

用免疫神经网络做数据挖掘^{*}

Using Immune Neural Network for Data Mining

郑建国^{1,2} 刘芳¹ 焦李成¹

(西安电子科技大学雷达信号处理国家重点实验室 西安 710071)¹

(湖北汽车工业学院 湖北十堰 442002)²

Abstract Based on analyzing the immune phenomena in nature and utilizing performances of the existent artificial neural network, a novel network structure, i.e., an immune neural network (INN), is proposed which integrates the immune mechanism and the structure of neural information processing. The learning algorithm of INN is also given which contains the method of selecting an activation function and the adaptive algorithm of the network. This model makes it easy for an user to directly utilize the characteristic information of a pending problem and to simplify the original structure through adjusting the activation function with the prior knowledge, and then the working efficiency and the searching accuracy are both improved. The analysis in theory and the simulating test for data mining problem show that, comparing with the artificial neural network, INN is not only effective but also feasible. INN can conclusively simplify the structure of the existent model and express good working performances when dealing with a pending problem.

Keywords Immune neural network, Artificial neural network, Data mining

1 引言

随着社会信息量的增大,在各种应用领域里的数据库中存储了大量的数据,这使得人们对这些数据进行分析并转化为有用知识的需求变得越来越迫切^[1]。于是知识发现与数据挖掘(Knowledge Discovery and Data Mining, KDD)自然成为近年来人们从大型数据库中获取信息的一个重要的研究领域。一般地,数据挖掘就是指从数据库或数据仓库中发现隐藏的、预先未知的、有趣的信息的过程,该过程可以看作是知识发现过程中的一个核心的步骤。目前,能够用于解决机器学习问题的方法主要有三种类型,即:模糊规则的学习方法、神经网络学习方法^[2~4]和遗传进化的学习方法。纵观数据挖掘中的规则提取方法,决策树规则提取方法不能实现多变量搜索^[5],因为它在建树时每一个节点只含有一个特征,故属于一种单变元算法,特征间的相关性强调不够,虽然它将多个特征用一棵树连接在一起,但这种联系是一种松散形式。另外,决策树推理方法对数据中的噪声较为敏感,尽管神经网络从数据中提取规则的最大困难在于其所表示的知识都隐含在网络连接中,但是神经网络具有一些独特的优点,如非线性映射、容错能力等,因而相对于其它方法,神经网络在数据挖掘中具有更大的前景。

在上述三种类型的方法中,人工神经网络是人们探索并模仿脑神经系统的智能信息处理机制的一个重要研究领域。特别是 Rosenblatt 和 Holland 等人^[6~8],先后提出了各种较为实用的网络模型,从而把这一领域的研究从理论分析推广到工程实践,并取得了许多有意义的成果和进展。然而,目前已有的这些神经网络模型都是在对生物神经系统认识的基础上,进行了高度简化与抽象的结果,这样做固然有利于其在工

程实践领域中的开发与应用,但也不可避免地损失了一些原有的功能。随着神经网络应用的不断推广与深化,其中的问题也不断暴露出来,比较典型的包括:在学习算法选择不当时系统易于陷入局部极优状态,网络的复杂度与泛化能力之间存在矛盾等等。

另一方面,在人工智能不断向生物智能学习的过程中,人们逐渐意识到生物免疫能力的重要性,并对其进行了一定的研究。其中,Farmer 等人首先在工程领域提出了“免疫”(Immunity)概念^[9]。之后,Varela、Stewart 和 Perelson 等人受生物学家 Jerne 的“免疫网络学说”(Immune network theory)^[10]的启发,进一步完善了生物免疫系统的结构(如提出了一种新的网络模型——免疫网络模型等)和相关机理^[11~13]。在他们看来,生物免疫的功能主要包括三个方面,即:免疫防御、免疫稳定和免疫监视。在生物免疫系统中主要有 B 和 T 两类淋巴细胞,其中 B 细胞由骨髓产生并执行体液免疫功能,而 T 细胞由胸腺产生并执行细胞免疫和免疫调节功能。后来,Ishiguro 等人提出了一种互联耦合免疫网络模型,即免疫系统是通过多个用于完成某一特定任务的局部免疫网络之间的相互通讯来构成一个大规模的免疫网络,该模型已被用于六足行走机器人的速度控制^[14];Kumak 等人将神经网络和免疫学理论相结合并用于复杂动态行为的建模和自适应控制^[15];Lee 等人基于细胞选择学说和独特型网络学说的控制机理,将每个机器人看成 B 细胞,将每个环境条件作为抗原以及将行为策略作为抗体,设计了一个分布式机器人的控制系统^[16];Mitsumoto 等人基于免疫系统中 B 细胞之间相互反应的机理设计一个多机器人系统的控制结构^[17];Takahashi 等人设计了一个 PID 型免疫反馈控制器,并用于控制带有非线性干扰的 DC 伺服马达系统,而且验证了其有

