基于贝叶斯算法的垃圾邮件过滤技术

林巧民[1], 许建真[1], 许棣华[1], 王 诚[2]

(1.南京邮电大学 信息网络中心, 江苏 南京 210003;
2.南京邮电大学 信息工程系, 江苏 南京 210003)

[摘要] 对基于朴素贝叶斯算法的垃圾邮件过滤技术进行了研究分析和实验验证. 介绍了向量空间模型(VSM)方法以及特征向量抽取方法, 推导和研究了引入"特征之间互相独立"假设的朴素贝叶斯分类算法. 采用 K次交叉验证的方法, 以收集的一些邮件为语料, 应用朴素贝叶斯分类算法, 通过训练集计算得到类别的先验概率和特征项的类条件概率, 并以此为基础对测试集中的邮件进行归属判断, 以正确率和召回率为指标给出了实验结果.

[关键词] 垃圾邮件, 文本分类, 向量空间模型, 贝叶斯算法
[中图分类号]TP181, [文献标识码]B, [文章编号]1672-1292(2005)04-0061-04

Research on Bayes-Based Spam Filtering

LIN Qiaomin[1], XU Jianzhen[1], XU Dihua[1], WANG Cheng[2]

(1. Campus Network Center, Nanjing University of Posts and Telecommunications, Jiangsu Nanjing 210003, China;
2. Department of Information Engineering, Nanjing University of Posts and Telecommunications, Jiangsu Nanjing 210003, China)

Abstract E-mail communications between people have been greatly affected by spam problem. In this paper, Naive Bayesian categorization algorithm is deduced and analyzed as well as its application and validation in the experiments of spam filtering. Firstly, the paper introduces Text categorization technique, including commonly used vector space model to represent the text and feature extraction methods, such as information gain and document frequency based method. What is more, the behavior of information gain method in the experiments is explained. Secondly, it deduces and analyzes Naive Bayesian with the premise of independence within features. Then, it uses mails collected before as corpus, utilize k-fold cross-validation, and applys the naive Bayesian in experiments. Based on probabilities and that of terms belonging to some category which are gained through training corpus, the paper categorizes mails from test corpus respectively. Finally, experimental result is shown by two indicators, precision and recall.

Key words spam, text categorization, vector space model, Bayes algorithm

目前互联网上的垃圾邮件已经泛滥成灾. 对于学校邮箱用户而言, 由于教师的对外联系较频繁, 垃圾邮件问题显得更为严重, 造成的损失也更大. 针对此现状, 本文讨论了基于叶斯算法的垃圾邮件过滤技术. 由于邮件本身就是文本, 因此过滤垃圾邮件主要是通过文本识别和分类技术来实现的.

1 文本分类技术

1.1 文本分类简介

所谓文本分类, 就是先给定分类体系, 然后将文本分到某个或者某几个类别中去. 这个分类体系通常都是由人工构建的. 分类的模式可能是两类的, 也可能是多类的, 这主要由建模的过程来决定. 在垃圾邮件过滤领域中, 一般采用两类模式, 即 Spam 或者 Non-spam. 文本分类方法一般有自动分类和人工分类, 本文只关注文本的自动分类问题.

1.2 文本分类方法

文本自动分类方法必须解决的首要问题就是如何在计算机中表示文本, 基本的步骤主要包括确定句子和段落的边界、过滤停用词、提取特征词(特征项 Term), 然后将文本转换成可以进行算法分析的特征向量. 假如用 C_i 代表某个特征项, 则一

收稿日期: 2005-06-02
基金项目: 江苏省自然科学基金资助项目(01KJD520005).
作者简介: 林巧民(1979-), 助教, 主要从事计算机网络安全以及嵌入式可配置实时操作系统的研究. E-mail: qmlin@njupt.edu.cn

篇文档就可表示为 $d = (C_1, C_2, \cdots C_j, \cdots, C_k)$. 但由于可能出现 $C_i = C_j (j <> k)$ 的情况, 即文挡中的某个特征词出现了不止一次, 为了计算方便, 通常采用向量空间模型(VSM)来表示文本, 一篇文本可以表示为一个 n 维向量 $(W_1, W_2, W_3, \cdots W_i, \cdots, W_n)$, 其中 $W_i (i = 1, 2, 3, \cdots, n)$ 代表特征项 C_i 的权值, 权值有多种计算方法, 最简单的是布尔权值. 即权值为0或1. 更多的情况下, VSM中的权值计算采用词频(TF)和文档频次(DF)的某种组合. 文本分类可以抽象为一般的描述: 设类别总数为 $|M|$, M_j 表示第 $j (0 < j < |M| - 1)$ 类, 提供给分类器的训练集(训练集中的文本都已经过人工分类)包含 $|D|$ 篇文本, 特征空间 $(F_1, F_2, F_3, \cdots, F_i, \cdots, F_n)$, n 为特征数量, 每篇文本表示为 $d_i = (W_{i1}, W_{i2}, \cdots, W_{in})$, $i = 1, 2, \cdots, |D|$, 一篇待分类新文档可表示为 $d_x = (W_{x1}, W_{x2}, \cdots, W_{xn})$, 目标为将 d_x 分到相应的 M_x 类中去. 图1为文本分类的过程示意图.

