文章编号:1001 - 9081(2006)03 - 0558 - 04

基于有监督 Bayesian 网络的垃圾邮件过滤

刘 震,周明天
(电子科技大学 计算机科学与工程学院,四川 成都 610054)
(liuzhen2004@126.com)

摘 要:对影响邮件特性的邮件报文格式作了仔细的分析并对垃圾邮件的特征进行了分类归纳,在此基础上构建了一个有监督的 Bayesian 邮件分类网络。通过对该网络作 Bayesian 参数估计,实现了判定邮件类别的不确定推理。对不同邮件测试集的在线学习试验结果表明,有监督 Bayesian 邮件分类网络能够有效地实现垃圾邮件的相对完备特征学习,改善邮件过滤的准确率。

关键词:垃圾邮件;Bayesian 网络;邮件过滤;参数估计
中图分类号: TP393.098; TP181 **文献标识码**:A

Spam filtering algorithm based on supervised Bayesian parameter estimation

LIU Zhen, ZHOU Ming-tian
(*College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan* 610054, *China*)

Abstract: To improve the reliability and completeness of spam filtering, the E-mail message format was carefully analyzed, and the spam characteristics were generalized and classified. Based on these analysis, a supervised Bayesian network for E-mail classifer was constructed. Parameter estimation on this network realized an uncertain inference to identify E-mail's sort. On-line learning for different E-mail testing sets shows that such a classifying network can ensure the classification and filtering efficiently. It practically provides a viable solution by building a supervised Bayesian classifying network to execute relatively complete characteristics learning and improve the accuracy of E-mail filtering.

Key words: spam; Bayesian network; E-mail filtering; parameter estimation

0 引言

随着 Internet 的普及,电子邮件由于其方便、快捷、低成本的特点逐渐成为现代社会主要的网络通信方式之一。然而近年来,垃圾邮件的日趋泛滥给电子邮件系统和用户带来了很多的危害甚至损失。垃圾邮件的传播不仅浪费网络资源,造成了邮件服务器负荷增大,而且也成为有害信息和病毒传播的重要途径[1]。为了保护邮件系统的正常运行和邮箱用户的利益,必须使邮件系统具有反垃圾邮件的能力。目前常见的反垃圾邮件技术主要有基于关键字的过滤、基于黑白名单的过滤、基于规则的过滤[2~4]等。然而,这些技术往往存在误报率和漏报率偏高、适用范围窄、不能训练学习等诸多问题。

由于判定垃圾邮件的过程是一个不确定推理过程,本文引入了 Baysian 参数估计理论,通过构建有监督 Bayesian 网络作不确定推理来实现对垃圾邮件的分类和过滤。Bayesian 网络是无环的 DAG 图,节点之间存在着因果的逻辑联系。通过学习与推理,Bayesian 网络可以利用先验知识对一些现象进行识别、分类和预测。作为一种基于概率的不确定性推理方法,贝叶斯网络在处理不确定信息的智能化系统中得到了重要的应用,已成功地应用于医疗诊断、统计决策、专家系统等领域[6]。Bayesian 参数估计作为基于统计学的不确定推理理论的一个重要研究方向,有着坚实完备的数学基础。将 Bayesian 参数估计引入到网络学习中,可以充分利用节点的先验知识作后验估计;因为节点之间逻辑上的因果关系,能够提高先验的可信度。

1 Bayesian 参数估计理论

Bayesian 参数估计也称为 Bayesian 参数学习。Bayesian 参数估计的思想是通过前 m 次的先验统计概率分布,估计第 $m+1$ 次事件发生的概率。它通过不断地概率学习,从而不断地适应和逼近变化的概率分布。通常利用多项式分布描述邮件的特性分布[5]。已知随机事件 X 在前 m 次的概率分布,要估计下一次 $X[m+1]$ 的概率,可以计算:

$$p(x[m+1] \mid x[1], x[2], \cdots, x[m]) = \int p(x[m+1] \mid \theta, x[1], x[2], \cdots, x[m]) p(\theta \mid x[1], x[2], \cdots, x[m]) \mathrm{d}\theta = \int p(x[m+1] \mid \theta) p(\theta \mid x[1], x[2], \cdots, x[m]) \mathrm{d}\theta \quad (1)$$

