基于改进的 Porter Stemmer 词干提取与核方法的垃圾邮件过滤算法

孙汉博 冯国灿
(中山大学数学学院 广州 510275)

摘 要 统计学习方法现已大量应用于垃圾邮件识别,其中表现突出的包括贝叶斯过滤器、支持向量机等。近年来,为应对日益严重的垃圾邮件问题,提出诸多改进算法或创新思路。通过改进 Porter Stemmer 并使之适用于垃圾邮件过滤,从而充分提取文本的有效特征,摒弃冗余信息,加强了过滤效果;将改进方法的 Porter Stemmer 与原方法分别应用于线性核、高斯核、多项式核支持向量机以及贝叶斯过滤器,对比实验结果可知,错误率分别下降了 63.7%,63.1%,61.3%和 11.4%,证明了改进方法的显著效果;另外,实验结果证明 SVM 过滤器显著优于贝叶斯过滤器,且能更大程度体现改进方法的优势;最后,给出多种定量评价和语义角度的分析,启发采用用户个性化定制的过滤器。

关键词 垃圾邮件,SVM,核方法,SMO 算法,Porter Stemmer
中图法分类号 TP181 **文献标识码** A

Spam Filter Algorithm with Improved Porter Stemmer and Kernels Methods

SUN Han-bo FENG Guo-can
(School of Mathematics, Sun Yat-sen University, Guangzhou 510275, China)

Abstract At present, statistical learning methods have been widely used in spam classification in which Bayesian classifier and SVM are favorable. To face the challenge of spams, a number of novel ideas and improved algorithms were proposed. We proposed improved Porter Stemmer algorithm to extract text features thoroughly and tailored it for spam classifiers. Compared with original algorithm, linear kernel SVM, gaussian kernel SVM, polynomial SVM and Naïve Bayes classifiers obtain 63.7%, 63.1%, 61.3% and 11.4% decrease of error rate respectively based on proposed improved Porter Stemmer. Besides, experimental results justify that SVM has significant advantages when applied to spam classification compared to Naïve Bayes, while SVMs also obtain greater improvements facilitated by improved Porter Stemmer. We also conducted a shallow analysis from the perspectives of linguistics and illustrated the potential value of spam classifier with personalized customization.

Keywords Spam, SVM, Kernel function, SMO algorithm, Porter Stemmer

1 引言

当前,电子邮件已成为世界范围内重要的沟通工具。然而,垃圾邮件却成为不可忽视的威胁,其中大部分垃圾邮件带有商业性质,也有些邮件包含病毒和其他具有危害的内容,给用户带来了精力和财产损失。近年来,垃圾邮件产业愈发成熟,甚至形成了完整的靶向选取策略[1]。

迄今为止,垃圾邮件(Spam/Junk mail)在国际上没有统一的定义。在《中国互联网协会反垃圾邮件规范》中垃圾邮件被界定为:

1)收件人事先没有提出要求或者不同意接收的广告、电子刊物以及各种宣传邮件。

2)收件人无法拒收的电子邮件。

3)隐藏发件人身份、地址、标题等信息的电子邮件。

4)含有虚假的信息源、发件人、路由等信息的电子邮件。

主流垃圾邮件过滤技术分为 3 类:基于行为、基于规则、基于内容。基于行为的方法根据发件人的发件行为预判垃圾邮件,如秦逸提出的 BJMD[2] 方法。基于规则的方法主要有黑名单、关键字匹配等。这些方法的主观性会造成大量误判,因此目前的垃圾邮件过滤系统逐渐倾向于引入基于内容的统计学习方法。

基于内容的垃圾邮件过滤可以将问题转化为文本分类问题,但与一般分类问题有所区别:

1)将正常邮件误判为垃圾邮件通常是更严重的错误;

2)垃圾邮件数量巨大,因此不仅重视过滤效果,也重视性能;

3)垃圾邮件的认定因人而异或随时间改变。

基于内容的垃圾邮件过滤方法可以细分为基于统计学习的方法和基于规则的方法,统计学习方法往往得到隐式规则;而基于规则的方法常常得到可以直观理解的显式规则。

Sahami[3] 最先将 Naive Bayes 引入垃圾邮件过滤,他采用自己收集的邮件作为实验数据,除了使用词汇作为特征外,还使用了词组特征和其他属性特征(如标题中非字母和数字字符所占的百分比),实验结果表明,其他属性特征能够较大幅度地提高过滤结果(精确率在 95%左右)。因为计算效率高、

本文受国家自然科学基金(61272338)部分资助。
孙汉博(1994-),男,CCF 学生会员,主要研究方向为统计学习、模式识别,E-mail:hanbosun@umich.edu;冯国灿(1962-),男,博士,教授,主要研究方向为图像处理、计算机视觉、模式识别,E-mail:mcsfgc@mail.sysu.edu.cn。

过滤性能良好，贝叶斯过滤器得到了广泛的应用[4]。文献[5]比较了 7 个版本的贝叶斯过滤器，得出布尔朴素贝叶斯(Bolean Naive Bayes)、多项式布尔朴素贝叶斯(Multinomial Bolean Naive Bayes)、基本朴素贝叶斯(Basic Naive Bayes)具有较好的过滤效果。

另一种公认较好的垃圾邮件分类方法是支持向量机，Drucker[6]最先将线性 SVM 用于垃圾邮件过滤，取得了当时最好的结果，并指出采用二值表示的 SVM 的效果稍好于采用多值表示的 SVM。Androutsopoulos[7]也在实验中引入了 SVM，与 Drucker 不同的是，他使用了实数值作为特征权重。Kolcz[8]则采用了多种 SVM 方法的变形进行垃圾邮件过滤。

Carreras[9]使用决策树来过滤垃圾邮件，但由于效果一般，因此其一般作为 Boosting 方法的弱学习器来使用。Carreras[9]和 Nicholas[10]将 AdaBoost 引入到垃圾邮件过滤。Androutsopoulos[7]在实验中引入了另外一种 Boosting 方法——LogitBoost，但结果略逊于 AdaBoost。Boosting 方法最主要的缺点是训练速度较慢。

