一种混合的垃圾邮件过滤算法研究

秦志光，罗 琴，张凤荔

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】贝叶斯邮件过滤器具有较强的分类能力，极高的准确率，在内容过滤领域占据主导地位。人工免疫系统具备强大的自学习、自适应，鲁棒性等能力，已发展成为计算智能研究的一个崭新的分支。该文在分析贝叶斯的原理和人工免疫的仿生机理的基础上，将贝叶斯与人工免疫相结合，设计和实现了一种基于贝叶斯和人工免疫的混合垃圾邮件过滤算法，并利用现有的垃圾邮件语料库得到预期的实验结果。

关 键 词 人工免疫; 垃圾邮件; 贝叶斯; 邮件过滤算法
中图分类号 TP311 文献标识码 A

Research of a Hybrid Spam Filtering Algorithm

QIN Zhi-guang，LUO Qin，ZHANG Feng-li

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract Bayes filtering has a dominant place in the area of spam filtering for its strong categorization and high precision. Artificial immune system has become a new embranchment in computing intelligence for its good self-learning, self-adaptability, and robusticity. This paper analyzes the basic principle of Bayes and artificial immune systems, proposes a hybrid spam filtering algorithm based on Bayes and artificial immune system, and then designs and develops the spam filtering system based on this algorithm. It is proved that this system is effective to filter spam in English and Chinese e-mail corpus.

Key words artificial immune; Bayes; spam; spam filtering

近年来，垃圾邮件的泛滥使网络用户在使用网络的过程中受到极大的影响[1]。调查数据显示，全球65%的电子邮件是垃圾邮件。美国更甚，90%的电子邮件为垃圾邮件。到2006年，我国每年收到的500亿封的邮件中，200多亿封为垃圾邮件。反垃圾邮件技术已成为全球所关注的热点。目前，基于内容的垃圾邮件的判别方法大体分成基于概率统计的方法和基于规则的方法[2]。贝叶斯邮件过滤器具有较强的分类能力和极高的准确率，但动态适应性较弱，对未知样本，特别是对已有样本的变异样本和利用Bayes模型生成的垃圾邮件样本，其判别往往不尽如人意[3]；人工免疫系统具备强大的自学习、自适应，鲁棒性等能力。本文在分析贝叶斯原理和人工免疫仿生机理的基础上，结合它们的优点，研究基于贝叶斯和人工免疫的混合垃圾邮件过滤算法，应用于垃圾邮件过滤系统中，并分别在中、英文语料集上做对比测试及分析，取得了较好的效果。

1 贝叶斯算法

朴素贝叶斯算法是通过假定各因素之间不存在任何联系，根据贝叶斯概率公式，对于给定的向量 $\boldsymbol{d}(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n)$ 属于第 C_k 类的概率为：

$$p(C_k|\boldsymbol{d})=\frac{p(C_k)\times p(\boldsymbol{d}|C_k)}{p(\boldsymbol{d})} \qquad k=1,2,\cdots,m \qquad (1)$$

式中 $p(\boldsymbol{d})=\sum_{k=1}^{n} p(\boldsymbol{d}|C_k)\times p(C_k)$。由式(1)可知，要判断一个待识别邮件的类别，可以通过计算 $p(C_k|\boldsymbol{d})$ 概率来完成，根据该文档中出现的单词与向量空间模型中特征项的匹配情况，决定该文档属于第C_k类的概率。假定 $\boldsymbol{w}_j$ 表示第j个特征项。基于文档中单词出现的概率相对独立的假设，有：

$$p(\boldsymbol{d}\,|\,C_k)=p(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\,|\,C_k)=\prod_{i=1,k=1}^{n} p(\boldsymbol{w}_i\,|\,C_k) \quad (2)$$

垃圾邮件过滤中涉及垃圾邮件类别和非垃圾邮件类别。C=0/1表示正常邮件类/垃圾邮件类，则：

收稿日期：2007 – 04 – 25
基金项目：国家242计划资助项目(2005C58)；国家863计划资助项目(2006AA01Z411)
作者简介：秦志光(1956 –), 男, 教授, 博士生导师, 主要从事信息安全和网络安全等方面的研究; 罗 琴(1982 –), 女, 硕士, 主要从事信息安全方面的研究; 张凤荔(1963 –), 女, 教授, 主要从事移动数据处理和网络安全方面的研究.

