CPDGA：基于一致性传播的DGA域名主动检测算法

刘双双[1]，王志[1]，董伊萌[1]，李万鹏[2]
（1.南开大学密码与网络空间安全学院，天津 300350；2.利物浦大学计算机学院，利物浦 L693BX）

摘　要：攻击者通过域名生成算法（DGA）动态注册域名以支持恶意软件活动，恶意域名不断演化导致概念漂移现象，使得现有依赖可持续性学习模型的检测技术时效性不足。针对这一威胁，结合一致性预测与一致性聚类方法，提出了一种基于一致性传播的DGA域名主动检测算法（CPDGA）。通过对2019—2023年恶意与良性域名数据集进行实验，证明CPDGA能够有效缓解概念漂移对机器学习检测模型性能的影响，并使检测准确率提升20.4%。此外，CPDGA在检测13种最新对抗模型生成域名时取得了96.42%的准确率，展现了强大的鲁棒性与适应性。

关键词：域名生成算法；概念漂移；一致性预测；一致性聚类；对抗模型
中图分类号：TP309.5
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025106

CPDGA: foresee future DGA using proactive conformal propagation

LIU Shuangshuang[1], WANG Zhi[1], DONG Yimeng[1], LI Wanpeng[2]
1. College of Cryptology and Cyber Science, Nankai University, Tianjin 300350, China
2. College of Computer Science, University of Liverpool, Liverpool L693BX, UK

Abstract: Attackers dynamically register domain names through the domain generation algorithm (DGA) to support malware activities. The continuous evolution of malicious domain names leads to the phenomenon of concept drift, rendering the existing detection techniques based on continual learning models less effective over time. To address this threat, by combining conformal prediction and conformal clustering, a foresee future DGA using proactive conformal propagation (CPDGA) was proposed. Experiments were conducted using datasets of malicious and benign domain names from 2019 to 2023. CPDGA was applied to mitigate the effect of concept drift. As a result, the impact of concept drift was effectively reduced. The detection accuracy was improved by 20.4%. Additionally, CPDGA achieves an accuracy rate of 96.42% in detecting the domain names generated by 13 latest adversarial models, showing its strong robustness and adaptability.

Keywords: domain generation algorithm, concept drift, conformal prediction, conformal clustering, adversarial model

0　引言

僵尸网络是由大量感染恶意软件的计算机组成的分布式网络，这些计算机受一个或多个远程控制者的指挥，执行各种恶意活动[1-4]。这些活动包括但不限于分布式拒绝服务（DDoS, distributed denial of service）攻击、垃圾邮件发送、数据盗窃和恶意软件分发。其影响涉及个人隐私、社会稳定以及全球网络生态系统。被僵尸网络感染的计算机可

收稿日期：2025–04–12；修回日期：2025–05–29
通信作者：王志，zwang@nankai.edu.cn
基金项目：CCF–绿盟科技“鲲鹏”科研基金资助项目（No.CCF-NSFOCUS 2024016）
Foundation Item：The CCF-NSFOCUS “kunpeng” Research Fund (No.CCF-NSFOCUS 2024016)

能被远程控制，导致用户的私人信息、浏览记录和敏感数据被盗或泄露，以及身份盗用和金融欺诈，给用户造成信用和经济损失。此外，攻击者还可能利用僵尸网络对关键基础设施和公共服务发起DDoS攻击，导致银行、医疗和交通等行业的服务中断，威胁社会稳定和公共安全。

针对僵尸网络多种威胁的有效对策至关重要。检测和应对僵尸网络的方法可总结为以下几类。

1)流量分析与行为分析：通过监控网络流量，分析主机和网络设备的异常行为模式，识别异常的请求模式和通信频率。

2)网络之间互联的协议（IP, Internet protocol）地址与恶意域名检查：通过检查网络流量中的IP地址和恶意域名，利用域名和IP黑名单服务阻断僵尸网络命令与控制服务器（C2, command and control server）的通信[5-8]。

3)主机行为监控技术：部署端点检测与响应软件，实时监控进程活动、文件操作和系统资源使用情况，实现对感染的及时检测与隔离。

4)针对域名生成算法（DGA, domain generation algorithm）的检测技术：通过分析和识别伪随机生成的域名来推断并阻断与僵尸网络C2通信相关的有效域名[9-14]。

DGA是僵尸网络控制者采用的一种策略，通过伪随机算法生成大量域名作为通信点。检测DGA恶意域名的技术包括多种方法。首先，通过分析域名的结构和模式，可以识别DGA生成的域名与正常域名之间的差异[15-19]。其次，利用统计分析和机器学习技术，如N-gram分析、频率分析，以及支持向量机（SVM, support vector machine）和随机森林（RF, random forest）等分类器，自动识别和分类DGA域名。此外，安全团队通过实时监控网络流量和域名系统（DNS, domain name system）查询，及时识别新的DGA活动，并借助威胁情报共享平台获取最新的威胁信息，从而调整和优化防御策略。

然而，随着DGA域名的概念漂移和对抗样本（AE, adversarial example）的出现，DGA的检测技术面临诸多挑战[20-26]。

概念漂移是指数据分布或特征随时间发生变化的现象。概念漂移导致模型在应用于新数据检测时性能下降，主要有2个方面的原因：1) 概念漂移可能导致训练过程中使用的模型无法有效泛化到新的数据分布，尤其是新出现的DGA变种可能无法被现有模型准确识别，从而降低检测准确率；2) 随着时间推移，用于区分正常域名与DGA域名的特征可能失效，某些DGA会调整其算法以规避已知的检测特征，从而使基于固定特征集的检测算法失效。

概念漂移可能导致模型更新滞后于实际情况，无法及时识别和阻止最新的恶意域名生成请求。应对概念漂移意味着训练数据集需要被持续更新和维护，以反映新出现的DGA变种及行为模式，DGA检测需要以实时或近实时的方式进行。然而，获取和标注大规模的实时数据集是一项艰巨的任务。

此外，基于深度学习的分类器对AE具有脆弱性。对抗样本是通过对输入添加微小扰动生成的，这些扰动对人类几乎不可察觉，却能够显著影响模型的判断，导致其输出错误分类结果。攻击者可能故意操纵域名的特征来规避机器学习模型的检测[27-30]。因此，概念漂移和对抗样本对基于DGA的恶意域名检测算法带来了重要挑战。使用动态自适应的模型和算法能够持续地监测和预测恶意域名的出现和演化，并帮助确定新出现的DGA域名是否与已知的恶意模式相关，从而提升检测效率和准确性。

为了解决上述挑战，本文提出了基于一致性传播的DGA域名主动检测算法（CPDGA），该算法通过对恶意域名进行后验分析[31-34]，能够有效提升已训练模型的准确性，提升幅度可达20%。本文的主要贡献如下。

1)本文提出了CPDGA。CPDGA包含训练阶段、数据增强阶段和应用阶段3个阶段。训练阶段利用一致性预测（CP, conformal prediction）方法，构建用于度量新样本一致性的p值网格；数据增强阶段基于一致性聚类（CC, conformal clustering）方法，生成与已知恶意域名分布一致的扩展样本集；应用阶段将CPDGA用于缓解概念漂移和对抗攻击下的DGA检测。该算法能够预测未来DGA域名的潜在分布范围，从而显著提高检测模型的鲁棒性和准确率。本文已经公开CPDGA的源代码和实验数据。

2)本文引入一致性预测（CP）与一次性聚类（CC）机制至DGA检测任务中，首次在域名数据增强过程中实现了对样本可靠性的一致性评估与聚

合，确保生成样本的统计特性与原始恶意样本保持一致。

3)本文构建了包含2019—2023年的恶意和正常域名的大规模数据集，同时还构建了包含13种对抗模型生成域名的数据集，以全面评估模型的泛化能力和对抗鲁棒性。

4)本文评估了CPDGA在减轻概念漂移影响方面的有效性。在仅使用2019年数据进行训练的情况下，CPDGA在2020—2023年的DGA域名检测任务中准确率最高提升了20.4%，显著优于现有主流算法。