还有一些其他方法，如刘洋等[11]将粗糙集(Rough Sets)引入到垃圾邮件过滤，采用了 11 种非文本属性(包括收信人个数等)进行邮件分类，在小规模的垃圾邮件样本上得到了 80%左右的正确率；潘文峰[12]将 Balanced Winnow 算法引入到垃圾邮件过滤，该方法的效果接近目前所发表的最好结果，而 Winnow 在训练速度和分类速度上具有较大的优势，具有较高的实用价值。另外，聚类方法[13]、人工免疫系统(Artificial Immune System)[14-15]、马尔科夫随机域模型(Markov Ran-dom Field Model)[16]、社会网络(Social Network)[17]方法也见诸于研究成果。

学者对不同方法的比较也有相当的成果，Drucker[6]比较了 Ripper 和 SVM，结论是两种方法相当，而 SVM 与 Boosting 相比训练时间更短；Androutsopoulos[18]比较了 Naive Bayes，KNN 以及基于关键词过滤(Outlook 邮箱所用)的方法，结论是基于关键词过滤的效果最差，KNN 和 Naive Bayes 效果相当，但是 KNN 的过滤时间较长；Carreras[9]比较了决策树、Naive Bayes 和 Boosting 方法，结论是 Boosting 方法最好，Naive Bayes 和决策树方法的性能相当，但 Naive Bayes 的正确率高于决策树方法。已经实现的算法中[19]，Naive Bayes 和 Rocchio 在训练和分类上速度占优但结果一般，而 Flexible Bayes，SVM，Winnow 方法分类速度较快且结果性能很好。Boosting 尽管结果很好，但在训练速度上处于下风。如果考虑扩展性(更新的方便程度)，Winnow 算法占据优势。

目前的成果集中在对经典统计学习的改进算法和一些创新方法。

然而，将统计学习方法应用于该领域时大量工作局限于算法的改进，较少着眼于文本处理。传统方法基于邮件语料库进行信息筛选后建立字典会造成信息冗余；对原始或简单处理后的邮件构建特征向量并作为分类器的输入会导致有效特征的流失。以上原因导致分类器既不准确又低效。本文根据英文语法特点，结合多种一致化处理技巧并改进 Martin Porter[20]提出的波特词干分析法(Porter Stemmer)，提出了改进的基于词干提取和还原的特征提取方法，并将其应用于过滤器(如 SVM，Naïve Bayes)，该方法有以下优势：

1)构建字典的过程中可以有效地排除无用信息，在同样过滤效果下，只需要更低维输入向量，减少了计算量，提升了过滤器性能。

2)从语义角度而非逐字符认定同义，从而加强了有效特征，使同样的字典长度包含更大的信息量。

3)避免区分同义异形词(如时态、单复数)，使特征向量更能表现邮件文本内容，有效提高了分类正确率。

4)解决了错拼、插入信息等问题。垃圾邮件发送者近年来翻新花样(如，插入超链接、邮箱、故意使用邮件接收者可以理解的错误拼写等)，该方法将一并解决。

2 改进的 Porter Stemmer 方法

“特征比算法更能影响分类效果”是一个共识。相较传统方法，本文改进了波特主干分析法(Porter Stemmer)并使其适用于邮件文本特征提取。首先选取高频词干并结合邮件领域的特殊表达建立字典。尽管在文本处理阶段增加了计算量，但处理后的文本更加还原和浓缩了有效特征，从而为后续训练和预测阶段带来分类效果和性能的大幅提升。

2.1 结合 Porter Stemmer 方法的意义

目前应用最为广泛的词干提取手段即波特词干算法，它基于后缀剥离，由大量规则和部分特例构成，复杂程度中等。此方法最初应用于信息检索系统索引词的范化。直到目前，研究者仍在不断改进和实现不同程序语言版本。

波特词干算法应用于邮件文本特征提取的可行性首先由其输出结果的确定性保证。

其次，Porter 方法将意义大致相同的词提取出相同的词干，如 “help”，“helps”，“helped”都提取出“help”，实现了语义下的一致化。否则，形式意义下的相同会刻意区分大体同义的单词(如仅有单复数、时态、词性等区别)，从而产生语义意义下有效词汇频率或信息量不足的问题，造成重要特征的流失。

再者，一些垃圾邮件发送者故意使用不影响人工理解的错误拼写，企图欺骗过滤器，例如，将“click”拼作“c1ick”，“watch” 拼作“wa * tch”等。通过波特词干提取，这些不同“单词”的语义词干(可能是原词)会被还原或提取出来。

最后，垃圾邮件发送者用于传递丰富商业信息的超链接、邮箱及其他信息的插入也会被转化成为有效特征，从而辅助过滤器做出判断。

2.2 改进 Porter Stemmer 方法

尽管 Porter Stemmer 方法应用于信息系统的技术已经非常成功，然而在文本分类方面的应用还不成熟。具体来说，对于不同性质的文本分类应该具有个性化设计。比如在邮件分类领域，如果忽略文本中插入的超链接、邮箱地址、货币符号及其他某些特殊表达，往往会造成大量重要特征的流失。

本文提取并筛选个性化特征并基于此改进 Porter Stemmer 方法，使得字典不局限于传统意义的单词。这一系统也很容易推广到其它文本识别或分类领域。改进算法流程图如图 1 所示。

