doi:10.3969/j.issn.1672-1292.2017.03.012

基于分类邮件代理 MCP 的垃圾邮件动态检测

陈　斌,东一舟,毛明荣

(南京师范大学信息化建设管理处,江苏 南京 210023)

[**摘要**]　针对互联网邮件中垃圾邮件占比暴增的问题,提出了一种基于分类代理 MCP 的动态检测算法. 该方法基于近半年时间对校园网邮件宿主机及各代理虚拟机间传输的会话日志的采集,针对记录中各类投递状态及状态消息集进行了行为分析,最终达到对垃圾邮件的有效检测,从而为分拣提供依据. 实验结果表明,在持续进行了若干频次的分类策略调节后,该检测算法的准确度可高达 96.1%. 该设计可对垃圾邮件宿主机及代理虚拟机的行为进行有效检测,从而彻底抑制垃圾邮件的产生.

[**关键词**]　垃圾邮件宿主机,代理虚拟机,简单邮件传输协议会话,分类代理,分类器,邮件状态信息
[**中图分类号**]TP39　[**文献标志码**]A　[**文章编号**]1672-1292(2017)03-0080-07

Dynamic Detection of Spam Based on Classified Mail Proxy MCP

Chen Bin, Dong Yizhou, Mao Mingrong

(Informatization Office of Nanjing Normal University, Nanjing 210023, China)

Abstract: In order to solve the problem of increasing the proportion of spam in Internet mail, a dynamic detection algorithm based on MCP is proposed. Based on the collection of the session logs collected from the campus network mail hosts and virtual agents in the past six months, the method analyzes all kinds of delivery status and status message set in the record, and achieves the result of effective spam detection finally, so as to provide the basis for sorting. The experimental results show that after a certain number of frequency classification strategy is adjusted, the highest accuracy of the detection algorithm is up to 96.1%. The design detects the behavior of spam host and virtual machine effectively, and completely suppresses the generation of spam.

Key words: spam host, proxy virtual machine, smtp session, classified agent, classifier, mail status message

随着电子邮件的普及,垃圾邮件便一直相伴相生,日益困扰着邮件使用者. 按照 Anti-Abuse 消息工作组 2011 年度发布的调查报告显示,互联网电子邮件总量中超过 90%的都是垃圾邮件[1]. 这极大地浪费了互联网有限的带宽和邮件服务商的存储,也对正常用户的权益造成了强烈的影响. 该问题最常用的解决方法,是在邮件服务器、代理或客户端对邮件进行过滤,该措施执行效果很大程度上受邮件服务商、代理或客户端过滤能力的限制[2]. 以终端过滤器的方式可相对准确地隔离或过滤垃圾邮件,但因其未能从根源对垃圾邮件的产出进行控制,故而还是吞噬着相当一部分的网络带宽,这对于资源利用要求极高,供给又较为有限的校园网来说,造成的影响更是明显. 对垃圾邮件源的遏制,目前已成为该领域急需解决的问题.

对于如何检测及认定垃圾邮件宿主机,有两个疑难问题需要解决:选择垃圾邮件宿主机的关键属性[3];建立所选定的关键属性的检测模型[4],该模型需要能匹配动态变化的垃圾邮件行为. 本文将从邮件服务器及代理虚拟机中部署的分类代理的运作细节,以代理服务端及客户端采集的海量数据集所产生的简单邮件传输协议(simple mail transfer protocol,SMTP)日志为基础,对各种情况下垃圾邮件行为特征进行归类学习和动态调整. 该设计可帮助校园网络的管理者检测垃圾邮件宿主机,从而抑制这些宿主机的行为.

