维普资讯 http://www.cqvip.com

基于 Bayes 参数估计的垃圾邮件过滤算法研究*)

刘 震 佘 堃 周明天

(电子科技大学计算机学院卫士通安全联合实验室 成都 610054)

摘 要 朴素 Bayes 邮件过滤算法由于简单、易于理解,已被人们广泛接受,并应用到一些商用邮件系统当中。但面对目前垃圾邮件问题依然严重的现状,人们逐渐开始认识到采用简单的朴素 Bayes 邮件过滤算法已不能满足现有邮件过滤的性能要求。Bayes 网络一直以来作为知识发现的一个重要分支,是人们研究的热点;邮件过滤问题也可以映射到一个 Bayes 决策网络模型中。通过构建针对邮件过滤的 Bayes 决策网络模型,并经过概率学习对关键节点作 Bayes 参数估计,可以实现邮件的概率分类发现。邮件样本试验结果表明新算法与朴素 Bayes 邮件过滤算法相比具有更快的收敛速度和更高的稳定性。

关键词 垃圾邮件,贝叶斯网络,邮件过滤,参数估计

Research on Advanced Filtering Algorithm for Spam Email Based on Bayes Parameter Estimation

LIU Zhen SHE Kun ZHOU Ming-Tian

(Westone United Lab. of College of Computer Science and Engineering, UESTC, Sichuan, Chengdu 610054)

Abstract Presently, naive Bayes algorithm for Email filtering has been accepted widely for its simplicity and plainness. At the same time, the algorithm has also been applied in many commercial Email system. However, facing with the persisted severe situation of spam Email, people are realizing gradually that depending on such a simple algorithm can't meet the practical needs for anti-spam. On the other hand, Bayes network, as an important branch in knowledge discovery area, has been researched and explored for a long time. The question of Email filter can be projected to a Bayes network model too. By doing so and utilizing Bayes parameter estimation for key node in the model, we can accomplish Email classification and discrimination on probabilistic condition. According to the testing results, new algorithm has much higher rate to converge and stabilization than the naive Bayes one.

Keywords Spam email, Bayes network, Email filtering, Parameter estimation

1 引言

垃圾邮件(Spam Email)的出现给电子邮件系统和用户带来了很多不便和危害,不仅造成了邮件服务器负荷增大,也成为有害信息和病毒传播的重要途径。为了保证邮件系统的正常运行和邮箱用户利益,必须使邮件系统具有反垃圾邮件的能力。目前,主流的邮件过滤策略是采用限制发送数量、过滤地址与简单关键字匹配等方式来阻止垃圾邮件的传播。针对关键字过滤的朴素 Bayes 算法由于形式简单,易于实现,被众多邮件过滤系统所采用[1,2]。很多邮件服务运营商也纷纷将朴素 Bayes 过滤功能整合到自己的网上邮局,如 yahoo、网易等。然而,当前垃圾邮件泛滥的形势仍然严峻。据权威机构统计,2003 年发往中国邮件服务器的垃圾邮件为 1500 亿封,国内邮件服务商过滤 60%至 80%,用户实际共计收到垃圾邮件 470 亿封,给中国经济造成 48 亿元的损失。研究高性能的反垃圾邮件算法已经成为迫切的形势要求[3~5]。

贝叶斯网络也称为信念网络,最早由 Judea Pearl 于 1988 年提出的贝叶斯网络实质上是一种基于概率的不确定性推理网络[6]。通过学习与推理,贝叶斯网络可以利用一些先验的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[8]。

本文首先介绍针对普遍 Bayes 网络的参数估计理论;然后将这种理论应用到垃圾邮件过滤中,构建针对关键字过滤的 Bayes 网络模型,并提出了一种基于 Bayes 参数估计的垃圾邮件过滤算法。通过实验,证明这种新的分类过滤算法与朴素 Bayes 过滤算法相比具有更稳定的性能和更快的收敛速度。

2 Bayes 参数估计理论

Bayes 参数估计也称为 Bayes 参数学习,是本文第 3 节构建邮件过滤算法的基础。Bayes 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。它通过不断的概率学习,从而不断适应和逼近变化的概率分布。假设 X 满足多项式分布,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,可以计算

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

*)基金项目:国家 863 项目 863-104-03-01 课题支持。刘 震 博士生,主要研究方向为智能安全,不确定推理,人工智能等。佘 堃 博士研究生,副教授,研究方向为智能安全。周明天 教授,博士生导师,主要研究方向为网络计算,网络安全技术。

维普资讯 http://www.cqvip.com

由 Bayes 公式，有

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

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

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

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

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 $\alpha_1+N_1,\alpha_2+N_2,\cdots,\alpha_k+N_k$。容易计算

