基于SMO和指纹技术在线邮件过滤方法与优化

祝庆荣[1]， 董守斌[2]， 陈 彬[1]

(1.华南理工大学计算机科学与工程学院 广州 510640;2.广东省计算机网络重点实验室 广州 510640)

摘要：研究了垃圾邮件的指纹特征向量表示和SVM过滤方法，设计实现了基于指纹特征和SMO的在线式邮件过滤器FSVM，在在线垃圾信息过滤上获得到了与传统方法相当的效果.在SVM过滤的运算速度方面，基于原始SMO算法，对上述在线方法提出了邮件样本动态集方法(DFSVM)进行条件减弱，在降低了计算量的同时能够保证指纹SMO获得相当的过滤效果.在标准测试集和真实邮件系统中进行了实验验证和对比，结果表明，该改进对提高SVM分类精读有一定的帮助.

关键词：垃圾邮件过滤；支持向量机；条件减弱；动态子集

中图分类号：TP 391 **文章编号**：1671－6841(2009)01－0090－04

0 引言

全球垃圾邮件泛滥，垃圾比例超过九成.当前垃圾信息过滤技术较多使用的是根据信息本身的内容特征判断其类别并加以过滤垃圾信息的检测过滤方法.目前使用较为广泛的方法有3种：基于统计的方法、KNN方法[1]和SVM方法.SVM(Support Vector Machines)方法以其较好的推广性能和较高的分类准确率成为文本分类中公认的较好的新型机器学习方法[2].

1 SVM原理

SVM是一种建立在统计学习理论基础上的机器学习方法，通过数据集学习获得分类决策函数.其基本思想是构造一个超平面作为决策平面，使正负模式之间的间隔最大.对给定的数据集 X，它包含 n 个标记样本向量$\{(x_1,y_1),\cdots,(x_n,y_n)\}$，其中 x_i 是第 y_i 所标示的类别某个特征表示.引入一个设定的向量 w 和偏项 b，预测函数为 $f(x)=\mathrm{sign}(<w,x>+b)$.

SVM的目标是通过训练找出 w 超平面作为对未知信息进行分类的依据.这个超平面满足最小化分类产生的错误总量和最小化正负类别间隔的要求，b 的作用就是调整两者的权重比例，使两者达到一个平衡点.具体求解原理可参见文献[3].应用SVM进行文本分类中一般要进行如图1的几个步骤.由图1可见，SVM是一种批处理应用模式.

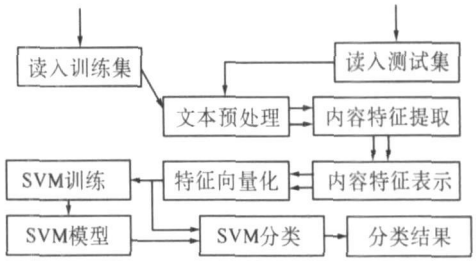

图1 SVM分类的一般步骤

Fig.1 SVM classification steps

2 在线式SVM垃圾邮件过滤方法

2.1 邮件特征表示模型

将SVM应用于垃圾邮件过滤问题上，首先要确定的是邮件特征的提取与表示方法.

采用Salton等人提出的经典的向量空间模型(Vector Space Model，VSM)，其定义为，给定一自然语言文档 D，在选定了特征项以后，用向量 $D=(t_1,w_1;t_2,w_2;\cdots;t_N,w_N)$ 来表示文档 D，其中，$t_i(i=1,\cdots,N)$ 为

收稿日期：2008-12-31

作者简介：祝庆荣(1984－)，男，硕士，主要从事计算机海量信息处理研究，E-mail：zqr.scut@163.com.

特征项,w_i 为 t_i 的权重,规定 $t_i(i=1,\cdots,N)$互不相同,称向量 $D(w_1,w_2,\cdots,w_N)$为文档 D 的向量表示或向量空间模型.

2.2　指纹特征提取

垃圾邮件的语言类型繁多,就中国而言,要正确处理简体中文、繁体中文和英文邮件是很重要的,要满足这样要求的邮件系统,如果提取自然词汇为特征通常遇到编码、分词方面的困难,因此引入指纹技术,指纹是用来标识对象的二进制位串.Broder[4]证明了在实际应用中,原文档的相似性可以体现为一部分指纹概要的相似性,以指纹为特征,可以避免语言处理问题.

本文采用了文献[5]中实现的电子邮件指纹计算方法,它的突出优点是具有很高的指纹精度,重复率控制在百万分之一,且具有很好的效率,平均每封邮件处理时间仅为 0.025 s.

