doi:10.11959/j.issn.1000-436x.2016208

基于历史数据的异常域名检测算法

袁福祥[1,2]，刘粉林[1,2]，芦斌[1,2]，巩道福[1,2]
(1. 解放军信息工程大学网络空间安全学院，河南 郑州 450001；2. 数学工程与先进计算国家重点实验室，河南 郑州 450001)

摘 要：提出一种基于域名历史数据的异常域名检测算法。该算法基于合法域名与恶意域名历史数据的统计差异，将域名已生存时间、whois 信息变更、whois 信息完整度、域名 IP 变更、同 IP 地址域名和域名 TTL 值等作为主要参量，给出了具体的分类特征表示；在此基础上，构建了用于异常域名检测的 SVM 分类器。特征分析和实验结果表明，算法对未知域名具有较高的检测正确率，尤其适合对生存时间较长的恶意域名进行检测。

关键词：异常域名；域名历史数据；特征；检测
中图分类号：TP309 **文献标识码**：A

Anomaly domains detection algorithm based on historical data

YUAN Fu-xiang[1,2], LIU Fen-lin[1,2], LU Bin[1,2], GONG Dao-fu[1,2]
(1. School of Cyberspace Security, PLA Information Engineering University, Zhengzhou 450001, China;
2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China)

Abstract: An anomaly domains detection algorithm was proposed based on domains' historical data. Based on statistical differences in historical data of legitimate domains and malicious domains, the proposed algorithm used domains' life-time, changes of whois information, whois information integrity, IP changes, domains that share same IP, TTL value, etc, as main parameters and concrete representations of features for classification were given. And on this basis the proposed algorithm constructed SVM classifier for detecting anomaly domains. Features analysis and experimental results show that the algorithm obtains high detection accuracy to unknown domains, especially suitable for detecting long lived malicious domains.

Key words: anomaly domain, domain historical data, feature, detection

1 引言

近年来，随着网络技术的不断发展，网络中出现的各种威胁也不断增加，如恶意软件[1]、僵尸网络[2]和木马[3]等。其中，僵尸网络和木马在发动诸如垃圾邮件、网络钓鱼[4]等恶意行为的过程中往往都通过域名系统，即 DNS 解析域名获取回连服务器的 IP 地址，从而隐藏躲避在僵尸代理身后的命令与控制服务器(C&C, command-and-control server)[5~8]，回连控制端接收控制消息或回传盗取的数据信息，躲避检测和封堵，提高自身的顽健性，延长生命周期。由此可见，域名在僵尸网络及木马发动攻击行为的过程中发挥了至关重要的作用，因此，如何对此类攻击中所使用的域名进行检测，对于发现并防范僵尸网络及木马的传播具有极为重要的意义。

目前，针对异常域名检测的研究大致包括基于域名自身特性、域名网络行为特性的检测方法等。如文献[9]主要从域名的字符构成角度，通过分析合法域名与算法产生的恶意域名在字符构成方面的差异，对恶意域名进行检测，实验表明，该方法能够检测出网络中出现的算法产生的恶意域名。文献[10]从域名的注册信息等特性出发，将输入的已知恶意

收稿日期：2015-12-21；**修回日期**：2016-09-12
基金项目：国家自然科学基金资助项目（No.61379151, No.61272489, No.61302159, No.61401512）；河南省杰出青年基金资助项目（No.144100510001）
Foundation Items: The National Natural Science Foundation of China (No.61379151, No.61272489, No.61302159, No.61401512), The Excellent Youth Foundation of Henan Province of China (No.144100510001)

域名作为种子，通过种子域名的域名服务器特征和注册信息特征推测出与该种子可能为同一批的恶意域名，并利用相关的黑名单对推测结果进行验证，结果表明，73%的推测域名最终出现在黑名单中。文献[11]基于木马使用域名进行回连这一事实，对木马域名进行分析，提取出域名使用时间、访问域名周期和域名 IP 地址所属国家变更等特征，实验结果表明，该检测准确率与之前的方法相当。文献[12]设计了一个域名信誉系统——Notos，该系统使用被动 DNS 查询的数据，分析域名的网络特征，为已知域名建立模型，并用该模型为新域名计算信誉分数，该检测方法准确率较高，并能够在恶意域名被列入黑名单几周甚至几个月前检测出恶意域名。文献[13]提出了 Kopis 检测系统，该系统通过对顶级域名服务器以及权威域名服务器进行监测获取数据提取特征，能够检测恶意软件相关的域名，该方法可以从全球的角度对域名的请求、解析等网络行为进行分析，相比于其他方法监测范围更广。文献[14]提出了 Exposure 检测系统，该系统通过对真实的 DNS 流量数据分析，从 DNS 应答以及域名构成等方面出发提取特征，该方法能够对真实网络中的域名进行高效的检测。文献[15]基于域名的长度、域名中存在的特殊字符、域名的被解析次数、被解析时间以及被解析出的 IP 的变化等来构建特征，该检测能够从真实的 DNS 数据中检测出恶意域名。

