文章编号:1001 - 9081(2005)11 - 2559 - 03

基于人工免疫系统的反垃圾邮件过滤机制

胡 可,张家树
(西南交通大学 信号与信息处理省重点实验室,四川 成都 610031)
(toplilis@ sina. com)

摘 要:提出了一种基于人工免疫系统(AIS)的垃圾邮件过滤机制。将邮件文本向量空间化并结合免疫识别算法进行过滤。对机制进行了建模与算法描述,对检测器的性能和系统的学习更新进行了向量空间上 r 模拟仿真并与 Naïve Bayes 方法进行比较。研究结果说明将人工免疫系统应用于垃圾邮件处理有动态性和自适应强的优点,同时为特定领域的信息分类问题提供了一种参考机制。

关键词:人工免疫系统;信息过滤;垃圾邮件;自然免疫系统
中图分类号: TP393.098 **文献标识码**:A

Anti-spam filtering mechanism based on artificial immune systems

HU Ke, ZHANG Jia-shu
(*Sichuan Province Key Laboratory of Signal and Information Processing, Southwest Jiaotong University, Chengdu Sichuan* 610031, *China*)

Abstract: Based on artificial immune systems, an anti-span E-mail filtering system, called EIS, was developed. The mechanism and algorithms of this system were described. The simulation results show that it performs as well as Naïve Bayes system, especially on recall, and is fit for information classification.

Key words: AIS(Artificial Immune System); information filtering; spam; NIS(Natural Immune System)

0 引言

电子邮件以其方便、快捷、高效的优势逐渐成为了人们日常生活中信息交互的重要手段之一[1]。但同时它也带来了许多负面影响,在人们每天收到的邮件中有相当一部分是我们不需要的,它们或者是广告,或者是病毒,也可能是一些反动不良信息。针对这种垃圾邮件泛滥的现状,出现了不少的垃圾邮件过滤系统[2~3]。它们所采用的判别方法可以大体分成基于规则的方法[2]和基于概率统计[3]两种。

基于规则的过滤机制方法简单,实现容易,可以体现用户个人需求的主观意识,但由于垃圾邮件的规律非常不明显,这种人为的主观性必然造成大量邮件的漏判与误判;同时,规则的制定往往是静态的,不能及时根据最新的要求自动更新。基于概率统计的过滤机制着重于寻求邮件内在的概率统计关系,不需较多的人为干预,可以实现客观的、规律的识别判断,但其判断依据是以往的统计数据,识别时很大程度上依靠前次训练结果,动态适应性较弱。

人工免疫系统(Artificial Immune Systems, AIS)[4,5]通过模仿生物抵御外来侵袭的构造机能,在人工领域实现对计算机安全、模式识别、智能信息处理、数据挖掘、自动控制等的优化,它的主动性、动态性、自适应性使其在信息处理中体现了极大的优势。

本文通过将 AIS 在 E-mail 的垃圾邮件过滤体系上建模,并对检测器的性能和动态学习进行了仿真验证。结果说明系统对初始训练不敏感,学习过程呈现动态,识别的重心和范围随每次检测结果自适应调节,同时也表明 AIS 对信息的动态、主动处理有着巨大优势。

