

基于朴素贝叶斯模型的邮件过滤技术

杨 赫¹, 孙广路^{1,2}, 何勇军^{1,2}

(1. 哈尔滨理工大学 计算机科学与技术学院, 黑龙江 哈尔滨 150080;

2. 哈尔滨理工大学 信息安全与智能技术研究中心, 黑龙江 哈尔滨 150080)

摘 要: 针对朴素贝叶斯算法应用于反垃圾邮件过滤时,其有效性十分依赖于对邮件内容的有效建模,而邮件内容建模方面研究尚不成熟限制了贝叶斯方法在垃圾邮件过滤中的性能. 采用了三种概率分布对邮件内容进行建模,据此提出了3种概率分布下的朴素贝叶斯算法. 为了提高训练效率,算法采用了一种增量式的垃圾邮件过滤方法. 在 trec05p-1、trec06p 两个公开数据集上对这3种贝叶斯算法进行了实验对比,分析出三种贝叶斯分布的适用范围. 从不同分布的邮件内容建模角度出发,为过滤垃圾邮件的方法选择提供了有效依据.

关键词: 邮件过滤; 朴素贝叶斯; 机器学习

中图分类号: TM391.1 **文献标志码:** A **文章编号:** 1007-2683(2014)01-0049-05

SPAM Filtering with Naive Bayes

YANG He¹, SUN Guang-lu^{1,2}, HE Yong-jun^{1,2}

(1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China;

2. Research Center of Information Security and Intelligent Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: The effectiveness of Naive Bayes in spam filtering depends on the modelling of the mail contents. However, mail content modelling is not mature, which limits the performance of Bayesian method in spam filtering. This paper presents three kinds of probability distribution to model email content, and proposes three Naïve Bayes algorithms based on different probability distributions. To improve training efficiency, the incremental training algorithm is utilized in the experimental procedure. Experiments on trec06p and trec05p-1 show that the three proposed algorithms can achieve good performance in different sceneries. Such a finding also provides effective basis for the selection of the filtering methods.

Key words: e-mail filtering; naive bayes; machine learning

0 引 言

垃圾邮件泛滥严重影响了电子邮件正常使用. 垃圾邮件消耗大量网络资源,损害用户利益,还会被

一些别有用心的人用来散播虚假消息,危害社会安定. 很多机器学习算法已经应用于反垃圾邮件过滤系统中,其中 Boosting 和 SVM^[1] 等算法在文本分类中有好的表现. 在商用和开源垃圾邮件过滤领域中,朴素贝叶斯分类器(naive bayes, NB)应用十分广

收稿日期: 2013-05-05

基金项目: 黑龙江省普通高等院校新世纪优秀人才培养计划(1155-ncet-008); 教育部人文社科项目(11YJC740048); 黑龙江省教育科学规划课题(GBC1211062); 黑龙江省高等教育教学改革项目(2011-NP33).

作者简介: 杨 赫(1987—),男,硕士研究生;

孙广路(1979—),男,博士,教授, E-mail: guanglu.sun@gmail.com;

何勇军(1980—),男,博士,副教授.

泛.与一些复杂算法的分类器相比,贝叶斯分类器的优势在于其易实现、高准确性,且具备较低的线性计算复杂度.目前有三种比较常用的朴素贝叶斯算法,分别是基于伯努力多变量贝叶斯,基于多项式词频贝叶斯,基于布尔类型属性多项式贝叶斯.本文主要研究三种贝叶斯模型用于邮件过滤时的分类效果,以及分析三种贝叶斯模型的使用范围.

本文组织形式如下:第1节是相关工作,简要介绍目前垃圾邮件过滤领域的方法和模型,第2节提出贝叶斯算法和基于贝叶斯的垃圾邮件过滤模型,第3节设计并实现了三种贝叶斯分类器,第4节是采用国际公开数据集进行实验,并对实验结果做了分析与评估,第5节对全文进行了总结.