纵观现有的异常域名检测方法都各有所长，且有各自适用的范围。相对而言，基于域名网络行为分析的检测方法(如文献[13])，其检测正确率较高，且适用范围较广，但该类方法需要从顶级域名服务器、权威域名服务器或者递归解析域名服务器获取大量的域名解析数据。然而，无论是顶级域名服务器、权威域名服务器还是递归解析域名服务器，其流量数据都很难获取。本文基于域名的自身特性及网络行为特性相关的历史数据，根据合法域名与恶意域名 whois、解析 IP 变更及 *TTL* 等信息存在的差异，提出了一种恶意域名检测算法，该算法通过对域名的 whois 信息、域名解析的 IP 变更信息、同 IP 地址域名数量以及域名的 *TTL* 值这些数据的统计分析，量化出用于分类的四维特征，使用已知的合法域名与恶意域名作为训练数据集，并对 SVM (support vector machine) [16]分类器进行训练，使用训练好的分类器对测试域名集合进行检测。特征分析与实验表明，算法在能获取到一定量的历史数据条件下，能够有效识别出具有一定生存时间的可疑域名，尤其对域名生存期较长的异常域名具有更高的可靠性。同时，本文所使用的数据较易获取，且处理数据量也较小。

2　域名相关数据分析与分类特征具体表示

大量的域名历史数据表明，恶意域名与合法域名在 whois 信息变更、域名 whois 信息完整度、IP 变更、同 IP 域名数量和 *TTL* 等方面表现出不同的性态，并且这种差异与域名的生存时间密切相关。本节将基于这种差异特性，首先给出相关数据对分类贡献的分析，然后给出分类特征的具体表示。

2.1　域名相关数据分析

域名的 whois 信息更新次数、whois 信息完整度、IP 变更、同 IP 域名数量、域名的 *TTL* 值等可以通过相关的域名信息网站及查询工具获得，这些数据对恶意域名与合法域名而言，会随域名生存时间的增长而表现出某种稳定的性态，这种稳定的性态对异常域名的检测有着不同的贡献，下面将给出这些数据的具体分析。

1) 域名 whois 信息更新次数。whois 是一个用来查询域名是否已经被注册，以及已注册域名的详细信息的数据库，这些信息包括域名的注册组织、域名的注册商以及注册时间、更新时间等。为保证域名的可用性，注册商或域名持有者可以对域名的相关注册信息加以更新。一般而言，合法域名经常被用户查询，为保证域名更好地服务于用户，域名持有者会对域名的 whois 信息及时更新，其 whois 信息更新次数往往较多；而恶意域名仅仅为恶意攻击服务，攻击者往往并不关心域名的 whois 信息，大部分恶意域名持有者并不需要及时更新 whois 信息，其 whois 信息更新次数往往较少。此外，域名的 whois 信息更新次数与域名已生存时间存在着相应的关系，就统计意义而言，生存时间越长恶意域名与合法域名在 whois 信息更新次数上的区别将会越明显。图 1 给出了 2 类域名共 2 000 个样本的 whois 信息更新次数随域名已生存时间变化的样本具体分布情况，分布结果表明，随着生存时间的增长，合法域名 whois 信息更新次数较多，更新速率较快，而大部分恶意域名 whois 信息更新次数几乎不变，二者有较为明显的差异。

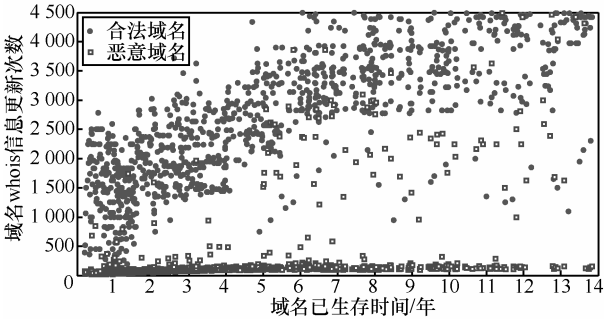

图 1　whois 信息更新次数随域名已生存时间变化的域名分布

2) 域名 whois 信息完整度。域名持有者在注册某个域名时，往往会提供相关的域名信息。合法域名为了能够提高域名知名度，方便用户查询域名信息、了解域名，在注册时一般会尽可能地将信息填写完整。而恶意域名为了掩盖其恶意目的，在注册域名时往往将很多信息随意填写，如注册人、联系方式等相关信息，甚至尽可能地减少填写这些信息。大量统计发现，域名 whois 信息在完整的情况下总条数约为 52。对 2 000 个域名样本进行统计，统计结果如表 1 和表 2 所示。结果表明：92.47%的合法域名其 whois 信息条数在 30 条之上，90.85%的恶意域名其 whois 信息条数在 20 条之下；有 9 条 whois 信息是合法域名和恶意域名共有的，且拥有这 9 条信息的域名数量在合法域名和恶意域名中分别占 99.52%和 98.85%。

表 1　　域名样本 whois 信息完整度统计

whois 信息数/条	合法域名	恶意域名
≤10	1.76%	79.28%
11~20	1.99%	11.57%
21~30	3.78%	4.75%
31~40	5.42%	3.11%
≥41	87.05%	1.29%

表 2　合法、恶意域名共有的 whois 信息项及其含义

whois 信息项	含义
registrar	域名注册机构
registrant name	注册人姓名
registrant city	注册人所在城市
registrant postal code	注册人所在城市邮政编码
registrant country	注册人所在国家
admin name	域名管理人姓名
admin city	域名管理人所在城市
admin country	域名管理人所在国家
name server	域名服务器

