

基于规则关联的安全数据采集策略生成

陈佩^{1,2}, 李风华^{1,2}, 李子孚^{1,2}, 郭云川^{1,2}, 成林³

(1. 中国科学院信息工程研究所, 北京 100093;
2. 中国科学院大学网络空间安全学院, 北京 100049;
3. 中国信息安全测评中心, 北京 100085)

摘要: 有效的安全数据采集是精准分析网络威胁的基础, 当前常用的全采集、概率采集和自适应采集等采集方法, 未考虑采集数据的有效性和采集数据的关联关系, 消耗过多的资源, 其采集收益和成本率低。针对该问题, 考虑影响采集收益和成本的因素(节点特征间关系、网络拓扑关系、系统威胁状况、节点资源情况、节点相似度等), 设计了一种基于规则关联的安全数据采集策略生成方法。该方法根据节点间的关联规则和系统中所发生安全事件间的关联规则, 构建备选采集项, 缩减数据采集范围; 综合考虑采集收益和采集成本, 设计最大化采集收益和最小化采集成本的多目标优化函数, 基于遗传算法求解该优化函数。与常用采集方法进行比较和分析, 实验结果表明所提方法 12 h 累计数据采集量较其他方案减少了 1 000~3 000 条数据记录, 数据有效性较其他数据采集方案提升约 4%~10%, 证明了所提方法的有效性。

关键词: 策略优化生成; 多目标优化; 数据协同采集; 多关联规则挖掘

中图分类号: TP393

文献标识码: A

DOI: 10.11959/j.issn.2096-109x.2021085

Using rule association to generate data collection policies

CHEN Pei^{1,2}, LI Fenghua^{1,2}, LI Zifu^{1,2}, GUO Yunchuan^{1,2}, CHENG Lin³

1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
3. China Information Technology Security Evaluation Center, Beijing 100085, China

Abstract: Collecting security-related data of devices effectively is the foundation of analyzing network threats accurately. Existing data collection methods (full data collection, sampling based data collection and adaptive data collection) do not consider the validity of the collected data and their correlation, which will consume too much col-

收稿日期: 2020-07-20; 修回日期: 2021-02-01

通信作者: 李子孚, lizifu@iie.ac.cn

基金项目: 国家重点研发计划(2016QY06X1203); 国家自然科学基金(U1836203); 山东省重点研发计划(重大科技创新工程)项目(2019JZZY020127)

Foundation Items: The National Key R&D Program of China (2016QY06X1203), The National Natural Science Foundation of China (U1836203), Shandong Provincial Key Research and Development Program (2019JZZY020127)

论文引用格式: 陈佩, 李风华, 李子孚, 等. 基于规则关联的安全数据采集策略生成[J]. 网络与信息安全学报, 2021, 7(5): 132-148.

CHEN P, LI F H, LI Z F, et al. Using rule association to generate data collection policies[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 132-148.

lection resources, resulting in low collection yield. To address this problem, considering the factors (relationship between node attributes, network topology relationship, threat status, node resource and node similarity) that impact collection costs and benefits, a rule association method to generate collection policies was designed. In the method, two types of association rules (inter-node association rules and inter-event association rules) were adopted to generate candidate data collection items and reduced the scope of data collection. Then, a multi-objective program was designed to maximize collection benefits and minimize collection costs. Further, a genetic algorithm was designed to solve this program. Proposed method was compared with existing data collection methods. The experimental results show that the number of the collected data records of proposed method is 1 000~3 000 less than that of others per 12 hours, and the validity of the collected data of proposed method is about 4%~10% higher than others, which proves the effectiveness of the proposed method.

