第 36 卷 第 9 期 计 算 机 工 程 2010 年 5 月
Vol.36 No.9 Computer Engineering May 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)09—0181—03 文献标识码：A 中图分类号：TP18

人工免疫系统中的抗体生成与匹配算法

徐 佳，张 卫
(华东师范大学信息科学技术学院，上海 200241)

摘 要：现有的人工免疫系统被应用于文本识别中时，检测器生成算法对不同基因等质化对待，不能最优反应基因在抗体中出现的频率。针对该问题，提出基因显性度的概念，通过在检测器生成算法及匹配算法中引入基因显性度的因子来提高算法效率。实验结果表明，显性度的引入可降低检测器生成算法约 30%的时间复杂度。

关键词：人工免疫系统；文本识别；匹配算法；检测器生成；显性度

Antibody Generation and Matching Algorithm in Artificial Immune System

XU Jia, ZHANG Wei
(School of Information Science and Technology, East China Normal University, Shanghai 200241)

【Abstract】 Existing Artificial Immune System(AIS) for recognition applications in the text, the detector generating algorithm for different genes, such as the quality of treatment, there can not be the optimal response gene frequency in the antibody deficiency. This paper proposes the concept of degree of gene dominant through the detector generation algorithm and the matching algorithm introduces genes dominant degree factors to improve the efficiency of the algorithm. Experimental results show that the introduction of dominant degree of the detector generating algorithm can be reduced by 30% of the time complexity.

【Key words】 Artificial Immune System(AIS); text recognition; matching algorithm; detector generation; dominant degree

1 概述

从计算角度来看，生物免疫系统是一个高度并行、分布、自适应和自组织的系统，具有很强的学习、识别、记忆和特征提取能力。研究者据此开发了面向应用的免疫系统计算模型——人工免疫系统(Artificial Immune System, AIS)，用于解决工程实际问题。目前，AIS 已发展成为计算智能研究的一个新的分支[1]。

在人工免疫系统的主要仿生机理有：免疫识别，免疫学习，免疫记忆，克隆选择，多样性，公布式，自适应，免疫网络[2]。免疫识别是重要的一个方面。人工免疫系统中的抗原识别，是通过计算抗原和抗体之间的亲和力来实现的，当亲和力达到设定的匹配阈值时，认为该抗体匹配抗原。

抗体抗原的亲和力和它们之间的距离相关[3]。在目前提出的匹配算法主要有 Euclidean 距离、Manhattan 距离、Hamming 距离等。在一些不适合转化为二进制字符串的领域，如 SPAM 识别、文本分类识别、垃圾短信识别等领域。目前通行的做法是根据骨髓模型，把字符作为基因，生成以字符为基因免疫细胞，使用 r-连续位匹配算法等进行亲和力计算。

但这类算法往往不考虑免疫细胞的基本组成—基因(字符)，它在识别对象中出现的概率是不一样的，出现概率高的显然更有代表性。在目前的算法中，它们在检测器的筛选、成熟和匹配过程中的计算权重是一样的，这显然没有达到最优。本文提出基因显性度的概念。基因显性度是指基因在抗原中出现频度与在自体中出现频度之差与自体中出现频度之比。该值可以有效反映基因的抗原特异性。根据显性度计算公式计算基因库里的每个基因的显性值，并引入新的抗体生成算法。使之更符合生物免疫的实际情况。该算法提供了对文本信息识别更好的性能和更低的算法时间复杂性。实验仿真验证了它能显著降低生成检测器的数量。

2 抗原基因库的生成

首先从抗原，即垃圾短信中提呈抗原的组成基因，对每个提呈的基因进行显性度计算。

2.1 基因的提取

通过对垃圾短信集合的语料分析，提取出所有的词，删除无意义的特定词，如“的”及阿拉伯数字。得到原始抗原基因库：

$Origin\_Gene=\{Gene_1,Gene_2,\cdots,Gene_n\}$

提取的基因在正常的短信中也可能存在，需对抗原基因进行否定检测，删除更能代表自体(正常短信)的基因，降低基因数量及检测器生成的算法复杂度。对每个基因，设置两个附加值：抗原匹配 *Ag_affinity*，自体匹配 *B_affinity*。当基因每匹配一个抗原时，将基因 *Ag_affinity* 值加 1，每匹配一个自体，将基因的 *B_affinity* 加 1。得到 2 个附加值能够反映该基因在抗原自体中出现的频度。算法如下：