由贝叶斯公式,有:

$$p(\theta \mid x[1], x[2], \cdots, x[m]) = \frac{p(x[1], x[2], \cdots, x[m] \mid \theta) p(\theta)}{p(x[1], x[2], \cdots, x[m])} \quad (2)$$

其中 $p(x[1], x[2], \cdots, x[m] \mid \theta)$ 是似然函数,$p(\theta)$ 是先验分布。

因为 X 满足多项式分布,那么有:

$$p(\Theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \quad (3)$$

其中 $\alpha_1, \alpha_2, \cdots, \alpha_K$ 是参数 $\theta_k (k = 1,2,\cdots,K)$ 对应的超

收稿日期:2005 - 09 - 20 **修订日期**:2005 - 12 - 28 **基金项目**:国家 863 计划项目(863-104-03-01)
作者简介:刘震(1976 -),男,吉林吉林人,博士研究生,主要研究方向:网络计算、网络安全技术; 周明天(1939 -),男,广西容县人,教授,博士生导师,主要研究方向:网络计算、网络安全技术.

参数。

由于$p(\Theta)$满足 Dirichlet 共轭先验条件,那么它的后验分布对应的超参数为α_1+N_1, α_2+N_2, …, α_K+N_K。容易计算:

$$p(\Theta \mid D) \propto p(\Theta)p(D \mid \Theta) \propto p(\Theta)p(D \mid \Theta) \propto \prod_{k=1}^{K}\theta_k^{\alpha_k-1}\prod_{k=1}^{K}\theta_k^{N_k}=\prod_{k=1}^{K}\theta_k^{\alpha_k+N_k-1} \tag{4}$$

根据式(1),X 的后验 Bayesian 参数估计值为:

$$p(x[m+1]=k \mid D)=\int\theta_k\, p(\Theta \mid D)\mathrm{d}\Theta=\frac{\alpha_k+N_k}{\sum_l(\alpha_l+N_l)} \tag{5}$$

2　完备 Bayesian 邮件分类网络

下面通过分析邮件的报文格式来构建 Bayesian 网络。根据 RFC2822 定义的 Internet 邮件报文格式,一封邮件由报头域和正文组成。其中报头必须存在,而正文是可选的。分析邮件的性质时,分析报头域是非常必要的,而这也是部分反垃圾系统片面强调内容过滤而忽略了的地方。报头由一系列由特殊语法构成的行组成。正文则仅仅由字符串组成。正文和报头由一空行分隔开。

报头域是由域名和域体组成,二者以一个冒号分开。域名必须是可打印的 US-ASCII 字符。域体可以是任意的 US-ASCII 字符。下面分析几个重要的报头域:

1)起始日期域

Orig-date = "Date: "date-time CRLF

起始日期代表的是邮件创建者完成邮件并且将邮件送交至递送系统的时间。这个域可以成为 Bayesian 网络中的一个节点的理由是因为在某些敏感日期,如节假日,病毒爆发日,垃圾邮件容易泛滥,系统应该对这些日期提高预警。

2)发件人地址域

from = "From: "mailbox-list CRLF

sender = "Sender: "mailbox CRLF

reply-to = "Reply-To: "address-list CRLF

发件人地址域包括 from 域,sender 域和 reply-to 域,它们指明了邮件的来源。Sender 域显然应该成为 Bayesian 网络的一个节点,对于发垃圾邮件发送者,他们的邮件地址是最直接的一个判据。基于黑名单的系统就是简单地拒绝来之黑名单列表中的地址发来的邮件实现过滤。但这可能导致误判,因为,垃圾邮件发送者不一定总是用这个邮箱地址发垃圾邮件,他们也可能会用这个邮箱发送正常的非垃圾邮件。

3)目的地址域

to = "To: "address-list CRLF

cc = "Cc: "address-list CRLF

bcc = "Bcc: "(address-list/[CFWS]) CRLF

目的地址域由三个可选的域构成:to 域,cc 域和 bcc 域。它们域名分别是“To”,“Cc”和“Bcc”,域体指明了邮件的收件人。通过 cc 域和 bcc 域可以作为判断垃圾邮件的一个依据。这是基于垃圾邮件发送者在发送垃圾邮件时往往采用群发方式的一个重要观察。