本文的改进如下：

1)加入对符号的处理流程，处理包括符号统一、删除和替换；

2)处理插入内容(如 HTML、图片)，并进行替换；

3)处理邮件类文本中常见的特殊表达(如邮箱、链接)，分析并进行替换、添加、删除、改写；

4)调整 Porter Stemmer 算法顺序，使之适用于邮件过滤。

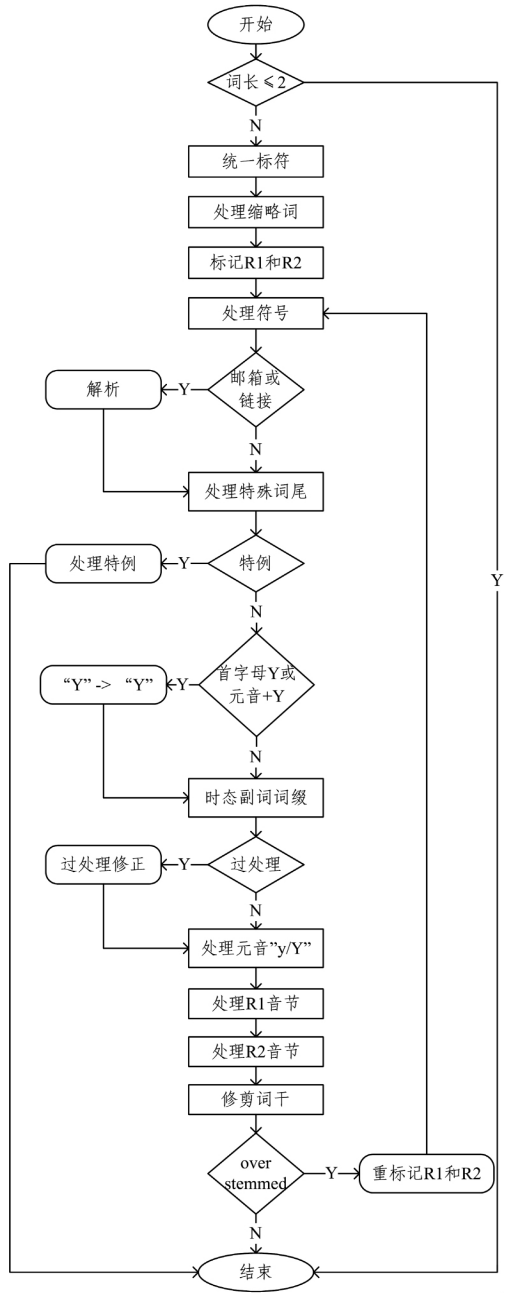

图 1　算法流程图[1)]

改进的 Porter Stemmer 与原方法处理前后对照示例如表 1 所列。

为叙述具体算法，首先引入以下定义 。

定义 1　R1：若存在"元音＋非元音"组合，则定义为第一个这样的组合之后的区域；若不存在，则定义为空。

定义 2　R2：在 R1 中，若存在"元音＋非元音"组合，则定义为第一个这样的组合之后的区域；若不存在，则定义为空。

定义 3　元音(Vovel)：a，e，i，o，u，y。

定义 4　例外(exception)：不符合算法规则的特例，如按算法规则，herring→her，outing→out，canning→can 等不合理，需要特殊处理。设合理映射集 $\Phi=\langle\phi_1,\phi_2\rangle$，其中：

$\phi_1: W\to X$ 是特殊变换映射，W 为原词空间，X 为变换后词空间，$\phi_1(W_{ij})=X_i$。

$\phi_2: V\to X$ 是不变映射，$\phi_2(V_i)=V_i$。

定义 5　叠字(double)：bb ddff gg mm nn pp rrtt。

定义 6　短音节(short syllable)："非元音＋元音＋非元音"且后一个非元音不是 w，x 或 Y；或"首字母元音＋非元音"。

定义 7　短词(short word)：以短音节结尾且 R1 为空的词。

定义 8　有效 li-ending(valid li-ending)：c d e g h k m n r t。

表 1　Porter Stemmer 与改进方法处理对照示例

原词	改进处理后	原方法
help/helps/helped	help	help
C1ICK/click/cl !ck	click	报错/click/报错
wat * ch/watches/watch	watch	报错/watch/ watch
something	someth	someth
www. bbc. com	httpaddr	报错
abcd@126. com	emailaddr	报错
abcd@yahoo. com. cn	cneeeothereeeyahoo	报错
$ 100	dollarnumb	报错
900＋	number	报错

算法具体内容如下。

如果该词仅有小于或等于两个字母，则不做处理；否则，进行以下操作：

步骤 0a：统一引号[1)]，并删除首尾引号，保留词中引号(如 we'll，John's)。

步骤 0b＋[2)]：处理引号伴随"s"。

搜索表 2，取其中最长可能的字符串并删除。

表 2

'	's	's'

步骤 0c：根据定义 1 和定义 2 确定 R1 和 R2。

步骤 1a：一致化处理，根据表 3 进行替换。

表 3　一致化前后对照表

原始内容	A－Z	0－9	各种 URLS	Email 地址	各种 HTML	$ 等货币符	标点符号
一致化后	a－z	number	httpaddress	emailaddress	html	dollar	N/A

步骤 1b：邮箱分析。若是 emailaddress，则按序进行如下分析：地域分析→邮箱性质分析→是否是常用邮箱。对照表如表 4 所列。

表 4　邮箱分析对照表

地域	. HK	. pk	. cn	. de	…	other
性质	. edu	. net	. gov	. org	…	other
常用	gmail	outlook	hotmail	yahoo	…	other

在"emailaddress"后插入词"loceeecateeegen"，其中"loc"，"cat"和"gen"分别做地域、性质和常用邮箱分析，"eee"是分隔符，同时标记该词原型是邮箱地址。例如 abcd@yahoo. com. cn→cneeeothereeeyahoo. ，地域和性质都将". com"等常规后缀和不能匹配的后缀归为"other"，若地域、性质都是"other"，则删除插入词"loceeecateeegen"，继续处理"emailaddress"。

步骤 1c：链接分析。若是"httpaddress"，则做地域和性质分析，分析方法与邮箱分析的基本相同，区别在于分隔符使用"uuu"。若地域和邮箱性质都是"other"，则类似地删除插入词并继续处理"httpaddress"。

步骤 2：处理"s"型和"ed"型。搜索以下字符串，并取其中最长可能的字符串进行下述操作。

1) 引号有两种编码方式，需统一

2) ＋：反复操作

sses→ss

ied+ies*[1]→i(如果之前有多于一个字母)或 ie(如果之前不多于一个字母)

s→Null[2](如果前一位之前的部分包含一个元音)

us+ss → NT[3]

