协同分类器及其在邮件过滤中的应用

路 梅[1,2],叶澄清[2]
LU Mei[1,2],YE Cheng-qing[2]

1.徐州师范大学 计算机科学与技术学院,江苏 徐州 221116
2.浙江大学 计算机科学与技术学院,杭州 310027
1.College of Computer Science and Technology,Xuzhou Normal University,Xuzhou,Jiangsu 221116,China
2.College of Computer Science and Technology,Zhejiang University,Hangzhou 310027,China
E-mail:lumei@xznu.edu.cn

LU Mei,YE Cheng-qing.Anti-spam technique with SVM and K-NN cooperation method.Computer Engineering and Applications,2008,44(4):135-137.

Abstract: This paper presents an improved classific algorithm which combines the Support Vector Machine with the K-Nearest Neighbor (K-NN).In this algorithm,the classification of samples in the critical zone is decided by the voting of K-NN set in the feature space and the sample space.Using the method,the filtering error rate will decrease obviously.Finally,lots of examples show that the method is feasible.

Key words: spam;intelligent filtering;Support Vector Machine

摘 要:提出了一种基于支持向量机的改进分类方法.该方法将特征空间分类超平面附近的样本分类,交由特征空间和样本空间中的K-近邻集体投票表决。方法应用于垃圾邮件的过滤之中,邮件合法性误判发生的概率可被有效降低。最后通过垃圾邮件过滤实例验证了该方法的有效性。

关键词:垃圾邮件;智能过滤;支持向量机
文章编号:1002-8331(2008)04-0135-03 **文献标识码**:A **中图分类号**:TP393

日益肆虐的垃圾邮件给电子邮件系统造成了严重的干扰,也给互联网络增加了额外的负载和潜在的危害,垃圾邮件处理和过滤技术已引起了众多研究者的关注,国际上也相继成立了专门的反垃圾邮件应急响应机构。但是,当前的邮件过滤技术仍不能很好满足应用需要,合法邮件被无辜误判、垃圾邮件侥幸漏网的情况时有发生。本文在详细分析现有过滤技术的基础上,讨论了垃圾邮件的技术特征,而后提出了同时考虑样本空间和特征空间因素、K-NN 协同 SVM 的垃圾邮件过滤算法。

1 引言

电子邮件系统的投递流程[1]可示于图 1。按照邮件过滤在邮件系统所处环节的不同,可分为:MTA(Mail Transfer Agent)过滤、MDA(Mail Delivery Agent)过滤、MUA(Mail User Agent)过滤等,它们分别对垃圾邮件源、传播途径和易感受体等环节进行监控,采用滤除、拒收和销毁垃圾邮件的方法,可有效遏制垃圾邮件的发生,降低垃圾邮件的危害。其中 MTA 过滤是健壮邮件传递系统必备;MDA 过滤因其善后处理方式而备受置疑;接受端的MUA 过滤因能方便用户而被广泛使用。无论在何种环节过滤垃圾邮件,精确命中的、高效的过滤算法都不可或缺。目前较为成熟的垃圾邮件过滤算法主要包括基于黑名单的、基于规则的和基于动态统计学习的过滤算法等。

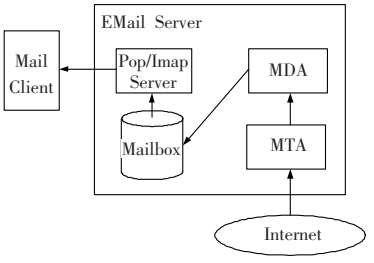

图 1 电子邮件的投递流程

依据邮件内容进行垃圾邮件过滤具有更强的适应性,因此逐渐成为当前反垃圾邮件的主流技术。基于内容的垃圾邮件判别方法大体可分为基于规则的方法和基于机器学习的方法等[1]。从提高垃圾邮件过滤速度出发,Cohen[2]等提出了 Ripper 方法,该方法的速度比原有方法快近两个数量级。Carreras[3]等将 Boosting 方法应用于垃圾邮件的过滤,也获得了较高的性能。此外,刘洋[4]等将粗糙集用于规则的制定和精简。垃圾邮件过滤规则易于理解和修改,但制定和维护困难,仅适合于高端用户。垃圾邮件内容的识别本质上是一个二分类问题(垃圾和非垃圾邮件),利用统计学习的方法可将识别垃圾邮件的知识

作者简介:路梅(1976-),女,硕士,讲师,主要研究方向:多媒体技术、图形图像处理等;叶澄清(1939-),男,教授,博士生导师,主要研究方向为高性能及智能计算机系统结构、分布式并行计算机、多媒体计算机技术与应用等。