Keywords: policy optimization generation, multi-objective optimization, collaborative data collection, multiple class-association rules mining

1 引言

安全数据指在入侵检测分析过程中通常使用的数据,该类数据可以协助发现系统遭受的威胁,安全数据可以是特定攻击行为的特征、签名或指纹^[1]。有效采集安全数据是精准分析网络威胁的基础,数据采集的内容决定了威胁分析的准确性和时效性。复杂网络环境(如大数据环境、云计算环境、天地一体化信息网络环境)中大量异构设备产生了不同类型的海量数据,网络拓扑复杂性、数据类型多样性和数据量巨大性,导致当前的全量无差异采集模式可能采集大量噪声数据和冗余数据(如重复的日志文件数据、低信息量的主机状态数据和海量网络流量数据),这将消耗过多计算、存储和带宽资源,威胁分析难度大,数据采集的有效性低,不适应复杂网络环境。

国内外学者在采集策略方面展开了大量研究^[1-5],策略可以分为单节点数据采集策略和多节点协同采集策略。单节点数据采集指在单个节点上采集主机运行数据、日志数据和流量数据等数据,其采集策略关注于单个节点内的采集内容及其采集频率^[6-9],根据数据变化平滑度、数据相关性、预测数据变化幅度等因素,自适应调整采集频率。多节点协同采集指在多个节点上协同采集安全数据^[7,10-14],其采集策略更多地关注网络拓扑、数据传输能耗和安全事件间关联等对数据采集的影响^[11,15-18]。目前采集策略的相关工作只考虑采集策略的效率,未考虑采集数据的有效性。单节点采集策略忽略了节点间隐含的位置、时空

等关联关系对安全威胁的影响,导致部分隐含的威胁预警信息缺失,安全威胁分析的精确度降低。对于多节点协同采集策略,目前的研究大多考虑了节点拓扑和数据传输能耗等影响因素,忽略了采集数据的物理含义,未考虑威胁严重状况、数据有效性等影响因素。

针对上述问题,本文设计了基于规则关联的安全数据采集策略生成方法,该方法考虑多个事件在时空属性或其他事件属性的隐含关联性,设计了备选采集项的确定方法;通过分析采集成本和收益,构建基于多目标优化的采集策略生成算法,支撑数据多节点协同采集。本文的贡献如下。

(1) 数据采集预处理。本文设计了安全事件描述方法,将节点运行状态数据、流量数据和日志记录等不同类型的原始数据进行归一化预处理,得到涵盖事件时间、发生地点、事件类型和事件属性等字段在内的安全事件,实现对异构节点不同类型原始数据的统一描述。

(2) 基于关联规则的备选采集项确定。针对同一节点存在多个安全事件、多个节点间存在同一安全事件这两种场景,基于关联规则挖掘技术,设计备选采集项确定算法。该算法依据实时安全状况和生成的关联规则,构建与实时威胁紧密相关的备选采集项,缩减了数据采集的范围。

(3) 基于多目标优化的策略生成。综合考虑节点间相似性、采集数据有效性和威胁严重程度对采集收益的影响,以及资源占用成本、数据隐私泄露成本和虚假预测风险等因素对采集成本的影响,设计最大化采集收益和最小化采集成本的

多目标优化函数;基于遗传算法求解该优化函数,减少采集数据量,提升采集数据有效性。

2 相关工作

2.1 关联规则挖掘

关联规则算法可以挖掘出数据集中各个变量之间隐含存在的关联关系,该算法广泛应用于数据挖掘的各个场景中,可以有效挖掘事件记录之间的关系,对网络系统中的入侵行为进行检测和发现。根据关联规则层次,将其划分为单层关联规则和多层关联规则。在关联规则层次中,根据不同场景下数据类型的区别,又划分为静态离散数据、动态离散数据和动态序列数据。