收稿日期:2017-01-19.
基金项目:中国高等教育学会教育信息化专项课题(2016XXYB02).
通讯联系人:陈斌,博士,工程师,研究方向:云计算技术. E-mail:njnuchenbin@ njnu.edu.cn

1　分类邮件代理 MCP 模型

对垃圾邮件的过滤检测,首先需要一套健全有效的解析归类和存储管理机制[5]. 本文提出了分类邮件代理 MCP 模型,如图 1 所示. 该模型划分为 3 个层次,其中 MCP 内核是本文的研究重点. 分类邮件代理 MCP 架构的客户端层汇集了各种类型终端及各类邮件编辑器,由它们发起的邮件报文从不同的源头在此汇集,继而同步发送给邮件分类代理 MCP 控制器. 邮件分类代理 MCP 控制器在 MCP 架构中处于核心位置,它由轻量级调节器、邮件代理虚拟机管理器及 MCP 内核所构成. 轻量级调节器首先向邮件代理虚拟机管理器发送“控制器空闲”报文,邮件代理虚拟机管理器中的虚拟机管理器将协调 Job 池和资源池通信,并将当前待处理邮件返回给调节器,继而调节器将利用邮件规则库及邮件历史库快速按既有规则拆分邮件头、邮件体及邮件地址. 拆分工作由不同拆分器分步实施,拆分结果将传递给 MCP 内核中的垃圾邮件解析器进行解析,解析结果直接回送给邮件代理虚拟机管理器,其将会对被通知为垃圾邮件的目标邮件进行标注,并阻止其继续向 SMTP 服务端层输送,同时向垃圾邮件历史库中添加拆分出的邮件地址、邮件宿主机 MAC 地址、邮件宿主机 IP 地址,并将结果向客户端层反馈. 只有由解析器认定为非垃圾邮件的,才会进一步向 SMTP 服务端层传输,将邮件存入 SMTP 服务器组下邮件服务器的邮件虚拟机中.

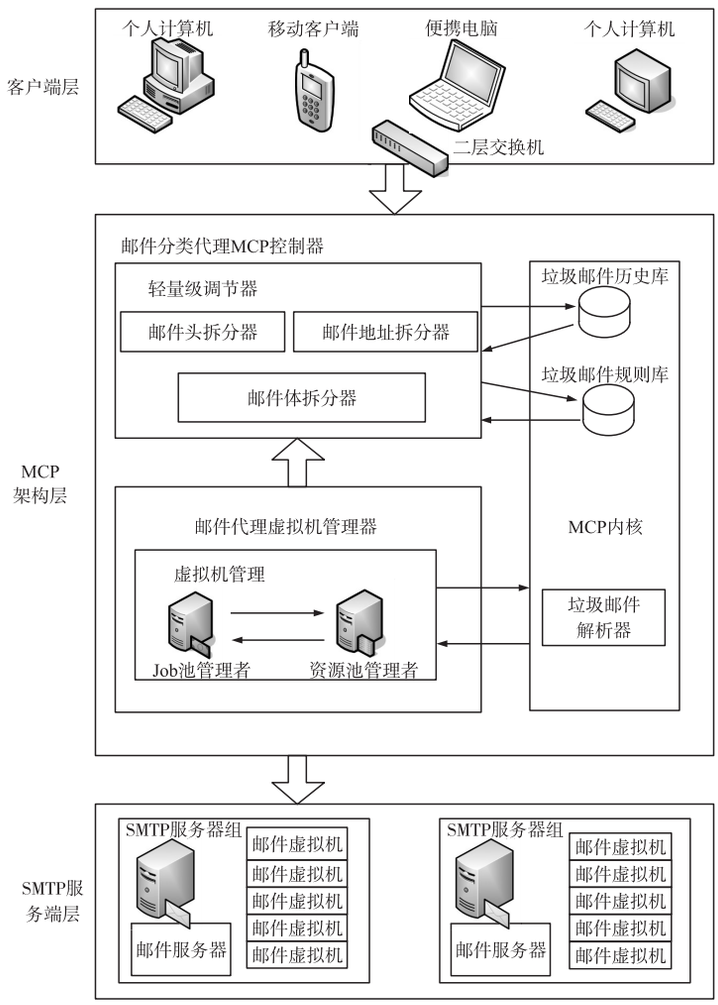

图 1　分类邮件代理 MCP 框架

Fig. 1　MCP framework based on classified mail agent

2　垃圾邮件动态检测分析

在图 1 的架构中,垃圾邮件动态检测分析主要是依托 MCP 内核中的垃圾邮件解析器,并按照一定规则,根据既定检测条件进行算法解析的. 其中垃圾邮件行为执行失败消息特征分析及检测分为 5 个阶段,即:(1)通过干扰检测系统对校园网络与互联网之间的 SMTP 交互日志进行捕获;(2)从日志中提取出校