显然，合法域名与恶意域名相比，域名 whois 信息完整度方面有着明显的差异，这种差异对二者的分类应有帮助。

3) 域名 IP 变更。域名在注册时会绑定 IP 地址，合法域名会根据解析 IP 是否可以提供正常的服务来决定是否需要更换 IP 地址，从而保证某一 IP 出现问题后使用更换后的 IP 仍然能够提供服务，提高了服务的可用性。一般而言，合法域名的 IP 地址更换，总在一个 IP 数量有限的 IP 池内更换 IP[17]，其更换 IP 地址的个数是有限的，且随着生存时间的增长，IP 变更的个数(使用过的 IP 个数)远小于变更次数。而恶意域名由于遭到检测及封堵，攻击者需要经常更换域名对应的 IP 地址，且每次 IP 变更几乎都将域名映射到一个新的 IP 地址，因此，其 IP 变更的个数与变更次数都会增加。图 2 和图 3 给出了 2 000 个域名样本的 IP 变更个数、变更次数随域名已生存时间变化的样本分布情况，结果表明：随着域名生存时间增长，合法域名与恶意域名的 IP 变更个数与次数都会增加，但合法域名与恶意域名相比，其 IP 变更个数与次数的增长要缓慢。

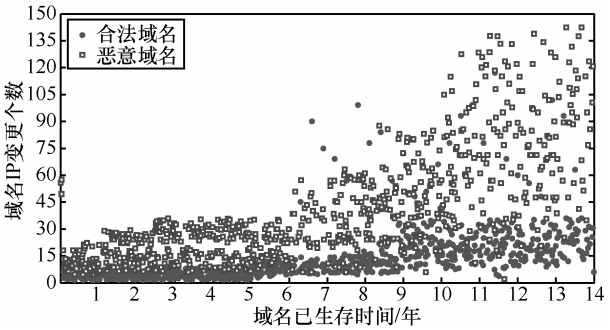

图 2　IP 变更个数随域名已生存时间变化的域名分布

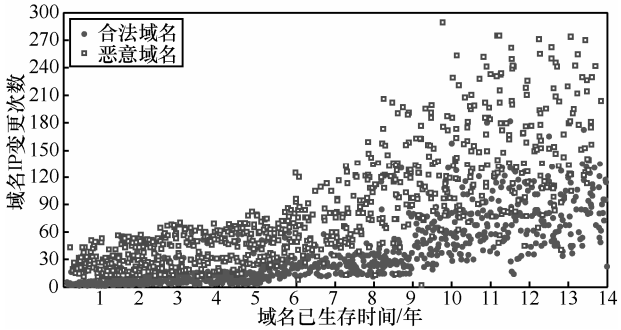

图 3　IP 变更次数随域名已生存时间变化的域名分布

4) 同 IP 地址域名数量。同 IP 地址域名，即与某个确定的域名共享其解析 IP 地址的域名。由

于同一台服务器可能同时为多个域名提供服务，所以会出现多个域名共享同一个 IP 地址的现象。合法域名其目的是为互联网用户提供网络服务，考虑到服务质量问题，同一台服务器一般不会同时为大量的域名提供服务，因此，与合法域名同 IP 地址的域名数量往往较少。而恶意域名其真正目的是用于攻击者发动攻击，并不是为用户提供服务，为了躲避检测和封堵，攻击者往往会将大量域名注册到同一 IP 地址。统计 2 000 个域名样本同 IP 域名数量结果发现，同 IP 域名数量小于 50 个的合法域名占域名样本总数的 79.20%，大于 50 个的合法域名仅占 20.80%，而恶意域名同 IP 域名数量大于 50 个的为 93.37%，小于 50 个的仅为 6.63%。图 4 所示为这 2 000 个域名样本的同 IP 域名数量分布情况，对同 IP 域名数量大于 100 个的统一视为定值 100。由图 4 可知，绝大多数恶意域名样本其同 IP 域名数量大于等于 100 个，几乎所有合法域名样本其同 IP 域名数量小于 100 个。

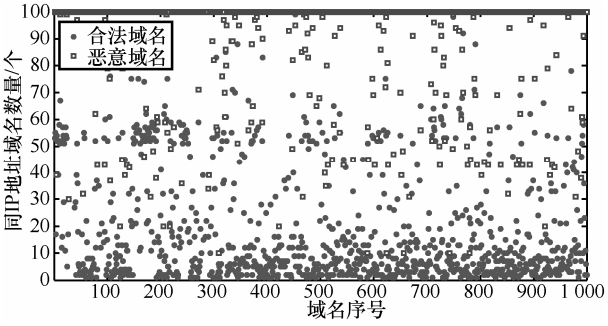

图 4　同 IP 地址域名数量的域名分布

5) 域名 *TTL* 值。域名 *TTL* 是指域名服务器将域名解析记录作为缓存保留的最长时间，以秒为单位。合法域名解析对应 IP 地址往往在一个固定的 IP 池内，为提供较为稳定的服务，*TTL* 值往往设置较大，通常被设置为 1~5 天[14]。恶意域名由于遭到封堵，域名解析对应 IP 地址经常变化，且每次变化往往映射到一个新的 IP 地址，所以其 *TTL* 值往往较小。统计 2 000 个域名样本得出，有 23%以上的合法域名将其 *TTL* 值设为 86 400 s 即 1 天，64%以上的合法域名其 *TTL* 值设置大于 1 000 s，而恶意域名中几乎有 40%其 *TTL* 值设置小于 300 s，小于 1 000 s 的更是占总量的 75%以上。

通过上述分析可以得出，合法域名与恶意域名在 whois 信息更新次数、whois 信息完整度、IP 变更个数及次数、同 IP 地址域名数量、*TTL* 值方面差异明显，且随着域名已生存时间的增长，二者在 whois 信息更新次数、IP 变更个数及次数方面差异变得更加显著，具体信息及其变化如表 3 所示。