^{*} 本文得到国家自然科学基金(No. 60073053)和国家“863”计划(No. 863-316-05-03-01)资助。郑建国 副教授,博士生,主要研究方向为数据挖掘与知识发现、智能决策支持系统、神经网络。焦李成 教授,博士生导师,IEEE 高级会员,主要研究方向为神经网络、智能信息处理、非线性科学等。

效性^[18];Ishida 等人研究了一种基于相关识别特性的免疫网络模型用于故障诊断的方法^[19];Chun 等人基于体细胞理论和免疫网络理论提出了一种免疫算法^[20],该算法将抗原作为目标函数,抗体作为求解结果,抗原和抗体之间的亲和力作为解答的联合强度;Forrest 等人使用了反面选择算法来检测受保护的数据和程序文件的变化,在 DOS 环境下进行了包括文件感染器和根扇区病毒等许多试验^[21,22],结果表明这种方法很容易检测到由于病毒感染而发生在数据文件中的变化。另外,文[23,24]将“免疫”思想与已有的进化算法相结合,从而形成一种基于免疫策略的进化算法并成功地用于一些疑难问题的求解。这种算法的核心思想是认为每一个待求的实际问题都会有自身一些基本的、显而易见的特征信息或知识,如果能有效利用这些信息与知识,则会有利于待解问题的求解。

我们在对原有一些网络模型或算法的深入分析过程中发现它们在结构与参数设置上相对固定,在求解问题时,可变的灵活程度较小,这无疑对网络和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时的帮助作用,特别是在求解一些复杂问题时,这种“忽视”所带来的损失往往就比较明显了。有鉴于此,本文旨在进一步分析人工神经网络所依据的生物学基础,探索改进与完善网络模型的途径。我们将“免疫”的思想及理论与已有的一些神经网络模型相结合,研究并设计了一种能够利用背景理论和先验知识的新型网络,即“免疫神经网络”(Immune Neural Network—INN),并依此来构造进行数据挖掘过程的机器学习机制,以求提高对问题的处理能力。

2 免疫神经网络模型

Varela 等人提出的免疫网络模型主要包括两部分:第一部分为系统的动力学部分,即由微分方程来控制用以模拟克隆淋巴细胞及其相应的免疫球蛋白(Immunoglobins—Ig)集合中元素的增减;第二部分为系统的元动力学部分(Meta-dynamics),即通过一定的算法来模拟群体中某些淋巴细胞的清除和由骨髓产生大量的、新的淋巴细胞,从而使系统得到恢复。不过到目前为止,有关免疫网络所具有的一般形式尚未形成一种普遍接受的认识。

2.1 免疫神经网络的生物学基础

人工神经网络的物质基础来自于生物系统中的神经元模型。一般认为,一个完整的神经元由细胞体、树突、轴突、突触和神经末梢等几大部分构成,其中细胞体是神经元的主体。另一方面,人的大脑系统大约由 10^{11} 个神经元组成。这些神经元虽然在物理结构上是基本一致的,但其功能和在系统中所发挥的作用是有明显差别的,究其原因有多方面的,其中之一在于不同神经元的细胞体,对来自外界或其它神经元的神经信号的敏感程度和处理能力不同。就以视觉神经细胞而言,可能会存在这样一种现象:某些神经元细胞对红色信号的反应较为敏感而对蓝色信号的反应很迟钝。而且,这种反应程度上的差别会因为大脑接受外界刺激的先后与记录而发生变化。换言之,上述现象很可能是因为对红色信号反应强烈的神经元自身存在或外界强加了一种对蓝色信号的免疫力所造成的。其实,生物自然界里的这种细胞免疫力影响神经信号处理的现象是普遍存在的。

