息工作者面临的巨大挑战。本章从免疫进化网络理论着手，研究了 aiNet 和 AIRS 等聚类和分类算法，在免疫系统形态空间理

论的基础上构建了一种基于免疫进化网络理论的分类器，并将其运用于 DNA 序列的模式识别中。由于 DNA 序列的特征提取和亲和力测量方法对分类性能有显著影响，在本分类器设计中，采用离散增量度量亲和力，从而更好地衡量序列之间的相似性。将该分类器用于模式生物基因的识别，仿真结果表明了其有效性。

7.1 基本概念及问题描述

生物信息学（Bioinformatics）是一门新兴的交叉学科，它的研究焦点主要集中于使用统计学和计算机科学工具，分析和解释海量分子生物学数据信息。生物信息学作为一门专门的学科，发起于 20 世纪 80 年代。“生物信息学”这一名称首先是由 Paulien Hogeweg 和 Ben Hesper 于 1978 年提出的，但直到 20 世纪 80 年代末才被广泛使用。在这一时期，由于测序技术的飞速发展，尤其是随着人类基因组计划的实施，海量的 DNA 序列信息被世界各地研究机构的测序仪源源不断地检测出来。尽管这些基因组信息只有 A、T、C、G 四个字母，然而其容量却是惊人的。正是以这些信息为蓝本，造就了自然界生命和物种的多样性。如何分析和解读这样海量的信息，就成为生命科学研究工作者需要解决的一道难题。随着现代计算机信息技术的长足发展，对海量数据处理的能力越来越强，同时，大量的计算机科学家和统计学家进入了这一领域，帮助生物学家来完成海量生物信息数据的处理工作，从而产生了生物信息学这一学科。到了 21 世纪，人类步入了信息化时代，分子生物学也有了更进一步的发展，包括基因组学、蛋白质组学、转录组学等学科都在分子生物学领域扮演了重要的角色。相对而言，各种组学数据所呈现的多维度、多粒度、海量庞杂等特点也让生物信息学作为一种分析和研究的手段有了不可替代的用武之地。在 21 世纪中期，随着人类对生命本身的不断了解，生物信息学将扮演更重要的角色。从信息角度看，生物信息学指利用计算机及相

关技术对各种生物信息数据进行提取、储存、处理和分析。但生物信息学的研究领域十分广泛，从序列比对到基因发现与功能研究，从基因表达分析到蛋白质结构与功能预测，乃至更复杂的调控网络、代谢网络及蛋白质相互作用网络等。其关键点主要表现在两方面：

（1）生物信息的数学解析，如由 A/G/C/T 四种碱基表达的 DNA 序列如何映射到数据空间，得到该序列的有效特征值；

（2）生物特征数据的处理，即从海量数据中挖掘出信息，寻找新的基因或预测基因的功能进而直到分子进化研究。

对于第一个方面，一般是采用统计计算方法获得 DNA 序列的特征信息，例如 A/G/C/T 四种碱基在序列中的含量、两碱基含量（四种碱基的两两组合，共 16 种）和三联体含量法（A/G/C/T 四种碱基每三个一组合，共 64 种）。还可以从信息角度考虑，计算其离散量和信息熵。现在多数学者采用智能计算的方法（如人工神经网络、支持向量机和模糊运算等）处理数据和发现信息。模式生物基因组计划在人类基因组的研究中占有极其重要的地位。利用模式生物基因组比较和鉴别不同进化阶段生物体的基因组信息，将有利于加深对高等生物特别是人类基因组结构和功能的了解, 揭示生命的本质规律。利用模式生物基因组与人类基因组之间编码顺序和组织结构上的同源性，可用单一或简单的生物模式阐明高等生物特别是人的基因组在结构及物种进化方面的内在联系，目前已从模式生物之间及人类之间发现了一些共性特征及各自的独特性。本节采用线虫、酵母和拟南芥三种模式生物，将其基因组中的内含子、外显子和基因间序列归为三类，滑动统计这些序列中 64 种三核苷的重复出现次数，作为离散源的状态参数。这样就得到了这些序列的 64 维特征值，并将这些数据分成训练样本集和测试样本集。然后根据免疫进化网络理论，用离散增量作为抗体-抗原间的亲和力函数，将训练样本集看成抗原，模拟免疫网络对抗原的一系列刺激过程，如抗体-抗原识别、免疫克隆增殖、亲和度成熟和网络抑制等，构造了一个基于离散增量的免疫分类器。最后，用该分类器对训练集和测试集进行测试，结果表明该分类器性能优良，分类预测准

确率达到了 85%以上。同时，也可尝试将此方法用于蛋白质的结构功能预测及其他分类应用领域。

7.2 人工免疫网络理论

人工免疫系统是在免疫学尤其是在理论免疫学的基础上发展起来的，因此离不开对生物免疫系统的理解和研究。目前较有影响的基于免疫网络系统的数据分析算法是 aiNet 和 RLAIS。在这两个算法中，训练集合由抗原集合确定，目的是产生一组抗体（B 细胞）来表示这些抗原。

7.2.1 aiNet

De Castro 研究了免疫系统的一些基本机制，提出一种名为 aiNet 的人工免疫网络算法。免疫网络是一个复杂结构，Jerne 曾指出：在一个完备的免疫系统中，不同抗体间也可以产生作用。免疫网络的关键是对自身的识别，其对外来抗原的应答建立在识别自身抗原的基础上，由细胞表面的抗体组成初始网络，通过抗体间的相互识别不断调整网络。De Castro 提出的 aiNet 是一种基于连接主义的调节免疫网络算法，主要研究无标签数据集合的压缩和聚类问题，算法表明人工免疫网络具有强大的计算能力。aiNet 模拟机体中免疫系统对抗原刺激的应答过程，主要包括抗原识别、亲和力成熟、克隆增殖和网络抑制等机制。这里将待处理数据看作抗原，抗体则是算法产生的反映抗原特征的数据。算法最终输出一个记忆细胞集合 M，M 体现了抗原数据的内部结构，不仅表明免疫系统本身具有强大的计算能力，而且可用免疫学原理发展数据处理工具。

aiNet 基本原理如 3.2.2 节的描述，这里仅给出 aiNet 算法的简单描述，如图 7-1 所示。

步骤 1：

（1）对于每一个抗原，计算随机产生的抗体亲和力，选出 n 个高亲和力抗体。

（2）在 n 个高亲和力抗体中，计算亲和力并产生克隆抗体集 D，对于每一个抗体，其亲和力越高，克隆数越多。

（3）对克隆抗体集 D 进行亲和力计算得到 D^*，即 D 中每个抗体根据公式 $C = C - a(C - X)$ 进行变异，亲和力越高，变异率越小；其中，C 是网络细胞矩阵，X 是抗原矩阵，a 是学习率或成熟率，根据 *Ag-Ab* 亲和力设定：亲和力越高，a 越小。