1 人工免疫系统与垃圾邮件过滤系统

1.1 自然免疫系统

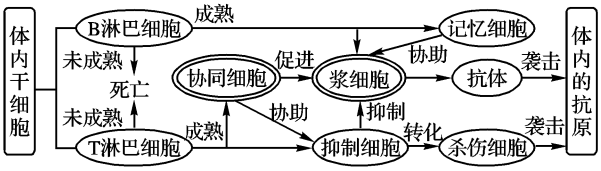

图 1 生物免疫系统免疫机理

自然免疫系统(Natural Immune Systems, NIS)是生物体用来保护其免受外部的病毒与细菌的侵入与感染的一个多层次、多结构的自适应防御体系,主要包括皮肤、体液、淋巴等部分。它的运行机理是:首先由体内特定区域随机生成检测病原体的淋巴细胞,即 B 淋巴细胞和 T 淋巴细胞[5],然后将这些细胞与自身细胞进行阴性选择[4],存活的淋巴细胞就可以离开这些特定区域成为成熟的细胞(表明附有抗体),并可以识别哪些是非自身细胞,从而对病原体(包含多种不同的抗原)进行消灭。这些成熟的淋巴细胞随体内循环系统遍布全身各个部分,倘若在其生命周期中未能黏附任何的病原体,就会自动死亡,而黏附了较多病原体(结合的抗原数目超过激活门限)的细胞则会被激活,同时被大量的克隆复制(变异概率较高),这个过程称为克隆选择[4]。最后 B 淋巴细胞会分化为两种:一种成为记忆细胞,记录曾遇到过的病原体,使其再次遇到时能快速反映识别;另一种则转化为浆细胞,产生大量的抗体以消灭抗原。T 淋巴细胞也分化为两种:一种成为协同细胞,为 B 细胞及其他细胞提供辅助信号;另一种则用来辅助其他细胞或直接消灭病原体。

1.2 人工免疫系统

人工免疫系统源于生物(主要指人)的自然免疫系统。早在 1974 年美国免疫学家 Jerne 就提出了免疫网络理论。随后 Forrest 根据生物免疫系统所具有的阴性选择机理提出了

收稿日期:2005 - 05 - 11;**修订日期**:2005 - 08 - 18

作者简介:胡可(1980 -),男,内蒙古包头人,硕士研究生,主要研究方向:智能信息处理、人工免疫系统; 张家树(1965 -),男,四川南充人,教授,博士生导师,主要研究方向:现代通信信号分析与处理、网络信息处理及其信息安全技术、多媒体信息传输与宽带网、非线性系统与混沌信息工程学.

阴性选择算法[6]，同时将其应用于网络安全，进行计算机异常入侵检测。Secker 等将 AIS 应用于邮件分类，构造了 AISEC[7] 系统，不过它所针对的只是邮件的主题和发信人。在应用 AIS 对信息进行处理，尤其是信息分类时，可以把需要的信息定义为内部信息（Self），不需要的定义为外部信息（Non-self），系统通过探究 Self 和 Non-self 内在规则来构造检测器。检测器最初可随机产生，然后在 Self 集合中通过反向选择的筛选，实现对 Non-self 的识别。除此之外检测器还要具有记忆功能（类似生物体的 B 细胞记忆功能）以便今后的高效判别，同时检测器的检测范围也可变（类似生物体中的细胞变异）。这样当外部信息进入时，由检测器来识别其是否为 Non-self，并做出相应的处理。

1.3 垃圾邮件过滤系统

垃圾邮件过滤是一个特定领域的信息分类问题[1]。它的用户各异性较强，常常是因人而异，因时而异的，有别于通常的信息分类。现有基于显式规则的垃圾邮件的过滤系统有决策树方法[2]、Rough Set 方法等；基于统计的有 KNN 方法、Naïve Bayes 方法[3]等。无论何种方法构造的系统，一般都包含 4 个步骤。首先是邮件预处理，主要进行邮件头信息和正文内容的分离与各种格式和编码的转换，同时如果是针对汉语，要进行分词处理；其次是邮件内容特征提取，主要是从处理过的邮件中提取反映其真实内容的关键词，一般对其进行向量空间化（VSM），用空间矢量来抽象邮件内容；然后是特征匹配过滤，将提取的邮件特征与特征库（经过训练的）进行比对，其相似度大于阈值的，将得到相应处理；最后是用户反馈信息处理，由用户对处理的结果给出评价和纠正，以便为今后的处理提高准确率。整个系统处理过程类似于生物的免疫防护系统，因此利用 AIS 方式对垃圾邮件进行处理必然可以取得较好的过滤效果。

2 AIS 在垃圾邮件过滤上的应用

2.1 EIS 系统

利用 AIS 在信息处理上的各种优势，将其应用到垃圾邮件过滤上，本文构造一个电子邮件免疫系统（E-mail Immune Systems，EIS）。这个系统的任务就是通过学习训练对新邮件进行判断识别，将其分为垃圾邮件和非垃圾邮件。

