

DOI: 10.3969/j.issn.1671-0673. 2026. 03. 010

一种时序驱动的 Web 攻击行为同源分析方法

高耀军, 张 震, 王文博, 王亚文

(信息工程大学, 河南 郑州 450001)

摘要: 针对现有网络攻击同源分析方法在攻击行为相似性度量上存在不足、且在多密度样本与高噪声场景下同源分析准确度低的问题, 提出一种时序驱动的 Web 攻击行为同源分析方法。首先, 提取 3 类 20 种 Web 攻击警报特征, 构建攻击序列; 其次, 提出一种攻击序列相似度算法, 计算序列间相似度并构建距离矩阵; 最后, 基于距离矩阵设计聚类算法, 对攻击序列进行聚类, 进而实现攻击同源分析。相较于仅依靠统计特征分析的传统方法, 该方法将攻击行为的时序关联性纳入同源分析范畴, 有效提升了 Web 攻击同源分析的准确度。实验结果显示, 在噪声环境中, 该方法的关键指标调整兰德指数 (ARI) 达 94.9%, 验证了同源分析效果优于其他方法, 同时在较宽参数范围内能够保持性能稳定。

关键词: 高级持续性威胁; Web 攻击; 同源分析; 动态时间规整; 聚类

中图分类号: TP393.08

文献标识码: A

文章编号: 1671-0673(2026)03-0316-07

A Time Series Driven Homology Analysis Method for Web Attack Behaviors

GAO Yaojun, ZHANG Zhen, WANG Wenbo, WANG Yawen

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To address the problems that existing network attack homology analysis methods suffer from insufficient measurement of attack behavior similarity and low accuracy in multi-density samples and high-noise scenarios, a time series driven homology analysis method for Web attack behaviors is proposed. Firstly, 3 categories of 20 Web attack alert features are extracted to construct attack series. Then, a similarity algorithm for attack series is proposed to calculate the similarity between sequences and construct a distance matrix. Finally, a clustering algorithm is designed based on the distance matrix to cluster attack sequences, thereby realizing attack homology analysis. Different from traditional methods relying only on statistical feature analysis, the proposed method incorporates the temporal correlation of attack behaviors into homology analysis and effectively improves the accuracy of Web attack homology analysis. Experimental results show that the key indicator ARI of the method reaches 94.9% in noisy environments, verifying that its homology analysis performance outperforms other methods. Meanwhile, the method can maintain stable performance within a wide range of parameters.

Key words: advanced persistent threat; Web attack; homology analysis; dynamic time warping; clustering

Web 应用是支撑各网络业务、数字服务的核心基础, 高级持续性威胁 (Advanced Persistent Threat, APT) 攻击组织常使用开源情报及网络测绘工具,

主动识别特定 Web 资产以策划后续攻击行动^[1]。同时 APT 组织发起 Web 攻击的过程中, 会通过各类技术手段完成自身身份的隐匿伪装, 如依托僵

收稿日期: 2026-02-05; 修回日期: 2026-03-18

基金项目: 国家科技重大专项 (2025ZD1501300)

作者简介: 高耀军 (1990—), 男, 硕士生, 主要研究方向为网络空间安全。E-mail: desmond080@163.com

尸网络干扰混淆前端流量、利用Tor匿名网络搭建隐蔽通信链路、结合跳板攻击与代理实现攻击路径隐藏等^[2-3],这也让Web攻击的溯源和防御工作面临诸多实际难题。基于此,开展面向APT场景的Web攻击同源分析研究,以实现Web攻击行为的精准溯源与早期预警,已成为迫切需求。

在网络攻击的同源分析领域,当前研究的主流思路有基于恶意样本的同源分析方法、基于流量的同源分析方法和基于网络攻击行为的同源分析方法。其中基于恶意样本的同源分析方法通过对比不同恶意软件的代码、结构、运行特征等,判断它们是否来自同一源头或具有相似的开发背景,具体可分为静态同源分析和动态同源分析。静态同源分析主要是对恶意样本代码进行分析,通过提取代码结构特征信息,找到代码间的相似性^[4]。动态同源分析通过在沙箱等环境中实时监控恶意样本的运行行为,捕捉网络通信、文件操作、应用程序接口(Application Program Interface, API)调用的模式、系统调用参数等动态特征,揭示恶意样本隐蔽的恶意行为,识别具有共同来源或相似攻击特征的恶意样本^[5]。

