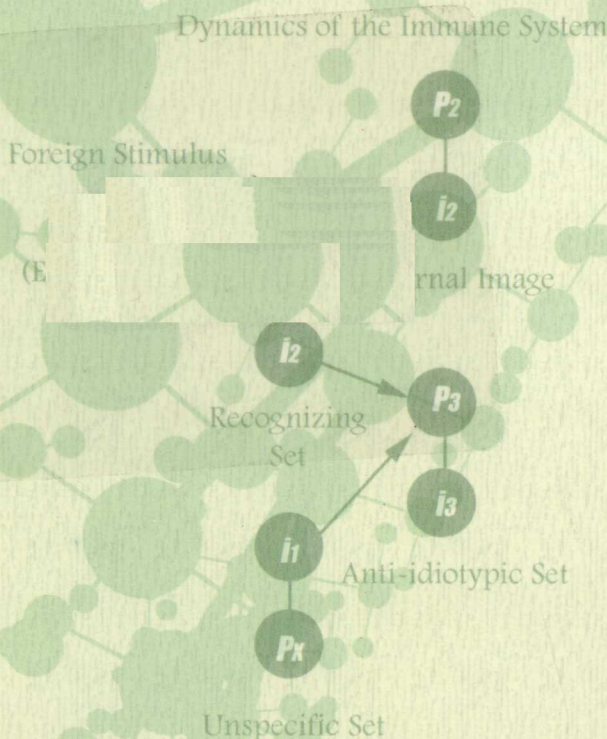


人工免疫系统及其 数据挖掘应用研究

刘 韬 著



中国矿业大学出版社

China University of Mining and Technology Press

人工免疫系统及其数据挖掘应用研究

刘 韬 著

中国矿业大学出版社

内 容 提 要

本书是作者在人工免疫系统及其数据挖掘应用领域研究工作的总结。首先讨论生物免疫学的机理,从信息处理的角度分析人工免疫系统的基本框架及其数据挖掘的基本模式,重点阐述基于距离浓度免疫算法及其数据分类、关联规则和聚类挖掘方法,探讨改进的资源有限人工免疫系统模型及其关联规则数据挖掘方法,设计自适应免疫应答网络模型及其聚类数据挖掘算法,讨论基于人工免疫系统的决策支持系统结构及其在煤炭生产决策中的应用。

本书可以为计算机学科、信息学科、人工智能和自动化技术等领域从事人工免疫系统研究的相关专业技术人员提供参考,也可以作为相关专业研究生和高年级本科生教材。

图书在版编目(CIP)数据

人工免疫系统及其数据挖掘应用研究/刘韬著. —徐州:

中国矿业大学出版社, 2010. 1

ISBN 978 - 7 - 5646 - 0515 - 5

I. 人… II. 刘… III. 免疫学—应用—人工智能—数据
采集—研究 IV. TP18

中国版本图书馆 CIP 数据核字(2009)第 199060 号

书 名 人工免疫系统及其数据挖掘应用研究
著 者 刘 韬
责任编辑 章 毅 付继娟
责任校对 周俊平
出版发行 中国矿业大学出版社
(江苏省徐州市解放南路 邮编 221008)
营销热线 (0516)83885307 83884995
网 址 <http://www.cumtp.com> E-mail: cumtpvip@cumtp.com
排 版 徐州中矿大印发科技有限公司排版中心
印 刷 徐州中矿大印发科技有限公司
经 销 新华书店
开 本 850×1168 1/32 印张 7.5 字数 192 千字
版次印次 2010 年 1 月第 1 版 2010 年 1 月第 1 次印刷
定 价 28.00 元

(图书出现印装质量问题,本社负责调换)

前 言

免疫系统是生物系统的重要子系统之一,它通过不同种类抗体的表现形式构造自体与非自体间的非线性网络,在动态变化环境中对维护生物体的健康起着重要作用。免疫系统内没有中心控制器官,而是由体内不断循环的分散免疫细胞集合构成的,能区别自体与非自体,并具有排除和记忆非自体的功能。从信息处理的观点看,生物免疫系统是一个并行分布式的自适应系统,它具有认知、特征抽取、多样性、学习和记忆等特点。人工免疫系统是研究、借鉴和利用生物免疫系统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的应用而产生的各种智能系统的统称。开展人工免疫系统的理论和应用研究,在理论和实践上都具有非常积极的意义。数据挖掘也称为数据库中的知识发现(Knowledge Discovery in Database, KDD),是指发现隐藏在数据库大量信息中有用的知识。开展人工免疫系统在数据挖掘领域的应用研究,可以丰富数据挖掘方法和理论,获得更有效的数据挖掘技术,更好地完成数据挖掘任务。此外,研究基于免疫原理的数据挖掘和推理策略无疑将会为决策支持系统提供新的技术和方法。