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

根据式(1)，X 的后验 Bayes 参数估计值为

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

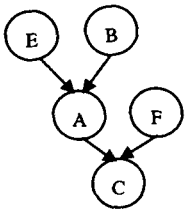

图 1　Bayes 网络

推广到一般的 *Bayes* 网络，同样可以对各个节点作 *Bayes* 参数估计。假设有一个五节点的 *Bayes* 网络(如图 1 所示)，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

$$D=\begin{bmatrix}E[1] & B[1] & A[1] & F[1] & C[1]\\ \vdots & \vdots & \vdots & & \\ E[m] & B[m] & A[m] & F[m] & C[m]\end{bmatrix}$$

那么该网络的似然函数分布为

$$\begin{aligned}L(\Theta:D)&=\prod_m p(E[m],B[m],A[m],F[m],C[m]:\Theta)\\&=\prod_m p(C[m]\mid A[m]:\Theta_{C\mid A})p(A[m]\mid E[m],B[m]:\Theta_{A\mid D,B})\\P(F[m]:\Theta_F)p(E[m]:\Theta_D)p(B[m]:\Theta_B)&=\prod_m p(C[m]\mid A[m]:\Theta_{C\mid A})\prod_m p(A[m]\mid E[m],B[m]:\Theta_{A\mid D,B})\prod_m p(F[m]:\Theta_F)\prod_m p(E[m]:\Theta_D)\prod_m p(B[m]:\Theta_B)\end{aligned} \tag{6}$$

对某一个节点，它的似然函数满足

$$\begin{aligned}L_i(\Theta_i:D)&=\prod_m p(x_i[m]\mid pa_i[m]:\Theta_i)=\prod_{pa_i}\prod_{pa_i[m]=pa_i}p(x_i[m]\mid pa_i:\Theta_i)\\&=\prod_{pa_i}\prod_{x_i}p(x_i\mid pa_i:\Theta_i)^{N(x_i,pa_i)}\\&=\prod_{pa_i}\prod_{x_i}\theta_{x_i\mid pa_i}^{N(x_i,pa_i)}\end{aligned} \tag{7}$$

式(7)中 pa_i 是 x_i 的父节点。将式(7)带入式(1)，可求得

$$\theta_{x_i\mid pa_i}=\frac{\alpha(x_i\mid pa_i)+N(x_i\mid pa_i)}{\alpha(pa_i)+N(pa_i)} \tag{8}$$

3　邮件分类过滤算法

针对垃圾邮件的 Bayes 参数估计首先要初始化先验的概率分布参数。我们需要初始化两个重要参数 $\theta_{keyword}$ 和 $\theta_{spam\mid keyword}$。$\theta_{keyword}$ 是邮件中某一关键字出现的概率，$\theta_{spam\mid keyword}$ 是邮件中该关键字出现的条件下邮件为垃圾邮件的概率。根据两个参数的依赖关系，构建图 2 所示的 Bayes 网络模型。网络中 X 表示邮件中出现某一关键字的事件；Y 表示邮件中存在该关键字的条件下，该邮件是垃圾邮件的事件。对 $m+1$ 次到达的新邮件作 $\theta_{keyword}$ 和 $\theta_{spam\mid keyword}$ 的 Bayes 参数估计要依赖前 m 次的输入邮件分布。

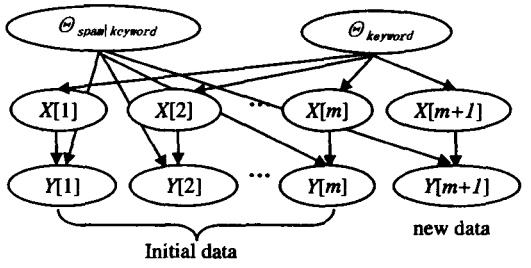

图 2　邮件分类 Bayes 网络

创建垃圾邮件的关键字列表。假设初始化 n 个过滤关键字，记为 $\{keyword_1,keyword_2,\cdots,keyword_n\}$。然后引入一批邮件训练样本，其中垃圾邮件 m 封，正常邮件 n 封。

①初始化：统计关键字和垃圾邮件在这批邮件中的概率分布情况。即对当前邮件样本，计算 $p\{x=keyword_i\}$，$p\{y=spam\mid x=keyword_i\}$，$p\{y=spam\}$，其中 $i=1,2,\cdots,n$。