除了以上的三个域,邮件格式中还有邮件标识域、邮件信息域、重发域、回溯域、扩展域等。经分析,我们认为这些域不是判断邮件性质的必要条件,所以没有把它们纳入后面的 Bayesian 网络。

对邮件体的分析目前仍然集中在某些关键词的概率估计上,这完全不同于简单的关键词匹配。因为很多垃圾邮件中出现的词汇,也可能会出现在正常邮件中。所以,利用关键词判断邮件的性质也是不确定的推理过程。

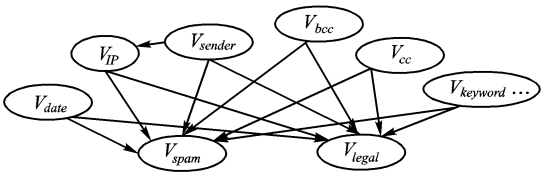

图 1　基于垃圾邮件特征的完备 Bayesian 网络

图 1 是根据垃圾邮件的基本特征构建的一个 Bayesian 网络。IP 可以通过域名作反向 DNS 查询来得到,这样可以有效地防止域名欺骗。由于需要通过 sender 的域名判定其 IP 是否是垃圾邮件发送者的 IP 的概率,所以存在一根网络连线从 sender 节点指向 IP 节点。关键词节点中所加省略号,表示关键词不唯一;图中是一种省略的表示法。Bayesian 网络都是 Causal 图,它描述了节点间的因果关系。图 1 建立的网络涵盖了导致邮件成为垃圾邮件的主要因素。通过概率关系来描述这个网络可以定量地研究邮件是垃圾邮件的可能性。因为邮件是否是垃圾邮件有共性的特征也有个性的特征,所以,对垃圾邮件的判断无论对人和机器而言都是一种不确定推理,Bayesian 网络正好可以满足对不确定推理的要求。

如果当前该网络已有 m 次的证据输入,输入分布矩阵表示为:

$$\boldsymbol{D}=\begin{bmatrix} V_{spam}[1] & V_{legal}[1] & V_{date}[1] & V_{IP}[1] & V_{sender}[1] & V_{bcc}[1] & V_{cc}[1] & V_{keyowrd_1}[1] & \cdots & V_{keyoword_n}[1] \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ V_{spam}[m] & V_{legal}[m] & V_{date}[m] & V_{IP}[m] & V_{sender}[m] & V_{bcc}[m] & V_{cc}[m] & V_{keyowrd_1}[m] & \cdots & V_{keyoword_n}[m] \end{bmatrix}$$

那么该网络的似然函数分布为:

$$\begin{aligned} &L(\Theta:D) \\ &=\prod_m p(V_{spam}, V_{legal}, s_{date}, s_{IP}, s_{sender}, s_{bcc}, s_{cc}, s_{keyword}:\Theta) \\ &=\prod_m \Theta_{spam\mid date}\Theta_{spam\mid IP}\Theta_{spam\mid sender}\Theta_{spam\mid bcc} \\ &\Theta_{spam\mid cc}\Theta_{spam\mid keyword_1}\cdots\Theta_{spam\mid keyword_m}\Theta_{legal\mid date}\Theta_{legal\mid IP}\Theta_{legal\mid sender} \\ &\Theta_{legal\mid bcc}\Theta_{legal\mid cc}\Theta_{legal\mid keyword_1}\cdots\Theta_{legal\mid keyword_m}\Theta_{IP\mid sender}\Theta_{date}\Theta_{IP}\Theta_{sender} \\ &\Theta_{bcc}\Theta_{cc}\Theta_{keyword_1}\cdots\Theta_{keyword_m} \end{aligned} \tag{6}$$

对于节点 V_{spam} ,它的似然函数满足:

$$\begin{aligned} L_{spam}(\Theta_{spam}:D) &=\prod_m p(V_{spam}[m] \mid pa_{spam}[m]:\Theta_{spam}) \\ &=\prod_{pa_{spam}}\prod_{pa_{spam}[m]=pa_{spam}} p\left(V_{spam} \mid pa_{spam}:\Theta_{spam}\right) \\ &=\prod_{pa_{spam}}\prod_{V_{spam}} p\left(X_{spam} \mid pa_{spam}:\Theta_{spam}\right)^{N(V_{spam},\, pa_{spam})} \\ &=\prod_{pa_{spam}}\prod_{V_{spam}} \theta_{V_{spam}\mid pa_{spam}}{}^{N(V_{spam},\, pa_{spam})} \end{aligned} \tag{7}$$