5)本文还验证了CPDGA对13种对抗样本的检测能力。实验结果表明，CPDGA对这些由对抗模型生成的域名的检测准确率超过90%，最高可达96.42%，有效增强了模型在对抗攻击下的安全性。

1 相关工作

在恶意域名检测相关的研究中，已经提出了许多方法，并取得了显著的成果。基于特征工程和机器学习技术的传统方法是常见的研究方向。大多数机器学习方法通过手动创建特征来实现恶意域名的监控。例如，Schüppen等[17]提出了基于特征的域名自动分类（FANCI）系统，通过监控DNS流量中的不存在域检测DGA恶意软件感染。FANCI系统采用机器学习的分类方法，将不存在域分类为与DGA相关或正常类别，其分类基于从不存在域中提取的21个特征。基于FANCI系统，Zhao等[35]优化了特征提取，提出了基于语言学特征的域名分类系统（DOLPHIN）。DOLPHIN利用语言学原理，结合上下文检测和发音与拼写的对应关系，设计了一种基于语音学的分类方案。

传统的机器学习方法通常需要手动选择特征，而恶意域名的动态性和多样性推动了深度学习方法的研究。Yu等[36]提出了一种基于真实流量数据和深度学习技术的DGA，克服了传统方法中训练数据的限制，提供了更好的性能。Tran等[37]提出了LSTM.MI的新算法，结合了二分类和多分类模型，将原始的长短期记忆（LSTM, long short-term memory）网络模型调整为成本敏感型模型，在反向传播的学习过程中引入成本项，以考虑类别间识别的重要性。然而，恶意域名通过不断变化逃避检测，传统的机器和深度学习方法难以跟上其变化速度，导致检测准确率下降[38-39]。

此外，随着一系列对抗性域名生成算法的出现，传统模型的检测性能受到了一定的限制。Anderson等[40]利用生成对抗网络（GAN, generative adversarial network）构建了一种基于深度学习的恶意域名生成算法。在一系列对抗性迭代中，生成器学会生成越来越难以检测的域名。因此，检测模型需要不断调整其参数以适应对抗性生成的域名。Yun等[27]提出了一种基于神经语言建模和Wasserstein生成对抗网络（WGAN）的新型DGA。它利用神经语言学建模调节音节和缩略词，生成与真实域名高度相似的对抗性域名，具有很强的反检测能力。Peck等[41]提出了一种基于字符的简单域名生成算法（CharBot），通过破坏Alexa顶级域名生成对抗性域名。Hu等[42]提出了一种基于双向长短期记忆（BiLSTM, bidirectional long short-term memory）网络的对抗性DGA。该算法建模正常域名中字符的位置及其前后字符序列之间的语义关系，生成对抗性域名，绕过动态DGA分类器的检测。上述算法的实验结果表明，传统模型在检测对抗性域名时存在局限性[43-44]，因此开发有效的对抗性域名检测算法尤为重要。

针对传统模型在恶意域名检测中的局限性，本文提出了一种结合一致性预测[45]和一致性聚类[46]方法的新算法。一致性预测方法能够提供可靠的置信度估计，评估检测结果的可信度，并且在处理未知和不确定数据时具有适应性。一致性聚类方法能够基于域名的特征和行为模式，将其分类到不同的簇中，从而识别具有相似属性的恶意域名。Barbero等[47]提出了基于一致性预测策略的框架（TRANSCEND），能够有效识别并拒绝漂移样本。Park等[48]提出了一种基于一致性预测和相似性度量的干净标签后门攻击，并结合子空间压缩策略提高了模型鲁棒性。

一致性预测和一致性聚类方法能够提供更丰富的数据表示，有助于揭示域名之间的潜在几何结构和相互关系。因此，将这些方法应用于僵尸网络检测具有重要的意义。

2 先验知识

2.1 一致性预测

CP是一种非参数预测方法，无须依赖分布假

设。具体而言，CP通过最小化先验假设，基于非参数方法生成统计上有效的预测集，即使在样本有限的情况下也能保持可靠性。CP依赖于先前积累的数据和经验，从中提取关于未来预测的信息，并在假设样本独立同分布的前提下生成预测集，这一特性为误差率提供了可靠的保证。通过这种方式，CP能够为每个新预测提供准确的置信度水平。

这里介绍CP的基本思想。假设有$n+1$个可交换的观测值，表示为$z_1,z_2,z_3,\cdots,z_{n+1}$。对于给定的新观测值$z$，定义非一致性度量$R_i$和$R_{n+1}$，如式(1)和式(2)所示，用于度量新观测值与过去观测值之间的非一致性，以计算p值，所有样本都属于一个对象空间Z。

$$R_i=A\left(\left\{z_1,z_2,\cdots,z_{i-1},z_{i+1},z_n,z\right\}\right)=A\left(\left\{z_1,\cdots,z_n,z\backslash z_i,z_i\right\}\right),\ \forall i\in 1,\cdots,n \tag{1}$$

$$R_{n+1}:=A(\{z_1,\cdots,z_n,z\}) \tag{2}$$

其中，:=为定义性赋值符号，A为非一致性度量函数。典型的度量方法包括距离度量、核函数、相似性度量等。p值衡量新观测值与过去观测值之间的不一致程度。较小的p值表示新观测值在经验中相对不常见，表明具有较高的非一致性。相反，较大的p值表示新观测值在经验中较为常见，表明一致性较高。可以通过式(3)和式(4)决定是否将新观测值纳入预测集。

$$\alpha_i=A\left(c\backslash z_i,z_i\right),\forall i\in C \tag{3}$$

$$p_z=\frac{\left\{i:\alpha_i>\alpha_n\right\}+\{i:\alpha_i=\alpha_n\}\tau}{n} \tag{4}$$

如果新观察值z的p值大于设定的阈值，则将z包含在预测集内。本文使用的CP方法的详细步骤在算法1中描述。

算法1　计算未知域名样本的p值

输入　样本集Z（包含已知标签的域名），未知域名样本z^*，非一致度量函数A

输出　未知域名样本z^*在每个类别样本集下的p值$p_{z^*}^C$

1)for 每个聚类标签l在Z中do

2)　初始化子集C'，包含Z中标签为l的样本；

3)　将未知样本z^*添加到C'中，即

$C'=\left\{z_1,\cdots,z_{n-1},z^*\right\}$；

4)　for i←1 to n do

5)　$\alpha_i\leftarrow A\left(c\backslash z_i,z_i\right)$；

6)　end for

7)　$\tau\leftarrow U(0,1)$；

8)　$p_{z^*}^C\leftarrow\dfrac{\#\left\{i:\alpha_i>\alpha_n\right\}+\#\{i:\alpha_i=\alpha_n\}\tau}{n}$；

9)end for

2.2　一致性聚类

CC的核心思想是使用CP来创建聚类，将CP方法应用于无监督聚类任务。传统聚类方法在样本分配过程中通常不提供置信度衡量，导致无法评估每个簇的可靠性或统计显著性。CC通过设置置信度水平的阈值来控制那些不属于任何簇的对象数量，从而提升聚类的可靠性。较高的置信度水平意味着更严格的条件，只有当对象和簇的样本高度一致时，才会被分配到该簇中。

CC的基本思想是将CP中的不一致度度量和p值应用于聚类任务。CC通过衡量新样本点与现有簇（或类别）之间的不一致性为新样本点分配一个p值。根据该p值，形成置信区间，以确定该样本点是否应被分配到现有簇中，或者将其归类为异常。本文使用算法2中描述的CC方法来对新样本进行聚类，以下是方法实现的详细过程。

1)本文将数据集表示为特征向量，并用于训练一致性预测器。一致性预测器在特征向量的覆盖范围内创建一个虚拟特征空间，可表示为一个具有一定分辨率的表格区域。

2)循环遍历虚拟特征空间中的每一个点，将它们视为测试对象，并使用CP方法计算每个点的p值，表示该点与训练数据的一致性程度。

3)本文使用预定义的显著性水平ε作为参数，调整簇的边界以控制簇的大小。较大的ε值会识别出更多的异常点，从而减小簇的大小。p值大于显著性水平ε的点被归类为一致性区域，这些区域与训练数据分布具有较高的相似度，通常由多个簇组成。

