

基于多层次特征融合的APT域名归因方法

马利昭¹, 黄诚¹, 黎明², 牟浩天¹, 彭佳仁¹, 谢启亮¹

(1. 四川大学网络空间安全学院, 四川 成都 610207; 2. 中国电子科技集团航空电子有限公司, 四川 成都 611731)

摘要: 高级持续性威胁 (advanced persistent threat, APT) 因其高度定向性、技术复杂性和长期潜伏性, 对国家关键基础设施和重点行业构成严峻的安全挑战。在现实场景中, APT 检测既需实现早期预警, 又面临 APT 域名数据稀缺的困境, 导致传统基于通信流量特征或依赖大规模训练数据的神经网络方法难以奏效。为此, 提出了一种基于多层次特征融合的 APT 域名归因方法。该方法融合了文本、统计、语义和关联属性 4 个层次的特征, 将归因任务分解为 APT 域名检测与归因两个阶段。在检测阶段, 通过提取域名的文本特征、统计特征和语义特征, 结合支持向量机 (support vector machine, SVM) 分类器, 实现从海量域名中高效筛选潜在 APT 域名。在归因阶段, 补充由 RoBERTa 模型提取的深层语义特征, 结合基于外部数据库查询的域名系统 (domain name system, DNS) 解析信息等关联属性特征, 构建多层次特征表示。接下来采用 K-means 聚类算法对检测出的 APT 域名进行关联分析, 实现向特定攻击组织的归因。实验结果表明, 该方法在检测任务中 F1 分数达 97.47%, 在归因任务中调整兰德指数 (ARI) 达 86.85%、标准化互信息 (NMI) 达 90.74%。在真实场景中, 该方法相较公开报告平均提前 112 天标记出 18 个 APT 域名, 实现了从被动响应到主动防御的转变。

关键词: 高级持续性威胁; 多层次特征融合; 恶意域名检测; APT 域名归因; 提前预警

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ25254

APT domain attribution based on multi-level feature fusion

Ma Lizhao¹, Huang Cheng¹, Li Ming², Mou Haotian¹, Peng Jiaren¹, Xie Qiliang¹

1. School of Cyber Science and Engineering, Sichuan University, Chengdu 610207, China

2. China Electronic Technology Avionics Co., Ltd., Chengdu 611731, China

Abstract: APT activities posed severe challenges to national critical infrastructure and key industries due to their highly targeted, technically complex, and persistent nature. However, early-warning capabilities were required in real-world APT detection, while APT domain data were scarce, which rendered conventional methods based on communication-traffic features or data-intensive neural networks ineffective. To address this issue, an APT domain attribution method based on multilevel feature fusion was proposed, in which textual, statistical, semantic, and associative attribute features were integrated and the overall problem was decomposed into two stages: detection and attribution. In the detection stage, potential APT domains were efficiently screened from massive data by extracting textual, statistical, and semantic features, and classification was performed using a SVM classifier. In the attribution stage, deep semantic features were extracted by the RoBERTa model, and associative attribute features such as DNS resolution were supplemented via external database queries to construct a multilevel representation; correlation analysis

收稿日期: 2025-10-14; 修回日期: 2026-01-17

通信作者: 黄诚, codesec@scu.edu.cn

基金项目: 国家自然科学基金面上项目 (No. 62472296); 四川省科技厅项目 (No. 2024ZHCG0197)

Foundation Items: The National Natural Science Foundation of China General Program (No. 62472296), Sichuan Provincial Department of Science and Technology Project (No. 2024ZHCG0197)

and attribution to specific threat actors were then conducted using K-means clustering. The experimental results show that this method achieves an F1-score of 97.47% in the detection task, and in the attribution task, it reaches an ARI of 86.85% and an NMI of 90.74%. In real-world testing, 18 APT domains were identified on average 112 days earlier than public reports, and proactive defense is enabled by shifting from passive response to early warning and defense.

Key words: advanced persistent threat, multilevel feature fusion, malicious domain name detection, APT domain attribution, early warning

0 引言

应对高级持续性威胁（APT）攻击已成为国家关键基础设施和重点行业面临的最严峻网络安全挑战之一。2025年，360数字安全集团高级威胁研究院累计捕获针对我国的APT攻击1300余起，受影响的单位主要分布于政府、教育、科研、交通运输等15个行业领域^[1]。与普通网络攻击不同，APT攻击具有高度定向性、技术复杂性和长期潜伏性特征，攻击者常借助域名生成算法批量生成随机域名，以躲避传统基于签名和规则匹配方法的检测^[2]。

早期针对恶意域名检测的研究^[3-6]，大多采用特征工程与传统机器学习相结合的方法。这类方法的核心在于提取出可以有效区分不同类别样本的特征。这些特征主要分为两类：一类是基于网络行为的动态特征，例如从域名系统（DNS）流量与日志记录中提取的特征，这类特征只能对已发生的攻击或网络活动进行分析，存在滞后性，难以满足APT检测的早期预警需求；另一类是基于域名字符串的静态特征，这类特征虽解决了滞后性问题，但仍未摆脱特征工程范式的固有局限（即特征设计高度依赖安全专家的领域知识和实战经验），导致泛化能力有限，难以应对APT攻击手法的持续演进。

为突破人工特征工程的瓶颈，后续研究^[7-9]逐渐转向以深度学习为代表的自动特征学习方法。通过构建端到端的神经网络模型，这类方法能够从原始数据中自动挖掘深层的恶意模式，从而降低对专家知识的依赖。不过，深度学习方法的有效性高度依赖大规模、高质量且类别均衡的训练数据。在APT攻击场景中，由于其具有高度定向性和隐蔽性，APT域名训练数据存在恶意样本稀缺和类别不平衡问题。这不仅容易导致模型过拟合，也使其难以在真实网络环境中有效部署。