表 1 NIS 与 EIS 的系统关系

Natural IS	Email IS
自身 Self	合法邮件
非自身 Non-self	待检测邮件
未成熟细胞 Naïve Cells	垃圾邮件检测器
记忆细胞 Memory Cells	垃圾邮件检测器
抗原 Antigens	垃圾邮件
共刺激 Co-stimulation	用户的反馈信息
抗体基因库 Antibody Gene Libraries	邮件标引词特征空间
细胞亲合度 Affinity	邮件间的相似度
细胞生命周期 Cells Lifecycle	检测器的存在有效期

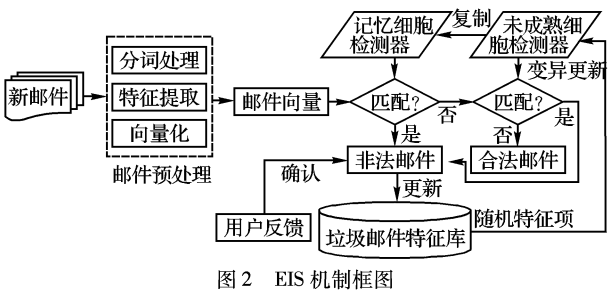

图 2 EIS 机制框图

EIS 是通过对垃圾邮件的训练建立垃圾邮件特征库来构造检测器，当有新邮件时，与特征库中的检测器进行相似度比对，大于阈值的会被认为是垃圾邮件并将被提取特征插入到特征库中，同时所有小于阈值的检测器其生命周期将被减 1，而大于阈值的检测器将被复制，一部分将被存储到记忆库中用于下次高效的快速检测，而在另一部分中按相似度排序后选取一定比例进行变异，整个系统更新完毕。

2.2 EIS 算法实现

2.2.1 邮件预处理

邮件预处理是整个系统的前提。当收到新邮件时，先进行语种分离，主要是为对中文分词，一般可以采用较为简单的前向最大匹配分词法（FMM），处理之后得到的是邮件的字符流，再通过禁用词表、虚词表、词根表去除意义不大的词条，得到表示邮件的项集，设为 q：

$$q = \{s_1, s_2, \cdots, s_i, \cdots, s_h\} \tag{1}$$

其中，s_i 表示邮件 q 的一个特征词条。

2.2.2 系统训练

首先，需要定义一个垃圾邮件的特征库 T：

$$T = \{t_1, t_2, \cdots, t_j, \cdots, t_n\} \tag{2}$$

其中，t_i 表示一个特征词，T 实质是从垃圾邮件中提取的一些短语、词条，或是手工描述项构成的一个标引词空间。如果把每个特征词看作是一个向量，代表空间中的一维，则样本中任一个垃圾邮件都可以表示为这个多维空间的一个向量，这里设为 D_i：

$$D_i = \{d_{i1}, d_{i2}, \cdots, d_{ij}, \cdots, d_{in}\} \tag{3}$$

其中，d_{ij} 表示第 j 个特征词在邮件 D_i 中的权重。权重的设定采用通常的 $tf \times idf$ 处理：

$tf \times idf$ 权重 / 邮件向量的欧氏长度

其中，tf 为项的词频，idf 为逆邮件频率，向量的欧氏长度为：

$$\sqrt{w_1^2 + w_2^2 + \cdots + w_n^2} \tag{4}$$

w_i 是第 i 个项在邮件中的 $tf \times idf$ 权重；特殊地，未出现的词，其值为 0。

然后根据所生成的特征库形成检测器空间 $\boldsymbol{A}$：

$$\boldsymbol{A} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1j} \\ d_{21} & d_{22} & \cdots & d_{2j} \\ \vdots & \vdots & & \vdots \\ d_{i1} & d_{i2} & \cdots & d_{ij} \end{bmatrix} \tag{5}$$