For each G ∈ Origin Gene
//计算基因的抗原匹配值

作者简介：徐 佳(1976－)，男，讲师、硕士，主研方向：人工免疫系统，计算机网络；张 卫，教授、博士生导师

收稿日期：2009-10-22 **E-mail**：xu9jia@qq.com

```
For i=1 to count(垃圾短信) //count(垃圾短信)为样本包含垃圾短
//信数量
        If G ⊂ 垃圾短信 i then
            Ag_affinity=Ag_affinity+1
        End if
    Next
    //计算基因的抗体匹配值
    For j=1 to count(正常短信 j) // count(正常短信 j)为样本包含正
//常短信数量
        If  ⊂ 正常短信 j then
            B_affinity=B_affinity+1
        End if
    Next
Next
```

//计算后，得到新的抗原基因矢量库 Ag Gene，每个元素包含
//3 个子量：基因 Gene，抗原匹配 Ag_affinity，自体匹配 B_affinity：

Ag Gene=<Gene, Ag_affinity, B_affinity>

Gene={$Gene_1$,$Gene_2$,…, $Gene_n$}

Ag_affinity={$Ag\_affinity_1$, $Ag\_affinity_2$,…, $Ag\_affinity_n$}

B_affinity={$B\_affinity_1$,$B\_affinity_2$,…, $B\_affinity_n$}

2.2 基因显性度计算及基因分类

根据基因在垃圾信息和正常信息中出现频度的不同，引入显性度计算公式：

基因显性度 $DD=(Ag\_affinity_m - B\_affinity_m)/B\_affinity_m$

显性度计算公式能够准确筛选中那些在垃圾邮件中出现的频度远高于在正常邮件中出现频度的基因。根据基因显性度计算的结果，对抗原基因矢量库进行分类，设定 3 个集合：高显性基因集合 *AbGene*，低显性基因集合 *LaGene*，自体基因集合 *BGene*，分类依据如下：

$$AgGene_m\begin{cases}\in AbGene & DD > G_{dominant}\\ \in LaGene & Ag\_affinity_m > B\_affinity_m \wedge DD < G_{dominant}\\ \in BGene & Ag\_affinity_m \leqslant B\_affinity_m\end{cases}$$

其中，$G_{dominant}$ 为显性阈值，且 $G_{dominant}$ >0。$G_{dominant}$ 取值越高，表明该基因的抗原特异性越高。

3 检测器的生成

分类后，在高匹配基因集合内的基因都是与抗体高度匹配的基因，使用该集合内的基因作为检测器的组成基因，基因数量减少，能够降低检测器生成算法的复杂性，提高检测器生成质量。

检测器随机从 *AbGene* 集合中抽取而得，不重复，数量为 N_{Gene}，N_{Gene} 的值的大小与检测器生成算法的复杂度有很大的关系，数值过大，检测器生成算法复杂，但检测效率可以提高，反之生成算法复杂度降低，但检测效率亦相应降低，在不同的应用环境中，N_{Gene} 的取值有所不同，因此，需要选择合适的值。根据应用中抗原的特点，N_{Gene} 的取值设定为 5。

检测器生成后，除了应当能够正常识别抗原(垃圾短信)外，还应当保证不能错误地将自体(正常信息)识别为抗原，因此，需要对生成的检测器进行自体耐受训练，经过训练的检测器才能成熟成为抗体。

3.1 检测器的成熟

成熟过程用检测器的成熟算法来实现。检测器的成熟需要把检测器和检测对象进行匹配，从而过滤掉识别自体的检测器，留下与抗原亲和力最高的一批检测器。

一些文献提出了如下的亲和力计算公式[4]：

$$Affinity(A,B)=\frac{A\cap B}{\min(A,B)}$$

上式在比较 2 类文本的相似度上简单有效，也是目前文本识别领域用得比较多的方法。但此公式对检测器中的任何基因包含于抗原时，它们的计算权重是一样的。

例如，当 2 个检测器中都有 3 个基因与抗原匹配，但一个检测器中是 3 个高显性的基因，另一个检测器中为 3 个低显性的基因，它们的亲和力是一样的。这显然无法正确反应检测器与抗体的匹配程度。为了解决这个问题，在 2.2 节中已入显性度的概念，在亲和力计算公式中也加入显性度的计算因子，以有效反映检测器与抗体的匹配程度，新的匹配算法公式如下：

$$Affinity(detector,Ag)=\sum_{i=1}^{m}DD\cdot match_m, match_m=\begin{cases}1 & Gene_m\subset Ag\\ 0 & Gene_m\not\subset Ag\end{cases}$$

其中，m 为检测器包含的基因数；DD 为基因的显性度 $\frac{Abaffinity_m - Baffinity_m}{Baffinity_m}$；*detector* 为检测器；*Ag* 为被检测的抗原对象。