针对上述问题，本文提出一种基于多层次特征融合的APT域名归因方法。该方法融合了文本、统计、语义、关联属性4个层次的特征，将归因任务分解为APT域名检测与APT域名归因两个阶段。在检测阶段，通过提取域名的文本特征、统计特征和语义特征，并采用支持向量机（SVM）作为分类器，可以从海量域名中快速筛选出潜在的APT域名，减轻后续归因阶段的计算负担。在归因阶段，本文引入预训练模型RoBERTa，利用其在预训练阶段学习到的通用知识对域名进行向量化表征，在数据稀缺的场景下有效补充域名的深层语义特征。此外，本文还提取了基于外部数据库查询的关联属性特征，包括DNS解析模式、地理位置、域名注册信息查询协议（WHOIS）注册信息、安全套接层（secure sockets layer, SSL）协议证书属性、虚拟专用服务器（virtual private server, VPS）服务商IP归属等。最终，融合上述特征构建多层次特征表示，并采用K-means聚类算法对检测出的APT域名进行关联分析，实现向特定攻击组织的归因。实验结果表明，本文所提方法在APT域名检测任务中F1分数达97.47%，在APT域名归因任务中调整兰德指数（ARI）达86.85%、归一化互信息（NMI）达90.74%。在真实场景下的测试结果表明，相较于国内安全厂商的公开报告，本文方法平均提前112天标记出18个APT域名，实现了从被动响应向主动防御的转变，为应对APT威胁提供了新的思路。

本文的主要贡献如下。

1) 提出一种融合多层次特征的APT域名归因模型，能够从海量域名中检测与APT攻击相关的域名，并将这些APT域名归因到所属的APT组织。

2) 在真实场景中，相较于国内安全厂商的公开报告，所提方法能够平均提前112天标记出潜在的APT域名，提高了APT域名检测的早期预警能力。

1 相关工作

1.1 恶意域名检测技术

作为 APT 域名归因体系的基础环节, APT 域名检测效果直接影响后续的归因效果。现有主流的恶意域名检测方法大致分为两类, 即基于特征工程的检测方法和基于深度学习的检测方法。

基于特征工程的检测方法通过人工设计特征, 并将提取的特征输入分类器, 由分类器输出分类结果来实现检测。在恶意域名检测中, 主流特征提取方式可以分为以下两类。

1) 从 DNS 流量中提取特征。Bilge 等^[3]从 DNS 查询流量中提取了 15 类特征, 用于刻画域名及其查询方式的异常模式, 能够自动识别各类恶意活动中的未知恶意域名。MahdaviFar 等^[4]从 DNS 流量中提取了基于词法的特征、基于 DNS 统计的特征、基于第三方应用程序接口 (application programming interface, API) 背景信息的特征, 并采用 K 近邻 (k-nearest neighbors, KNN) 算法在两类数据集上取得了良好的检测效果。然而, 这类检测方法的有效性依赖于已发生的攻击行为和网络活动, 无法在 APT 域名投入使用之前就将其标记为潜在的恶意域名, 故难以满足 APT 检测的早期预警需求。

2) 从域名字符串中提取特征。此为本文在 APT 域名检测阶段所采用的方法。Kidmose 等^[5]将域名特征划分为 3 组 (一组为通用特征, 其余两组为词汇特征), 使用随机森林作为分类器进行检测, 结果显示, 词汇特征可以有效提升恶意域名的检测性能。Ghalati 等^[6]提出一种基于静态词汇特征的随机森林分类模型, 用于区分恶意统一资源定位符 (uniform resource locator, URL) 与良性 URL。该模型从 URL 中提取了 3 类特征, 即黑名单特征、词汇特征和基于主机的特征。

基于深度学习的检测方法无须人工设计特征, 只需将域名输入训练好的深度学习模型, 即可自动学习并生成特征表示。根据是否需要针对特定任务数据集进行训练, 可以将深度学习方法分为以下两类。

1) 基于从零开始训练的深度模型检测方法。此类方法采用的模型主要包括卷积神经网络 (convolutional neural network, CNN)、长短期记忆

网络 (long short-term memory network, LSTM)、图神经网络 (graph neural network, GNN) 等。Vinayakumar 等^[7]对多种深度学习方法进行了评估, 结果表明 LSTM 在恶意域名检测任务中表现更优。Chen 等^[8]提出了融合注意力机制的 LSTM 模型, 通过注意力机制聚焦域名中的关键字字符串, 从而提升域名表征能力。这类方法需要依赖大规模的训练数据来优化网络参数, 从而提升域名特征表征能力。然而, 由于 APT 域名数据稀缺, 这类方法难以直接用于 APT 域名检测场景。

2) 基于预训练模型的检测方法。预训练模型通过在大规模通用语料上进行预训练, 可以习得语言结构的通用表示。在 APT 域名检测中, 预训练模型能够利用学习到的词汇先验知识对域名进行向量化表征, 解决了数据稀缺难以训练深度学习模型的问题。Su 等^[9]基于双向编码器表示转换模型 (BERT) 的标记化方法及自注意力机制构建了一种面向恶意 URL 检测的深度学习框架, 在 3 类数据集上均取得较高的分类精度。相较于采用静态掩码预训练策略的 BERT 模型^[10], 本文所采用的 RoBERTa 模型通过动态掩码、更大规模的训练数据和去除下一句预测任务的预训练策略^[11], 能够更精准地提取深层语义特征。

1.2 APT 域名归因研究

在域名归因领域, 域名和 IP 的 WHOIS 信息曾被用于检测域名的所有者^[12]。然而, 受隐私保护服务和数据脱敏的限制, WHOIS 信息的实用性大打折扣, 仅凭此类信息难以有效支撑归因任务。Sebastián 等^[12]构建了融合被动 DNS、传输层安全 (transport layer security, TLS) 协议证书、网页内容等数据的域名归属分析框架, 在真实场景下取得了 94% 的 F1 分数。Xiang 等^[13]提出一种基于 IPv4 的威胁归因模型 IPAttributor, 通过融合真实入侵行为数据与异构商业情报, 结合动态加权社区发现算法, 实现了对同源攻击者的追踪。