步骤 3a:忽略特殊词:

sky, news, howe, Inning, outing, canning, herring, earring, proceed, exceed, succeed

即:define exception0 as {

[substring] atlimit[4] among(

/* invariant forms: */

'sky'howe'news'

'atlas'cosmos'bias'andes' //并非复数

'inning'outing'canning'herring'//并非进行时

'proceed'exceed'succeed'//并非过去时

//…)}

步骤 3b:若是如下词,则进行替换后退出。

define exception1 as {

[substring] atlimitamong(

/* special changes: */

'skis'(<-'ski')

'skies' (<-'sky')

'dying' (<-'die')

'lying' (<-'lie')

'tying' (<-'tie')

/* special-LY cases */

'idly' (<-'idl')

'gently' (<-'gentl')

'ugly' (<-'ugli')

'early' (<-'earli')

'only' (<-'onli')

'singly' (<-'singl')

//…tensions possible here…)}

步骤 4:若首字母是 y 或有"元音+y"的组合,则将 y 改写为"Y"。

步骤 5a:处理时态和副词。搜索以下字符串,并取其中最长可能的字符串进行下述操作:

eedeedly+→ee(若 eedeedly+$\in R_1$)

ededly+ingingly+→ Null(若之前部分含一个元音)

步骤 5b:若 5a,则 5b,若以 at,bl 或 iz 结尾,则在结尾补 e;若以叠字(定义 5)结尾,删除最后一个字母;若是短词(定义 7),结尾加 e。

步骤 6*:处理元音 y。y 或 Y →i,若前一位是非首字母的非元音。

步骤 7a:R1 区域的第一优先级音节处理。搜索以下最长可能的字符串,若存在于 R1,则进行下述操作:

tional → tion

enci → ence

anci → ance

abli → able

entli → ent

izerization → ize

ationalationator → ate

alismalitialli → al

fulness → ful

ousliousness → ous

ivenessiviti → ive

bilitibli+→ ble

ogi+→ og(如果前一位是 l)

fulli+→ ful

lessli+→ less

li+→ Null(若之前是一个 valid li-ending(定义 8))

步骤 7b:R1 区域的第二优先级音节处理。搜索以下最长可能的字符串,若存在于 R1,则进行下述操作:

tional+→ tion

ational+→ ate

alize → al

icateicitiical → ic

fulness → Null

ative* → Null(若 ative 属于 R2)

步骤 8:R2 区域音节处理。搜索以下最长可能的字符串,若存在于 R2,则进行下述操作:

al anceenceeric ableibleantementmentent ism ate itiousiveize → Null

ion → Null (若前一位是 s 或 t)

步骤 9*:修剪词干。搜索以下字符串,若存在则进行下述操作:

e → Null(若属于 R2 或属于 R1 且之前不是短音节(定义 6))

l → Null(若属于 R2 且前一位是 l)

步骤 10:Y → y。步骤 11:修正 over-stemmed。

参考 Lovins stemmer[5],若处理后是 gener, commun 或 arsen,将 R1 的标记点(R1 左边第一位)设在处理后的词尾对应的原词位再从步骤 2 重复。

算法至此结束。

由于采用特例方法(piecemeal)而非系统(systematic)方法,因此得到确定结果并避免了大量低效工作。另外,由于 Porter Stemmer 方法对词长等的限制,实现过程中采用 try-catch 块。

2.3 构造特征向量

首先基于改进的算法,结合文档频率和步骤 1 中的特殊表达生成字典。之后,为将邮件样本输入支持向量机,需要对样本进行预处理,步骤如下:

步骤 1 移除邮件头。

---

[1] *:一次性操作

[2] Null:置空

[3] NT:Do Noting,不操作

[4] atlimit:完整搜索,如果发现匹配,则进行相应操作

[5] Lovins J B. Development of a Stemming Algorithm[J]. Mechanical Translation and Computational Linguistics, 1968:11:22-31

步骤 2　对邮件文本应用 2.2 节中的算法进行处理。

步骤 3　建立从文本到特征向量的映射。

$\Psi: W \to X$

设字典集 $d=\{d_1, d_2, \cdots, d_n\}$，处理后的邮件词集 $w=\{w_1, w_2, \cdots, w_m\} \in W$，$dw=\{d_{I_1}, d_{I_2}, \cdots, d_{I_m}\} \in d \cap w$，$I=\{I_1, I_2, \cdots, I_m\}$，则特征向量 $x=\{x_1, x_2, \cdots, x_n\} \in X$，其中 $x_i \in \{0,1\}$，$x_i=1$ 当且仅当 $i \in I$。

3　核方法理论及实现方法

本文采用核方法 SVM 实现垃圾邮件过滤，分别采用线性核、高斯核和多项式核。

核方法 SVM 归结于求解最优化问题，如(1)所示：

$$\min \omega(\alpha)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \cdot K(x_i, x_j)-\frac{C}{2}\sum_{j=1}^{m}\alpha_j$$
$$\text{s.t.} \sum_{i=1}^{m} y_i \cdot \alpha_i=0, 0 \leqslant \alpha_i \leqslant c \tag{1}$$

其中，$k(x,z)=\langle \phi(x), \phi(z) \rangle$ 是核函数，α_j 是松弛变量，C 是惩罚因子。

最优解记为 $\alpha^*=(\alpha_1^*, \alpha_2^*, \cdots, \alpha_m^*)^{\mathrm{T}}$，选择 α^* 中一个小于 C 的正分量 α_i^*，并据此计算式(2)，求得决策函数(式(3))。

$$b^*=y_j-\sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x_j) \tag{2}$$

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i^* K(x_i, x_j)+b^*\right) \tag{3}$$

由于数据的稀疏性，为提高效率，本文 SVM 的实现采用 SMO(Sequential Minimal Optimization)算法，将工作集大小限定为 2，即每次只根据两点进行优化，从而每个二次规划(Quadratic Programming，QP)问题可得到一个解析解。虽然把工作集限制到最小导致迭代次数增加，但是由于两个变量的问题的 QP 子问题求解很快，整个 QP 子问题也得以快速求解。