（4）求出抗原和 D^*中每一个抗体的亲和力。

（5）从 D^*中选出一定比例高亲和力的抗体，放入克隆记忆集中。

（6）求出记忆集中抗体的相似度进行克隆抑制。

（7）将记忆集中抗体存入总的记忆抗体集中。

步骤 2：求总记忆抗体相似度，进行网络抑制。

步骤 3：免疫网络抗体生成。

步骤 4：终止条件——抗体达到某一指定数或指定迭代次数完成。

图 7-1　aiNet 算法基本流程图

将 aiNet 算法用于数据分析具有以下优点：对二维、三维数据通过形成的免疫记忆数据以较好的可视化效果反映原始数据之间的聚类结构，其性能超过传统的分级聚类方法；可通过抑制阈值参数调节控制生成的记忆细胞数目；产生的记忆细胞质量较高。通过研究不难发现，传统的 aiNet 算法主要存在以下不足：

（1）aiNet 算法中的大多参数都采用定值策略，这是根据决策者的经验决定的，aiNet 模型的网络结构、抗体数目、聚类准确性等受抑制阈值的影响非常大，且阈值参数不能随网络的进化而动态改变。

（2）在现有的 aiNet 算法中，初始抗体细胞记忆集均随机产生，这并不能很好地反映原始数据的特征，存在一定的盲目性，因此 aiNet 模型的最终聚类结果受人为主观因素影响较大。

（3）在现有的克隆选择算法和人工免疫网络算法中，抗体和抗原间的亲

和度均采用简单的距离函数来评价。由于种群中的个体之间存在联系，这种简单的距离度量并不能得到全局评价的效果，导致在克隆选择时漏掉优秀的个体。

（4）传统 aiNet 算法只能处理低维和简单数据，数据规模较小时可以获得理想的效果，但对四维以上高维数据聚类无法实现网络可视化。

7.2.2 AIRS

Wakins 受 Tinimis 提出 RLAIS 方法的启发，尤其是 RLAIS 中 ARBs 概念的影响，提出了基于资源有限人工免疫系统的免疫分类器 AIRS 解决数据分类问题。其算法流程图如图 7-2 所示。

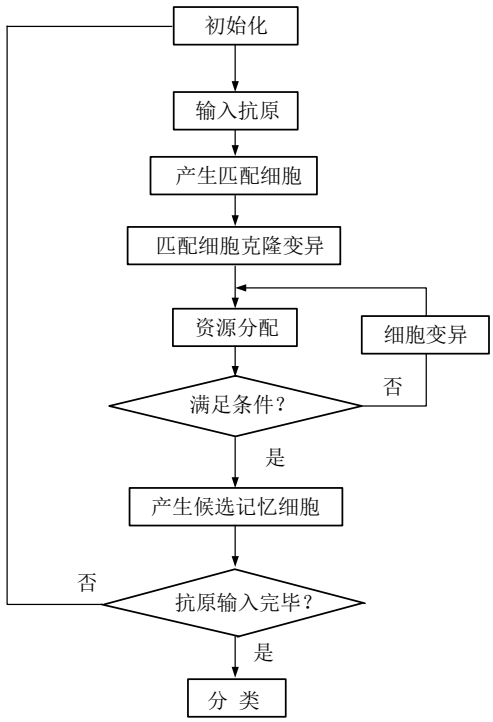

图 7-2 AIRS 算法流程图

AIRS 模拟免疫系统 B 细胞的主要机制。AIRS 中的抗原初始化为特征向量形式，在训练和学习期间提呈给系统。AIRS 中的 B 细胞的表示与抗原的表示相同，即都是形态空间中具有相同维数的特征向量。所有具有类似特征的 B 细胞表示为 ARBs。AIRS 是一个资源有限的监督学习系统。ARBs 需要竞争固定数目的资源，这有助于接近训练抗原的 ARBs 逐渐进化。另一个要素是记忆细胞池，类似 B 细胞，但在系统中存活时间长，用于测试抗原的实际分类。ARBs 池用于哺育候选记忆细胞。作为一个监督学习系统，AIRS 中的每一个抗原（训练集合中的特征向量）都具有类别，而产生对该抗原应答的 ARBs。应答强烈的 ARBs 进一步处理进入记忆细胞池，经过资源分配过程，从记忆细胞池中删除不具竞争力的 ARBs，在竞争中获胜的 ARBs 在训练结束后保留下来，成为最终的分类工具。算法中的变异能够产生成功的 ARBs，是由于变异不产生与训练向量同样的抗体，而是与训练向量充分相似，极具竞争力。这是 AIRS 具有数据泛化能力的重要原因。AIRS 从用户初始化设置的记忆细胞集合开始培养记忆细胞池。一般的，AIRS 在记忆细胞中产生的记忆细胞数量大约是提呈给系统的训练细胞数量的一半。一旦训练完成，记忆细胞作为 kNN 分类系统对测试数据分类。由于 AIRS 用 Euclidean 距离计算亲和力，所以用实数值向量测试分类。

AIRS 的优点是不需要事先知道分类器的合理设置，该分类器的重要参数可由用户确定。一旦训练完成后，分类器本身就是一个 kNN 分类器的延伸。但是，它具有传统 kNN 所不具备的泛化能力。

综上所述，aiNet 和 AIRS 都是通过免疫记忆、克隆和变异原理产生记忆细胞或抗体去模拟逼近原始数据集合，整个过程是一个通过免疫算法逼近数据模式的过程。二者只是在产生记忆抗体或细胞的过程上有所不同，在对记忆细胞或抗体矩阵的最终应用目的上有所不同：一个用于聚类分析，一个用于分类分析。

7.3 基于免疫进化网络理论的分类器

分类是有监督学习，通过学习可以对未知的数据进行预测。要构造分类器，需要有一个训练样本数据集作为输入。训练集由一组数据库记录或元组构成，每一个元组是一个由有关字段（又称属性或特征）值组成的特征向量。此外，训练样本还有一个类别标记。一个具体样本的形式可为$(v_1,v_2,\cdots,v_n; c_j)$，其中，$v_i$表示属性值，为实数；$c_j$（$j$=1,2,⋯,$m$）表示类别。因此，抗体与抗原采用实数的形态空间描述。为了充分体现免疫系统的自学习和自适应能力，根据克隆选择原理、形态空间理论和独特型网络理论，构造了一个基于免疫网络理论的分类器（aiENC）。

在本书中，分类算法将训练数据集看作抗原，将算法中产生反映抗原属性特征的数据看作抗体，然后模拟免疫网络抗体−抗原之间的相互刺激和作用来优化网络结构，完成数据的处理。最后保留对应 m 个类别的记忆细胞池 M_j (j=1,2,⋯,m)对未知的数据进行预测分类。算法步骤如图 7-3 所示。

具体算法说明如下：

采用式（7.1）对抗原 Ag 的每个属性特征值作标准化处理，但类属性值不作标准化处理。每类随机产生（0,1）之间的 N 个初始化抗体 Ab，并设置训练代数 P。

$$Ag_{ij}=\frac{Ag_{ij}-\min(Ag_{\cdot j})}{\max(Ag_{\cdot j})-\min(Ag_{\cdot j})} \tag{7.1}$$

