基于核方法的贝叶斯邮件分类网络研究

刘 震，周明天

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】提出一种包含核函数的Bayesian参数估计方法，提高了Bayesian参数估计的实用性。结合邮件内容和报文格式两个方面分析和提取邮件的重要特征，建立了对应的Bayesian邮件分类网络。将包含核函数的Bayesian参数估计方法应用到邮件分类网络，在对不同邮件测试集的在线学习试验结果证明，这种新的分类模型能够有效地实现垃圾邮件的分类过滤。

关 键 词 Bayesian网络; 高斯核; 参数估计; 垃圾邮件;
中图分类号 TP393 文献标识码 A

Research on Bayesian Classification Network for Spam Based on Kernel Method

LIU Zhen，ZHOU Ming-tian

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

Abstract A kernel function based Bayesian parameter estimation approach is proposed in this paper which is able to make the algorithm more applicable. Combined with the both sides of email content and format, a Bayesian network for spam classification is well constructed. The testing results by on-line learning for different email testing sets prove that the new model can ensure the classification and filtering efficiently by applying the kernel function based Bayesian parameter estimation approach into the classification network.

Key words Bayesian network; Gaussian kernel; parameter estimation; spam

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[1]。将Bayesian参数估计引入到贝叶斯网络学习中，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。但由于概率密度函数通常是未知的，限制了经典Bayesian参数估计方法的应用。本文通过引入核方法，实现了对概率密度函数的近似估计，从而提高了Bayesian参数估计方法的实用性。在文献[2]工作的基础上，本文根据对垃圾邮件所作的特征属性分析，构建了有监督Bayesian网络；提出的垃圾邮件分类过滤算法充分利用了网络所建立的节点关系来实现不确定特征学习，采用统计推理的方法确保了对垃圾邮件和正常邮件准确和有效的分类识别。

1 Bayesian参数估计理论

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。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，可计算X的后验Bayesian参数估计概率：

$$p(x[m+1]=k \mid D)=\int \theta\, p(\Theta \mid D)\mathrm{d}\Theta \qquad (1)$$

然而，式(1)求解的前提需要知道概率密度函数 $p(\Theta \mid D)$ 的形式，如果预先无法得到精确的概率分布函数，则不能按照式(1)作概率参数学习。所以在实际的基于统计学习的模式分类问题中，需要研究如何得到概率密度函数。先假设从概率密度函数 $f_X(x)$ 提取随机样本 $x_1, x_2, \cdots, x_N$ ，一种自然的局部估计近似具有如下形式：

$$f(x_0)=\frac{\#x \in N(x_0)}{N\lambda} \qquad (2)$$

式中 $N(x_0)$ 是 x_0 周围宽度为 λ 的较小度量邻域。KNN和最小二乘回归分析是传统的研究近似概率密度函数的方法，但这些方法得到的估计是起伏的[1]。所以本文采用光滑的Parzen估计：

$$\hat{f}(x)=\frac{1}{N\lambda}\sum_{i=1}^{N}K_\lambda(x_0,x_i) \qquad (3)$$

因为式(3)使用随 x_0 的距离递减的权处理邻近

收稿日期：2005－03－07
作者简介：刘 震(1976－)，男，博士生，主要从事智能安全、不确定推理、人工智能等方面的研究.

x_0 的观测。所以本文选择具有类似特征的高斯核 $K_\lambda(x_0,x)=\phi(|x-x_0|/\lambda)$。设 ϕ_λ 表示具有均值0和标准差 λ 的高斯密度，则概率密度函数为：

$$\hat{f}(x)=\frac{1}{N\lambda}\sum_{i=1}^{N}\phi_\lambda(x-x_i)=(\hat{F}\phi_\lambda)(x) \tag{4}$$

利用式(4)，可以直接使用贝叶斯定理进行分类。针对 J 类问题，分别在类别上拟合非参数密度估计 $\hat{f}_j(x)$，$j=1,2,\cdots,J$，以及类的先验 $\hat{\pi}_j$ 的估计(通常是样本的比例)，那么边界判定式为：

$$\hat{\Pr}(G=j\mid X=x_0)=\frac{\hat{\pi}_j\hat{f}_j(x_0)}{\sum_{k=1}^{J}\hat{\pi}_j\hat{f}_k(x_0)} \tag{5}$$