表 3　　各类信息变化趋势

信息	合法域名	恶意域名	是否与域名已生存时间相关
whois 信息更新次数	明显增长	变化较小	是
whois 信息完整度	几乎不变	几乎不变	否
IP 变更个数	数量较少	明显增长	是
IP 变更次数	增长较缓	明显增长	是
同 IP 地址域名数量	数量较少	数量较多	否
域名 *TTL* 值	较大	较小	否

2.2　分类特征具体表示

通过对合法域名与恶意域名的上述信息随域名生存时间变化趋势的分析，本文将域名 whois 信息更新次数作为一维特征，其余三维特征表示如下。

2.2.1　域名 IP 变更个数与次数的比值

由 2.1 节的统计分析表明，随着域名生存时间的增长，域名 IP 变更个数与次数都有所增长，但合法域名 IP 变更个数相对固定，恶意域名 IP 变更个数却逐渐增长。从这种变化趋势来看，合法域名 IP 变更个数与变更次数之比会随域名生存时间的增长而不断减小；而恶意域名该比值会随域名生存时间的增长而不断增大，且该比值将逐渐趋于 1。为此，这种变化趋势使将域名 IP 变更个数与次数之比作为一维特征更有利于异常域名的检测。记域名 IP 变更的个数为 *IPCN*，IP 变更的次数为 *IPCT*，则域名 IP 变更个数与次数之比可表示为：$\frac{IPCN}{IPCT}$。

本节仍用 2 000 个合法域名与恶意域名作为样本，图 5 显示了域名 IP 变更个数与次数之比随域名生存时间增长的变化情况，其中，纵轴为特征值。由图 2、图 3 和图 5 可知，随着域名生存时间的增长，使用域名 IP 变更个数与次数之比的分类效果要远好于使用域名 IP 变更个数和次数的分类效果，且确实存在随生存时间的增长分类效果越好的趋势。

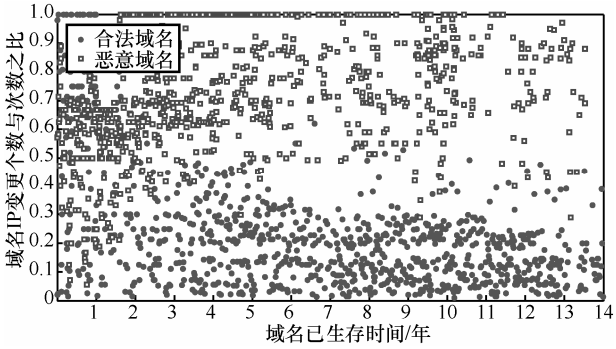

图 5 IP 变更个数与次数比随域名已生存时间变化的域名分布

2.2.2 域名同 IP 域名数量与其 whois 信息完整度总和的比值

由 2.1 节的分析可知，合法域名为了保证网络服务质量，往往不会与大量的域名共享同一个 IP；而恶意域名会仅仅为达到某种单一的目的而出现大量域名共享 IP 地址的情况，如僵尸网络中会将大量的域名关联到同一个 IP 地址。此外，大部分合法域名的 whois 信息都较为完整，而大量的恶意域名其 whois 信息完整度都较低。从域名的同 IP 域名数量与相应的同 IP 域名 whois 信息完整度之和的比值来看，如果同 IP 的域名均为合法域名，该比值较小，而如果同 IP 的域名均为恶意域名，则该比值往往较大。因此，本文将域名同 IP 地址域名数量与其相应 whois 信息完整度总和的比值作为一维特征。下面给出该一维特征的一种具体表示方法并验证其分类效果。

记域名 M 的第 $n(1 \leqslant n \leqslant N)$条 whois 信息为 $i(n)$，其值用 $v(i(n))$表示。如果 $i(n)$在域名 M 的 whois 信息中存在，则置 $v(i(n))$为 1；否则置 $v(i(n))$为 0。

由 2.1 节的分析可知，域名 whois 信息在完整的情况下总条数约为 52 条，故取 N=52；由表 2 可知，在大量的合法域名与恶意域名 whois 信息中，有 9 条信息项是共有的，显然共有项对分类的贡献要小于非共有项，故可对域名 whois 信息的每一项加以赋权。记共有的 9 条 whois 信息项构成的集合为 I，域名的第 n 条 whois 信息项 $i(n)$的权值为 $W(i(n))$，则

$$W(i(n))=\begin{cases}a, i(n)\in I\\ b, i(n)\notin I\end{cases} \tag{1}$$

其中，$a<b$。

设域名的 whois 信息完整度为 WCR，则对于域名 M，其 whois 信息完整度可表示为

$$WCR(M)=\sum_{1}^{52}\left(W(i(n))v(i(n))\right) \tag{2}$$

设与域名 M 同 IP 的域名为 $M_1, M_2, M_3, \cdots, M_K$，则域名 M 同 IP 域名数量与其 whois 信息完整度总和的比值可表示为

$$\frac{K}{\sum_{1}^{K}(WCR(M_K))} \tag{3}$$

仍采用上述 2 000 个域名为样本，图 6 显示了同 IP 域名数量与其 whois 信息完整度总和的比值的域名分布(实验中有大量的比值大于 0.5，本文将大于 0.5 的比值统一视为定值 0.5)。验证过程中，置式(1)的参数 a=0.1，b=0.9。由图 4 和图 6 可知，基于式(3)的这维特征能够更好地区分合法域名与恶意域名。但由于约 6.7%的合法域名与其同 IP 的域名中存在部分恶意域名，约 3.9%的恶意域名与其共享 IP 的域名中存在大量合法域名，所以导致合法域名与恶意域名基于该特征的特征值出现部分交叉。

