

免疫优化理论及其应用

张瑜 主编

湖北科学技术出版社



前言

自然界一直是人类各种技术思想、工程原理及重大发明的灵感源泉。层次各异、种类繁多的生物界经过长期的“优胜劣汰、适者生存”的自然进化过程,逐渐适应自然环境的变化,并得以生存与发展。人类也是自然界长期进化的产物。一部人类与人类智能的发展史更是一部仿生学发展史。

在仿生学研究中,人工智能是其研究的主攻方向。在人工智能研究领域,计算智能是一个主要研究方向,它以生物进化的观点认识和模拟智能,以数据为基础,通过训练建立联系,进行问题求解。而在计算智能领域,人工免疫系统理论与算法又是一个研究热点,已经引起了国内外学者的广泛关注和极大兴趣。

人工免疫系统以生物免疫系统的原理、机制和特性为理论基础,实现类似于生物免疫系统的抗原识别、细胞分化、记忆和自我调节的功能,主要应用于网络安全、计算机病毒检测、入侵检测、优化计算、模式识别等领域。本书汇总了作者在免疫优化算法、模型和理论等方面的基础研究,以及在恶意代码检测等网络安全应用方面的研究成果。

本书的研究内容和主要成果归纳如下:

(1)提出了抗体、MHC 和抗原三位一体的免疫进化算法体系。本算法体系突破了传统免疫算法的抗体/抗原二元结构,更全面地模拟生物免疫系统中抗体在 MHC(Major Histocompat-

ibility Complex, 主要组织相容性复合体) 分子辅助下学习、记忆、识别和清除抗原的机制, 提高了免疫进化算法的性能和效率, 拓展了人工免疫系统算法研究思路。

(2) 提出了基于 MHC 的抗体克隆算法通用框架。该算法通用框架的主要特点是: ①利用 MHC 分子的单倍型特性来保留优秀的父代抗体基因, 并利用疫苗接种遗传给子代, 以增强局部搜索能力并提高收敛速度; ②利用 MHC 分子的多基因性和多态性以及随机引入部分新抗体来生成多样性的抗体群体, 以防止群体早熟, 从而增强全局搜索能力。

在此算法通用框架的基础上, 针对多模态函数优化问题, 设计了新的免疫优化算法 MOAMHC。定义了抗体、MHC、抗原、亲和力和相似性等基本概念, 设计了免疫选择算子、接种疫苗算子、变异算子以及 MHC 移位算子、MHC 替换算子、MHC 交叉算子及 MHC 变异算子。利用马尔可夫链理论分析了该算法的收敛性问题。MOAMHC 算法的特点: ①利用 MHC 单倍型特性来保存优秀抗体基因片段, 以使优势基因得以在群体中扩散, 可搜索局部最优值、提高群体的收敛速度; ②利用 MHC 的多态性和多基因性直接促进抗体群的多样性, 防止群体早熟, 从而确保算法结束时搜索到全局最优值。通过 14 个基准测试函数, 对比了 MOAMHC 算法和遗传算法 SGA、典型克隆选择算法 CLONALG 的性能。实验表明, MOAMHC 算法在求解精度、收敛速度、稳定性和解的多样性等方面, 要优于 SGA 算法和 CLONALG 算法。

针对组合优化问题, 设计了基于 MHC 的免疫优化算法 COAMHC。定义了抗体、MHC、抗原、亲和力、相似性等基本概念, 设计了免疫选择算子、接种疫苗算子、变异算子以及 MHC 系列算子。利用马尔可夫理论分析了 COAMHC 算法的收敛性问题。通过 5 个典型的 TSP 问题的求解, 对比了 COAMHC 算法和

克隆选择算法 CLONALG 的性能。实验表明,COAMHC 算法在收敛速度和求解精度方面,要优于 CLONALG 算法。

(3)提出了基于 MHC 的恶意代码检测模型 MCDMHC。给出了模型中自体、非自体、MHC 基因、抗体、检测器、抗原提呈、免疫识别等数学定义,建立了自体/非自体、MHC 基因库、检测器自体耐受等动态演化模型。本检测模型的特点:①建立自体与非自体集动态演化模型,以解决自体集合完整性问题;②构建 MHC 基因库,利用基于 MHC 的抗体克隆算法动态生成检测器中的抗体;③建立检测器集合的动态演化模型,以解决检测器集合完备性问题。本模型利用疫苗接种来初始化未成熟检测器,提高检测器生成效率。在 MHC 基因库进化的基础上利用抗体克隆算法生成有限的获得非自体空间大规模覆盖的检测器集,从而克服了传统的基于免疫的恶意代码检测方法中因检测器集的完备性和自体集的完整性不够而导致的检测率低和误报率高的缺陷。与相关恶意代码检测方法进行的对比实验表明,本模型能对未知恶意代码进行高效检测,从而为恶意代码检测研究提供了一种新方法,为将生物免疫机制和特性应用于解决具体网络安全问题提供了一种新思路。