2 有监督Bayesian邮件分类网络

为了构建有监督的Bayesian邮件分类网络，需要分析邮件的报文格式。根据RFC2822定义的Internet邮件报文格式(Internet Message Format)，一封邮件由报头域(Header Fields)和正文(Body)组成。其中报头必须存在，而正文是可选的。报头是一系列由特殊语法构成的文本行组成，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(Field Name)和域体(Field Body)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。下面分析三个重要的报头域：

(1) 起始日期域(The Origination Date Field)：

Orig-date=”Date:”date-time CRLF

这个域可以成为Bayesian网络中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(Originator Fields)：

from=”From:”mailbox-listCRLF，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，对于垃圾邮件发送者，他们的邮件地址是最直接的一个判据。

(3) 目的地址域(Destination Address Fields)：

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Bc域和Bcc域可以作为判断垃圾邮件的一个依据。

经分析认为邮件格式中的其他域不是判断邮件性质的必要条件，所以本文没有把它们纳入Bayesian网络的结构中。

对邮件体的分析目前仍然集中在某些关键词出现的概率估计上，这是基于内容的过滤技术常常关注的分类特征。本文研究关键字并不是采用简单的关键词匹配技术。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，所以应该用概率的方法对关键字做必要的取舍。

图1所示为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，表示网络中关键词不唯一，图1只是一种省略的表示法。由于Bayesian网络都是Causal图，箭头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该网络可以定量地研究邮件是垃圾邮件的可能性。

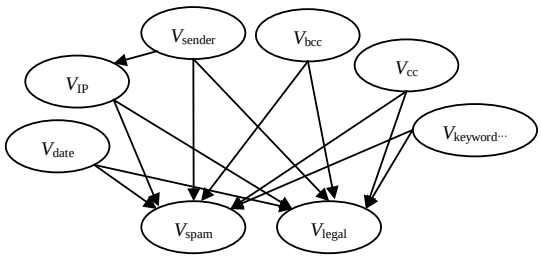
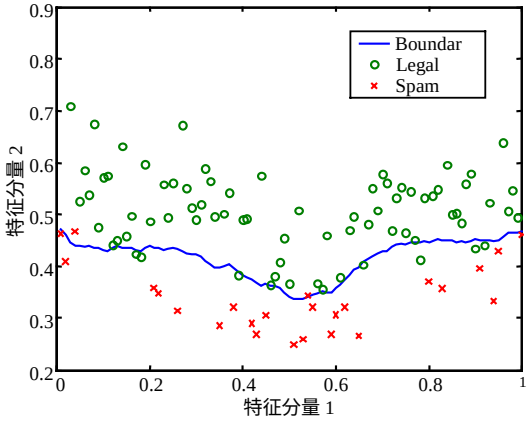


图1 基于垃圾邮件特征的完备Bayesian网络

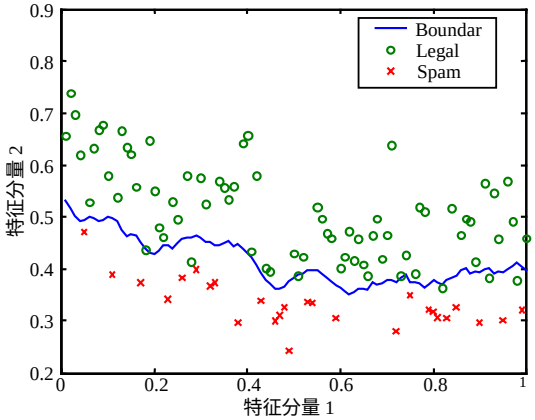
3 训练邮件过滤器

本文以四个邮件样本集为例，进行邮件分类器的测试实验。其中EN、PU1、Ling-Spam集是网络上可以下载的公共测试集[2]，而CH集是本文构建的中文邮件测试集。设输入向量定义为：$\boldsymbol{X}=(x_{\text{date}},x_{\text{IP}},x_{\text{sender}},x_{\text{IP|sender}},x_{\text{bcc}},x_{\text{cc}},x_{\text{keyword}_1},x_{\text{keyword}_2},\cdots,x_{\text{keyword}_n})$，以第2节构建的Bayesian分类网络所描述的分类特征关系为分类依据，按照第1节引入的核函数方法对初始邮件样本集做近似的概率密度函数估计，最终可以得到Spam类和Legal类邮件的判定边界，即得到集合 $\{x\mid p(G=S_{\text{spam}}\mid X=x)=1/2\}$。图2分别展示了在四个样本集上的判定边界。当有新的待分类邮件到达时，首先要根据Bayesian分类网络对邮件的输入特征向量作特征值的映射，本文对所有特征值都做了归一化预处理。如果满足 $\{x\mid p(G=S_{\text{spam}}\mid X=x)>1/2\}$，该邮件判断为垃圾邮件；如果 $\{x\mid p(G=S_{\text{spam}}\mid$