4)实际观测值被投影到虚拟特征空间中，并根据它们在一致性区域中的位置进行分类，形成簇并标记异常点。

算法2　域名的一致性聚类

输入　D_{mal}、D_{ben}（恶意域名和良性域名数据集），D_{unknown}（未知域名数据集）

输出　D_{unknown}的预测结果，网格上预测的恶意域名区域

1)收集D_{mal}、D_{ben}，包含特征；
2)收集D_{unknown}；
3)从每个域名提取特征向量；
4)对特征F进行预处理（如归一化、标准化）；
5)对F应用t分布随机邻居嵌入（t-SNE），得到F_{2D};
6)使用D_{mal}中的F_{2D}训练一致性聚类；
7)在F_{2D}空间上构建二维网格；
8)for 每个网格点 G_i do
9)　将 G_i视为未知域名；
10)　计算$p_{G_i}^{D_{\text{mal}}}$；
11)　if $p_{G_i}^{D_{\text{mal}}} > \varepsilon$ then
12)　　标记G_i为恶意；
13)　end if
14)end for
15)对F_{2D}使用基于密度的聚类算法（DBSCAN）进行聚类，识别子群（如DGA家族）；
16)for 每个域名$d \in D_{\text{unknown}}$ do
17)　初始化变量 associated为false;
18)　for 每个聚类 do
19)　　if distance$(d,\text{cluster}) \leqslant$ grid distance then
20)　　将d与聚类关联；
21)　　将 associated设置为true；
22)　　end if
23)　end for
24)　if associated为 false then
25)　　标记d不属于任何聚类；
26)　end if
27)　if d与多个聚类关联then
28)　　合并聚类
29)　end if
30)end for

2.3　显著性水平

CC方法需要设置显著性水平，显著性水平是假设检验中的一个概念，表示原假设正确但被错误拒绝的概率，用ε表示，$\varepsilon \in (0,1)$。置信水平为$1-\varepsilon$，表示对所获得置信区间的信任程度。在该方法中，ε被引入来量化模型对训练数据预测的不确定性。CC通过分析数据集的模式，构建一个置信区间，涵盖对未来的预测。如果进行多次预测，方法有$1-\varepsilon$的概率使预测值落在该置信区间内。因此，在实验中，可以通过ε来控制预测的置信水平。

3　系统模型

本节介绍CPDGA的基本框架和原理。如图1所示，CPDGA由3个模块组成。第一个是训练模块，在原始算法生成的域名（AGD，algorithmically-generated domain）进行特征提取后执行，并输出用于数据增强的模型；第二个是数据增强模块，基于设定的阈值和计算出的样本p值，生成恶意域名的传播范围；最后一个是应用模块，用来缓解概念漂移问题和检测对抗恶意域名。

3.1　CPDGA原理

图2为CPDGA的设计原理。在CPDGA模型中，黑点、浅灰点和深灰点分别表示恶意域名样本、良性域名样本和标记为Q的未知域名。基于恶意域名特征的初始分布，本文采用CPDGA构建了恶意域名的预测范围，表示未来可能的分布。图2中的曲线线条大致勾勒出预测范围的边界，落在该范围内的域名被认为是恶意域名。

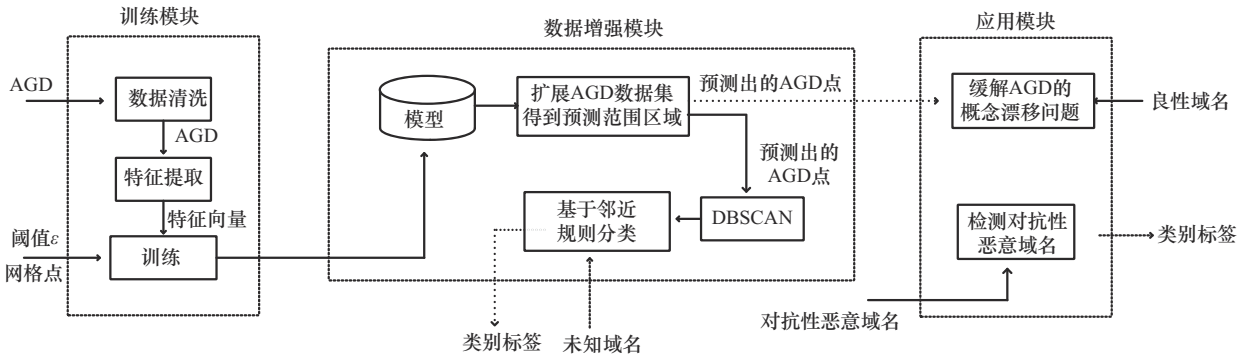

图1　CPDGA基本框架

选择预测范围的具体策略在右侧放大显示：“×”点表示得到的预测范围，标记了网格单元的p值，颜色越深表示p值越高，表明与原始恶意域名样本的一致性越强。本文根据这些p值和设定的显著性水平ε来过滤网格单元，选择的网格单元是恶意域名未来极有可能分布的区域。图2中的3条曲线线条表示通过设定不同ε值得到的预测范围，其中$\varepsilon_0 > \varepsilon_1 > \varepsilon_2$。较小的$\varepsilon$值会导致预测范围的边界更加平滑，涵盖更多的样本点。通过控制显著性水平ε，本文可以控制预测范围的大小，确保最大范围地覆盖恶意域名的分布范围，同时避免错误涵盖良性域名的分布，从而提升对未来域名的分类准确度。以ε_2为例，图中的Q点被包含在恶意域名的预测范围内，因此被分类为恶意域名。

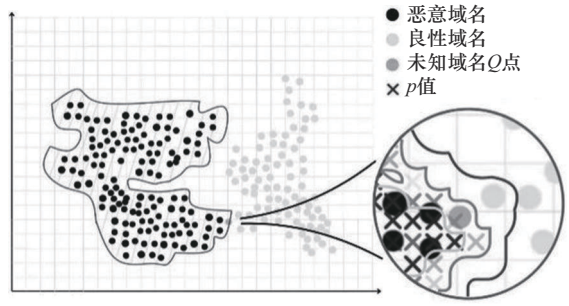

图2　CPDGA的设计原理

3.2　数据预处理

由于大多数DGA仅生成中间域名字符串，本文只关注二级域名和顶级域名。对于收集的域名，本文从结构特征、语言特征和统计特征3个方面提取了34个特征。这些特征不仅从字符串结构的角度捕捉域名的特征，还从语言模式和字符分布的角度关注良性域名与恶意域名之间的差异。具体信息如表1所示，其中1~6表示结构特征（如域名长度、顶级域名长度等），7~21表示语言特征（如元音字符的比例、辅音字符的比例等），22~34表示统计特征（如N-gram字符出现频率的中位数和标准差等）。

在CC的训练过程中，需要计算n^d个p值，其中d是特征维度，n表示d维域名特征网格中的网格点数量。为了直观展示实验结果，本文在实验部分采用t-SNE[49]进行可视化。降维过程中常常面临维度灾难的问题，即高维空间中分明的样本点在低维空间中可能变得密集，导致信息丢失。t-SNE通过保持样本点的分离性并保留数据中的结构和关系，有效解决了这一问题。

