基于人工免疫的中文垃圾短信识别①

徐　佳　张　卫 (华东师范大学 信息科学技术学院 上海 200241)

摘　要： 垃圾短信问题越来越严重，为了识别中文垃圾短信，将人工免疫系统的方法引入中文垃圾短信识别领域。针对中文短信息系统本身的一些特点，提出了适应这些特点的人工免疫算法，并在实验中得到验证。实验结果表明，基于人工免疫的中文垃圾短信识别方法具有较低的错误否定率和错误肯定率。

关键词： 人工免疫；垃圾短信；文本分类识别；匹配算法；检测器生成

Recognition of Chinese Junk SMS Based on Artificial Immune

XU Jia, ZHANG Wei

(East China Normal University, Shanghai 200241, China)

Abstract: Junk SMS problem is getting worse. To identify Chinese junk SMS, an artificial immune system is introduced to areas of the Chinese junk messages recognition. In light of Chinese SMS characteristics, this paper proposes the artificial immune algorithm, which is verified in experiments. The experimental results show that the Chinese junk SMS method based on artificial immune recognition has a lower false negative rate and false positive rate.

Keywords: artificial immune; junk SMS; text classification; matching algorithm; detector generates

短消息服务(SMS, Short Messages Service)是移动通信网络上的一项基础服务。由欧洲电信标准协会 ETIS 提出并规定，应用在 GSM、CDMA、TDMA 无线网络以及一些诸如 ISND 等基于有线技术的网络上。在 ETIS 最初规定，短消息服务最长支持 140 个字节的信息[1]。因短消息服务便捷、价格低廉等原因得到了广泛的使用。但垃圾短信也随之产生，并且日益成为移动通信领域一项公害。短信息包含的信息只有发送号码和信息内容，内容通常非常简短，按照 ETIS 的标准最长是 70 个汉字。所含信息量相对较少。使得在反垃圾邮件的一些技术，象电子邮戳、来源认证等技术无法使用在垃圾短信的识别上。

人工免疫系统是基于生物体免疫系统原理而仿生的一种人工智能方法。从计算的角度来看，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。人们从中获取灵感，开发面向应用的免疫系统计算模型—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崭新的分支[2]。

将生物体免疫系统的一些原理引入垃圾短信识别领域：垃圾短信可以认为是抗原，正常短信认为是自体，建立模仿人体的免疫系统的检测器集合，承担免疫识别的任务，根据生物免疫系统的特点，这个检测器集合还具有免疫学习、记忆、克隆选择等等功能，能够适应垃圾短信类型的变化，保持较高的效率。

本文的贡献在于，根据人工免疫系统的原理，结合短信息系统的实际情况，提出了基于人工免疫系统的，针对于中文垃圾短信的识别算法，并在实验中对该识别算法进行了验证。实验表明，基于人工免疫的中文垃圾短信识别是高效可行的。

1　定义及算法流程

生物体中，基因是指携带有遗传信息的 DNA 序

① 基金项目:台州广播电视大学课题(08KT-012)
收稿时间:2009-06-28

列，是病毒、细菌的基因组成，是进行自体/非自体识别的基本依据。在垃圾短信识别中，基因指垃圾短信和正常短信中所包含的字或词。抗原在生物体中指侵入生物体的、非自体的病毒、细菌等，是免疫系统识别、攻击的对象，在垃圾短信识别中，抗原指的是垃圾短信。抗体是生物体的免疫系统，用来识别并排除非本体的病毒、细菌等，在垃圾短信识别中，抗体指的是根据基因生成的检测器的集合，用来检测一条短信是否垃圾短信。生物体中的自体在这里则是指正常的短信。表 1 为生物体与人工免疫系统定义对照：

表 1　定义对照表

	生物体	人工免疫
基因	DNA 序列	字、词
抗原	入侵的病毒、细菌	垃圾短信
抗体	免疫系统的各种细胞	检测器
自体	生物体本身	正常短信

定义基因集合：

Gene={Gene1,Gene2, … ,Genen};　m,Genen Agm

抗原集合 Ag={Ag1,Ag2,…,Agn};

抗体集合 Ab={Ab1,Ab2,…,Abn};