$$p(\boldsymbol{d})=p(C=1)p(\boldsymbol{d}|C=1)+(1-p(C=1))p(\boldsymbol{d}|C=0) \quad (3)$$

训练样本集中，假定N为邮件总数；N_L为正常邮件数；N_S为垃圾邮件数；n_1包含特征向量$\boldsymbol{w}_i$的正常邮件数；n_s包含特征向量$\boldsymbol{w}_i$的垃圾邮件数，则：

$$\begin{cases} p(C=1)=\dfrac{N_S}{N} \\ p(\boldsymbol{d}\mid C=1)=\prod_{i=1}^{n} p(\boldsymbol{w}_i|C=1) \end{cases} \quad (4)$$

$p(C=0)$和$p(\boldsymbol{w}_i\mid C=0)$等式的含义与上类同。

文本向量是布尔权重，如果特征词在文本中出现，权重为1，否则为0。设特征数量为n，将文本看作一个事件，通过n重贝努里实验产生。设B_{xi}=1/0表示特征$\boldsymbol{w}_i$在文本$\boldsymbol{d}$中的出现情况，B_{xi}=1/0表示出现/不出现，则有：

$$p(\boldsymbol{d}|C=1)=\prod_{i=1}^{n}(B_{xi}\,p(\boldsymbol{w}_i|C=1)+(1-B_{xi})(1-p(\boldsymbol{w}_i|C=1)) \quad (5)$$

$$p(\boldsymbol{d}|C=0)=\prod_{i=1}^{n}(B_{xi}\,p(\boldsymbol{w}_i|C=0)+(1-B_{xi})(1-p(\boldsymbol{w}_i|C=0)) \quad (6)$$

根据式(1)、(3)、(5)和(6)就可以计算待分类邮件属于垃圾邮件类别的概率。从式(2)可以看出，文本是所有特征类的条件概率之积，如果特征在文本中出现，乘的项是$p(\boldsymbol{w}_i|C=1)$；若不出现，乘的项是$1-p(\boldsymbol{w}_i|C=1)$。$p(\boldsymbol{w}_i|C=1)$的估计也采用文档频次，即：

$$p(\boldsymbol{w}_i|C=1)=\frac{\text{垃圾邮件中特征}\boldsymbol{w}_i\text{出现的文本数量}}{\text{垃圾邮件的文本数量}} \quad (7)$$

对式(7)进行简单的平滑处理得：

$$p(\boldsymbol{w}_i|C=1)=\frac{1+\text{垃圾邮件中特征}\boldsymbol{w}_i\text{出现的文本数量}}{2+\text{垃圾邮件的文本数量}}$$

考虑特征词的出现次数，将每个特征词的出现看作“事件”，文本是这些事件的集合。假设这些事件之间是相互独立的，对式(7)进行简单的平滑处理可得：

$$p(\boldsymbol{w}_i|C=1)=\frac{1+\text{垃圾邮件中特征}\boldsymbol{w}_i\text{出现的文本数量}}{n+\text{垃圾邮件的文本数量}}$$

