一种基于动态特征词典的SVM中文电子邮件过滤方法*)

侯 岩 王文剑

(山西大学计算机与信息技术学院 计算智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 随着电子邮件的广泛应用,泛滥成灾的垃圾邮件对人们的生活和网络安全带来了严重的威胁,反垃圾邮件问题已成为全球性的具有现实意义的问题。本文提出了一种基于动态特征词典的SVM中文邮件过滤方法,通过动态构造特征词典以及选择合适的支持向量机(Support Vector Machine,SVM)核参数,有效地提高了垃圾邮件的过滤精度,实验结果超过了网易免费邮所公布的过滤指标。

关键词 支持向量机,中文电子邮件,过滤,动态特征词典

A SVM Chinese E-mail Filtering Approach Based on Dynamic Feature Dictionary

HOU Yan WANG Wen-Jian

(School of Computer & Information Technology, Key Laboratory of Computational Intelligence & Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006)

Abstract With the widely use of email, amounts of spams are fiercely threatening the safety of the internet and the lives of people. Anti-spam problem has become an international, significant and practical topic now. This paper presents a SVM Chinese email filtering approach based on dynamic feature dictionary. By constructing a dynamic feature dictionary and selecting the parameter of SVM kernel, the presented approach can improve the correct rejection rate of filtering spams dramatically. The simulation results show that the filtering factors of the presented approach can beyond those published by NetEase Free E-mail.

Keywords Support vector machine, Chinese E-mail, Filter, Dynamic feature dictionary

1 引言

当今世界是信息的世界,随着Internet的广泛应用,电子邮件已经深入到人们的生活,成为人们互相交流、获取信息的重要渠道。然而,伴随着电子邮件的广泛应用,垃圾邮件、邮件病毒、邮件攻击也越来越猖獗,这些垃圾邮件的存在占用了大量的资源,如网络带宽资源、服务器存储资源等。同时,日益增长的垃圾邮件附载大量虚假、不健康,甚至危害社会稳定与安全的信息,它的发送处于非受控状态,对国家的信息安全造成很大的威胁。因此建立有效的垃圾邮件过滤系统是非常必要的。

早期的垃圾邮件过滤工具中常采用黑名单-白名单或者手工制订规则的方法[1,2],黑白名单技术存在很大的局限性;而基于手工制定规则的方法需要人为地制定垃圾邮件过滤规则,往往给没有相关经验的用户造成很大的不便,同时,手工制定过滤规则需要耗费较多时间,准确率也不高,而且,随着垃圾邮件的发展,其相关特征也在发生变化,用户在对过滤规则进行维护时就显得更加困难了。因此,目前的垃圾邮件过滤工具逐渐倾向于采用基于内容的机器学习判别方法[3]。

目前基于内容的垃圾邮件过滤方法可以大体分成基于规则的方法和基于概率统计的方法。前者通过训练得到人们可以理解的显式规则,其中以Ripper、Decision Tree和Boosting等方法[4,5]为代表。这种方法虽然能够处理邮件头和正文,但是实质上还是简单的二值判断,缺少可信度的知识。而且用户的兴趣和垃圾邮件发送者所开发出的产品和诡计都在不断变化,所以规则也要进行不断更新,要求用户有丰富经验和充裕时间来训练和调整这些过滤规则,另外,这种方法误判的发生率也比较高。而基于统计的方法则是指根据统计理论,先对已分类的邮件训练样本进行学习,提取出能表征各类邮件的特征向量及特征值,再根据这些值对新的邮件进行分类过滤。本质上,这种方法可以看作是基于规则方法的推广,只不过统计方法中得到的规则是一种不被人轻易理解的"隐式规则"。目前基于统计的方法主要有:贝叶斯方法和神经网络方法等[6]。

