贝叶斯网络在过滤垃圾邮件算法中的应用研究

崔 超[1], 杨 威[2], 张宪忠[1], 张志军[3]

(1. 齐齐哈尔大学 应用技术学院,161006 黑龙江 齐齐哈尔, cv63@163.com; 2. 齐齐哈尔市信息中心, 161006 黑龙江 齐齐哈尔;
3. 齐齐哈尔市信息技术研究所 161006 黑龙江 齐齐哈尔)

摘 要: 为在用户数据流中删除垃圾邮件,研究了具有自我学习能力的自适应邮件过滤系统. 在对正常和垃圾2类邮件误分类成本分析的基础上,利用概率性的学习方法创建满足过滤任务需要的过滤器,且讨论使用邮件域名特征变量进行特定邮件过滤并设计了过滤器,最后对实际邮件组进行操作,验证了算法的可靠性.

关键词: 贝叶斯理论方法;概率;特征变量;邮件过滤

中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2011)11-0145-04

Bayesian application study on arithmetic for filtering junk e-mail

CUI Chao[1], YANG Wei[2], ZHANG Xian-zhong[1], ZHANG Zhi-jun[3]

(1. College of Applied Technology, Qiqihar University, 161006 Qiqihar, China, cv63@163.com;
2. Qiqihar Information Center, 161006 Qiqihar, China; 3. Qiqihar information Technology Institute, 161006 Qiqihar, China)

Abstract: To delete spams from User data stream, an adaptive self-learning spam filtering system has been studied and presented. On the basis of cost analysis mistakenly classified of normal and spam e-mail, a filter to meet the requirements of filtering tasks with learning methods of probabilistic is created, and the use of mail domain name characteristic variables for a particular e-mail filtering is studied. Finally, experiments verify the reliability of the algorithm.

Key words: bayesian theory; probability; feature-variable; filtering e-mail

早期对邮件的过滤只是对邮件正文进行过滤,逻辑过滤规则只是机械地使用二进制决策来决定取舍,而这种规则是有局限性的. 研究能根据过滤环境特征变量[1]变化进行自我调整,并能进行自我学习的邮件过滤系统是一种解决办法. 由于垃圾邮件特征变化的多样性和实时性,逻辑过滤规则的优化对系统提出了更高的要求. 理想的过滤系统应具备随过滤目标对象的特征变化而自动调整过滤规则的能力,对邮件进行分析,能直接从系统邮件缓冲区读取数据并进行学习,个性化地为每个用户架构有自身特征的过滤系统,过滤系统完成对目标的操作. 依据贝叶斯定理设计的分类器,本文设计一个功能模块,并制订了一个可行的过滤方案,并对邮件错误成本进行定量分析,完成垃圾邮件过滤[2].

收稿日期: 2010-11-10.
作者简介: 崔 超(1961—),男,教授.

1 贝叶斯模型

采用贝叶斯定理的分类器,在自动生成概率性的文本分类模型的过程中,将贝叶斯定理应用到分类技术的算法中,利用贝叶斯定理建模,既能解决传统文本过滤问题,又可以在贝叶斯过滤器中引入具体的特征变量,将邮件域名信息同要完成的指定任务结合起来,联合使用研究模型和系统损失模型,对目标文件作出是否为垃圾邮件的判定.

1.1 贝叶斯网络

为完成垃圾邮件的检测需要建立一个概率性的分类器,利用贝叶斯定理的网络形式和其直接的、非循环形式的对应图表达一个紧密的几率分布,图1中任一个随机变量 x_i 都用一个节点来表示,两节点间的线段表示由父、子节点分别确定的

两变量间概率相关性. 网络结构表明网络中每一个节点 x_i 都有条件地独立于由其父节点给出的其他节点. 为描述几率分布,将网络中的节点 x_i 组成一个条件几率图,图 1 用来说明对应的概率分布.

