基于贝叶斯过滤算法的反垃圾邮件策略

李闻天
(昆明理工大学 信息与网络中心,云南 昆明 650093)

摘要: 结合传统贝叶斯算法的数学定理,给出了朴素贝叶斯过滤算法以及该算法在反垃圾邮件中的六个处理步骤,算法通过渐进学习的方式分别建立三个哈希表,并预置相应的阀值来判定收到的邮件是否为垃圾邮件.以实例方式说明了此算法为基础的反垃圾邮件策略,最后指出了朴素贝叶斯算法存在的问题及可能改进的方向,对防范和处理垃圾邮件提供理论依据和实践参考.

关键词: 贝叶斯定理; 朴素贝叶斯算法; 反垃圾邮件; 哈希表; 网络管理
中图分类号: TP393.08 文献标识码: A 文章编号: 1007-855X(2005)03-0068-04

Strategy of Anti-Spam Based on Bayesian Filtering Algorithm

LI Wen-tian
(Center of Information and Network, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: Based on the traditional Bayesian theory, Naive Bayesian filtering algorithm and its six processing steps in anti-spam are presented. The algorithm creates three Hash Tables respectively by gradual study of mails, and presets threshold accordingly to determine whether a mail received is spam or not. Examples are given to prove this algorithm is a basic strategy to anti-spam. Finally, problems and future improvement of Naive Bayesian filtering algorithm are pointed out, which provides theoretical basis and practical reference to keep away and deal with spam.

Key words: Bayes theorem; Bayesian Naive Algorithm; anti-spam; Hash Table; network management

0 引言

随着网络以超摩尔定理的速度迅猛增长,电子邮件以其快捷、经济的通信优点也得到飞速发展.但是,许多广告、反动信息、色情信息、病毒或蠕虫引起的垃圾邮件也在网络中深度扩散.因垃圾邮件数量多,具有反复性、强制性、欺骗性、不健康性或传播速度快等特点,严重干扰了人们正常生活,浪费用户的时间、精力,甚至造成很多额外的经济支出和信息安全隐患.尽管一些商业化产品允许用户人工建立垃圾邮件的过滤规则,但是系统要求用户应具有丰富的经验,且需花销许多时间,况且垃圾邮件在不断改变,用户必须经常调整这些规则.因此,研究邮件自动过滤方法具有重要意义.邮件自动过滤方法研究主要有基于规则和基于概率两种,基于概率的自动过滤算法已逐渐成为主要研究趋势,贝叶斯加权统计分析算法可以根据用户认为的垃圾邮件和非垃圾邮件进行统计计算,具有学习渐进的功能,可以逐渐取得好的效果,可达到较高的判断准确率.

1 贝叶斯定理的数学定义

贝叶斯定理是计算概率的一种方法,即认为一个事件会不会发生取决于该事件在先验分布中已经发生过的次数.贝叶斯定理指出,对于事件 X 和 Y,已知 Y 的概率时,X 发生的概率(用 $p\{X \mid Y\}$ 表示),等于已知 X 的概率时 Y 发生的概率(用 $p\{Y \mid X\}$ 表示)乘以 X 的概率($p\{X\}$),再除以 Y 的概率($p\{Y\}$).或者用公式表述如下:

收稿日期: 2004-12-20.
作者简介: 李闻天(1973~),男,硕士,助理研究员.主要研究方向: 计算机网络管理. E-mail: lw@kmust.edu.cn

$$P\{X/Y\} = \frac{P\{X\} \times P\{Y/X\}}{P\{Y\}}$$

这一公式更形式化的定义为: 设试验 E 的样本空间为 S. Y为 E 的事件, $X_1, X_2, \cdots, X_n$为 S的一个划分, 即构成一完备事件组, 且 $P(Y) > 0, P(X_i) > 0 (i = 1, 2, \cdots, n)$ 则

$$P\{X_i/Y\} = \frac{P\{X_i\}P\{Y/X_i\}}{\sum_{j=1}^{n} P\{X_j\}P\{Y/X_j\}}, \quad i = 1, 2, \cdots, n$$

贝叶斯定理已经成为垃圾邮件过滤系统的基础, 很多过滤策略都是基于贝叶斯定理而提出的.