表1　基于结构、语言学和统计学的特征

序号	特征	含义
1	domain_len	域名的总长度
2	sld_len	二级域名的长度
3	tld_len	顶级域名的长度
4	tld_dga	是否包含恶意顶级域名
5	tokens_sld	被“-”分割的标记数量
6	flag_dig_tld	是否以数字开头
7	uni_domain	域名中唯一字符的数量
8	uni_sld	二级域名中唯一字符的数量
9	uni_tld	顶级域名中唯一字符的数量
10	digits_sld	二级域名中数字的数量
11	sym_sld	特殊字符的比例
12	hex_sld	十六进制字符的比例
13	dig_sld	数字数量
14	vow_sld	元音字母的比例
15	con_sld	辅音字母的比例
16	repeat_letter_sld	重复字母的比例
17	rep_char_ratio_sld	重复字符与唯一字符的比例
18	cons_con_ratio_sld	连续辅音的比例
19	cons_dig_ratio_sld	连续数字的比例
20	gib_value_sld	Gib检测
21	hmm_log_prob_sld	隐马尔可夫模型书写检测
22	entropy_sld	香农熵
23	gram2_med_sld	2阶元字符频率
24	gram3_med_sld	3阶元字符频率
25	gram2_cmed_sld	2阶条件元字符频率
26	gram3_cmed_sld	3阶条件元字符频率
27	avg_gram1_rank_sld	1阶元字符频率排序的平均值
28	avg_gram2_rank_sld	2阶元字符频率排序的平均值
29	avg_gram3_rank_sld	3阶元字符频率排序的平均值
30	std_gram1_rank_sld	1阶元字符频率排序的标准差
31	std_gram2_rank_sld	2阶元字符频率排序的标准差
32	std_gram3_rank_sld	3阶元字符频率排序的标准差
33	gini	Gini值
34	cer	字符分类错误

3.3 训练模块

本节通过使用已知恶意域名进行训练，创建了一组DGA域名预测模型，用于预测和划分DGA域名的分布。首先，清洗数据集并执行特征提取操作，详细过程见3.2节。其次，使用训练集数据构建一个d维网格，其中每个网格单元表示特征空间中的一个区域。计算每个网格单元内样本的p值，并设置一个阈值，详细计算过程见2.2节。一致性度量函数采用K最近邻（KNN, K-nearest neighbor）算法。最后，训练模块输出用于数据增强的模型，准备扩展数据集，以缓解概念漂移。

3.4 数据增强模块

数据增强模块首先基于训练模块获得的模型扩展原始数据集，增加用于范围预测的数据量（该数据集可用于缓解概念漂移现象）。其次，对预测的域名执行DBSCAN聚类过程。最后，基于数据点之间的相似性对未知域名进行分类（用于检测对抗性恶意域名）。具体规则是，如果测试对象与某个域名簇中的点之间的距离小于或等于网格单元的大小，则认为该测试对象属于该簇。如果测试对象与多个簇相关联，则这些簇会被合并。具体过程如下：首先，设训练数据集为$D_{\text{train}}=\{x_1,x_2,\cdots,x_n\}$，一致性预测器根据$D_{\text{train}}$建立虚拟特征空间$S\subset R^d$，通过在$S$内均匀采样生成候选点集合$S'=\{s_1,s_2,\cdots,s_m\}$，并计算每个候选点的$p$值$p(s_i)$。对于每个候选点$s_i$，若其$p$值满足$p(s_i)\geqslant\varepsilon$，其中，$\varepsilon$为设定的置信水平阈值，则将$s_i$选入增强数据集$D_{\text{aug}}$，否则舍弃。最终扩展得到的新数据集为$D_{\text{new}}=D_{\text{train}}\cup D_{\text{aug}}$。然后，在$D_{\text{new}}$上应用DBSCAN，设DBSCAN的半径参数为$\delta$，最小簇大小为Minpts，则将满足以下条件的数据点划分为同一簇：对于任意点对(x_i,x_j)，若$\text{dist}(x_i,x_j)\leqslant\delta$，其中$\text{dist}(\cdot,\cdot)$是欧氏距离或其他相似性度量函数，则认为$x_i$和$x_j$在同一密度可达区域内。同时，每个聚簇至少包含Minpts个点。在测试阶段，对于未知域名样本x_{test}，其分类规则如下：若存在聚簇C_k，使$\exists x_i\in C_k,\text{dist}(x_{\text{test}},x_i)\leqslant\Delta$。其中，$\Delta$为网格单元大小（即虚拟特征空间的离散尺度），则将$x_{\text{test}}$归入聚簇$C_k$。若测试对象同时与多个簇满足上述条件，则这些簇将被合并，x_{test}被分配至合并后的新簇。若测试对象不满足任何已有聚簇的条件，则将其标记为异常点或新类别。通过上述流程，数据增强模块有效扩展了原始样本的分布范围，提升了模型对概念漂移及对抗性恶意域名的适应能力，并且利用一致性预测生成的p值筛选增强样本，确保扩展数据具有分布合理性。

3.5 应用模块

基于数据增强模块的输出，本文开发了2个应用程序。第1个应用程序旨在缓解恶意域名检测中的概念漂移现象。实验结果表明，使用CPDGA能够有效地预测未来的恶意域名分布。第2个应用程序旨在预测由最新对抗性模型生成的域名。该应用程序处理输入数据，包括由对抗性模型生成的恶意域名样本和良性域名。结果表明，该应用程序在识别对抗性模型生成的恶意域名方面表现良好。

4 实验分析

本节利用2个实验来评估CPDGA。第1个实验评估CPDGA在缓解训练模型中概念漂移问题方面的效率。第2个实验评估CPDGA识别由13个对抗模型生成的恶意域名的能力。

4.1 数据集

本文为2个实验分别构建了独立的训练集和测试集。在第1个实验中，本文构建了一个包含恶意和良性域名的数据集。恶意域名数据集来自恶意域名文档（DGArchive），其中恶意域名通过重新实现逆向工程的变量生成算法并使用已知种子生成，是一个高度可靠的恶意域名数据来源。本文的数据集涵盖了从2019年10月到2023年5月的时间段，能够捕捉并分析恶意DGA域名的长期演变趋势。良性域名数据集来自Tranco排名前百万网站的随机样本。在第2个实验中，恶意域名数据集由13个已知对抗恶意算法生成，每个算法生成的域名数量为10 000个。良性域名集同样来自Tranco排名前百万网站的随机样本。本文使用的恶意域名数据集和良性域名数据集示例如表2所示。

表2 恶意域名和良性域名数据集简单示例

恶意域名数据集	良性域名数据集
decidemanager.com	google.com
mrmonopol.de	baidu.com
Hartromboblood.club	youtube.com

4.2　评估CPDGA缓解概念漂移的能力

本节展示了CPDGA在预测恶意域名范围方面的有效性。本文从2019—2023年收集的恶意域名数据中随机选择10 000个域名样本。在对选定的样本进行特征提取和标准化后，使用t-SNE进行降维。2019年的数据作为训练集，用于预测域名的分布；2020—2023年的数据作为单独的测试集，映射到预测区域，DGA域名的预测范围结果如图3所示。颜色深浅表示不同的p值，即预测点与训练集的一致性概率。颜色越深，p值越高，表示该点与训练集中的数据一致性更强，与训练数据的相似度更高。

随着时间的推移，域名的特征分布逐渐发生变化，年份相隔越远，恶意域名分布范围之间的差异越明显，这是概念漂移的显著表现。此外，本文发现，最初聚集在一起的恶意域名变得更加分散。这表明恶意域名的生成方式趋于多样化，导致特征分布逐渐扩散。因此，恶意域名的特征变得越来越复杂，与良性域名的区分难度也随之增加。

基于2019年数据训练的恶意域名范围预测模型显示，2020—2023年的测试数据在预测范围内被不同程度覆盖。通过邻近规则，本文评估了预测范围对后续年份（2020—2023年）测试集的覆盖能力。结果表明，2020年恶意域名数据的覆盖率