本书主要对人工免疫系统及其在数据挖掘领域中的应用问题进行阐述。讨论人工免疫算法和人工免疫网络模型的设计,探讨基于人工免疫系统的数据挖掘方法和基于免疫原理的推理方法,并将人工免疫系统理论应用于决策支持系统的工程实践中,以此丰富人工免疫系统及其应用的内涵。本书共分为8章,主要内容如下:第1章是绪论,阐明相关研究依据及其意义。第2章概括阐述生物免疫系统的组成、机理和特性。第3章分析现有人工免疫系统及其数据挖掘方法,根据人工免疫算法形式和网络结构的特

点,在免疫系统信息处理特性启示的基础上,概括提出人工免疫系统及其数据挖掘方法的基本框架。第4章通过对免疫算法设计方法的研究和对现有免疫算法不足的分析,提出基于距离浓度的免疫算法(Distance Concentration Artificial Immune Algorithm, DCAIA),给出该算法的收敛性证明,通过与遗传算法和传统免疫算法的实验对比,验证所提出算法的良好性能。第5章提出基于DCAIA的聚类、关联、分类的数据挖掘方法,通过实验给出这些数据挖掘方法的应用实例,验证所提出算法的可行性。第6章分析资源有限人工免疫系统存在的不足,提出一种改进的资源有限人工免疫系统,并设计基于改进模型的关联规则数据挖掘算法。第7章利用免疫应答机理和生物免疫网络理论,提出一种自适应免疫应答网络(Adaptive Immune Response Network, AIRN),给出具体网络的表示方式和构建过程,分析讨论AIRN的自适应性和收敛性,将AIRN应用于聚类分析,该模型在数据划分和模式识别上效果良好。第8章在基于数据仓库和数据挖掘支持的决策支持系统体系结构框架基础上,提出一种基于免疫原理的决策支持系统模型,阐述基于动态克隆策略的推理系统策略,并分析基于免疫原理推理系统的特点,讨论基于免疫原理的数据挖掘系统模型结构,通过对煤矿企业数据仓库中瓦斯数据的挖掘和通风机故障的问题求解,把基于免疫原理的数据挖掘和推理策略应用于煤矿生产的决策支持分析中。

本书是作者在江苏省现代企业信息化应用支撑软件工程技术研发中心研究工作的总结。在国家博士后科学基金(项目编号:20060390326)、国家自然科学基金(项目编号:60970058)、江苏省自然科学基金(项目编号:BK2009131)、江苏省现代企业信息化应用支撑软件工程技术研发中心开放基金(项目编号:eisecSX200806)和江苏省青蓝工程和福建省科技基金(项目编号:2008F5046)等项目的资助下,作者对人工免疫系统及其应用进行

了探讨,尤其对距离浓度的免疫算法及其数据挖掘应用进行了较深入的探索。

本书在撰写过程中参阅了大量的文献,对所引用的文献都尽量予以表明。限于篇幅和可能的疏漏,列出的参考文献尚不够全面,谨请有关作者谅解。如果有不妥之处,请通过电子邮件(banklt@163.com)告知本人改正,在此表示歉意和谢意。

本书力求比较全面地介绍作者在人工免疫系统及其数据挖掘应用研究的工作进展情况,但由于水平和精力所限,缺点和错误也在所难免,诚恳请求读者不吝指正。

刘 韬

江苏省现代企业信息化应用支撑软件

工程技术研发中心

2009 年 7 月

目 录

第 1 章 绪论	1
1.1 选题依据及意义	1
1.2 人工免疫系统及其应用研究综述	3
参考文献	10
第 2 章 生物免疫系统	17
2.1 免疫系统简介	17
2.2 免疫系统组成	18
2.3 免疫系统机制	38
2.4 免疫系统基本特性	57
2.5 小结	63
参考文献	64
第 3 章 人工免疫系统及其数据挖掘应用分析	66
3.1 人工免疫系统的仿生理理	66
3.2 人工免疫系统相关模型分析	70
3.3 人工免疫系统数据挖掘应用分析	82
3.4 小结	89
参考文献	90
第 4 章 距离浓度免疫算法(DCAIA)	94
4.1 免疫算法设计分析	94
4.2 基于信息熵免疫算法分析	97
4.3 基于距离浓度免疫算法设计	101