由Vapnik提出的基于统计学习理论的支持向量机[7]是一种有效的机器学习方法,可以解决许多实际问题[8],尤其在文本分类领域中,SVM是公认的较好的方法之一[9]。本文提出一种基于内容的SVM中文电子邮件过滤方法,此方法首先将邮件内容进行分词,然后根据自动构建的垃圾邮件特征词典对邮件进行向量化处理,再应用SVM进行垃圾邮件过滤,以期提高邮件的过滤效率。

2 本文提出的邮件过滤方法

本文提出一种基于内容的SVM中文电子邮件过滤方法,首先对中文邮件向量化,其中包括邮件分词处理、特征词典的动态生成和项的权重计算,然后利用SVM对邮件进行过滤。

2.1 邮件内容分词处理

不同于英文邮件,中文电子邮件的信体部分每个词条间

*)本文受到国家自然科学基金(60673095),山西省高校科技研究开发项目(200611001),山西省留学人员科技活动择优资助项目,山西省高校青年学术带头人基金资助。侯 岩 硕士研究生,主要研究方向为机器学习与计算机网络;王文剑 教授,博士生导师,主要研究领域为计算智能与机器学习。

没有固定的空格分隔符，为了将中文电子邮件向量化，首先需要进行分词。在实际的应用中，人们常选择最大匹配法作为分词的方法，但最大匹配法是一种基于分词词典的机械分词法，不能根据文档上下文的语义特征来切分词语，对词典的依赖性较大，所以在应用时，常常会造成一些分词错误。为了提高系统分词的准确度，本文采用正向最大匹配法和逆向最大匹配法相结合的分词方案，先根据标点对文档进行粗切分，把文档分解成若干个句子，然后再对这些句子用正向最大匹配法和逆向最大匹配法进行扫描切分，如果两种分词方法得到的匹配结果相同，则认为分词正确，否则，按最小交集处理。

2.2 邮件特征表示

本文采用向量空间模型(Vector Space Model，VSM)来表示邮件。在邮件向量空间模型中，一封邮件就是一篇文档(Document) d，邮件中的每个词就是一个特征项(Term) t，所有的特征项构成特征词典(Term List) $\{t_k \mid 1 \leqslant k \leqslant n\}$。这样，邮件可以用特征词典表示为 $d(t_1, t_2, \cdots, t_n)$。

在邮件过滤中，特征项的选择会在很大程度上影响邮件分类的效果，所以构造一个合适的特征词典显得非常重要。但是对于中文邮件来讲，一般中文的词汇高达十万多条，如果从这些词汇中选择对分类贡献较大的词，将需要花费大量的时间和资源，而与邮件过滤最相关的那些词往往出现在已知的邮件中。本文采用动态构造特征词典的方法，将训练集上进行分词后得到的互不相同的特征词作为特征词典，因为训练集不同，特征词典也不同，这种动态词典的构造方法更符合实际应用，可以有效地提高邮件的过滤效率。

对特征词典中的任一特征项而言，由于它在邮件中出现的位置和出现的频率不同，对邮件分类结果的影响也是不同的，因此，对含有 n 个特征项的邮件文档而言，应该给每个特征项赋予一定的权重表示其重要程度，可将邮件表示为 $d=d(t_1, w_1; t_2; \cdots; t_n, w_n)$，简记为 $d=d(w_1, w_2, \cdots, w_n)$，其中 w_i 是 t_k 的权重，$1 \leqslant k \leqslant n$。目前给权重 w_i 赋值的方法主要有两种：一种是由专家或者用户根据自己的经验与所掌握的领域知识人为地赋权，这种办法随意性很大，而且效率也很低，很难适用于大规模真实文本的处理。另一种是运用统计的方法，也就是用文本的统计信息(如词频、词之间的同现频率等)来计算邮件特征词的权重。目前常采用的方法是 w_i 表示为特征项 t_i 在邮件 d 中的出现频率 $tf_i(d)$(Term Frequency)或者是 $tf_i(d)$ 的函数，常见的有布尔函数、平方根函数、对数函数和 TF-IDF(Inverse Document Frequency)函数等[10]。