(7)式中 pa_{spam} 代表 V_{spam} 的父属性节点。将式(7)带入式(1),可求得:

$$\theta_{V_{spam}\mid pa_{spam}}=\frac{\alpha(V_{spam} \mid pa_{spam})+N(V_{spam} \mid pa_{spam})}{\alpha(pa_{spam})+N(pa_{spam})} \tag{8}$$

同理,可以得到:

$$\theta_{V_{legal}\mid pa_{legal}}=\frac{\alpha(V_{legal} \mid pa_{legal})+N(V_{legal} \mid pa_{legal})}{\alpha(pa_{legal})+N(pa_{legal})} \tag{9}$$

式(8)中的$\alpha(pa_{spam})$和$\alpha(V_{spam} \mid pa_{spam})$是网络节点的先验超参数,它们代表了网络节点的先验水平。不同的先验水平

(EOP) 对参数估计分布也有不同的影响。图 2 是不同先验力度条件下由式(8) 计算得到的参数估计分布;图 3 是不同先验比率条件下由式(9) 得到的参数估计分布。其中 $M = \alpha(pa_{spam})$ 代表先验力度,$K = \alpha(V_{spam} \mid pa_{spam})/\alpha(pa_{spam})$ 代表先验比率。从图中可以看出选择适当的先验水平,对参数估计值收敛的快慢有着直接的影响。本文采用文献[5]的 Gibbs 采样方法可以得到近似优化的先验力度。

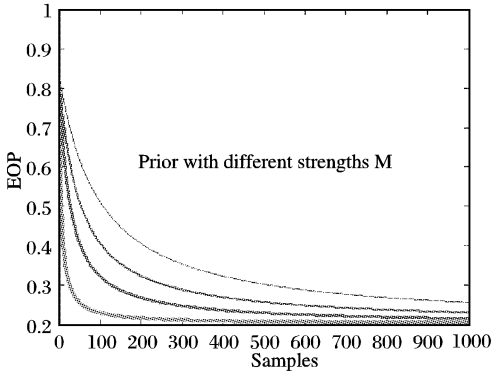

图 2 不同先验力度(M)条件下的参数估计分布

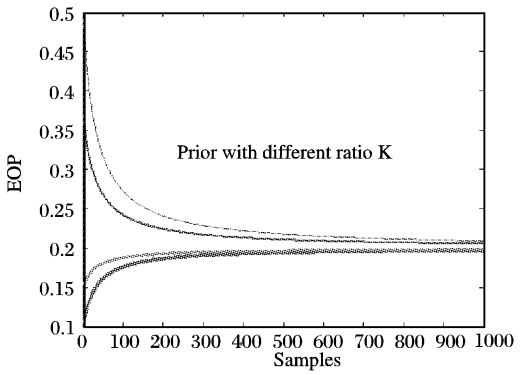

图 3 不同先验比率(K)条件下的参数估计分布

3 训练邮件过滤器

针对垃圾邮件的 Bayesian 参数估计首先要初始化先验的概率分布参数。本文利用了部分已知的邮件样本集初始化两个重要参数 $\theta_{spam\mid attribute_i}$ 和 $\theta_{legal\mid attribute_i}$。其中 $\theta_{spam\mid pa_{spam}}$ 是在满足某个邮件的父属性条件下,邮件是垃圾邮件的概率;$\theta_{legal\mid pa_{legal}}$ 是在满足某个邮件的父属性条件下,邮件是正常邮件的概率。网络中 X 表示邮件中出现某一属性的事件;Y 表示邮件中存在该属性的条件下,该邮件是垃圾或者正常邮件的事件。对 $k+1$ 次到达的新邮件做 $\theta_{pa_{spam}}$ 和 $\theta_{spam\mid pa_{spam}}$ 的 Bayesian 参数估计要依赖前 k 次的输入邮件分布。