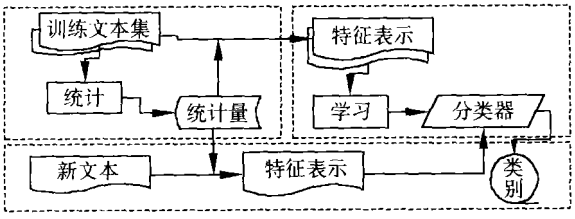

图1 文本分类的过程图

从图1可见, 特征提取是文本分类过程中的一个重要环节. 在文本分类过程中, 不同的特征选择方法将直接影响特征空间的形成, 并对最终的文本分类结果产生某种影响. 常见的特征选择方法有:

(1) 信息增益(Information Gain, IG)[1]. 定义如下:

$$IG(t) = -\sum_{i=1}^{|c|} p(c_i) \lg p(c_i) + p(t) \sum_{i=1}^{|c|} p(c_i \mid t) \lg p(c_i \mid t) + p(\bar{t}) \sum_{i=1}^{|c|} p(c_i \mid \bar{t}) \lg p(c_i \mid \bar{t}) \quad (1)$$

$IG(t)$ 反映了该词为整个分类所能提供的信息量. 式(1)中, $p(\bar{t})$ 表示词 t 不出现的概率, $p(c_i \mid t)$ 表示包含词 t 的文本(正例样本)属于 c_i 类的概率, $p(c_i \mid \bar{t})$ 表示不包含词 t 的文本(反例样本)属于 c_i 类的概率.

信息增益用于度量给定的属性对于训练样例的区分能力, 通过计算信息增益可以得到在正例样本中出现频率高而在反例样本中出现频率低的特征, 以及在反例样本中出现频率高而在正例样本中出现频率低的特征.

(2) 基于文档频次. 一般认为特征项的DF太大的词没有区分度, 而DF太小的词又没有代表性, 因此基于DF的特征选择方法只留下那些DF介于中间的词作为特征.

此外还有不少特征选择方法, 如互信息(MI)、优势率、χ^2 统计量以及相对熵等等.

一般常用于垃圾邮件过滤的文本分类方法和机器学习理论[2], 主要可以分成两类, 一是基于统计的方法, 训练过程为一个统计学习的过程, 得到相应分类器, 如贝叶斯分类算法、k 近邻等; 另一类是基于规则的方法, 这类方法的训练过程是从训练文本集合中学习分类规则, 如决策树、Boosting方法、粗糙集等.

2 贝叶斯分类算法

Bayes公式表述如下: 设 B 是样本空间 S 的一个事件, $P(B) > 0$, $A_1, A_2, \cdots, A_n$ 为 S 的一个事件组, 且满足:

(1) $A_1, A_2, \cdots, A_n$ 互不相容, 且 $P(A_i) > 0 (i = 1, 2, \cdots, n)$;

(2) $A_1 \cup A_2 \cup \cdots \cup A_n = S$.

则

$$P(A_k \mid B) = P(A_k B) / P(B) = P(A_k) P(B \mid A_k) / \sum_{i=1}^{n} P(A_i) P(B \mid A_i) \quad (2)$$

上述贝叶斯公式也被称为后验公式, 它将事件的先验概率与后验概率结合起来, 利用已知信息来确定新样本的后验概率. 假定样本空间 S 为特征空间, 样本点即为特征项(Term), B 事件为某篇新文本, 而 $A_i (i = 1, 2, \cdots, n)$ 为分类体系中的不同类别, 则分类过程可以表述如下:

(1) 待分类文本用一个 n 维特征向量 $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ 来表示.