2 本文方法

2.1 整体框架

本文提出了一种基于多层次特征融合的 APT 域名归因方法。考虑到 APT 域名在海量域名集合中占比极低, 若直接对所有域名进行归因分析, 需向外部数据库发送大量的查询请求, 导致

计算开销和响应时延增加。为此,本文设计了一种双阶段架构,将APT域名归因任务分解为检测和归因两个阶段。检测阶段的目标是从海量域名中快速筛选出潜在的APT域名;归因阶段的目标是对这些潜在APT域名进行关联分析,将其归因至所属的APT组织。这种双阶段的架构设计能够有效减轻归因分析的计算负担,提升系统响应速度。所提方法的整体框架如图1所示。

整体框架包括数据采集和预处理、APT域名检测、APT域名归因3个模块。

1) 数据采集和预处理模块:通过整合 Alexa 排名靠前的良性域名和开源威胁情报平台获取的已知APT域名构建初始数据集,并利用生成式大模型进行数据增强以解决样本不均衡问题。该模块输出包含正负样本的数据集。

2) APT域名检测模块:对数据集进行特征提取,得到包含文本、统计和语义3个层次的文本内容特征组。利用提取出的特征和样本标签,有监督地训练二分类模型,用于从海量域名列表中检测出APT域名。APT域名检测模块输出数据集的二分类结果,其中被分类为APT域名的集合将被输入APT域名归因模块,以便分析其归属。

3) APT域名归因模块:针对检测出的潜在APT域名,通过向外部数据库发送查询请求,获取域名的DNS解析模式、WHOIS注册、SSL证书等信息,并从中提取出域名的关联属性特征组。随后,通过融合文本内容特征组、关联属性特征组与RoBERTa提取的深层语义特征,采用K-means聚类算法将检测出的APT域名关联到特定攻击组织。

2.2 数据预处理

2.2.1 数据增强

针对APT域名检测任务中正样本稀缺的问题,传统的数据增强方法主要侧重于字符层面的扰动^[14-15],容易破坏APT域名中隐含的组织命名习惯或特定偏好,导致生成的样本偏离真实分布,形成无效噪声^[16]。为此,本文提出一种领域知识引导的生成式数据增强方法。该方法通过在提示词中融入专家先验知识、域名构造模板和特定的统计约束,诱导模型模拟攻击者的思维逻辑生成合成域名。这种方法不仅可以保证生成样本的文本多样性,还能够在深层语义层面保持与真实APT域名样本的一致性。用于生成域名数据的提示词如图2所示。

【角色设定】
你将扮演威胁情报分析师与APT域名IOC逆向工程专家的角色,具备以下能力:
掌握已知APT组织的域名生成模式
精通Unicode混淆、同形异义字替换等隐蔽技术
熟悉ICANN域名注册规则与WHOIS隐私保护机制

【任务解析】
核心目标:生成符合APT特征的拟真域名数据集
关键要求:
1. 生成策略多样性:需包含以下生成维度
字符分布:熵值控制在3.2到4.1之间
TLD分布:.info/.top/.xyz占比60%,伪装型.com/.net占比40%
结构模板:至少应用5种模式
◦日期嵌套型:2024.api0723.support
◦单词混淆型:adm1n-p4nel.xyz
◦哈希截断型:d4e8a1b.download
◦地理伪装型:sydney-backup.org
◦服务仿冒型:aws-cdn-update.pw
2. 约束条件:
严格遵循RFC 1035域名规范
不与已有公开威胁情报库重复
包含可解析的生成路径说明

【思维引导】
请按以下 workflow 生成:
① 特征提取:分析输入域名的
字符集分布(元音/辅音/数字比例)
分词模式(n -gram 概率)
TLD使用规律
② 模式扩展:应用以下增强技术
位置变异:如将"mail-"前缀改为"-server"后缀
长度抖动:基础域名长度 ± 3 字符

图2 数据增强提示词

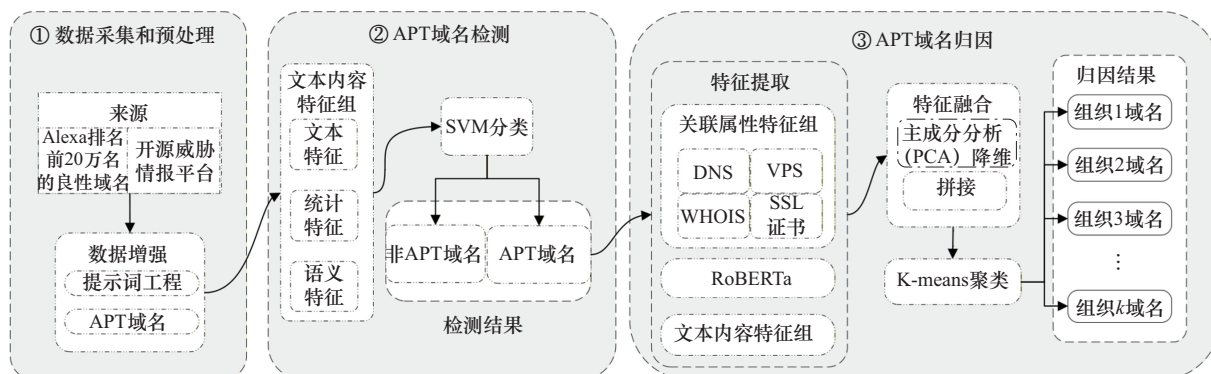


图1 所提方法的整体框架

2.2.2 域名分词

为提取域名的语义结构与局部模式, 本文设计了如下分词预处理流程。

分隔符分词: 依据 “.” “-” “—” 3 个分隔符对域名字符串进行切分。例如, “windows-verifyprotection.org” 将被分解为词元列表[‘windows’, ‘verifyprotection’, ‘org’], 保留了各部分的层级顺序, 称这一结果为初步分词结果。