上式在初始检测器状态计算抗体与抗原的亲和力时有效。但在实验中发现，在进行检测器克隆增生及变异后，由于检测器中引入了低显性基因，使其可能包含较多的与自体匹配的基因，但如该检测器包含有显性度特高的基因时，其依然具有与其他不包含自体匹配基因的检测器近似的亲和力值。为了避免这些缺陷，引入检测器与自体的匹配度公式：

$$BodyAffinity(detecter,B)=\sum_{i=1}^{m}match_m, match_m=\begin{cases}1 & Gene_m\subset Ab\\ 0 & Gene_m\not\subset Ab\end{cases}$$

其中，m 为检测器包含的基因数；B 为自体集合。

使用上述这 2 个公式对所有检测器进行成熟训练，*BodyAffinity* 值达到自体识别阈值 *Db*，*Affinity* 未达到抗体激活阈值 *Da* 的检测器予以删除。余下的检测器集合为与抗体具有高亲和力，且不会错误识别自体的检测器，激活这些细胞后得到的检测器集合即为合格的抗体集合。

3.2 检测器的克隆选择与变异

在生物体免疫系统中，识别抗原的细胞进行克隆扩增，以产生大量抗体。在克隆过程中，免疫细胞会发生变异现象。变异对维护和完善检测器的多样性和保持检测器的高亲和力具有重要的作用。为此引入克隆与变异机制。免疫系统克隆选择过程使用克隆选择算法[5]。

对所有检测器设定生命周期，以匹配次数为时间单位，检测器集合内的高亲和力检测器每经过一轮检测如果被匹配，则将该检测器的生命周期设置为永久，如未被匹配，则生命周期减去一个时间单位。

对一定时期内未被匹配部分检测器进行克隆变异，变异时采用遗传变异算法，新生成的检测器由父辈的基因通过遗传算子而来，通常使用的遗传算子有交叉、复制、变异。这里采用类似于变异的遗传算子，即在检测器的基因中随机选择一个，进行定向变异，定向变异的范围选择为：*AbGene*∪*LaGene*。在生成检测器时将基因选择范围设定为 *AbGene*，而在克隆变异时加上 *LaGene*，可以在保持检测器“先天”具有较高抗体亲和力的基础上，保持检测器的多样性，维持检测器较高的识别率。

当原检测器集合经过若干次检测后，可以激活克隆变异机制，产生一定的新检测器，新检测器也经过成熟算法，成为成熟的检测器细胞，加入到检测器集合中。

为提高系统效率，也可引入诸如基于种群划分及杂交的免疫遗传算法等一些新的研究成果[6]。

4　人工免疫系统的更新

当新的抗原认定达到一定数量时(包括人工协同加入)，则对检测器重新进行亲和力计算，设置达到匹配阀值的检测器生命周期为永久，同时再次进行基因提呈与归类，以筛选出新的高显性基因，加入到 *AbGene* 及 *LaGene* 中，以便进行新的检测器生成时随机选用。同时在检测器集合中删除生命周期终结的检测器，以维持检测器集合的新鲜状态。

检测器对象集合与抗原集合在匹配计算时通常难以完全重合。错误肯定与错误否定的情况就难以避免。

当检测器对象的覆盖达到抗原集合外时，这部分对象被检测会发生错误肯定。错误肯定发生时，采用人工协同刺激的办法，人工将成熟检测器集合中的所有检测器与发生错误肯定的检测对象一一进行亲和力计算，删除所有匹配的检测器。

当检测器对象的集合未完全覆盖抗原集合时，检测这部分抗原时就会发生错误否定，有两种方法能够处理：一是加大检测器集合数量，这是以降低整个系统的效率以获得更低的错误否定率。

更有效的方法是进行人工协同刺激，针对该抗体人工提呈基因，进行基因的显性度计算，并根据计算结果产生的高显性基因随机生成针对性的一些检测器，这些检测器经过成熟算法激活后，加入到检测器库中。这种方法对当前抗原和以后极相似的抗原有效。

5　实验仿真

为了验证理论的可行性与可靠性，对浙江某移动运营商提供的垃圾短信进行实验。共提取到用户投诉的 2 830 个垃圾短信。同时从短信网关提取非群发点对点短信 10 000 个，这些短信被认为是非垃圾短信。

对垃圾短信进行基因提呈、人工加工后得 2 358 个基因，对基因进行显性计算，高显性阈值 $G_{dominant}$ 设置为 0.5，得到高显性基因 1 636 个。随机生成检测器，检测器包含基因数 N_{Gene} 设置为 5，生成初始检测器数目 600 000 个，进行检测器成熟算法筛选，自体识别阈值 *Db* 设置为 2，抗体激活阈值 *Da* 设置为 0.22，得到有效的检测器 39 220 个。