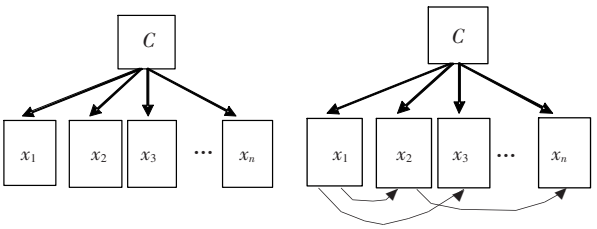
对每个过滤任务,一个贝叶斯过滤器就是一个贝叶斯网络,网络中的节点 C 表示级别,每个节点 x_i 用于表示一个特征变量[3],将一个指定的范例 X(将 $x_i(i = 1,2,3,\cdots)$)的值分配给对应的特征变量,网络允许为每一个级别 C_K 计算几率 $P(C = c_K \mid X = x)$ 为

$$P(C = c_k \mid X = x) = \frac{P(X = x \mid C = c_k)P(C = c_k)}{P(X = x)}. \quad (1)$$

式中:变量 $P(C = c_k \mid X = x)$ 在未加上域名变量前无实际意义,式(1) 中的每一个特征变量 x_i 都有条件地独立于其他任一特征变量,在给定级别变量 C 后,计算为

$$P(X = x \mid C = c_k) = \prod P(X_i = X_i \mid C = c_k). \quad (2)$$

为降低贝叶斯过滤器对特征变量 x_i 的限制,过滤方法允许特征变量 x_i 间有限形式的相关. 图 1(a) 展示了贝叶斯过滤器在具有同其他过滤器相同的上述特征变量 x_i 间有限形式时的系统结构;图 1(b) 则表现了特征变量间具备有限相关性的复杂贝叶斯分类器的系统结构.



(a)特征变量 x_i 间有限形式时的系统结构　(b)有限相关性的复杂贝叶斯分类器的系统结构

图 1　同贝叶斯过滤器相对应的贝叶斯网络

1.2　变量空间模型

采用贝叶斯定理的垃圾邮件过滤系统在对邮件文本进行分类时,要依据相应的邮件信息形成特征变量. 用一个变量空间模型将变量空间维数的定义同整个邮件库中给定的词对应起来,每个具体的单词用一个二进制变量表示,表明单词出现与否.

2　邮件域名的作用

具体邮件过滤时,首先要考虑邮件正文内容在判定本文是否为垃圾时所起的作用,同时还要考虑邮件的其他特征变量,例如一些有诱惑性的短语或特殊符号的非常规使用,如“快速致富绝招”、“!!!!”等,此外,邮件中还经常包含许多非正文属性的特征,如邮件域名信息等,对垃圾邮件判定也非常有用.

2.1　使用特征变量的贝叶斯过滤器

将有具体邮件特点的特征变量引入到贝叶斯过滤器中是很简单的,把邮件提供的各种原始信息一律无变化地引入到分类模型中,不需要对算式做任何修改,但需要对在分类模型中使用的具体变量的表达形式进行分析.

首先完成对邮件外部表述文字有无特定短语匹配的检测,如“免费”、“只是为了 money”等,在实验中引入了一个有 40 个特征短语变量的集合,使用概率性的过滤方案,在当前规则限定下,利用用户手动建立起来的,由短语组成的特征变量,快速完成对当前邮件的过滤.

其次再考虑邮件域名类型等有关域的特征变量,常见的垃圾邮件经常隐藏本身域名,如邮件是一个具体用户发出的,对一封垃圾邮件非文本特征标识的判定是很容易实现的. 一般垃圾邮件是不带附件的,接收的时间也常在夜间,标题中出现非字母和数字的字符所占的百分比也是判定垃圾邮件的重要标识之一,如“￥￥￥人民币￥￥￥”等,类似上述判定依据在概率性的分类器中效果良好,而在基于规则的系统中误差率则较高,如图 2所示,在垃圾邮件和正常邮件的标题中所含的非字母和数字的分布有明显的差异[4].