4.4	基于距离浓度免疫算法的收敛分析	110
4.5	实验仿真	115
4.6	小结	120
	参考文献	122
第5章	基于 DCAIA 的数据挖掘方法研究	125
5.1	基于 DCAIA 的分类挖掘方法	125
5.2	基于 DCAIA 的关联规则挖掘方法	136
5.3	基于 DCAIA 的 K-均值聚类挖掘方法	145
5.4	小结	151
	参考文献	152
第6章	改进的资源有限人工免疫系统及其关联规则 挖掘算法	155
6.1	RLAIS 人工免疫系统模型分析	155
6.2	改进的资源有限人工免疫系统(MRLAIS)	160
6.3	基于 MRLAIS 的关联规则挖掘算法	166
6.4	小结	172
	参考文献	173
第7章	自适应免疫应答网络(AIRN)及其聚类 挖掘方法	177
7.1	免疫应答和免疫网络	177
7.2	自适应免疫应答网络(AIRN)模型	182
7.3	自适应免疫应答网络(AIRN)模型分析	189
7.4	基于 AIRN 的聚类挖掘方法	193
7.5	小结	199
	参考文献	200

第 8 章 基于免疫原理的煤矿决策支持系统·····	203
8.1 决策支持系统(DSS)的发展 ·····	203
8.2 基于免疫原理的煤矿 DSS 系统框架 ·····	206
8.3 基于免疫原理的推理系统 ·····	207
8.4 基于免疫原理的数据挖掘系统 ·····	214
8.5 煤矿企业决策支持实例 ·····	216
8.6 小结 ·····	226
参考文献·····	227

第1章 绪 论

1.1 选题依据及意义

很多生物机体在动态竞争的环境中展现出了强大而复杂的学习和问题处理能力,生物机体的这些特性,激发了人的思维,促进了人类科学技术的发展。在过去的几十年中,受生物机体启发而产生的方法和理论不断涌现和发展。目前在信息领域有明确研究内容的四个子系统分别是:脑神经系统、遗传系统、免疫系统与内分泌系统。其中免疫系统和内分泌系统已被证明具有高度的信息处理能力,已经开始受到研究者的重视。最近几年,人工免疫系统(Artificial Immune System, AIS)的研究迅速发展,已经成为智能系统中继模糊逻辑、神经网络、遗传算法之后的又一研究热点^[1~4]。人们基于生物免疫原理提出了多种人工免疫模型和算法,并应用于自动控制^[5]、故障诊断^[6~8]、优化计算^[9~11]、模式识别^[12]、机器学习^[13,14]、数据分析^[15~18]等领域。

人工免疫系统是研究、借鉴和利用生物免疫系统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的应用而产生的各种智能系统的统称^[19]。生物免疫系统结构复杂、功能独特。免疫系统内没有中心控制器官,它是由体内不断循环的分散免疫细胞群构成的,能区别自体与非自体,并具有排除与记忆非自体的功能。从信息处理的观点看,生物免疫系统是一个并行分布式的自适应系统,它具有认知、特征抽取、多样性、学习和记忆等特点^[2]。目前虽然在医学免疫学上取得了有关实验的成果

和基于非线性模型的理论研究成果,但人工免疫系统理论及其在工程中的应用研究还处在起步阶段。研究免疫系统的良好特性和机理,可以获得新的理论、新的思想和新的方法,用于解决日益复杂的工程问题。尤其是加强免疫系统在信息处理系统中的理论和应用研究,可以有效地提高对信息的智能处理能力。例如,免疫系统可以有效地用于构建模式识别分类器或探测器,而且这种模式识别分类器具有高效的特征抽取、快速进化、学习和记忆性质,这恰是工程实践中所需要的。因此,加强人工免疫系统的理论和应用研究不仅意义重大,而且是非常必要的,借鉴生物免疫系统的特性,在人工智能计算方面有着巨大的潜能。

