

基于 ROCCH 方法的垃圾邮件过滤系统的评估¹

项涛^{1,2} 龚俭^{1,2} 丁伟^{1,2} 马薇³

(¹东南大学计算机科学与工程学院, 南京, 210096)

(²江苏省计算机网络技术重点实验室, 南京, 210096)

(³东南大学生物电子学国家重点实验室, 南京, 210096)

摘要: 在分析现实中收到的垃圾邮件数量与正常邮件数量的比例以及垃圾邮件误报代价与正常邮件的误报代价的比例是不断变化的基础上, 本文采用基于 ROC 曲线的 ROCCH 方法来判定不同邮件比例和误报代价比例下的最优的垃圾邮件过滤系统。实验表明 ROCCH 方法能够有效的判定潜在的最优垃圾邮件过滤系统。

关键字: 评估, 垃圾邮件过滤, ROCCH 方法, ROC 曲线

An Evaluation of Spam Filtering Based on ROCCH

Xiang Tao^{1,2} Gong Jian^{1,2} Ding Wei^{1,2} Ma Wei³

(¹School of Computer Science and Engineering, Southeast University, Nanjing 210096, china)

(²Key Laboratory of Computer Network Technology of Jiangsu Province, Nanjing 210096, china)

(³State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, china)

Abstract: After analyzing the varying class distribution and misclassification cost in real world, the ROCCH method based on ROC curve is applied to identify the optimal Spam Filtering during the varying class distribution and misclassification cost in this paper. The result demonstrates that the ROCCH method determine the underlying optimal Spam Filtering effectively.

Key words: Evaluation; Spam Filtering; ROCCH method; ROC curve

1 引言

科学的, 客观的, 公正的评估垃圾邮件过滤系统, 不仅可以给用户选择适合自己需求又有良好过滤效果的垃圾邮件过滤系统提供依据, 还能为垃圾邮件过滤系统的厂商提供好的判定其产品效果的工具。垃圾邮件过滤系统的评估研究具有很强的研究意义和商业意义。

目前已有一些评估垃圾邮件过滤系统的指标和方法, Sahami et al^[1]使用正常邮件的查全率, 查准率和垃圾邮件的查全率, 查准率四个指标评估了一组基于贝叶斯的垃圾邮件过滤系统。Ion Androutsopoulos^[2]使用加权正确率, 加权错误率以及总代价比率 (TCR) 等代价敏感指标, 评估了基于简单贝叶斯和基于关键字的垃圾邮件过滤系统。除此之外, 近些年来, 一些研究者开始使用 ROC 曲线来评估垃圾邮件过滤系统的过滤效果。Gordon Cormack^[3], Andrew P. Bradley^[4]使用 ROC 曲线下的面积评估了多个开源的过滤算法或系统。

然而, 不幸的是, 现实中邮件服务器收到的垃圾邮件与正常邮件的比例是不断变化的, 且不同的服务器收到的垃圾邮件与正常邮件的比例也是不同的 (Tom Fawcett^[5])。除此之外, 在现实中, 不同的服务器, 不同的时间段, 收到的垃圾邮件的误报代价与正常邮件的误

¹作者简介: 项涛 (1983.10-) 男, 江西人, 东南大学在读硕士生, 主要研究方向为网络安全, 网络应用。

报代价之比也是不同的。要客观，公正的评估垃圾邮件过滤系统就必须考虑在不同的邮件比例和误报代价比例下进行评估。**Sahami, Gordon Cormack** 等人的研究使用的是静态指标，没有考虑邮件比例和误报代价比例变化的情况。**Ion Androutsopoulos** 虽然考虑了误报的代价比，但其代价比是给定的 **1, 9, 999** 几个值，不具有通用性；而且也没有考虑垃圾邮件与正常邮件的比例变化的情况。

从前人的研究可以看出，针对现实中不断变化的邮件比例和误报代价比例，目前还没有一种好的办法评估在此条件下的垃圾邮件过滤系统。在本文中，将介绍一种在机器学习研究中用来判定变化目标条件下分类器性能的 **ROCCH (The ROC Convex hull)** 方法，并用该方法来判定变化的邮件比例和误报代价下的最优垃圾邮件过滤系统。

本文的组织如下：首先介绍 ROC 曲线的原理以及如何绘制垃圾邮件过滤系统的 ROC 曲线，在第二节介绍如何使用 ROCCH 方法来判定不同邮件比例和误报代价下的最优垃圾邮件过滤系统，于第三节给出使用 ROCCH 方法评估的几个开源垃圾邮件过滤系统的评估结果，第四节是本文的结论与展望。