(2) 设有 $|M|$ 个类 $A_i (i = 1, 2, \cdots, |M|)$, 则贝叶斯分类算法的目标就是求待分类文本 $\boldsymbol{X}$ 在不同类 $A_i (i = 1, 2, \cdots, |M|)$ 中的最大后验概率, 并将 $\boldsymbol{X}$ 归纳入具有最大后验概率的类 $A_k (P(A_k \mid \boldsymbol{X}) > P(A_i \mid \boldsymbol{X}) \; 0 < i < |M| + 1, i \neq k)$, 根据公式(2)及条件概率公式可以有:

$$P(A_k \mid \boldsymbol{X}) = P(A_k) P(\boldsymbol{X} \mid A_k) / P(\boldsymbol{X}) \quad (3)$$

对于任何类别而言, $P(\boldsymbol{X})$ 是相同的, 且可由全概率公式求得. 此外, $P(A_k)$ 可通过 A_k 类中的文本数除以总文本数求出. 因此, 该算法的关键就是求 $P(X \mid A_k)$, 即假定属于 A_k 类条件下求待分类文本的概率 $P(\boldsymbol{X})$, 而 $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$. 为了降低

问题复杂性, 且便于统计计算, 假设特征空间中的特征项 (Term) 之间是不关联的, 即它们是事件 X 的互相对立的"子事件", 此时 $P(X)$ 的计算就可以通过事件的独立性定理求得.

(3) 贝叶斯分类算法的核心就是求 $\max\{P(A_k \mid X),\ 0 < k < |M| + 1\}$.

上述贝叶斯分类算法由于在计算过程中引入了"特征之间互相独立"的假设, 因此也被称为朴素贝叶斯算法, 或简单贝叶斯算法 (Naive Bayes). 将该算法应用于垃圾邮件的过滤系统中时, 若 $|M| = 2$, 则计算时就是判断 $P(\text{Spam} \mid X)$ 与 $P(\text{Non-Spam} \mid X)$ 的差值:

$$\Delta p = P(\text{Spam} \mid X) - P(\text{Non-Spam} \mid X) \quad (4)$$

$\Delta p > 0$, 则表明文本 X 属于 Spam 的可能性大些; 反之 $\Delta p < 0$, X 属于正常邮件的概率大些. 如果 Δp 的绝对值过小, 也就意味着区分该邮件归属的风险过大. 因而在实际邮件过滤系统中, 通常会设有阈值 Step, 只有差值 Δp 满足: $|\Delta p| > \text{Step}$ 时才进行分类, 而差值小于阈值的情况下, 系统将其默认为正常邮件或者进行隔离 (留待邮件管理员人工分类).

朴素贝叶斯是基于先验概率的算法, 因此过滤系统存在一个学习 (训练) 的过程, 这就需要一个语料库 (corpus). 目前, 常用的语料有英文的路透社语料 (Reuters corpus)、TREC 语料、中文的人民日报语料等等. 而针对实际的邮件过滤系统, 语料更多的是来源于本系统日常处理的邮件. 因而邮件过滤网关随着时间推移, 其性能会有所提高.

邮件过滤系统的学习和分类过程如图 2和图 3 所示.

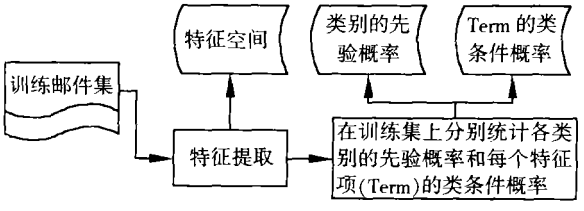

图 2 贝叶斯邮件分类算法的学习过程

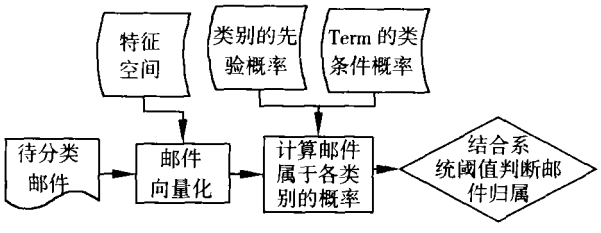

图 3 贝叶斯邮件分类算法的分类过程

3 贝叶斯邮件分类性能评价体系

由于贝叶斯邮件分类算法是基于数理统计原理的, 因此有必要对其进行测试, 并给出一个性能评价体系. 假定测试邮件集 (事先已经人工分好类) 中有 T 封邮件, 其中 Spam 为 $T1$ 封, Non-Spam 为 $T2$ 封, 而系统的贝叶斯分类结果如表 1所示.

表 1 贝叶斯分类结果

	人工分类结果为 Spam	人工分类结果为 Non-Spam
系统判定结果为 Spam	$T1$	$T2$
系统判定结果为 Non-Spam	$T3$	$T4$

假设人工分类的正确率是 100%, 则有以下评价指标:

(1) 正确率 (Precision) (即垃圾邮件检对率). Precision = $T1/(T1+T2)$* 100%, 正确率反应了过滤系统"找对"垃圾邮件的能力, 正确率越大, 将非垃圾邮件误判为垃圾邮件的数量越少, 因此是过滤系统的一个非常重要的评价指标.

(2) 召回率 (Recall) (即垃圾邮件检出率). Recall= $T1/(T1+T3)$ * 100%, 这个指标反映了过滤系统"找出"垃圾邮件的能力, 召回率越高, "漏网"的垃圾邮件就越少.