语义分词: 部分域名中存在多个单词邻接出现形成一个字符串的情况 (如 verifyprotection), 本文将这种字符序列称为邻接词, 将针对邻接词的分词方法称为语义分词。经过语义分词后, 上述域名最终被分解为[‘windows’, ‘verify’, ‘protection’, ‘org’], 称这一步的结果为最终分词结果。

2.3 APT 域名检测

APT 域名检测模块旨在高效地从海量域名集合中检测出潜在的 APT 域名, 检测出的域名将用于后续的归因分析。该模块采用基于多特征融合的机器学习方法, 通过提取域名的文本特征、统计特征和语义特征, 构建高效的分类模型, 实现对 APT 域名的自动化检测。

2.3.1 文本内容特征提取

为实现 APT 域名检测, 本文设计了如表 1 所示的文本内容特征组, 用于对域名进行向量化表

征。文本内容特征的定义如下: 直接从域名字符串文本内容中提取、无须依赖外部数据或协议, 聚焦于域名的结构、字符分布和统计规律的特征。文本内容特征组包含文本特征、统计特征^[17]和语义特征。文本特征是指直接描述域名本身的结构、组成和显性文本属性的特征, 聚焦于可直接观察的文本信息。统计特征是指基于域名的分词结果, 通过数学计算或量化分析得出的特征, 反映域名的字符分布、组合规律、随机性。语义特征是指用于描述域名在语言意义层面的特征, 关注域名与特定关键词的关联性。这里的关键词与编辑距离特征中的关键词相同, 均为数据样本中恶意域名分词结果中出现频次最高的前 5 个单词。

下面对表 1 中的部分特征进行详细介绍。

(1) n -gram 熵

APT 组织常采用域名生成算法批量生成域名, 导致其域名在字符组合上呈现出一定的随机性^[18]。在 APT 域名检测任务中, n -gram 是指域名字符串中长度为 n 的连续子串。 n -gram 熵特征通过计算各 n -gram 在域名中的频率分布 $P(s_i)$, 并利用香农熵公式计算特征值, 从而捕捉 APT 域名的这种随机性特征。

$$H_n(D) = -\sum P(s_i) \log_2 P(s_i) \quad (1)$$

其中, $H_n(D)$ 表示域名 D 的 n -gram 熵特征, D 表

表 1 文本内容特征组

特征归类	特征	特征描述
文本特征	长度 (3)	域名、顶级域名和子域名的长度
	域名层级计数	域名中由点号分隔的层级数量
	www, com (2)	在非顶级域名部分是否出现 “www” “com”
	连字符数量	“-” “—” 出现的次数
	顶级域名	域名的顶级域名是否属于以下类型: “www” “com” “net” “gov” “edu”
	数字 (2)	域名中出现的数字数量、域名首个字符是否为数字
	元音与辅音 (4)	域名中元音字符的数量、辅音字符的数量、元音字符占域名总字符数的比重、辅音字符占域名总字符数的比重
统计特征	唯一字符 (2)	域名中出现次数唯一的字符数量、唯一字符占域名总字符数的比重
	初步分词结果 (5)	初步分词结果中的单词数、单词长度的平均值、单词长度的最大值、单词长度的最小值、单词长度的标准差
	邻接词数 (2)	域名中邻接词的数量、邻接词长度的平均值
	最终单词数	最终分词结果中的单词数
	n -gram 熵	域名中所有 n -gram 的香农熵
语义特征	n -gram 的 TF-IDF (5)	域名中所有 n -gram 的 TF-IDF 值的平均值、标准差、最大值、上四分位数、90% 分位数
	编辑距离 (5)	最终分词结果中各词与关键词的编辑距离最小值

注: 特征栏中括号内数字表示该特征包含的子特征数量。

示待检测的域名字符串, s_i 表示从域名 D 中提取出的第 i 个长度为 n 的连续子串, $P(s_i)$ 表示子串 s_i 在域名 D 的所有 n -gram 子串出现的概率, $\log_2$ 表示以 2 为底的对数。

(2) n -gram 的 TF-IDF

在 APT 域名检测任务中, n -gram 的词频-逆文档频率 (term frequency-inverse document frequency, TF-IDF) 特征的有效性源于如下机制: 首先, APT 组织在生成域名时往往表现出特定的词汇选择偏好, 这些偏好使得某些 n -gram 在恶意样本中呈现高频特性, 其词频 (term frequency, TF) 分量能够捕捉此类重复出现的特征; 其次, 逆文档频率 (inverse document frequency, IDF) 通过 $\log\left(\frac{N}{df_t + 1}\right)$ 的计算方式, 显著提升那些在恶意域名中频繁出现但在正常域名中罕见的 n -gram 的权重, 从而放大具有组织特异性的特征。TF-IDF 的数学表达式如式(2)所示。

$$N_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t + 1}\right) \quad (2)$$

其中, $tf_{t,d}$ 为 n -gram t 在域名 d 中出现的频率, N 为数据集中域名总数, df_t 为包含 t 的域名数量。

2.3.2 域名分类

基于提取出的多层次特征, 本文选用 SVM 作为域名分类器, 其设计考量如下: 首先, SVM 通过最大化分类间隔优化决策边界^[19], 能够有效处理恶意样本与良性样本在高维特征空间中的非线性交织问题, 结合径向基核函数的非线性映射能力, 能够捕获特征间的隐含关联; 其次, SVM 基于结构风险最小化^[20]的理论框架, 在有限标注样本下具有较高的泛化性能, 可以降低 APT 域名样本分布稀疏带来的过拟合风险。相较于神经网络等复杂模型, SVM 的分类结果具有更强的可解释性。此外, SVM 训练效率更高, 支持快速模型迭代和响应, 能够满足实际场景中威胁检测模型的动态更新和实时检测需求。

2.4 APT 域名归因

在完成 APT 域名检测后, 接下来将聚焦 APT 域名归因。尽管 APT 域名检测模块能够有效区分 APT 域名与非 APT 域名, 但单纯依靠二分类结果无法满足威胁情报的战术级溯源与响应