收稿日期:2007-05-28 **修回日期**:2007-08-03

学习到分类器之中,即所谓基于机器学习的过滤方法,具体又可分为贝叶斯(Bayes)方法、K 近邻(K-NN)方法、支持向量机(Support Vector Machine,SVM)方法等。基于统计学习的过滤技术具有自适应性、智能性和动态扩展性等特点,是垃圾邮件过滤最具潜力的研究方向。但是,当前该类技术仍然存在诸多不足:如选取的邮件判定特征仍有待改进,高命中率的智能算法仍在寻找之中等。此外,由于各特征量间通常并非完全相互独立,而上述方法将各特征量视为不存在依赖关系的特征,因此,该类过滤技术对垃圾邮件的判定失当就难以根本避免。

2 核函数和支持向量机

支持向量机[5](Support Vector Machine,SVM)的基本思想可概括为:首先通过非线性变换将样本空间变换到一个高维空间,然后在新空间中求取最优线性分类面,此非线性变换通过定义适当的核函数实现。

给定一组样本$\{(x_i,y_i)|x_i\in\Re^m,y_i\in\{-1,+1\},1\leqslant i\leqslant n\}$,若设 ϕ 是输入空间 S 到特征空间 F 的一个映射,核函数 κ 对应 F 中向量内积运算,即

$$\kappa(x,x')=<\varphi(x),\varphi(x')> \tag{1}$$

则非线性可分下的最优分类问题可转化为一个带约束的二次优化问题:

$$\max\left\{\sum_{j=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i y_i\alpha_j y_j\kappa(x_i,x_j))\right\} \tag{2}$$

$$\text{s.t.}\quad \sum_{j=1}^{n}\alpha_i y_i=0,a_i\geqslant 0,0\leqslant i\leqslant n \tag{3}$$

若设 α_i^* 为最优解(通常 α_i^* 中只有少部分不为零),其中非零 α_i^* 对应的样本称为支持向量。

则最佳分类函数[6]:

$$f(x)=\text{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i\kappa(x_i,x)+b^*\right) \tag{4}$$

其中

$$b^*=-\frac{1}{2}\left[\max_{\{i|y_i=-1\}}\left(\sum\nolimits_{j=1}^{n}y_j\alpha_j\kappa(x_i,x_j)\right)+\max_{\{i|y_i=+1\}}\left(\sum\nolimits_{j=1}^{n}y_j\alpha_j\kappa(x_i,x_j)\right)\right] \tag{5}$$

根据 Hilbert-Schmidt 原理,运算满足 Mercer 条件的函数皆可作为此处的内积使用[7],即所谓核函数。常用的此类函数有径向基函数(RBF,Radial Basis Function)

$$\kappa(x,x_i)=e^{-|x-x_i|^2/\sigma^2} \tag{6}$$

SVM 分类器具有较好的泛化能力,衡量其泛化能力的公式为[8]:

$$\overline{p(\text{error})}\leqslant\frac{\overline{svnum}}{(tnum+1)} \tag{7}$$

其中 $\overline{svnum}$ 表示支持向量平均个数,$tnum$ 表示训练样本数。即对于测试样本分类错误率的期望的上界是训练样本中平均的支持向量数占总训练样本数的比率。

SVM 的性能受到核函数形式及其参数、问题本身的复杂程度、分类面附近的噪声点、输入向量参数选择、样本数量、样本分布等因素的影响,其中前三个因素起着关键性作用。若能够量化分析这些因素对 SVM 性能的影响程度,便能确定支持向量占训练样本总数的比率,从而找到降低 SVM 错分率的办法。

3 K-NN 协同 SVM 分类器

SVM 可用一平面分类问题示意,如图 2 所示,其中 L_1 和 L_2 为分界面左右限,L_0 为实际分界面,L 为理想中的最佳分类面。误判通常发生在 L_1、L_2 区间内的样本。训练样本不足是错误发生的主要原因,此外,特征空间线性可分特性的微小出入,都会导致分界面附近的误判。KSVM 算法[9]结合了 K-NN 方法与 SVM 的优点,可有效降低判别器的错分率,KSVM 算法在特征空间的分类面附近采用 K-NN 进行分类,未能充分利用训练样本集中多数样本的信息。

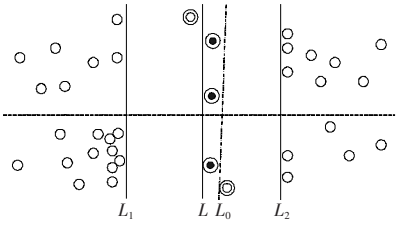

图 2 SVM 错分示意图

3.1 K 最近邻法

最近邻法(Nearest Neighbor)将所有训练样本都作为代表点,在分类时计算待识别样本 x 到所有初始训练样本的距离,x 的类别确定为最近邻训练样本所属类别。