在单层关联规则方面,针对单层次的静态离散数据,文献[12]对静态离散数据集 KDDCUP99 提取关联规则,然后根据关联规则的支持度来确定合适的特征,最后采用混合人工神经网络和 AdaBoost 算法对数据集进行异常检测;文献[13]利用矩阵和位运算,设计 BV-Apriori 关联规则算法,结合模糊集技术生成高效的匹配规则库,降低了漏检率和误检率。针对单层次的动态序列数据,文献[14]设计了基于序列种群的遗传关联规则挖掘算法,其关联分析过程可解决入侵模型的特征拟合,遗传进化过程可解决异常模式的增量式预测。文献[19]对企业资源计划管理系统中的日志文件进行分析,基于关联规则算法对商业行为序列进行异常检测,以发现其中的欺诈行为。文献[20]针对云计算平台中的日志审计数据的安全分析需求,通过 Apriori 算法进行轻量级的改进,以删除弱关联项和可调节最小置信度的策略,提高了算法效率。文献[21]综合多个卫星传感器数据,挖掘多个时间序列模式中存在的关联规则,通过主成分分析得出测度空间的结构并识别异常,通过关联规则的变化来确定异常的原因。

在多层关联规则方面,针对多层次的动态离散数据,文献[15]基于分布式的关联规则算法分层次地对异构网络海量安全数据构建关联规则图,首先在异构网络下各个子网中分别对告警数据构建关联规则,然后通过子网间的交互信息进一步确定全局关联规则。针对多层次的动态序列

数据,文献[22]根据运行进程动态行为的关联规则进行异常检测,将各个进程的执行序列存储在不同的关联规则中以便修改,因此该方案适用于数据集动态变动的场景。

虽然这些方案通过在各异的场景下挖掘不同异常行为之间的关联规则来进行异常检测,但缺乏对同一异常行为在不同场景间关联规则的挖掘。因此,本文设计了同一节点内发生多个不同安全事件和相同安全事件发生在不同节点间两种关联规则场景,通过同一事件在多节点的隐含关系推测对安全事件可能发生的节点位置,达到多节点协同采集数据的目的。

2.2 采集策略生成

安全数据的采集策略决定了节点如何进行数据采集,包括节点是否采集某个采集项和以何频率采集两个部分。本文将采集策略生成分为单节点采集策略生成和多节点协同采集策略生成。

针对单节点采集策略生成,文献[6]提出了基于数据变化平滑度的时间间隔调整算法,该算法可得到较为精确的数据拟合曲线。文献[7]利用单节点内数据时间相关性,设计了基于二次指数平滑法的采样频率调整算法,降低了错误丢失率和采样率。文献[8]采用“数据变化平滑时加法增大、数据变化剧烈时乘法减小”的思想,设计了基于旋转门的采集间隔自适应调整算法,该算法能大幅度降低数据采集量。总体上这些方法简单高效,但数据变化平滑度与数据数值不具备明确物理含义,无法确保所采集的数据与系统安全状况紧密相关。文献[9]设计了一种上下文感知的自适应数据采集方案,该方案采集手机主机数据、流量数据和 LTE 数据 3 种异构数据,并提出了基于预测数据变化幅度来调整采集频率的自适应数据采集算法。文献[23]根据高层监测需求和采集项贡献度,将采集策略生成转为采集的成本和收益平衡的非线性优化问题,使生成的采集策略与威胁监测需求紧密相关。虽然这些方法提升了采集效率,但仅限于单节点采集策略生成,未考虑复杂网络环境下多节点的协同采集。