其中，D_i 就是垃圾邮件的检测器。

2.2.3 相似度识别

待检测邮件项集 q 经过特征库标引，形成邮件向量 Q：

$$Q = \{q_1, q_2, \cdots, q_n\} \tag{6}$$

其中 q_i 表示特征词在邮件中的权重。识别时，Q 分别与检测器 D_i 匹配，其相似度定义为 Aff：

$$Aff = sim(D_i, Q) = D_i \cdot Q \tag{7}$$

设定门限阈值为 K_a，则当 $Aff \geqslant K_a$ 时，判定邮件 q 为垃圾邮件（要求用户的确定应答）。

2.2.4 系统学习更新

被检测为垃圾邮件的 q，其特征词条 s_i 将加入到特征库 T 中，成为 T_g，同时在检测器集合 $D = \{D_1, D_2, \cdots, D_i, \cdots, D_m\}$ 中，设满足 $Aff \geqslant K_a$ 的集合为 D_d，不满足的为 D_x。D_x 中的检测器生命周期将递减，而集合 D_d 将被复制，一些拷贝到具有更长生命周期的记忆检测器集合 M_c 中，另一些按相似度排序，从大到小选取一定数目（比例为 $K_g \in (0,1)$）的检测器，设其集合为 D_{dk}，对其中每个检测器 D_{dki} 的特征项用 T_g 重新计算权重，得到变异后的 D_{dki}'，再将其加入到集合中，同时从中删除相同数目（比例为 K_g）的相似度最小的检测器，这样系统完成一次更新。

另外，在整个过程中需要考虑两个刺激信号：一个是内部的检测器生命周期信号，这里设其为常整数，在未成熟检测器

集中为 L_n,记忆检测器集中为 $L_m = K_l L_n (K_l \geqslant 10)$。每次检测中,如果检测器没有被激活,则其生命周期减 1,为 $L' = L - 1$,当值为 0 时,检测器将自动消灭;如果检测器被激活,则其生命周期加 1,为 $L'' = L + 1$。这样做可以使无效的检测器不占有系统资源,有效的检测器更有效地为系统服务,提高整个系统的效率。另一个信号是用户的反馈信号,表示为布尔变量 U,被识别为垃圾的邮件会要求用户给出应答,得到肯定确认,则返回值为 1,否则为 0。若为 $U = 0$,则该封邮件将不被系统学习。

3　实验结果

3.1　仿真

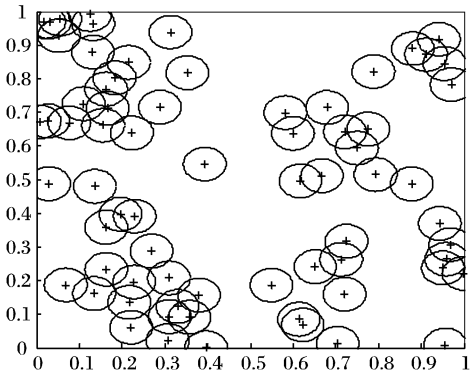

图 3　样本训练生成检测器(图中 + 表示)