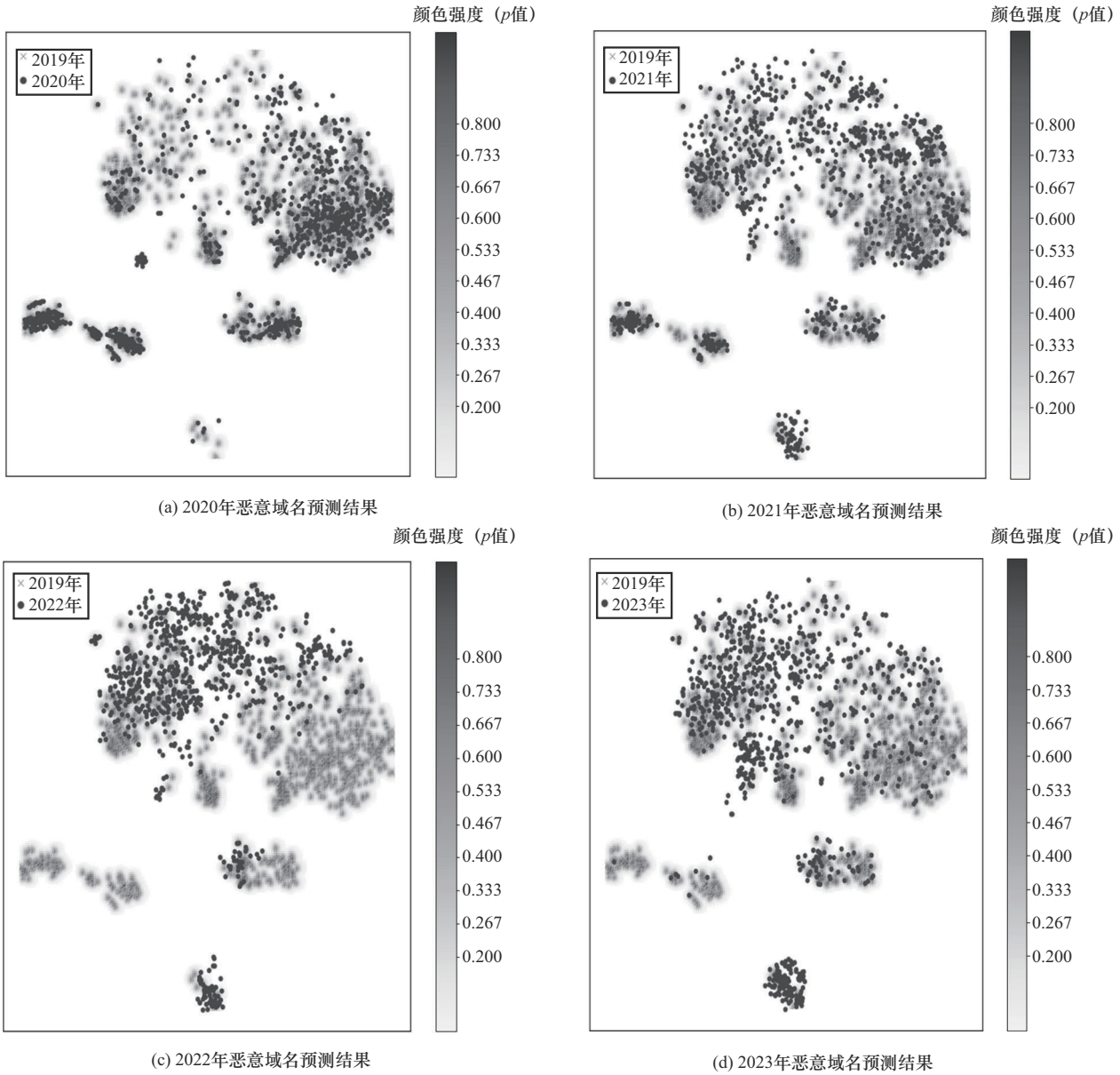

图3　DGA域名的预测范围结果

为95.1%，2021年为91.2%，2022年为71.0%，2023年为77.6%。CPDGA在预测未来年份样本方面展示了一定的有效性，能够基于现有的恶意域名特征传播更多潜在的特征点，代表未来可能的恶意活动。这形成了未来恶意域名分布的预测范围，展示了CPDGA预见概念漂移趋势的能力。然而，覆盖率随着时间呈下降趋势，表明本文算法存在一定的时间依赖性。

4.2.1 消融实验

为了评估CPDGA及其各个组成部分对恶意域名检测性能的影响，本文设计了以下消融实验。实验使用了由2019年10 000个良性域名和恶意域名组成的训练集，测试集则包含从2023年随机抽取的2 500个良性和恶意域名。

本文比较了4种不同配置的模型，包括完整的CPDGA及其去除不同模块后的版本。具体配置和实验结果如下。

1)CPDGA（完整）

CPDGA包含CP和CC 2个核心模块，能够综合利用两者的优势进行恶意域名的检测。在所有6种模型中，使用完整CPDGA的检测准确率普遍较高，尤其在决策树（DT, decision tree）模型中，准确率达到了78.0%。

2)去除CP

该配置去除了CP部分，仅保留CC。去除CP后，检测准确率略有下降，尤其在线性判别分析（LDA, linear discriminant analysis）和自适应增强（AdaBoost，adaptive boosting）模型上，准确率分别下降了1.6%和2.2%。

3)去除CC

该配置去除了CC部分，仅保留CP。去除CC后，虽然准确率仍然较高，但比完整CPDGA略低。在DT和极限梯度提升（XGBoost, extreme gradient boosting）模型中，分别下降了1.2%和1.1%。

4)去除CP和CC（传统方法）

该配置完全去除CP和CC，采用传统的恶意域名检测算法。传统的恶意域名检测算法的检测准确率明显较低，尤其在LDA和DT模型中，准确率仅为49.0%和57.6%，相比完整CPDGA下降约10%。

表3所示的实验结果表明，CPDGA在所有6种模型中均表现出较强的检测能力，尤其是在DT模型中，准确率达到78.0%。去除CP或CC部分都会导致检测性能有所下降，特别是在去除两者后，准确率显著低于完整算法。传统方法的检测准确率明显较低，证明了CPDGA在应对恶意域名检测任务时的显著优势。

4.2.2 CPDGA对传统检测模型性能的提升

本文测试了XGBoot、LSTM、随机森林（B-RF）和CNN 4种常用的DGA检测模型，使用2019年的域名数据对这些模型进行训练，评估它们在2019—2023年检测恶意域名的性能，如表4~表7所示。由表4~表7可知，这些模型的准确率、精确度、召回率和F1分数从2019—2023年逐年下降。并且随着时间的推移，最初有效的检测模型可能会逐渐失效，因为它们未能适应新的数据分布，无法应对域名概念漂移现象。

为了进一步评估CPDGA在预测未来恶意域名方面的有效性，并衡量预测得到的恶意域名点的价值，本文选择6种传统的机器学习模型，包括LDA、DT、XGBoost、RF、GBDT和AdaBoost，并设置4种训练场景进行对比。对于训练集中的恶意域名数据，第1种未使用任何增强数据作为训练集；第2种使用KNN生成的增强数据作为训练集；第3种使用高斯混合模型（GMM, Gaussian mixture model）生成的增强数据作为训练集；第4种使用CPDGA生成的增强数据做训练集。测试集则由从随后的年份（2020—2023年）中随机选择的2 500个恶意域名和良性域名组成。

表3　　消融实验

方法	LDA	DT	XGBoost	RF	GBDT	AdaBoost
使用CPDGA（完整）	59.2%	78.0%	76.4%	71.8%	72.0%	75.4%
去除CP	57.6%	75.2%	74.1%	69.5%	70.3%	73.2%
去除CC	58.5%	76.8%	75.3%	70.2%	71.1%	74.3%
去除CP和CC	49.0%	57.6%	58.8%	61.4%	60.4%	62.8%

表4　2019—2023年传统恶意域名检测模型XGBoost的性能演变情况

年份	准确率	精确率	召回率	F1分数
2019年	79.04%	94.89%	62.40%	75.29%
2020年	75.85%	93.87%	56.52%	70.55%
2021年	78.16%	64.80%	41.85%	50.85%
2022年	68.86%	63.02%	43.41%	58.79%
2023年	68.02%	51.39%	22.23%	31.04%

表5　2019—2023年传统恶意域名检测模型LSTM的性能演变情况

年份	准确率	精确率	召回率	F1分数
2019年	83.14%	99.99%	83.18%	90.82%
2020年	85.00%	92.95%	76.48%	83.92%
2021年	80.37%	72.85%	43.44%	54.42%
2022年	73.06%	79.85%	63.35%	70.65%
2023年	60.53%	74.06%	20.80%	32.48%