以上是邮件过滤系统性能评价的两个最重要的指标, 而在实际的系统中, 正确率其实比召回率更重要. 除此以外, 还有其他一些评价指标, 如精确率 (Accuracy)、F 值以及虚报率 (Fallout) 等.

4 实验验证

本文的实验工具为 Bow[3]. 邮件语料一类来源于学校网关所拦截的垃圾邮件 (1200 封), 另一类则是自己积累的私人正常邮件 (370 封) 以及 PU1 语料[4]中的部分非垃圾邮件 (610 封), 总共有 2180封邮件样本, 均分为 6份, 每次取其中的 5份作为训练集, 另一份为测试集, 如此交叉做 6次, 结果取平均值, 这种实验方法也被称为 K 次交叉验证 (K-fold cross validation). 特征选择采用信息增益 (IG)方法: 将训练集中的所有词按照信息增益计算值的大小排序, 选取排在前面约 80% 的词作为特征集. 实验结果如表 2所示.

以上的实验结果是在对样本邮件进行了预处理之后取得的, 预处理包括去停用词和词汇还原, 并且实验时没有考虑设置阈值 Step, 若设置 Step 值 (去除不易区分的未分类邮件), 则贝叶斯分类性能将得到进一步提高. 从表 2已经可以看出, 基

于简单贝叶斯分类算法的垃圾邮件过滤效果在实验中得到了验证和肯定 Sahami[5,6]也曾做过这方面的深入研究.

表 2 实验结果

次序	正确率(Precision)	召回率(Recall)
1	97.1	80.3
2	96.6	83.4
3	98.5	69.1
4	95.3	84.4
5	93.0	86.5
6	95.2	85.0
Avg	96.0	81.5

在本实验过程中,随着特征数量的增加,召回率开始逐渐增大,以后逐渐减小,特征数量为 330附近有个极大值.针对这种现象解释如下:特征太少,不能全面表现邮件的内容,区分度不够;特征太多,又有一些无关的特征,引入了分类噪声.且朴素贝叶斯方法的前提是假设特征之间相互独立,特征数量增加后,特征之间相互依赖的机会增大,独立性变小,因为邮件文本内容中各个词汇之间本来就不是完全独立的.

5 结语

本文主要研究了属于文本分类技术范畴的贝叶斯分类算法,并对其在垃圾邮件过滤领域的应用做了实验验证.贝叶斯算法是基于概率论中的先验概率和条件概率的,在推导朴素贝叶斯算法的过程中引入了"特征之间互相独立"的假设.而在实际邮件文本中,特征属性之间往往存有一定关联,如果把特征间的这种依赖性考虑进来,例如 Friedman和 Goldszmidt就研究了具有树结构的 TAN(tree augmented naive bayes)分类器[7],它放松了朴素贝叶斯算法中的独立性假设条件,扩展了朴素贝叶斯的结构,允许每个属性结点最多可以依赖于 1个非类结点,这能在一定程度上提高算法的分类性能(尤其是召回率).但实验同时也证明了朴素贝叶斯算法在垃圾邮件过滤中是有效的,因为它可以达到相当高的正确率(尽管召回率低一些),因而目前许多邮件网关产品中主要采用的就是基于朴素贝叶斯分类的过滤技术.

[参考文献]

[1] 许洪波,程学旗,王斌,等. 文本挖掘与机器学习 [J]. 信息技术快报, 2005, 3(2): 1-14.

[2] Androutsopoulos I, Paliouras G, Michelakis E. Learning to Filter Unsolicited Commercial E-Mail [R]. Technical Report 2004/2, NCSR "Demokritos", 2004.

[3] McCallum, Andrew Kachites. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering [EB/OL]. http://www.cs.cmu.edu/~mccallum/bow, 1996.

[4] Androutsopoulos I, Koutsias J, Chandrinos K V, *et al*. An evaluation of naive bayesian anti-spam filtering[C]// Potamias G, Moustakis V, Someren Van M, *et al*. Proceedings of the Workshop on Machine Learning in the New Information Age. Barcelona: 11th European Conference on Machine Learning (ECML 2000), 2000: 9-17.

[5] Sahami M. Using Machine Learning to Improve Information Access [EB/OL]. http://ai.stanford.edu/~sahami/bio.html, 1998.

[6] Sahami M, Dumais S, Heckerman D, *et al*. A bayesian approach to filtering junk e-mail[C]// Sahami Mehran, Craven Mark, Joachims Thorsten, *et al*. Learning for Text Categorization: Papers from the 1998 Workshop [s.l.]: AAAI, 1998.

[7] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers [J]. Machine Learning, 1997, 29: 131-163.

[责任编辑:刘　健]