将所有过滤属性记为 $\{attribute_1, attribute_2, \cdots, attribute_n\}$,属性之间的关系满足 Bayesian 网络中的节点关系。引入一批邮件训练样本,已知其中垃圾邮件 m 封,正常邮件 n 封。

1) 初始化

统计属性和垃圾邮件在初始邮件样本集中的样本分布情况。首先根据事件发生的频率初始化 $p\{x = attribute_i\}$, $p\{y = attribute_i \mid x = spam\}$, $p\{y = spam\}$, 其中 $i = 1, 2, \cdots, n$。

那么对于每个关键字 $x = attribute_i$,由贝叶斯公式容易得到:

$$p\{y = spam \mid x = attribute_i\} = \frac{p(x = attribute_i \mid y = spam)p(y = spam)}{p(x = attribute_i)} \tag{9}$$

令 $\theta_{spam\mid attribute_i} = p\{y = spam \mid x = attribute_i\}$。同理,由 $p = \{y = legal \mid x = attribute_i\}$,可以初始化计算 $p = \{y = legal \mid x = attribute_i\}$。

2) 训练学习

由式(8),对每个属性有:

$$\theta_{spam\mid attribute_i} = \frac{\alpha(y = spam \mid x = attribute_i) + N(y = spam \mid x = attribute_i)}{\alpha(x = attribute_i) + N(x = attribute_i)} \tag{10}$$

同理有:

$$\theta_{legal\mid attribute_i} = \frac{\alpha(y = legal \mid x = attribute_i) + N(y = legal \mid x = attribute_i)}{\alpha(x = attribute_i) + N(x = attribute_i)} \tag{11}$$

其中,$\alpha(y = spam \mid x = attribute_i) = M\theta_{spam\mid attribute_i}$,$\alpha(y = legal \mid x = attribute_i) = M\theta_{legal\mid attribute_i}$;$\alpha(x = attribute_i) = M\theta_{attribute_i}$,$M = m + n$。

当有新的邮件到达时,首先分析邮件中的属性。一般情况下可能会有多个过滤属性出现在该邮件中。此时,要针对每一个出现在邮件中的 $attribute_i$,更新相应的 $\theta_{attribute_i}$。如果满足条件 $\sum_{attribute_i} \theta_{spam\mid attribute_i}\theta_{attribute_i} > \sum_{attribute_i} \theta_{legal\mid attribute_i}\theta_{attribute_i}$,那么判定该邮件为垃圾邮件;如果满足条件 $\sum_{attribute_i} \theta_{spam\mid attribute_i}\theta_{attribute_i} < \sum_{attribute_i} \theta_{legal\mid attribute_i}\theta_{attribute_i}$,则判定该邮件为合法邮件;根据以上判断结果更新 θ_{spam} 或者 θ_{legal}。如果出现 $\sum_{attribute_i} \theta_{spam\mid attribute_i}\theta_{attribute_i} = \sum_{attribute_i} \theta_{legal\mid attribute_i}\theta_{attribute_i}$ 的情况,则将该邮件放入标识邮件类型无法识别的缓存队列,如果下一次新邮件到达以后并作出了正常的判断,则可以根据更新的 θ_{spam} 或者 θ_{legal} 从缓存队列中取出邮件重复以上判断过程。每次有新的邮件到达时,反复以上步骤,就可以实现图 1 所示有监督 Bayesian 网络的在线学习和过滤。

4 邮件过滤器的性能分析

在分析邮件过滤器的性能之前,首先需要引入误报和漏报的概念[7]。误报是指误将合法邮件判断为垃圾邮件($L \rightarrow S$) 的情况,漏报则恰好相反,是将垃圾邮件误判为合法邮件($S \rightarrow L$) 的情况。误报率的定义式为:$R_{err} = \frac{n_{L\rightarrow S}}{N_S + n_{L\rightarrow S}}$,漏报率的定义式为:$R_{miss} = \frac{n_{S\rightarrow L}}{N_L + n_{S\rightarrow L}}$。其中 N_L 和 N_S 分别表示合法邮件和非法邮件的数量,$n_{L\rightarrow S}$ 和 $n_{S\rightarrow L}$ 分别表示误报邮件和漏报邮件的数量。整体评价一个分类器的好坏,需要综合看它在漏报和误报两方面的性能表现。整体准确率定义式为 $Acc = \frac{n_{L\rightarrow L} + n_{S\rightarrow S}}{N_L + N_S}$,整体错误率定义式为 $Err = \frac{n_{L\rightarrow S} + n_{S\rightarrow L}}{N_L + N_S}$。