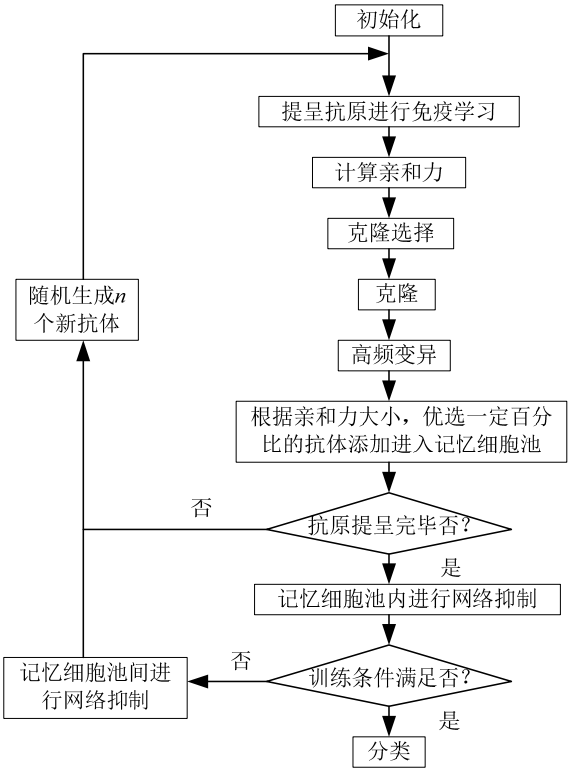

图 7-3　分类算法流程图

对标准化后所得的抗原进行免疫学习与识别，并采用欧氏距离公式计算抗原与抗体之间的亲和度 d_{ij}，如式（7.2）。然后根据亲和度排序，选择亲和度高的 m 个抗体作为网络细胞，对 m 个网络细胞进行克隆操作，其克隆数与亲和度成正比。此时可获得增殖后的抗体网络细胞群 C。

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (Ab_{ik} - Ag'_{kj})^2} \tag{7.2}$$

根据式（7.3）对克隆后的抗体网络细胞群进行变异操作。其中，C 代表有网络细胞群，C_{Ag} 表示克隆抗原细胞，α 为变异率矩阵。变异使抗体朝向识别抗原的方向进化。C_{Ag} 与 α 的规模与 C 一样。

$$C^* = C - \alpha * (C - C_{Ag}) \tag{7.3}$$

计算变异后的网络细胞群 C^*与本次提呈抗原的亲和度。根据亲和度排序，优选一定比例（η%，一般选择 25%左右）的网络细胞作为本次提呈抗原的记忆细胞 M_{Agi}，并根据抗原的类别将 M_{Agi} 添加进相应的记忆细胞池 M_i。根据独特型网络理论，对 m 个记忆细胞池 M_i 单独实行网络抑制操作。该操作分别计算每个记忆细胞池中抗体间的相似度，如式（7.4）。清除相似度小于阈值 σ_{s1} 的记忆细胞，最后得到压缩后的记忆细胞池 M_i。

$$s_{ij} = \sqrt{\sum_{k=1}^{n} (Ab_{ik} - Ab'_{kj})^2} \tag{7.4}$$

然后计算每类记忆细胞池中的抗体与其他类记忆细胞池中抗体的相似度，清除相似度小于阈值 σ_{s2} 的记忆细胞，并将每类的记忆细胞池中的记忆细胞与随机生成的 N 个抗体构成新的该类抗体。最后提呈待分类的数据集，与 m 个记忆细胞池中的每个抗体计算欧氏距离，然后采用最近邻法则判断每个抗原的类属性。输出最终结果。

aiENC 分类算法与 aiNet 算法之间的区别如下:

（1）aiENC 算法相对 aiNet 算法而言，其抗原有类别标记。

（2）aiENC 算法对应提呈抗原所产生的记忆细胞带有类标记，而 aiNet 算法没有。

（3）aiENC 算法由于具有类别标记，通过监督学习机制分类抗原，而 aiNet 算法使用传统的无监督学习聚类方法分析记忆细胞中的聚类信息，间接反映原始抗原中的信息。

（4）与 aiNet 算法中抗原和记忆细胞没有类别不同，aiENC 算法抗原与 M_j 相互匹配时，只与 M_j 中同类的记忆细胞相匹配，不同类的记忆细胞保留下来适应将来的同类抗原匹配。

（5）aiNet 算法中只采用了一次网络抑制操作，而 aiENC 算法不但对记忆细胞池内的细胞进行网络抑制，而且对不同类别的记忆细胞池之间的抗体进行网络抑制，这更有利于保留记忆细胞的特异性，提高分类准确率。

aiENC 分类算法与 AIRS 算法的区别如下：

（1）两者产生记忆细胞的机制不同。aiENC 算法的记忆细胞是通过人工免疫网络的促进和抑制原理、亲和力成熟等机制产生的。AIRS 算法是通过资源有限机制产生记忆细胞的。

（2）每次算法运行产生的记忆细胞数目不同。aiENC 每次提呈抗原后，可以根据设置产生多个记忆细胞，经过免疫操作算子的多次循环后产生记忆细胞矩阵。而 AIRS 算法每次至多只能产生一个记忆细胞，最后累计产生记忆细胞池用于对新抗原分类。

7.4 仿真实验及应用

现在，科学家们发现了 DNA 序列中的一些规律性与结构。由 A、T、G、C 四种碱基按一定顺序排成的基因草图序列中，在每三个字符一组构成的 64 种表达中，大多数用于编码构成蛋白质的 20 种氨基酸，而在不用于编码蛋白质的序列片段中，A 与 T 的含量多些。由此可看出，DNA 序列中存在着局部性与全局性的结构。充分挖掘序列的结构对理解和破译 DNA 序列非常有意义。因此，有必要对 DNA 序列片段进行分类，然后归类分析。现在，已有许多研究者采用传统方法和智能计算方法对其进行聚类和分类研究。

每个 DNA 序列都是由腺嘌呤 A、胞嘧啶 C、鸟嘌呤 G、胸腺嘧啶 T 四个字符随机组成的字符串，如：TGACCTCTTGTCCTGTATAGCAA。DNA 序列的特征提取方法很多，不同方法对分类的影响很大。本书采用二维、四维和五维数字化 DNA 序列进行讨论。

7.4.1 数据准备与处理

数据采用 Art-model-data 中的前 40 号样本和 Nat-model-data 数据。其中，

Art-model-data 中的 1～10 号序列为 A 类，11～20 号序列为 B 类，21～40 号样本的类别未知；Nat-model-data 为 182 个长自然 DNA 序列，其所有序列类别也是未知的。所以，在分类算法中，Art-model-data 中的 1～20 号序列为训练数据样本，Art-model-data 中的 20～40 号序列和 Nat-model-data 数据为待测数据样本。

由于 DNA 序列由四字符组成，分别统计每个 DNA 序列中 A、G、C、T 出现的频率即可获得四维向量标记的 DNA 序列。再加上类属性，训练数据样本为一个 20×5 的矩阵；而待测数据样本为 20×4 与 182×4 的矩阵。由于腺嘌呤 A 和鸟嘌呤 G 同属嘌呤类碱基，结构相似，在某些方面具有共性，可归为一类。同理，胞嘧啶 C 和胸腺嘧啶 T 同属嘧啶类碱基，具有共性，也可归为一类。统计 DNA 序列中（A+G）和（C+T）的出现频率就可获得二维向量标记的 DNA 序列。对于五维 DNA 序列，我们采用如下方式得到。根据生物学中密码子的概念，A、C、G、T 四个碱基可以组成 4^3=64 种密码子，其中 61 种对应各种氨基酸的编码，另外三种 TAA、TGA、TAG 是多钛链合成的种植信号。61 种密码子所表示的氨基酸分类如表 7-1 所示。计算序列中四类氨基酸出现的频率，再加上种植信号出现的频率即可获得五维 DNA 序列向量。