园网内部宿主机与外部宿主机初始会话中的 SMTP 报文;(3)计算来自于每台内部宿主机个体的多种类型 SMTP 会话投递成功及失败消息的数量及类型;(4)按照宿主机状态,通过行为检测的方式,针对训练集中的内部宿主机打上垃圾邮件源标签或正常宿主机标签;(5)在此基础上,可通过增量学习算法持续检测校园中的垃圾邮件宿主机. 特征属性分析和检测常用的两种分类方法分别为基于规则的方法和基于机器学习的方法[6]. 针对分类后的特征结果集合,淘汰历史样本集中的非支持向量,将支持向量同新增样本一起训练,以达到增量学习的目的. 一旦通过检测确认当前存在垃圾邮件行为,则会将发现的垃圾邮件宿主机列表发送给管理者从而对其加以抑制.

当一个 SMTP 响应消息从服务器到达本地后,其报文头中会携带着一个 SMTP 邮件分发会话的成功与失败状态标志位[7],若总是发生失败的情况,则说明异常的可能等级较高. 起初试图通过 SMTP 响应码对日志中的失败消息进行分类,但事实上响应码和真正的失败原因的关系是多元的[8],无法满足一对一关系的需求. 表 1 中列出了一些实验过程中的实例,其中列举了很多比响应码更合理的键值选项,这些键值选项都是从多种失败消息中提取并汇总的. 此外,针对同样的失败原因邮件服务器给出的响应消息可以不同,例如由于黑名单而造成的邮件阻塞在表 1 中就给出了多种列举,所以在实验过程中通过人工识别键值的方式,根据语义分析对垃圾邮件进行了标注,并基于失败原因的键值域对响应消息进行组织. 作为一个检测系统,在默认 SMTP 策略脚本中并没有将其期望回复结果列出[9],只是在默认脚本中添加了附加键值信息,以帮助其识别垃圾邮件会话. 附加信息包括了日志是否成功转发、邮件头的格式、邮件回复路径、发送者以及接收者的地址和主题等.

表 1　相同 SMTP 响应码情况下不同响应消息示例

Table 1　Instance of different response messages in the same SMTP response code

[命令,响应码]	响应消息
[DENY,602]	被垃圾邮件监控拒绝/由于接入受限,被垃圾邮件过滤器拒绝/由于 IP 地址黑名单而未被接收
[DENY,605]	接入 IP 地址未知异常/发送者的地址域不存在/接入受限(含黑名单控制)
[RECPT,318]	域未找到/接收者地址被拒绝/本地接收表中用户未知错误/由于本地策略导致的邮箱错误
[RECPT,319]	超出发送率的限制/邮箱忙,稍后重试/地址在黑名单中/达到了单用户每小时邮件最大限制数量
[RECPT,602]	未知用户被拒绝/访问邮件服务器被拒绝/访问邮件系统服务被拒绝

从 SMTP 会话日志记录开始,会话状态统计可以启动,主要是指会话成功或失败情况消息统计[10]. 基于 SMTP 转发规则定义的消息语义有多种类型,按照实际情况可分为六大类,键值是以人工的方式设置的. 关键的键值子集在表 2 中列出,在实验的分类处理过程中以失败消息里的模糊匹配结合正则表达式的方式进行键值的匹配和查找. 由于海量日志中的隐性键值是没有办法穷举的[11],故实验当中的键值并不是完备的.