本书研究和探索生物免疫系统中抗体在 MHC 分子辅助下学习、记忆、识别并清除外来入侵抗原的机制和 MHC 分子所具有的特性(单倍型特性、多态性和多基因性)的基础上,针对数值优化、组合优化以及恶意代码检测等问题展开了深入研究,提出了基于 MHC 的抗体克隆算法和基于 MHC 的恶意代码检测模型,为人工免疫系统算法和模型研究提供了新思路。探索将生物免疫系统原理和免疫智能优化算法应用于解决恶意代码检测等网络安全实际问题,对解决目前的网络安全所面临的困境,具有重要的理论价值和广阔的应用前景。

本书共分为 6 章,每章的主要内容如下:

第1章绪论。介绍本书的研究背景、恶意代码检测研究现状及存在的问题、研究目的、研究意义、课题来源以及主要研究内容。

第2章人工免疫系统概况。首先,介绍了生物免疫系统的组成、机理,MHC分子特性以及克隆选择学说;然后,介绍了人工免疫系统的模型、算法和应用。

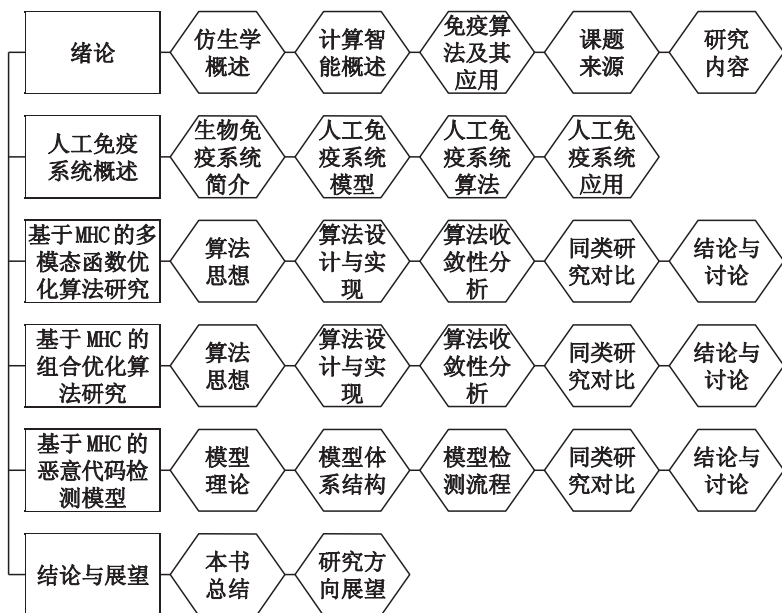
第3章基于MHC的多模态函数优化算法研究。首先,介绍了目前解决多模态函数优化问题所使用的方法及其优、缺点;其次,介绍了基于MHC的多模态函数优化算法的思想与具体设计实现;再次,理论分析了该算法的收敛性;最后,进行了数值仿真实验并与遗传算法和克隆选择算法做了对比,并对实验结果做了分析和讨论。

第4章基于MHC的组合优化算法研究。首先,介绍了目前组合优化算法所存在的问题;其次,介绍了基于MHC的组合优化算法的设计思想与具体的算法实现,包括算法的基本定义、算子的构造、算法的基本流程;然后,从理论上证明了算法的收敛性;最后,就TSP问题进行了同类算法实验对比,并对实验结果做了分析和讨论。

第5章基于MHC的恶意代码检测模型。本章详细介绍了模型的基本定义,包括自体、非自体、MHC基因、MHC基因库、抗体、检测器、抗原提呈、免疫识别等;然后,介绍了模型的体系结构、模型的检测流程以及模型的性能分析;最后,进行了恶意代码检测实验,并进行了同类研究对比,验证了该模型在未知恶意代码检测方面的有效性。