本文采用词频权重方法作为计算权重的方法。词频权重的定义为：

$$w_i(d)=tf_i(d) \tag{1}$$

考虑到文本长度对权值的影响，本文对项的权值进行归一化处理：

$$w_i(d)=\frac{tf_i(d)}{\sqrt{\sum_{t_i \in d}[tf_i(d)]^2}} \tag{2}$$

2.3 SVM 邮件分类

目前，SVM 方法已成功应用于分类、回归、时序预测等领域[8,9]。SVM 是一种基于核函数的学习方法，常用的核函数有线性核、多项式核、sigmoid 核和高斯核等，本文采用最常用的线性核和高斯核来进行邮件的过滤。在众多的 SVM 学习算法中，Thorsten Joachims 提出了一种解决大型数据集学习的算法[11]，它运行速度快、操作简单。本文采用 SVM^{light} 软件包[12]对电子邮件进行过滤。

3 仿真实验

3.1 实验环境及评价指标

本文采用中国教育和科研网紧急响应组(CCERT) 2005 年 8 月公布的电子邮件数据集[13]，该数据集包含从 2005 年 6 月 1 日至 6 月 31 日收到的垃圾邮件和正常邮件。其中垃圾邮件包括 10000 封，正常邮件包括 10000 封。实验中，本文分别从垃圾邮件样本集和正常邮件样本集中各随机选取 1000 个样本作为训练集，从剩余的垃圾邮件样本集和正常邮件样本集中各随机选取 1000 个样本作为测试样本，共生成五组测试样本。

本文采用正常邮件通过率(Normal Mail Rate，NMR)、虚警率(False Alarm Rate，FAR)、漏检率(Miss Rate，MR)和正确过滤率(Correct Rejection Rate，CRR)来检验邮件过滤的效率，它们分别定义如下：

$$NMR=\frac{N_{H \to H}}{N_{H \to S}+N_{H \to H}} \tag{3}$$

$$FAR=\frac{N_{H \to S}}{N_{H \to S}+N_{H \to H}} \tag{4}$$

$$MR=\frac{N_{S \to H}}{N_{S \to H}+N_{S \to S}} \tag{5}$$

$$CRR=\frac{N_{S \to S}}{N_{S \to H}+N_{S \to S}} \tag{6}$$

其中，$N_{H \to H}$ 表示将正常邮件判断为正常邮件的样例数；$N_{H \to S}$ 表示将正常邮件判断为垃圾邮件的样例数；$N_{S \to H}$ 表示将垃圾邮件判断为正常邮件的样例数；$N_{S \to S}$ 表示将垃圾邮件判断为垃圾邮件的样例数。

3.2 实验结果及分析

实验中首先将训练集中的 2000 个样本(其中包括 1000 封垃圾邮件和 1000 封正常邮件)分词，一共得到 12460 个互不相同的特征词，由这些特征词组成新的特征词典，然后，在此基础上进行邮件过滤。

1) 基于线性核的邮件过滤

实验中，选择惩罚因子 $C=1.0$。训练结果和测试结果分别见表 1 和表 2。

表 1 显示了基于线性核的邮件过滤的训练结果，其中，有效样例数为 1995，支持向量数为 321，训练时间为 0.36s。从表 2 中可以看出，使用基于动态特征词典的邮件过滤方法可以取得很好的过滤效果，正常邮件的通过率和垃圾邮件的过滤率分别达到了 98.5%和 96.27%以上，所得到的虚警率最大不超过 1.9%。由于线性核比较简单，垃圾邮件的过滤率还未达到理想的效果。