针对多节点协同采集策略生成,文献[16]将压缩感知和群集控制算法相结合来构建标量场图,多个传感器采用群集控制算法在传感区域内

移动, 并和邻居传感器交换采集的数据, 以较低的数据采样率得到目标数据, 降低了数据采集的功耗。为了提高传感器网络的数据采集效率, 文献[17]提出了一种基于能量平衡树的移动 Sink 采集策略, 该策略逐层构建数据采集树, Sink 遍历各个子树根节点收集数据, 有效平衡了不同节点的工作量, 提高了数据采集的实时性和数据上传的效率。文献[18]提出了一种负载均衡和满足有效传输时延的多传感器部署方案, 该方案具有较低的部署成本和能量消耗, 在满足数据采集传输需求的前提下, 延长了网络寿命。文献[10]设计了基于数据聚合的主动监测机制, 解决了能耗控制和数据采集主动性不足的问题。文献[11]针对长细移动自组织网络的带宽占用和时延问题, 设计了一种由分布式分组机制和数据聚合方案组成的协作数据采集传输框架, 该框架可以显著减少数据传输量和网络连接数。虽然这些方案考虑到网络拓扑和能量功耗等因素, 通过调整采集节点和接收节点部署方案等方式来优化采集策略, 但这些方案忽略了采集数据间的关联, 未考虑节点特征和数据特征对采集策略的影响。因此, 本文通过多目标优化方法综合考虑多项采集收益和成本的影响因素, 构建多节点协同采集策略。

3 基于关联规则的备选采集项确定

在大规模网络环境下, 多个节点上的安全事件通常存在两类关联: 节点间的关联和节点内安全事件间的关联。节点间关联是指具有相同特征(如设备厂商、设备类型、部署的网络域、运行状态)的多个节点可能遭受相同类型的威胁; 节点内安全事件间的关联是指同一节点内部的多个安全事件可能伴随出现。因此在采集数据时, 可以利用这两类隐含关联缩减采集的范围。基于上述观测, 本节将 Apriori 关联规则生成算法应用于这两种场景, 设计了备选采集项确定算法, 该算法用于确定可能发现威胁的备选采集项, 为后续的采集策略生成提供输入。

3.1 关联规则生成

3.1.1 基本概念

本节首先介绍 Apriori 算法的相关概念。假设 $I = \{i_1, i_2, \dots, i_i, \dots, i_n\}$ 是包含所有安全事件的项

集, i_i 表示一个安全事件。规定安全事件发生时间差不超过 T_{timediff} 为同时发生。设 T 为同时发生的安全事件集合, 它是项集 I 的子集, 即 $T \subseteq I$ 。每个 T 有唯一标识的 ID, 记为 T_{id} , T 的集合构成了所有时刻同时发生的安全事件集合 D 。关联规则是形如 $X \Rightarrow Y$ 的蕴涵式, 表示当安全事件 X 发生时安全事件 Y 发生, 其中, X 和 Y 均为安全事件项集 I 中的项, 且 $X \cap Y = \emptyset$ 。

对于安全事件 X , 设定 $\text{count}(X \subseteq T)$ 为包含安全事件 X 的 T 集合数量, $\text{support}(X)$ 包含事件 X 的 T 集合数量与所有的 T 集合数量之比, 表示安全事件集合 D 对安全事件 X 的支持度, 如式(1)所示。

$$\text{support}(X) = \frac{\text{count}(X \subseteq T)}{|D|} = \frac{|\{T : X \subseteq T, T \in D\}|}{|D|} \quad (1)$$

$\text{support}(X \Rightarrow Y)$ 表示 D 中包含安全事件 X 和安全事件 Y 的 T 集合数量与所有 T 集合数量之比, 即安全事件 X 和 Y 同时发生的概率, 该值用来衡量安全事件 X 和安全事件 Y 同时发生的支持度, 如式(2)所示。

$$\text{support}(X \Rightarrow Y) = \frac{\text{count}(X \cup Y)}{|D|} = \frac{|\{T : X \cup Y \subseteq T, T \in D\}|}{|D|} \quad (2)$$

支持度具有对称性, 如式(3)所示。

$$\text{support}(X \Rightarrow Y) = \text{support}(Y \Rightarrow X) \quad (3)$$

置信度 $\text{confidence}(X \Rightarrow Y)$ 指同时包含安全事件 X 和安全事件 Y 的 T 集合数量与包含安全事件 X 的 T 集合数量之比, 该值用来衡量安全事件 X 发生的前提下, Y 发生的概率大小, 如式(4)所示。

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{support}(X \Rightarrow Y)}{\text{support}(X)} \quad (4)$$