基于流量的同源分析方法通过多种方法对网络流量进行监测和分析,利用数据包标记、日志记录、协议特征提取等方式追踪攻击路径,识别攻击源,从而发现具有同源性的恶意流量。王建华等^[6]提出一种基于工控系统功能码特征的同源攻击分析方法,使用工控协议功能码的粗粒度统计特征和细粒度序列特征来量化攻击行为,识别攻击行为相似的组织。王跃达等^[7]提出了一种基于高速网络的WebShell综合检测溯源技术,通过特征码匹配的方式实现对同族WebShell的高效检测。

基于网络攻击行为的同源分析通过分析访问模式、攻击手段等,判断不同网络行为是否由同一攻击者或攻击组织发起。由于攻击者通常存在较为固定的行为模式,难以被刻意改变,该方法更加广泛地应用于网络攻击源性分析研究。刘浩^[8]提出了一种基于多源威胁情报的网络攻击源性分析方法,通过融合日志数据和外部威胁情报,构建全面的IP画像并完成同源分析。Kiseleva等^[9]利用聚类算法,通过密度特征将离散网络流量自动分组,实现攻击簇识别与同源关联。白波等^[10]提出一种基于网络行为的攻击同源分析方法,通过细致分析攻击者在不同阶段的行为特征,并提取相关行为特征矩阵,利用社区发现算法将不同的攻

击行为划分到不同的攻击源社区,实现攻击行为源性分析。林晨等^[11]提出了一种基于N-gram序列比对与ATT&CK映射的关联分析方法,通过将设备日志转化为有向图并结合网络流量五元组信息,实现了对跨设备攻击链条的精准溯源与可视化复现。Burton^[12]提取多维行为特征构建静态特征向量,利用聚类算法识别不同生成机制的稳定攻击模式。吴朝雄等^[13]提出了一种基于电子指纹的同源分析技术,通过构建原生网络攻击树,利用同源推理与剪枝算法,实现了从碎片化日志到完整攻击链的逻辑还原。王婷等^[14]基于图注意力机制与威胁情报知识图谱,实现攻击报告所属组织的精准同源分析。

然而,以上方法还存在一些缺陷。在基于恶意样本的同源性分析方法中,攻击者能够通过借用或更换恶意工具来逃避检测,导致同源分析的精准度较低;基于流量的同源分析方法只适用于特定场景,具有应用的局限性;现有基于网络攻击行为的同源分析方法虽引入时间维度,但多依赖统计特征对比,难以有效捕捉网络攻击的动态行为关联特性,在攻击相似性度量方面存在局限,且难以在高噪声环境下有效分离攻击信号与背景噪声,对多密度的攻击样本适应性较差,导致同源分析准确率较低。

鉴于上述问题,本文提出一种时序驱动的Web攻击行为同源分析方法。该方法通过对攻击序列进行分析比较,捕捉攻击行为的动态变化和时序规律,并在此基础上进行后续的同源分析。首先,提取多维度攻击特征,并将其整合形成攻击序列;其次,提出了基于动态时间规整(Dynamic Time Warping, DTW)的攻击序列相似度算法,精准地衡量不同攻击序列之间的相似程度;最后,提出基于层次密度聚类算法(Hierarchical Density-Based Spatial Clustering of Applications with Noise, HDBSCAN)的攻击序列聚类算法,完成对攻击序列的同源分析,以及攻击数据与噪声数据的分离。

1 本文方法架构

本文方法流程如图1所示。首先,对攻击警报数据进行预处理;其次,通过基于DTW的攻击序列相似度算法计算攻击序列的相似距离矩阵;最后,利用基于HDBSCAN的攻击序列聚类算法完成攻击序列聚类与噪声分离,实现同源攻击识别。