需求。为此, 本文提出多层次特征融合的归因框架, 通过融合 DNS 解析模式信息、WHOIS 注册信息等外部关联属性特征, 以及 RoBERTa 提取的深层语义表征, 构建具备组织特异性的聚类模型。APT 域名归因模块以 APT 检测模块的输出作为输入, 通过层次化特征分析与无监督学习, 实现从“判断是否恶意”到“识别攻击组织”的跃迁。

2.4.1 关联属性特征提取

为实现 APT 域名归因, 本文设计了如表 2 所示的关联属性特征组, 依托多源外部数据构建 APT 域名的关联属性特征。关联属性特征组的定义如下: 依赖外部数据源或网络协议获取的关联信息, 或从这些关联信息中衍生的其他信息, 能够反映域名的注册和解析、服务器属性等背景信息的特征。本文将关联属性特征组中的特征划分为 5 类。在构建关联属性特征组时, 除了域名自身的属性特征外, 还考虑了域名与 IP 地址之间的关联关系, 并将其抽象为地理位置与 VPS 两类特征, 前者从空间维度刻画了 IP 地址的分布特性, 后者则从基础设施维度揭示了 IP 地址的服务环境。

表 2 关联属性特征组

特征归类	特征	解释
DNS	IP 地址数量	根据域名解析出的 IP 地址数量
	TTL 值	DNS 记录在缓存中的存活时间
地理位置	国家数量	解析出的 IP 地址对应的国家数量
	国家	出现频次最高的国家
WHOIS	域名存活时间	域名到期日期-创建日期
	域名活跃时间	域名更新日期-创建日期
SSL 证书	SSL 证书颁发服务类型	证书是否为免费签发
	SSL 证书到期时间	证书签发时设定的有效截止时间
VPS	VPS 服务商 IP 地址归属	解析的 IP 地址是否属于已知 VPS 服务商的 IP 地址段

在关联属性特征组中, DNS 是指通过 DNS 查询服务获取域名对应的 IP 地址及每条 DNS 记录的 TTL 值。WHOIS 是指通过 WHOIS 接口查询域名的创建日期、到期日期和更新日期, 并基于这 3 个日期计算出域名的存活时间和活跃时间。VPS 是指应用虚拟化技术在物理服务器上划分出的独立虚拟服务器, 用户可以远程部署和管理自

定义服务。这类特征可以辅助追踪 APT 攻击者的资源调度模式,提升域名归因效果。以下为表 2 中部分特征的详细介绍。

(1) TTL 值

域名的存活时间 (time to live, TTL) 值是指域名 DNS 解析记录在缓存中保留的有效时长,攻击者通常会设置极小的 TTL 值以实现命令与控制 (command and control, C2) 服务器的快速切换^[21]。

(2) SSL 证书颁发服务类型

出于降低攻击成本、快速部署攻击基础设施的目的,APT 组织在搭建 C2 通信域名时,通常倾向于使用免费 SSL 证书。基于这一行为特征,可以通过解析域名数字证书的元数据属性,判断该证书是否由免费 SSL 证书分发页机构签发。

(3) VPS 服务商 IP 地址归属

张慧菲等^[22]指出,APT 组织在部署 C2 基础设施时,倾向于使用 VPS 服务商提供的 IP 地址。相较于企业自建数据中心,VPS 服务具备匿名性强、部署灵活、覆盖范围广的优势,这是 APT 组织偏好使用 VPS 服务的核心原因。针对这一特征,可以采用动态更新的 VPS IP 库,通过对比域名解析出的 IP 地址和 VPS IP 库中的 IP 地址来判断域名是否使用 VPS 服务商提供的 IP 地址资源。

2.4.2 RoBERTa 特征提取

在 APT 域名归因模块中,除了提取域名的关联属性特征外,本文还利用预训练的 RoBERTa 模型提取 APT 域名的语义特征。考虑到 APT 域名检测阶段需要从海量域名中筛选出少量 APT 域名,若在此阶段引入预训练模型将会增加模型的计算开销和时延,故 RoBERTa 模型仅应用于数据量更小的第二阶段归因任务。具体而言,提取 RoBERTa 模型最后一层隐藏层中起始字符 <s> 对应的向量,将其作为域名的深层语义特征。

2.4.3 特征融合

为实现 APT 域名多层次特征的高效融合,本文采用直接特征拼接方法融合 4 个层次的特征,具体流程如下:首先,采用主成分分析 (principal component analysis, PCA) 将 RoBERTa 提取的深层语义特征投影至 128 维低维空间,并

对关联属性特征进行分位数归一化与离群值截断处理,消除不同特征间的量纲差异,同时过滤噪声干扰;然后,将文本内容特征、降维后的 RoBERTa 深层语义特征与标准化后的关联属性特征沿特征维度直接拼接,生成域名的联合表征向量。该方法在最大程度地保留原始特征的同时,具备较高的计算效率,为轻量化实时检测场景提供了兼顾检测效率与归因精度的技术方案。

2.4.4 APT 域名聚类

本节基于无监督聚类方法对检测出的潜在 APT 域名进行关联挖掘与归因分析。在算法方面,本文选用 K-means 聚类算法,其核心在于通过迭代优化质心位置,将高维特征空间中的域名样本划分为若干簇内紧密度高、簇间差异显著的子集。相较于密度聚类或层次聚类,K-means 算法的优势体现在两个方面:其一,针对 APT 域名特征工程提取的多层次结构化特征,K-means 采用的欧氏距离度量方式,能够有效捕捉特征维度间的全局相似性;其二,算法支持通过预定义簇数约束聚类规模,适用于 APT 攻击中基础设施分属有限组织的假设。