表 7-1　氨基酸的分类

种　类	三字符串	种　类	三字符串
非极性 疏水性氨基酸	GGA、GGG、GGC、GGT、GCA、GCG、GCC、GCT、GTA、GTG、GTC、GTT、CTA、CTG、CTC、CTT、TTA、TTG、ATA、ATC、ATT、TTC、TTT、CCA、CCG、CCC、CCT	极性 中性氨基酸	TGG、TCA、TCG、TCC、TCT、AGC、AGT、TAC、TAT、TGC、TGT、ATG、AAC、AAT、CAA、CAG、ACA、ACG、ACC、ACT
酸性氨基酸	GAC、GAT、GAA、GAG	碱性氨基酸	AAA、AAG、CGA、CGG、CGC、CGT、AGA、AGG、CAC、CAT

7.4.2 仿真结果

首先，我们采用 Art-model-data 中带有类属性的 1～20 号序列为训练数据样本，根据本章第二节中的算法，运用留一交叉验证法来训练分类器。在 20 个训练样本中，每次留下一个训练样本作为测试样本，对分类器学习的结果进行验证，结果如表 7-2 所示。

表 7-2 留一交叉验证法分类的准确率

	σ_s=0.05	σ_s=0.1	σ_s=0.15	σ_s =0.2	σ_s=0.25	σ_s=0.3
二维样本	100%	100%	95%	95%	95%	95%
四维样本	95%	95%	95%	90%	90%	90%
五维样本	90%	90%	90%	90%	95%	90%

根据表 7-2 中的结果，二维与四维数据选择抑制阈值 σ_s=0.1 比较合适，对于五维数据选择 σ_s=0.25。然后用 Art-model-data 中 1～20 号的训练样本按本章第二节中的步骤进行指导训练，由此得到的分类器再对 Art-model-data 中的 20～40 号序列和 Nat-model-data 数据进行分类预测，其结果如表 7-3 与表 7-4 所示。

表 7-3 Art-model-data 中 20～40 号序列分类结果

	Art-model-data 中 20～40 号序列	
	A 类	B 类
二维数据样本	2 3 5 7 9 10 14 15	1 4 6 8 11 12 13 16 17 18 19 20
四维数据样本	2 3 5 7 9 10 14 15 17	1 4 6 8 11 12 13 16 18 19 20
五维数据样本	2 3 5 6 8 10 15 16 19	1 4 7 9 11 12 13 14 17 18 20

表 7-4 Nat-model-data 中 182 个序列的分类结果

	Nat-model-data 中 182 个长序列	
	A 类	B 类
二维数据样本	7 9 10 12 17 20 28 33 34 36 44 47 51 53 58 60 65 74 75 77 80 83 84 87 89 93 97 98 99 110 114 127 128 129 130 138 142 143 146 147 150 151 152 155 171	1 2 3 4 5 6 8 11 13 14 15 16 18 19 21 22 23 24 25 26 27 29 30 31 32 35 37 38 39 40 41 42 43 45 46 48 49 50 52 54 55 56 57 59 61 62 63 64 66 67 68 69 70 71 72 73 76 78 79 81 82 85 86 88 90 91 92 94 95 96 100 101 102 103 104 105 106 107 108 109 111 112 113 115 116 117 118 119 120 121 122 123 124 125 126 131 132 133 134 135 136 137 139 140 141 144 145 148 149 153 154 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 172 173 174 175 176 177 178 179 180 181 182
四维数据样本	5 8 13 14 15 16 18 20 27 32 42 49 55 58 59 61 62 67 68 69 70 71 73 79 82 89 90 91 100 102 104 109 112 115 117 118 120 124 134 136 141 155 158 162 171 176	1 2 3 4 6 7 9 10 11 12 17 19 21 22 23 24 25 26 28 29 30 31 33 34 35 36 37 38 39 40 41 43 44 45 46 47 48 50 51 52 53 54 56 57 60 63 64 65 66 72 74 75 76 77 78 80 81 83 84 85 86 87 88 92 93 94 95 96 97 98 99 101 103 105 106 107 108 110 111 113 114 116 119 121 122 123 125 126 127 128 129 130 131 132 133 135 137 138 139 140 142 143 144 145 146 147 148 149 150 151 152 153 154 156 157 159 160 161 163 164 165 166 167 168 169 170 172 173 174 175 177 178 179 180 181 182

续表

	Nat-model-data 中 182 个长序列	
	A 类	B 类
五维数据样本	1 2 3 5 7 10 12 13 14 15 16 17 18 20 21 25 26 27 28 31 32 33 34 35 36 37 38 39 40 41 42 44 45 46 47 49 50 51 52 53 58 60 61 62 63 64 65 66 67 68 71 72 73 74 77 79 80 82 83 84 85 86 87 88 89 92 93 94 95 96 97 98 99 100 101 102 104 105 106 109 110 111 112 115 116 122 124 125 128 129 130 131 132 133 134 136138 139 140 141 142 144 145 146 147 148 151 154 155 157 160 161 162 166 167 170 172 175 176 177 178 179 180 181 182	4 6 8 9 11 19 22 23 24 29 43 48 54 55 56 57 59 69 70 75 76 78 81 90 91 103 107 108 113 114 117 118 119 120 121 123 126 127 135 137 143 149 150 152 153 156 158 159 163 164 165 168 169 171 173 174

由表 7-3 与表 7-4 中的结果可知，分类结果与 DNA 序列的表达有关。根据留一交叉验证法的结果，采用人工免疫系统方法构造出的分类器具有比较高的分类准确率。同时，可以很方便地将该方法推广用于生物信息学中的模式识别和其他分类任务。怎样提取 DNA 序列的特征值，更准确地表达出 DNA 序列的特点对分类结果将有帮助。

7.5 免疫进化网络分类器改进及应用

7.5.1 基本概念

离散量（Measure of Diversity）和信息熵（Information Entropy）都是从信息的角度对状态空间的一种描述，度量的基础都是根据信息量度的对数函数。信息熵是对一个信息符号不确定性的度量，也是对状态不确定性或紊乱性的一种描述，而离散量是对整体不确定性多少的度量，也是离散多少的度量。信息熵大，表示不确定性的程度大，但具有的离散量并不一定多；反过来，离散量多并不意味着紊乱性的程度大。

对于 s 个信息符号的状态空间，用 n_i 表示第 i 个状态出现的个数，如此离散源 X：$(n_1, n_2, \cdots, n_s)$ 的离散量为式（7.5）

$$D(X) = D(n_1, n_2, \cdots, n_s) = N \log_b N - \sum_{i=1}^{s} n_i \log_b n_i \qquad (7.5)$$