2 垃圾邮件过滤系统的 ROC 曲线

ROC 曲线是以 1-特异度为 x 轴，灵敏度为 y 轴的二维曲线。当给定一个分类阈值时，分类器对标准样本集测试后计算特异度和灵敏度可得到 ROC 空间上的一点。通过不断的由小到大的调整分类阈值，得到 ROC 空间上的一系列点，这些点拟合成的曲线就称为该分类器的 ROC 曲线。

在垃圾邮件过滤中，当一封新的邮件到达时，垃圾邮件过滤系统计算该邮件的垃圾值，依据系统设置好的分类阈值，如果该邮件的垃圾值大于分类阈值，就判定该邮件为垃圾邮件，否则判定其为正常邮件。故当设置好分类阈值时，使用该垃圾邮件过滤系统对标准邮件集进行分类测试后，将判定的结果与标准答案比较就会得到一个复合矩阵。通过该复合矩阵，可以计算垃圾邮件过滤系统的灵敏度和特异度。计算方法如下：（其中“Golden Standards”^[3]是标准答案，一般由用户反馈得到；Filter 表示某垃圾邮件过滤系统）。

Golden Standards		
Filter		Spam
	Spam	A
	Ham	C
		Ham
	B	D

其中 A 代表过滤系统判定为垃圾邮件且“Golden Standards”也判定为垃圾邮件的数量

B 代表过滤系统判定为垃圾邮件但“Golden Standards”判定为正常邮件的数量

C 代表过滤系统判定为正常邮件但“Golden Standards”判定为垃圾邮件的数量

D 代表过滤系统判定为正常邮件且“Golden Standards”也判定为正常邮件的数量

I 垃圾邮件查全率(Spam Recall) $SR = \frac{A}{A+C}$ 。也叫做垃圾邮件过滤系统的灵敏度，它体现垃圾邮件过滤系统检测垃圾邮件的敏感程度，敏感度高的过滤系统不容易漏报垃圾邮件。

I 正常邮件查全率(Ham Recall) $HR = \frac{D}{B+D}$ 。也叫做垃圾邮件过滤系统的特异度，它体现垃圾邮件过滤系统检测垃圾邮件的特异程度，特异度高的过滤系统不容易将正常邮件误报为垃圾邮件。

逐步增大分类阈值，生成一系列 $(1 - HR, SR)$ 点对，它们拟合成的曲线就是垃圾邮件

过滤系统的 ROC 曲线。它是以 $1-HR$ 为 x 轴, SR 为 y 轴的两维曲线。由 $1-HR$ 和 SR 确定的 ROC 空间中有几个重要的点。左下角的 $(0, 0)$ 点表示垃圾邮件过滤系统将所有的邮件都判定为正常邮件, 其效果相当于不设置过滤系统。与它相反的另一个极端是, 右上角的 $(1, 1)$ 点表示垃圾邮件过滤系统将所有到来的邮件都判定为垃圾邮件, 此时分类阈值设置为最大。除此之外, 左上角的 $(0, 1)$ 点表示所有的邮件都判定正确, 它代表最理想的垃圾邮件过滤系统的过滤效果, 这在现实中很难达到。在 ROC 空间中, 如果一点的在另一点的西北方 (即: SR 更大或 $1-HR$ 更小或两者皆是), 那么该点的过滤效果好于另一点。

垃圾邮件过滤系统的 ROC 曲线描绘了垃圾邮件过滤系统在垃圾邮件的查全率和正常邮件的误报率之间的平衡关系, 反映垃圾邮件过滤系统的整体过滤效果, 其曲线下的面积越大表示其过滤准确度越好。

3 ROCCH 方法

ROC 曲线是评估垃圾邮件过滤系统的重要手段, 目前使用 ROC 曲线评估垃圾邮件过滤系统的过滤准确度的指标主要是 ROC 曲线下的面积, 但它反映的是系统的整体过滤效果, 不能评估变化邮件比例以及误报代价比例条件下的垃圾邮件过滤系统的过滤效果。本节将使用基于 ROC 曲线的 ROCCH 方法来评估邮件比例和误报代价比例变化情况下的垃圾邮件过滤系统的过滤效果, 针对不同的邮件比例和误报代价比例判断出最优的垃圾邮件过滤系统。

3.1 ISO-性能线

假设新到来的邮件是垃圾邮件的概率为 $p(Y)$, 是正常邮件的概率是 $p(N) = 1 - p(Y)$;