3 实验

为验证各模块的理论性能,实验阶段对检测与归因任务分别进行分析。

3.1 数据集

本文使用的 APT 域名数据集通过系统化采集并整合多源开源威胁情报构建而成。数据主要来源于标准化威胁情报平台和专业安全厂商发布的攻击报告。为平衡数据集的类别分布,本文从 Alexa 排名前 20 万名的良性域名中随机下采样,按 1:1 的比例将良性样本作为负样本纳入数据集。最终构建的数据集包含 1 799 个 APT 域名样本和等量的良性域名样本,覆盖 25 个 APT 组织。为保障模型评估结果的可靠性,本文采用分层抽样策略,将数据集按 7:2:1 的比例划分为训练集、验证集和测试集。该划分方式能够确保各 APT 组织的样本在 3 个子集中的分布比例与原数据集保持一致。

3.2 评价指标

本文针对 APT 域名检测与 APT 域名归因两个阶段的任务特性,构建了差异化的评价指标体系,

以实现对模型性能的量化评估。对于APT域名检测任务，基于其分类问题的本质，本文采用分类任务中常用的精确率、召回率和F1分数作为评价指标。对于APT域名归因任务，本文选取被广泛用于评估聚类任务的调整兰德指数（adjusted rand index, ARI）和标准化互信息（normalized mutual information, NMI）作为评价指标。ARI^[23]用于量化聚类结果与真实APT标签之间的全局一致性，评估簇内标签分布的匹配程度，其取值范围为[-1,1]。NMI^[24]用于衡量两个数据分布的吻合程度，其取值范围为[0,1]。这两个指标的值越大，表明聚类结果与原始数据的划分越吻合。

3.3 实验结果

3.3.1 对比实验

为验证本文所提方法的有效性，研究人员设计了对比实验。本文选取的基线方法均在恶意域名检测领域具有影响力或直接相关性。文献[25]提出了一种基于特征工程和机器学习的恶意域名检测方法 MaldomDetector。MaldomDetector 虽主要面向域名生成算法（domain generation algorithm, DGA）域名检测，但其特征工程思路与本文方法高度相关。因此，研究人员在本文构建的数据集上对其进行了测试。此外，文献[26]所提大语言模型（large language model, LLM）方法虽应用于钓鱼域名检测，但其提示词框架可以直接迁移至APT域名检测任务。基于该思路，本文设计了图3所示的提示词，并调用 DeepSeek-R1 模型，对比了LLM和本文方法在APT域名检测任务中的性能。上述两种方法与本文方法检测性能对比见表3。从表3可以看出，无论是与同类型方法对比^[25]，还是与LLM方法^[26]对比，本文方法在精确率、召回率与F1分数3项评价指标上均取得最优的分值。

你是一名网络安全分析师，专精于恶意域名检测。
你的任务是分析以下域名，并判定其为恶意域名还是合法域名，
最终结论仅输出一个词：恶意域名 或 合法域名。
请按照以下步骤谨慎推理：
- 检查域名结构（如长度、字符分布、子域名层级）
- 域名拼写错误（如"paypal.com"代替"paypal.com"）
- 存在高危关键词（如"login" "update"等）
- 与知名域名对比视觉相似度
- 查询域名的注册时间、注册商等信息的完整性
- 评估域名所指向的IP、所属自治系统编号（ASN）
域名：{domain}

图3 APT域名检测提示词

表3 检测性能对比

检测方法	精确率	召回率	F1 分数
文献[25]方法	91.14%	85.71%	88.34%
文献[26]方法	94.26%	93.81%	94.03%
本文方法	98.72%	96.25%	97.47%

为验证本文针对APT域名检测任务设计的模型架构的有效性，本文选取多种主流机器学习与深度学习模型进行对比实验。实验涉及的模型包括随机森林（random forest, RF）、KNN、朴素贝叶斯（naive Bayes, NB）、决策树（decision tree, DT）、LSTM、BiLSTM，以及系列增强算法模型。各类模型的APT域名检测对比实验结果见表4。从表4可以看出，SVM模型在APT域名检测任务中以97.47%的F1分数优于其他模型，验证了本文设计的文本内容特征的有效性和SVM模型的非线性分类优势。值得注意的是，机器学习模型的性能普遍优于LSTM和BiLSTM模型。这一结果表明，结构简单、参数量少的机器学习模型能够更好地适应训练数据稀缺的场景。

表4 APT域名检测对比实验结果

模型	精确率	召回率	F1 分数
RF	98.68%	93.75%	96.15%
KNN	97.33%	91.25%	94.19%
NB	86.42%	87.50%	86.96%
DT	93.59%	91.25%	92.41%
AdaBoost	97.37%	92.50%	94.87%
GBM	98.70%	95.00%	96.82%
XGBoost	98.70%	95.00%	96.82%
LightGBM	97.44%	95.00%	96.20%
CatBoost	97.40%	93.75%	95.54%
LSTM	91.39%	91.15%	91.11%
BiLSTM	91.73%	91.72%	91.72%
SVM	98.72%	96.25%	97.47%

为提高本文所提方法在APT域名检测任务中的可解释性，研究人员采用沙普利加性解释（Shapley additive explanations, SHAP）方法对APT域名检测任务中的特征进行了重要性分析，按特征重要性从高到低排序，输出了排名前10名的

特征, 结果如图4所示。从图4可以看出, 重要性排名最高的特征对模型决策的贡献度仅接近7%, 这说明该模型的决策机制并非依赖单一特征的独立作用, 而是依靠多种特征的协同交互效应。