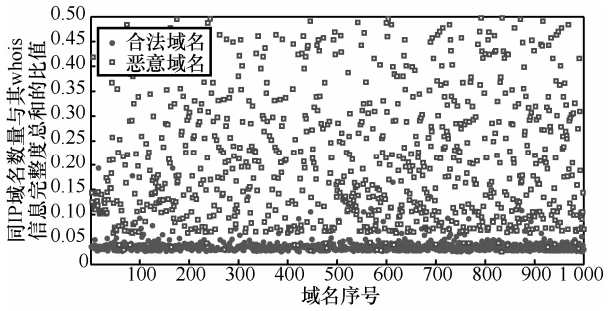

图 6 同 IP 域名数量与其 whois 信息完整度总和的比值特征的域名分布

2.2.3 基于域名 IP 变更速率与 *TTL* 的二元函数值

由 2.1 节关于域名 IP 变更个数和变更次数的分析可知，合法域名 IP 总是在一个数量有限的 IP 池内变更，而恶意域名的 IP 变更数量是逐渐增加的，因而就统计而言，合法域名 IP 变更速率会随域名生存时间的增长而减小，而恶意域名对应的 IP 地址经常变化，且每次变化几乎都映射到新的 IP 地址，因而其 IP 变更速率并不具有合法域名那样明显的特性。此外，从域名提供服务的角度而言，合法域名为提供稳定的公共服务，便于用户访问，其解析记录在 DNS 服务器中缓存的时间比较固定，即域名的 *TTL* 值设置固定(统计上该值设置一般较大)；而恶意域名只是为攻击者达到某种恶意的目的，为避免相关的检测和封堵，其域名所对应的 IP 地址需要经常变更，这就使域名的解析记录在 DNS 服务器中的缓存时间较短，即 *TTL* 值设置往往较小。就合

法域名与恶意域名二者的特性而言，域名 IP 变更的速率与 *TTL* 值的设置会呈现一定程度的负相关性。为此，本节将用基于域名 IP 变更速率与 *TTL* 值的二元相关函数的函数值作为特征来刻画这种负相关性。

设域名 *M* 已生存时间为 *T*，则 *M* 已生存时间内的IP变更速率*IPCR*可表示为 $\frac{IPCN}{T}$。其中，*IPCN* 在 2.2.1 节中已提到。

设域名 *M* 的 *TTL* 值为 *ttl*，域名 *M* 的 IP 变更速率与 *TTL* 值的二元相关函数记为 *f*(*IPCR*，*ttl*)，根据域名 IP 变更速率与 *TTL* 值的关系，本文将这种二元关系表示为

$$f(IPCR,ttl)=\frac{IPCN}{T}\cdot\frac{1}{ttl} \tag{4}$$

随着生成时间的增长，合法域名该函数值将会减小，而恶意域名该函数值将增大，通过基于域名IP变更速率与*TTL*的二元函数能够较好地区分合法域名与恶意域名。部分大型合法网站为了提高网站可用性及服务质量，使用内容分发网络（CDN，content delivery network）或轮转 DNS（rrDNS，round robin DNS）技术，其域名 *TTL* 值设置可能较小，但生存时间内 IP 变更速率较小，与恶意域名的 IP 变更速率依然有较大差异，因此，基于域名 IP 变更速率与 *TTL* 的二元函数值仍然能够将这类合法域名与恶意域名区分开。

仍采用上述 2 000 个域名为样本，图 7 显示了基于域名 IP 变更速率与 *TTL* 的二元函数值为特征的域名分布(部分恶意域名的特征值远大于 0.1，为便于观察，实验中将大于 0.1 的特征值统一视为 0.1)。从图 7 可知，通过域名 IP 变更的速率与 *TTL* 值呈现出的负相关性，可以良好地刻画合法域名与恶意域名之间的差异。

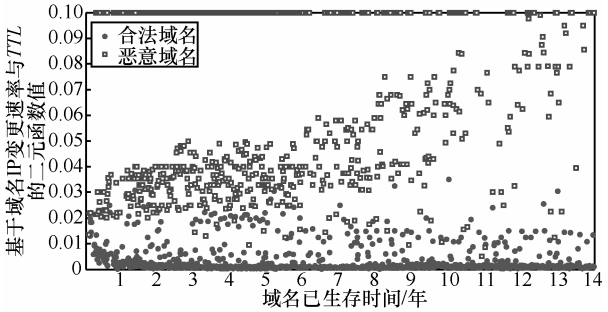

图 7　基于域名 IP 变更速率与 *TTL* 的二元函数值随域名已生存时间变化的域名分布

3　检测算法

本文提出的异常域名检测算法的主要思想是通过域名信息网站获取域名相关的各类历史数据信息，从中收集合法域名与恶意域名在域名 whois 信息更新次数、whois 信息完整度、域名解析对应的 IP 变更、同 IP 地址域名数量、域名的 *TTL* 值方面的统计差异，构建域名 whois 信息更新次数、域名 IP 变更个数与次数之比、同 IP 域名数量与其 whois 信息完整度总和的比值、基于域名 IP 变更速率与 *TTL* 的二元函数值这些四维分类特征，如表 4 所示。

表 4　特征集合

特征序号	特征名称
F_1	域名 whois 信息更新次数
F_2	域名 IP 变更个数与次数之比
F_3	域名同 IP 域名数量与其 whois 信息完整度总和的比值
F_4	基于域名 IP 变更速率与 *TTL* 的二元函数值