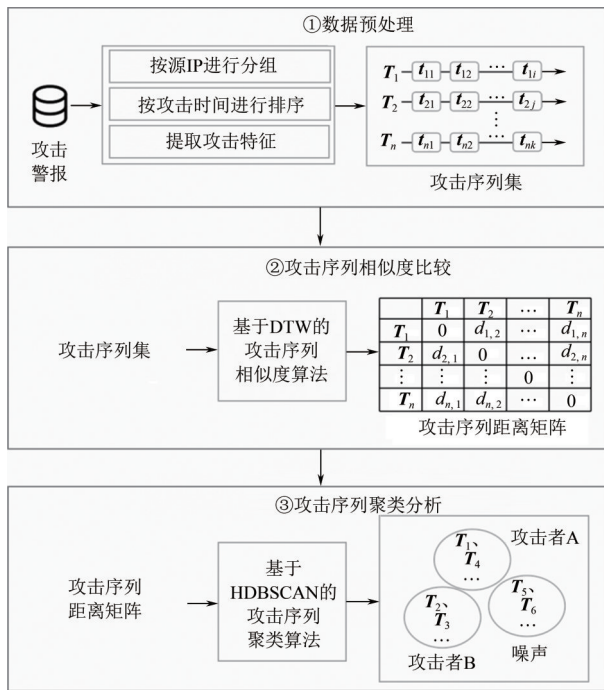


图1 方法流程示意图

1.1 数据预处理

本文方法主要针对IDS(Intrusion Detection System)、WAF(Web Application Firewall)等设备攻击警报的通用警报内容进行分析。首先,将所有警报按源IP地址进行分组,组内的警报按告警时间进行排序,并从原始攻击警报中提取攻击特征。此时,一组警报即为一条攻击序列,以此多条攻击序列形成攻击序列集。根据实践经验,以及参照以往相关研究^[7-10],本文提取了时空、Web指纹、攻击行为3类共20种特征。特征具体内容如下。

1)时空特征。该类特征共包括4项内容:告警时间,即攻击告警时间,单位为h;告警时间间隔,即相邻告警的时间间隔,单位为s;源IP地址,即告警信息中的源IP地址,考虑到攻击者很有可能使用同一网段中的其他地址,本文将同一C类网段地址的源攻击IP视为同一攻击者;源IP归属地,即源IP地址对应的国家或地区。

2)Web指纹特征。选择8种Web指纹信息:Hyper Text Transfer Protocol(HTTP) Method、Host、Uniform Resource Locator(URL)、User Agent(UA)、Accept Language、Cookie、X-Forwarded-For(XFF)和Post Body。这些指纹信息都是常见的HTTP协议内容。一方面在实践中这些HTTP字段信息较为常见且容易提取,更重要的是在SQL注入(Structured Query Language Injection)、远程命令执行等Web攻击中,其攻击载荷可能存在于以上这些指纹字段中(HTTP Method除外),提取这些信息可以更

好地对攻击载荷进行分析研究。

3)攻击行为特征。该类特征共提取8项内容,用来表征攻击的细节和个性化特征,包括:攻击类型,即攻击者所使用的Web攻击类型,涵盖漏洞扫描、SQL注入攻击、XSS(Cross-Site Scripting)攻击、路径穿越、远程命令执行等常见的Web服务攻击类型;攻击阶段,即在攻击链中当前攻击所处的阶段(如信息收集、漏洞利用等);攻击载荷,即指原始攻击信息中执行实际破坏或未经授权操作的代码部分;攻击载荷位置,即攻击载荷所处的HTTP协议字段;检测绕过方式,即攻击者在攻击载荷中可能使用的4种检测绕过方式,包括URL编码绕过、Unicode编码绕过、Base64编码绕过和HTML注释绕过。

对数据集预处理后,得到了攻击序列集 $T = \{T_1, T_2, \dots, T_n\}$, n 为序列的数量,对应总的攻击IP数量。 T_i 表示第 i 条攻击序列($i \in \{1, 2, \dots, n\}$),每条序列对应同一攻击IP产生的所有攻击告警按时间排序后的特征向量集合 $T_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$, m 为该序列告警的数量,序列中的项 t 是一个包含20个攻击特征的向量,表征一次攻击行为。

1.2 攻击序列相似度比较

为比较特征的相似性,定义特征相似性比较方法集合 $S = \{s_1, s_2, \dots, s_{20}\}$, 包含针对这20个特征各自的相似性比较方法,其中 s_i 是第 i 个特征的相似性比较方法,相似性度量方法包括以下5类。

1)告警时间相似度。计算方式如下:

$$R_{\text{Sim-a}} = 1 - \frac{1}{12} \left(\min(|H_1 - H_2|, 24 - |H_1 - H_2|) \right) \quad (1)$$