4　实验与分析

4.1　数据集说明

实验数据集[1)]来自于反垃圾邮件联盟(SpamAssassin)。样本由 5000 封邮件构成，按照 3∶1∶1 的比例随机分作训练集(Training Set)、验证集(Validation Set)、测试集(Test Set)，如表 5 所列。

表 5　样本组成

	合法邮件/封	垃圾邮件/封	总数/封	垃圾邮件占比
训练集	2036	964	3000	0.321
验证集	687	313	1000	0.313
测试集	692	308	1000	0.308
合计	3415	1585	5000	0.317

4.2　评价指标

过滤性能的评价通常借用文本分类的相关指标。设测试集中共有 N 封邮件，先定义变量，如表 6 所列，其中，$N=A+B+C+D$。

表 6　变量定义

	垃圾邮件	合法邮件
判定为垃圾邮件	A	B
判定为合法邮件	C	D

本文引入如下评价指标。

1)垃圾邮件召回率(Recall)

$$Recall=\frac{A}{A+C}$$

即垃圾邮件检出率。这个指标反映了过滤系统发现垃圾邮件的能力，召回率越高，“漏网”的垃圾邮件越少。

2)垃圾邮件识别准确率(Precision of Spam)

$$Precision=\frac{A}{A+B}$$

即垃圾邮件检对率。准确率反映了过滤系统“找对”垃圾邮件的能力，准确率越高，将非垃圾邮件误判为垃圾邮件的比例越低。

3)正确率(Accuracy)

$$Accuracy=\frac{A+D}{N}$$

即对所有邮件的判对率。

4)F 值(F score，也称 F1 值)

$$F=2\frac{Precision \times Recall}{Precision+Recall}$$

该值实际上是准确率和召回率的调和平均，取值范围为 $[0,1]$。准确率和召回率分别从不同角度反映了分类质量。一般来说，这两个标准是互补的，单纯提高准确率或召回率都是不恰当的。

5)AUC 值(Area Under Curve)

即 ROC(Receiver Ope-rating Characteristic)曲线在轴方向上的积分。

这里有两个问题值得注意：

1)过滤垃圾邮件的目标是保证在低错纠率的情况下尽量降低漏纠率，因为在通常情况下错误地阻断一封合法邮件要比漏掉一封垃圾邮件的代价大得多，这也就是很多用户不愿轻易使用垃圾邮件过滤设备的原因，因此对垃圾邮件的识别准确率(Precision)应格外重视。

2)垃圾邮件和合法邮件的量常常是不均衡的，这会形成不平衡的类(skew class)。设想极端情况，如果垃圾邮件的总量只有 1%，只要将所有邮件都判定为合法邮件，就可以获得 99%的正确率，但这样的分类器显然是不合格的，此时，F 值相比于正确率更有意义。

4.3　实验结果

4.3.1　参数选取

本文分别使用线性核函数(LSVM)、高斯核函数(GSVM)、多项式核函数(PSVM)基于 SMO 算法迭代求解。实验对比改进的 Porter Stemmer(APS)和未改进的方法(PS)，并选择贝叶斯过滤器作为对照(NB)。

首先确定 SVM 所需参数，依据验证集上的表现选取最优参数，并以测试集结果避免乐观估计，如表 7 所列。

表 7　参数选取与正确率

	LSVM+APS	GSVM+APS	PSVM+APS
参数值	$C=0.03$	$C=1.65, \sigma=4.5$	$C=0.03, c=1, d=2$
正确率	99.10%	97.20%	99.10%
	LSVM+PS	GSVM+PS	PSVM+PS
参数值	$C=0.07$	$C=1.5, \sigma=4.45$	$C=0.07, c=1, d=2$
正确率	97.80%	96.10%	97.90%

注：正确率是 10 次实验的均值。

[1)] 语料可以从 http://spamassassin.apache.org/publiccorpus/下载

4.3.2 性能评价

训练和预测完全向量化使性能大幅提升(见表 8),同时兼容其他核函数。

表 8 性能测试

核函数	嵌套循环训练	向量化训练	嵌套循环预测	向量化预测
线性核 SVM	290.324	24.809	16.981	0.003
高斯核 SVM	425.871	35.778	429.602	0.873
多项式核 SVM	319.128	27.689	154.303	0.134

注:测试环境为 OSX 10.10.5,1.6GHz Intel Core i5,4GB 1600MHz DDR3;样本量为 3000;特征维度为 1837;时间是 30 次实验均值,单位为秒。

4.3.3 学习曲线

根据 4.2 节的评价指标,分别绘出正确率(见图 2)、准确率(见图 3)、召回率(见图 4)和 F 值(见图 5)的学习曲线(Learning Curve),以便分析样本量的影响,并比较不同核函数 SVM 以及贝叶斯方法。