设有 m 个类别 $c_1,c_2,\cdots,c_m$ 的分类问题,每类含有样本 n_i (i=1,2,$\cdots$,m)个。样本 x 与类别 c_i 的距离函数为:

$$d_i(x)=\min\|x-x_i^l\|,l=1,2,\cdots,n_i \tag{8}$$

如果 $d_j(x)=\min\{d_i(x)|1\leqslant i\leqslant m\}$ 则有 $x\in c_j$ 该决策方法称为最近邻法。

K 最近邻法(K-Nearest Neighbor,K-NN)是最近邻法的推广,分类时选 x 的 K 个近邻,再次考察 x 属于 K 类中的哪一类。当样本数量趋于无穷时,K-NN 表现出了良好的性能[8]。K-NN 方法的时空复杂性较大,同时也容易受分类边界样本分布的影响。为此,单独使用的 K-NN 垃圾邮件过滤方法并非一个很好的选择,将其与 SVM 相配合可显著减少分类器的误判率。

3.2 协同分类器

本文对原有 SVM 方法进一步改进,提出了 K-NN 和 SVM 协同的分类新方法,其主要特点有:当测试样本 x 落在 L_1 左侧或 L_2 右侧时完全由 SVM 决定;当测试样本 x 落在 L_1 与 L_2 之间时,采用同时考虑原始样本和特征向量的 K-NN 分类方法完成。算法主要步骤为:

(1)传统 SVM 的求解。求解式(2)、(3)构成的二次规划问题,可得最优解 α_i^* 及 b^*,利用此最优解可构造传统支持向量分类器。对于者,按照传统 SVM 方法确定其类别。对于$|f(x)>T|$者,按照下述部分完成分类。此处 T 为事先确定的阈值。

(2)特征空间的 K-NN 表决。在特征空间内寻找 x 的 K-NN 集合 S。其中,垃圾邮件子集合 S_1,正常邮件子集合为 S_2。同时按下式计算属于垃圾邮件的概率:

$$p_f=1-\frac{\sum_{i\in S_1}\|\phi(x)-\phi(x_i)\|}{\sum_{i\in S}\|\phi(x)-\phi(x_i)\|} \tag{9}$$

由于在特征空间内有

$$\|\phi(x)-\phi(x_i)\|^2=\kappa(x,x)-2\kappa(x,x_i)+\kappa(x_i,x_i) \tag{10}$$

则综合式(9)和式(10)可知

$$p_f=1-\frac{\sum_{i\in S_1}\|x-x_i\|}{\sum_{i\in S}\|x-x_i\|}=1-\frac{\sum_{i\in S_1}\sqrt{|\kappa(x,x)-2\kappa(x,x_i)+\kappa(x_i,x_i)|}}{\sum_{i\in S}\sqrt{|\kappa(x,x)-2\kappa(x,x_i)+\kappa(x_i,x_i)|}} \quad (11)$$

(3)样本空间 K-NN 的表决以及综合分类。同时在样本空间内寻找 x 的 K-NN,同上步可得属于垃圾邮件的概率 p_s。综合特征空间和样本空间所得测试样本从属垃圾邮件的概率,可得

$$p=\lambda p_f+(1-\lambda)p_s \quad (12)$$

其中 λ 为一待定系数,可以根据实际情况动态确定,这里取 $\lambda=0.5$,当 $p\geqslant 0.5$ 则判定为垃圾邮件。根据上述分析,可设计改进 SVM 方法的完整算法。

3.3　协同分类器的算法加速

样本空间 K-NN 算法需要将所有样本保留,并在判别器使用时寻找 K 个最近邻。该过程计算量较大,并且需消耗大量的内存空间。为此,本文进一步提出了协同分类器的加速算法。

(1)寻找特征空间临界训练样本集合 S_c,并构造 Kd-tree 加速搜索[13]。对于未知类别的样本 x,若有$|f(x)|<T$,则按照 K-NN 方法分类,且可采用下述步骤加速搜索:将支持向量中所有符合该条件的样本汇集成 S_c',并将距该集合不超过 T 的所有支持向量记录于集合 S_c''之中。这样,按照式(9)、(10)、(11)表决所需的搜索空间减少为 S_c 而非所有支持向量,计算量可以显著减少。

(2)摒除无效样本,减少非支持向量的有效样本数目。分类超平面在样本空间通常不再是超平面(而是超曲面)。对于未知类别的样本 x,若有$|f(x)|<T$ 则按照 K-NN 方法分类,需同时在特征空间和样本空间寻找 K-最近邻,并据此判别是否垃圾邮件。设$|f(x)|<T$ 的特征向量集合为 S_f,若令

$$S_s=\{x|\ \|z-x\|<2T, z\in S_f\} \quad (13)$$

则在 K-最近邻搜索中,可将整个样本空间减少为 S_s。如果设支持向量集合为 S,则集合

$$\Im=\{x|x\in S\wedge x\notin S_v\wedge x\notin S_c\wedge x\notin S_f\}$$

内的样本可以直接舍弃,从而减少空间消耗和时间消耗。

(3)为 S_s 构造 Kd-tree 加速搜索。上步中将整个样本空间减少到了 S_s,进一步对 S_s 构造 Kd-tree,可减少样本空间中 K-最近邻搜索的时间。