表 1 基于线性核的邮件过滤的训练结果

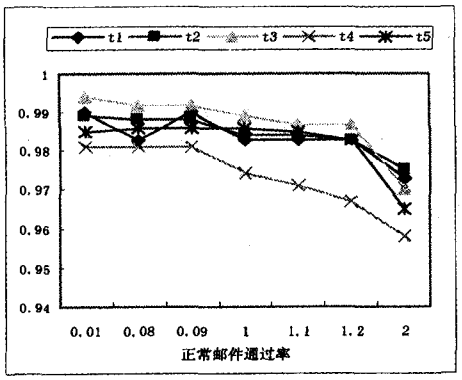
权重	有效样例数	#SV	CPU Time
TF 权重	1995	321	0.36s

表 2 基于线性核的邮件过滤的测试结果

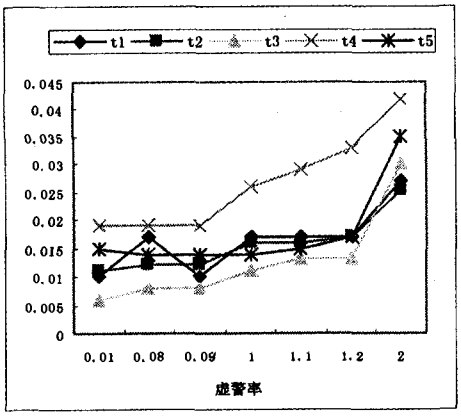
测试集	NMR	FAR	MR	CRR
test1	0.9899	0.0101	0.0170	0.9830
test2	0.9890	0.0110	0.0291	0.9709
test3	0.9940	0.0060	0.0142	0.9858
test4	0.9810	0.0190	0.0373	0.9627
test5	0.9850	0.0150	0.0351	0.9649
平均值	0.9878	0.0122	0.0265	0.9735

2)基于高斯核的邮件过滤

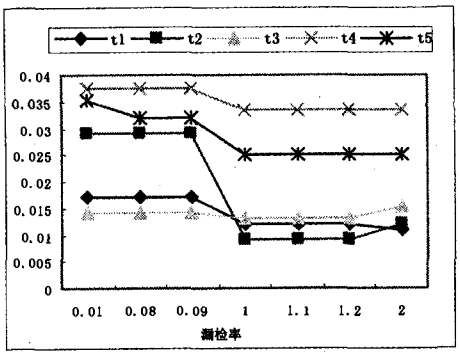
实验中，通过调节参数 C 和高斯核参数 r 进行邮件过滤，不同的 C 和 r 得到的训练结果和测试结果分别见表 3 和图 1。



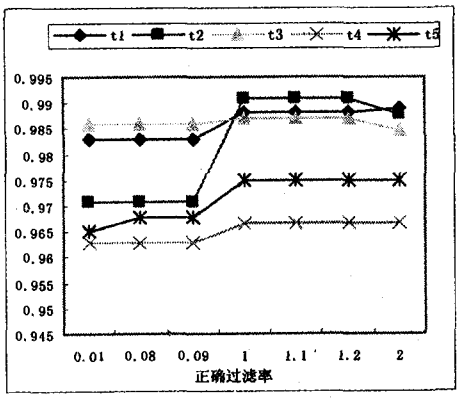
(a)



(b)



(c)



(d)

图 1　基于高斯核的邮件过滤的测试结果在四个测试指标上的比较

从表 3 可以看出，随着高斯核参数 r 从 0.01 增加到 2.0，支持向量数增加了一倍多，由 333 增加到了 679，而随着核参数的增加，训练时间也不断增加，在 r 等于 2.0 时达到了 3.8s，CPU 时间比使用线性核时增加了一个数量级。图 1 给出了高斯核参数 r 从 0.01 增加到 2.0 时四个指标的变化情况。当 r 等于 0.01 时，正常邮件的通过率和垃圾邮件的过滤率与使用线性核所得到的结果一致。当 r 达到 1.0 时，平均正确过滤率达到了 98.15%，平均虚警率保持在 1.68%。当 r 继续增加到 2.0 时，正确过滤率下降到 96%附近，而虚警率却上升到 4%左右。因此我们可以得出，当 r 达到 1.0 时，取得的过滤效果最好，实验结果超过了网易 163 免费邮公布的垃圾邮件过滤指标[14]，即正确率为 98%。