影响我们生活方方面面的信息技术正在飞速发展,它积累了大量的信息数据,随着数据库技术的迅速发展,数据存量急剧增加。但是,挖掘海量数据中隐藏知识的手段远远不足,从而导致了“数据爆炸但知识贫乏”的现象。计算机技术的另一领域——人工智能(Artificial Intelligence, AI)自20世纪60年代诞生之后取得了重大进展。该领域目前的研究热点是机器学习,即用计算机模拟人类学习的过程。用机器学习的方法来分析数据,挖掘大量数据背后的知识,这两者的结合促成了数据挖掘的产生。处理这些信息,把它变成有用知识的数据挖掘(Data Mining)技术则是我们提取知识的有效手段。数据挖掘也称为数据库中的知识发现(Knowledge Discovery in Database, KDD),是指发现隐藏在数据库大量信息中有用的知识。其目的是帮助有关人员寻找信息数据之间的内在关系,发现数据项之间新的要素^[20]。实际上,数据挖掘是一门交叉学科,涉及机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、高性能计算、专家系统等多个领域。数据挖掘的成果可以用在信息管理、过程控制、科学研究、决策支持等许多方面。

数据挖掘的目的是帮助决策者寻找数据间潜在的相互关系和

模式特征,而从生物免疫系统中接受启发,利用其特征抽取、快速进化、学习和记忆等特性,正是人们在数据挖掘时所孜孜以求的。因此,开展人工免疫系统在数据挖掘领域的应用研究,可以丰富数据挖掘方法和理论,获得更有效的数据挖掘技术,更好地完成数据挖掘任务。

此外,自 20 世纪 90 年代以来,在信息技术的推动下,新的决策理论和方法不断涌出,特别是数据挖掘技术和智能推理方法使决策支持系统(Decision Support System, DSS)得到了长足发展,因此,研究基于免疫原理的数据挖掘和推理策略无疑将会为决策支持系统提供新的技术和方法。

1.2 人工免疫系统及其应用研究综述

1.2.1 人工免疫算法和人工免疫网络模型

随着对生物免疫系统特性和机理的认识深入,人们构造了一系列人工免疫系统,用来解决一些复杂的工程实际问题。在这些人工免疫系统中可以分为人工免疫算法和人工免疫网络以及基于网络结构的框架模型。

1.2.1.1 免疫算法

人工免疫算法(Artificial Immune Algorithm, AIA),简称免疫算法,是由免疫学理论和观察到的免疫功能、原理和模型启发而设计出的计算方法。这方面的研究最初从 20 世纪 80 年代中期的免疫学研究发展而来。1990 年, H. Bersini 等^[21]首次使用免疫算法来解决问题。20 世纪末, D. Dasgupta 和 S. Forrest^[22]开始将免疫算法应用于计算机安全领域。同期, D. E. Cooke 等^[23]开始将免疫算法应用于机器学习领域。自然免疫系统所具有的显著信息处理能力对计算技术有很多重要的启发。近年来,特别是 1997 年人

工免疫学研究在国际上兴起之后,国内外研究人员提出很多免疫算法^[24~26],主要有以下几种算法类型:

(1) 阴性选择算法: S. Forrest 等^[27] 基于阴性选择原理提出阴性选择算法来检测异常,其算法主要包括以下两步:首先产生监测器集合,其中每个监测器与被保护的数据都不匹配;然后不断地将集合中每个监测器与被保护的数据相比较,如果监测器与被保护的数据相匹配,则判定该数据发生了变化。P. D'haeseleer 等^[28] 基于阴性选择原理给出了一种阴性选择算法,用于监测数据改变。其中抗体(问题解答)与抗原(问题)的匹配采用 S. Forrest 提出的部分匹配规则。

(2) 免疫遗传算法: J. S. Chun 等^[9] 提出一种免疫算法,实质上是改进的遗传算法。它采用类似遗传算法的搜索策略,借用遗传算法的选择、交叉和变异算子(有的只用变异算子),但在群体搜索策略、解的表示和记忆单元设置等方面与遗传算法有所不同。

(3) 免疫学习算法: J. D. Farmer 等^[29] 首先对免疫系统的学习特性进行了探索性研究,指出免疫系统具有分类器系统的性质。随后 J. E. Hunt 等^[13] 基于免疫网络理论给出一种有监督的机器学习算法。Y. Ishida^[31] 也提出一种基于 PDP 模型的人工免疫系统学习算法,并将其应用到故障诊断中。