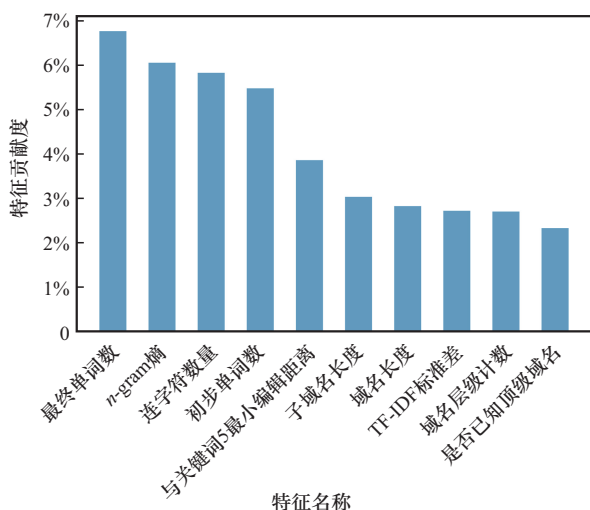


图4 APT域名检测特征重要性排序

在 APT 域名归因任务中, 为验证本文提出的多层次特征融合框架及所选 K-means 聚类算法的有效性, 研究人员选取了多种主流聚类算法进行对比实验, 包括层次聚类 (hierarchical clustering, HC)、基于密度的带噪声应用空间聚类 (density-based spatial clustering of applications with noise, DBSCAN) 算法、高斯混合模型 (Gaussian mixture model, GMM)、狄利克雷过程混合模型 (Dirichlet process mixture model, DPMM)。各类聚类算法的 APT 域名归因对比实验结果见表 5。从表 5 可以看出, K-means 聚类算法以 86.85% 的 ARI 值和 90.74% 的 NMI 值, 显著优于其他对比聚类算法, 这一结果充分验证了本文方法的有效性。该结果证明, 本文针对 APT 域名提取的 3 类特征能够准确刻画不同 APT 组织生成域名的偏好, 且这些特征分布能被 K-means 聚类算法有效利用。相较之下, DBSCAN 算法表现最差, 其 ARI 值仅为 22.53%。分析其原因, 主要是 DBSCAN 算法具有显著的密度敏感特性^[27], 而本文构建的 APT 域名特征空间存在密度不均衡问题, 该算法无法适应特征空间的密度变化, 导致在区分不同 APT 组织的域名时出现严重的簇重叠现象。

表5 APT域名归因对比实验结果

算法	ARI	NMI
HC	83.21%	88.93%
DBSCAN	22.53%	53.30%
GMM	80.54%	90.39%
DPMM	86.66%	90.61%
K-means	86.85%	90.74%

3.3.2 真实场景

为探究本文方法在真实场景中的提前预警能力, 研究人员设计了一项主动检测实验。具体来说, 首先使用本文提出的方法持续对全球每日新注册域名进行恶意性分析, 筛选并标记潜在的 APT 域名。随后从国内权威安全厂商发布的 APT 攻击报告中提取公开披露的 APT 域名共计 26 个, 对其进行去重与精准匹配验证后发现, 其中有 18 个已被本文方法前期标记为潜在 APT 域名。

上述 18 个提前发现的域名部分示例见表 6, 表中明确标注了 3 个关键信息: 标记日期 (本文方法将该域名标记为潜在 APT 域名的日期)、披露日期 (该域名被权威安全厂商公开披露的日期)、提前周期 (域名从被标记到被公开披露的时间间隔)。经统计, 18 个域名的平均提前预警周期达到 112 天。该实验验证了本文方法在对抗 APT 的真实网络场景中所具备的早期预警能力。

表6 提前发现的APT域名部分示例

域名	标记日期	披露日期	提前周期/天
ottawadesignlab.com	2024-09-02	2024-10-21	49
theoysservices.info	2024-09-21	2025-02-28	160
bijoyshare.buzz	2024-11-02	2025-02-28	118
sharetobijoy.buzz	2024-12-22	2025-02-28	68
totalservices.info	2024-12-27	2025-02-28	63
jlu-edu.org	2025-06-15	2025-07-11	26

4 结束语

为提高 APT 域名检测方法的早期预警能力, 应对 APT 域名数据稀缺和类别不均衡的问题, 本文提出一种基于多层次特征融合的 APT 域名归因方法。该方法融合了文本、统计、语义和关联属性 4 个层次的特征, 并将归因任务分解为

APT域名检测和APT域名归因两个阶段。在检测阶段，通过提取域名的文本特征、统计特征及语义特征，并采用SVM作为分类器，实现了从海量域名中高效筛选潜在的APT域名。在归因阶段，通过融合文本内容特征、由预训练模型RoBERTa提取的深层语义特征，以及基于外部数据库查询获取的DNS解析模式、WHOIS注册信息、SSL证书属性、VPS服务商IP地址归属等相关属性特征，构建多层次的特征表示，并利用K-means聚类算法实现将恶意域名归因到攻击组织。实验结果表明，本文方法不仅在APT域名检测和归因任务中取得了较高的精度，还在真实场景中展现出平均提前112天的早期预警能力，验证了本文方法的有效性，为应对APT攻击提供了新的思路。

本文所提方法虽然在APT域名检测与归因方面具有良好的性能，但仍存在一定的局限性。首先，当前APT域名检测模型依赖于静态的特征工程，这些基于预定义规则提取的特征在面对APT组织快速演变的战术、技术与程序（tactics techniques procedure, TTP）时，缺乏自适应性和鲁棒性。攻击者一旦改变域名生成策略或基础设施使用模式，模型性能可能出现下降。其次，本文采用的通用RoBERTa模型在捕捉APT组织独特命名习惯的语义方面可能存在局限性。这主要是因为公开的APT域名数据规模极为有限，若在数据样本少的场景下对RoBERTa进行领域微调，极易导致模型过拟合，反而削弱其从预训练中获得泛化能力。

针对上述挑战，未来可以从4个方面开展深入研究：①探索在线学习与增量学习框架，使模型能够依据新到达的数据动态调整决策边界，从而及时响应APT组织TTP的演变；②增加对抗训练，通过在模型训练阶段注入扰动样本来提升模型的鲁棒性；③致力于构建更大规模的APT域名语料库，为预训练模型进行领域微调提供数据基础，进而更精准地捕捉APT组织的语义特征；④探索基于图结构的归因建模方法，利用图神经网络挖掘更深层次的归因线索，以期构建一个更全面的APT组织归因系统。