以上定义反映的是误报和漏报给用户带来的后果相同的情况,但实际情况通常并非如此。因为,用户一般能够容忍把少数几封垃圾邮件误判为正常邮件,但用户很难容忍一封正常邮件误判为垃圾邮件而被过滤掉,尤其对用户非常重要的邮件。针对这一实际情况,就必须对上面的定义作修正,本文解决的方法是引入权值。权重准确率和权重错误率定义式为:

$$WAcc = \frac{\lambda n_{L\to L} + n_{S\to S}}{\lambda N_L + N_S}$$

$$WErr = \frac{\lambda n_{L\to S} + n_{S\to L}}{\lambda N_L + N_S} \quad (12)$$

以上定义式表示将 1 封正常邮件误判为垃圾邮件等价于将 λ 封垃圾邮件误判为正常邮件。换言之,如果误报和漏报的邮件一样多,那么误报对邮件过滤系统优劣评价的影响更负面。

假设系统没有邮件过滤器;那么正常邮件都能正确地识别;而垃圾邮件全部漏判为正常邮件。则有:

$$WAcc_S = \frac{\lambda N_{L\to L}}{\lambda N_L + N_S}$$

$$WErr_S = \frac{N_S}{\lambda N_L + N_S} \quad (13)$$

本文同时引入比值 *TCR* 用于描述邮件过滤器的过滤性能的另一指标, *TCR* 越大,过滤器的过滤性能越好。

$$TCR = \frac{WErr_S}{WErr} = \frac{N_S}{\lambda n_{L\to S} + n_{S\to L}} \quad (14)$$

下面以 4 个邮件样本集为例,进行邮件过滤器的性能实验。其中 EN, PU1, Ling-Spam 集是网络上可以下载的公共测试集[8],而 CH 集是我们自己构建的中文邮件测试集。图 4,图 5 分别是 $\lambda = 1$ 和 $\lambda = 9$ 条件下的权重准确率分布曲线。