2.3　文档向量化表示

文档指纹抽样需要确定取指窗口大小和取样数量.文献[5]的研究表明,实际系统的应用中窗口大小取 6～8 时效果最佳,指纹数量取大于 60 时,对 95%相似的文档,检测相似度可以达到 90%.适量的指纹有助于控制向量机维数大小,本文加入文档长度信息确定指纹抽样的最大数量,计算为 $\mathrm{Finger}_{\max}=1\ 000+1\ 000^{*}\mathrm{fileSize}/20\ 000$,20 000 为邮件近似平均长度.由此可以进一步进行邮件向量化表示,定义指纹向量 X 为:

定义 1　已有全体指纹向量 $\mathrm{GlobeFp}\{g_1,g_2,\cdots,g_n\}(n=1,2,\cdots)$,对当前邮件指纹 $\mathrm{CurrtFp}\{c_1,c_2,\cdots,c_m\}(m=1,2,\cdots)$,若 $g_i=c_j(g_i\in\mathrm{GlobeFp},c_j\in\mathrm{CurrtFp})$,则 $i\in X$.

2.4　在线式 SVM 过滤器

SVM 是批处理模式应用,而 SVM 需要的训练时间是以训练样本数量为基础呈二次方指数增加.近年来不少学者针对传统 SVM 算法,提出了许多改进算法,其中,文[6]提出的 SMO(Sequential Minimal Optimization)算法在实际应用中取得了良好效果.SMO 基于原始的 SVM 模型,优化时使用了块与分解技术,并将分解算法思想推向极致.每次迭代仅优化两个点的最小子集,其威力在于两个数据点的优化问题可以获得解析解,从而不需要将二次规划优化算法作为算法一部分.尽管需要更多的迭代才收敛,但每次迭代需要很少的操作,且不需存储核矩阵,因此在整体速度上有数量级的提高.

基于 SMO 算法的迭代思想进行在线式设计,定义在线子集 T,它包含新到的 m 个邮件样本:

定义 2　大小确定的集合 $T\{t_1,t_2,\cdots,t_m\}$,$m=1,2,\cdots$,新到邮件样本 t_k,则对 T 更新为 $T\{t_2,t_3,\cdots,t_m,t_k\}$.

```
Train (small set);
T_s={}; MaxTrain=m; n=0;
While (1):
Sleep till next email [ex_{n+1}] arrives;
    n=n+1;
SVM_Classify (email [ex_n])
If (T.SIZE()+1>MaxTrain)
T_s={ex_2, ex_3, ..., ex_n};
    Else
T_s={ex_1, ex_2, ..., ex_n}
    SVM_Train (T) start at current
                    SMO model;
Do
```

图 2　在线式 SVM 一般流程伪代码

Fig.2　Online SVM pseudo-code

在线式 SVM 过滤方法一般流程的伪代码如图 2 所示,上述在线模型的优势很明显,过滤器的训练从上一模型开始,更新样本优化信息,大大减少了训练时间.

3　过滤器的性能优化

对于邮件过滤这样的二元分类系统,更关心的是分类的正确性,只要能够正确判别即可,多次迭代并非十分必要,平衡因子也应向降低错误倾斜.文献[7]的研究指出,C 取大于 10 时,SMO 迭代 1 次即可达到良好的分类正确性,在实际测试中也得到了类似结果.

上述条件的减弱可以在一定程度上减少过滤器的运算时间,但过滤的时间耗费还是主要集中在训练优化超平面上.图 3 所示过滤方法中,T 值过小,优化效果不佳,过大又起不到减少运算的作用,而且当连续的同类(同是正常或者垃圾)邮件到来时,T 中某一类别所占比例过重,超平面优化会过于偏向某一边而导致连片错误.

为了解决 T 过小时带来的问题,得益于在线式松弛(ROSVM)[7]的思想,对FSVM子集 T 的更新设置训练集减弱条件,对于分类结果十分明确的邮件样本不加入 T 中,仅对离边界较接近的最近若干邮件样本进行优化,并仅在这样的邮件出现时才训练更新.

对于 T 过大带来的问题,在ROSVM基础上对子集进行比例优化,统计正常与垃圾邮件比例 S,设置 T 中的正常与垃圾邮件样本比例也等于 S,动态均衡训练集,使超平面优化具有全局性,并且能使训练集在较小的 T 样本集条件下,保证训练集的代表性,而不至于因过多同类样本进入训练集而使分割平面过分偏移,这就是动态的子集条件.