表 2　针对各种失败原因的语义键值归类

Table 2　Classification of the semantic key for the various failure reasons

失败原因	键值域
域认证错误	DNS/发送认证域/主机名等
IP 黑名单	黑名单/阻塞单元/灰名单等
接收者未找到	无效用户/异常地址/接收者账户/受限用户等
比率受限	收发率受限/连接受限/单位时间访问受限等
复杂未明原因	异常拒绝/策略异常/未经批准拒绝等
异常命令响应	900

基于语义检索的自然语言处理技术,可协助上下文语句的检索匹配,针对邮件不同区域的检索,包括邮件标题和正文,进行概念分析、分类、标引、描述和处理,形成具有语义关联的资源元数据集合,并使用 RDF 和 OWL 语言进行语义层面的表述和描述[12],通过适应于邮件类型的自然语言关系模型学习处理,结合针对邮件上下文的语义分析,形成用以与分类器预定义分拣数据集相比较匹配的语义关键词或语句. 回复消息通常由若干短句构成,消息中只可能出现一次失败标志的键值,其中隐含的失败原因将对垃圾邮件宿主机行为产生潜在的影响. 在失败消息中,标识为邮件接收者未找到的类别通常有以下 3 种情况:目标邮件地址已过期停用,但垃圾邮件制造源依然在持续向其发送邮件;目标邮件地址由于解析错误而造成拼写问题,这种解析错误是由于垃圾邮件制造源的恶意探测器在网络上扫描目标源后,对其地址试探性轮询分析产生的过程结果;另外,垃圾邮件制造源也会随机产生邮件地址作为目标邮件地址,对于之前已发送过垃圾邮件的宿主机而言,对端服务器可能会将其列入 IP 黑名单,在每次接收到邮件后的检索过程中,可能会对这些宿主机的邮件进行退信处理.

针对外部邮件服务器,由于其响应消息中的一些情况及其状态并不确定,分类时将其列入单独的类

别. 另有一类不常用命令对,例如 SMTP 会话数据无响应,这类错误通常与响应码 900 相关联,所以将该响应码作为该分类的键值. 对于校园里的每一个 IP 地址,均可按照相应键值计算其回复消息数量,该统计结果可用以识别疑似垃圾邮件宿主机. 实验中使用了八维特征向量来描绘内部宿主机的每一个实例的 SMTP 会话,其中第一元记录的是成功投递情况,第二至第七元记录的是失败消息的 6 种分类(见表 2),第八元为宿主机是否为邮件服务器标志位. 在表 2 中列举的只是导致失败的一部分原因,而垃圾邮件会话的动态检测机制对象是多样的,其随着实际情况的演变也会动态变化.

通过使用内部宿主机行为结果作为训练集,并手工检测以下邮件头区域,以建立基本的垃圾邮件宿主机判断机制. 主要邮件头区域为:主题,通过检查邮件消息的主题,判断其是否疑似为垃圾邮件,例如其是否包含了攻击性关键词;发送者,垃圾邮件发送者通常都会对自己进行伪装,例如使用随机产生的邮件地址或者域名;接收者,该域可以在垃圾邮件中被随机产生,所以一旦检测到有序列化的随机目标的行为产生,则可断定其来源为垃圾邮件宿主机. 同时再通过扫描传输控制协议(Transmission Control Protocol, TCP)绑定的 25 号端口,并检查发送源宿主机域名称,从而判断宿主机是否为 SMTP 服务器.

邮件分类器需保持不断更新以识别最新的垃圾邮件行为,在此使用了被动攻击增量学习算法,用以对当前分类器依赖的邮件规则库进行调整. 针对样本实例的更新,需纠正当前分类器的预测错误,并通过主动调整来更新当前分类器,当前已被最小化错误处理后的分类器将作为下一次数据采集所选择的分类器使用,从而达到提高分类精确度的目的. 在对上述方法进行建模之前先进行标记的定义,数据集 P_t 在周期 t 时被采集,$|P_t|$ 的标签都是成对的,$\{(u_1,v_1),(u_2,v_2),\cdots,(u_{|P^t|},v_{|P^t|})\}$,实例数组中的 u_n 是宿主机在八元组周期观测值条件下的 SMTP 行为,相应的类标签 v_n 是垃圾邮件或非垃圾邮件标识符. k^t 为周期 t 下分类器组成向量的权重,当每一个实例 $u_n \in P^t$ 产生时,被更新的分类器 k^{t+1} 都会修正之前 k^t 分类器的错误,k^t 则进行最小化修正. 若 u_n 从 k^t 获得了不正确的预测值,则 k^t 的调节将被 u_n 的自身边界值所取代. Q 为 k^t 的基于(u_n,v_n)键值对的更新模型,分类器优化调整建模可表示为:

$$Q(k^t,(u_n,v_n),P^t)=\operatorname{argmin}\left\{\frac{1}{2}\|\bar{k}-k^t\|^2+E_0\sum_{u_m\in P^t,u_m\neq u_K}s(k,(u_m,v_m))\right\},$$