那么对于每个关键字 $x=keyword_i$，由 Bayes 公式可以得到

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

令 $\theta_{spam\mid keyword_i}=p\{y=spam\mid x=keyword_i\}$。

②训练学习：由式(8)，对每个关键字有

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

其中，$\alpha(y=spam\mid x=keyword_i)=M\theta_{spam\mid keyword_i}$；

$\alpha(x=keyword_i)=M\theta_{keyword_i}$；$M=m+n$。

当有新的邮件到达时，首先分析邮件中的关键字。一般情况下可能会有多个过滤关键字出现在该邮件中。此时，要针对每一个出现在邮件中的 $keyword_i$，更新相应的 $\theta_{keyword_i}$；并且根据上次的 $\theta_{spam\mid keyword_i}$ 判断该邮件的类别；如果 $\sum_{keyword_i}\theta_{spam\mid keyword_i}\theta_{keyword_i}>\theta_{spam}$，那么判定该邮件为垃圾邮件，否则为合法邮件。由判断结果更新 θ_{spam}。修改 $N(y=spam\mid x=keyword_i)$，代入式(10)更新 $\theta_{spam\mid keyword_i}$。每次有新的邮件到达时，重复以上步骤。

4　邮件过滤试验和性能分析

常规的邮件过滤采用的是朴素 Bayes 分类算法，利用极大似然法估计后验概率。采用这种算法虽然分类速度较快，但也有误报率高的问题。下面，我们对两种方法加以比较。首先，利用从哥伦比亚大学 IDS 实验室网站下载的邮件样本初始化先验参数[9]，设定邮件过滤关键字列表。然后对新到达的邮件作分类训练。图 3 是在相同的邮件样本输入情况下，采用不同 Dirichlet 先验参数估计和极大似然法估计的后

维普资讯 http://www.cqvip.com

验分布变化对比图。从图中可以看出，基于 Dirichlet 先验的参数估计受分布变化的影响更小。

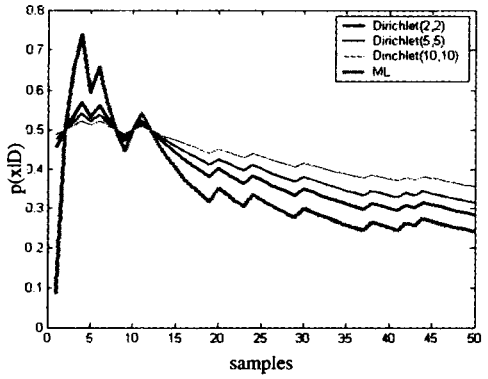

图 3　后验分布变化对比图

计算 KL 散度可以比较估计的参数分布与真实分布之间的误差。KL 散度的定义式为

$$KL(P \| Q)=\sum_x P(X)\log\frac{P(X)}{Q(X)} \tag{11}$$

其中，$P(X)$是真实分布，$Q(X)$是估计的参数分布。图 4 是利用基于 Dirichlet 先验的 Bayes 估计算法和基于极大似然后验估计的朴素 Bayes 分类算法的 KL 散度比较图。从图 4 中可以看出，虽然在小样本邮件输入时，采用基于 Dirichlet 先验的 Bayes 估计得到的 KL 值震荡较大，但随着样本数的增加，KL 值收敛得更快，也即是说 Bayes 估计与极大似然后验估计相比更符合真实的邮件类别分布。

结束语　针对当前严重的垃圾邮件问题，本文提出了一种基于 Bayes 参数估计的邮件分类新算法，算法充分利用了 Bayes 网络的决策分析能力和参数学习能力。实验表明，本文提出的邮件分类过滤算法与传统朴素贝叶斯分类算法相比抗分布噪声能力更强，同时分类精度也更高。基于 Bayes 参数估计的邮件分类算法为解决目前困扰人们的垃圾邮件问题，提供了一种新的解决方法和途径。

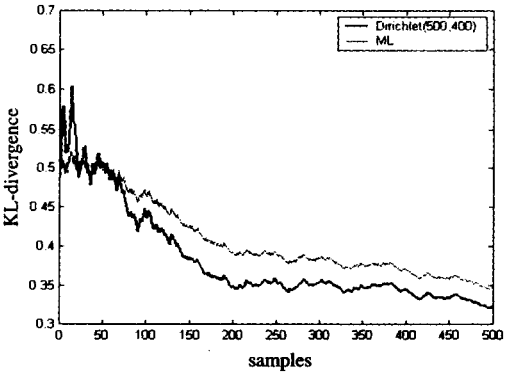

图 4　KL 散度比较图

参 考 文 献

1　Androutsopoulos I, et al. An Evaluation of Naive Bayesian Anti-Spam Filtering [C]. In: Workshop on Machine Learning in the New Information Age. C. 2000. 578～584