定义 1 最小支持度: 安全事件集合 D 对安全事件 X 的支持度的最小常数阈值。

定义 2 频繁集: 大于最小支持度的项集称为频繁集, 项的个数为 k 的频繁集被称为频繁 k 项集。

定理 1 如果一个项集是非频繁集, 那么其所有超集也是非频繁集。

Apriori 算法基于定理 1, 通过候选项集的笛卡尔积运算和最小支持度剪枝生成频繁集,

其中, 候选 1 项集为所有仅包含单个项的项集, 候选 k 项集为频繁 $k-1$ 项集的笛卡尔积运算结果。

通过 Apriori 算法构建关联规则的具体流程如下。

(1) 搜索候选 1 项集及其对应的支持度, 剪枝去掉支持度小于最小支持度的 1 项集, 得到频繁 1 项集。

(2) 对频繁 i 项集进行笛卡尔积计算, 若满足定理 1 则添加到候选 $i+1$ 项集中, 剪枝去掉支持度小于最小支持度的 i 项集, 得到频繁 $i+1$ 项集。迭代进行步骤 2, 直到无法找到频繁 $k+1$ 项集。

(3) 由频繁 k 项集生成关联规则, 若频繁 k 项集中某一项大于最小置信度, 则将该项加入关联规则中。

Apriori 关联规则生成算法如下。

算法 1 Apriori 关联规则生成算法

输入 事务数据库 D , 最小支持度阈值 $\min_support$

输出 频繁集 L

```

1) def buildAsscaition( $D, \min\_support$ )
2)  $C_1 = \text{create\_freq1}(D)$  // 构建候选 1 项集
3)  $L_1 = \text{gen\_Lk\_by\_Ck}(D, C_1, \min\_support)$ 
// 构建频繁 1 项集
4)  $L_i = L_1$ 
5) for  $i$  in  $\text{range}(2, k+1)$ 
6)  $C_i = \text{create\_Ck}(L_i, i)$  // 构建候选  $k$  项集
7)  $L_i = \text{gen\_Lk\_by\_Ck}(D, C_i, \min\_support)$  // 剪枝构建频繁  $k$  项集
8)  $L.add(L_i)$ 
9) end for
10) return  $L$ 
11) def gen_Lk_by_Ck( $D, C_k, \min\_support$ )
12) for  $d$  in  $D$ 
13) for  $\text{item}$  in  $C_k$ 
14) if isSubset( $\text{item}, d$ ) // 满足定理 1
15)  $\text{count}[\text{item}] += 1$ 
16) end if
17) end for
```

```

18) end for
19) for  $\text{item}$  in  $\text{count}$ 
20) if ( $\text{count}[\text{item}] / \text{len}(D) \geq \min\_support$ ) // 最小支持度剪枝
21)  $L_k.add(\text{item})$ 
22) end if
23) end for
24) return  $L_k$ 
```

3.1.2 节点内安全事件关联规则生成

一般地, 单个节点遭受的威胁往往由若干连续发生的安全事件组成。例如, 攻击者采用漏洞 CVE-2018-2628 对应用程序 Oracle WebLogic Server 执行远程命令攻击, 发动该攻击前, 首先通过 Nmap 进行主机探测, 确认 Windows 主机的存在, 然后使用工具 WeblogicScan 扫描该主机上是否存在 WebLogic 服务和 CVE-2018-2628 漏洞, 若存在则针对该漏洞构建 T3 协议进行握手, 并发送 payload 进行入侵, 从而破坏活动。从这个例子中可以看出, 在单个节点上发生的多个安全事件间在时序上具有关联性。为了利用这种隐含的关系, 需要对发生在单节点上的多个安全事件生成关联规则。