1 相关工作

垃圾邮件过滤技术主要分为三种,基于黑白名单的过滤技术,基于规则或特征词匹配的邮件过滤技术,基于机器学习的邮件过滤技术.早期过滤系统主要采用基于黑白名单的过滤技术,根据发件人信息、源IP地址信息,主题关键字等过滤邮件.这种方法虽然运行速度快,实现简单,但准确率不高.随着垃圾邮件发送技术的更新,垃圾邮件的特征大部分体现在邮件的内容上,因此垃圾邮件发送者发送的垃圾邮件容易躲避基于黑白名单过滤技术的检测.研究人员采用分析邮件内容来判别垃圾邮件.比如,建立垃圾邮件特征词的词库,将邮件内容分词处理后和词库中特征词进行匹配.若匹配成功的词达到一定数量,判断为垃圾邮件.该技术实用性强,词库便于更新,称为基于规则或特征词匹配的垃圾邮件过滤技术.但该方法的缺点也很明显.由于垃圾邮件数量巨大,种类繁多,导致规则库或词库变得非常庞大,导致系统匹配速度变慢、过滤规则冲突等问题^[1].针对内容的邮件过滤问题,机器学习算法在其中扮演着重要的角色,基于机器学习的方法以其成本较低,准确率高,已成为目前广泛使用的垃圾邮件过滤方法.该方法针对邮件内容进行过滤,从邮件的内容获取特征,并通过这些特征和对应的标注信息来在线训练垃圾邮件和正常邮件的模型,进而优化过滤器^[2],实现对垃圾邮件的过滤.

为了测试目前研究的方法在实际网络邮件环境中的性能,使垃圾邮件研究者们更好地合作交流,国内外举办了多次邮件评测及会议,积累了完整的评价指标和大量的测试数据集,有效促进垃圾邮件过

滤技术的发展.例如:TREC(text retrieval conference)^[3-5]会议于2005、2006、2007年分别举办了垃圾邮件评测、CEAS(conference on email and anti-spam)^[6]会议于2007年和2008年举行了垃圾邮件评测和国内的由华南理工大学组织的SEWM(search engine and web mining)会议于2007、2008、2010以及2011年分别举行了垃圾邮件评测.

基于机器学习的垃圾邮件过滤方法是往往采用在线学习方法^[7]进行系统实现和测试.机器学习方法运用在垃圾邮件过滤器中有两类模型:生成模型^[8],如朴素贝叶斯^[9];判别模型,如SVM^[10-11]和逻辑回归模型^[12].国外的垃圾邮件过滤技术研究起步较早并取得了许多重要成果.Sahami^[13]在1998年第一次提出了将朴素贝叶斯方法运用于解决垃圾邮件过滤问题,为贝叶斯算法在垃圾邮件过滤领域的应用打下坚实基础.Metsis^[14]在2006年提出了五种贝叶斯改进模型,分别阐述并且对邮件分类效果进行说明分析.Ciltik^[15]提出基于n-gram特征提取方法的朴素贝叶斯分类器,这种分类器只提取每封被测邮件的前一部分特征,极大程度上提升了邮件过滤器的准确度和效率.Kim^[16]提出根据邮件中的URL链接来判断邮件类别的垃圾邮件过滤方法.Schneider^[17]提出了两种贝叶斯改进模型,提出针对多项式模型的特征选择方法.Graham提出适合在线特征选取的模型并且对基于词的特征选择做了系统深入的研究,对贝叶斯模型做了一定改进,从而使贝叶斯模型实现简单,计算效率高,而且在不同邮件数据集上均能取得很好的分类效果.