使用成熟检测器集合对原短信库进行检验，错误否定 48 个，错误否定率为 1.73%，未发生错误肯定，错误肯定率为 0%，具有极高的灵敏度。

对第 2 批提供的 474 个垃圾短信和 1 754 个正常短信组成的集合进行检测，错误否定率为 1.9%，错误肯定率为 0%。效果令人满意

为了比较引入基因显性度后系统的优越性，仍以文献[4]的公式作为亲和力计算公式，在检测器的成熟算法和抗原抗体亲和力计算中去除显性度因子。检测器包含基因数量依然为 5，亲和力阈值为 0.6。为了便于比较，在引入显性度的亲和力公式中，将抗体激活阈值 *Da* 增大到 2.4，以便产生多于文献[4]公式的抗体，并按照该公式产生抗体的数量随机抽取相同数量的抗体。比较数据如表 1 所示，比较曲线如图 1 所示。

假设用人工免疫的方法对该样本的最大错误否定率为 1.9%，传统算法在产生 160 000 个初始检测器时趋于收敛，而引入显性度计算因子后，在产生 110 000 个初始检测器时即趋于收敛，仅就检测器生成而言，意味着降低了约 31%的时间复杂度，效率的提高显而易见。

表 1　样本检测数据

原始抗体数	成熟抗体数	传统算法		本文算法	
		错误否定数	错误否定率/(%)	错误否定数	错误否定率/(%)
10 000	644	111	23.42	76	16.03
20 000	1 284	58	12.24	40	8.44
30 000	1 913	31	6.54	26	5.49
40 000	2 583	25	5.27	18	3.80
50 000	3 246	19	4.01	16	3.38
60 000	3 920	18	3.80	13	2.74
70 000	4 569	15	3.16	12	2.53
80 000	5 215	15	3.16	10	2.11
90 000	5 871	13	2.74	10	2.11
100 000	6 554	13	2.74	10	2.11
110 000	7 204	10	2.11	9	1.90
120 000	7 852	10	2.11	9	1.90
130 000	8 526	10	2.11	9	1.90
140 000	9 167	10	2.11	9	1.90
150 000	9 807	10	2.11	9	1.90
160 000	10 395	9	1.90	9	1.90

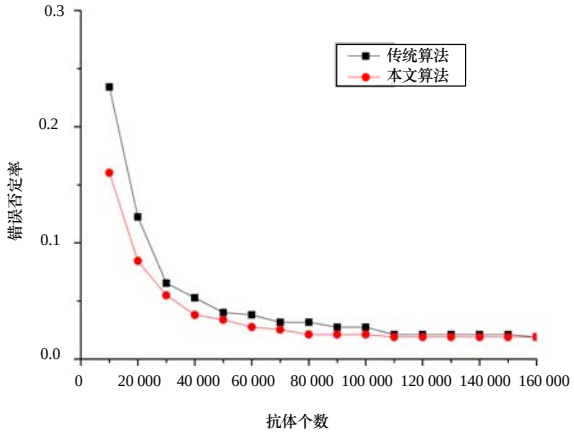

图 1　错误否定率比较

6　结束语

本文分析了人工免疫系统中传统抗体生成算法的缺点，提出了引入基因显性度的新算法，在抗体生成前对组成基因进行优化选择，以提高生成抗体的质量。经过实验，使用新算法生成的抗体具有更优越的识别性能，达到同样错误否定率所需的抗体数量减少了 31%左右，具备一定的实用价值。

参考文献

[1] Dasgupta D, Attoh-Okine N. Immunity Based Systems: A Survey[C]//Proc. of IEEE International Conference on Systems, Man, and Cybernetics. Orlando, Florida, USA: [s. n.], 1997: 369-374.

[2] 肖人彬, 王　磊. 人工免疫系统: 原理、模型、分析及展望[J]. 计算机学报, 2002, 25(12): 1283-1285.

[3] de Castro L N, von Zuben F J. Artificial Immune Systems, Part I: Basic Theory and Applications[R]. Campinas, Brazil: School of Computing and Electrical Engineering, State University of Campinas, Tech. Rep.: DCA-RT 01/99, 1999.

[4] 张泽明, 罗文坚, 王煦法. 一种基于人工免疫的多层垃圾邮件过滤算法[J]. 电子学报, 2006, 34(9): 1616-1620.

[5] de Castro L N, von Zuben F J. Learning and Optimization Using the Clonal Selection Principle[J]. IEEE Trans. on Evolutionary Computation Special Issue on Artificial Immune Systems, 2002, 6(3): 239-251.

[6] 武　妍, 李儒耘. 一种基于种群划分及杂交的免疫遗传算法[J]. 计算机工程, 2008, 34(3): 220-222.

编辑　金胡考