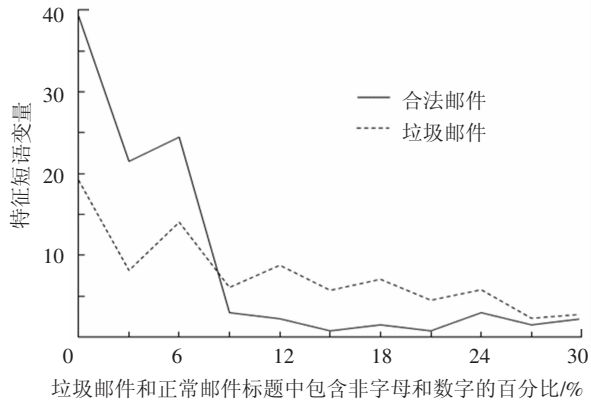


图 2　垃圾邮件和正常邮件标题中包含非字母和数字的百分率变化曲线

利用标题中包含的字母和数字占标题的百分率和域名方面其他一些特征变量对邮件的判定经常是有效的,在对邮件信息进行垃圾判定时,概率性的分类器可以使用上述特征变量作为依据对目标邮件作出判定,在实验中引入了 40 个特征变量,包括域和编辑的非短语等.

3　垃圾邮件总体划分和实验室测试

为确认研究结果，对垃圾邮件进行过滤实验，以信件正文和特征变量为对象进行基本分类操作，对性能进行了测定，同时对过滤器在操作环境设定领域的学习效率进行了评估.

3.1　建立实验用样本分类器

由于特征变量空间广阔，要进行变量的选择，减少维数可以实现对评估系统参数模型变化结果进行精确的控制，降低过滤器对独立变量的干扰. 实验首先选择 50 个特征变量组成一个特征变量组来建立分类器，它包括了以单词、编辑组成的短语和具体域变量，排除文中出现次数 < 4 次的单词，再计算每个特征变量 x_i 和级别变量 C 间的信息影响因子 MI 为

$$MI(X_i;C) = \sum_{X_i=x_i,C=c} P(X_i,C)\log\frac{P(X_i,C)}{P(X_i)P(c)}. \quad (3)$$

尽管实验中 50 个变量不是最佳值，开始阶段的多个实验还是表明了由 50 个特征变量提供的实验结果是可靠的.

3.2　添加域名特征变量

为判定使用特征变量的效率，利用为邮件过滤而编辑建立的特征变量，以一个有 1 578 封垃圾邮件，211 封合法邮件的邮件库为样本，将邮件库分成包括 1 538 封邮件组成的训练组和由 251 封邮件组成的测试组，以邮件标题、邮件体为实验对象，采用基于单词的标记作为特征变量，然后用 35 个特征变量（为完成这项任务而专门收集的）去增加特征变量的数量，再用 20 个非文本、有具体域特点的变量来扩大变量组，将测试组同变量组相联，完成变量选择和贝叶斯分类器的建立，最后实现对测试组垃圾的分类.

3.3　邮件分类敏感成本理论和实际测试结果

设邮件为垃圾邮件可能性 >98% 时，只能被划分为垃圾邮件. 实验中得到的一组数值显示了 98% 的基本值是满足实验任务要求的，根据变量结构的变化对垃圾和正常邮件进行再调用和分类的有关数据统计如表 1 所示.

表 1　采用不同变量组的分类结果　%

变量结构	垃圾邮件		合法邮件	
	准确度	重调率	准确率	重调率
单词	98.5	95.3	87.9	93.6
单词 + 短语	98.8	95.5	88.9	94.9
单词 + 短语 + 域变量	100.0	97.6	96.7	100.0

垃圾邮件准确度是真实垃圾邮件数和本组邮件数量的比值[5]；垃圾邮件重调率表示在经过测试分类器分类以后的测试组中包含垃圾邮件数量和本组邮件数量的比例，合法邮件的重调率算法亦同此. 为避免将合法邮件误判为垃圾邮件，在只采用单词类变量进行过滤的基础上，短语类信息增加对过滤精度有较小提高，而域变量的微小增加都会对过滤效果产生重大的影响，如表 1 所示.