式中: H_1 、 H_2 为需要进行比较的告警时间; $\min(\cdot)$ 表示取最小值。

2)IP地址相似度。若两个IP地址A、B段不完全相同,相似度为0;若两个IP地址A、B段相同,相似度 $R_{\text{Sim-i}}$ 计算方式如下:

$$R_{\text{Sim-i}} = 1 - \frac{1}{255} \left(|N_{\text{IP-C1}} - N_{\text{IP-C2}}| \right) \quad (2)$$

式中, $N_{\text{IP-C1}}$ 、 $N_{\text{IP-C2}}$ 为两个IP地址的C段。

3)数值型特征相似度。该类特征包括告警时间间隔和URL编码绕过(攻击者可能使用多次URL编码绕过,用数值表示URL编码绕过的次数),相似度 $R_{\text{Sim-n}}$ 计算方式如下:

$$R_{\text{Sim-n}} = 1 - \frac{\max(n_1, n_2) - \min(n_1, n_2)}{\max(n_1, n_2)} \quad (3)$$

式中: n_1 、 n_2 表示数值型特征的值; $\max(\cdot)$ 表示取最大值。

4)字符串型特征相似度。该类特征包括URL、UA、Accept Language、Cookie、XFF、Post Body、攻击载荷和源IP归属地,相似度 $R_{\text{Sim-s}}$ 计算方式如下:

$$R_{\text{Sim-s}} = 1 - \frac{\text{Levenshtein}(S_1, S_2)}{\max(\text{len}(S_1), \text{len}(S_2))} \quad (4)$$

式中: S_1, S_2 表示字符串类型特征的值; $\text{len}(\cdot)$ 表示字符串的长度; $\text{Levenshtein}(\cdot)$ 表示两个字符串的编辑距离。

5)其他特征相似度:除以上4类以外的其他特征直接进行比较,若完全相同则相似度为1,否则相似度为0。

定义权重向量 $W = (w_1, w_2, \dots, w_{20})$,用于调整不同特征在最终相似性评估中的重要性。权重的确定主要基于重要性和变更难度这两个因素。1)重要性用于衡量该特征在区分不同攻击者过程中的重要性,越能表征攻击者行为个性化的特征,权重越高;2)变更难度用于衡量攻击者修改该特征的难易程度,越难被更改的特征,权重越高。先对各特征从重要性、变更难度两个维度进行0~3分定级,以维度得分相乘得到单特征综合评分;再依据综合评分构建特征两两对比的优先级矩阵,行特征综合评分高赋值1、二者相当赋值0.5,列特征综合评分高赋值0,对矩阵每行求和得到各特征的总优先级值;最后将总优先级值做归一化处理,即可得到各特征基于重要性和变更难度的最终权重。

构建一个 $n \times n$ 的距离矩阵 D ,其中 n 是攻击序列的数量,该矩阵将存储每对攻击序列间的相似性得分。对于任意一对攻击序列 T_i 和 T_j ,应用本文提出的基于DTW的攻击序列相似度算法来计算它们之间的序列距离得分,并将该得分存入序列间的距离矩阵 D 的对应位置中(即 D_{ij})。其中,基于DTW的攻击序列相似度算法如下所示。

算法1 基于DTW的攻击序列相似度算法

输入:时间序列集合 $T = \{T_1, T_2, \dots, T_n\}$,特征权重向量 $W = (w_1, w_2, \dots, w_{20})$,特征相似性比较方法集合 $S = \{s_1, s_2, \dots, s_{20}\}$
输出:距离矩阵 D

```

1. 初始化距离矩阵 $D$ 为 $n \times n$ 的矩阵,所有元素初始化为0
2. for  $i=1$  to  $n$  do
3.   for  $j=i$  to  $n$  do
4.     if  $i=j$  then
5.        $D[i][j]=0$ 
6.     else
7.       //计算序列 $T_i$ 和 $T_j$ 的加权DTW距离
8.       初始化矩阵 $C$ ,所有元素初始化为无穷大,大小为 $(\text{len}(T_i)+1) \times (\text{len}(T_j)+1)$ 