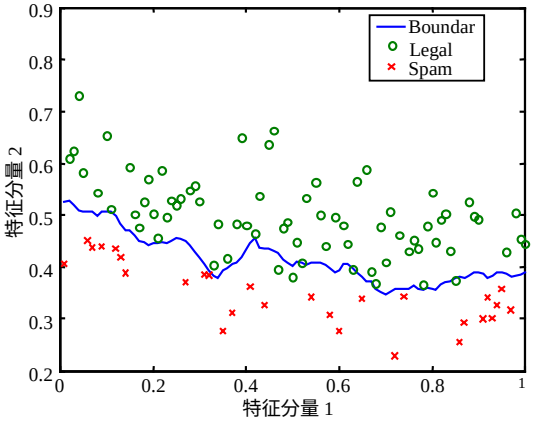
$X = x) < 1/2\}$，则把该邮件判断为正常邮件；如果正好处于边界，则将该邮件放入未知类别缓存队列，留到判定边界更新以后再作二次判断。将已分好类的邮件样本加入样本训练集，取一个适当的时间间隔更新一次判定边界。每次有新的邮件到达时，反复以上步骤，就可以实现基于有监督Bayesian网络的在线学习和分类过滤。



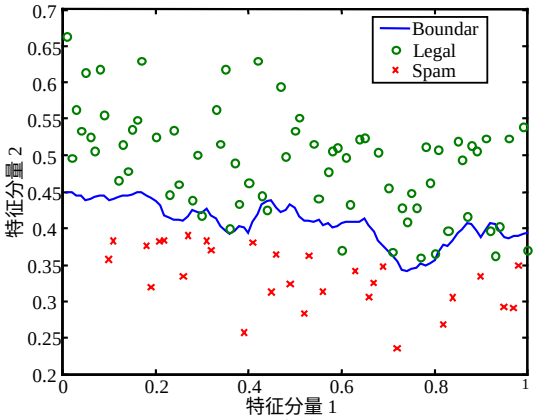
a. EN样本集



b. PU1样本集



c. Ling-Spam样本集



d. CH样本集

图2　在EN、PU1、Ling-Spam、CH集中产生的Bayesian判定边界

4　性能测试

在分析邮件分类网络的性能之前，需要引入误报和漏报的概念[3]。误报是指误将合法邮件判断为垃圾邮件(Legal → Spam)的情况；漏报则恰好相反，是将垃圾邮件误判为合法邮件(Spam → Legal)的情况。整体评价一个分类器的好坏时，需要综合看它在漏报和误报两方面的性能表现。

用户一般能够容忍把少数几封垃圾邮件误判为正常邮件的情况，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，本文解决的方法是引入权值校正。权重准确率的定义式为：

$$W_{\mathrm{Acc}} = \frac{\lambda n_{\mathrm{L}\to\mathrm{L}} + n_{\mathrm{S}\to\mathrm{S}}}{\lambda N_{\mathrm{L}} + N_{\mathrm{S}}} \tag{6}$$

式(6)表示将一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。

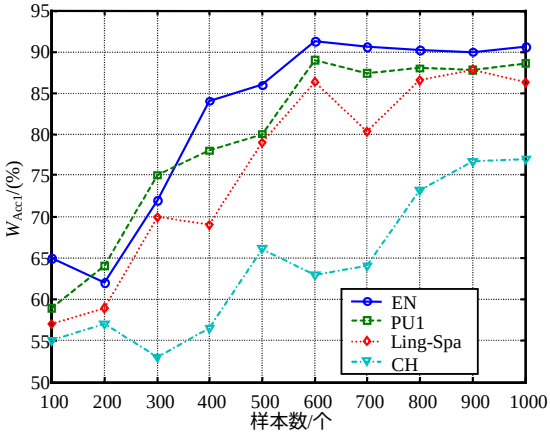


图3　$\lambda = 1$ 时过滤不同邮件集的 W_{Acc1} 对比图

(下转第593页)

