一种基于SVM算法的垃圾邮件过滤方法

范婕婷,赖惠成
FAN Jie-ting,LAI Hui-cheng

新疆大学 信息科学与工程学院,乌鲁木齐 830046
College of Information Science & Engineering,Xinjiang University,Urumuqi 830046,China
E-mail:fjt1212@163.com

FAN Jie-ting,LAI Hui-cheng.Spam filter approach based on Support Vector Machine.Computer Engineering and Applications,2008,44(28):95-97.

Abstract: Content-based filtering is one of the mainstream technologies used so far.In view of which the essence of spam filter is the classify problem,this paper aims to provide a kind of filter method against spam based on SVM,and tries to adopt SMO algorithm to categorize the spam.Through the experiment,SMO algorithm can perform better,and the classification time of SVM classifier is reduced greatly.

Key words: spam;Support Vector Machine(SVM);Sequential Minimal Optimization(SMO);classification time

摘 要: 基于邮件内容的过滤是当前解决垃圾邮件问题的主流技术之一。针对垃圾邮件过滤本质是分类问题,提出了一种基于支持向量机对垃圾邮件过滤的方法,并且将SMO分类算法结合到垃圾邮件分类中。通过实验,SMO算法能够取得较好的分类效果,缩短了支持向量机分类器的分类时间。

关键词: 垃圾邮件;支持向量机;序列最小优化算法;分类时间
DOI: 10.3778/j.issn.1002-8331.2008.28.033 **文章编号:** 1002-8331(2008)28-0095-03 **文献标识码:** A **中图分类号:** TP391

1 引言

随着互联网的快速发展,电子邮件逐渐成为人们普遍采用的通信方式。CNNIC在第21次报告中指出它在网络应用中的使用率为56.5%,而将近1/4的邮件为垃圾邮件。迄今为止,国际上对垃圾邮件并没有一个标准的定义。但垃圾邮件的基本特征是“不请自来”,主要是一些商业广告或有其他目的的宣传。

早期对垃圾邮件的过滤常采用黑白名单。虽然具有速度快和简单的特点,但存在着需要用户不断更新垃圾邮件的过滤规则和维护黑名单邮件列表的缺点。而垃圾邮件过滤的本质是一个二分类问题(垃圾邮件和合法邮件),即可以根据邮件的内容进行分类。所以,目前反垃圾邮件的主流技术倾向于基于邮件内容的过滤技术[1]。基于邮件内容的过滤技术又可分为基于规则的方法和基于概率统计的方法。基于规则的方法有:Ripper、决策树(Decision Tree)、粗糙集(Rough Sets)、Boosting等方法。基于概率统计的主要方法有:贝叶斯(Navie Bayes)[2]、K近邻(K-Nearest Neighbor)[3]、神经网络(Neural Network)和支持向量机(Support Vector Machine,SVM)[4]等方法。而SVM的分类效果优于其他的分类方法。

支持向量机是Vapnik等人[5]提出的一种基于统计学习理论的机器学习方法,它以最大化分类间隔构造最优分类超平面来提高分类器的泛化能力,具有训练样本小、学习速度快、易于扩展等特点。影响基于邮件内容过滤器的因素主要是邮件的特征表示和分类器的分类速度,所以当训练样本数目非常大的时候,SVM分类器的速度会有所下降。本文为了提高分类器在样本数目非常大的情况下的分类速度,提出了SMO(Sequential Minimal Optimization)算法对样本进行训练。通过实验,SMO算法能够取得较好的分类效果,并且提高了分类器的分类速度。

本文的章节结构如下:以上为引言部分,简要介绍了目前垃圾邮件过滤的主要技术;第2章为垃圾邮件过滤的预处理技术,为垃圾邮件分类做准备;3章是分类器的理论说明;4章是该系统的实验结果和分析部分;5章是总结。

2 垃圾邮件过滤预处理

2.1 邮件分块

电子邮件的一般格式包括信头和信体两部分。有时候仅仅根据信头信息就可以判断一封邮件是否是垃圾邮件,所以首先要分离信头和信体,然后分别进行基于信头和信体的过滤。这样可以提高工作效率,避免一些不必要的工作。

2.2 中文分词和去停用词

中文邮件不同于英文邮件,邮件的信体部分每个词条间没有固定的空格分隔符,为了将中文电子邮件向量化,让机器识

基金项目: 新疆高校科研重点项目(the Key Research Project for University of Xinjiang No.XJEDU2007I07)。
作者简介: 范婕婷(1983-),女,硕士研究生,研究方向:网络与通信系统;赖惠成(1963-),通讯作者,男,教授,硕士生导师,研究方向:网络与通信系统。
收稿日期: 2008-04-10 **修回日期:** 2008-06-23