```

```

9.        $C[0][0]=0$ 
10.      for  $a=1$  to  $\text{len}(T_i)$  do
11.        for  $b=1$  to  $\text{len}(T_j)$  do
12.          //计算项 $t_{ia}$ 和 $t_{jb}$ 的加权距离 $d_{\text{weighted}}$ , $t_{ia}$ 为序列 $T_i$ 的第 $a$ 项, $t_{jb}$ 为序列 $T_j$ 的第 $b$ 项
13.           $d_{\text{weighted}}=0$ 
14.          for  $k=1$  to 20 do
15.            //计算项 $t_{ia}$ 和 $t_{jb}$ 第 $k$ 个特征 $t_{ia}[k]$ 、 $t_{jb}[k]$ 间的相似性 $s_{\text{item}}$ 
16.             $s_{\text{item}}=s_k(t_{ia}[k], t_{jb}[k])$ 
17.            //计算加权距离
18.             $d_{\text{weighted}}+= (1-s_{\text{item}})*w_k$ 
19.          end for
20.          //计算 $C$ 矩阵的值
21.           $C[a][b]=d_{\text{weighted}}+\min(C[a-1][b], C[a][b-1], C[a-1][b-1])$ 
22.        end for
23.      end for
24.       $D[i][j]=C[\text{len}(T_i)][\text{len}(T_j)]$ 
25.       $D[j][i]=D[i][j]$ 
26.    end if
27.  end for
28. end for
29. return  $D$ 

```

1.3 攻击序列聚类分析

本文使用基于HDBSCAN的攻击序列聚类算法对攻击序列进行聚类分析,该算法通过构建相互可达距离空间,将原始数据转换为基于密度的层次化结构,能自适应地识别不同密度的攻击簇。同时,该算法在树状图压缩与簇提取过程中,将无法稳定形成簇的孤立点判定为噪声,从而在高噪声环境下依然保持对攻击样本的有效聚类。

该算法以距离矩阵 D 、最小簇大小 n_{size} 和最小样本数 n_{sample} 作为输入,首先调用HDBSCAN算法构建简化的HDBSCAN聚类树,并调用ExtractOptimalClusters过程^[15]对树中每个聚类计算稳定性,并求解最大总稳定性下的非嵌套聚类集合。最终,将选中的聚类映射为正整数标签,未被包含的对象统一标记为噪声(-1),最终输出聚类标签向量。算法时间复杂度为 $O(n^2)$, n 为序列数量。

2 实验与评估

2.1 数据集

当前公开的APT攻击数据集较少,仅有的数据集缺少从侦察到攻击实施的完整流程,攻击细节信息不够丰富,且缺少真实攻击者信息,在完成

同源分析后无法评判分析结果的准确性,无法支撑完成本文实验。因此,本文参考数据集 Warden^[16]和 ECML/PKDD 2007^[17]内容信息,模拟产生 APT 攻击数据:为不同攻击者构建包含 IP 地址池、Web 指纹特征及攻击行为偏好的画像,并结合攻击者各自攻击阶段与时间频率策略偏好,生成包含多样化攻击载荷的结构化 Web 攻击日志数据。数据集包含 459 个 IP 地址共 10 623 条攻击数据,每条数据有 21 个可共聚类分析的字段,其中含有 30% 噪声数据。数据集生成时参考 ATT&CK 模型,模拟了 10 个组织的攻击行为,攻击活动包含侦察、漏洞利用、持久化等攻击阶段,使用了 SQL 注入、XSS 攻击等典型攻击方法。数据集及数据生成代码在知名开源社区 Gitee 上开源^[18]。与真实警报数据相比,该数据集最重要的特点是增加了“真实攻击者”数据标签,可以有效验证同源分析的准确性,为评估同源分析方法的有效性提供了依据。并将该数据集应用到现有的成熟同源分析方法中,其输出结果调整兰德指数(Adjusted Rand Index, ARI)值整体超过 0.8,证明模拟数据集足以被现有方法识别并处理分析,具备作为测试集的有效性。

2.2 实验评估指标

为全面评估本文算法的性能,采用 ARI、归一化互信息(Normalized Mutual Information, NMI)、轮廓系数和 Noise F1 等 4 个互补的评估指标。