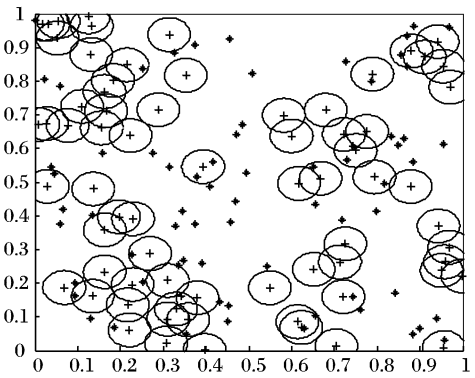

图 4　待检测量(图中 * 表示)输入检测器

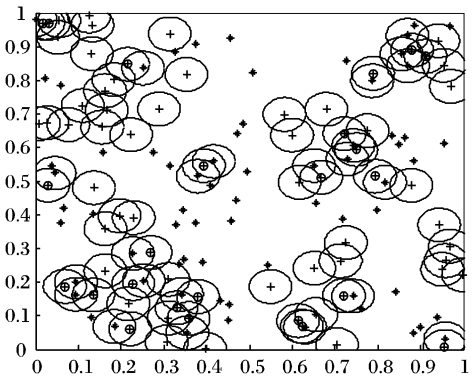

图 5　检测器识别与新检测器(图中虚线表示)生成过程

由于在经过预处理之后,系统的免疫算法主要针对的是向量的匹配及空间的重构,因此本文针对向量来模拟实际的邮件检测,说明系统学习的机理与效果。

在规范化的([0,1];[0,1])二维平面内,先取一定数目(大于 70)的初始训练样本,如图 3 所示,每个样本所生成的检测范围如图上的圆圈所示,这里设检测范围的半径为 0.05。检测器比较集中的地方表明其包含的特征词较多。这时将随机检测向量输入,如图 4 所示,当这些新邮件特征落入到检测器的范围内时,检测器将被激活,同时发生变异,生成新的检测器,新检测器的识别范围如图 5 所示。可以看到新检测器的检测范围包括了未成熟检测器不能检测的范围,实现了系统的动态学习。经过数次检测以后,如果未成熟检测器一直没被激活,到达其生命周期后就会自我消灭,如图 6 所示,形成了新的检测空间。学习后的检测在特征上更加集中,同时摒弃掉了许多由于初始条件不恰当造成的"伪特征"检测器。

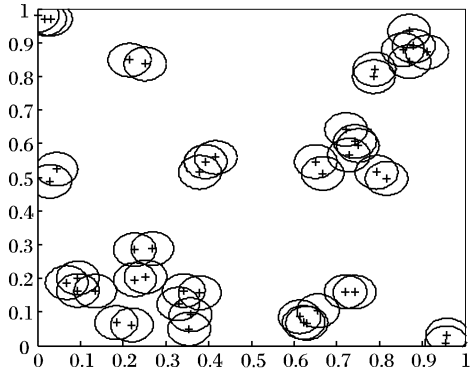

图 6　消除未成熟检测器后更新的系统

3.2　性能分析

通过以上的仿真结果,可以看到 EIS 系统的识别过程中也是一个动态学习过程,体现了系统自动更新的自适应能力。为了说明它在垃圾邮件过滤中的实际效果,这里将它与 Naïve Bayes 方法进行了实验比较,结果见表 2。

表 2　Naïve Bayes 与 EIS 的系统实验结果

过滤系统	正确率(%)	召回率(%)	精确率(%)
Naïve Bayes	89.77	71.01	85.33
EIS	84.50	83.41	85.66

表 2 是对 800 多封邮件进行实验的结果,可以看出 EIS 在正确率比 Naïve Bayes 差,精确率上相近,在召回率上则明显高出,原因在于 AIS 有动态主动检测机能,使得被漏检的邮件有了明显的下降。同时可以看到影响结果的因素主要有两个:其一是相似度门限阈值 K_a 的设定,它体现了检测器的检测范围。K_a 过大,会造成特征的误判,降低精确率;K_a 过小,又会造成特征的漏判,降低了召回率,这是典型的"历史效应"问题。这里考虑到邮件的误判要比漏判危害大,因此 K_a 要设得相对较小。其二是生命周期 L 的设定,它体现了系统的动态性。L 过大,会造成冗余的"伪检测器",加大系统负担,降低效率,也增加了错误率;L 过小,又会造成特征的遗漏和丢失,降低判别准确度。具体到邮件应用时,可以考虑由用户根据需求设定 L,如果处理邮件不多可以将 L 设得大些,以提高准确度;反之,则设得小些,以提高效率。

参考文献:

[1] COHEN WW. Learning Rules that Classify E-mail[A]. AAAI Spring Symposium on Machine Learning in information access[C], 1996.

[2] CARRERAS X, MARQUEZ L. Boosting Trees for Anti-Spam Email Filtering[A]. Proceedings of Euro Conference Recent Advances in NLP (RANLP-2001)[C], 2001. 58-64.

[3] SAHAMI M, DUMAIS S, HECKERMAN D, *et al*. A Bayesian approach to filtering junk e-mail[A]. Proceedings of AAAI Workshop on Learning for Text Categorization[C], 1998.55-62.

[4] HOFMEYR SA, FORREST S. Immunity by Design: An Artificial Immune System[A]. Proceedings of GECCO-99[C], 1999.1289-1296.

[5] MANNIE MD. Immunological Self-Nonself Discrimination[J]. Immunologic Research, 1999, 19(1):65-87.

[6] FORREST S, HOFMEYR SA, SOMAYAJI A. Computer immunology[J]. Communications of the ACM, 1997, 40(10): 88-96.

[7] SECKER A, FREITAS AA, TIMMIS J. AISEC: an artificial immune system for e-mail classification[A]. The 2003 Congress on Evolutionary Computation[C], 2003, 1.131-138.