2 贝叶斯算法和基于贝叶斯的垃圾邮件过滤模型

2.1 贝叶斯模型

贝叶斯定理是通过对某一事件过去发生概率情况来推断当前这一事件发生的概率的计算方法,它可以被形式化表述为:设实验 E 的样本空间为 M , A 为 E 一个事件, $B_1, B_2, \dots, B_n$ 为 M 的一个划分, $P(A) > 0$, $P(B_i) > 0, i = 1, 2, \dots, n$,则有

$$P(B_i | A) = \frac{P(A | B_i) P(B_i)}{\sum_{j=1}^n P(A | B_j) P(B_j)} \quad (1)$$

2.2 基于贝叶斯的垃圾邮件过滤器模型

垃圾邮件过滤问题可以看成是垃圾邮件文本的二值分类问题,实际上是贝叶斯分类器的一种具体

实现方式. 它的优点是模型所需估计的参数很少, 对缺失数据不太敏感, 算法也比较简单, 缺点是模型假设属性之间相互独立, 这个假设在实际应用中往往是不成立的, 这给模型的正确分类带来了一定影响. 每一封邮件样本, 使用向量空间对其进行数学形式化描述, $d = (x_1, w_1; \dots; x_n, w_n)$, x_i 是邮件选取特征项, w_i 是 x_i 的权值, w_i 取值为 0 或 1, 邮件过滤器的任务就是计算待分类邮件是垃圾邮件的概率和正常邮件的概率, 判断结果. 邮件判定中的邮件类别: $C \in \{\text{Spam}, \text{Ham}\}$. 根据贝叶斯理论, 一封邮件属于垃圾邮件的概率计算公式为

$$P(C = \text{Spam} | d_x) = \frac{P(d_x | C = \text{Spam}) P(C = \text{Spam})}{P(d_x)} \quad (2)$$

同理得 $P(C = \text{Ham} | d_x)$. $P(C = \text{Spam} | d_x)$ 大于 $P(C = \text{Ham} | d_x)$, 或两者的商值大于指定阈值则可以判定这封邮件为垃圾邮件. $P(C = \text{Spam})$ 指选取样本中垃圾邮件的先验概率, $P(d_x | C = \text{Spam})$ 指一封垃圾邮件中所有特征 $(x_1, x_2, \dots, x_n)$ 同时出现的概率.

3 3 种贝叶斯分类器

3.1 基于伯努力的多变量贝叶斯

设邮件的特征向量是 $F = (t_1, \dots, t_m)$, t_i 是对每个特征给出的一个标记, 基于伯努力的贝叶斯将每一个邮件样本看作是一个经过去重处理的标记集合 $X = (t_1, \dots, t_m)$, 其中每个 t_i 是一个二元变量, 其取值是 0 或 1, 取 0 时表示 t_i 特征没有出现在样本 d 中, 取 1 时表示 t_i 特征出现在样本 d 中. 判断一个邮件样本 d 是属于类别 C 可以看作是 m 次伯努力实验的结果, 每次实验我们判断特征 t_i 是否出现在样本 d 中. 对于某一次伯努力实验 i 的概率值 $p(t_i | C)$, 基于伯努力的多变量贝叶斯算法做了一个附加的假设条件, 每个伯努力实验的结果是相互独立的^[8]. 这是一个理想化的假设, 因为单词之间出现时间对分类的判断是非独立的. 相似的假设在所有的贝叶斯模型中都被应用. 尽管在大多数的案例中这些假设过于简单, 但是这些贝叶斯算法在很多的分类任务中仍有非常好的表现, 即

$$P(\vec{X} | C) = \prod_{i=1}^m p(t_i | C)^{x_i} (1 - p(t_i | C))^{(1-x_i)} \quad (3)$$

$$P(t | c) = \frac{1 + M_{t,c}}{k + M_c} \quad (4)$$

其中用拉普拉斯算子进行平滑处理, $M_{t,c}$ 是有属于类别 c 有特征 t_i 的文本数, k 为类别数量, 对于二值分类时, $k=2$, M_c 是属于类别 c 的文本数.