表6　2019—2023年传统恶意域名检测模型B-RF的性能演变情况

年份	准确率	精确率	召回率	F1分数
2019年	80.01%	91.95%	66.79%	77.38%
2020年	77.19%	90.97%	61.54%	73.41%
2021年	73.75%	51.69%	42.43%	46.61%
2022年	66.50%	85.06%	41.91%	56.15%
2023年	64.46%	42.02%	25.79%	31.97%

表7　2019—2023年传统恶意域名检测模型CNN的性能演变情况

年份	准确率	精确率	召回率	F1分数
2019年	95.97%	99.98%	100.00%	99.99%
2020年	90.81%	91.30%	98.64%	84.12%
2021年	87.04%	89.88%	80.17%	82.16%
2022年	85.24%	88.64%	72.11%	81.18%
2023年	84.84%	83.24%	70.46%	80.89%

4种场景的准确率如表8所示。“2019→2020”表示未使用增强数据作为训练集检测2020年域名数据的准确率；“KNN → 2020”表示使用KNN生成的恶意域名数据作为训练集检测2020年域名数据的准确率；“GMM→ 2020”表示使用GMM生成的恶意域名数据作为训练集检测2020年域名数据的准确率；“CPDGA→ 2020”表示使用CPDGA生成的恶意域名数据作为训练集检测2020年域名数据的准确率。

根据表8的实验结果，DT作为基础模型，使用KNN时，检测2023年域名的准确率较未使用算法下降了2.4%；使用GMM时，检测2023年域名的准确率同样较未使用算法时下降了2.4%。与此不同，采用CPDGA时，检测2023年域名的准确率相比未使用算法时提高了20.4%。这一结果表明，CPDGA在DGA域名检测任务中相较于常见的KNN和GMM具有显著的优势。

此外，尽管3种算法在前几年的表现提升较为有限，甚至在某些年份出现准确率下降的趋势，可归因于训练数据仅从预测范围内随机抽样，并未完全覆盖整个可能的预测区域。这种现象反映了传统模型在面对数据分布变化时的局限性，特别是在存在概念漂移的情况下。然而，到了2023年，CPDGA相较于其他2种算法显示了显著的性能提升。本文认为这一结果是由于2019年与2023年数据之间显著的概念漂移，传统机器学习模型未能有效应对这种变化，从而导致准确率大幅下降。实验结果表明，CPDGA相较于KNN和GMM，在应对概念漂移时表现出了更强的适应性，其预测结果在5年期间持续稳定地提高，并且能够有效覆盖未来恶意域名的分布，证明了CPDGA在长期检测任务中的可持续性和实用价值。

4.2.3　CPDGA对先进模型性能的提升

本节与当前最先进的僵尸网络检测模型（异构域名检测模型HAGDetector[50]、基于特征的域名检测模型FANCI[17]、基于字符的域名分类n-CBDC[51]、基于词频–逆向文件频率的检测模型TF-IDF[52]、基于残差网络的检测模型ResNet[53]）进行了对比实验。如表9所示，这5个模型的检测准确率随时间逐年下降，这进一步证明了概念漂移现象的存在。这些模型的预测性能随着新变种和新僵尸网络家族的出现显著下降。此外，当使用CPDGA生成的数据作为训练集进行检测时，检测准确率显著提高，尤其是在检测2023年的恶意域名时，各模型检测准确率达到了最大的提升：相较于HAGDetector模

表8 相较于其他算法对传统检测算法的性能提升

训练集/测试集 (Alexa+DGA)	LDA	DT	XGBoost	RF	GBDT	AdaBoost
2019→2020	83.4%	79.6%	85.4%	86.0%	81.8%	82.4%
KNN→2020	83.0% (↓0.4%)	79.4% (↓0.2%)	70.8% (↓16.8%)	70.4% (↓15.6%)	66.2% (↓15.6%)	66.4% (↓16.0%)
GMM→2020	87.4% (↑4.0%)	90.4% (↑10.8%)	89.6% (↑4.2%)	88.6% (↑2.6%)	87.4% (↑5.6%)	87.2% (↑4.8%)
CPDGA→2020	87.8% (↑4.4%)	82.6% (↑3%)	87.4% (↑2.0%)	90.6% (↑4.6%)	82.6% (↑0.8%)	83.0% (↑6.0%)
2019→2021	66.6%	77.6%	76.4%	77.6%	76.2%	78.4%
KNN→2021	71.2% (↑4.6%)	63.2% (↓14.4%)	58.8% (↓17.6%)	58.6% (↓19.0%)	55.8% (↓20.4%)	55.8% (↓22.6%)
GMM→2021	72.2% (↑5.6%)	73.4% (↓4.2%)	73.2% (↓3.2%)	70.8% (↓6.8%)	70.2% (↓6.0%)	71.6% (↓6.8%)
CPDGA→2021	69.0% (↑2.4%)	82.2% (↑4.6%)	82.6% (↑6.2%)	84.6% (↑7.0%)	85.2% (↑9.0%)	81.4% (↑3.0%)
2019→2022	73.2%	80.4%	79.4%	81.4%	81.0%	75.0%
KNN→2022	52.6% (↓20.6%)	48.2% (↓32.2%)	49.0% (↓30.4%)	50.0%(↓31.4%)	49.4%(↓31.6%)	48.2%(↓26.8%)
GMM→2022	54.0% (↓19.2%)	52.6% (↓27.8%)	51.8% (↓27.6%)	50.8% (↓30.6%)	50.8% (↓30.2%)	51.8% (↓23.2%)
CPDGA→2022	75.6%(↓2.4%)	83.4% (↑3.0%)	87.2% (↑7.8%)	85.6% (↑4.2%)	85.2% (↑4.2%)	79.2% (↑4.2%)
2019→2023	49.0%	57.65	58.8%	61.4%	60.4%	62.8%
KNN→2023	65.2% (↑16.2%)	55.2% (↓2.4%)	51.8%(↓7.0%)	52.6%(↓8.8)	51.2% (↓9.2%)	55.2%(↓7.6%)
GMM→2023	59.8% (↑10.8%)	60.6% (↑3.0%)	60.4% (↑1.6%)	52.6% (↓8.8%)	57.6% (↓2.8%)	58.2% (↓4.6%)
CPDGA→2023	59.2% (↑10.2%)	78.0% (↑20.4%)	76.4% (↑17.6%)	71.8%	72.0% (↑11.6%)	75.4% (↑12.6%)

表9 相较于最先进检测算法的性能提升

训练集/测试集 (Alexa+DGA)	HAGDetector[50]	FANCI[17]	n-CBDC[51]	TF-IDF[52]	ResNet[53]
2019→2020	89.8%	75.1%	77.5%	92.1%	91.9%
CPDGA→2020	92.1% (↑2.3%)	78.6% (↑3.5%)	81.6% (↑4.1%)	94.6% (↑2.1%)	93.2% (↑2.8%)
2019→2021	85.4%	73.8%	74.2%	90.4%	85.4%
CPDGA→2021	91.2% (↑5.8%)	76.3% (↑2.5%)	78.7% (↑4.5%)	93.6% (↑3.2%)	88.6% (↑3.3%)
2019→2022	80.5%	69.4%	78.6%	81.2%	83.1%
CPDGA→2022	86.3% (↑6.3%)	74.6% (↑5.2%)	80.5% (↑1.9%)	89.7% (↑8.5%)	87.6% (↑4.5%)
2019→2023	70.8%	60.9%	64.2%	84.5%	82.9%
CPDGA→2023	85.1% (↑14.3%)	78.6% (↑17.7%)	79.6% (↑15.4%)	90.2% (↑5.7%)	89.2% (↑6.3%)

型提高了14.3%，相较于FANCI模型提高了17.7%，相较于n-CBDC模型提高了15.4%，相较于TF-IDF模型提高了5.7%，相较于ResNet模型提高了6.3%。这一结果与传统模型下的检测结果一致，进一步验证了CPDGA相较于最新的检测模型具有显著的优势。