2 朴素贝叶斯算法

朴素贝叶斯算法是在一般贝叶斯算法的基础上, 通过假定各因素之间不存在任何联系, 即完全独立而得到的一种简化贝叶斯算法. 这种算法在文本分类中得到非常广泛的应用. 根据贝叶斯概率公式, 对于给定的向量 $d(\omega_1, \omega_2, \cdots, \omega_n)$ 属于第 $C_k (k = 1, 2, \cdots, m)$ 类的概率为:

$$P\{C_k/d\} = \frac{P\{C_k\} \times P\{d/C_k\}}{P\{d\}}$$

其中, $P\{d\} = \sum_{k=1}^{m} P\{d/C_k'\} \times P\{C_k'\}$

由上式可知, 要判断一个待识别邮件的类别, 可以通过计算 $P\{C_k/d\}$ 概率来完成, 它表示出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, 而决定该文档属于第 C_k 类的概率. 我们可通过先验概率 $P\{C_k\}$ 和条件概率 $P\{d/C_k\}$ 来得到后验概率 $P\{C_k/d\}$.

假定 ω_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:

$$P\{d/C_k\} = P\{\omega_1, \omega_2, \cdots, \omega_n/C_k\} = \prod_{j=1}^{n} P\{\omega_j/C_k\}$$

假定 N_k 表示训练样本集中属于第 C_k 类的邮件总数, N 表示训练样本集中的邮件总数, 先验概率 $P\{C_k\}$ 为:

$$P\{C_k\} = \frac{N_k}{N}$$

3 贝叶斯过滤算法反垃圾邮件的基本步骤

第一步: 通过收集大量的邮件, 按规则定义分为垃圾邮件和非垃圾邮件, 建立垃圾邮件集和非垃圾邮件集, 相当于两个数据库;

第二步: 提取邮件主题和邮件体中的独立字串, 例如 A3B4CD2, $#156等作为 TOKEN串, 并统计提取出的 TOKEN串出现的次数, 即字频. 按照上述方法分别处理垃圾邮件集和非垃圾邮件集中的所有邮件;

第三步: 每一个邮件集对应一个哈希散列表, hash_nospam对应非垃圾邮件集, 而 hash_spam对应垃圾邮件集. 表中存储 TOKEN串到字频的映射关系, 如下所示:

TOKEN串	出现次数
A3B4CD	n1
$#156	n2
SSH43	n3
KUKU	n4

第四步: 计算每个哈希表中 TOKEN串出现的概率 P= {(某 TOKEN串的字频) /(对应哈希表的长度)};

第五步: 综合考虑 hash_nospam和 hash_spam, 推断出当新来的邮件中出现某个 TOKEN串时, 该新邮件为垃圾邮件的概率, 数学表达式为:

事件 S: 该邮件为垃圾邮件, t_1, t_2, …, t_n 代表 TOKEN 串, 则 $P\{S/t_i\}$表示在邮件中出现 TOKEN 串 t_i 时, 该邮件为垃圾邮件的概率.

设 $P_1(t_i)$ = (t_i 在 hash_nospam 中的值), $P_2(t_i)$ = (t_i 在 hash_spam 中的值)则

$P\{S/t_i\} = P_1(t_i)/[(P_1(t_i) + P_2(t_i)], i = 1, 2, \cdots, n$

第六步: 建立新的哈希表 hash_possible存储 TOKEN 串 t_i 到 $P\{S/t_i\}$的映射, 如下所示:

TOKEN 串	垃圾邮件的概率
A3B4CD2	$p\{S/t_1\}$
$#156	$p\{S/t_2\}$
SSH43	$p\{S/t_3\}$
KUKU	$p\{S/t_4\}$