1) ARI,基于兰德指数改进而来,取值范围为 $[-1,1]$,值越接近 1 表示聚类结果与真实标签一致性越高,聚类效果越好。

2) NMI,基于信息论中的互信息改进而来,如式(5)所示,取值范围为 $[0,1]$,值越接近 1 表示聚类结果与真实标签的共享信息越多,聚类效果越好。

$$NMI(U,V) = \frac{2 \cdot I(U;V)}{H(U) + H(V)} \quad (5)$$

式中: $I(U;V)$ 是真实标签 U 与预测标签 V 的互信息; $H(U)$ 和 $H(V)$ 分别是 U 和 V 的熵。

3) 轮廓系数用于评估聚类的内部质量,衡量样本和其所属簇的紧密程度(内聚性)与其他簇的分离程度(分离性)之间的平衡。取值范围是 $[-1,1]$,接近 1 表示簇内紧凑、簇间分离,接近-1 表示样本可能被错误分配。对于每个样本 i ,其轮廓系数 $S(i)$ 定义如式(6)所示,整体轮廓系数为所有样本轮廓系数的均值。

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

式中: $a(i)$ 是样本 i 到同簇其他样本之间的平均距离; $b(i)$ 是样本 i 到最近其他簇所有样本的平均距离; $\max(\cdot)$ 表示取最大值。

4) Noise F1 值评估算法识别噪声的准确性,是精确率 $R_{\text{precision}}$ 和召回率 R_{recall} 的调和平均。取值范围为 $[0,1]$,值越高表示噪声识别越准确,评价指标计算公式如下:

$$R_{\text{precision}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (7)$$

$$R_{\text{recall}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (8)$$

$$R_{\text{Noise F1}} = \frac{2R_{\text{precision}}R_{\text{recall}}}{R_{\text{precision}} + R_{\text{recall}}} \quad (9)$$

式中: N_{TP} 表示噪声样本被识别为噪声的样本数; N_{FP} 表示非噪声样本被识别为噪声的样本数; N_{FN} 表示噪声样本被识别为非噪声的样本数。

2.3 实验环境及参数设置

实验操作系统为 Windows 11,处理器为 AMD Ryzen 7 5800H,内存为 16 GB,显卡为单块 NVIDIA RTX 3060(6 GB 显存),编程语言为 Python 3.8。基于 HDBSCAN 的攻击序列聚类算法中最小样本数设置为 2,最小簇大小取值范围为 $[2,12]$ 。

2.4 结果分析

2.4.1 与其他方法的对比

为评估本文方法在噪声干扰下的性能表现,在不同噪声比例的数据上,将本文方法与另外 5 种同源分析方法^[9-12,14]作对比,实验结果如表 1 所示。

表 1 同源分析方法对比结果

方法	ARI			NMI			Noise F1			轮廓系数
	10% 噪声	20% 噪声	30% 噪声	10% 噪声	20% 噪声	30% 噪声	10% 噪声	20% 噪声	30% 噪声	
本文方法	0.966 4	0.945 4	0.949 0	0.965 5	0.963 2	0.968 6	0.871 0	0.939 6	0.968 0	0.396 8
方法 1 ^[9]	0.817 4	0.797 9	0.745 8	0.911 2	0.907 8	0.875 1	0.833 3	0.882 8	0.882 1	0.801 3
方法 2 ^[10]	0.895 6	0.873 7	0.822 9	0.936 3	0.931 4	0.910 3	0.832 3	0.898 0	0.894 7	0.776 5
方法 3 ^[11]	0.848 2	0.784 6	0.597 1	0.915 5	0.857 8	0.817 9	0.855 1	0.645 5	0.644 9	0.296 2
方法 4 ^[12]	0.955 4	0.899 9	0.854 5	0.960 3	0.935 0	0.925 6	0.834 2	0.879 4	0.901 5	0.193 2
方法 5 ^[14]	0.902 0	0.890 7	0.749 5	0.938 8	0.934 6	0.889 2	0.864 2	0.903 2	0.872 9	0.297 3

由表1可得,本文方法在ARI、NMI、Noise F1值3项指标上均显著优于其他方法,且在不同噪声水平下保持高度稳定,其优势源于基于密度的层次聚类机制能够自适应地识别簇结构和离群点,能准确地发现不同攻击行为偏好的团体并分离噪声数据。相比之下,其他方法均随噪声增加表现出明显的性能退化趋势。需说明的是,该算法生成的簇几何结构不够规则,导致轮廓系数偏低,但不影响语义层面的聚类正确性。需要注意的是,以上4种指标都是单独计算,并没有综合考虑其他指标的高低,任何一种聚类方法都无法同时使得4种评价指标取得最高值。