自体集合 B={B1,B2,…,Bn}

对整个人工免疫系统，构造流程图如图 1：

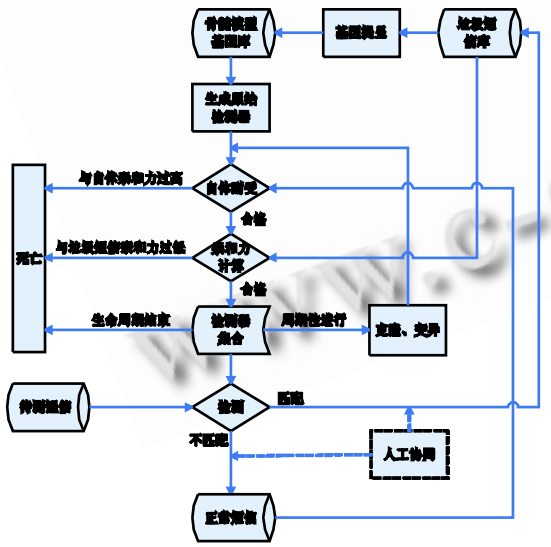

图 1　基于人工免疫的垃圾短信过滤流程图

2　抗原基因库的生成

生物体中抗体识别抗原是根据抗原所包含的基因与抗体的匹配程度。而在基于人工免疫的垃圾短信识别时，基因指的是垃圾短信所包含的文字。作为生成检测器(抗体细胞)的基础工作，首先从抗原，也就是垃圾短信中提呈抗原基因，找出垃圾短信所包含的所有基因。

2.1 基因的提呈

通过对垃圾短信集合进行汉字的提取，将其中一些无意义的特定词，如“的”、及阿拉伯数字、英文字母等人工删除。是后得到原始抗原基因库：

Gene= {Gene1,Gene2,…,Genen}

2.2 基因的更新

抗原基因集合中的基因是动态的，当有新的抗原加入时，必须对新抗原作一次基因提呈工作，将其中新的基因加入到基因库中。在垃圾短信检测时，如果发生了错误否定，在人工协同刺激下，也会加入新的抗原，达到一定数量时同样需要进行基因提呈，这一点在下文会具体探讨。

3　检测器的生成

经过对抗原提呈的基因进行预处理后，可以用生成的基因集合库来生成检测器。在垃圾邮件检测中，对于检测器的组成，有采用随机长度的；也有采用垃圾邮件社区概念的，由一组垃圾邮件组成社区作为检测器来进行匹配；也有采用比较长的固定长度的。根据短信信息量小的特点，采用固定长度基因组成检测器的方法来随机产生检测器集合。

在此采用骨髓模型[3]，检测器由数量为 N_{Gene} 的基因组成，基因随机函数从 Gene 集合中抽取而得，不重复。N_{Gene} 的取值设定为 5。由于基因库是一个有限集合，所以检测器集合的空间也是有限的，根据 NGene 的取值和基因库集合空间的大小，检测器集合的空间上限为：

$$O(\text{detector}) = \prod_{i=1}^{N_{gene}} count(Gene)^2 + (1-i)\,count(Gene)$$

检测器的数量在错误否定率与错误肯定率可以接受的情况下保持最少的数量。此外 R J Boer 和 A S Perelson 提出初始检测器与自体集合的大小是成指数关系的：

$$N_{R_0} = \frac{-\ln(P_f)}{P_m \cdot (1-P_m)^{N_s}}$$ [4]

Pm 为匹配率，Pf 为错误否定率，NR0 为初始检测器数，NS 为自体集合。这个公式主要是针对二进制

串的匹配，在字符匹配中会有所不同，检测器数量的实际情况，还需根据实验结果来确定，在第 7 节实验中会有论述。

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(正常短)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

这个过程用检测器的成熟算法来实现。把检测器和正常短信进行匹配，过滤掉与正常短信匹配度过高的检测器。

在人工免疫系统中，匹配程度的计算，是通过计算它们之间的亲和力来实现的，当亲和力达到设定的亲和力阈值时，认为它们相匹配。抗体抗原的亲和力和它们之间的距离相关[5]。目前在入侵检测、病毒检测等领域应用的 Euclidean 距离、Manhattan 距离、Hamming 距离等以及 r-连续位匹配等，用来计算能用二进制数列表达的抗原亲和力时，有较好的效果。但在不适合转化为二进制字符串的领域，如垃圾邮件识别，垃圾短信识别等领域，则不适合使用。因此需要提出新的匹配算法。