$$p(\boldsymbol{d}|C=0)=\prod_{i=1}^{n} p(\boldsymbol{w}_i\mid C=0) \quad (8)$$

$$p(\boldsymbol{d}|C=1)=\prod_{i=1}^{n} p(\boldsymbol{w}_i|C=1) \quad (9)$$

式中n为特征向量的数量。根据式(1)、(3)、(7)和(8)就可以计算待分类邮件属于垃圾邮件类别的概率。

2 人工免疫算法

从计算的角度来看，生物免疫系统是一个高度并行、分布、自适应和自组织的完整运行系统，具有很强的学习、识别、记忆和特征提取能力。人工免疫的机理包括：(1) 免疫识别通过淋巴细胞上的抗原识别受体与抗原的结合的强度(亲合度)实现；(2) 免疫记忆是初次遇到抗原时，以最优抗体的形式保留对该抗原的记忆信息；(3) 克隆选择是在抗原的刺激下产生克隆增殖，通过遗传变异，分化为多样性效应细胞和记忆细胞；(4) 多样性包括免疫细胞多样性和抗体多样性；(5) 分布性包含潜在的效率和错误耐受性；(6) 免疫系统是在一个有机体内进行自然选择。本文引用的是基于群体的免疫算法，其步骤包括：(1) 定义抗原；(2) 产生初始解群体；(3) 计算亲和力；(4) 克隆选择；(5) 评估新的抗体群体，若满足终止条件，当前的抗体群体为问题的最优解，否则重新开始计算。在生物免疫系统中集成了垃圾邮件过滤器要求的许多特性：(1) 将垃圾邮件看作抗原，可利用生物免疫系统的模式识别能力对垃圾邮件进行分类设计；(2) 邮件的过滤体现出多样性特征，如同生物时常面对全新的病毒侵袭一样，垃圾邮件的模式和定义处在不断的变化中，邮件过滤器要跟踪和适应这些变化；(3) 生物和人工免疫系统对噪声具有耐受特性，对于邮件过滤器非常重要。

3 基于贝叶斯的人工免疫垃圾邮件过滤算法

贝叶斯分类对未知样本(特别是已有样本的变异样本和利用Bayes模型生成的垃圾邮件样本，如单词“Viagra”的变形词“Via＊gra”、“Via\gra!”、“/\iagra”等)其准确性难以保证。人工免疫系统的自体/非自体识别能力正是识别垃圾邮件良好而又天然的解决方法，把人工免疫系统的各种机理应用于垃圾邮件过滤，其基本思想是：当一封邮件到达时，提取邮件的文本特征向量(VSM)，以此生成“入侵”抗原，抗原首先通过记忆细胞检测器，与记忆细胞进行匹配，若出现匹配，则直接确认该邮件为垃圾邮件；若没有出现匹配，则将抗原与未成熟细胞检测器中的抗体进行匹配，匹配的程度用亲和力表示。若亲和力达到一设定的阈值，则抗原为“非自我”，邮件被判为垃圾邮件；若亲和力小于阈值，则抗原为“自我”，邮件被判为正常邮件。亲和力的计算是最为关键的问题，将直接影响到邮件过滤效果的好坏。依据抗体和抗原的结构，结合贝叶斯(Bayes)网络，可设计系统本身的抗原抗体亲和力计算公式。系统包括训练模块、过滤模块和种族更新模块三个部分，系统的机制图如图1所示。

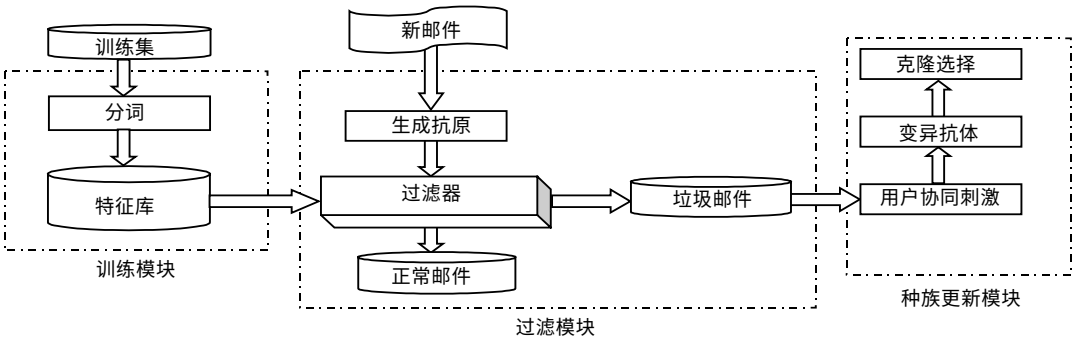

图1　邮件免疫系统的机制框图

4　算法试验

4.1　评价方法

基于贝叶斯的垃圾邮件过滤系统，主要的评价参数[2]有：(1) 召回率：对垃圾邮件检出率；(2) 正确率：反映过滤系统“找对”垃圾邮件的能力；(3)精确率：对邮件的判对率；(4) 错误率：对邮件的判错率。用$L\rightarrow S$表示把正常邮件判断为垃圾邮件，$S\rightarrow L$表示把垃圾邮件判断为正常邮件，$L\rightarrow S$的代价是$S\rightarrow L$的λ倍，则判断一封邮件为垃圾邮件时要满足$\frac{p(C=1|\boldsymbol{d})}{p(C=0|\boldsymbol{d})}>\lambda$，即$p(C=1|\boldsymbol{d})>t$、$t=\frac{\lambda}{1+\lambda}$和$\lambda=\frac{t}{1-t}$。