参考文献：

- [1] 王晨阳, 彭国军, 杨秀璋, 等. 恶意代码动态分析对抗与反对抗技术综述[J]. 网络与信息安全学报, 2025, 11(6): 1-16.
- [2] 张维维, 龚俭, 刘茜, 等. 基于词素特征的轻量级域名检测算法[J]. 软件学报, 2016, 27(9): 2348-2364.
- [3] Bilge L, Kirda E, Kruegel C, et al. Exposure: finding malicious domains using passive DNS analysis[C]//Proceedings of the 18th Annual Network and Distributed System Security Symposium. 2011: 195-211.
- [4] Mahdavi S, Maleki N, Lashkari A H, et al. Classifying malicious domains using DNS traffic analysis[C]//Proceedings of the 2021 IEEE International Conference on Dependable, Autonomic and Secure Computing. Piscataway: IEEE Press, 2021: 60-67.
- [5] Kidmose E, Stevanovic M, Pedersen J M. Detection of malicious domains through lexical analysis[C]//Proceedings of the 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). Piscataway: IEEE Press, 2018: 1-5.
- [6] Ghalati N F, Ghalaty N F, Barata J. Towards the detection of malicious URL and domain names using machine learning[C]//Proceedings of the Doctoral Conference on Computing, Electrical and Industrial Systems. Berlin: Springer, 2020: 109-117.
- [7] Vinayakumar R, Soman K P, Poornachandran P. Detecting malicious domain names using deep learning approaches at scale[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(3): 1355-1367.
- [8] Chen Y, Zhang S, Liu J, et al. Towards a deep learning approach for detecting malicious domains[C]//Proceedings of the 2018 IEEE International Conference on Smart Cloud (SmartCloud). Piscataway: IEEE Press, 2018: 190-195.
- [9] Su M Y, Su K L. BERT-based approaches to identifying malicious URLs[J]. Sensors, 2023, 23(20): 8499.
- [10] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Kerrville: Association for Computational Linguistics, 2019: 4171-4186.
- [11] Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach[PP]. arXiv[2025-10-14]. arXiv: arXiv. 1907. 11692.
- [12] Sebastián S, Diugan R G, Caballero J, et al. Domain and website attribution beyond WHOIS[C]//Proceedings of the Annual Computer Security Applications Conference. New York: ACM, 2023: 124-137.
- [13] Xiang X Y, Liu H, Zeng L Y, et al. IPAttributor: cyber attacker attri-

- bution with threat intelligence-enriched intrusion data[J]. Mathematics, 2024, 12(9): 1364.
- [14] Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv: 1901.11196, 2019.
- [15] Xie Q, Dai Z, Hovy E, et al. Unsupervised data augmentation for consistency training[J]. Advances in Neural Information Processing Systems, 2020, 33: 6256-6268.
- [16] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation[J]. arXiv:1711.02173, 2017.
- [17] Sahingoz O K, Buber E, Demir O, et al. Machine learning based phishing detection from URLs[J]. Expert Systems with Applications, 2019, 117: 345-357.
- [18] 宋文纳, 彭国军, 傅建明, 等. 恶意代码演化与溯源技术研究[J]. 软件学报, 2019, 30(8): 2229-2267.
- Song W N, Peng G J, Fu J M, et al. Research on malicious code evolution and traceability technology[J]. Journal of Software, 2019, 30(8): 2229-2267.
- [19] Tharwat A. Parameter investigation of support vector machine classifier with kernel functions[J]. Knowledge and Information Systems, 2019, 61(3): 1269-1302.
- [20] Li H, Lü Z, Yue Z. Support vector machine for structural reliability analysis[J]. Applied Mathematics and Mechanics, 2006, 27(10): 1295-1303.
- [21] Afek Y, Litmanovich A. Decoupling DNS update timing from TTL values[C]//Proceedings of the NOMS 2025-2025 IEEE Network Operations and Management Symposium. Piscataway: IEEE Press, 2025.
- [22] 张慧菲, 杨秀璋, 彭国军. APEA: 一种恶意 URL 新型识别方法[J]. 网络与信息安全学报, 2025, 11(6): 77-91.
- Zhang H F, Yang X Z, Peng G J. APEA: a novel identification method for malicious URLs[J]. Chinese Journal of Network and Information Security, 2025, 11(6): 77-91.
- [23] Hubert L, Arabie P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.
- [24] Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3(3): 583-617.
- [25] Almashhadani A O, Kaiiali M, Carlin D, et al. MaldomDetector: a system for detecting algorithmically generated domain names with machine learning[J]. Computers & Security, 2020, 93: 101787.
- [26] La O R L, Catania C A, Parlanti T. LLMs for domain generation algorithm detection[PP]. arXiv[2025-10-14]. arXiv: arXiv. 2411. 03307.
- [27] 张唯一, 罗启瑞, 张琳. 基于动态聚类和信誉评估的抗攻击联邦框架[J]. 网络与信息安全学报, 2025, 11(6): 120-130.
- Zhang W Y, Luo Q R, Zhang L. Anti-attack federated framework based on dynamic clustering and reputation evaluation[J]. Chinese Journal of Network and Information Security, 2025, 11(6): 120-130.

[作者简介]



马利昭 (2001-), 男, 四川大学硕士生, 主要研究方向为威胁检测、机器学习和自然语言处理。



黄诚 (1987-), 男, 四川大学教授, 主要研究方向为网络空间安全、攻击检测、威胁溯源、数据挖掘、社交网络、机器学习和自然语言处理。



黎明 (1977-), 男, 中国电子科技集团航空电子有限公司创新中心主任, 主要研究方向为航空计算技术、航空通信安全和人工智能。



牟浩天 (2002-), 男, 四川大学硕士生, 主要研究方向为机器学习、高级持续性威胁检测。



彭佳仁 (2000-), 男, 四川大学博士生, 主要研究方向为网络威胁情报与大语言模型。



谢启亮 (1998-), 男, 四川大学博士生, 主要研究方向为流量检测和区块链。