人们一般认为:免疫系统是生物,特别是脊椎动物和人类所必备的一种防御系统。它由具有免疫功能的器官、组织、细胞和免疫效应分子等组成,可以保护机体抗御病原体、有害异

物和癌变细胞等致病因子的侵害^[23]。生物体的免疫功能一般包括三个方面,即免疫防御、免疫稳定和免疫监视。在生物免疫系统中,免疫反应主要是通过抗体与抗原之间的相互作用而进行的。其中,抗原(Antigen)是各种非己器官或组织的统称,它由载体和表位组成;抗体(Antibody)是一种由免疫细胞所分泌出的免疫分子,它可以识别和结合抗原,在有其它免疫细胞和分子的合作情况下可以进一步杀死抗原。另外,我们将具有外来抗原的主要特征但并不对生物机体构成侵害的物质称为疫苗(Vaccine)。抗原与抗体在机体内的相互存在是有一定比例的(即浓度);如果生物体内针对某一抗原的抗体数量较少,则在这种抗原侵入生物体时,抗体这时没有足够的能力来结合并清除抗原,但可以记住抗原的特性;如果这种抗原再次侵入该生物体,则生物体的性质会发生一定的变化,使抗体的数量迅速增加,从而杀死抗原。所以,人们可以通过向生物机体注射疫苗来提高其对某些外来抗原入侵的免疫力。由上述内容可以看出,免疫系统具有记忆功能以及自学习、自组织和自适应的能力。

2.2 人工免疫神经网络的设计及其模型

在已有的人工神经网络模型中,人们只是将细胞体简单地等价为一个对输入神经信号进行总体求和,在与阈值对比后产生一种神经输出信号的单元。这种模型的突出特点是结构简单,通用性强,但随之带来的不足是没有考虑到具体问题的特征信息在求解问题时可能产生的积极的、辅助的作用,其具体表现在于模型中没有给特征信息留有接口。有鉴于此,我们在原有神经网络模型的基础上,添加了一个“疫苗接种单元”以利用待解问题的特征信息和人们的先验知识,旨在提高网络针对问题的处理能力,其具体结构如图1所示。

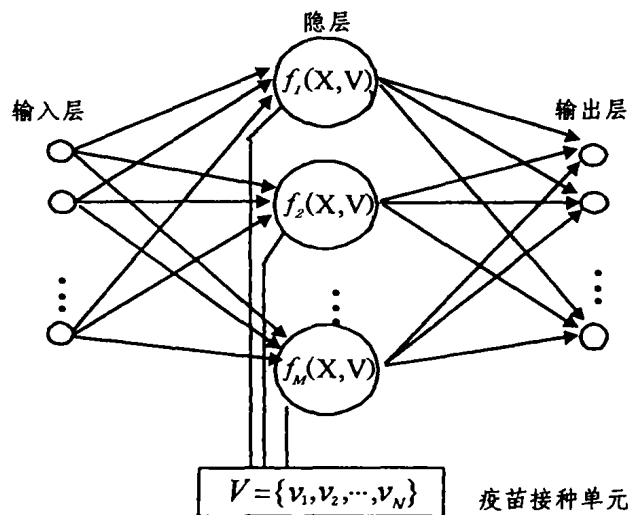


图1 免疫神经网络模型

在此模型的设计方面,我们首先认为细胞体在信息的加工与处理过程中起到了非常重要的作用;其次,各个细胞体之间虽然在基本性质上相似,但在具体的表现形式上存在差异。所以,我们将细胞体的激励函数设计为一种可变的形式,即函数的基本特性保持不变,但可以通过某些参数的调整而达到改变其具体形式的目的。具体而言,我们可以将任一神经元 i 的激励函数设计为如下的形式:

$$u_i = f_i(X, V) \quad (2.1)$$

其中, $f_i(\cdot)$ 是一个带有一系列参数 V 的函数族,而该函数形式的选取和其中参数的确定均与待解问题有关。这与原

有模型中激励函数的选取基本上为通用的阈值型或S型等函数的情况是不同的。另一方面,由于网络的中间处理层含有实际问题的某些特征,因此在结构上往往比较简单,从而使网络连接权值在训练时的工作量减少,进而提高学习的速度。

3 免疫神经网络的学习算法

免疫神经网络的学习训练包括两个方面,即网络权值的学习和网络激励函数中相关参数的学习。这里所要考虑的重点是激励函数的选取和函数中待定参数的学习方法。

3.1 网络激励函数的选取方法

免疫神经网络模型中的激励函数在一般情况下是不能通用的,它是根据待解问题的特征信息和人们的先验知识而专门设计的。同一网络的不同激励函数之间虽然性质相同,但在具体表现形式方面存在差异。一般而言,在网络模型的设计过程中应遵循以下步骤:

1. 提取疫苗。此处所谓的疫苗是引用生物学的概念,在这里指待解问题的特征信息和人们对其所具备的先验知识。而提取疫苗是指分析待解问题的过程,从中搜集特征信息,然后再根据先验知识找出输入变量之间的相互约束关系。而且,这种约束关系应简单易求,而不要求全面而且精确。

2. 设计激励单元的基本类型。即根据上述约束关系,选取一种适当的含有待定参数的函数族。这些函数的基本类型和特性是一致的,但随其中参数的不同而在具体的表现形式上存在差异。

3. 接种疫苗。根据第1步所提取的疫苗填充图1中所示的“疫苗接种单元”,即确定疫苗向量 $V = \{v_1, v_2, \dots, v_N\}$,其用以确定激励单元中相关函数的参数。这里需要说明的是:由于待解问题的特征信息往往不止一个,也就是说针对不同特征信息所能提取的疫苗也可能不止一种,因此“疫苗接种单元”中的疫苗向量可以是简单的一个,也可以是多个或所有这种向量按照一定逻辑关系的组合。

4. 网络训练。选取一种网络学习算法,如BP、L-M、径向基函数算法等,利用训练样本来修正网络中的权值矩阵和阈值等相关参数。

3.2 网络的自学习算法

在某些情况下,我们一方面会因为对待解问题一时难以形成较为成熟的先验知识,而无法从分析问题的过程中提取出合适的特征信息,故得不到有效的免疫疫苗;另一方面,为寻求用于全局求解的局部方案需要付出的代价超出其应占的比例,使计算成本增加、效率降低,从而使提取疫苗的工作失去了意义。所以,当我们无法设计网络中的“疫苗接种单元”时,图1所示的网络就恢复成一般的人工神经网络。不过,这时的网络激励单元仍然可以采用激励函数可调的形式,只是需要将其中的参数当作网络训练的目标之一。不妨以单隐层结构的免疫神经网络为例,设输入层—隐层、隐层—输出层间的权值矩阵分别为 $W^{(1)}$ 和 $W^{(2)}$,则输出向量 Y 为:

$$Y = W^{(2)} f(W^{(1)} X, V) \quad (3.1)$$

设训练样本中的实际输出为 Z ,则定义误差评价函数为:

$$E = \frac{1}{2} \sum_{i=1}^P (z_i - y_i)^2 \quad (3.2)$$

为方便起见,我们可以采用梯度下降法来对网络进行训练。但在具体的训练过程中,为平滑学习路径和提高学习速度,我们最好采用成批训练和添加动量项的综合方法。其中,成批训练方法主要是针对逐一学习方法导致学习速度缓慢而

提出的,它是将一批学习样本所产生的修正值累加后统一进行一次修正;添加动量项法的思想是通过引入前一步的修正值来平滑学习路径、避免陷入局部极小。具体而言,权值 W 和参数 V 的修正公式表达如下:

$$\begin{cases} \Delta W^{(i)}(t+1) = \eta^{(i)} \sum_{j=1}^M \frac{\partial E}{\partial W^{(i)}}(t) + \alpha [W^{(i)}(t) - W^{(i)}(t-1)] \\ \Delta V^{(i)}(t+1) = \mu^{(i)} \sum_{j=1}^M \frac{\partial E}{\partial V^{(i)}}(t) + \beta [V^{(i)}(t) - V^{(i)}(t-1)] \end{cases} \quad (3.3)$$

$i = (1, 2)$

式中 α 和 β 为动量因子, $\eta^{(i)}$ 和 $\mu^{(i)}$ 为学习速率。其中,学习速率 $\eta^{(i)}$ 和 $\mu^{(i)}$ 在训练过程中影响较为明显, $\eta^{(i)}$ 、 $\mu^{(i)}$ 大则有利于收敛,但到后期会引起算法振荡。所以一般是在学习初期将采用较大的值,到后期再令其随学习次数递减。

4 针对数据挖掘的仿真实例

我们用中国大陆各省区农业统计数据的挖掘和知识发现的实例来说明免疫神经网络方法的可行性和有效性。从多年的中国统计年鉴中收集中国大陆30个省区1982年至1999年间农业人口、耕地面积、农业投资、农业总产出等有关数据,目的是分析和发现其中的普遍特征以及多种因素与农业总产出的关系。