为了使准确率和错误率能体现损失的不同[4]，把每一封正常邮件看成λ封邮件，即当一个正常邮件被误判为垃圾邮件，就相当于λ个错误；否则就相当于λ个正确。所以，定义精确加权$W_{\text{Accuracy}}=\frac{\lambda n_L\rightarrow L+n_S\rightarrow S}{\lambda N_L+N_{Ss}}$和错误加权$W_{\text{Error}}=\frac{\lambda n_L\rightarrow S+n_S\rightarrow L}{\lambda N_L+N_S}$；而系统的加权精确率和加权错误率为$W^{b}_{\text{Accuracy}}=\frac{\lambda N_L}{\lambda N_L+N_S}$和$W^{b}_{\text{Error}}=\frac{N_S}{\lambda N_L+N_S}$；代价因子(Total Cost Ratio，TCR)$F_{\text{TCR}}=\frac{W^{b}_{\text{Error}}}{W^{b}_{\text{Error}}}=\frac{N_S}{\lambda n_L\rightarrow S+n_S\rightarrow L}$。代价因子表示：系统没有使用过滤器时人工删除所有垃圾邮件的代价与使用过滤器时人工删除被误判为正常邮件的垃圾邮件的代价和重发那些被误判为垃圾邮件的正常邮件的代价之和的比率。代价因子越高越好。

4.2　测试结果

人工免疫系统中有多个因素会影响分类的结果：进化函数、未成熟细胞集合、判断垃圾邮件的阈值、生命周期L的设定、未成熟抗体集合所拥有的抗体数量和记忆细胞集合拥有的细胞数量等。在测试中，演化率(进化函数)的取值为0.97。图2～7分别给出了较为成熟细胞集合抗体数量、记忆细胞集合抗体数量、阈值以及生命周期对该系统测试结果的影响。测试所用的英文语料库[5]包含了1 897条垃圾邮件和4 150条正常邮件。中文语料库[6]包含20 308条垃圾邮件和9 042条正常邮件。所有实验都是把语料库中邮件分为10份，其中9份作为训练集，1份作为测试集，如此交叉做10次实验，最后取10次实验的平均值作为最后的实验数据。