进行匹配计算前首先要确定抗原的表示问题，抗原也是由基因组成的，也可以表示为一组基因集合：

Ag={Gene1,Gene2,⋯,Genem}

由于短信息在长度方面的限制，为了进一步降低处理时的计算复杂度，可以考虑对抗原不提取基因，直接用原字符串进行处理。匹配度计算公式为：

$$Affinity(\text{detector},Ag)=\sum_{n=1}^{N_{gen\in c}} match_n, match_n=\begin{cases}1, Gene_n\subset Ag, Gene_n\in Ag\\ 0, Gene_n\not\subset Ag, Gene_n\in Ag\end{cases}$$

当 Affinity 函数的值大于或者等于匹配阀值 Affi 时，则认为检测器 detector 与抗原 Ag 是匹配的。使用这个公式对检测器进行成熟训练，预先设定自体匹配阀值，所有与 Ag 的匹配度超过阀值的检测器将被删除我。检测器经过训练后成为成熟的检测器，组成检测器集合 Detector。

对所有检测器，设定生命周期 Gtime，以匹配次数为时间单位，检测器集合内的检测器每经过一次检测时，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。首先应用样本垃圾短信对所有检测器进行匹配检测，如获得匹配，设置该检测器的生命周期 Gtime 为 0，表示永久保留，该检测器相当于生物免疫系统中的记忆细胞，具有强烈的二次应答能力。如未被匹配，设置该检测器的生命周期 Gtime 为一固定值。

4 垃圾信息的检测

当新的被检测对象送达后，用检测器集合中的元素逐一用亲和力计算公式进行亲和力计算，只要有一个检测器的亲和力阀值达到抗体匹配阀值 Affi，该检测对象即被判定为垃圾短信，对象同时被送入垃圾短信社区作为下一轮基因的备选。与该对象相匹配的检测器的生命周期 Gtime 则被置为 0，未被匹配的检测器的则被减去 1，当被减到 1 时，则该检测器被为无效的检测器，其生命周期终结，从检测器集合中删除。这个机制能够保证检测器集合中的数量保持在一定的范围内，不会被无限地扩大。

生物体中的免疫细胞对入侵者的检测分别由记忆细胞应答和非记忆细胞应答，前者产生强烈的二次应答反应，能够更加有效地识别抗原并协助免疫系统杀死入侵者。它是生物体能够对二次感染的病毒、细胞产生免疫力的原因。从计算机这个工具的本身特点来看，未必要模拟生物体的这一特性，只需在上述算法中，在抗原与检测器一一匹配时，按照各个检测器在成熟过程中匹配公式 Affinity()得到的值从高到低进行匹配计算，也能够取得同样的效果，且算法更加简单。因此在基于人工免疫的垃圾短信识别中不区分检测器是否为记忆细胞，原始检测器经过耐受训练后直接成熟。

5 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，变异后的抗原细胞由与抗原的亲和力来决定扩增和删除。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[6]。

对检测器集合中，生命周期为永久的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[3]。这里采用类似于变异的遗传算子，即在检测器的基因中选择一个没有匹配的

基因，进行定向变异，定向变异的范围选择为：基因集合 Gene，随机抽取一个新基因，代替原检测器中没有被检测到的基因组成新的检测器。

生成的检测器均保持了其父代检测器的高亲和力，经过变异后，可能取得更高的匹配程度。经过成熟算法的训练，变成成熟的检测器加入到检测器集合中。

6 人工免疫系统的动态更新

人工免疫系统在运行过程中，抗原集合不断扩大，检测过程中的人工协同刺激也可能不断把新的抗原加入到抗原集合中，抗体必须具备不断地动态更新能力才能适应抗原类型不断新增和变化。当新的抗原认定达到一定数量时，为了保持检测器集合的广谱性，必需在基因集合中反应抗原的变化。因此对新的抗原，进行新的基因提呈，将提呈的新基因加入到基因集合 Gene 中。再按比例生成一定的新检测器进行成熟训练，以维持检测器集合的新鲜状态。

6.1 错误肯定与错误否定的处理

从集合上看，当检测器对象集合的覆盖达到垃圾集合外时，这部分对象被检测会发生错误肯定。错误肯定发生时，一般采用比较简单的处理办法，即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进行一一亲和力计算，删除所有匹配的检测器。

当检测器对象的集合未完全覆盖垃圾短信集合时，检测这部分垃圾短信时就会发生错误否定，可以加大检测器集合数量来应对，但这种方法会加大系统的负担，降低系统运行效率且未必能达到预期目标。可以进行人工协同刺激，把被错误否定的垃圾短信加入到垃圾短信集合中。