定义3 大小确定的集合 $T_1\{t_1,t_2,\cdots,t_m\}$, $m=1,2,\cdots$,和 $T_2\{t_1,t_2,\cdots,t_n\}$, $n=1,2,\cdots$,新到邮件样本 t_k,若 t_k 属于类1,则对 T_1 更新为 $T_1\{t_1,t_2,\cdots,t_m,t_k\}$,若此时 T_1 大于设定最大训练值与该类别的比例和乘积,删去多余邮件样本,对 t_k 属于类2时作类似处理.

```
Train (small set);
T(spam):={}; T(ham):={}; n=0; C(spam)=0;
C(ham)=0; MaxTrain=m; MinMargin=k;
While(1):
    Sleep till next emil [ex_{n+1}] arrives;
    n=n+1;
    res:=SVM_Classify (email [ex_n]);
    C(res)++;
    If (ex_n's margin<MinMargin)
        If (sub=T(res).SIZE()+1-
        MaxTrain* C(res)/(C(spam)+C(ham))>0)
            T(res):={ex_{sub+1}, ex_{sub+2}⋯, ex_n};
        Else
            T(res):={ex1, ex2, ⋯, ex_n}
        Smo:=SVM_Train (Mix(T(ham), T(spam)))
            start at Smo;
Do
```

图3 DFSVM过滤方法伪代码

Fig·3 DFSVM filtering method pseudo-code

优化的过滤流程的伪代码如图3,本方法能使优化子集的分布更具有代表性.对比之前的性能有了大幅提升,过滤效果有明显改善,对比ROSVM则具有了更强的邮件过滤推广能力和真实环境适应性.

4 实验结果

图4对Trec07p数据集按名称排序,用随机数抽样(随机抽样方法下同)1 000封邮件样本并打乱次序,取不同指纹数量进行过滤性能对比,其中,B.S表示原始批量式FSVM过滤方法,O.S表示子集大小为100的在线式FSVM过滤方法,由图4可以看出,后者的正确性并没有因为训练集的减少而降低,时间耗费却大为减少,经过在线式FSVM的开销对指纹向量的敏感度也减少了.

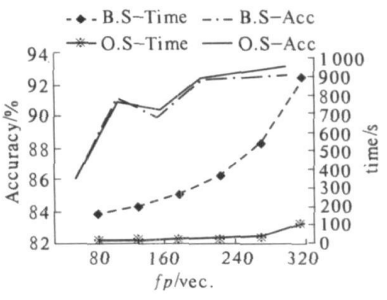

图4 正确率与时间耗费比较

Fig·4 Accuracy & time cost comparison

表1 各种过滤方法对比

Tab·1 Filtering test comparison of some methods %

	Trec07p	SEWM08Corpus	real mail System
Bayes[8]	99.03	98.80	96.50
Bogo	99.10	97.60	92.47
FSVM	99.26	98.72	96.00
DFSVM	99.29	99.01	98.70

表1为各种过滤方法的效果对比.其中,Trec07P,SEWM08数据集随机抽取30 000封样本并随机排序,在真实的邮件系统中,抽取某时刻开始的连续30 000封真实邮件,以到达时间为序,由系统内建过滤器结果(保证99.95%正确率)作为评判标准.对于标准的测试集,SVM的过滤效果基本与传统的过滤方法持平,说明本过滤方法的过滤效果还是十分理想的.在真实邮件系统中,可以看出DFSVM对实际环境的适应性,突出表现了SVM的良好推广性和自主学习性,DFSVM在实际应用中优于FSVM是因为实际中往往很多正常或垃圾邮件连续到来,当集合不太大时,FSVM的分割平面受影响偏移较大,前者则很好应对了这种情况.

对表1中的测试结果绘制了对应ROCA图(ROCA=1-ROC%),见图5,ROCA值越小,性能越高.不

难看出,FSVM 和 DFSVM 的可信度普遍高于同条件下的其他过滤方法.

图 6 的测试环境为上面提到的真实邮件系统,对比测试了动态集优化对 SVM 过滤减少子集大小的作用.测试以时间为序连续测试.对比结果表明,优化后的 DFSVM 比原来的简单条件松弛方法更具稳定性,子集的大小对过滤效果影响平缓,显示了优于 ROSVM 的实际环境适应性.