式中,E_0 是一个用来决策控制分类器偏差与预测错误校正之间如何权衡的变量,$s(k,(u_m,v_m))$是其关键偏差函数:

$$s(k,(u_m,v_m))=\begin{cases}0 & v(k\cdot u_m)\geqslant 1;\\ 1-v_m(k\cdot u_m) & v(k\cdot u_m)<1.\end{cases}$$

按上式对 k^t 对应的分类器进行更新时,将$\{Q(k^t,(u_K,v_K),P^t):1\leqslant n\leqslant|P^t|\}$作为新分类器的备选键值组对,为防止新的分类器被当前分类器过多影响,选择策略会在 P^t 中挑选最合适的分类器,当具有同等高分类准确性的分类器大于一个时,则可选择该分类器中与 k^t 差别最小的,因此新的分类器 k^{t+1}可按照该策略从备选分类器中进行选择. 按上述对基础过程的描述,垃圾邮件过滤器的更新是使用增量学习算法来完成的,其流程为:

步骤 1　初始化数据集 P^t、分类器 k^t 以及分类优化调整内核函数 Q;

步骤 2　在每一个周期 t,按照所采集数据的具体不同情况对数据集 P^t 进行更新,用于增量学习;

步骤 3　在生成新样本实例时,将首先使用内核函数 Q 将现有样本映射到更高维度空间,计算 k^t,并更新 $\bar{k}_n$;

步骤 4　后续更新增样本时,均采用步骤 3 中的方式计算 k^t,并更新 $\bar{k}_n$.

上述增量学习算法的形式语义建模描述如下:

(1) Initialize:$k^1=(0,0,\cdots,0)$;

(2) for $t=1,2,\cdots$do

(3) Recpt_Collect_data(P^t);

(4) Predict($\hat{v}_u$) = sign($k^t\hat{u}_n$) $u_n\in P^t$;

(5) Get $P^t=\{u_n\,|\,u_n\in P^t,v_u\neq\hat{v}_u\}$;

(6) for each $u_n\in P^t$ do

(7) $\tau_n = \dfrac{1 - v_n(k^t \cdot u_n) - v_n u_n \sum\limits_{u_m \in P^t, u_m \neq u_n} v_m u_m}{\| u_n \|^2}$;

(8) $\bar{k}_n = k^t + \sum\limits_{u_m \in P^t, u_m \neq u_n} v_m u_m + \tau v_n u_n$;

(9) end

(10) choose

(11) $k^{t+1} = \arg \sum\limits_{u_m \in P^t} s(k,(u_m,v_m)) + \| k - k^t \|$;

(12) end

在每一个周期 t,数据集 P^t 都会采集数据用以更新当前分类器 k^t,预测标签显示为错误的 SMTP 行为在算法中的(4)~(5)行中由 k^t 和 P^t 给出标识,对于 SMTP 行为实例 $u_n \in P^t$,当前分类 k^t 是单独作为备选分类器 $\bar{k}_n$ 进行更新的,更新策略是按照前述公式进行的. 若其在算法第(10)行选择并获得了最小的预测错误个体,最终将如(7)~(8)行展示的,分类器 $\bar{k}_n$ 本身将被选择为 k^{t+1}分类器. 特别是在首轮周期,k^1 被初始化为(0,0,…,0),其预测结果往往都是正向的. 继而 k^1 被更新调整为第一个已更新过的分类器 k^2,该更新的准确性依赖于可以导致最小结果的正向实例 $\| \boldsymbol{k}^2 - \boldsymbol{k}^1 \|$ 的结果. 此外,除最小化分类器偏差之外,同时对此前错误的分类器进行了修正.