为了防止下溢, 在实际使用中, 一般对概率值采用对数形式. 邮件样本分类标准 $T=0$, 判断一封邮件样本是垃圾邮件, 即

$$\log \frac{P(C = \text{Spam}) P(d_x | C = \text{Spam})}{P(C = \text{Ham}) P(d_x | C = \text{Ham})} > T \quad (5)$$

3.2 基于词频的多项式贝叶斯

基于词频的多项式贝叶斯考虑词频的因素, 每个特征词出现看作一个事件, 邮件样本看作是这些独立事件的集合. $P(d_x | C_j)$ 符合多项式分布, 指一封已知类别的邮件中所有特征 $(x_1, x_2, \dots, x_n)$ 同时出现的概率. 文本 d_x 包含的所有词的总数为 $|d|$, 邮件判定中的邮件类别: $C \in \{\text{Spam}, \text{Ham}\}$, $|d|$ 不受类别 C 的影响, 这是一个简单化的假设, 因为实际邮件情况中接收一封较长正常邮件的概率要比收到一封同样长度的垃圾邮件的概率要大的多, m 为特征词的数量, N_{t_i} 表示特征词 t_i 在邮件样本 d_x 中出现的总次数, 则

$$P(d_x | C_j) = P(d_x) |d|! \prod_{i=1}^m \frac{P(t_i | C_j)^{N_{t_i}}}{N_{t_i}!} \quad (6)$$

垃圾邮件过滤中的具体实现为

$$P(d_x | C = \text{Spam}) = P(d_x) |d_x|! \prod_{i=1}^m \frac{P(t_i | C = \text{Spam})^{N_{t_i}}}{N_{t_i}!} \quad (7)$$

$$P(t_i | C = \text{Spam}) = \frac{M_{t_i, \text{Spam}}}{M_s} \quad (8)$$

其中, $P(d_x | C = \text{Ham})$ 和 $P(t_i | C = \text{Ham})$ 同理求得, M_s 是所有垃圾邮件文本中特征词总数, $M_{t_i, \text{Spam}}$ 是所有垃圾邮件文本中特征项 t_i 出现的次数, 为了避免分母为 0, 需对上式进行拉普拉斯平滑处理^[9], 其中 k 为类别数量, 对于二值分类时 $k=2$.

$$P(t_i | C = \text{Spam}) = \frac{1 + M_{t_i, \text{Spam}}}{k + M_s} \quad (9)$$

根据全概率公式

$$P(C = \text{Spam} | d_x) = \frac{P(C = \text{Spam}) P(d_x | C = \text{Spam})}{P(C = \text{Spam}) P(d_x | C = \text{Spam}) + P(C = \text{Ham}) P(d_x | C = \text{Ham})} \quad (10)$$

其中,

$$P(d_x | C = \text{Ham}) = \prod_{i=1}^m P(t_i | C = \text{Ham})^{N_{t_i}} \quad (11)$$

$$P(d_x | C = \text{Spam}) = \prod_{i=1}^m P(t_i | C = \text{Spam})^{N_{t_i}} \quad (12)$$

$P(C = \text{Ham} | d_x)$ 同理求得, $P(C = \text{Spam} | d_x)$ 大于 $P(C = \text{Ham} | d_x)$ 或两者的商值大于指定阈值则可以判定这封邮件为垃圾邮件。

3.3 基于布尔属性的多项式贝叶斯

基于布尔属性的多项式贝叶斯和基于词频的贝叶斯多项式十分类似,都包含条件概率 $p(t_i | C)$,只是属性值为布尔类型而已。这种方法和基于伯努力贝叶斯的不同是对特征 $t_i = 0$ 时,这个特征项就不加入对条件概率的计算,此时有

$$P(\vec{X} | C) = \prod_{i=1}^m p(t_i | C)^{x_i} \quad (13)$$

而且对 $p(t_i | C)$ 的拉普拉斯平滑处理方法不同。

对于基于布尔属性的贝叶斯算法在给出较少的特征关联信息的情况下,反而比基于词频的贝叶斯方法表现更好。目前已经证明在假设邮件长度和邮件类别无关的前提下,基于词频属性的贝叶斯等价于基于泊松分布属性的贝叶斯。因此,当词频属性不服从泊松分布时,基于布尔属性的贝叶斯要比基于词频属性的贝叶斯的表现要好。