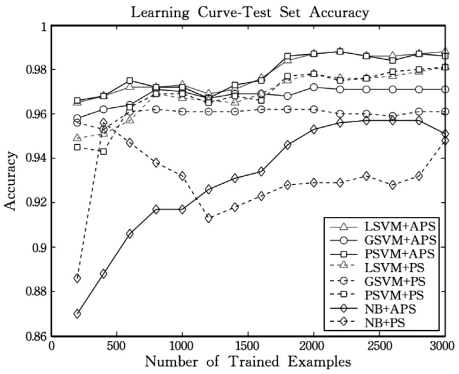

图 2 正确率学习曲线

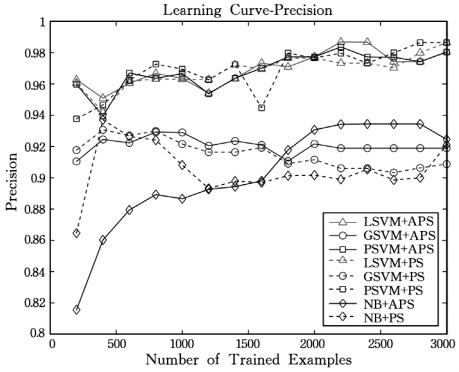

图 3 准确率学习曲线

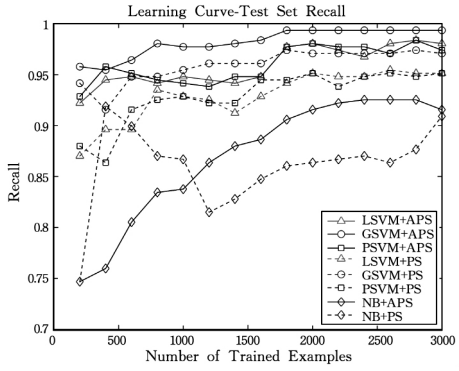

图 4 召回率学习曲线

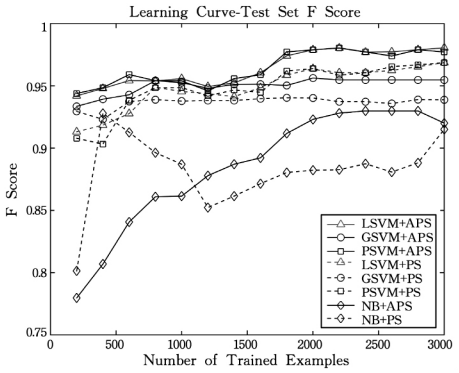

图 5 F 值学习曲线

结合改进的 Porter Stemmer 的各种分类器的各项指标(除 precision 略优于外)都明显优于未改进的方法。

SVM 过滤器全面优于贝叶斯过滤器,且学习曲线相对稳定,即对样本量的依赖程度较低。

3 种核函数 SVM 各有优势:尽管多项式核 SVM 的评价指标全面略优于线性核 SVM,但性能却逊于线性核;另一方面,线性核和多项式核 SVM 从时间开销上看都优于高斯核,但高斯核所需训练样本更少,且召回率明显占优。

文本分类通常是高维问题,而高斯核的优势在于将特征空间映射到无穷维,因此难以发挥优势。仅从正确率角度考虑,往往用多项式核或线性核会取得更好的预测结果。

4.3.4 ROC 图

为进一步对比和评价 SVM 分类器,图 6 示出 3 种核函数 SVM 分别结合改进的 Porter Stemmer(+APS)和原始方法(+PS)的 ROC 曲线;表 9 列出 6 条 ROC 曲线分别对应的 AUC 值。3 种核函数 SVM 结合改进 Porter Stemmer SVM 的 AUC 值分别比原方法提高 0.45%,0.78%和 0.47%,提升明显。其中,PSVM+APS 取得最好结果。

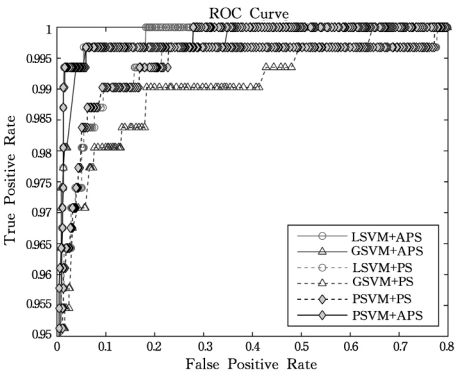

图 6 ROC 曲线局部图

表 9 AUC 值/%

LSVM+PS	GSVM+PS	PSVM+PS	LSVM+APS	GSVM+APS	PSVM+APS
99.28	98.87	99.34	99.73	99.65	99.81

4.4 对比评价

最终评价结果通过五折交叉验证得到,并以贝叶斯方法进行对照,如表 10 所列。改进的 Porter Stemmer 方法对各种核函数的 SVM 以及贝叶斯过滤器都有明显的提升效果,对 SVM 过滤器的提升效果尤为显著,错误率分别下降了 63.7%,63.1%和 61.3%。另外,SVM 的过滤效果明显优于贝叶斯过滤器;各种核函数在各项指标的表现各有优劣(如高斯核函数在垃圾邮件准确率上表现略逊,而召回率更好,这后

发了根据需求的个性化分类器，将在稍后略加讨论），但 3 种核 SVM 的综合指标（正确率和 F 值）都非常接近。

另外，在稳定性方面，GSVM＋APS 优于 LSVM＋APS 和 PSVM＋APS。

最后，表 10 列出了一些最新研究成果，将本文方法应用于特征提取并与之结合可能取得性能的进一步提升。

表 10　交叉验证结果对比

方法 \ 指标	正确率/%	准确率/%	召回率/%	F 值	AUC
NB＋PS	93.68	92.28	87.26	0.8967	\
NB＋APS	94.40	92.73	89.16	0.9090	\
LSVM＋PS	97.19	97.42	93.49	0.9538	0.9928
GSVM＋PS	96.86	92.70	97.79	0.9518	0.9887
PSVM＋PS	97.18	97.49	93.49	0.9544	0.9934
LSVM＋APS	98.98 (0.21%)	99.14 (0.33%)	97.62 (0.17)	0.9837 (0.25%)	0.9973 (0.18%)
GSVM＋APS	98.84 (<0.1%)	96.87 (<0.1%)	99.62 (<0.1%)	0.9820 (<0.1%)	0.9953 (<0.1%)
PSVM＋APS	98.91 (0.13%)	99.18 (0.27%)	97.37 (0.16%)	0.9826 (0.22%)	0.9985 (0.15%)

注 1：每折验证都是 10 次均值。

注 2：括号内是标准差；为保证标准差结果的有效性，使用同样顺序的随机排列数据，因此，贝叶斯方法方差为 0，故没列出。

最后，表 11 列出了其他算法的正确率的对比结果，以供对比。

表 11　其它方法的正确率的对比结果/%

SOAP[21]	RHT-SVM[22]	J48[23]	Three-way Decision[24]	MBPSO[25]	Shifted-1D-LBP[26]
97.00	98.69 (0.03)	92.76	92.96	94.27 (3.31)	92.34