3　实验及数字化论证

实验环境基础配置为:8 核 4.8GHz×8CPU,128GB 内存,32TB 硬盘,双 200GB/s 网卡的机架型服务器. 虚拟机操作系统选择 64 位的 Linux,虚拟机最大并发数为 1024 台. 实验采用基于径向基内核(Radial Basis Function,RBF)的支持向量机(Support Vector Machine,SVM)以实现分类器的设计,同时使用 MATLAB 算法分析包对读取参数与内核参数进行有效开采和识别. 在实验中,MCP 内核分类器是定期增量更新的,更新周期为 2 h,更新对象为打了标签的数据集,分类器 k^t 在周期 t 中由实例标签键值对 P^t 进行更新. 增量学习算法的性能在不同设置条件下对 MCP 内核分类器错误修正的实际效果是不同的,在选择潜在分类器时起到了最小化评估错误的作用. 按照分类器性能进行评估时,需要同时强调垃圾邮件和非垃圾邮件宿主机的分类效果,因此测量平均分类准确率也是由这两大类别共同计算得出的. 表 3 中列出了周期为月计的实验数据集,每行中的数字是具有邮件行为的宿主机数量、垃圾邮件宿主机数量以及非垃圾邮件宿主机数量. 对于每一个实例来说,数据集中的 u_n 包含了八元组向量中的 SMTP 行为,每一个 u_n 的实例都被打上了垃圾邮件或非垃圾邮件标签.

表 3　201605 至 201610 校园网内垃圾邮件宿主机统计表

Table 3　Campus network spam host statistics from 201605 to 201610

样本序号	样本名	宿主机数量	垃圾邮件宿主机数量	非垃圾邮件宿主机数量
1	201605	329	44	285
2	201606	334	49	285
3	201607	318	43	275
4	201608	218	49	169
5	201609	347	58	289
6	201610	302	46	256

在不同 E_0 和 E 调节系数情况下,针对混合邮件集的增量学习算法检测结果如表 4 所示. 实验中尝试了多种 E_0 和 E 值情况下的调节效果,在表 4 中只列出了部分有代表性结果. 根据调节效果显示,大多数分类精确度都是通过 $t=2$ 或 $t=3$ 情况下的增量学习分类调节后提升的,增量学习算法当 $E=1$ 时有着最优的检测能力,根据结果显示,从第二个周期开始平均分拣准确度在 85%以上,并保持在稳定水平. 此外,当 $E_0=0$,$E=1$ 以及 $E_0=0.25$,$E=1$ 时,较 $E_0=0.5$,$E=1$ 时准确性更稳定. 对于分类器产出者来说,当一个新的分类器衍生出之后,产出者错误检测修正权重将会变小以避免过拟合问题的出现,增量学习算法在本实验中保守地采取了最小化调节效果.

表 4　增量学习算法在不同参数情况下的检测结果

Table 4　Detection results of incremental learning algorithm under different parameters

增量参数(E_0,E)	P_1/%	P_2/%	P_3/%	P_4/%	P_5/%	P_6/%
(0,1)	78.53	86.32	88.13	87.49	85.57	86.11
(0.25,1)	78.53	83.10	82.32	83.31	83.59	82.37
(0.5,1)	78.53	81.04	82.20	83.76	79.11	83.25
(0.25,0)	78.53	70.12	73.41	75.92	76.01	73.35
(0,0)	78.53	78.53	78.53	78.53	78.53	78.53

表 5 列出了在不同增量学习配置类条件下的细节区别,主要为 $E_0=0.25$,$E=1$,以及 $E_0=0$,$E=1$ 两种情况. 由于优化改进了关键偏差函数,从表 5 中可以看出非垃圾邮件宿主机(none spam host,NSPH)的识别准确度普遍低于 74%,有一些属性不确定的宿主机由于接收消息中包含了失败响应标记因而也被认定为垃圾邮件宿主机,非垃圾邮件宿主机可能会误导预测结果并降低综合检测准确度. 在实践中,类似的错误识别情况已通过白名单的方式进行了纠正,有效提升了综合准确度,对垃圾邮件宿主机(spam host,SPH)的 3-4 个周期的平均检测识别准确度提升至 93%以上,最高值可达 96.1%. 而据各类文献记载分析,相关垃圾邮件检测系统准确度一般低于 90%,文献[10]提出的基于改进堆叠自动编码机的垃圾邮件分类技术最高准确度可达 97.66%,但其强烈依赖于熄火率动态函数的选择,而熄火率动态函数与实现环境及算法的选择又有牵制关系,普适性受到一定的影响. 垃圾邮件宿主机与非垃圾邮件宿主机基于不同增量学习配置条件下的调节预测准确度结果如图 2 所示.