用各省区农业统计数据的挖掘和知识发现过程来考查免疫神经网络的工作能力。具体的方法是:首先,对表1进行属性泛化,将省份和城市名属性泛化为所在区域位置,分为东北部、北部、西北部、东部等;将其它属性分别泛化为“小”、“中”、“大”或“低”、“中”、“高”等;用最大方差法分别将这些数值属性离散化,如表2所示。选取农业数据样本点,共1800个(表1为其中的一部分),以区域、耕地面积、农业人口、农业投资作为输入模式,以农业产出为输出目标值,保持一些基本的参数条件不变,分别采用BP网络、L-M前向网络、径向基函数网络和本文的INN对上述数据库进行了仿真研究(实际数据库的记录数共1800条,以其中的1000条为实验数据,其余的800条为测试数据)。

表1 农业统计信息属性表

编号	区域	耕地面积	农业人口	农业投资	农业产出
1	Heilongjiang	8826.53	20084.6	343	24540
2	Liaoning	3470.4	22724.4	327	27380
3	Jiling	3935.53	14883.2	195	18910
4	Shandong	6867.87	68459.6	216	64750
5	Hebei	6560.47	52315.5	279	35760
6	beijing	414.47	3952.96	132	7020
7	tianjing	432.27	3830.47	120	5490
8	henan	6944.40	74522.5	479	50200
9	shanxi	3701.80	22177.7	167	12480
10	shanxi	3541.07	26659.6	123	17000
11	ganshu	3477.13	18682.5	286	10310
12	neimonggu	4911.53	14877.7	156	15690
13	ningxia	795.00	3542.94	94	2470
14	xinjiang	3072.93	10005.9	325	14470
15	qinghai	572.00	3101.09	72	2450
16	zejiang	1731.13	35426.6	173	33680
17	jiangshu	4562.33	52707.6	192	58050
18	shanghai	324.00	4188.87	176	6820
19	anhui	4373.00	48175.6	333	37090
20	hubei	3486.60	41728.7	395	40220
...	...	...	...	...	...

在具体的实现方面,对于本文的INN方法,采用前面第

3.2节介绍的方法提取特征信息,确定疫苗向量 $V=\{v_1, v_2\}$ 和训练网络。采用MatLab5.3来编写实验程序,其中,在对网络中的可调参数 v_1, v_2 进行优化时,迭代步长为0.1,将训练好的网络用于对测试样本进行推理;以5次计算结果的平均值为最终统计结果,四种算法的对比情况如表3所示。

从表3可以看出,本文所提出的INN方法,进行推理的准确率优于现有的BP网络、L-M前向网络、径向基函数网络等算法;这是因为免疫网络可以直接利用待解问题的特征信息,并通过注入先验知识来调节隐层单元的激励函数,以达到简化网络的结构,提高其工作的效率和准确性。

表2 泛化的农业信息决策表

编号	区域	耕地面积	农业人口	农业投资	农业产出
1	东北	大	中	大	中
2	东北	中	中	大	中
3	东北	中	少	中	低
4	北	大	多	中	高
5	北	大	多	中	中
6	北	小	少	小	低
7	北	小	少	小	低
8	北	大	多	大	高
9	北	中	中	中	低
10	西北	中	中	小	低
11	西北	中	中	中	低
12	西北	中	少	中	低
13	西北	小	少	小	低
14	西北	中	少	大	低
15	西北	小	少	小	低
16	东	小	中	中	中
17	东	中	多	中	高
18	东	小	少	中	低
19	东	中	多	大	中
20	中	中	中	大	中
...	...	...	...	...	...

表3 四种算法计算结果对照表

算法种类	准确率
BP网络	68.2
L-M前向网络	73.4
径向基函数网络	84.7
INN	93.1

结论与讨论 本文在探讨与分析生物自然界中免疫现象的基础上,结合当前人工神经网络的特点,提出了一种集免疫机制和神经信息处理机制于一体的新型网络结构——免疫神经网络,及相应的学习算法。该网络可以使人们直接利用待解问题的特征信息,并通过注入先验知识来调节隐层单元的激励函数,以达到简化网络的结构,提高其工作的效率和准确性。理论分析和针对数据挖掘的仿真实例的实验结果表明,与一般的人工神经网络相比,免疫神经网络不仅是有效的,而且也是可行的。它可以明显简化原模型在应用于具体问题时的结构,并同时具有良好的工作性能。当然,对该模型的设计与优化在理论上还留有广阔的空间,比如隐层单元激励函数的优化算法问题、激励函数的参数调节与网络权值调节的协调问题等都需要进一步的研究与完善。