上述研究表明，CPDGA生成的数据可以有效应对概念漂移现象，并提升各种最新检测模型对新兴恶意域名的检测性能，进一步证明了其在僵尸网络检测应用中的潜力和有效性。

4.3 CPDGA检测对抗模型能力的评估

鉴于CPDGA在恶意域名检测中强大的概念漂移缓解能力，本文进一步评估了它检测由对抗模型生成的恶意域名的有效性。本节使用的数据集包含恶意域名和良性域名，恶意域名数据集分别由13种对抗模型生成的域名组成，良性域名数据集则通过从Tranco排名前百万的网站中随机选择获得。

本文对13种对抗性模型进行了一系列实验，并将CPDGA与当前最先进的检测算法进行了比较。实验结果如表10所示。在实验设置中，训练

集占总数据集的5%，测试集占总数据集的95%。这种设置使模型在测试时会遇到大量未知数据，能够很好地评估模型的泛化能力。在实际应用中，训练数据可能相对较少，模型需要对大量新数据进行预测。通过这种测试方式，可以模拟真实场景，验证模型在数据有限的情况下是否仍能取得良好的预测结果。如表10所示，CPDGA在检测对抗模型生成的域名方面表现良好。具体而言，它在检测13种对抗模型中的8种所生成的域名时表现尤为突出。这表明，CPDGA在处理由多种对抗模型生成的恶意域名时具有一定的优势，特别是在8种特定算法的检测中取得了较高的准确率。尽管在检测其余5种对抗模型（文献[27]、文献[54]、文献[55]、文献[56]和文献[42]）时，CPDGA的表现稍逊色于HAGDetector和TF-IDF，但其检测准确率仍然超过90%。因此，CPDGA在对抗性恶意域名检测中展现了良好的泛化能力和鲁棒性，能够有效应对多种复杂的对抗生成算法。

表10　当前最先进检测模型在检测对抗模型生成域名方面的准确率

对抗模型	HAG Detection[50]	n-CBDC[51]	TF-IDF[52]	ResNet[53]	CPDGA
文献[27]	94.00%	58.30%	50.80%	52.60%	93.58%
文献[40]	52.00%	50.00%	52.00%	50.00%	90.74%
文献[41]	66.00%	57.90%	62.40%	42.10%	95.47%
文献[42]	89.00%	72.90%	99.60%	54.60%	96.42%
文献[54]	96.00%	50.30%	62.70%	85.30%	85.63%
文献[55]	53.00%	50.00%	94.50%	52.30%	91.89%
文献[56]	85.00%	85.80%	96.40%	50.00%	94.21%
文献[57]	59.00%	58.00%	61.10%	42.00%	92.00%
文献[58]	74.00%	66.40%	79.60%	42.10%	86.53%
文献[59]	65.00%	54.30%	50.80%	50.00%	92.21%
文献[60]	76.00%	50.00%	57.70%	57.30%	91.79%
文献[61]	61.00%	50.90%	51.10%	50.00%	90.74%
文献[62]	81.00%	53.10%	57.30%	50.00%	95.26%

5　结束语

通过分析和建模恶意域名的演化过程，本文发现其概念漂移是渐进式的，主要特征随着时间的推移逐步且系统地演变。为此，本文提出了一种基于一致性传播的DGA域名主动检测算法。

在对2019—2023年的恶意域名数据进行概念漂移分析过程中，CPDGA成功地识别出恶意域名分布的潜在特征。本文算法在缓解概念漂移问题上取得了显著进展，相较于基础分类模型，恶意域名检测准确率提升幅度可达20.4%。通过将CPDGA与最先进的检测模型对比，展示了其在对抗恶意域名检测方面的优异性能，在检测13种最新对抗性域名时最高准确率可达96.42%。上述实验结果充分表明了CPDGA在恶意域名威胁的主动防御方面具有显著有效性，能够为网络安全防护提供有力支持。

参考文献：

[1] MOCKAPETRIS P, DUNLAP K J. Development of the domain name system[J]. ACM SIGCOMM Computer Communication Review, 1988, 18(4): 123-133.

[2] EASTLAKE D, KAUFMAN C. Domain name system security extensions[R]. 2016.

[3] PANG J, HENDRICKS J, AKELLA A, et al. Availability, usage, and deployment characteristics of the domain name system[C]//Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2004: 1-14.

[4] WANG J Y, CHITSAZ F, DERBYSHIRE M K, et al. The conserved domain database in 2023[J]. Nucleic Acids Research, 2023, 51(1): 384-388.

[5] GHEORGHIȚĂ C A, SMADA D, VEVERA A V, et al. Blacklists and whitelists in the framework of a domain reputation system[J]. Romanian Journal of Information Technology and Automatic Control, 2023, 33(4): 33-46.

[6] ARORA A. Improving the efficiency of a new malicious domain prediction system[R]. 2023.

[7] SKULA I, KVET M. Domain blacklist efficacy for phishing web-page detection over an extended time period[C]//Proceedings of the 33rd Conference of Open Innovations Association (FRUCT). Piscataway: IEEE Press, 2023: 257-263.

[8] SUN X J, LIU Z F. Domain generation algorithms detection with feature extraction and domain center construction[J]. PLoS One, 2023, 18 (1): e0279866.

[9] PEREIRA M, COLEMAN S, YU B, et al. Dictionary extraction and detection of algorithmically generated domain names in passive DNS traffic[C]//Proceedings of the 21st International Symposium Research in Attacks, Intrusions, and Defenses. Berlin: Springer, 2018: 295-314.

[10] YADAV S, REDDY A K K, NARASIMHA REDDY A L, et al. Detecting algorithmically generated domain-flux attacks with DNS traffic analysis[J]. IEEE/ACM Transactions on Networking, 2012, 20(5): 1663-1677.

[11] ANTONAKAKIS M, PERDISCI R, NADJI Y. From throw-away traffic to bots: Detecting the rise of DGA-based malware [C]//Proceedings of the 21st USENIX Security Symposium (USENIX Security 12). Berkeley: USENIX Association, 2012:491-506.

[12] YUAN J T, CHEN G X, TIAN S W, et al. Malicious URL detection based on a parallel neural joint model[J]. IEEE Access, 2021, 9: 9464-9472.

[13] SCHIAVONI S, MAGGI F, CAVALLARO L, et al. Phoenix: DGA-based botnet tracking and intelligence[C]//International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin: Springer, 2014: 192-211.

[14] HASSAOUI M, HANINI M, EL KAFHALI S. Domain generated algorithms detection applying a combination of a deep feature selection and traditional machine learning models[J]. Journal of Computer Security, 2023, 31(1): 85-105.

[15] LAU S Q. Domain analysis of e-commerce systems using feature-based model templates[D]. Waterloo: University of Waterloo, 2006.

[16] HARIRI N, CASTRO-HERRERA C, MIRAKHORLI M, et al. Supporting domain analysis through mining and recommending features from online product listings[J]. IEEE Transactions on Software Engineering, 2013, 39(12): 1736-1752.

[17] SCHÜPPEN S, TEUBERT D, HERRMANN P, et al. FANCI: Feature-based automated NXDomain classification and intelligence[C]//Proceedings of the 27th USENIX Security Symposium (USENIX Security 18). Berkeley: USENIX Association, 2018:1165-1181.

[18] ZHAO C, ZHANG Y Z, ZANG T N, et al. A multi-feature-based approach to malicious domain name identification from DNS traffic[C]// Proceedings of the 2020 27th International Conference on Telecommunications (ICT). Piscataway: IEEE Press, 2020: 1-5.

[19] CHENG Y N, CHAI T T, ZHANG Z X, et al. Detecting malicious domain names with abnormal WHOIS records using feature-based rules[J]. The Computer Journal, 2022, 65(9): 2262-2275.

[20] CHOW T, KAN Z L, LINHARDT L, et al. Drift forensics of malware classifiers[C]//Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2023: 197-207.

[21] KAN Z L, MCFADDEN S, ARP D, et al. TESSERACT: eliminating experimental bias in malware classification across space and time (extended version)[J]. arXiv Preprint, arXiv: 2402.01359, 2024.

[22] GIBERT D. Machine learning for windows malware detection and classification: methods, challenges and ongoing research[J]. arXiv Preprint, arXiv: 2404.18541, 2024.