具体检测算法如下。

1) 通过筛选后的合法域名与恶意域名构造域名样本集合 *S*。

2) 对$\forall M\in S$，获取域名 *M* 的历史数据 D_{history}，对 D_{history} 进行分析，提取特征构造域名 *M* 的特征向量 $\boldsymbol{F}_M(F_1,F_2,F_3,F_4)$。

3) 设 $\boldsymbol{F}_M\in F$，*F* 为所有域名样本的特征向量构成的集合。将特征向量集 *F* 分为训练集 F_{train} 与测试集 F_{test}，并使用 F_{train} 训练分类器，得出已训练好的分类器模型(*Model*)。

4) 使用测试集 F_{test} 对 *Model* 进行测试，得出检测结果。

4　实验设计及结果

为了验证本文所设计的恶意域名检测算法对生存时间较长的恶意域名良好的检测效果，以及与现有一些检测方法的比较，本文设计 2 组实验，其中，4.1 节为验证性实验，4.2 节为对比性实验。

4.1　本文检测算法检测效果验证

4.1.1　域名样本来源与构造

该组实验中，合法域名通过网站 Domains5[注1]获取。该网站提供有 Alexa 排名的域名，并按域名排

注1：Domains5. http://www.domains5.cn/。

序列出，从该网站上共获得合法域名样本 4 773 个。

恶意域名主要通过 Malwr [注2]网站获取。为了进一步确定网站上提供的恶意域名的性质，通过 McAfee [注3]对其进行进一步的筛选。McAfee 网站提供对域名性质的判别，能够识别出域名是否与恶意行为相关，并且较为精确。根据 McAfee 的验证结果筛选后，获得恶意域名样本 2 318 个。

4.1.2 特征获取

对于获取到的域名样本，通过 Robtex [注4]、Domaintools [注5]网站及 whois、nslookup 命令获取用于生成各维特征的域名历史数据信息。通过对 Domaintools 网站数据统计获得域名的创建时间、域名 whois 信息更新次数、域名 IP 变更个数及 IP 变更次数；通过 Robtex 网站数据统计获取与某个域名样本具有相同 IP 地址的域名及其数量，并通过 whois 命令获取同 IP 域名的 whois 信息；通过 nslookup 命令查询并获取每个域名样本的 *TTL* 值。在获取到各类数据信息后，将信息按照本文所述方式进行组合进而得到各维特征。

4.1.3 实验及结果分析

本文检测算法中所使用的特征与域名的已生存时间有较大的相关性，域名已生存时间越长，分类效果应越好。为了验证本文算法的该特点，本节将域名样本集合按已生存时间的长短进行划分，构建相应的训练集和测试集，分类器运用 SVM 分类器。

根据域名已生存时间对样本集合进行划分。其中，S_1 表示所有已生存时间为 0~3 年的域名样本集；S_2 表示所有已生存时间为 3~6 年的域名样本集；S_3 表示所有已生存时间为 6~9 年的域名样本集；S_4 表示所有已生存时间为 9~12 年的域名样本集；S_5 表示所有已生存时间为 12 年以上的域名样本集。$S_r\left(\frac{3}{4}\right)$ 表示 S_r 中 $\frac{3}{4}$ 的域名构成的集合；$S_r\left(\frac{1}{4}\right)$ 表示 S_r 中另外 $\frac{1}{4}$ 的域名构成的集合（r=1, 2, 3, 4）。实验中用第 x 年至第 y 年的样本构成训练集（如 x=0，y=3），分别用第 x 年后的样本与第 y 年后的样本作为测试集，依据上述样本划分，实验可分为 4 个小组。具体实验数据如表 5 所示。

注2：Malwr. https://malwr.com/。

注3：McAfee. http://www.siteadvisor.com/sites/。

注4：Robtex. https://www.robtex.com/。

注5：Domaintools. http://www.domaintools.com/。

由于域名样本的特征向量中分量值出现过大或过小的现象，如 whois 信息更新次数相对于其他 3 个分量值较大，而这些奇异分量可能引起训练时间增加，并可能引起网络无法收敛，因此需要对训练数据与测试数据进行归一化处理，本节中将训练集与测试集的特征向量分量均归一化到[0,1]。实验结果如表 6 所示。其中，分类正确率、漏报率、虚警率计算如下，分类正确率：$\frac{TT+FF}{X+Y}$，漏报率：$\frac{TF}{Y}$，虚警率：$\frac{FT}{X}$，其中，X 表示测试集中合法域名数量，Y 表示测试集中恶意域名数量，TT 表示被正确分类的合法域名数量，FF 表示被正确分类的恶意域名数量，FT 表示被错误分类的合法域名数量，TF 表示被错误分类的恶意域名数量。

表 5 训练集、测试集构造

实验序号	训练集构成及样本数量			测试集构成及样本数量		
	训练集构成	合法样本/个	恶意样本/个	测试集构成	合法样本/个	恶意样本/个
1	$S_1\left(\frac{3}{4}\right)$	562	447	$S_1\left(\frac{1}{4}\right)$、$S_2$、$S_3$、$S_4$、$S_5$	4 211	1 871
2				S_2、S_3、S_4、S_5	4 024	1 723
3	$S_2\left(\frac{3}{4}\right)$	745	522	$S_2\left(\frac{1}{4}\right)$、$S_3$、$S_4$、$S_5$	3 279	1 201
4				S_3、S_4、S_5	3 031	1 028
5	$S_3\left(\frac{3}{4}\right)$	809	393	$S_3\left(\frac{1}{4}\right)$、$S_4$、$S_5$	2 222	635
6				S_4、S_5	1 953	505
7	$S_4\left(\frac{3}{4}\right)$	1 074	276	$S_4\left(\frac{1}{4}\right)$、$S_5$	879	229
8				S_5	522	138