在同一威胁导致的多个安全事件中, 关联性包括时间属性上的关联性和事件属性上的关联性两类。例如, 固定顺序的攻击步骤, 导致了事件发生时间是固定顺序的; 攻击由同一攻击者发起, 导致安全事件源 IP 属性是相同的。因此本文主要依靠时间属性和事件属性来对在同一时间段内发生的多个事件进行关联, 或者对含有相同源 IP 的多个事件进行关联。这种事件间的关联通过 Apriori 算法, 迭代进行笛卡尔积运算和最小支持度剪枝, 生成事件间的关联规则。

通过时间属性生成关联规则的过程如图 1 所示。左图表示节点 1 在不同的时间区间内同时发生的安全事件集合, 例如, 节点 1 在时间区间 1 同时发生了安全事件 A、B、C 和 D。右图表示对左边的安全事件应用关联规则生成算法, 对项集中的安全事件 A、B、C、D、E、F 迭代地进行两次笛卡尔积运算和剪枝的过程, 最终得到节点 1 上 6 个安全事件的关联规则和频繁集。将事件和事件发生时间作为事务数据库 $D_{\text{event_time}}$, 输入最

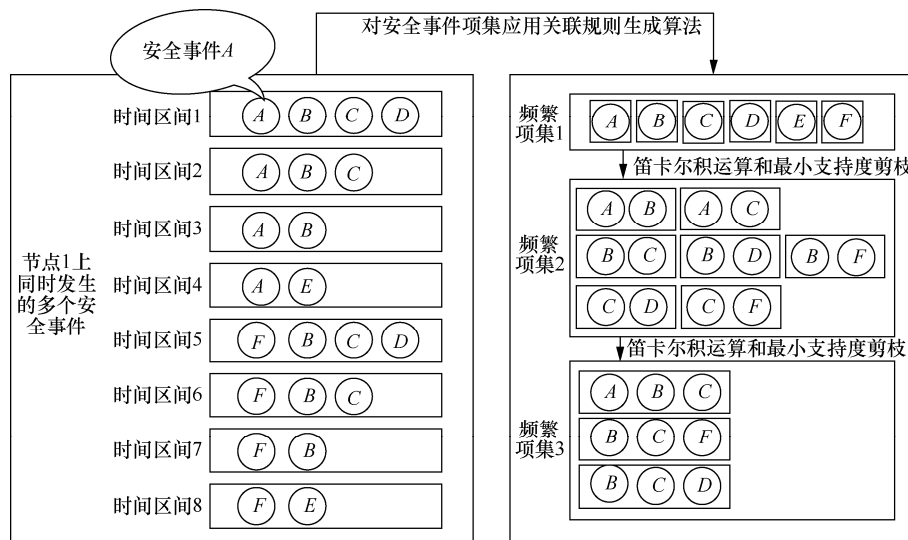


图1 安全事件关联规则生成示意
Figure 1 Schematic diagram of security event correlation rule generation

小支持度 $\min_support$ ，利用 Apriori 算法，得到基于时间属性的节点内事件关联规则频繁集生成算法，如下所示。

算法 2 基于时间属性的节点内关联规则生成算法

输入 事件集合 $event$ 和每个事件发生的时间 D_{event_time} ，最小支持度 $\min_support$

输出 事件关联规则频繁集 $Rule_{event}$

1) **if** $\text{count}(event) > Event_{Threshold}$

2) $Rule_{event} = \text{buildAssociation}(D_{event_time},$

$\min_support)$

3) **end if**

通过事件属性也可以生成关联规则。例如，定义具有相同源 IP 为关联条件，如果事件 A 具有源 IP 属性，同时该源 IP 属性值与频繁项集中某个安全事件的源 IP 属性值相同，则将事件 A 添加到候选项集中，按照相同的方式迭代进行笛卡尔积运算和最小支持度剪枝，生成关联规则。

3.1.3 相同安全事件在节点间的关联规则生成

节点间的相似性导致多个节点可能同时遭到同一入侵者的攻击，产生同样的安全事件，因此可将 Apriori 算法应用于发生相同安全事件的节点，生成节点间关联规则，从而利用多节点的关联协助确定备选采集项。具体的应用方式如下。