4 实验结果与分析

4.1 评价体系

共使用 4 个评测指标,分别是:

- 1) 正常邮件错判率 ham%;
- 2) 垃圾邮件错判率 spam%;
- 3) 逻辑平均错误率 lam%;

$$\text{lam\%} = \text{logit}^{-1} \left(\frac{\text{logit}(\text{ham\%}) + \text{logit}(\text{sm\%})}{2} \right) \quad (14)$$

其中

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right) \quad (15)$$

- 4) ROC 曲线上面部分的面积 $(1 - \text{ROCA})\%$ 。

以 ham% 为横坐标,以 spam% 为纵坐标,做 ROC 曲线,对于 trec05p-1 的 ROC 曲线如图 1 所示,对于 trec06p 的 ROC 曲线如图 2 所示,求 ROC 曲线上方面积,即 $(1 - \text{ROCA})\%$ 值,其值越小说明邮件分类效果越好^[2]。

4.2 评测结果

用 3 种贝叶斯模型分别对 trec05p-1、trec06p 进行分类,并对结果文件进行评测。对于正常邮件判

错率,评测结果如表 1 所示,对于垃圾邮件判错率如表 2 所示,对于 $(1 - \text{ROCA})\%$ 值评测结果如表 3 所示。

表 1 ham% 实验结果

ham%	Multi-variate Bernoulli NB/%	Multinomial NB,TF/%	Multinomial NB,Boolean/%
trec05p-1	10.58	3.06	4.41
trec06p	9.16	1.91	1.82

表 2 spam% 实验结果

spam%	Multi-variate Bernoulli NB/%	Multinomial NB,TF/%	Multinomial NB,Boolean/%
trec05p-1	19.39	11.83	16.02
trec06p	6.16	11.60	10.53

表 3 1-ROCA% 实验结果

$(1 - \text{ROCA})\%$	Multi-variate Bernoulli NB/%	Multinomial NB,TF/%	Multinomial NB,Boolean/%
trec05p-1	8.095 6	2.668 6	4.200 5
trec06p	4.708 7	4.174 2	1.809 3

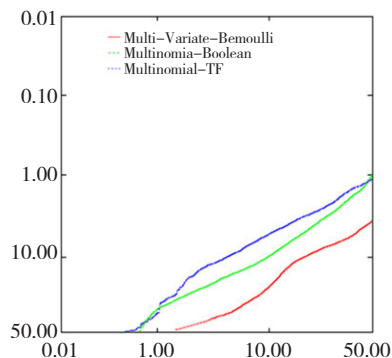


图 1 trec05p-1 ROC 曲线

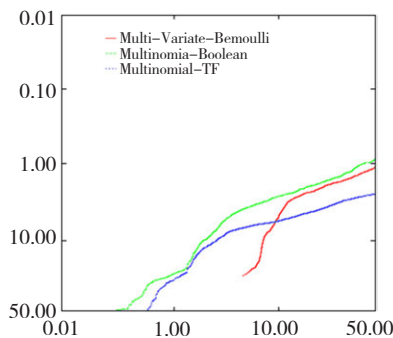


图 2 trec06p ROC 曲线

4.3 评测结果分析

从测试实验的评测结果以及生成的 ROC 曲线可以得出结论:对于正常邮件的判错率,基于词频的

多项式贝叶斯模型过滤效果最好,对于垃圾邮件的判错率,基于词频的多项式贝叶斯模型平均过滤效果更好. 对于 $(1 - \text{ROCA})\%$ 值,当词频属性服从泊松分布时基于词频的多项式贝叶斯模型总体上的邮件过滤效果最好,当词频属性不服从泊松分布时基于布尔属性的贝叶斯模型过滤效果要更优.