第6章结论与展望。本章对全书的研究成果进行了总结,分析了研究中存在的问题和不足,展望了今后的研究方向。

本书的结构如下图所示:



本书的出版是基于我的博士课题研究和近年来主持的海南省自然科学基金项目(610220)、浙江省信息安全重点实验室项目(2010ZISKL007)、海南师范大学博士科研项目(00203020214)等研究项目资助。尤其要感谢我的博士生导师、四川大学计算机学院李涛教授,感谢海南师范大学信息学院吴丽华教授、马生全教授,感谢浙江省电子产品检验所宋丽萍女士,感谢湖北科学技术出版社,感谢海南师范大学信息学院的领导与同事对我的大力支持与帮助。

作 者

2010 年 11 月

目 录

第 1 章 绪论	1
1.1 仿生学概述	1
1.2 计算智能概述	3
1.3 免疫算法及其应用	4
1.3.1 免疫算法的发展	4
1.3.2 免疫算法的类型	6
1.3.3 免疫算法的应用	7
1.4 研究内容	10
第 2 章 人工免疫系统	16
2.1 生物免疫系统简介	16
2.1.1 生物免疫系统组成	16
2.1.2 生物免疫系统机理	20
2.1.3 抗体克隆选择原理	23
2.1.4 MHC 的概念与特性	24
2.2 人工免疫系统概况	26
2.2.1 人工免疫系统模型	28
2.2.2 人工免疫系统算法	30
2.2.3 人工免疫系统应用	35

2.2.4	人工免疫系统展望	37
2.3	本章小结	39
第3章	基于 MHC 的多模态函数优化算法研究	40
3.1	基于 MHC 的抗体克隆算法	40
3.2	多模态函数优化问题概述	43
3.3	MOAMHC 算法设计与实现	44
3.3.1	MOAMHC 算法思想	44
3.3.2	MOAMHC 算法的基本定义	44
3.3.3	MOAMHC 算法的算子构建	46
3.3.4	MOAMHC 算法流程	48
3.3.5	MOAMHC 算法收敛性分析	50
3.4	仿真实验与结果分析	52
3.4.1	基准测试函数	52
3.4.2	实验参数设置	55
3.4.3	对比实验结果及分析	55
3.5	结论与讨论	67
3.6	本章小结	69
第4章	基于 MHC 的组合优化算法研究	70
4.1	组合优化问题概述	70
4.2	COAMHC 算法设计	71
4.2.1	COAMHC 算法思想	71
4.2.2	COAMHC 算法的基本定义	72
4.2.3	COAMHC 算法的算子构建	74
4.2.4	COAMHC 算法流程	76
4.2.5	COAMHC 算法收敛性证明	78
4.3	仿真实验与结果分析	80

4.3.1	基准测试问题	80
4.3.2	实验参数设置	81
4.3.3	对比实验结果及分析	81
4.4	结论与讨论	86
4.5	本章小结	86
第5章 基于MHC的恶意代码检测模型		88
5.1	研究背景	88
5.2	恶意代码检测研究现状	93
5.3	存在的主要问题	96
5.4	发展趋势	98
5.5	研究目的	99
5.6	研究意义	101
5.7	恶意代码定义	101
5.7.1	计算机病毒	101
5.7.2	蠕虫	103
5.7.3	木马	103
5.8	模型理论	104
5.8.1	问题域	104
5.8.2	模型定义	106
5.9	模型体系结构与检测流程	118
5.9.1	模型体系结构	118
5.9.2	恶意代码检测流程	120
5.9.3	模型性能分析	121
5.10	实验与结果分析	123
5.10.1	实验数据集	123
5.10.2	实验结果及分析	124
5.10.3	同类研究比较	125

5. 11 本章小结	127
第 6 章 结论与展望	128
6. 1 本书主要贡献	128
6. 1. 1 基础研究	128
6. 1. 2 应用研究	129
6. 2 本书创新点	129
6. 3 进一步研究展望	130
参考文献	132
作者论著及科研项目	150