实际应用中，核函数的选取可以针对需求进行。如办公邮箱因为不能承担错将合法邮件分类成垃圾邮件的风险，一般对垃圾邮件识别准确率要求更高，这时选取多项式核更恰当；如果同时对性能有所要求，则线性核即为首选；反之，如果是专门用来接收订阅杂志、歌单的邮箱，则其对垃圾邮件的识别准确率要求不高，且这类邮箱往往充斥着垃圾邮件，不妨使用高斯核函数，即采取更强力的过滤手段。3 种核函数模型的综合性能十分接近

4.5　文本和语义角度的分析

4.5.1　特征权重分析

3 种核函数所判断的前十位敏感词和合法词如表 12 所列。

表 12　前十位敏感词和合法词

敏感/合法词	线性核	多项式核	高斯核
No.1	click/spamassassin	click/spamassassin	pleas/ but
No.2	basenumb/ wrote	basenumb/ wrote	our/ wrote
No.3	remov/url	remov/url	remov/ user
No.4	guarantee/ date	guarantee/ date	receiv/ seem
No.5	our/httpaddr	our/httpaddr	your/sai
No.6	pleas/ the	pleas/ the	click/ group
No.7	free/numbertnumb	free/ author	money/thei
No.8	most/ author	most/numbertnumb	dollarnumb/ were
No.9	your/ re	your/newslett	email/ version
No.10	here/newslett	you/ re	busi/httpaddr

其中，敏感词有很好的直观解释，而合法词稍差，从语言学角度看这是自然的，即直觉对“click”，“free”，“please”这类垃圾邮件常用词更敏感。

事实上，从语言学角度深析，合法邮件的词更多为正式和严肃的词，如：“date”，“author”，“but”等，这些词很少会出现在垃圾邮件中。

另外，接近 50%的词汇的权重接近于 0，对分类效果的影响小，这也给未来工作指出了一个方向：改进字典的构建方法，如结合停词表或停用该实验中权重接近于 0 的词。

4.5.2　错分邮件分析

训练集上的错分邮件是短邮件，合法邮件被错分为垃圾邮件的情况居多，这往往是因为在很短的内容中包含一些敏感词；而测试集上被错分的邮件则恰好相反，几乎全部是长邮件，或是因为合法邮件包含大量敏感词，或是因为垃圾邮件发送者的精心包装而被错分。因此，改进方案应该从增加有效特征、增加训练样本、提升抗好词能力等角度思考。

结束语　本文改进了 Porter Stemmer 并提出了结合此方法的 SVM 垃圾邮件过滤器。通过改进 Porter Stemmer 方法并使之适用于垃圾邮件过滤领域，充分选取有效特征，提升了垃圾邮件的过滤效果和性能。根据多种评价指标，本文结果接近目前最好水平。另外，通过对多种核函数 SVM 效果和性能的对比，启发我们采用个性化定制的过滤器。最后，本文还从语言学角度对有效特征和错分邮件进行了一些浅显的探讨，启发改进思路。

本研究还有一些不足，可以尝试加入文本单词外的其他特征，如恶意表达（malicious）[27]、邮件发送者的行为[28]等；最后，如何识别经过合法词包装的垃圾邮件仍需进一步研究。

参 考 文 献

[1] WANG D，IRANI D，PU C. A study on evolution of email spam over fifteen years[C]// 2013 9th International Conference Conference on Collaborative Computing：Networking，Applications and Worksharing （Collaboratecom），2013. Austin，TX，USA：IEEE，2013：1-10.

[2] 秦逸. 基于行为的垃圾邮件检测技术[J]. 计算机科学，2012，39(11)：86-89.

[3] SAHAMI M，DUMAIS S，HECHERMAN D，et al. A Bayeslan approach to filtering junk E-Mail[C]// Proceeding of Learning for Text Categorization Workshop-held in Conjunction with ICML/AAAI-98. Madison，WI，USA，1998：3256-3260.

[4] 王青松，魏如玉. 基于短语的贝叶斯中文垃圾邮件过滤方法[J]. 计算机科学，2016，43(4)：256-259.

[5] ALMEIDA T A，YAMAKAMI A. Advances in spam filtering techniques[J]. Computational Intelligence for Privacy and Security，2012，394：199-214.

[6] DRUCKER H，D W，VAPNIK V N. Support Vector Machines for Spam Categorization[J]. IEEE Transactions on Neural Networks and Learning Systems，1999，20(5)：1048-1054.

[7] ANDROUTSOPOULOS I，PALIOURAS G，et al. Learning to Filter Unsolicited Commercial E-Mail[J]. International Proceedings of Computer Science & Information Tech，2004(2)：1-52.

[8] KOLCZ A，ALSPECTOR J. SVM-based Filtering of E-mail Spam with Content-specific Misclassification Costs[C]// Proc of Textdm01 Workshop on Text Mining-held at the 2001 IEEE International Conference on Data Mining，2001. San Jose CA USA：IEEE，2001：1-14.

[9] CARRERAS X，MARQUEZ L. Boosting Trees for Anti-Spam Email Filtering[C]// Proceedings of Euro Conference Recent Advances in NLP(RANLP-2001). TzigovChark，Bulgari：RANLP，2001：58-64.

（下转第 79 页）

参 考 文 献

[1] VLACHOS I K，SERGIADIS G D. Intuitionistic fuzzy information：Applications to pattern recognition [J]. Pattern Recognition Letters，2007，28(2)：180-210.

[2] ATANASSOV K T. Intuitionistic fuzzy sets[J]. Fuzzy Sets & Systems，1986，20(1)：87-96.

[3] ÇOKER，DOGAN. Fuzzy rough sets are intuitionistic L-fuzzy sets [J]. Fuzzy Sets & Systems，1998，96(3)：381-383.

[4] NANDA S，MAJUMDAR S. Fuzzy rough sets [J]. Fuzzy Sets & Systems，1992，45(2)：157-160.

[5] JENSEN R，SHEN Q. Fuzzy-rough attribute reduction with application to web categorization[J]. Fuzzy Sets & Systems，2004，141(3)：469-485.