在采用不同变量组的情况下，图 3 给出了垃圾邮件准确率和重调率曲线，为更清晰地表示曲线的变化，重调率取 0.85 ~1.0，通过曲线可以看出，随着变量组合种类的增加，整个曲线不断上移，同只采用单词模式的过滤手段相比，包括域变量的过滤精度最高.

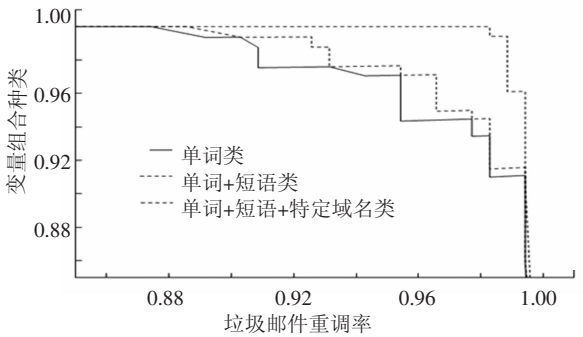


图 3　垃圾邮件准确率和重调率曲线

表明对于指定文本过滤问题，域名信息在过滤过程中对提高过滤效果有非常大的帮助.

3.4　实际垃圾邮件的分类

垃圾邮件常分为包含黄色信息和所谓“目标信息”2 类，根据实验要求，用合法邮件、黄色邮件和其他垃圾邮件为匹配模型，创建邮件过滤器.

利用 2 类不同的垃圾邮件可使学习型分类器有更多的自由度，为实验需要建立一个由 972 个垃圾邮件，211 个合法邮件组成的一个邮件组，并暂时将其分成一个有 916 个信息组成的训练组，267 个信息组成的测试组. 为测试不同垃圾邮件分类的效率，用 2 种不同的方法来标识邮件组，首先完成对每个邮件是否为垃圾邮件的标记，然后再对分类出的垃圾邮件进行内部划分.

表 2　依据垃圾邮件内部类型划分的邮件分类结果　%

信息分类	垃圾邮件		合法邮件	
	准确率	重调率	准确率	重调率
合法邮件和垃圾邮件	98.9	94.3	87.2	97.5
合法邮件、黄色和其他垃圾邮件	95.6	77.3	61.2	90.9

根据研究分类器的需要，在实验中将短语和域名类变量引入变量组，通过变量选择生成 50 个

变量,再设定 98% 为垃圾邮件分类极限,以反映实验中出现错误的不对称成本. 为过滤出垃圾邮件,可将测试组分成黄色和其他垃圾邮件 2 类,完成了对“垃圾”邮件初步的操作,数据显示基于 3 种选择的垃圾邮件分类的效果要优于 2 种的.

由表 2 可以发现,垃圾邮件内部种类的划分不提高邮件分类效率,反而使精度降低,在给定的准确率和重调率的范围内,图 4 曲线也反映了这点.

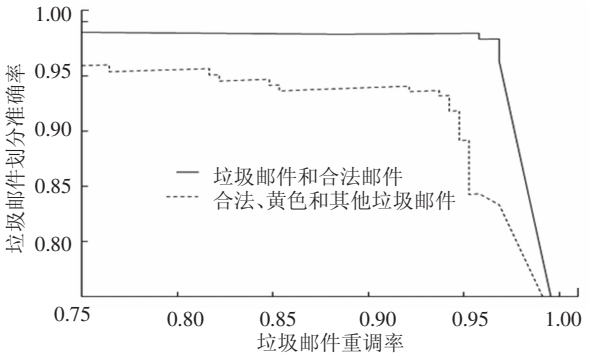


图 4 垃圾邮件准确率和重调率曲线

图 3 垃圾邮件内部种类划分精度原因之一是同 3 种分类结果的方案相比较,在 2 种分类结果的方案中有明显指示效应的变量在前者的指示作用非常有限,更重要的一点是随着自由度的增加,同每类结果对应的、基于分类的变化度在增大,所以同 2 种分类结果的方案相比,3 种分类结果的方案在操作中要完成更多的变量匹配,所以设定参数的变化对成熟的分类器性能的影响非常大.