参考文献

- Fayyad U M, et al. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. 1~34
- Omlin C W, Giles, C L. Rule revision with recurrent neural net-

- works. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(1): 183~188
- Lu H J, Setiono R, Liu H. Effective data mining using neural networks. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 957~961
- Fu L M. A neural-network model for learning domain rules based on its activation function characteristics. IEEE Transactions on Neural Networks, 1998, 9(5): 787~795
- Quinlan J R. Comparing connectionist and symbolic learning methods. In: Hanson, S. J., Drastall, G. A., Rivest, R. L. eds. Computational Learning Theory and Natural Learning Systems. MIT Press, Vol 1. 1994. 445~456
- Rosenblatt F. Principles of Neurodynamics. Spartan, New York, 1962
- Hopfield J J. Learning algorithms and probability distributions in feed forward and feedback networks. Proc. Natl. Acad. Sci. USA, 1987, 84: 8429~8433
- Grossberg S. Nonlinear neural networks: principles, machines and architectures. Neural Networks, 1988, 1: 15~57
- Farmer J D, Packard N H, Perelson A S. The immune system, adaptation and machine learning. Physica D, 1986, 22: 187~204
- Jerne N K. Towards a network theory of the immune system. Annual Immunology, 1974, 125C: 373~389
- Varela F J, Stewart J. Dynamics of a class of immune networks. I) Global behavior. J. Theo. Biol., 1990, 144: 93~101
- Stewart J, Varela F J. Dynamics of a class of immune networks. II) Oscillatory activity of cellular and humoral component. J. Theo. Biol., 1990, 144: 103~115
- Perelson A S. Immune Network Theory. Immunological Review, 1989, 10: 5~36
- Ishiguro A, Kondo T, Watanabe Y. Emergent Construction of Artificial Immune Networks for Autonomous Mobile Robots. In: Proc. 1997 IEEE Intl. Conf. on System Man and Cybernetics, Orlando, FL, USA, 1997. 1222~1228
- Kumak K K, Neidhoefer J. Immunized Neurocontrol. Expert Systems with Applications, 1997, 13(3): 201~214
- Lee D W, Sim K B. Artificial Immune Network-based Cooperative Control in Collective Autonomous Mobile Robots. In: Proc. 6th IEEE Intl. Workshop on Robot and Human Communication, Sendai, 1997. 58~63
- Mitsumoto N, Fukuda T, Shimojima K. Control of the Distributed Autonomous Robotic System Based on the Biologically Inspired Immunological Architecture. In: Proc. the 1997 IEEE Intl. Conf. on Robotics and Automation, Albuquerque, NM, USA, 1997. 3551~3556
- Takahashi K, Yamada T. Self-tuning Immune Feedback Controller for Controlling Mechanical Systems. In: Proc. 1997 1st IEEE/ASME Intl. Conf. on Advanced Intelligent Mechatronics, Tokyo, Japan, 1997. 101~105
- Ishida Y, Mizessyn F. Learning Algorithms on an Immune Network Model: Application to Sensor Diagnosis. In: Proc. Intl. Conf. Neural Networks, China, 1992. 33~38
- Chun J-S, Kim M K, Jung H K. Shape Optimization of Electromagnetic Drives Using Immune Algorithm. IEEE Trans on Magnetics, 1997, 33(2): 1876~1879
- Forrest S, Perelson A S, Allen L, Cherukuri R. Self-nonsell Discrimination in a Computer. In: Proc. IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, 1994. 202~212
- D'haeseleer P, Forrest S, Herman P. An Immunological Approach to Change Detection: Algorithms, Analysis and Applications. In: Proc. IEEE Symposium on Research in Security and Privacy, Oakland, CA, USA, 1996. 110~119
- Jiao L C, Wang L. A Novel Genetic Algorithm Based on Immunity. IEEE Trans. on Systems, Man, And Cybernetics-Part A: Systems and Humans, 2000, 30(5): 552~561
- 王磊, 潘进, 焦李成. 基于免疫策略的进化算法. 自然科学进展, 2000, 10(5): 451~455