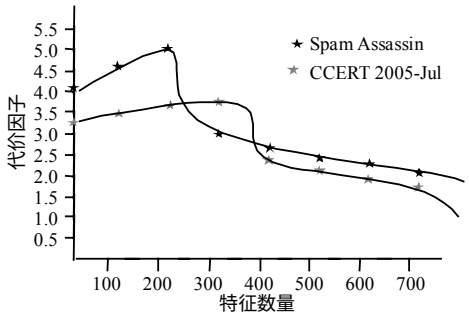

图2　λ=1未成熟细胞集合抗体数量对代价因子

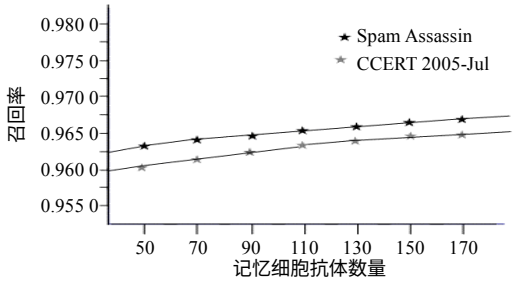

图3　记忆细胞集合抗体数量对系统召回率的影响

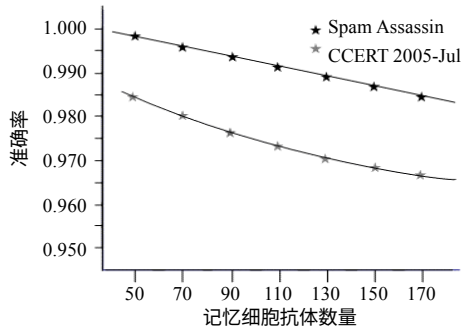

图4 记忆细胞集合抗体数量对系统准确率的影响

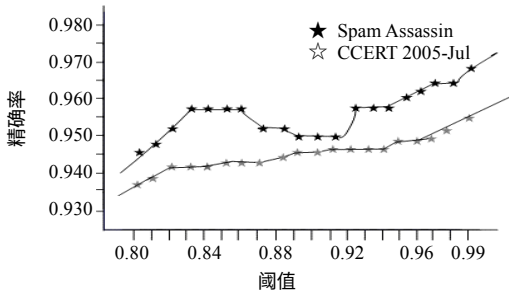

图5 阈值对系统精确率的影响

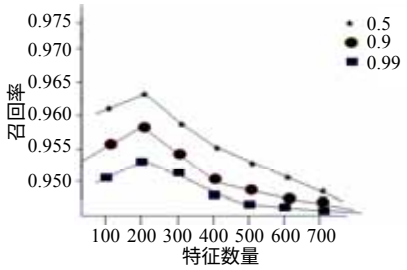

图6 阈值对召回率的影响

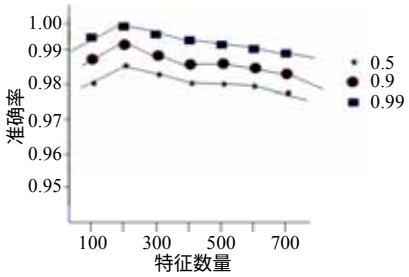

图7 阈值对准确率的影响

4.3 算法比较

在算法比较测试中使用Spam Assassin邮件样本集的4 000封邮件(其中垃圾邮件2 000封)对两个分类器进行分类训练，得到初始的检测器。然后利用Spam Assassin邮件样本集的2 000封邮件(其中垃圾邮件1 500封，包含人为对类似“Viagra”作变形词“\/iagra”等变异的垃圾邮件)进行连续识别。系统运行参数设置为：未成熟细胞集合抗体数量200、记忆细胞集合抗体数量110、阈值0.99。图8所示为算法比较结果，可从识别的起始、中期和后期三个阶段观察比较结果。在识别的后期，贝叶斯和人工免疫混合算法优于贝叶斯算法，结果表明基于贝叶斯和人工免疫混合算法的垃圾邮件过滤系统不但一开始就具有很好的分类能力，而且具有免疫系统的自学习、自适应和鲁棒性等特点。

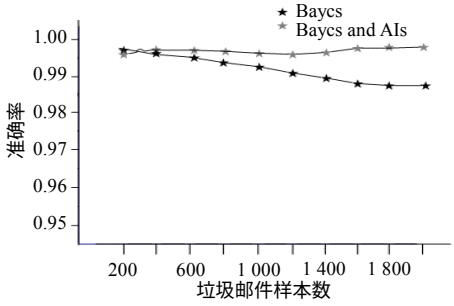

图8 两个算法的比较

5 总 结

本文提出了基于贝叶斯和人工免疫的混合垃圾邮件过滤算法，设计了邮件过滤系统，并通过垃圾邮件过滤仿真测试，比较了未成熟细胞集合抗体数量、记忆细胞集合细胞数量及阈值对测试结果的影响。此外，在相同训练集的基础上比较了基于贝叶斯和人工免疫的混合垃圾邮件过滤与基于Bayes算法的垃圾邮件过滤的准确率，体现了该系统的自学习、自适应和鲁棒性等特点。

参 考 文 献

[1] 中国互联网协会反垃圾邮件中心. 年度反垃圾邮件报告[DB/OL]. http://www.anti-spam.cn/, 2007-04-05.

[2] ANDROUTSOPOULOS I, PALIOURAS G, KARKALETSIS V, et al. Learning to filter spam E-mail: a comparison of a naive Bayesian and a memory-based approach[C]// In Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. Athens: IEEE Press, 2000.

[3] ANDROUTSOPOULOS I.G, PELIOUMS E M. Leaming to filter unsolicited commercial E-mail[R]. Technical report 2004.

[4] ANDROUTSOPOULOS I, KOUTSIAS J, CHAN-Drinos KV, et al. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with encrypted personal E-mail messages[C]// SIGIR Forum(ACM Special Interest Group Information Retrieval), Association for Computing Machinery. New York: [s. n.], 2000.

[5] DEERSOFT. SpamAssassin Project[DB/OL]. http://www.spamassassin.org, 2007-05-10.

[6] 中国教育和科研计算机网紧急相应组. CCERT 2005-Jul数据集[R]. 2005.

编 辑 熊思亮