上述三步保证留用的样本皆为必需,同时对样本空间和特征空间中的 K-最近邻搜索进行了加速。

4　电子邮件的协同过滤器构造

完整的电子邮件过滤器需要选取电子邮件的特征并对其量化,同时,需要选择合适的评价指标对过滤器进行评价。

4.1　语料库和特征向量的选取

与普通信件类似,电子邮件也有自己的"信封"和"信文",分别被称为邮件首部和邮件主体。许多邮件在尾部还有"签名"部分。邮件首部主要有收信人电子邮箱地址、发信人电子邮箱地址、传送邮件经过的路径以及信件标题等。邮件主体为实际要传送的信件内容。特征提取就是选取最能区分出是否为垃圾邮件的一些特征,这些特征除了包括对邮件正文文本部分常见的一些词或是词组,还包括垃圾邮件经常会出现的一些特殊特征,如域名中的敏感词汇等。

本文采用 Hopkins、Reeber 等人提供 Spambase 语料库[10]进行实验验证。Spambase 预先选择出来了 57 个词汇作为特征,将每封邮件都表示为向量的形式,以词频作为权重。Spambase 共包含 1 813 封垃圾邮件,2 788 封非垃圾邮件,这些邮件皆来自提供者的私人邮件,已经成为验证过滤算法有效性的常用语料库。

4.2　垃圾邮件过滤评价指标

垃圾邮件过滤的性能评价通常借用文本分类的相关方法,具体指标包括招回率、正确率、精确率、错误率等[11]。上述指标均视垃圾邮件和合法邮件的误判损失相等,事实上,人们往往对合法邮件的误判更为敏感。为此,文献[12]将合法邮件误判损失设置为垃圾邮件误判损失的 λ 倍,提出了更为灵活实用的过滤评价指标——代价因子。

若假设待测试的邮件集合中共有 N 封邮件,垃圾邮件 $A+C$ 封,合法邮件 $B+D$ 封,其中的 A、B 分别对应被正确识别的垃圾邮件和合法邮件,C、D 分别对应被误判的垃圾邮件和合法邮件。若令

$$W=\frac{\lambda D+C}{\lambda(B+D)+(A+C)},\ W_s=\frac{A+C}{\lambda(B+D)+(A+C)}$$

则代价因子 TCR 可表示为

$$TCR=\frac{W_s}{W}=\frac{A+C}{\lambda D+C} \quad (14)$$

TCR 越大,表明当前垃圾邮件过滤系统的损失越低。

此外,垃圾邮件检出率与 TCR 同为用户最关心的过滤系统的评价指标,其计算公式可表示为:$R=A/(A+C)$,该指标可反映过滤系统发现垃圾邮件的能力。垃圾邮件检出率越高,"漏网"的垃圾邮件就越少。

5　算法实现与实验分析

本文所述算法均利用 C++实现完成。同时,利用文献[10]提供的语料库,对上述算法进行了验证。具体步骤是:将 Spambase 语料库的实验样本分为大致相等的 4 等份,每组又分为训练样本和测试样本,并对本文算法进行了训练和测试验证。重复上述操作 4 次,可得实验结果如表 1 所示。经多次实验,SVM 方法中的折衷系数 C 和径向基函数的参数 σ 最终选取为 100 和 0.75。此外,这里的 λ 取 2。

表 1　过滤算法实验对比结果

次数	训练样本/封		测试样本/封		误判合法/垃圾邮件/封				R/%		TCR()		SV/个
	合法邮件	垃圾邮件	合法邮件	垃圾邮件	M1	M2	M1	M2	M1	M2	M1	M2	
1	670	420	30	29	1	1	1	2	0.966	0.931	9.67	7.25	217
2	675	425	29	23	0	2	2	3	0.913	0.870	11.50	3.29	219
3	640	420	30	20	0	1	0	2	1.000	0.900	∞	5.00	213
4	680	445	34	31	1	1	1	1	0.970	0.970	10.30	10.30	221

注:M1 为本文方法,M2 为文献[9]的方法。

6　结语

近年来,电子邮件的日益普及给当今社会的通信联系带来了前所未有的便利,但是,其副产品垃圾邮件却象一颗毒瘤,给使用者和网络系统产生了严重的危害。本文针对垃圾邮件的过滤技术展开研究,针对 SVM 过滤算法在分界面附近误判率高的特点,提出了能有效解决该问题的崭新方法,并通过大量实例对本文方法进行了验证,实验结果表明该方法是切实可行的。

(下转 168 页)