6.2 人工协同刺激

当错误否定发生时，手工在垃圾短信集合中加入目标短信，进行人工协同刺激是一种高效的解决问题的方法。做法是，专门对加入的垃圾短信进行基因提呈，并根据提呈的基因生成若干数量的检测器，对检测器进行成熟训练，成熟后的检测器加入到检测器集合中。这些检测器对于该垃圾短信及类似的垃圾短信具有极高的亲和力。

7 实验

从某移动运营商提取到 2830 个垃圾短信，和点对点短信 10000 个，这些短信被认为是非垃圾短信。对 2830 个垃圾短信进行基因提呈，人工删除一些无用基因，如“的”、数字、符号后，余下 2358 个基因，作为基因集合。根据基因集合生成检测器，检测器包含基因数量 Ngene=5，亲和力阀值设置为 3，基因生命周期 Gtime 设置为 2000，并对生成的检测器对照原垃圾短信与正常短信集合进行自体耐受的成熟训练，形成检测器集合。 对第二批提供的 474 个垃圾短信和 1754 个正常短信组成的集合进行检测，实验数据见表 2：

表 2 实验数据

原始检测器数量	成熟检测器数量	错误肯定数	错误肯定率	错误否定数	错误否定率
10000	644	0	0%	111	23.42%
20000	1284	0	0%	58	12.24%
30000	1913	0	0%	31	6.54%
40000	2583	0	0%	25	5.27%
50000	3246	0	0%	19	4.01%
60000	3920	0	0%	18	3.80%
70000	4569	0	0%	15	3.16%
80000	5215	0	0%	15	3.16%
90000	5871	0	0%	13	2.74%
100000	6554	0	0%	13	2.74%
110000	7204	0	0%	10	2.11%
120000	7852	0	0%	10	2.11%
130000	8526	0	0%	10	2.11%
140000	9167	0	0%	10	2.11%
150000	9807	0	0%	10	2.11%
160000	10395	0	0%	9	1.90%

当生成原始检测器数 160000 个时，产生的成熟检测器数为 10395，此时错误否定率为 1.9%，错误肯定率为 0%。效果令人满意。由于样本的提取时间、类型比较接近，可能对效果产生一定的影响，如果在实际中应用，可能错误否定率与错误肯定率有所降低。

为了分析原始检测器数量与错误否定率的关系，根据图 2 数据曲线进行函数拟合，得如下公式：

$$N_{\text{detector}} = In\frac{P_f - 0.018}{0.32}\cdot 2.63$$

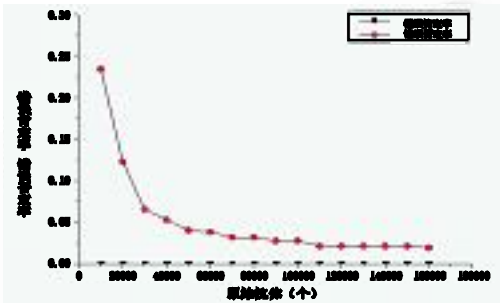

图 2　错误肯定率与错误否定率

其中 $N_{detector}$ 为初始检测器数，单位为万个；Pf 为错误否定率。可以看出，当产生 120000 个检测器时，已经趋于收敛。是考虑效率与功能平衡的较优状态。在实际应用中，检测器数量还需考虑垃圾短信规模以及文献[4]初始检测器与自体规模的关系动态调整初始检测器的数量，以求达到检测器数量、错误否定率、错误肯定率间的最优组合。

参考文献

1 ETIS. ETS90301 Digital cellular telecommunications system (Phase 2+); Technical realization of the Short Message Service (SMS); Point-to-Point (PP). Sophia Antipolice: European Telecommunications Standards Institute, 1998.

2 Dasgupta D, Attoh-Okine N. Immunity based systems: A survey: IEEE International Conference on Systems, Man, and Cybernetics. Orlando, 1997. 369－374.

3 李涛.计算机免疫学.北京:电子工业出版社, 2004. 44,47.

4 De Boer RJ, Perelson AS. How diverse should the immune system be? Proc. of the royal Society London B, v.252, 1993. London,S.N.

5 De Castro LN, Von Zuben FJ. Artificial Immune Systems: Part I-Basic Theory and Applications. Technical Report RT DCA. Brazil: Campinas, 1999.

6 De Castro LN, Von Zuben FJ. Learning and optimization using the clonal selection principle. IEEE Trans on Evolutionary Computation Special Issue on Artificial Immune Systems, 2002,6(3):239－251.