其中，$N = \sum_{i=1}^{s} n_i$，如此确立的离散量具有以下性质。

（1）非负性：$D(n_1, n_2, \cdots, n_s) \geq 0$。

（2）对称性：$D(n_1, n_2, \cdots, n_i, \cdots, n_j, \cdots, n_s) = D(n_1, n_2, \cdots, n_i^*, \cdots, n_j^* \cdots, n_s)$，其中，$n_i^* = n_j, n_j^* = n_i, (i \neq j)$，表示离散量的任意两个变量 n_i 和 n_j 变换位置以后离散量不变。

（3）扩展性：$D(n_1, n_2, \cdots, n_s) = D(n_1, n_2, \cdots, n_s, 0)$。

（4）可加性：$D(n_{11}, n_{12}, \cdots, n_{1s}; n_{21}, n_{22}, \cdots, n_{2s}; \cdots; n_{r1}, n_{r2}, \cdots, n_{rs}) = D(m_1, m_2, \cdots, m_r) + \sum_{i=1}^{r} D(n_{i1}, n_{i2}, \cdots, n_{is})$，其中，$m_i = \sum_{k=1}^{s} n_{ik}, (i = 1, 2, \cdots, r)$。

（5）极值性：如果离散源的 s 个数量 n_i 相等，即 $n\times s=N=\sum_{i=1}^{s}n_i$ 时，离散量达到极大值，则有 $D(n_1,n_2,\cdots,n_s)\leq D_s(n,n,\cdots,n)=sn\log s$ 。

（6）等倍增性：离散量与离散源以相同倍数增长时，则 $D(kn_1,kn_2,\cdots,kn_s)=kD(n_1,n_2,\cdots,n_s)$ 或写成 $D(kX)=kD(X)$ 。

如果有两个离散源 $X:(n_1,n_2,\cdots,n_s),Y:(m_1,m_2,\cdots,m_s)$，则离散增量定义如式（7.6）

$$\Delta(X,Y)=D(X+Y)-D(X)-D(Y)=D(M+N)-\sum_{i=1}^{s}D(m_i+n_i) \quad (7.6)$$

其中，$M=\sum_{i=1}^{s}m_i,N=\sum_{i=1}^{s}n_i$，$D(M+N)=(M+N)\log_b(M+N)-M\log_b M-N\log_b N$，$D(m_i+n_i)=(m_i+n_i)\log(m_i+n_i)-m_i\log m_i-n_i\log n_i$。

由离散量与离散增量的定义，离散增量的取值范围：$0\leq\Delta(X,Y)\leq D(M+N)$。离散量从 0 增加到 $D(M+N)$ 体现了两组数据 X 与 Y 之间的相似程度：$\Delta(X,Y)$ 越小，则两组数据越相似。

7.5.2 免疫离散增量分类器设计

在本节中，将采用离散增量作为抗体-抗原之间的亲和力函数和抗体-抗体之间的相似度函数来训练分类器，进行分类预测。这里，仍采用实数形态空间模型，抗原 $Ag=(v_1,v_2,\cdots,v_s;\ c_j)$，其中 v_i 表示属性值（特征值），采用 $c_j(j=1,2,\cdots,m)$表示类别。若有 r 个训练样本，则训练集（抗原空间）大小为 $r\times s$。在分类算法中所产生的反映抗原属性特征的数据看作抗体 Ab，然后模拟免疫网络抗体-抗原之间的相互刺激和作用来实现数据的处理，最后保留对应 m 个类别的记忆细胞池 M_j ($j=1,2,\cdots,m$)对未知的数据进行预测分类。算法步骤如下。

步骤 1：每类随机产生 N 个初始化抗体 $Ab_j=(w_1,w_2,\cdots,w_s)$，$j=1,2,\cdots,N$，即

有 m 个大小为 N（本算法中 N=100）的初始抗体群，并同时设置训练代数 P=100 作为训练结束标志。

步骤 2：把训练样本集看成抗原（Ag）。根据类别，每次提呈该类训练集中的一个抗原 $Ag_i(i=1,2,\cdots,r)$进行免疫学习与识别。

步骤 3：采用离散增量计算抗原与抗体之间的亲和度 $\Delta(Ag_i, Ab_j)$，见式（7.7）和式（7.8）

$$d_{ij} = \frac{1}{\Delta(Ag_i, Ab_j)} \tag{7.7}$$

$$\Delta(Ag_i, Ab_j) = D(Ag_i + Ab_j) - D(Ag_i) - D(Ab_j) = D(V+W) - \sum_{i=1}^{s} D(v_i + w_i) \tag{7.8}$$

其中，$V = \sum_{i=1}^{s} v_i, W = \sum_{i=1}^{s} w_i$，$D(V+W) = (V+W)\log_b(V+W) - V\log_b V - W\log_b W$，$D(v_i + w_i) = (v_i + w_i)\log(v_i + w_i) - v_i \log v_i - w_i \log w_i$。

然后根据亲和力 d_{ij} 排序，选择亲和度高的 m 个抗体作为网络细胞，对 m 个网络细胞进行克隆操作，其克隆数与亲和度成正比。此时可获得增殖后的网络细胞群 C。

步骤 4：对克隆后的网络细胞群进行变异操作，其方式如式（7.9）

$$C^* = C - \alpha^*(C - C_{Ag}) \tag{7.9}$$

C 代表由步骤 3 得到的网络细胞群；C_{Ag} 表示克隆抗原细胞；α 为变异率矩阵，变异使抗体朝向识别抗原的方向进化；C_{Ag} 与 α 的规模与 C 一样。

步骤 5：计算变异后的网络细胞群 C^*与本次提呈抗原的亲和度。根据亲和度排序，优选一定百分比（η%，一般选择 25%左右）的网络细胞作为本次提呈抗原的记忆细胞 M_{Agi}，并根据抗原的类别将 M_{Agi} 添加进相应的记忆细胞池 M_i。

步骤 6：判断所有抗原刺激结束否，如果没有，返回步骤 2。

步骤 7：根据独特型网络理论，将 m 个记忆细胞池 M_i 实行网络抑制操作。该操作计算每个记忆细胞池中抗体间的相似度，这里仍采用离散增量，见公式（7.10）。清除相似度小于阈值 σ_s 的记忆细胞。最后得到压缩后的记忆细胞池 M_i。

$$S_{ij} = \frac{1}{\Delta(Ab_i, Ab_j)} \tag{7.10}$$

$$\Delta(Ab_i, Ab_j) = D(Ab_i + Ab_j) - D(Ab_i) - D(Ab_j) \tag{7.11}$$

步骤 8：判断训练代数是否结束。如果没有，先将每类的记忆细胞池中的记忆细胞与随机生成的 N 个抗体构成新的该类抗体；然后返回步骤 2。

步骤 9：提呈待分类的数据集，与 m 个记忆细胞池中的每个抗体计算离散增量，然后采用最近邻法则判断每个抗原的类属性，输出最终结果。

7.5.3 分类器在模式生物识别中的应用

目前大肠杆菌、酵母、拟南芥、果蝇和线虫在基因组序列信息研究上取得了重大进展，并且成为后基因组研究的主要模式生物材料，在基因功能、转录组、蛋白质组等方面获得了重要的成果，为高等生物及人基因组的研究提供了很好的借鉴，并为深入认识它们及生命进化提供了基本信息。