5 结 语

本文重点介绍了三种贝叶斯分类模型: 基于伯努力多变量贝叶斯模型,基于词频的多项式贝叶斯模型,基于布尔属性的多项式贝叶斯模型,并分别用三种贝叶斯模型建立垃圾邮件过滤器. 基于数据集 trec06p、trec05p-1 进行评测,并对三种贝叶斯分类器的过滤结果进行分析.

参 考 文 献:

- [1] 王斌,潘文锋. 基于内容的垃圾邮件过滤技术综述 [J]. 中文信息学报,2005,19(5): 1-10.
- [2] 刘海韬,阳洁. 云计算平台下一种新型反垃圾邮件系统的研究 [J]. 中南大学学报: 自然科学版,2013,44(5): 20-31.
- [3] SUN Guanglu, SUN Hongyue, MA Yingcai, et al. Spam Filtering: Online naïve Bayes Based on TONE [J] //ZTE Communications, 2013: 51-54.
- [4] CORMACK G, LYNAM T. TREC 2005 Spam Track Overview [C] //Proceedings of the Fourteenth text Retrieval Conference Proceedings. US: Text REtrieval Conference, 2005: 123-130.
- [5] CORMACK G. TREC 2006 Spam Track Overview [C] //Proceedings of the Fifteenth Text Retrieval Conference Proceedings, US: Text REtrieval Conference, 2006: 117-128.
- [6] SCULLEY D. Online Active Learning Methods for Fast Label-Efficient Spam Filtering [C] //CEAS, 2007: 1-4.
- [7] 孙广路,齐浩亮. 基于在线顺序逻辑回归的垃圾邮件过滤 [J]. 清华大学学报,2013,53(5): 734-740.
- [8] SCULLEY D. Practical Learning From One-sided Feedback [C] //Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007: 609-618.
- [9] CHEN C, TIAN Y, ZHANG C. Spam Filtering with Several Novel Bayesian Classifiers [C] //Pattern Recognition, 2008, ICPR, 19th International Conference on IEEE, 2008: 1-4.
- [10] 全亮亮,吴卫东. 基于支持向量机和贝叶斯分类的异常检测模型 [J]. 计算机应用,2012,32(6): 1632-1635.
- [11] TONG S, KOLLER D. Support Vector Machine Active Learning with Applications to Text Classification [J]. The Journal of Machine Learning Research, 2002, 2: 45-66.
- [12] GOODMAN J, YIH W. Online Discriminative Spam Filter Training [C] //CEAS, 2006: 1-4.
- [13] SAHAMI M, DUMAIS S, HECHERMAN D. A Bayesian Approach to Filtering Junk e-mail [C] //Learning for Text Categorization: Papers from the 1998 Workshop, 1998, 62: 98-105.
- [14] METSIS V, PALIOURAS G. Spam Filtering with Naïve Bayes-which Naïve Bayes [C] //Third Conference on Email and Anti-Spam, US: 2006: 103-112.
- [15] CILTIK A, GUNGOR T. Time-efficient Spam E-mail Filtering Using N-gram Models [J]. Pattern Recognition Letters, 2008, 29(1): 19-33.
- [16] KIM J, CHUNG K, CHOI K. Spam Filtering with Dynamically Updated URL Statistics [C] //US: IEEE Security and Privacy, 2007, 5(4): 33-39.
- [17] SCHNEIDER M. A Comparison of Event Models for Naive Bayes Anti-spam e-mail Filtering [C] //Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003: 307-314.
- [18] SCULLEY D. Advances in Online Learning-based Spam Filtering [M]. ProQuest, 2008: 60-65.
- [19] 张付志,伍朝辉,姚芳. 基于贝叶斯算法的垃圾邮件过滤技术的研究与改进 [J]. 燕山大学学报,2009, 33(1): 47-52.
- [20] 赵静,刘培玉,许明英. 邮件过滤中特征选择方法的性能评价与分析 [J]. 计算机应用研究,2012,29(2): 693-697.

(编辑: 付长缨)