第 1 章 绪 论

1.1 仿生学概述

古往今来,自然界一直是人类各种技术思想、工程原理及重大发明的灵感源泉。层次各异、种类繁多的生物界经过长期的“优胜劣汰、适者生存”的自然进化过程,逐渐适应自然环境的变化,并得以生存与发展。人类也是自然界长期进化的产物。在长期的生产实践中,人类以自己直立的身躯、能劳动创造的双手、交流情感和思想的语言,促进了神经系统尤其是大脑的高度发展。然而,人类的智慧不仅仅停留在观察和认识生物界的现象与规律上,而且还能运用人类所独有的思维和设计能力模仿生物的形态结构与功能,通过创造性的劳动来延展自身智能和扩展自身的能力。因此,我们可以这样来描述:人类源于自然,人类智能源于自然,人类创造能力更是源于自然。这也是仿生学一直不断发展、不断拓展新的应用领域的深层次根源。一部人类与人类智能的发展史更是一部仿生学发展史。

人类模仿鱼类的形体造船,以木桨仿鳍。相传早在大禹时期,我国古代劳动人民观察鱼在水中用尾巴的摇摆而游动、转弯,就在船尾上安置木桨。通过反复地观察、模仿和实践,逐渐改成橹和舵,增加了船的动力,掌握了使船转弯的手段。这样,即使在波涛滚滚的江河中,人们也能让船只航行自如。

人类模仿鸟类的形体制造飞行器,以机翼仿翅。据《韩非子》记载,鲁班用竹木作鸟,“成而飞之,三日不下”。意大利著名科学家兼画家达·芬奇和他的助手对鸟类进行仔细解剖,研究鸟的身体结构并认真观察鸟类的飞行,设计和制造了世界上第一架人造飞行器——扑翼机。

人类模仿鲸鱼的体形和鱼类的沉浮系统——鱼鳔,制造了潜水艇;模仿蝙蝠、海豚的回音定位系统而发明了声呐系统;模仿苍蝇的楫翅(又叫平衡棒)而制成了广泛应用于火箭和高速飞机上实现自动驾驶功能的振动陀螺仪;利用蝴蝶的色彩在花丛中不易被发现的道理,发明了迷彩服,在军事设施上覆盖蝴蝶花纹般的伪装;模仿蜻蜓的飞行原理而发明了直升机;模仿蜂类的蜂窝构造原理而发明了覆盖面广、节省材料的蜂窝式 GSM 移动通信系统;根据响尾蛇的颊窝能感觉到 0.001°C 的温度变化的原理,人类发明了跟踪追击的响尾蛇导弹;人类还利用蛙跳的原理设计了蛤蟆夯;人类模仿警犬的高灵敏嗅觉制成了用于侦缉的“电子警犬”;科学家根据野猪的鼻子测毒的奇特本领制成了世界上第一批防毒面具。

以上这些模仿生物形态构造和功能的发明与尝试,是人类仿生学的具体应用。随着科学技术的发展和人类社会的进步,将会出现越来越多的仿生产品应用于生产活动中。那么,什么是仿生学?

仿生学是研究生物系统的结构和性质,为工程技术提供新的设计思想及工作原理的科学。它是在 20 世纪中期才出现的一门新的边缘科学。仿生学研究生物体的结构、功能和工作原理,并将这些原理移植于工程技术之中,发明性能优越的仪器、装置和机器,创造新技术。

仿生学 Bionics 一词是 1960 年美国斯蒂尔根据拉丁文“bios”(生命方式的意思)和字尾“nic”(“具有……的性质”的意思)构成

的。仿生学的研究主要包括力学仿生、分子仿生、能量仿生、信息与控制仿生等。其中,信息与控制仿生是仿生学的主要领域。一方面由于自动化向智能控制发展的需要;另一方面是由于生物科学已发展到这样一个阶段,使研究大脑已成为对神经科学最大的挑战。人工智能和智能机器人研究是仿生学研究的主攻方向。

1.2 计算智能概述

人工智能主要研究用人工的方法和技术,模仿、延伸和扩展人的智能,实现机器智能。可把人工智能分成两大类:一类是符号智能,一类是计算智能。符号智能是以知识为基础,通过推理进行问题求解,即所谓的传统人工智能。计算智能是以数据为基础,通过训练建立联系,进行问题求解。