别文本,所以要进行分词处理。常见的分词方法有正向/逆向最大匹配算法、正向/逆向最小匹配算法和基于概率方法的无词典分词方法等。本文采用正向/逆向最大匹配算法相结合的分词方案。首先根据标点对文本进行粗分,分成若干句子后,再对这些句子分别用正向最大匹配算法和逆向最大匹配算法进行扫描划分,如果两种分词方法结果相同,则判为分词正确,反之按最小交集处理。

分词处理完成之后,得到一系列文本单词所组成的表列,特征辞典是由表列中的单词所构成的集合。为了缩小文本特征辞典,提高邮件分类器的训练分类效率,通常需要对辞典进行去停用词处理。停用词通常是指在各类文本中都频繁出现,因而被认为很少带有助于分类的词,如代词、介词、连词等高频虚词。

2.3 邮件特征表示

邮件的表示主要采用向量空间模型(Vector Space Model,VSM),在邮件向量空间模型中,一封邮件就是一篇文档 d,邮件中的每个词就是一个特征项 t,用 $d=d(t_1,w_1;t_2,w_2;\cdots;t_N,w_N)$ 来表示文档 d。其中 $t_i(i=1,2,\cdots,N)$为特征项,w_i 为 t_i 的权重。对特征辞典中的任意特征项而言,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不同。所以应该给每个特征项赋予一定的权重来表示其重要程度,且规定 $t_i(i=1,2,\cdots,N)$互不相同,为此文档 d 可重新表示为 $d=d(w_1,w_2,\cdots,w_N)$。w_i 一般被定义为 t_i 在 d 中出现频率 $tf(t,d)$的函数,常见的有布尔函数、平方根函数、对数函数和 TFIDF 函数。本文采用 TFIDF 函数,一般比较普遍的 TFIDF 公式为:

$$w(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_t+0.01)]^2}}$$

其中 $w(t,d)$为特征项 t 在文档 d 中的权重,$tf(t,d)$为特征项 t 在文档 d 中的词频,N 为训练样本的总数,n_t 为训练样本集中出现特征项 t 的文本数,分母为归一化因子。

3 支持向量机

3.1 SVM 理论基础

设给定训练样本集$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中 $x_i\in R^d,y_i\in(-1,+1),i=1,\cdots,n$。$R^d$ 表示 d 维欧氏空间,y_i 为两类训练样本的类别标识。如图 1 所示,H 为超平面,而 H_1、H_2 为支持超平面,H_1、H_2 上的训练样本点就称作支持向量,H_1 与 H_2 之间的距离叫做分类间隔(*Margin*),当 *Margin* 达到最大时,H称为最优超平面。其中 $Margin=\frac{2}{\|\omega\|}$,*Margin* 最大等价于 $\frac{1}{2}\|\omega\|^2$ 最小。所以最优超平面应满足如下条件:

$$\min_{\omega,b}\left(\frac{1}{2}\|\omega\|^2\right) \tag{1}$$

$$y_i(\omega^*x_i+b)\geqslant 1 \quad i=1,\cdots,n \tag{2}$$

引入 Lagrange 函数:

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{N}\alpha_i[y_i(\omega^{\mathrm{T}}x_i+b)-1],\alpha_i\geqslant 0 \tag{3}$$

对 ω、b 微分,求式(3)的极小值可得:

$$\omega=\sum_{i=1}^{N}\alpha_i y_i x_i,\alpha_i\geqslant 0 \tag{4}$$

每个样本向量对应一个 α_i,支持向量即为 $\alpha_i>0$ 的训练样本,N 为支持向量的个数。将式(4)代入式(2)得到样本分类函数:

$$f(x)=\mathrm{sgn}(\omega^*x+b)=\mathrm{sgn}\left[\sum_{i=1}^{n}\alpha_i^*y_i(x_i^*x)+b^*\right] \tag{5}$$

其中 sgn()为符号函数,α_i^* 为最优解,b^* 为阈值。

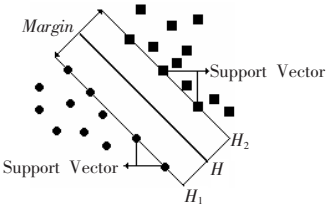

图 1 最优分类面

3.2 核函数

核函数的基本目的是将非线性可分输入空间等价映射到线性可分输入空间。核函数 $K(x_i,x_j)=\Phi(x_i)\Phi(x_j)$。常见核函数如下:

(1)线性内积核函数:$K(x,y)=x^*y$。

(2)多项式核函数(Poly):$K(x,y)=[(x^*y)+1]^q,q\in N,q$ 阶多项式分类器。

(3)径向基内积核函数(RBF):$K(x,y)=\exp\{|x-y|^2/\delta^2\}$。

(4)S 型核函数(Sigmoid):$K(x,y)=\tanh[\nu(x^*y)+c]$。

一般对于文本的非线性可分情况较为常用的核函数为 RBF 核函数和 Poly 核函数。

3.3 SMO 算法

分解算法是解决大量样本下 SVM 训练问题的一类有效方法。SMO[6]算法是目前较为有效的 SVM 训练方法。

SVM 的分类规则函数为式(5),α_i 通过优化如下的目标函数来求解:

$$\begin{cases}\underset{\alpha}{\mathrm{Maximise}}\ L_D=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i y_i\alpha_j y_j K(x_i,x_j)\\ \mathrm{Subjectto}\sum_{i=1}^{N}\alpha_i y_i=0,0\leqslant\alpha_i\leqslant C(i=1,2,\cdots,N)\end{cases} \tag{6}$$

目标函数的最优解必须满足 Karush-Kuhn-Tucher 条件。通过优化目标函数,寻找支持向量,从而训练得到一个 SVM 的分类规则。

不失一般性,假设优化 α_1、α_2,其他 α_i 固定,由线性约束条件,式(6)中第 2 式可知:

$$\alpha_1^{old}+s\alpha_2^{old}=\alpha_1^{new}+s\alpha_2^{new}=r \tag{7}$$

其中 $s=y_1y_2$,r 为常数。α_1^{old}、α_2^{old} 为 α_1、α_2 变化前的值,α_1^{new}、α_2^{new} 为变化后的值。将上式代入式(6)并优化目标函数可以得到下式:

$$\alpha_2^{new}=\alpha_2^{old}+\frac{y_2(E_2-E_1)}{k} \tag{8}$$

其中 $k=2K(x_1,x_2)-K(x_2,x_2),E_1=f(x_1)-y_1,E_2=f(x_2)-y_2$。为使 α_2^{new} 满足 $0\leqslant\alpha_i\leqslant C(i=1,2,\cdots,N)$和条件(7),必须对 α_2^{new} 进行修剪,将 α_2^{new} 的解限制在更为严格的取值范围内。

当 $s=-1$ 时,α_2^{new} 的取值范围为:

$$\max(0,\alpha_2^{old}-\alpha_1^{old})\leqslant\alpha_2^{new}\leqslant\min(C,C-\alpha_1^{old}+\alpha_2^{old}) \tag{9}$$

当 $s=+1$ 时,α_2^{new} 的取值范围为:

$$\max(0,\alpha_2^{old}+\alpha_1^{old}-C)\leqslant\alpha_2^{new}\leqslant\min(C,\alpha_1^{old}+\alpha_2^{old}) \tag{10}$$

设 α_2^{new} 取值范围的最小值为 L，最大值为 H，通过式(9)、(10)便可以计算出 L、H 的值。因此,α_2^{new} 修剪后的值为:

$$\alpha_2^{new}=\begin{cases}L & \alpha_2^{new}<L\\ \alpha_1^{new}+\dfrac{\gamma_2(E_2-E_1)}{k} & L\leqslant\alpha_2^{new}\leqslant H\\ H & \alpha_2^{new}>H\end{cases} \tag{11}$$

求得 α_2^{new} 的值后,由式(7)可求出 α_1^{new}。

SMO 每次迭代时，从训练集中启发式地选择最可能违返 KKT 条件的两点进行优化，并通过监视满足 KKT 条件的允许偏差来判断算法是否可终止。

4　实验与结果

4.1　评价指标

垃圾邮件过滤的性能评价通常借用文本分类相关指标,以下是本文所选用的三种评价指标:

召回率(Recall):$R=\dfrac{N_{s\to s}}{N_{s\to s}+N_{s\to h}}$,即垃圾邮件检出率。

正确率(Precision):$P=\dfrac{N_{s\to s}}{N_{s\to s}+N_{h\to s}}$,即垃圾邮件检对率。

F 值:$F=\dfrac{2PR}{R+P}$,F 是召回率和正确率的调和平均，它将召回率和正确率综合成一个指标。

其中,$N_{s\to s}$ 表示将垃圾邮件判为垃圾邮件的样例数,$N_{s\to h}$ 表示将垃圾邮件判为合法邮件的样例数,$N_{h\to s}$ 表示将合法邮件判为垃圾邮件的样例数。