4 仿真过滤

为测试实验过滤器的最终效率,提供一个由 2 600 个邮件组成的集合,并已完成它的垃圾邮件的划分,过滤近 10 d 某用户所收到的 224 信件,依据测试结果,其中有 46 封为垃圾邮件,用表 1 中的变量结构来构建学习型过滤器,以 98% 作为垃圾邮件划分精度的极限. 实验结果如表 3 所示.

表 3 采用最新训练组过滤效率表

邮件分类	已分类的“垃圾”邮件	已分类的“合法”邮件	总计
真实的垃圾邮件	37	9	46
真实的合法邮件	3	175	178
总计	40	184	224

在 3 封被误判的垃圾邮件中,有 1 封包含有部分主观的内容,其余内容均为垃圾内容,在进一步的研究中可划定为垃圾邮件;其他 2 封邮件包含来自服务器匿名投递的产品信息,非用户所需要的,所以从根本上来讲,表 3 中的划分结果对邮件用户无损失,且有比较高的精度的精度,基于上述研究中的结果绘制了效果曲线图,如图 4 所示.

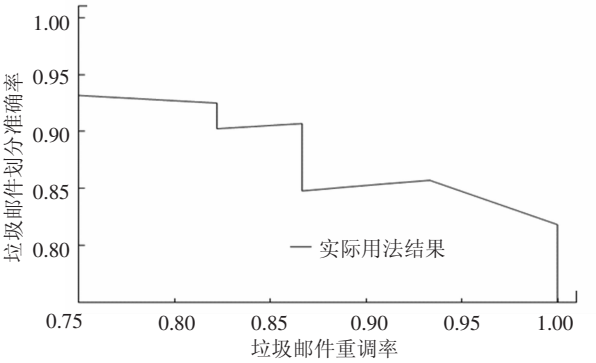


图 5 仿真过滤效果曲线图

5 结 语

一个具备自我学习能力的过滤系统能够排除绝大部分垃圾邮件,邮件正文、人工短语尤其是域名变量,在具体邮件过滤中对准确性的提高作用很大,用学习型分类模式去完善贝叶斯网络,对于建立高集成度、有域名变量特点的过滤器有指导意义,实际应用中的曲线图(图 4)证明了系统的可用性.

参考文献:

[1] COHE W W. Learning rule that classify E-Mail[C]// Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access. California: AAAI Press, 1996: 18 – 25.

[2] HALL R J. How to avoid unwanted email[J]. Communication of ACM, 1998: 41(3): 88 – 95.

[3] KUSHMERICK N. Learning to remove internet advertisements[C]//Proceedings of the 3rd Annual Conference on Autonomous Agents. New York, NY: ACM, 1999: 175 – 181.

[4] 段宏斌, 张健. 改进的 Naive Bayes 技术在反垃圾邮件系统中的应用[J]. 西北大学学报(自然科学版), 2006, 36(5): 737 – 740.

[5] 王宁, 张建忠. 基于改进贝叶斯模型的中文邮件分类算法[J]. 计算机工程与应用, 2006, 42(31): 97 – 100.

[6] GRAHAM P. Better Bayesian Filtering [EB/OL]. [2003-01-10]. http://http://paulgraham.com/better.html.

[7] RICHARD O D. 贝叶斯决策论模式分类 [M]. 北京: 机械工业出版社, 2003.

(编辑 张 红)