表 5　垃圾邮件宿主机与非垃圾邮件宿主机的检测结果

Table 5　Detection results for spam host and non-spam hosts

增量参数(E_0,E)	P_1/%	P_2/%	P_3/%	P_4/%	P_5/%	P_6/%
(0,1),SPH	54.7	90.1	93.4	94.9	96.1	95.2
(0,1),NSPH	93.2	76.3	74.2	71.7	73.1	71.1
(0.25,1),SPH	54.7	88.7	90.5	93.0	93.9	92.7
(0.25,1),NSPH	93.2	74.5	73.2	72.1	73.7	72.8

除了对特征权重的重要性和影响性进行了研究,实验还通过手工检测 SMTP 日志的方法,对可能误导检测结果的因素进行了分析,主要有接收者未响应应答、邮件服务器黑名单应答、垃圾邮件宿主机接收到新的失败响应几类情况. 此外,一个新的宿主机在观测周期内,只会初始化少量的 SMTP 会话,其观测行为的缺乏可能是错分类中偶然的结果. 垃圾邮件发送者往往会规避检测手段,但规避不可能总成功,因为其无法控制外部邮件服务器,根据表 2 中的失败原因键值归类情况,垃圾邮件发送者需要通过域认证,在垃圾邮件会话中避开非正常的命令,并频繁拒绝传递垃圾邮件. 此外,邮件服务器列出了一个黑名单以阻止垃圾邮件的进入企图,宿主机在控制垃圾邮件转发的同时,也会不断补充更新黑名单内容.

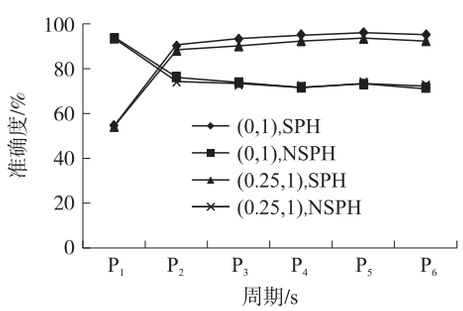

图 2　SPH 与 NSPH 基于不同增量学习配置调节的预测准确度

Fig. 2　Predictive accuracy based on different incremental learning configurations in SPH and NSPH

本实验与同类垃圾邮件分拣实验相比,最根本的不同是使用了以增量学习算法为基础的分类器,而其他实验主要以堆叠器编码机为主,将该分拣器植入分类代理 MCP 以构成动态检测核心. 相比较而言,使用堆叠器编码机的分类器其优点是分拣速度快且稳定,准确度在有条件的背景下能快速达到较高值,但其缺点在于通常需要与分拣对象的数据集属性强关联,其针对类似典型的 Enron 数据集的效率很高,但关联其他类型数据集则效果并不明显. 而使用以增量学习算法为基础的分类器,其与数据集属性无需强关联关系,且适应于各类数据集,效果差异不明显,但分拣准确度提升和稳定需要经历一定周期.

本实验的检测工作依赖于对独立宿主机的统计,这些独立宿主机以 IP 地址为识别符号,因而对主机地址做过网络地址转换(network address translation,NAT)映射的内网地址,或对使用了动态主机配置协议(dynamic host configuration protocol,DHCP)获取的地址而言,可能会存在不确定性. 对于前者来说,网络管

理员仍可识别近似源地址继而分析其垃圾邮件行为,但需对 NAT 所对应的真实设备进行处理;对于后者而言,垃圾邮件宿主机可能被认为来自于多个源,网络管理员同样可对实际分配 IP 地址的 DHCP 服务器进行分析,以查找到真实的地址源. 最难处理的情况是由移动终端获取到一个动态 IP 地址,且该地址又是做过 NAT 映射的. 当一个移动终端在某一个点稍作停留,其垃圾邮件发送行为就可能导致网络拥堵,造成非常严重的结果. 因此,一个有效的解决方案是,通过灰名单的方式仅仅阻塞该 IP 地址接收拥堵失败消息,若该发送源是一个正常的邮件服务器,一段时间后它将会再次发起请求,该方式至少遏制了垃圾邮件移动终端对其停留区域其他终端的垃圾邮件的恶意转发行为.