4.2　实验系统框图

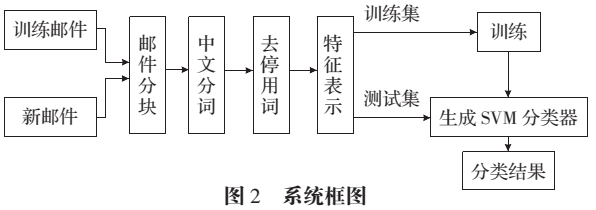

图 2　系统框图

训练邮件用于生成 SVM 分类器,新邮件则通过 SVM 分类器进行分类。

4.3　实验环境及实验结果

实验软件环境为:Windows XP Professional,Matlab 7.0。实验硬件环境为:PentiumIIICPU 为 1.7 GHz,40 G 硬盘,256 M 内存。

实验中电子邮件样本总数为 8 000 封，其中垃圾邮件为 4 000 封,合法邮件为 4 000 封,均来自中国教育和科研网紧急响应组(CCERT)于 2005 年 8 月公布的电子邮件数据集。实验共进行了 8 次，每次实验都从数据集中随机抽取同等数量邮件。训练集共选取 4 000 封,其中垃圾邮件为 2 000 封,合法邮件为 2 000 封。测试集共选取 4 000 封,其中垃圾邮件为 2 000 封,合法邮件为 2 000 封。前 4 次实验选取多项式核函数(Poly)进行邮件过滤,分别得到评价指标 P、R、F 的值,并且求得 4 次实验数据的平均值。后 4 次选取径向基内积核函数(RBF)进行邮件过滤,其实验过程同 Poly 核函数过程。最后与贝叶斯方法(Naïve Bayes)进行了对比。评价指标测试结果如表 1:

表 1　评价指标结果

	SVM(Poly)/(%)					SVM(RBF)/(%)					Bayes/
	1	2	3	4	平均值	5	6	7	8	平均值	(%)
P	93.32	94.61	88.47	89.76	91.47	90.68	88.27	91.17	88.45	89.64	83.49
R	90.27	90.53	89.87	86.47	89.29	89.33	87.92	90.13	86.23	88.40	81.57
F	91.77	92.53	89.16	88.08	90.39	90.00	88.09	90.65	87.33	89.02	82.52

由表 1 可知,*SVM* 方法能够取得更好的分类效果。R、P 和 F 的测试值均比 Bayes 方法的测试值高出八个百分点到六个百分点。其中 Poly 核函数的分类效果略好于 RBF 核函数的分类效果。

时间测试结果表示平均分类单个样本所需用时。测试结果如表 2。

表 2　平均分类单个样本的时间结果

	SVM(Poly)/s					SVM(RBF)/s					Bayes/
	1	2	3	4	平均值	5	6	7	8	平均值	s
X	0.274	0.343	0.264	0.279	0.290	0.110	0.107	0.099	0.101	0.104	0.815
G	0.263	0.292	0.281	0.298	0.284	0.106	0.120	0.118	0.093	0.109	0.798

注:X 表示为训练时间;G 表示为过滤时间。

由表 2 可知 RBF 核函数的单个样本训练时间和单个样本过滤时间的平均用时均小于 Poly 核函数的平均用时，说明 RBF 核函数的分类速度要快于 Poly 核函数的分类速度。而 Bayes 方法用时最长,是 SVM(SMO)方法的 3~8 倍,所以本文采用 SVM(SMO)分类器缩短了在大量样本情况下分类器的总体分类时间。但在样本数量不太大的情况下(如样本数量小于 200),SVM(SMO)分类器的分类时间与 Bayes 分类器的分类时间上的差别就不大了，甚至 Bayes 分类器的分类时间要短于 SVM(SMO)分类器的分类时间。

5　总结

本文将 SMO 分类算法引入到基于支持向量机的垃圾邮件过滤技术中。在样本数量非常大的情况下,通过与 Naïve Bayes 方法进行对比实验可知,本文提出的方法表现出了较好的分类效果,并且提高了分类器的分类时间,具有一定的可行性。

参考文献:

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

[2] 李雯,刘培玉.基于贝叶斯的垃圾邮件过滤算法的研究[J].计算机工程与应用,2007,43(23):174-176.

[3] 田泽,颜松远,徐敬东.基于改进 K 近邻的垃圾邮件过滤技术[J].计算机工程与应用,2007,43(25):178-181.

[4] Anguitad,Bona,Rdellas.Evaluating the generalization ability of Support Vector Machines through the bootstrap[J].Neural Processing Letters,2000,11(1):51-58.

(下转 145 页)