[6] JENSEN R，SHEN Q. Fuzzy-rough data reduction with ant colony optimization[J]. Fuzzy Sets & Systems，2005，149(1)：5-20.

[7] ZHOU L，WU W，ZHANG W. On characterization of intuitionistic fuzzy rough sets based on intuitionistic fuzzy implicators [J]. Information Sciences，2009，179(7)：883-898.

[8] 徐伟华. 序信息系统与粗糙集[M]. 北京：科学出版社，2013.

[9] 袁修久，张文修. 模糊目标信息系统的属性约简[J]. 系统工程理论与实践，2004，24(5)：116-120.

[10] 徐伟华，柴昱洲，李坚，等. 优势关系下分配约简矩阵算法的程序实现[J]. 重庆理工大学学报(自然科学版)，2011，25(4)：117-122.

[11] 张晓燕，徐伟华，张文修. 序目标信息系统中分布约简的矩阵算法[J]. 重庆理工大学学报(自然科学版)，2010，24(3)：56-61.

[12] 苟光磊，王国胤. 基于不协调置信优势原理关系的知识约简[J]. 计算机科学，2016，43(6)：204-207.

[13] 刘芳，李天瑞. 基于边界域的不完备信息系统属性约简方法[J]. 计算机科学，2016，43(3)：242-245.

[14] 徐伟华，张先韬，王巧荣. 序信息系统中变精度粗糙集属性约简的 Matlab 实现[J]. 重庆理工大学学报(自然科学版)，2013，27(1)：107-115.

[15] JING Y，LI T，HUANG J，et al. An incremental attribute reduction approach based on knowledge granularity under the attribute generalization [J]. International Journal of Approximate Reasoning，2016，76：80-95.

[16] 鞠恒荣，杨习贝，戚湧，等. 量化粗糙集的单调性属性约简方法[J]. 计算机科学，2015，42(8)：36-39.

(上接第 67 页)

[10] NICHOLAS T. Using AdaBoost and Decision Stumps to Identify Spam E-mail[J]. Natural Language Processing，2003：1-7.

[11] 刘洋，杜孝平，周二胜，等. “垃圾邮件”的智能分析、过滤及 Rough 集讨论[C]// 中国计算机学会网络与数据通信学术会议，2002. 武汉，2002：515-521.

[12] 潘文锋. 基于内容的垃圾邮件过滤研究[D]. 北京：中国科学院计算技术研究所，2004.

[13] SOONTHORNPHISAJ N，CHAIKULSERIWAT K，TANG O P. Anti-Spam Filtering A Centroid-Based classification Approach [C]// Proceedings of International Conference on Signal Processing (ICSP)，2002. Pattaya Thailand：ICSP，2002：1096-1099.

[14] ODA T，WHITE T. Increasing the accuracy of a spam-detecting artificial immune system[J]. IEEE Transactions on Evolutionary Computation，2004，1：390-396.

[15] 张泽明，罗文坚，王煦法. 一种基于人工免疫的多层垃圾邮件过滤算法[J]. 电子学报，2006，34(9)：1616-1620.

[16] CHHABRA S，YERAZUNIS W，SIEFKES C. Spam filtering using a Markov random field model with variable weighting schemas[C]// Proceedings of 4th IEEE International Conference on Data Mining，2014. Hong Kong，China：IEEE，2014：347-350.

[17] 李渊，廖闻剑，彭艳兵，等. 复杂网络性质探讨及在垃圾邮件过滤中的运用[J]. 计算机科学，2013，40(S1)：145-148.

[18] ANDROUTSOPOULOS I，KOUTSIAS J，CHANDRINOS K，et al. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal E-mail Messages[C]// Proceedings of the 23rd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (SIGIR)，2000. Athens Greece：ACM，2000：160-167.

[19] RENUKA D，HAMSAPRIYA T，CHAKKARAVARTHI M R，et al. Spam Classification Based on Supervised Learning Using Machine Learning Techniques[C]// Proceedings of Process Automation，Control and Computing (PACC)，2011. Coimbatore，India：PACC，2011：1-7.

[20] RIJSBERGEN C J V，ROBERTSON S E，PORTER M F. New models in probabilistic information retrieval：British Library Research and Development Report，no. 5587[R]. Cambridge：Computer Laboratory University of Cambridge，1980.

[21] SHEN H Y，LI Z. Leveraging Social Networks for Effective Spam Filtering[J]. IEEE Transactions on Computers，2013，63(11)：2743-2759.

[22] DEBARR D，SUN H，WECHSLER H. Adversarial Spam Detection Using the Randomized Hough Transform Support Vector Machine[C]// Proceedings of 2013 12th International Conference on Machine Learning and Applications (ICMLA'12). Miami，FL，USA，2013：299-304.

[23] SHARAM A，SAHNI S. A Comparative Study of Classification Algorithms for Spam Email Data Analysis [J]. International Journal on Computer Science & Engineering，2011，3(5)：111-117.

[24] ZHOU B，YAO Y Y，LUO J G. A Three-Way Decision Approach to Email Spam Filtering[C]// Advances in Artificial Intelligence，Canadian Conference on Artificial Intelligence. Canadian，Ottawa，Canada，2010：28-39.

[25] ZHANG Y D，WANG S G，PHILLIPS P，et al. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection[J]. Knowledge-Based Systems，2014，64：22-31.

[26] KAYA Y，ERTUĞRUL Ö F. A novel approach for spam email detection based on shifted binary patterns[J]. Security & Communication Networks，2016，9(10)：1216-1225.

[27] ALQATAWNA J，FARIS H，JARADAT K，et al. Improving Knowledge Based Spam Detection Methods：The Effect of Malicious Related Features in Imbalance Data Distribution[J]. International Journal of Communications，Network and System Sciences，2015，8(5)：118-129.

[28] NAKSOMBOON S，WATTANAPONGSAKORN N. Considering behavior of sender in spam mail detection[C]// Proceedings of International Conference on Networked Computing (INC). Gyeongju，Korea (South)，2010：1-5.