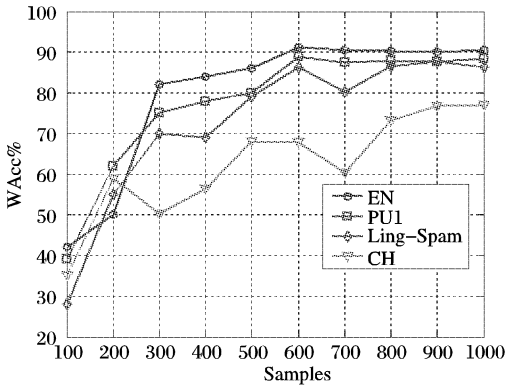

图 4　$\lambda = 1$ 时过滤不同邮件集的 *WAcc%* 对比

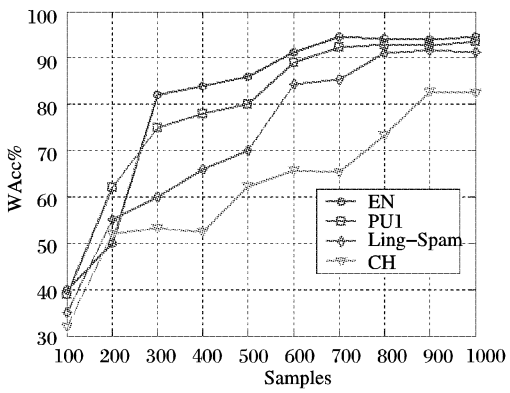

图 5　$\lambda = 9$ 时过滤不同邮件集的 WAcc% 对比

从图 4,图 5 可以看出,过滤器对英文邮件集的过滤性能比较好,无论 $\lambda = 1$ 或者 $\lambda = 9$, 当邮件样本超过 600 以后,过滤器的权重准确率都已接近 90%,明显优于对中文邮件的过滤性能。中文复杂的构词法和太多的转义是过滤器的性能较低的原因,而在我们构建的 Bayesian 网络中并没有设置专门与此相对应的节点逻辑,这个问题我们将在以后作更深入的研究。由于不同的权重对准确率的值也有一定的影响,式(12)的权重强调了正确判断对邮件过滤器性能的正面影响,所以权重 $\lambda = 9$ 的曲线相对于 $\lambda = 1$ 的曲线更平滑。适当的权重便于研究者更好地分析过滤器的整体特点,而不会因为曲线局部的起伏变化作出对过滤器整体性能的片面判断。表 1 的 *TCR* 值对比可以得出和前面权重准确率比较图相同的结论,两种测试指标的结果均能客观地反映邮件过滤器的性能。

表 1　不同测试集上的 *TCR* 值对比

Instance	$\lambda = 1$				$\lambda = 9$			
	TCR_{EN}	TCR_{PU1}	$TCR_{Ling\text{-}Spam}$	TCR_{CH}	TCR_{EN}	TCR_{PU1}	$TCR_{Ling\text{-}Spam}$	TCR_{CH}
100	2.97	2.90	2.86	2.73	0.99	0.94	0.91	0.95
200	3.59	3.51	3.45	3.35	1.19	1.21	0.99	1.01
300	4.59	4.47	4.58	4.02	1.38	1.36	1.30	1.12
400	6.31	6.23	6.02	5.86	1.41	1.35	1.32	1.32
500	7.11	6.89	6.78	6.12	1.52	1.49	1.39	1.35
600	7.33	6.99	6.89	6.23	1.52	1.48	1.43	1.29
700	7.51	7.05	6.97	5.98	1.56	1.51	1.39	1.38
800	7.97	7.41	7.12	6.52	1.59	1.55	1.47	1.39
900	8.02	7.77	7.58	6.55	1.60	1.57	1.46	1.38
1000	8.01	7.75	7.56	6.60	1.61	1.56	1.47	1.42

5　结语

发现垃圾邮件的过程是不确定推理的过程。提高不确定推理的可靠性,需要提取尽可能完备的样本特征属性。通过吸取和综合传统邮件过滤技术的思想,提取邮件特征,构建相对完备的有监督 Bayesian 网络可减少邮件漏报和误报,提高垃圾邮件过滤系统的广普性。过滤实验结果表明基于本文提出的有监督 Bayesian 网络的垃圾邮件过滤器,能够较好地实现英文邮件的分类和过滤,但对中文邮件的过滤效果还较差,说明网络的节点逻辑构建还有需要完善的地方,未来需要进一步加强对中文邮件特征的分析,调整和增加相应的 Bayesian 网络节点,从而更好地发挥 Bayesian 网络的概率学习能力。

参考文献:

[1] 陶卓彬, 邓元庆. 反垃圾邮件技术[J]. 信息网络安全, 2003, 9(5): 41 - 43.

[2] 刘震, 余堃, 周明天. 基于多级属性集的垃圾邮件过滤技术[J]. 计算机应用研究, 2005, 22(7): 122 - 126.

[3] Spam Is A Thousand Times More Horrible Than You Can Imagine [EB/OL]. http://www.internetweek.com/story/INW20021219S0003. (2002a), 2005.

[4] Study: e-mail viruses up[EB/OL]. http://www.internetweek.com/story/INW20021109S0002. (2002b), 2005.

[5] JENSEN FV. An Introduction to Bayesian Networks[M], London: UCL Press, 1996. 196 - 202.

[6] LAURITZEN SL, SPIEGELHALTER DJ. Local computations with probabilities on graphical structures and their application to expert systems[J]. Journal of the Royal Statistical Society, Series B (Methodological), 1988, 50(2): 157 - 224.

[7] O'BRIEN C, VOGEL C. Spam filters: bayes vs. chi-squared; letters vs. words[A]. Proceeding/Series-Proceeding-Section-Article [C], 2003. 291 - 296.

[8] ANDROUTSOPOULOS I, KOUTSIAS J, CHANDRINOS V, *et al.* An Evaluation of Naive Bayesian Anti-Spam Filtering[A]. Workshop on Machine Learning in the New Information Age[C], 2000. 578 - 584.