垃圾邮件误报为正常邮件的代价为 $C(Y, N)$, 正常邮件误报为垃圾邮件的代价为 $C(N, Y)$ 。

对于给定的 ROC 空间上的一点 $(1-HR, SR)$, 定义该“离散”过滤系统的预期代价为 $R = p(Y) \cdot SR \cdot C(Y, N) + p(N) \cdot \{1 - (1-HR)\} \cdot C(N, Y)$ 。为使其代价最小, 对 R 求导并

令导数等于 0。得到 $\frac{d_{SR}}{d_{(1-HR)}} = \frac{C(N, Y) \cdot p(N)}{C(Y, N) \cdot p(Y)} = m$ (1)。公式 (1) 得到的是 ROC 曲线上

某点的斜率, 称斜率为 m 的直线为 ISO-性能线。

3.2 ROCCH 方法

由于现实中的邮件比例和误报代价比例在不断变化, 需要标识出潜在的最优的垃圾邮件过滤系统的集合。对于给定的邮件比例和误报代价比例, 依据公式(1)可以计算出 ISO-性能曲线的斜率, 在 ROC 空间中绘制 ISO-性能曲线族。这种条件下最优的垃圾邮件过滤系统是与 ISO-性能曲线相切的切点最靠近左上角的那条 ROC 曲线对应的垃圾邮件过滤系统。

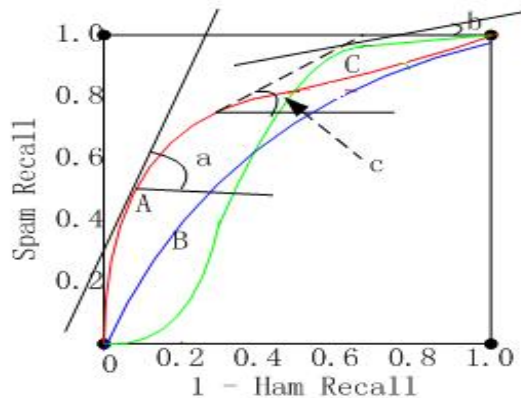


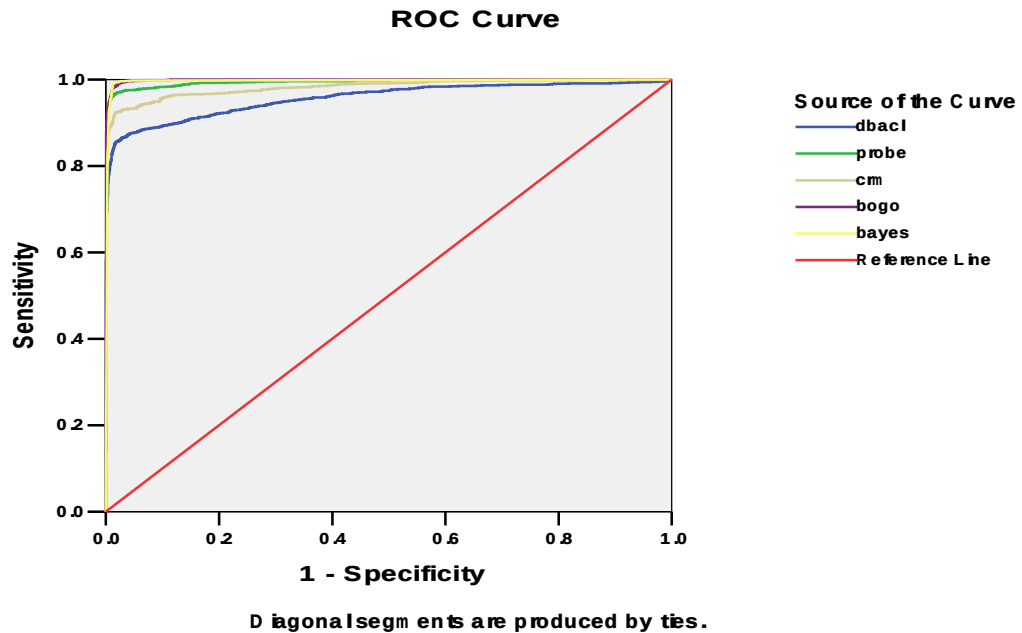
图1: 标识最优的垃圾邮件过滤系统

如上图所示，有红，蓝，绿色三条不同垃圾邮件过滤系统的 ROC 曲线，分别记作系统 A，B，C。从图中所示 ROC 曲线可以看出，当 ISO-性能线的斜率 $m=a$ 时，系统 A 是最优；同样，ISO-性能线的斜率 $m=b$ 时，系统 C 最优。系统 A 和系统 C 是潜在的最优的垃圾邮件过滤系统，从图中可以看到，当 $0 \leq m \leq c$ 时，系统 C 为最优的垃圾邮件过滤系统；当 $c \leq m \leq +\infty$ 时，系统 A 为最优的垃圾邮件过滤系统。当多个垃圾邮件过滤系统需要判定的时候也可以采取同样的方法。