表 6 分类结果

实验序号	正确率	漏报数/个	漏报率	虚警数/个	虚警率
1	89.9%	297	15.9%	318	7.6%
2	91.6%	232	13.5%	251	6.2%
3	94.3%	119	9.9%	137	4.2%
4	95.8%	82	8.0%	89	2.9%
5	97.8%	37	5.8%	26	1.2%
6	98.9%	15	3.0%	13	0.7%
7	99.3%	4	1.7%	4	0.5%
8	99.5%	1	0.7%	2	0.4%

由表 6 可知，实验 2、实验 4、实验 6、实验 8 中的分类效果要好于实验 1、实验 3、实验 5、实验

7 中的分类效果；从实验 1~实验 8，分类正确率不断提高，漏报率与虚警率不断下降。这一结果验证了本文算法的理论分析与算法特点。此外，分析上述 8 个实验测试集的数据可发现如下 4 类异常域名样本。

A 类：域名 whois 信息更新次数与已生存时间相关性不强，不足以区分恶意域名与合法域名，导致不同程度的漏报与虚警。

B 类：恶意域名 IP 变更个数与次数之比随域名已生存时间增加、合法域名 IP 变更个数与次数之比随域名已生存时间减少的规律不明显。

C 类：同 IP 的域名中存在合法域名与恶意域名共享 IP 的现象，导致部分域名样本出现异常。

D 类：随着域名已生存时间的增长，合法域名应呈现的域名 IP 变更速率与 *TTL* 值负相关性不明显，恶意域名并不存在这种负相关性。

表 7 给出了使用多类测试集对基于不同训练集所构分类器进行测试时，分类结果中出现的各类异常域名样本及其数量统计。

表 7　测试集恶意、合法样本中各类异常域名样本数量统计

实验序号	恶意样本中各类异常域名样本数量/个				合法样本中各类异常域名样本数量/个			
	A 类	B 类	C 类	D 类	A 类	B 类	C 类	D 类
1	47	120	79	51	57	138	65	58
2	31	107	53	41	42	88	70	51
3	25	41	29	24	25	44	41	27
4	19	12	31	20	21	25	25	18
5	8	9	12	8	6	5	8	7
6	4	3	6	2	4	1	5	3
7	1	0	2	1	0	0	3	1
8	0	0	0	1	0	0	1	1

就表 7 的统计结果而言，异常域名样本尽管与特征的表示相关，但就统计趋势而言，随着域名已生存时间的增长，域名样本集合中引起漏报及虚警的各类异常域名样本数量在不断下降，因而本检测算法对于生存时间较长的恶意域名检测更为可靠。

4.2　同类相关工作对比

本节主要介绍本文检测算法与文献[11]、文献[14]检测方法的对比。样本取文献[11]、文献[14]中的同源样本，合法域名样本为 Alexa 排名靠前的域名，恶意域名样本均取自 malwaredomains [注6]网站。取 2 000 个合法域名样本，从 malwaredomains 网站取 2 000 个恶意域名样本，通过这 2 类共 4 000 个域名样本构造测试集，并将测试集按生存时间如 4.1.3 节所述方式进行划分，得到域名样本集 S_1'~S_5'。使用划分后的测试集对 4.1.3 节中实验 1~实验 8 对应训练好的分类器进行测试，得出分类结果。其中，由于文献[11]、文献[14]中的检测方法未对域名样本集合按生存时间划分，而是对测试样本集合总体进行的分类，因此，在对比分类结果时，将本文检测算法对不同生存期的域名样本集的分类结果分别与这 2 种方法对测试样本集合总体的分类结果相比较。表 8 给出了实验 9~实验 16 中具体的测试集合构造与相应的 3 种检测方法的分类正确率。

由表 8 可以看出，本文检测算法对其他来源的恶意域名的检测效果也较好，并且依然对生存时间较长的域名拥有更高的分类正确率。由实验 9~实验 12 的分类结果可以得出，域名生存时间短时，本文检测算法的检测效果不如文献[11,14]中检测方法所获得的检测效果；由实验 13~实验 16 的分类结果可以得出，随着域名生存时间的增长，本文检测算法的检测效果与文献[11]中检测方法的检测效果相当；由实验 15 和实验 16 的分类结果可以得出，当域名生存时间更长时，本文检测算法的检测效果与文献[14]中检测方法的检测效果相当。表 6 和表 8 的分类结果说明相比于其他检测方法，本文检测算法对生存时间较长的恶意域名的检测能力较强，这也进一步验证了本文检测算法针对长期生存的恶意域名检测的优势及可靠性。

表 8　3 种检测方法对相同来源的域名样本分类结果对比

实验序号	测试集构成及样本数量			分类正确率		
	测试集构成	合法样本/个	恶意样本/个	本文方法	文献[11]方法	文献[14]方法
9	S_1'、S_2'、S_3'、S_4'、S_5'	2 000	2 000	89.8%	97.8%	99.5%
10	S_2'、S_3'、S_4'、S_5'	1 687	1 486	92.5%	97.8%	99.5%
11	S_2'、S_3'、S_4'、S_5'	1 687	1 486	95.7%	97.8%	99.5%
12	S_3'、S_4'、S_5'	1 271	885	96.9%	97.8%	99.5%
13	S_3'、S_4'、S_5'	1 271	885	98.1%	97.8%	99.5%
14	S_4'、S_5'	820	433	99.0%	97.8%	99.5%
15	S_4'、S_5'	820	433	99.6%	97.8%	99.5%
16	S_5'	220	116	99.8%	97.8%	99.5%