给定安全事件 $Event_i$ ，以及事件发生的节点

$Host_a$ 、 $Host_b$ ，在节点 $Host_a$ 和节点 $Host_b$ 上同时发生安全事件 $Event_i$ 的历史数据会影响这两个节点对事件 $Event_i$ 的关联性。将同时发生安全事件 $Event_i$ 的节点项集作为关联规则生成算法的输入，可以获得安全事件 $Event_i$ 在不同节点间的关联规则。如图 2 所示，左图表示在 4 个时间区间，多个节点同时发生安全事件 A，如节点 1、2、3 于时间区间 1 同时发生安全事件 A；右图表示对左图的节点项集应用关联规则生成算法，对节点 1、2、3、4、5 迭代地进行两次笛卡尔积运算和最小支持度剪枝的过程，最终得到事件 A 在 5 个节点上的关联规则集合。将事件和事件发生节点作为事务数据库 D_{event_host} ，输入最小支持度 $\min_support$ ，利用 Apriori 算法，可得到节点间关联规则频繁集生成算法，如算法 3 所示。

算法 3 节点间关联规则生成算法

输入 事件集合和每个事件发生的节点 D_{event_host} ，最小支持度 $\min_support$

输出 节点关联规则频繁集 $Rule_{host}$

1) **if** $\text{count}(event) > Event_{Threshold}$

2) $Rule_{host} = \text{buildAssociation}(D_{event_host},$

$\min_support)$

3) **end if**

3.2 备选采集项生成

本节利用 3.1 节生成的单节点中安全事件间

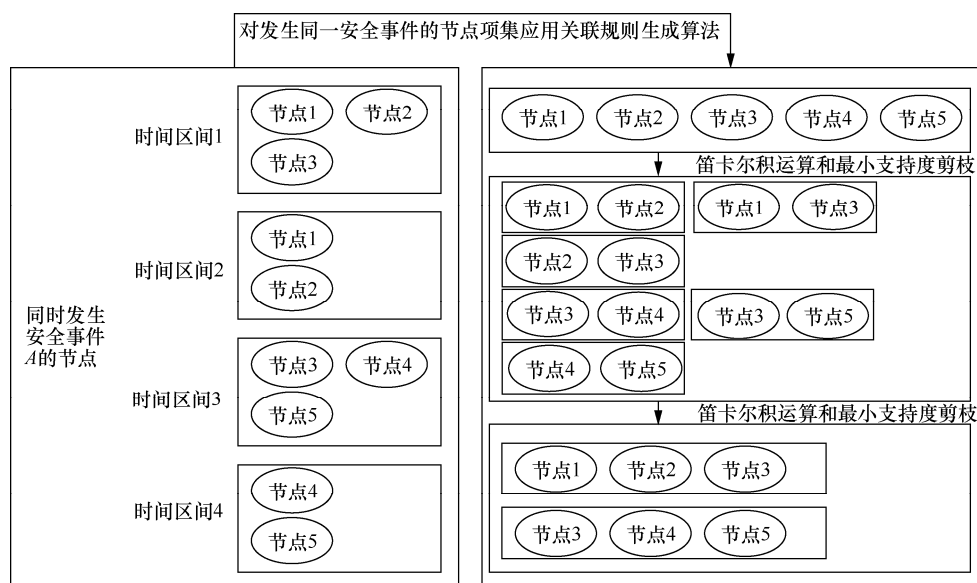


图 2 节点关联规则生成示意

Figure 2 Schematic diagram of node association rule generation

关联规则和发生相同安全事件的节点间关联规则，结合实时的安全状态，生成备选采集项，达到缩减采集范围的目的。该算法的核心思路为依次将两种关联规则中与实时事件有关联的事件加入备选安全事件集，然后递归搜索是否有与新加入事件具有关联的事件，通过最远规则距离的约束来终止递归，最终将备选安全事件集映射为备选采集项。