计算智能是以生物进化的观点认识和模拟智能。按照此观点,智能是在生物的遗传、变异、生长以及外部环境的自然选择中产生的。在用进废退、优胜劣汰的自然进化过程中,适应度高的(头脑)结构被保存下来,智能水平也随之提高。因此,可认为计算智能就是基于结构演化的智能。计算智能的主要方法有人工神经网络、遗传算法、遗传程序、演化程序、局部搜索、模拟退火等等。这些方法具有以下共同的要素:自适应的结构、随机产生的或指定的初始状态、适应度的评测函数、修改结构的操作、系统状态存储器、终止计算的条件、指示结果的方法、控制过程的参数。计算智能的这些方法具有自学习、自组织、自适应的特征和简单、通用、鲁棒性强、适于并行处理的优点,在并行搜索、联想记忆、模式识别、知识自动获取等方面得到了广泛的应用。

典型的代表,如遗传算法、免疫算法、模拟退火算法、蚁群算法、微粒群算法等仿生算法,都是基于“智慧源于自然”的理念,通过人们对自然界独特规律的认知,提取出适合获取知识的一套计算工具。而免疫算法是近年来在计算智能领域新兴的一类仿生

算法,它基于生物免疫系统基本机制,从体细胞理论和网络理论得到启发,实现类似于生物免疫系统的抗原识别、细胞分化、记忆和自我调节的功能。免疫算法具有自组选择学习、全息容错记忆、辩证克隆仿真和协同免疫优化的启发式人工智能。由于该算法具有收敛速度快、求解精度高、稳定性能好,并有效克服了早熟和欺骗的问题,成为一种新兴的实用智能算法。

1.3 免疫算法及其应用

生物免疫系统(Biology Immune System, BIS)是一个分布式、自组织和具有动态平衡能力的自适应复杂系统。它对外界入侵抗原(Antigen, Ag),可由分布全身的不同种类的淋巴细胞产生相应的抗体(Antibody, Ab),其目标是尽可能保证整个生物系统的基本生理功能得到正常运转。人工免疫系统(Artificial Immune System, AIS)就是研究、借鉴、利用生物免疫系统的原理、机制而发展起来的各种信息处理技术、计算技术及其在工程和科学中的应用而产生的多种智能系统的统称。从生物信息处理的角度看,可归为信息科学范畴,是计算智能研究领域一个新的热点内容与研究方向。

免疫算法是基于免疫系统的学习算法,是人工免疫系统研究的主要内容之一。它具有良好的系统应答性和自主性,对干扰具有较强维持系统自平衡的能力。免疫算法模拟免疫系统独有的学习、记忆、识别等功能,具有较强模式分类能力,尤其对多模态问题的分析、处理和求解表现出较高的智能性和鲁棒性。

1.3.1 免疫算法的发展

免疫算法和免疫系统的理论与应用的研究历史较短。相关理论可追溯到1958年澳大利亚学者 Burnet^[1]率先提出了克隆选择原理,1960年因此获得诺贝尔奖。1974年,诺贝尔奖得主、生

物学家、医学家、免疫学家 Jerne^[2] 提出了免疫网络理论并建立了数学模型,奠定了免疫计算的基础。之后 Farmer、Perelson、Bersini、Varela 等学者分别在 1986 年、1989 年和 1990 年发表了有关论文,在免疫系统启发实际工程应用方面做出了突出贡献,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开辟了道路。1990 年,Bersini^[3,4] 首次使用免疫算法来解决问题。20 世纪末,Forrest 等^[5,6] 开始将免疫算法应用于计算机安全领域。近年来,免疫理论和算法已经引起了相关研究人员的极大关注。

1996 年,在日本举行了基于免疫系统的国际专题讨论会,首次提出了“人工免疫系统”的概念。1997 年和 1998 年 IEEE Systems, Man and Cybernetics 国际会议还组织了相关专题讨论,并成立了“人工免疫系统及应用分会”。目前,世界上绝大多数人工免疫系统研究成果出自美国、英国、中国和日本。而巴西 Campinas 大学的 De Castro^[7-9] 博士最早在其博士论文中总结了人工免疫系统,并试图建立人工免疫系统的统一框架结构。取得显著成绩的主要有:利用免疫系统原理研究计算机安全的美国 New Mexico 大学计算机科学系的 Forrest^[5,6] 博士;研究基于免疫原理的计算机安全和异常检测及工业应用的 Missouri 大学计算机与数学系的 Dasgupta^[10,11] 博士;研究数据分析的英国 Kent 大学的 Timmis^[12-16] 博士;研究计算机网路入侵的 King's 学院的 Kim^[17] 博士;威尔士大学 Hunt 和 Cooke^[18-20] 领导的 ISYS 研究小组等。