5　结束语

本文给出了一种对异常域名进行检测的算法，该检测算法通过对各类域名信息网站提供的域名

注6：http://mirror2.malwaredomains.com/files/justdomains。

网络历史数据的分析，生成特征，利用机器学习进而对可疑域名进行检测。本文算法的主要贡献是在获取到域名历史数据的情况下，对长期活跃在网络中的恶意域名具有较为可靠的检测准确率，并且随着域名生存时间的增长，其检测效果尤为明显。算法对生存时间较长的可疑恶意域名的发现能力，为发现那些尚未检测到且长期存在的可疑恶意域名提供了一种新的方法，这是本文工作的一个特色，也是与其他工作的一个重要不同。此外，本文工作可离线进行，不需要通过对顶级域名服务器或者权威域名服务器、本地域名服务器的监测获取 DNS 流量等数据，数据获取容易、计算量小。

参考文献：

[1] ROSSOW C, DIETRICH C, BOS H. Detection of intrusions and malware, and vulnerability assessment[M]. Berlin: Springer, 2013.

[2] MAHMOUD M, NIR M, MATRAWY A. A survey on botnet architectures, detection and defences[J]. International Journal of Network Security, 2015, 17(3): 272-289.

[3] PU Y, CHEN X, CUI X, et al. Data stolen trojan detection based on network behaviors[J]. Procedia Computer Science, 2013, 17: 828-835.

[4] NIRMAL K, JANET B, KUMAR R. Phishing-the threat that still exists[C]//International Conference on Computing and Communications Technologies(ICCCT). IEEE, 2015: 139-143.

[5] CHEN C M, CHENG S T, CHOU J H. Detection of fast-flux domains[J]. Journal of Advances in Computer Networks, 2013, 1(2): 148-152.

[6] VANIA J, MENIYA A, JETHVA H B. A review on botnet and detection technique[J]. International Journal of Computer Trends and Technology, 2013, 4(1): 23-29.

[7] KHATTAK S, RAMAY N R, KHAN K R, et al. A taxonomy of botnet behavior, detection and defense[J]. Communications Surveys & Tutorials, IEEE, 2014, 16(2): 898-924.

[8] GARCÍA S, UHLÍŘ V, REHAK M. Identifying and modeling botnet C&C behaviors[C]//The 1st International Workshop on Agents and CyberSecurity. ACM, 2014.

[9] YADAV S, REDDY A K K, REDDY A L, et al. Detecting algorithmically generated malicious domain names[C]//The 10th ACM SIGCOMM Conference on Internet Measurement. Melbourne, Australia, 2010: 48-61.

[10] FELEGYHAZI M, KREIBICH C, PAXSON V. On the potential of proactive domain blacklisting[C]//The 3rd USENIX Conference on Large-Scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More. San Jose, CA, USA, 2010.

[11] 刘爱江, 黄长慧, 胡光俊. 基于改进神经网络算法的木马控制域名检测方法[J]. 电信科学, 2014, 30(7): 39-42.
LIU A J, HUANG C H, HU G J, Detection method of trojan's control domain based on improved neural network algorithm[J]. Telecommunications Science, 2014, 30(7): 39-42.

[12] ANTONAKAKIS M, PERDISCI R, DAGON D, et al. Building a dynamic reputation system for DNS[C]//USENIX Security Symposium. Washington, DC, USA, 2010: 273-290.

[13] ANTONAKAKIS M, PERDISCI R, LEE W, et al. Detecting malware domains at the upper DNS hierarchy[C]//USENIX Security Symposium. San Francisco, CA, USA, 2011: 23-46.

[14] BILGE L, SEN S, BALZAROTTI D, et al. Exposure: a passive DNS analysis service to detect and report malicious domains[J]. ACM Transactions on Information and System Security (TISSEC), 2014, 16(4): 14-41.

[15] 周勇林, 由林麟, 张永铮. 基于命名及解析行为特征的异常域名检测方法[J]. 计算机工程与应用, 2011, 47(20): 50-52.
ZHOU Y L, YOU L L, ZHANG Y Z. Anomaly domain name detection method based on characteristics of name and resolution behavior[J]. Computer Engineering and Applications, 2011, 47(20): 50-52.

[16] LENG Y, XU X, QI G. Combining active learning and semi-supervised learning to construct SVM classifier[J]. Knowledge-Based Systems, 2013, 44: 121-131.

[17] YU B, SMITH L, THREEFOOT M. Machine learning and data mining in pattern recognition[M]. Berlin: Springer, 2014.

作者简介：

袁福祥（1991-），男，山东济宁人，解放军信息工程大学硕士生，主要研究方向为网络信息处理。

刘粉林（1964-），男，江苏溧阳人，解放军信息工程大学教授、博士生导师，主要研究方向为网络信息安全、信息隐藏与检测。

芦斌（1982-），男，山西灵石人，解放军信息工程大学讲师，主要研究方向为数字水印、软件工程。

巩道福（1984-），男，山东淄博人，解放军信息工程大学讲师，主要研究方向为数字水印、网络信息安全。