一直得到出现某字串的邮件为垃圾邮件的概率时, 垃圾邮件集和非垃圾邮件集的学习过程就算结束. 根据建立的哈希表 hash_possible可以估计一封新到的邮件为垃圾邮件的可能性. 当新到一封邮件时, 按照第二步生成 TOKEN 串. 查询 hash_possible得到该 TOKEN 串的键值. 假设由该邮件共得到 N 个 TOKEN 串, t_1, t_2, …, t_n, hash_possible中对应的值为 $P_1, P_2, \cdots, P_N$, $P\{S/t_1, t_2, t_3, \cdots, t_n\}$表示在邮件中同时出现多个 TOKEN 串 t_1, t_2, …, t_n 时, 该邮件为垃圾邮件的概率. 由复合概率公式可得 $P\{S/t_1, t_2, t_3, \cdots, t_n\} = (P_1 * P_2 * \cdots * P_N)/[P_1 * P_2 * \cdots * P_N + (1-P_1) * (1-P_2) * \cdots * (1-P_N)]$, 当 $P\{S/t_1, t_2, t_3, \cdots, t_n\}$超过预定阀值时, 就可以判断邮件为垃圾邮件.

4 贝叶斯过滤算法举例

例如: 一封含有“法轮功”字样的垃圾邮件 A 和一封含有“法律”字样的非垃圾邮件 B, 根据邮件 A 生成 hash_ spam, 该哈希表中的记录为:

法: 1次, 轮: 1次, 功: 1次, 计算得在本表中: 法出现的概率为 0.33, 轮出现的概率为 0.33, 功出现的概率为 0.33 根据邮件 B生成 hash_nospam, 该哈希表中的记录为:

法: 1次, 律: 1次, 计算得在本表中: 法出现的概率为 0.5, 律出现的概率为 0.5.

综合考虑两个哈希表, 共有四个 TOKEN 串: TOKEN1: 法, TOKEN2: 轮, TOKEN3: 功, TOKEN4: 律, 当邮件中出现“法”时, 该邮件为垃圾邮件的概率为: $P = 0.33/(0.33 + 0.5) = 0.398$, 出现“轮”时: $P = 0.33/(0.33 + 0) = 1$, 出现“功”时: $P = 0.3/(0.3 + 0) = 1$, 出现“律”时: $P = 0/(0 + 0.5) = 0$, 由此可得第三个哈希表: hash_possible, 其数据为:

法: 0.398, 轮: 1, 功: 1, 律: 0

当新到一封含有“功律”的邮件时, 我们可得到两个 TOKEN 串, TOKEN1: 功, TOKEN2: 律, 查询哈希表 hash_possible可得

P(垃圾邮件 | 功) = 1, P(垃圾邮件 | 律) = 0

此时该邮件为垃圾邮件的可能性为: $P = (0 * 1)/[0 * 1 + (1-0) * (1-1)] = 0$, 由此可推出该邮件为非垃圾邮件.

5 存在问题及改进

垃圾邮件过滤的性能评价通常由多个指标来测定, 如召回率(垃圾邮件检出率)、正确率(垃圾邮件检对率)、精确率(所有邮件的判对率)、错误率(所有邮件的判错率)、漏报率、虚报率等. 以上策略所使用的朴素贝叶斯算法, 在计算条件概率时, 当文档中单词出现次数较少的, 会导致分子或分母为零而出现系统不能正常运行, 可以对不同的应用场合对算法予以修正; 还有朴素贝叶斯算法没有考虑合法邮件被错判为垃圾邮件的情况, 不少学者使用了其他贝叶斯模型, IBM 的 Mertz不是采用独立性假设, 而是考虑使用 N 元语言模型来估计相关的概率, 有的在朴素贝叶斯算法的基础上作改进, 如引入代价因子的最小风险算法

等, 但有时一味追求某个指标的提高, 往往又会引起其他指标的下降, 如引入代价因子的算法, 合法邮件被系统误判为垃圾邮件的可能性减少, 但同时垃圾邮件被系统误判为合法邮件的可能性增加, 我们只能根据实际的应用情况, 综合权衡各个指标, 适当改进算法, 建立适于各自系统特色应用的垃圾邮件过滤系统.