ROCCH 方法标识出潜在的最优垃圾邮件过滤系统，并给出了邮件比例和误报代价比例乘积在不同的范围下的最优垃圾邮件过滤系统。

4 实验结果及分析

使用以上介绍的 ROCCH 方法，本节计算 Bogo, DbacI, CRM-114, SpamProbe, SpamBayes 五个开源的垃圾邮件过滤系统，在不同性能曲线斜率（即不同的邮件比例和误报代价比例的乘积）下的最优垃圾邮件过滤系统。采用的邮件集是 Spamassassin-Corpus，测试方法是交叉交叉法。其实现流程是：首先绘制各垃圾邮件过滤系统的 ROC 曲线；其次使用 ROCCH 方法确定潜在的最优的垃圾邮件过滤系统。ROC 曲线的绘制流程为：首先将垃圾邮件过滤系统采用交叉交叉法训练和测试标准邮件集 Spamassassin-Corpus，并记录过滤系统对邮件测试的分值；其次使用 SPSS 软件根据系统的测试分值绘制 ROC 曲线。绘制各垃圾邮件过滤系统的 ROC 曲线如下：



使用 Tom Fawcett^[6]在 HP 实验室开发的 ROCCH 软件来标识不同邮件比例和误报代价比例乘积下的最优垃圾邮件过滤系统。通过计算得到，当 $0.0 \leq m \leq 0.028$ 或 $5.667 \leq m \leq 830$ 时，**bogo** 是最优的垃圾邮件过滤系统；当 $0.028 \leq m \leq 5.667$ 时，**Spambayes** 是最优的垃圾邮件过滤系统；当 $m \geq 830$ 时，五个过滤系统都一样好。由此，只要给定了邮件比例和误报代价比例，就可以通过计算 m 值来找出在此条件下最优的垃圾邮件过滤系统。假设某邮件服务器在某段时间内收到的垃圾邮件占 **40%**，正常邮件占 **60%**。垃圾邮件误报代价与正常邮件误报代价比例是 **1/9**；那么依据公式（1）计算 $m = \frac{60\%}{40\%} * 9 = 13.5$ 。依据上面的结论，可以

得到此邮件服务器在该段时间内最优的垃圾邮件过滤系统是 **bogo**。

5 结束语

针对现实中不断变化的邮件比例和误报代价,在此条件下目前还没有好的评估垃圾邮件过滤系统的方法。本文使用机器学习中常见的 ROCCH 方法来标识在变化的邮件比例和误报代价比例下的最优垃圾邮件过滤系统,实验表明 ROCCH 方法能够有效的应用于垃圾邮件过滤系统的评估。

但是,在现实中,误报代价比例是很难确定的,不同的人有不同的标准,而且不同邮件之间的误报代价比例也是不一样的。作为本论文的后续工作,在误报代价比例的确定上还有待进一步研究。

参考文献

- [1] Sahani M, Dumais S, Heckerman D, Horvitz E. A Bayesian Approach to Filtering Junk Email. In Learning for Text Categorization [A]: the AAAI Workshop. Proceedings of AAAI-98 Workshop on Learning for Text Categorization[C]. Madison Wisconsin. May. 1998. Pages 55 - 62.
- [2] Androutsopoulos I, Koutsias J, Chandrinos K, et al. An evaluation of naive Bayesian Anti-Spam filtering [A]: Potamias G, Moustakis and van Someren M (eds.). Proceedings of the workshop on Machine Learning in the New Information Age[C]. Barcelona. 2000. P 9-17.
- [3] Gordon Cormack, Thomas Lynam. A Study of Supervised Spam Detection applied to Eight Months of Personal E-Mail [A]: Proceedings of Conference on Email and Anti-Spam (CEAS) 2004[C], Mountain View, CA, July 30 and 31, 2004.
- [4] Andrew P. Bradley. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms.
- [5] Tom Fawcett. Evaluating Classifiers for Changing Environment.
http://home.comcast.net/~tom.fawcett/public_html/ROCCH/ROCCH-talk-public.html
- [6] Foster Provost, Tom Fawcett. Robust Classification for Imprecise Environments. Machine Learning, 42,203-231,2001