本节中的三种模式生物线虫、酵母和拟南芥的基因组数据来自 GenBank 数据库。线虫、酵母和拟南芥全基因组序列按照外显子、内含子和基因间序列分成三类，即分类器中的类别数 m=3。其中，从线虫的 6 条染色体中取用了 35823 条内含子、34796 条外显子、15784 条基因间序列；酵母从 16 条染色体中取用了 121 条内含子、2953 条外显子、5772 条基因间序列；拟南芥从 4 条染色体中取用了 40785 条内含子、44995 条外显子、20084 条基因间序列。每种生物的三类数据的训练样本集和测试样本集大小如表 7-5 所示。训练样本集用来对分类器进行训练；测试样本集用来检验该分类器的预测性能。

表 7-5 三种模式生物基因序列样本数据分布

		训 练 集	测 试 集	合 计
C. elegans	Intron	17354	18469	35823
	Exon	16739	18057	34796
	Intergenic DNA	7617	8167	15784
S. cerevisiae	Intron	58	63	121
	Exon	1484	1469	2953
	Intergenic DNA	2899	2873	5772
A.thaliana	Intron	20329	20456	40785
	Exon	22728	22267	44995
	Intergenic DNA	9747	10337	20084

由于不同模式的三核苷（三联体）内含子、外显子和基因间序列中的重复情形有所不同，所以取三核苷的 64（4^3）个模式作为状态空间的参量，滑动统计其所有样本上各个模式在每条序列中的重复次数作为特征参量值，比如在一条内含子中三核苷 AAG 重复出现了 n 次，则参量 AAG 的取值为 n。这样就得到一个 64 位的离散量 $Ag=(v_1,v_2,\cdots,v_s;\ c_j)$；这里，$s$=64，表明每条内含子、外显子和基因间序列是用 64 位的特征矢量表示；j=1,2,3，分别对应内含子、外显子和基因间序列三类。首先，用程序对选择的基因序列做统计处理，分别得到三种模式生物的所有内含子、外显子和基因间序列中的 64 位矢量特征值；然后，将其分成训练集和测试集两部分，用训练集训练免疫分类器；最后，采用训练所得的分类器预测测试集的类别。

采用 Guigo 的程序预测性能评价指标：敏感性、特异性和准确率。其定义如下：如果待测序列中有 M_1 条序列是内含子，M_2 条序列是外显子，M_3 条序列是基因间序列，用程序对序列进行预测的结果是，N_I 条序列被识别为内

含子(N_I=N_{11}+N_{12}，$N_{11}\in M_1$，$N_{12}\in M_2$ 或 M_3)，N_E 条序列被识别为外显子(N_E=N_{21}+N_{22}, $N_{21}\in M_2$, $N_{22}\in M_1$ 或 M_3)，则对内含子预测的敏感性为 S_n=N_{11}/M_1，对外显子预测的敏感 S_n=N_{21}/M_2，它表示程序的预测能力。而内含子预测的特异性为 T_n=N_{11}/N_I，对外显子预测的特异性为 T_n=N_{21}/N_E，它表示预测结果的可信赖程度。预测结果的准确率是敏感性和特异性的平均值。对基因间序列的敏感性、特异性及准确率的定义同上。采用本章中的免疫分类器，三种模式基因序列处理结果如表 7-6 所示。

表 7-6　三种模式生物预测结果

		Training set			Testing set		
		sensitivity	specificity	accuracy	sensitivity	specificity	accuracy
C. elegans	Intron	93.3%	96.4%	94.9%	89.5%	91.6%	90.6%
	Exon	94.2%	95.2%	94.7%	90.3%	89.4%	89.9%
	Intergenic DNA	92.4%	91.5%	92.0%	87.6%	86.7%	87.2%
S. cerevisiae	Intron	92.7%	93.1%	92.9%	85.6%	84.4%	85%
	Exon	92.2%	94.3%	93.3%	86.4%	87.3%	86.9%
	Intergenic DNA	94.8%	93.9%	94.4%	88.2%	89.5%	88.9%
A.thaliana	Intron	95.1%	96.7%	95.9%	90.2%	91.9%	91.1%
	Exon	93.1%	94.5%	93.8%	87.5%	90.1%	88.8%
	Intergenic DNA	94.2%	95.3%	94.8%	88.7%	91.8%	90.3%

由表 7-6 可知，训练集的预测性能普遍高于测试集，这是因为分类器本身是采用训练集进行有指导的学习，采用文中免疫分类算法所构造的分类器能

更准确地反映该训练类样本特征。该分类算法能实现抗体群自我调节，经过抗原（训练样本）反复刺激，产生免疫反应后保留的抗体群，即记忆细胞池 M_i 不但能体现训练集中样本的一致性，还能体现样本的多样性和特异性。而且，将该分类器用于测试集进行检验，同样获得了比较满意的结果。但该分类算法计算量比较大，程序执行时间比较长。

参考文献

[1] Nair A S. Computational biology & bioinformatics: a gentle overview. Communications of Computer Society of India, 2007, 30: 7-12.

[2] Benson D A, Karsch-Mizrachi I, Lipman D J, et al. GenBank. Nucleic Acids Res, 2011, 39(Database issue): D32-7.

[3] http://expasy.org/sprot/. 2011-09-21.

[4] 孙宝法．DNA 序列数据的聚类挖掘．河南科学，2004，22（5）：600-604．

[5] 由伟．用人工神经网络模型对 DNA 序列进行分类．科技信息，2007，25：89-90．

[6] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215 (3): 403-410.

[7] Thompson J D, Higgins D G, Gibson T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, 22 (22): 4673-4680.

[8] Vallenet D, Labarre L, Rouy Z, et al. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Research, 2006, 34(1): 53-65.

[9] Bocs S, Cruveiller S, Vallenet D, et al. AMIGene: annotation of microbial genes. Nucleic Acids Research, 2003, 31: 3723-3726.

[10] 张焕萍，宋晓峰，王惠南．基于离散粒子群和支持向量机的特征基因选择算法．计算机与应用化学，2007，24（9）：1159-1162.

[11] 徐克学．生物数学．北京：科学出版社，2001.

[12] Suzek B E, Ermolaeva M D, Schreiber M, et al. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics, 2001, 17: 1123-1130.

[13] Lowe T M, Eddy S R. tRNAScan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 1997, 25: 955-964.

[14] Gaesterland T, Sensen C W. MAGPIE: automated genome interpretation. Trends Genet, 1996, 12: 76-78.

[15] 李银山，杨春燕，张伟．DNA 序列分类的神经网络方法．计算机仿真，2003，20（2）：65-68.

[16] Overbeek R, Begley T, Butler R M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research, 2005, 33(17): 5691-5702.

[17] Overbeek R, Disz T, Stevens R. The SEED: a peer-to-peer environment for genome annotation. Communications of ACM, 2004, 47(11): 46-51.

[18] Van Domselaar G H, Stothard P, Shrivastava S, et al. BASys: a web server for automated bacterial genome annotation. Nucleic Acids Research, 2005, 33(Web server issue): 455-459.

[19] 王炼红．人工免疫优化与分类算法及其应用研究．湖南大学，2009.

[20] Frishman D, Albermann K, Hani J, et al. Functional and structural genomics using PEDANT. Bioinformatics, 2001, 17(1): 44-57.

[21] Tatusov R L, Fedorova N D, Jackson J D, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 2003, 4:41.