2.4.2 不同距离度量方法的对比

为评估距离度量方式对同源分析性能的影响,分别使用DTW距离与平均点对距离(Mean Pairwise Distance, MPD)矩阵作为基于HDBSCAN的攻击序列聚类算法输入,对比性能差异,其中MPD是一种基于统计特征分析的序列相似性度量方法。实验在30%噪声水平的数据集上进行,实验结果如表2所示。

表2 不同距离度量下聚类对比结果

距离度量方式	ARI	NMI	Noise F1
DTW	0.949 0	0.968 6	0.968 0
MPD	0.854 5	0.925 6	0.901 5

DTW距离相较于MPD距离,在攻击序列的语义相似性度量中表现更优,能够更好地利用时序信息,捕捉序列动态特征。如图2所示,同类攻击序列呈现明显的低距离聚集特征(深色块),验证了DTW距离在行为模式区分中的核心作用。

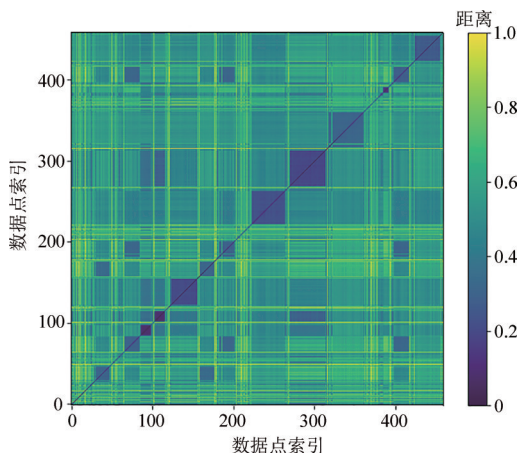


图2 DTW距离矩阵热力图

2.4.3 参数敏感性分析

本节开展参数敏感性分析,探讨关键参数对

同源分析性能与结构稳定性的影响。图3展示了基于密度的聚类算法(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)在不同邻域半径下的指标(最小样本数固定为2)的变化情况。图4展示了基于HDBSCAN的攻击序列聚类算法在不同最小簇大小值下的指标。

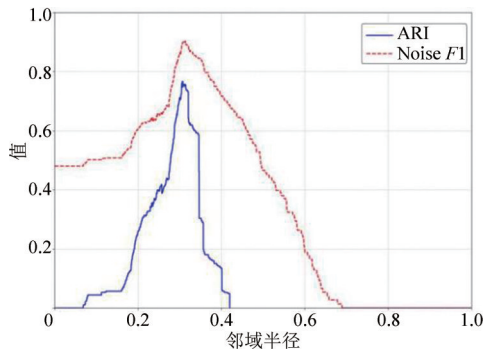


图3 DBSCAN聚类评估指标随邻域半径变化情况

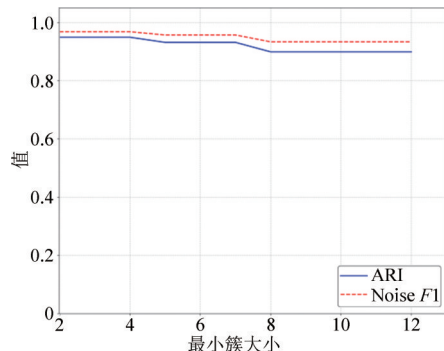


图4 HDBSCAN聚类评估指标随最小簇大小变化情况

由图3可得,随着邻域半径的增加,ARI和Noise F1呈现出明显的先升后降的趋势,表明性能对参数选择高度敏感,仅存在一个较窄的邻域半径取值区间可使DBSCAN取得最优聚类效果,在实际应用中,需要不断优化关键参数才可以达到理想同源分析效果。

由图4可得,可以发现在区间[2,12]内该算法稳定性较高,总体来说HDBSCAN对参数最小簇大小的敏感性较低,在较宽取值范围内性能稳定,具有更好的鲁棒性,更适合应用于实际的Web攻击同源分析任务中。

综上所述,通过实验验证了本文所提方法在Web攻击同源分析中的优越性。该方法可有效捕捉时间序列的语义相似性,在不同噪声水平下均表现出更优的聚类一致性与噪声识别能力。尽管该方法生成的簇轮廓系数偏低,但其在语义正确性和同源分析关键指标上的良好表现,充分证明了该方法在对高噪声、多密度的攻击样本进行同源分析时具有更强的适应性与实用性,为Web攻