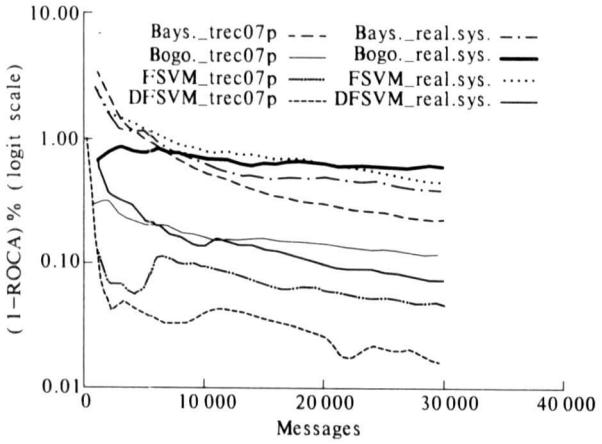

图 5　ROCA 曲线图对比

Fig. 5　ROCA curve comparison

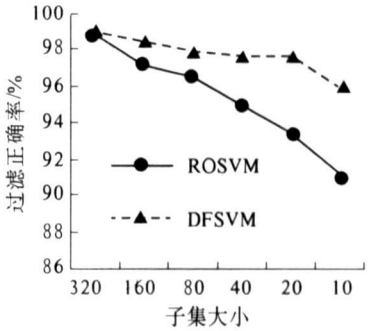

图 6　真实邮件环境下适应性比较

Fig. 6　The adaptability comparison under real system

5　结论

研究实现了一个适用于实际邮件系统的垃圾邮件过滤器 FSVM,并作了优化与测试.测试表明,以邮件指纹为特征的 DFSVM 具有较强的自动学习能力,对于真实系统的变化情况具有更好的推广性.但是在运算性能上还有很多值得研究的方向,例如邮件指纹向量的大小.从图 6 中的测试可以看出,时间开销随着向量维数的增加还是有明显的上升,高维制约着训练集不能过大.如何有效降低指纹向量维数是进一步降低开销所面临的问题.再如临界样本的问题,解决好过滤中的不平等容错性是值得研究的方向之一.

参考文献:

[1] Mitchell T M. Machine Learning[M]. Boston: McGraw-Hill, 1997:52-77.

[2] 王国胜, 钟义信. 支持向量机的若干新进展[J]. 电子学报, 2001, 29(10): 1397-1400.

[3] 刘霞,卢苇.SVM 在文本分类中的应用研究[J].计算机教育,2007(2):72-75.

[4] Broder A Z. On the resemblance and containment of documents[C]//Proceedings of Compression and Complexity of SEQUENCES 1997. Salerno, 1997:21-29.

[5] Fang Weidong. Network spam detection and filtering technologies[D]. Guangzhou: South China University of Technology,2007.

[6] 栾江, 唐常杰,黄晓冬,等.一种增量式支持向量机文本分类模型[J]. 计算机科学,2003,30(B):244-247.

[7] Sculley D, Wachman G. Relaxed online Support Vector Machines for spam filtering[C]//The 13th Annual ACM SIGIR Conference Proceedings. Amsterdam,2007.

[8] Chen Bin, Dong Shoubin, Fang Weidong. Introduction of fingerprint vector based Bayesian method for spam filering [C]//The 4th Conference on Email and Anti-Spam, Mountain View. California,2007.

(下转第 98 页)

[5] Horrocks I, Patel-Schneider P E. A proposal for an OWL rules language[C]//Proceedings of the 13th Int'l World Wide Web Conference. New York: ACM Press, 2004.

[6] Levy A Y, Rousset M C. Combining Horn rules and description logics in CARIN[J]. Artificial Intelligence, 1998, 104(1/2): 165-209.

[7] Motik B, Sattler U, Studer R. Query answering for OWL-DL with rules[C]//Proceedings of the 3rd International Semantic Web Conference. Berlin: Springer, 2004: 549-563.

Formalization and Consistency Checking of UML-Statechart Based on DL-Safe Rule

HE Hong-yue, SONG Zi-lin, ZHOU Bo

(*Institute of Command Automation, PLA University of Science & Technology, Nanjing* 210007, *China*)

Abstract: The semantic information of UML-Statechart is divided into static aspect and dynamic aspect. The static aspect is expressed by a knowledge base of description logics, and the dynamic aspect is expressed by DL-Safe rule. An algorithm is proposed for checking the consistency of UML-Statechart, which can use the DL-Safe rule to reason the knowledge base. Finally, the feasibility of the algorithm is analyzed theoretically.

Key words: UML-Statechart; description logic; DL-Safe rule; consistency

---

(上接第93页)

Online Filtering Method and Optimization Based on SMO and E-mail Fingerprint

ZHU Qing-rong[1], DONG Shou-bin[2], CHEN Bin[1]

(1. *School of Computer Science and Engineering, South China University of Technology, Guangdong* 510640, *China*; 2. *Guangdong Computer Network Key Lab, Guangdong* 510640, *China*)

Abstract: The finger features vectoring and the SVM filtering method are proposed on the spam, and an online SVM spam filter called FSVM is designed and implemented, which attains the corresponding performance as the classic method in the online spam filtering. In view of the computing speed of SVM filtering, a dynamic example set reduction method(DFSVM) is given out for SVM filtering method based on the original SMO algorithm, which can greatly reduce the computing cost and keep the corresponding performance. The experiment and comparison test running on the standard corpus is given out in the actual mail system. It is proved that the optimization can improve the accuracy of SVM classification.

Key words: spam filtering; SVM; conditional relax; dynamic subset