[22] Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25(1): 25-29.

[23] Salzberg S L, Delcher A L, Kasif S, et al. Microbial gene identification using interpolated markov models. Nucleic Acids Res, 1998, 26(2): 544-548.

[24] Wu C H, Yeh L S, Huang H, et al. The protein information resource. Nucleic Acids Res, 2003, 31(1): 345-347.

[25] QI Yutao, LIU Fang,GONG Maoguo,et al.Multi –objective immune algorithm with Baldwinian learning[J].Applied Soft Computing,2012,12(8): 2654-2674.

[26] Perelson A.S. Immune Network Theory, Imm.Rev., 1989, 110: 5-36.

[27] Xiaobo Zhou, Xiaodong Wang, Edward R. Dougherty. A Bayesian approach to nonlinear probit gene selection and classification. Journal of the Franklin Institute , 2004,341: 137-156.

[28] Cathy H. Wu. Artificial neural networks for molecular sequence analysis. Computer Chem., 1997, 21(4): 237-256.

[29] Jie Liang, Seman Kachalo. Computational analysis of microarray gene expresssion profiles: clustering, classification ,and beyond. Chemometrics and Intelligent Laboratory Systems,2002,62: 199-216.

[30] 吕志清，李前忠．用离散量预测蛋白质的结构型．生物物理学报，2001，17（4）：703- 711．

[31] CHEN Cuixia etc. The Identification of Exon Intron and Intergenic DNA in the Model Species Genomes. Acta Scientiarum Naturalium Universitatis NeiMongol, 2005, 7: 166-172.

[32] 鲍卫华，李前忠．预测线虫和酵母基因组中内含子、外显子及基因间序列的离散增量方法．内蒙古大学学报（自然科学版），2004，35（1）：60-64．

[33] 陈翠霞．五种模式生物基因序列的识别研究．内蒙古大学硕士学位论文，20040101．

[34] BursetM , Guigo R. Evaluat ion of gene structure prediction program. Genom ics, 1996, 34: 353-367.

[35] Heydt G T, Fjeld P S, Liu C C, et al. Applications of the window FFT to electric power quality assessment. IEEE Trans.on Power Delivery, 1999, 14(3): 1411-1416.

[36] Hinton, G.E., Krizhevsky, A., Srivastava, N., Sutskever, I., & Salakhutdinov, R. (2014). Journal of Machine Learning Research, 15, 1929-1958.

总结及展望

机器学习是人工智能的一个分支，能让计算机直接从样本、数据和经验中进行学习。通过让计算机智能地完成特定任务，机器学习系统能通过学习数据执行复杂的流程，而不是提前编程规则。近年来，我们看到了机器学习的惊人发展，有能力完成各种应用。数据可用性的增加使得机器学习系统能在大型的样本池上进行训练，计算处理能力的增加支撑了这些系统的分析能力。在此领域内，算法的进步也赋予了机器学习更强大的能力。这些进步带来的结果就是，几年前还低于人类能力的系统，现在在特定任务上已经超过了人类的水平。如今，许多人每天都会和基于机器学习的系统进行交互，例如社交媒体中使用的图像识别系统；虚拟助手使用的语音识别系统；在线零售商使用的推荐系统。随着该领域的进一步发展，机器学习展现出了能够支持大多领域转型、发展的潜力，带来的社会与经济机遇是巨大的。在医疗领域，机器学习正在创造能够帮助医生进行高效、准确诊断的系统；在交通领域，它支持了自动驾驶的开发，助力让现有交通网络更高效。对公共服务而言，它有潜力进行更高效的目标定位，以及零售服务的目标定位。在科学领域，机器学习正在帮助研究人员理解大量的数据，提供对生物学、物理学、医疗和社会科学等学科的新洞见。

机器学习是一个充满活力的研究领域，具有一系列令人兴奋的研究方向，在未来会通过不同的方法和应用进一步发展下去。除了纯技术问题的研究以

外，在机器学习领域里还有一些公众非常关心的议题，或是对其广泛使用的约束。因此，支持对于机器学习的研究，可以确保公众对于部署机器学习系统的信心。可以开展的研究包括算法的可解释性、鲁棒性、隐私、公平性、因果关系推理、人机交互和安全等方面。近年来，人们不断从生物系统中获得灵感，提出了若干仿生智能计算系统，包括人工神经网络、遗传算法、蚁群系统及 DNA 计算等。生物免疫系统同样是一个高度进化的生物系统，其通过分布式任务处理能力和局部采取行动的智能，也通过起交流作用的化学信息构成网络，进而形成全局观念。人工免疫系统正是在研究借鉴、利用生物免疫系统信息处理机制的基础上发展而来的。它通过学习自然免疫系统防御机理的学习计算技术，提供噪声忍耐，无教师学习、自组织、记忆等进化学习机理，基于免疫学的智能计算方法结合先验知识和免疫系统适应能力。它具备了分类器、神经网络和机器学习等系统的优点，因此具有提供新颖的解决问题的方法的潜力，其研究成果已经涉及了控制、数据处理、模式识别、优化学习和故障诊断等许多领域。算法设计是免疫机器学习方法的核心，从人工免疫系统产生至今，以及将来很长一段时间内，算法设计都将是免疫计算研究者所关注的主要方向之一。本书内容主要围绕免疫计算在机器学习、数据挖掘等领域的算法设计及工程应用问题，对免疫计算智能的生物学原理、系统模型、应用原理及工作机制进行了探讨，在此基础上重点对免疫计算机器学习进行有益探索，结合应用问题提出多种改进的机器学习软计算方法，对其进行了理论阐述及应用分析。针对现实复杂系统中的数据复杂、异构及不完备性，结合人工免疫系统中优异的数据聚类、分类方法，研究了多源属性约简、知识发现、含糊知识模糊化的免疫计算方法，将免疫计算智能应用于工程实践，进一步拓展理论与应用的结合。由此而生的机器学习方法，为解决现实问题及数据分析提供了一个强有力、很有前途的工具。本书以免疫计算智能这一新兴的生物启发式智能技术为研究基础，其多样性及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是很有效的情况，还能有效避免早熟现象，在数据处理及优化方面显示出良好的潜力。从工程上讲，它具有结合先验知识和免疫系统的适应能力；从信息科学上讲，