击同源分析提供了一种高效可靠的解决方案。

3 结束语

本文围绕当前网络攻击同源分析工作存在的问题,提出一种时序驱动的 Web 攻击行为同源分析方法。该方法首先使用多维度攻击特征体系提取攻击特征,并将其整合形成攻击序列,然后利用基于 DTW 的攻击序列相似度算法得到序列距离矩阵,再利用基于 HDBSCAN 的攻击序列聚类算法,完成对攻击序列的同源分析。实验结果显示,本文方法在 30% 噪声环境下,ARI 值最高可达 94.9%, Noise $F1$ 最高可达 96.8%,证明了本文方法的同源分析有效性。同时,该方法对参数敏感性不强,在 Web 攻击同源分析中具有更好的实践应用能力。未来可从聚类算法的效率优化、拓展数据维度、融合多源数据等方面进一步优化与拓展。

参考文献:

- [1] 王邳伟,何晞杰,易鑫,等.基于 APT 活动全生命周期的攻击与检测综述[J].通信学报,2024,45(9):206-228.
- [2] 杨秀璋,彭国军,刘思德,等.面向 APT 攻击的溯源和推理研究综述[J].软件学报,2025,36(1):203-252.
- [3] 王子晨,汤艳君,潘奕扬.面向取证的网络攻击者溯源分析技术研究综述[J].信息安全研究,2024,10(4):302-310.
- [4] CALISKAN-ISLAM A, LIU A, VOSS C, et al. De-anonymizing programmers via code stylometry[C]//Proceedings of the 24th USENIX Security Symposium. Berkeley, USA: USENIX Association, 2015:255-270.
- [5] WU S Y, WANG P, LI X, et al. Effective detection of Android malware based on the usage of data flow APIs and machine learning[J]. Information and Software Technology, 2016,75:17-25.
- [6] 王建华,陈永乐,张壮壮,等.基于工控系统功能码特征的同源攻击分析[J].计算机工程,2020,46(7):36-42.
- [7] 王跃达,黄潘,荆涛,等.一种基于高速网络的 WebShell 综合检测溯源技术研究[实现[J].信息安全,2021,21(1):65-71.
- [8] 刘浩.基于威胁情报的网络攻击同源性分析研究[D].哈尔滨:哈尔滨工业大学,2025:26.
- [9] KISELEVA O M, PRYTOMANOVA O M, FILAT O A. Application of clustering methods for detecting network threats[J]. Problems of Applied Mathematics and Mathematic Modeling, 2023,23:101-108.
- [10] 白波,冯云,刘宝旭,等.基于网络行为的攻击同源分析方法研究[J].信息安全学报,2023,8(2):66-80.
- [11] 林晨,刚占慧,韦彦,等.面向工业互联网设备的异常行为关联分析攻击溯源技术研究[J].信息安全研究,2024,10(6):532-538.
- [12] BURTON R. Characterizing certain DNS DDoS attacks [DB/OL]. (2019-07-23) [2026-03-18]. <https://arxiv.org/pdf/1905.09958>.
- [13] 吴朝雄,马书磊,石波,等.基于电子指纹的网络攻击溯源技术[J].计算机工程与设计,2020,41(11):3036-3041.
- [14] 王婷,严寒冰,郎波.一种基于图注意力机制的威胁情报归因方法[J].北京航空航天大学学报,2024,50(7):2293-2303.
- [15] CAMPello R J G B, MOULAVI D, SANDER J. Density-based clustering based on hierarchical density estimates [C]//Advances in knowledge discovery and data mining. Berlin, Germany: Springer, 2013:160-172.
- [16] HUSÁK M, ŽÁDNÍK M, BARTOŠ V, et al. Dataset of intrusion detection alerts from a sharing platform [DB/OL]. (2019-06-10) [2026-03-06]. <https://data.mendeley.com/datasets/p6tym3fghz/1>.
- [17] GSI. Web application attack datasets [DB/OL]. (2018-06-13) [2026-03-06]. <https://gitlab.fing.edu.uy/gsi/web-application-attacks-datasets>.
- [18] GAO. Generation of anonymous web attack database [EB/OL]. (2025-10-10) [2026-03-06]. https://gitee.com/haha55566/WEB_attack_database.

(编辑:邵 华)