在我国,西南交通大学的靳蕃教授等^[21] 在 1990 年前后就已经指出“免疫系统所具有的信息处理与肌体防卫功能,从工程角度来看,具有非常深远的意义”。目前,国内在免疫优化算法理论与应用方面的研究成果具有鲜明特色。其中,中国科学技术大学王煦法^[22-24] 教授在国内较早开展免疫优化算法方面的研究,并

扩展到基于免疫计算的人工生命、进化硬件领域;西安电子科技大学焦李成^[25-31]教授的科研团队提出了具有较完备理论基础的免疫克隆算法及一系列改进,也在国际上较早提出了新颖的免疫遗传算法;四川大学李涛教授^[32-53]的科研团队提出了基于免疫动态克隆选择算法。此外,哈尔滨工程大学莫宏伟^[54,55]博士科研团队、东华大学的丁永生教授^[56-59]科研团队,都先后对免疫算法及理论进行了研究,并提出了一系列免疫算法;重庆大学的黄席樾^[60-63]教授系统总结了人工免疫系统理论;华中科技大学肖人彬教授提出了工程免疫计算^[64,65];北京信息科技大学杨孔雨教授总结了免疫进化理论^[66,67]。笔者提出了基于 MHC 的抗体克隆算法等免疫优化算法。

1.3.2 免疫算法的类型

免疫算法是基于免疫机理提出的高效学习和优化算法,是人工免疫系统理论研究的重要内容之一。常见的免疫算法类型有:

1. 克隆选择算法(CSA;Clone Selection Algorithm)

由于生物免疫系统本身的复杂性,有关算法机理的描述还不多见,相关算法还比较少。Castro L D、Kim J、杜海峰、焦李成等基于抗体克隆选择机理相继提出了克隆选择算法。Nohara 等基于抗体单元的功能提出了一种非网络的人工免疫系统模型。而目前两个比较有影响的人工免疫网络模型是 Timmis 等基于人工识别球(Artificial Recognition Ball, AR)概念提出的资源受限人工免疫系统(Resource Limited Artificial Immune System, RLAIS)和 Leandro 等模拟免疫网络响应抗原刺激过程提出的 aiNet 算法。

2. B 细胞网络算法(B-cell Algorithm)

以独特型网络理论和克隆选择理论为基础。Hunt 等人模拟生物免疫系统的自学习、自组织机理提出一种人工免疫网络模

型——B细胞网络及其算法。

算法将未知问题的解看作抗原,认为只要找到能产生最高亲和度抗体的B细胞,也就找到了未知问题的解。实验结果证明,该算法具有较强的寻优能力,并保持网络中多种模式和谐并存,有比人工神经网络更快、更好的模式识别能力。

3. 否定选择算法(NSA: Negative Selection Algorithm)

否定选择算法基于生物免疫系统的特异性,借鉴生物免疫系统T细胞生成时的“否定选择”(Negative Selection)过程。Forrest研究了一种用于检测数据变化的阴性选择算法,用于解决计算机安全领域的问题。该算法通过系统对异常变化的成功监测而使免疫系统发挥作用,而监测成功的关键是系统能够分清自体和非自体的信息;随机产生检测器,删除那些测到自体的检测器,以使那些测到非自体的检测器保留下来。

4. 免疫遗传算法(IGA: Immune Genetic Algorithm)

免疫遗传算法可以看做一种新型融合算法,是一种改进的遗传算法,是具有免疫功能的遗传算法。该算法由西安电子科技大学焦李成教授提出。

1.3.3 免疫算法的应用

免疫算法是模仿免疫系统抗原识别、抗原与抗体结合及抗体产生过程,并利用免疫系统多样性和记忆机理抽象得到的一种智能算法。

人工免疫系统(Artificial Immune System, AIS)是受生物免疫系统启发,模拟生物免疫系统功能、机理和特性来解决复杂问题的自适应系统。

生物免疫系统所具有的学习与记忆、特征提取、模式识别、模糊与概率检测以及多样性、鲁棒性、自适应性、动态性等有效机制与特性,对于人工信息系统的安全保障显然具有重要的借鉴意