(4) 克隆选择算法: D. E. Castro 等^[32] 基于免疫系统的克隆选择理论提出克隆选择算法,该算法是模拟免疫系统学习过程的进化算法。

(5) 基于疫苗的免疫算法: 焦李成等^[24] 利用免疫系统的理论提出基于疫苗的免疫算法。

(6) 基于免疫网络的免疫算法: N. Toma 等^[33] 基于主要组织相容性复合体(MHC)和免疫网络理论提出一种自适应优化的免疫算法,用于解决多智能体中每个智能体的工作域分配问题。

由于基于免疫系统特性的应用探索还处于起步阶段,上述已

有的免疫算法的问题针对性比较强,应用范围有局限性,还有许多方面考虑不到。因此,算法的设计也有许多有待改进之处。

1.2.1.2 人工免疫网络模型

N. K. Jerne 在免疫学上提出的独特型网络模型^[34]以及其他研究人员提出的多种免疫网络学说^[35],这些都可借鉴用于建立人工免疫网络模型。总的来看,人工免疫网络模型将系统视为一个由节点(淋巴细胞)组成的网络结构,通过节点之间的信息传递和相互作用达到识别、效应、记忆等免疫系统功能。

在人工免疫网络研究中,A. Ishiguro 等^[36]提出一种互联耦合人工免疫网络模型,即免疫系统是通过多个完成某一特定任务的局部免疫网络之间的相互通信来形成大规模免疫网络,该模型已用于机器人的行走控制和路径规划之中。Z. Yamaguchi 等^[37]提出了一种多值免疫网络模型,通过字符识别显示出多值免疫网络在记忆模式、记忆容量方面,特别是在噪声免疫能力方面要优于二进制网络模型。D. E. Castro 等^[38]提出一个抗体网络模型 AB-NET,该模型以克隆选择原理为基础,其核心是网络的生成算法,生成的抗体网络用于识别给定数量的抗原(问题)。F. T. Vertosick 和 R. H. Kelly^[39]将免疫系统表示为一种搜索神经网络结构的替代方法,基于并行分布处理理论,提出 PDP 免疫网络。

这些网络模型从不同的侧面利用生物免疫系统的机理来解决工程问题,推动了人工免疫系统研究的发展,但这些模型大都针对某一特定的工程问题而设计,其应用的领域和范围受到限制。随着人工免疫系统探讨的深入,研究人员致力于开发新的网络模型和以网络结构为主的人工免疫系统框架模型。在随后的一节里,将结合它们在数据挖掘领域的应用,对这些人工免疫网络和系统框架模型给予阐述。

1.2.2 基于人工免疫原理数据挖掘综述

近年来,基于免疫系统原理开发的各种模型和算法广泛地应用在科学研究和工程实践中。鉴于免疫系统具有保护机体的强大能力,相应的 AIS 模型在信息安全方面得到了广泛应用^[40,41];利用其识别能力在模式识别方面的研究也取得了进展^[14,42,43];作为一种智能优化搜索策略,在函数优化、组合优化、调度问题等方面也得到应用并取得了很好的效果;在机器人学、故障诊断等方面的应用,很好地体现了 AIS 的智能突现与群体智能特性^[31,36,44,45]。人工免疫系统作为一种新兴智能系统,也已经被引入数据挖掘领域,用于完成数据挖掘的分类、聚类和数据浓缩等任务,其挖掘的对象多种多样,目前涉及的主要是数据库、Web 页等。

1.2.2.1 基于人工免疫原理的分类

在基于免疫系统开发的分类技术方面,J. D. Farmer 等^[29]最早证明了分类器和独特型免疫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^[9]的分类器系统的类似性,奠定了人工免疫系统用于数据挖掘技术的理论基础。S. Forrest 等在文献[47]中描述了用于研究人类免疫系统中发生的学习和模式识别过程的免疫系统模型。模型对检测器应用示意性表示法,在检测噪声输入数据中一般模式和保持群体内多样性的能力上进行了评估。J. E. Hunt 和 D. E. Cooke 在文献[13]中描述了一种使用突现适应度共享的遗传算法来发现数据模型,证明在人工免疫系统意义上,从正例样本学习是确实可能的。

J. H. Carter 提出的 Immuno-81^[12]是基于免疫原理的监督学习算法,该系统模拟 B 细胞和 T 细胞性质,以及它们产生初次和二次免疫应答的相互作用机制等几个方面。J. H. Carter 把 Immuno-81 和几个广泛使用的机器学习算法用于冠心病人数分