的考虑，该模型更加符合当前实际的网络状况。

4　结 束 语

蠕虫病毒对网络的危害性十分巨大，因此，准确描述出其传播规律对于病毒检测和消除至关重要。本文对传统的基于传染病模型理论的蠕虫病毒传播模型和基于概率分析描述的AAWP模型进行了概述，并针对当前因特网中的实际情况，提出了一种对AAWP模型进行改进的方法。仿真分析表明，改进后的模型能够更准确地描述蠕虫病毒在网络中的传播情况。在下一步的工作中，可以考虑由于蠕虫病毒传播而导致的网络流量异常等因素，从而更准确地判别出病毒的传播规律。

参　考　文　献

[1] SONG D, MALAN R, STONE R. A snapshot of global internet worm activity[R]. Arbor Networks, 2001.

[2] CHEN Z, GAO L, KWIAT K. Modeling the spread of active worms[C]//In: Proceedings of the IEEE INFOCOM 2003. San Francisco: IEEE Computer Society, 2003.

[3] MOORE D, SHANNON C. The Spread of code-red worm (CRv2)[EB/OL]. http://www.caida.org/analysis/security/code-red/ coderedv2_analysis.xml, 2005- 01-12.

[4] ZOU C, GONG W, TOWSLEY D. Code red worm propagation modeling and analysis[C]//In: Proceedings of the 9th ACM Symp on Computer and Communication Security. Washington: ACM Press, 2002.

[5] KIENZLE D, ELDER M. Recent worms: A survey and trends[C]//In: Proceedings of the ACM CCS Workshop on Rapid Malcode (WORM 2003). Washington: ACM Press, 2003.

[6] ZOU C, GONG W, TOWSLEY D. On the performance of Internet worm scanning strategies[R]. Electrical and Computer Engineering Department, University of Massachusetts, 2003.

[7] WANG Y, WANG C. Modeling the effects of timing parameters on virus propagation[C]//In: Proceedings of the ACM CCS Workshop on Rapid Malcode (WORM 2003). Washington: ACM Press, 2003.

[8] STEVE W. Open problems in computer virus research [EB/OL]. http://www.research.ibm.com/antivirus/SciPapers/White/Problems/ Problems.html, 2005-01-12.

[9] 文伟平, 卿斯汉, 蒋建春, 等. 网络蠕虫研究与进展[J]. 软件学报, 2004, 15(8):1208-1219.

编　辑　熊思亮

(上接第589页)

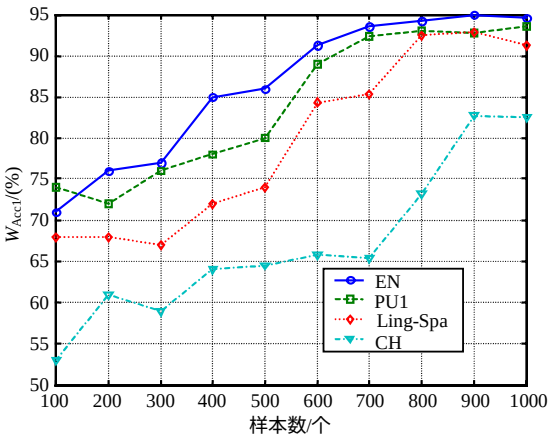


图4　$\lambda=9$ 时过滤不同邮件集的W_{Acc1}对比图

从图3、4中可看出，分类网络对英文邮件样本集的过滤性能比较好，无论 $\lambda=1$ 或 $\lambda=9$，当邮件样本超过600以后，分类网络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对分类网络的性能构成瓶颈的原因，而在本文构建的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(6)的权重强调了正确判断对分类网络性能的正面影响，所以权重 $\lambda=9$ 的曲线相对于 $\lambda=1$ 的曲线更平滑。

5　结　论

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。本文通过吸取和综合传统邮件过滤技术的思想，从报文格式和邮件内容两方面提取邮件特征，构建相对完备的有监督Bayesian网络，为降低邮件漏报和误报发生的数量，提高垃圾邮件过滤算法的广普性提供了一种有益的解决思路。

参　考　文　献

[1] JENSEN F. An Introduction to Bayesian networks[M]. London: UCL Press, 1996: 196-202.

[2] 刘　震, 佘　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] ANDROUTSOPOULOS I, KOUTSIAS J, CHANDRINOS V, et al. An evaluation of naive Bayesian anti-spam filtering[C]//Proceedings on Machine Learning in the New Information Age. Georgia, USA: [s. n.], 2000: 578-584.

编　辑　孙晓丹