2　Hidalgo J M G. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization[C]. In: Proc. of the 6th Intl. Conf. on the Statistical Analysis of Textual Data. 2003. 256～260

3　Internetweek. com. Spam Is A Thousand Times More Horrible Than You Can Imagine. http://www. internetweek. com/story/INW20021219S0003. (2002a)

4　Internetweek. com. Study: e-mail viruses up. http://www. internetweek. com/story/INW20021109S0002. (2002b)

5　Mulligan G. Removing the Spam: E-Mail Processing and Filtering [M]. London: Addison-Wesley Longman Publishing Co., Inc. 2002. 102～115

6　Han Jiawei, Kamber M. Data mining-concept and technique[M]. BeiJing: machinery industry press, 2001. 218～222

7　O'Brien C, Vogel C. Spam filters : bayes vs. chi-squared; letters vs. words [C]. Proceeding/Series-Proceeding-Section-Article. 2003. 291～96

8　Jensen F V. An Introduction to Bayesian Networks. London: UCL Press, 1996

9　http://www1. cs. columbia. edu/ids/

(上接第 20 页)

步的时间是 $O(|P|\times|T|\times(|P|+|T|+|F|))$。

对于一个极小死锁，最多有 $|P|$ 个库所，每个库所最多有 $|T|$ 条弧，因此求一个极小死锁的归约子网的时间是 $O(|P|\times|T|)$，于是第 2 步的时间是 $O(|P|\times|P|\times|T|)$。

由文[10,11] 用 FC 网的活性判定算法判定 N_D^{pre}，N_D^{post} 的时间是多项式时间的，而 N_D^{pre}、N_D^{post} 的个数最多是 $|P|$ 个，因此第 3 步的时间是 $O(|P|^2\times$(用 FC 网的活性判定算法判定 $N_{\overline{D}}^{pre}$ 所需时间))。

综上所述算法 2 的时间复杂性是多项式时间的。 □

用算法 2 可以对图 1 所示的 AC 网系统 Σ_0 进行判定。由于 $N_{\overline{D}}^{pre}$ 是活的但 $N_{\overline{D}}^{post}$ 不是活的，因此 Σ_0 不满足活性单调性。

定理 4.2 与定理 4.4 的结果取决于 FC 网活性判定的时间，文[12]给出的时间是 $O(n^6)$，$n=\max(|P|,|T|)$。

结论　上面我们给出的算法 1，仍然是充分性的判定算法，它们不能完全解决 AC 网的活性判定问题，当输出是“No”时，还需用其它方法进一步判定，但它可以解决文[11]不能解决的判定问题。我们将继续努力，去追求更加完美的结果。算法 2 是一个充分必要的判定，我们非常满意这个结果。在 AC 网中，找到在多项式时间内可判定其活性的充分必要的判定是我们的目标和追求。

参 考 文 献

1　陆维明，甄强. Petri 网的活性研究[J]. 计算机科学，1999,26(4):1～4

2　Reisig W. Petri Nets: an Introduction [M]. Germany: Springer EATCS Monographs in Theoretical Computer Scinence(1985)

3　Petri C A. Forgotten Topics [A]. LNCS, Springer-Verlag, 1987, 255:500～514

4　Best E. Structure Theory of Petri Nets: the Free Choice Hiatus, Petri Nets [A]. LNCS, Springer-Verlag, 1987, 255:168～206

5　焦莉. Petri 网的活性研究[D]:[博士论文 2001]. 中国科学研究院数学与系统科学研究院数学研究所

6　焦莉，陆维明. 扩展强化非对称选择网的活性和有界性[J]. 软件学报，2001,12(9):1312～1317

7　Esparza J, Silva M. A Polynomial-time Algorithm to Decide Liveness of Bounded Free Choice Nets [J]. Theoretical Computer Science, 1992, 12:185～205

8　Jiao Li, Wei-ming Lu. Notes on Boundeedness of Extended Strong Asymmetric Choice Nets Ⅱ [J]. J. Comput. Sci. & Technol, 2001, 16(5):426～433

9　Zhen Qiang, Wei-ming Lu. On Liveness and Safeness of Asymmetric Choice Net [J], 软件学报，2000,11(5):590～605

10　Kemper P, Bause F. An Efficient Polynomial-time Algorithm to Decide Liveness Boundedness of Free Choice Nets [A]. LNCS, Springer-Verlag, 1992, 616:263～278

11　Barkaoui K, Pradat-Peyre J. On Liveness and Controlled Siphons in Petri Nets [A]. LNCS, Springer-Verlag, 1996, 1091:57～72