它具有强壮的鲁棒性和预处理能力；因此，对免疫计算的生物启发式智能技术的深入研究将在我们的机器学习过程中提供新的思路和解决办法。

目前，大数据浪潮正对人类社会生活、科学研究的方方面面产生深刻影响。早期机器学习研究通常假设数据具有相对简单的特性，如数据来源单一、概念语义明确、数据规模适中、结构静态稳定等。当数据具有以上简单特性时，基于现有的基于免疫计算等仿生智能的机器学习理论与方法可以有效实现数据的智能化处理。然而，在大数据时代背景下，数据往往体现出多源异构、语义复杂、规模巨大、动态多变等特殊性质，为传统机器学习技术带来了新的挑战。为应对这一挑战，国内外科技企业巨头，如谷歌、微软、亚马逊、华为、百度等，纷纷成立以机器学习技术为核心的研究院，以充分挖掘大数据中蕴含的巨大商业与应用价值。可以预见，在未来相当长的一段时期内，机器学习领域的研究将以更广泛、更紧密的方式与工业界深度耦合，推动信息技术及产业的快速发展。同时，国际上关于机器学习的主要学术会议包括每年定期举行的国际机器学习会议（ICML）、国际神经信息处理系统会议（NIPS）、欧洲机器学习会议（ECML）及亚洲机器学习会议（ACML）等，主要学术期刊包括 *Machine Learning*、*Journal of Machine Learning Research*、*IEEE Transactions on Neural Networks and Learning Systems* 等。此外，人工智能领域的一些主要国际会议（如 IJCAI、AAAI 等）和国际期刊（如 *Artificial Intelligence*、*IEEE Transactions on Pattern Analysis and Machine Intelligence* 等）也经常发表与机器学习相关的最新研究成果。国内机器学习的重要学术活动包括每两年举行一次的中国机器学习会议（China Conference on Machine Learning, CCML）和每年举行的中国机器学习及其应用研讨会（Chinese Workshop on Machine Learning and Applications，MLA），该会议遵循“学术至上，其余从简”的原则，每届会议邀请海内外从事机器学习及相关领域研究的多位专家与会进行学术交流，包括特邀报告、与会交流及 Top Conference Review 等部分。产业发展与学术研究共同推进，大大促进了机器学习技术的普及及应用。

经过多年的发展，互联网已获得巨大的成功。由此，人们可以在不同时间与地域获取自己希望获得的数据。随着数据量的激增，如何有效获得并通过机器学习技术来更好地利用这些数据已成为信息产业继续兴旺发展的关键。因此，机器学习算法和技术就成为解决这类问题的有力工具。在中小规模问题上，机器学习已经从理论研究阶段逐渐上升到了实际应用阶段。但是在大规模的实际应用中，特别是在大数据环境下的大数据体量大、结构多样、增长速度快、整体价值大而部分价值稀疏等特点，对数据的实时获取、存储、传输、处理、计算与应用等诸多方面提出了全新挑战。传统的面向小数据的机器学习技术已很难满足大数据时代下的种种需求，并且使用单个计算单元进行运算的集中式机器学习算法难以在大规模的运算平台上执行。因此，在大数据时代，突破传统的思维定式和技术局限，研究和发展革命性的、可满足时代需求的并行机器学习的新方法和新技术，从大数据中萃取大价值，具有重要的学术和应用价值。目前，很多机器学习应用非常广泛的领域都已经面临了大数据的挑战。如互联网和金融领域，训练实例的数量是非常大的，每天会有几十亿事件的数据集。另外，越来越多的设备包括传感器，持续记录观察的数据可以作为训练数据，这样的数据集可以轻易地达到几百 TB。再如亚马逊或淘宝上的商品推荐系统，每天都有很多用户看到很多推荐的商品，并且进行了点击操作。这些用户点击推荐商品的行为会被亚马逊和淘宝的服务器记录下来，作为机器学习系统的输入。输出是一个数学模型，可以预测一个用户喜欢看到哪些商品，从而在下一次展示推荐商品的时候多展示用户喜欢的。类似的，在互联网广告系统中，展示给用户的广告及用户点击的广告也都会被记录下来，作为机器学习系统的数据，训练点击率预估模型。在下一次展示推荐商品时，这些模型会被用来预估每个商品如果被展示之后有多大的概率被用户点击。从这些例子我们可以看出来，这些大数据之所以大，是因为它们记录的是数十亿互联网用户的行为。而人们每天都会产生行为，以至于百度、阿里、腾讯、奇虎、搜狗这些公司的互联网服务每天都收集到很多块硬盘才能装下的数据。而且这些数据随时间增加，永无止境。传统机器学习技术在大数据环境下的低效率，以及大数据分布式存储的特点，使得

现代机器学习技术成为了解决从大规模、海量数据中学习的重要途径。因此，机器学习近期的成功很大一部分归因于一些领域的数据爆炸，例如图像或语音识别。这些数据提供了大量的样本，机器学习可使用它们改进自己的表现。作为回报，通过先进的数据分析提取有价值的信息，机器学习能帮助人们获得从所谓的“大数据”中期望的社会与经济收益。围绕着免疫计算智能方法研究，今后将在下述三个方向继续深入下去。

（1）发现新的思路，构造新的算法。从方法论上讲，算法研究主要是从更宏观、更本质的角度模拟自然免疫原理与机制，模拟生物智能的形成过程，并求解问题，进而融合数学、生物、计算机技术等各领域的原理与技巧，使所设计的算法执行策略有预期的特性，更加有效。

（2）免疫计算智能算法与进化算法等其他智能算法的对比研究。本书虽然在一定程度上进行了这方面的研究和实践，并应用于具体实践且取得了满意效果，但还需加强免疫系统与其他人工智能之间的融合与发展，研究基于免疫系统机制的智能系统理论和技术，开发新型自然计算技术和软计算技术，尤其是面对大数据环境下新的数据处理方法的设计和开发。

（3）应用研究与实现。面对大数据时代的到来，传统的数据处理方式面临着新的严峻挑战，大数据时代的大量化、多样化、快速化和价值密度低等特点让传统的搜索方法和工具有时只能望“数据”兴叹。只有面向大数据的技术不断发展，才能将大数据时代带来的挑战变为机遇，更好地运用这个重大战略资源，并有效构建相适应的数学模型和工具，真正将海量数据变化为有效信息。因此，将有关免疫系统的应用研究扩大到大数据应用领域，根据新的需求改进策略，使之体现出更大的经济效益和社会效益。

同时，面向未来更多的应用挑战和技术调整，新的机器学习理论、方法和算法会不断发展。例如，分布式机器学习、并行机器学习、哈希学习、深度学习等先进机器学习技术和方法是随着“大数据”概念和“云计算”的普及而得到迅速发展的。大数据给机器学习带来了需求；云计算给机器学习带来了条件。可以看到，面向数值计算的统计学习和以神经元网络为代表的深度

学习是现代人工智能的两个主要分支。而在大数据+云计算的时代，这两大分支都进入了新的发展黄金期。因此，随着机器学习系统在某些领域变得越来越普及与重要，我们需要有三种技能。首先，随着与机器学习的日常互动成为大多数人的常态，对数据和机器学习系统的了解与使用成为所有人群和背景所需要的重要工具。在学校介绍机器学习的关键概念有助于保障这一点。其次，为了确保各个领域和职业有能力以一种对它们有用的方式使用吸收和使用机器学习，我们需要新的机制来使用户或实践者获得足够的信息。最后，我们需要进一步的支持来让人们获得机器学习的高级技能。机器学习应用可以在特定任务上实现良好的表现。在许多案例中，人类都可以使用机器学习来增强自己的能力。尽管机器学习的发展很显然将会改变就业，但预测其实际的发生方式却并不简单，现有的研究也都给出了各自不同的预测。尽管机器学习有望给社会经济发展带来新的业务或领域，但其颠覆性的影响也将给社会带来挑战及关于其社会后果的质疑。其中一些挑战涉及数据的新兴使用方式，将重塑隐私和许可的传统概念，而其他一些挑战还涉及人们与机器的交互方式。我们需要谨慎的管理工作来确保社会中的所有人都能受益于机器学习所带来的生产力红利。