6 结 语

贝叶斯过滤算法是用户根据自己所接受的垃圾邮件和非垃圾邮件的统计数据来创建的, 这意味着垃圾邮件发送者无法猜测出过滤器是如何配置的, 从而有效阻止垃圾邮件. 贝叶斯过滤算法能够学习分辨垃圾邮件与非邮件之间的差别, 差别是用概率来表示的, 并且自动应用到以后的检测中. 在收到一定数量的信件后, 一个好的贝叶斯过滤算法就可以自动识别各种垃圾邮件. 垃圾邮件是全球性的问题, 且已经成为一种社会现象, 单靠反垃圾邮件技术的发展或是纯粹的技术手段是无法解决的 , 还应当采用管理与技术相结合的方式, 以先进的技术手段为基础, 以完善的管理制度和法律法规为依托, 对社会各主体的邮件活动进行规范, 通过建立国家级的反垃圾邮件公共服务体系, 完善国内的垃圾邮件举报平台, 促进各运营商和邮件服务商的协调合作, 进而推动反垃圾邮件技术的更新和快速发展.

参考文献:

[1] James O• Berger. 统计决策论及贝叶斯分析 [M]. 贾乃光, 吴喜之译. 北京: 中国统计出版社, 1998. 17~ 19, 130~ 146.
[2] 盛骤. 概率论与数量统计 • 第二版 [M]. 北京: 高等教育出版社, 1994. 18~ 25.
[3] 石霞军. 基于最小风险的贝叶斯邮件过滤算法 [J]. 计算机科学, 2002, 29(8): 50~ 51.
[4] 潘文锋. 基于内容的垃圾邮件过滤研究 [EB/OL]. http: //www. nlp. org. cn/docs/doclist. php? cat_id= 17& type= 10, 2004 - 11- 20.

(上接第 67页)

$$\begin{bmatrix} (A_{\alpha}X + B_rW) + (A_{\alpha}X + B_rW)^T + \varepsilon D_rD_r^T & (A_{\alpha}X + B_rW)^T + \varepsilon\sqrt{T}D_rD_r^T & (E_{1r}X + E_{2r}W)^T \\ (A_{\alpha}X + B_rW) + \varepsilon\sqrt{T}D_rD_r^T & -X + \varepsilon TD_rD_r^T & 0 \\ E_{1r}X + E_{2r}W & 0 & -\varepsilon I \end{bmatrix}$$

成立, 进而, 当该矩阵不等式有解 ε, W和矩阵 X 时, 则增益矩阵 $K = WX^{-1}$.

证明　由定理 1不难证明本定理, 注意到式 (14), 事实上只要在式 (6) 中用 $A_{\alpha} + B_rK$ 置换 A_{α}, 用 $E_{1r} + E_{2r}K$ 置换 E_{1r}, 并考虑到 $\gamma^2 = 1$和 $W = KX$ 就可以了.

3 结 论

本文研究了 Delta算子描述下的具有圆形区域极点约束的不确定线性系统的鲁棒稳定性分析和鲁棒控制问题, 利用线性矩阵不等式给出了系统鲁棒 D 稳定的充要条件, 通过建立和求解一个凸优化问题, 给出了系统 D 稳定的摄动参数的尽可能大允许摄动界. 提出 Delta算子不确定系统状态反馈控制器的设计, 得到了该问题可解的充要条件.

参考文献:

[1] 张端金, 杨成梧. 反馈控制系统 Delta算子理论的研究与发展 [J]. 控制理论与应用, 1998, 15(2): 153~ 160.
[2] Piou J E, Sobel K M. A time domain approach to performance robustness of sampled data systems using the delta operator [C]. In: Proc. 31st IEEE Conf. Decis. Contr., 1992, 1968~ 1969.
[3] 张端金, 杨成梧. Delta算子系统的鲁棒性能分析 [J], 自动化学报, 2000, 26(6): 848~ 852.
[4] 张端金, 吴捷, 杨成梧. Delta算子系统圆形区域极点配置的鲁棒性 [J]. 控制与决策, 2001, 16(3): 337~ 340.
[5] Khargonekar P P, Petersen I R, Zhou K. Robust stabilization of uncertain systems and optimal control[J]. IEEE Transaction on Automatic Control, 1990, 35(3): 331~ 361.