表 3　基于高斯核的邮件过滤的训练结果

r	C	#SV	有效样例	CPU Time
0.01	50.2504	333	1995	1.90s
0.08	6.5033	337	1995	2.00s
0.09	5.8093	337	1995	2.09s
1.0	0.7910	481	1995	2.35s
1.1	0.7495	505	1995	2.75s
1.2	0.7155	522	1995	3.20s
2.0	0.5783	679	1995	3.80s

以上实验结果表明，由于本文提出的方法可以实时更新特征词典，随时加入与邮件过滤密切相关的特征词，不断扩大邮件向量的维数，使得向量能充分地表示每封电子邮件，因而基于动态特征词典的邮件过滤方法在各项指标都有了很大提高，同时，由于使用的 SVM 分类方法具有可以避免“维数灾难”的特点，训练时间和测试时间都很短，平均每秒可以处理 1000 封电子邮件。

结论　本文提出了一种基于内容的 SVM 中文电子邮件过滤方法，通过采用动态构造特征词典有效地提高了邮件的过滤效率。同时，通过对核参数的调节，不仅有效地提高了垃圾邮件的过滤精度，使虚警率达到 1%左右，实验结果超过了网易 163 免费邮公布的垃圾邮件过滤指标，而且过滤速度很快。因此，本文提出的中文邮件过滤方法具有较高的实用价值。另外，本文构造的特征词典与训练集有关，如何保证稳定的过滤效果还值得进一步研究。

参 考 文 献

1　Kaplan S. How antispam software works. Wired M-agazine, 2003, 11(4):43
2　Vaughan-Nichols S J. Saving private e-mail. IEEE Spectrum Magazine, 2003, 40(8):40～44
3　William, Cohen W. Learning Rules that Classify E-mail. In: Proceedings of the 1996 AAAI Spring Sympo-sium in Information Access, 1996
4　Freund Y, SchapireR E. Game Theory, Ol-line Prediction, and Boosting. In: Proceedings of the Ni-nth Annual Conference on Computational Learning Theory, 1996
5　William, Cohen W. Fast effective rule induction. M-achine Learning. In: Proceeding of the 12th Int. Conf. , 1995
6　Sahami M, Dumais S, Heckerman D. A Bayesian Approach to Filtering Junk E-Mail. AAAI'98 Workshop on Learning for Text-Categorization, Madison, 1998
7　Cristianini N, Shawe-Talor J. 支持向量机导论. 李国正，王猛，曾华军. 电子工业出版社，2004
8　Vapnik V, Golowich S, Smola A. Support vector me-thod for function approximation, regression estimat-imation, and signal processing. In: M. Mozer, M. Jordan, T. Petsche, eds. Advanced in Neural Information Pro-cessing Systems 9. Cambridge, MA: MIT Press, 1997
9　Joachims T. Text Categorization with Suppo-rt Vector Machine: Learning with Many Relevant F-eatures. In: European Conference on MachineLearning, 1998
10　Salton G, Buckley C. Term weighting approaches in automatic text retrieval. Information Processing and Management, 1988, 24(5):513～523
11　Joachims T. Making large-Scale SVM Learning Pr-actical. In: B. Schölkopf, C. Burges, A. Smola, eds. Adv-ances in Kernel Methods - Support Vector Learning. MA: MIT-Press, 1999
12　Joachims T. svm_light. http://download. joac hims. org/svm_light/current/svm_light. tar. gz
13　CCERT 中文邮件样本集. http://www. ccert. edu. cn/ spam/sa/datasets. htm
14　网易 163 免费邮. http://mail. 163. com