4 结语

本文使用了基于分类代理 MCP 的动态检测模型来对垃圾邮件进行识别分拣及过滤,该工作基于大量的 SMTP 会话中嵌套的成功及失败转发消息,以及其中嵌入的邮件服务器信息. 实验结果显示,该方法可有效检测到垃圾邮件宿主机并对其行为加以抑制,故而从源头对邮件的产生进行干预及控制. 前文阐述中也提及了该检测模型针对非垃圾邮件的检测准确度较低,作为实际应用中的检测模型而言,仅有单向的垃圾邮件高识别度是不够的,往往还需要双向检测对比结果加以保证,这也是接下来的研究方向.

[参考文献](References)

[1] ALBERTO C,TERESA G V. Activity recommendation in intelligent campus environments based on the Eduroam federation[J]. Journal of ambient intelligence and smart environments,2016,8(2):35-46.

[2] STEPHEN M A,MATTHEW F,CLIVE G. Comparison of a cost-effective virtual cloud cluster with an existing campus cluster[J]. Future generation computer systems,2014,41(1):65-78.

[3] DEBELE F G,LI N,MEO M,et al. Experimenting resource on demand strategies for green WLANs[J]. Sigmetrics Perform Eval Rev,2014,42(3):61-66.

[4] 张绍成,刘威,程子傲,等. 代价敏感多主题学习的邮件过滤算法[J]. 华中科技大学学报(自然科学版),2016,10(44):176-180.
ZHANG S C,LIU W,CHENG Z A,et al. A spam filtering algorithm based on cost ensitive learning for multiple topics[J]. Journal of Huazhong university of science and technology(natural science edition),2016,10(44):176-180.(in Chinese)

[5] 梅峥,厉启鹏,李西太,等. 电力消息邮件体系架构及关键技术[J]. 电力系统自动化,2016,20(40):126-132.
MEI Z,LI Q P,LI X T,et al. Architecture and key techniques of message mail in electric power systems[J]. Automation of electric power systems,2016,20(40):126-132.(in Chinese)

[6] 王友卫,刘元宁,凤丽洲,等. 基于用户兴趣度的垃圾邮件在线识别新方法[J]. 华南理工大学学报(自然科学版),2014,42(7):21-27.
WANG Y W,LIU Y N,FENG L Z,et al. A novel online spam identification method based on user interest degree[J]. Journal of South China university of technology (natural science edition),2014,42(7):21-27.(in Chinese)

[7] PAN D R,FU M,SUN J J. A campus community-based mobility model for routing in opportunistic networks[J]. KSII transactions on internet and information systems,2016,10(3):1 034-1 050.

[8] KOSTA S,MEI A,STEFA J. Large-scale synthetic social mobile networks with swim[J]. IEEE transactions on mobile computing,2014,13(1):116-129.

[9] AHMED A,RAMI L,RAOUF B. Flow-based management for energy efficient campus networks[J]. IEEE transactions on network and service management,2015,12(4):565-579.

[10] 沈承恩,何军,邓扬. 基于改进堆叠自动编码机的垃圾邮件分类[J]. 计算机应用,2016,36(1):158-162.
SHEN C E,HE J,DENG Y. Spam filtering based on modified stack auto-encoder[J]. Journal of computer applications,2016,36(1):158-162.(in Chinese)

[11] WEN T H,WEI C B C. Incorporation of spatial interactions in location networks to identify critical geo-referenced routes for assessing disease control measures on a large-scale campus[J]. International journal of environmental research and public health,2015,12(1):4 170-4 184.

[12] ELISA R A,GUILLERMO I A,JOSE M G G. All-path bridging:path exploration protocols for data center and campus networks[J]. Computer networks,2015,79(1):120-132.

[责任编辑:严海琳]