类,互相比较来评估系统的性能,算法表现出良好的性能。

M. A. Potter 等^[41]描述了使用协同进化遗传算法 AIS 模型的进化抗体,用于概念学习,并在人工数据集合和议会投票记录数据进行分类能力测试,取得较好效果。

Andrew Watkins 等^[48]受资源有限人工免疫系统 (RLAIS)^[17]启发,提出的资源有限人工免疫分类器 AIRS (A Resource Limited Artificial Immune Classifier) 是一种主要的人工免疫监督学习方法。AIRS 模拟免疫系统 B 细胞相互作用机制。AIRS 中的抗原初始化为特征向量,在训练和学习期间提呈给系统。从 AIRS 中的 B 细胞与抗原都表示为向量 $V_i = (v_{i1}, v_{i2}, \dots, v_{ir}; c)$, v_{ij} ($j=1, 2, \dots, r$) 是 B 细胞或抗原的属性特征, c 为类别。所有同类 B 细胞在一起表示为 ARBs (ARB—Artificial Recognition Ball, 人工识别球^[17])。

1.2.2.2 基于人工免疫原理的聚类

在属于无监督学习任务的聚类技术研究方面,一些研究人员提出基于免疫原理的数据聚类方法。J. E. Hunt 和 J. Timmis 等在文献[14]中描述了一个称为 Jisys 的系统,用于在抵押贷款欺骗诊断系统中聚类数据,D. E. Castro 和 J. Timmis^[49]使用该系统成功地在标准测试数据 Iris 数据进行聚类分析。C. Hung 等^[50]实现结合记忆细胞的人工免疫系统,该模型在大规模和动态系统实现聚类。在聚类关于土壤湿度水平的远程感知数据(一个海量数据样本)的应用中,研究者发现其性能与其他分类器相比要差,并把性能较差的原因归咎于所使用的数据表示机制和匹配算法的选择。

J. Timmis 和 M. Neal 等^[51]给出三种不同的数据分析技术:单连接聚类、Kohonen 网络和人工免疫系统。研究者使用每一个聚类技术识别一些标准测试数据集合中的信息聚类。这些技术的前两个可看做单独聚类分析技术。他们认为 AIS 不仅是简单聚

类分析技术,也能提供更多样化的数据表示,另外认为 AIS 在允许用户全面理解数据集合的组成方面是有效的,这样有助于数据挖掘和分析过程。

J. Timmis 等利用免疫网络假说开发了资源有限人工免疫系统 (Resource Limited Artificial Immune System, RLAIS) 模型^[17,49],该模型用于数据聚类并在标准测试数据集 Iris 上实现聚类网络。系统发展的一个最有趣的概念是“资源有限”。该概念主要是指用 ARB 表示具有同样特征的 B 细胞的集合,每次免疫细胞暴露给一个给定抗原(或者抗原群体,即提呈给系统的数据),每一个细胞都试图基于刺激水平而获得更多的资源(指单个 B 细胞)。但是,系统范围内 B 细胞资源是有限的,如果某细胞消耗的资源比系统中规定的资源还多,则从该 ARB 中移除资源。首先从最少受刺激的细胞开始,直到细胞持有的总资源量等于系统中规定的资源量。资源的竞争给系统加入适当的进化压力,确保只有最适应的细胞生存,最终对这些生存下来的细胞进行克隆变异,并通过一定可视化方法反映和表示提呈的抗原数据聚类。此后,在 J. Timmis 本人对 RLAIS 系统的改进中^[52],提出一个无监督学习算法,用于进行数据分析和数据可视化,其中用实数向量表示抗体和抗原数据。

D. E. Castro 和 V. Zuben^[43]检验了免疫系统内克隆选择机制的作用,提出一个称为克隆选择算法(clonal selection algorithm)的无监督学习算法,此后基于免疫网络亚动力学的数据分析和数据简化算法^[1,49]开发了 aiNet 免疫网络模型^[53],并使用了与标准聚类技术混合的方法,用最小生成树(MST),借用图论的思想,自动地检测和分离网络聚类,利用双螺旋问题和链式连接数据测试 aiNet 的性能,得到较好的结果。

E. Hart 和 P. Ross^[54]结合免疫系统和松散分布记忆模型,提出用于聚类移动数据集合的免疫系统模型 SOSDM(Self-Organis-