[23] PENDLEBURY F, PIERAZZI F, JORDANEY R, et al. TESSERACT: eliminating experimental bias in malware classification across space and time [C]//28th USENIX Security Symposium. Berkeley: USENIX Association, 2019: 729-746.

[24] ŽLIOBAITĖ I, PECHENIZKIY M, GAMA J. An overview of concept drift applications[C]//Big Data Analysis: New Algorithms for a New Society. Berlin: Springer, 2015: 91-114.

[25] SINGHAL S, CHAWLA U, SHOREY R. Machine learning & concept drift based approach for malicious website detection[C]//Proceedings of the 2020 International Conference on COMmunication Systems & NETworks (COMSNETS). Piscataway: IEEE Press, 2020: 582-585.

[26] RUANO-ORDÁS D, FDEZ-RIVEROLA F, MÉNDEZ J R. Concept drift in e-mail datasets: an empirical study with practical implications[J]. Information Sciences, 2018, 428: 120-135.

[27] YUN X C, HUANG J, WANG Y P, et al. Khaos: an adversarial neural network DGA with high anti-detection ability[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 2225-2240.

[28] GEFFNER J. End-to-end analysis of a domain generating algorithm malware family[R]. 2013.

[29] NIE L H, ZHAO L P, LI K Q, et al. A game-based adversarial DGA detection scheme using multi-level incremental random forest[J]. IEEE Transactions on Network Science and Engineering, 2024, 11(1): 779-792.

[30] BEHRENDS R, DILLON L, FLEMING S, et al. On the kraken and bobax botnets[R]. 2008.

[31] DEMŠAR J, BOSNIĆ Z. Detecting concept drift in data streams using model explanation[J]. Expert Systems with Applications, 2018, 92: 546-559.

[32] SHAN S, BHAGOJI A. N, ZHENG H. et al. Poison forensics: traceback of data poisoning attacks in neural networks[C]//Proceedings of the 31st USENIX Security Symposium (USENIX Security 22). Berkeley: USENIX Association, 2022: 3575-3592.

[33] YANG W K, LI Z, LIU M C, et al. Diagnosing concept drift with visual analytics[C]//Proceedings of the 2020 IEEE Conference on Visual Analytics Science and Technology (VAST). Piscataway: IEEE Press, 2020: 12-23.

[34] ZOLA F, BRUSE J L, GALAR M. Temporal analysis of distribution shifts in malware classification for digital forensics[C]//Proceedings of the 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). Piscataway: IEEE Press, 2023: 439-450.

[35] ZHAO D, LI H, SUN X W, et al. Detecting DGA-based botnets through effective phonics-based features[J]. Future Generation Computer Systems, 2023, 143: 105-117.

[36] YU B, GRAY D L, PAN J, et al. Inline DGA detection with deep networks[C]//Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2017: 683-692.

[37] TRAN D, MAC H, TONG V, et al. A LSTM based framework for handling multiclass imbalance in DGA botnet detection[J]. Neurocomputing, 2018, 275: 2401-2413.

[38] AONZO S, HAN Y, MANTOVANI A, et al. Humans vs. machines in malware classification[C]//Proceedings of the 32nd USENIX Security Symposium. Berkeley: USENIX Association, 2023: 1145-1162.

[39] BITAAB M, CHO H, OEST A, et al. Beyond phish: toward detecting fraudulent e-commerce websites at scale[C]//Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2023: 2566-2583.

[40] ANDERSON H S, WOODBRIDGE J, FILAR B. DeepDGA: adversarially-tuned domain generation and detection[C]//Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2016: 13-21.

[41] PECK J, NIE C, SIVAGURU R, et al. CharBot: a simple and effective method for evading DGA classifiers[J]. IEEE Access, 2019, 7: 91759-91771.

[42] HU X Y, CHEN H, LI M, et al. ReplaceDGA: BiLSTM-based adversarial DGA with high anti-detection ability[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4406-4421.

[43] ZAGO M, GIL PÉREZ M, MARTÍNEZ PÉREZ G. UMUDGA: a dataset for profiling DGA-based botnet[J]. Computers & Security, 2020, 92: 101719.

[44] TUAN T A, LONG H V, TANIAR D. On detecting and classifying DGA botnets and their families[J]. Computers & Security, 2022, 113: 102549.

[45] SHAFER G, VOVK V. A tutorial on conformal prediction[J]. arXiv Preprint, arXiv: 0706.3188, 2007.

[46] CHERUBIN G, NOURETDINOV I, GAMMERMAN A, et al. Conformal clustering and its application to botnet traffic[C]//Proceedings of the 3rd International on Statistical Learning and Data Sciences. Berlin: Springer, 2015: 313-322.

[47] BARBERO F, PENDLEBURY F, PIERAZZI F, et al. Transcending TRANSCEND: revisiting malware classification in the presence of concept drift[C]//Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2022: 805-823.

[48] PARK S, BASTANI O, KIM T. ACon²: adaptive conformal consensus for provable blockchain oracles[C]//Proceedings of the 32nd USENIX Security Symposium. Berkeley: USENIX Association, 2023: 3313-3330.

[49] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2479-2605.

[50] LIANG J B, CHEN S H, WEI Z L, et al. HAGDetector: heterogeneous DGA domain Name detection model[J]. Computers & Security, 2022, 120: 102803.

[51] XU C Y, SHEN J Z, DU X. Detection method of domain names generated by DGAs based on semantic representation and deep neural network[J]. Computers & Security, 2019, 85: 77-88.

[52] VRANKEN H, ALIZADEH H. Detection of DGA-generated domain names with TF-IDF[J]. Electronics, 2022, 11(3): 414.

[53] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[54] SHU X, CAO C J, WANG L J, et al. GWDGA: an effective adversarial DGA[C]//International Conference on Frontiers in Cyber Security. Berlin: Springer, 2022: 30-48.

[55] SPOOREN J, PREUVENEERS D, DESMET L, et al. Detection of algorithmically generated domain names used by botnets: a dual arms race[C]//Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. New York: ACM Press, 2019: 1916-1923.

[56] LIU Q H, YU G, WANG Y Y, et al. A novel DGA domain adversarial sample generation method by geometric perturbation[C]//Proceedings of the 3rd International Conference on Advanced Information Science and System. New York: ACM Press, 2021: 1-10.

[57] NIE L H, SHAN X Y, ZHAO L P, et al. PKDGA: a partial knowledge-based domain generation algorithm for botnets[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4854-4869.

[58] SIDI L, NADLER A, SHABTAI A. MaskDGA: an evasion attack against DGA classifiers and adversarial defenses[J]. IEEE Access, 2020, 8: 161580-161592.

[59] ZHAI Y, YANG J, WANG Z X, et al. Cdga: a GAN-based controllable domain generation algorithm[C]//Proceedings of the 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Piscataway: IEEE Press, 2022: 352-360.

[60] LIU W P, ZHANG Z L, HUANG C, et al. CLETer: a character-level evasion technique against deep learning DGA classifiers[J]. ICST Transactions on Security and Safety, 2021, 7(24): 168723.

[61] CORLEY I, LWOWSKI J, HOFFMAN J. DomainGAN: generating adversarial examples to attack domain generation algorithm classifiers[J]. arXiv Preprint, arXiv: 1911.06285, 2019.

[62] ZHENG Y, YANG C, YANG Y Z, et al. ShadowDGA: toward evading DGA detectors with GANs[C]//Proceedings of the 2021 International Conference on Computer Communications and Networks (ICCCN). Piscataway: IEEE Press, 2021: 1-8.

[作者简介]

刘双双（1999– ），女，山东菏泽人，南开大学博士生，主要研究方向为恶意域名、恶意代码检测、后门攻击和密码学等。

王志（1981– ），男，山西长治人，博士，南开大学副教授，主要研究方向为计算机病毒分析与防治、二进制代码逆向分析等。

董伊萌（2002– ），女，天津人，南开大学硕士生，主要研究方向为机器学习与深度学习在网络安全中的应用。

李万鹏（1988– ），男，四川达州人，博士，利物浦大学助理教授，主要研究方向为信息安全、身份管理系统、漏洞挖掘、恶意代码检测等。