在上述过程中，多个关联规则集造成了安全事件间可能具有间接的关联规则，由此产生了规则距离，本文用规则距离表示关联规则的间接次数。例如，在节点 $Host_i$ 上事件 A 与事件 B 在节点内具有关联规则，事件 B 与事件 C 在节点 $Host_i$ 和 $Host_2$ 间具有关联规则，则节点 $Host_i$ 和 $Host_2$ 上事件 A 与事件 C 间接地具有关联规则，且事件 A 与事件 C 的规则距离为 2。规定规则距离为 0 表示实时发生的安全事件，规则距离为 1 表示与实时发生的安全事件具有直接的关联规则，规则距离大于 1 表示和实时发生的安全事件具有间接的关联规则。规则距离越大，则与实时发生安全事件的关联越低。需要注意的是，具有间接关联规则的安全事件在加入备选安全事件集时需要标注规则距离。

接下来介绍备选采集项构建算法，输入包括关联规则 $Rule_{event}$ 和 $Rule_{host}$ ，实时发生在节点 i

的安全事件 j 集合 $event$ ，最远规则距离为 $distance_{max}$ ，安全事件和采集项的映射关系 $EIMap$ 。其中， $Rule_{event}$ 和 $Rule_{host}$ 为 3.1 节的输出，分别表示在单个节点上的安全事件关联规则集合和发生相同安全事件的节点关联规则集合。备选采集项构建如算法 4 所示。

算法 4 备选采集项构建算法

输入 事件集合 $event$ ，节点关联规则集 $Rule_{host}$ ，事件关联规则集 $Rule_{event}$ ，最远规则距离 $distance_{max}$ ，安全事件和采集项的映射关系 $EIMap$

输出 备选采集项集合 AI

```

1) for  $e$  in  $event$ 
2)   addToAEAndRecordDistance( $e$ ,  $AE$ , 0) //将安全事件  $e$  添加到备选安全事件集合  $AE$  中，记录距离为 0
3) end for
4) for  $e$  in  $event$ 
5)    $curDistance = 0$ 
6)    $e1, distance = getNextAssociationEventAndDis(e, Rule_{host})$  //获取安全事件  $e$  在节点关联规则集  $Rule_{host}$  中的一个关联项和该关联项的记录距离
7)   while  $e1 \neq NULL$ 
8)      $curDistance += 1$ 

```

```

9)      if curDistance < distancemax do
10)      addToAEAndRecordDistance (e1,
AE, curDistance)
11)      end if
12)      e1, distance = getNextAssociation-
nEventAndDis(e, Rulehost)
13)      end while
14)      e2, distance = getNextAssociationE-
ventAndDis(e, Ruleevent)
15)      while e2 != NULL
16)      curDistance += 1
17)      if curDistance < distancemax do
18)      addToAEAndRecordDistance(e2,
AE, curDistance)
19)      end if
20)      e2, distance = getNextAssociation-
nEventAndDis(e, Ruleevent)
21)      end while
22)      end for
23)      AI = mapEventToItem(AE, EIMap)

```

算法3分为3个部分，首先将实时发生的安全事件加入备选安全事件集；然后在两个关联规则中迭代搜索备选安全事件集中相关联的事件，将当前规则距离和最远规则距离作为加入备选安全事件集的判断条件，以决定是否将搜索到的事件加入备选安全事件集并记录规则距离；最后，通过安全事件和采集项之间的映射关系，将备选安全事件集映射为备选采集项。算法的具体流程如下。

(1) 在备选安全事件集中增加当前已发生事件，新增备选安全事件项的记录距离为0，该步骤对应算法第1~3行。

(2) 设置当前距离 curDistance 为0，该步骤对应算法第5行。

(3) 将当前距离 curDistance 加1，搜索 Rule_{hosti} 的安全事件 j ，如果存在关联规则，且备选安全事件集中不存在安全事件 j ，则将安全事件 j 加入备选安全事件集，安全事件 j 的记录距离值为当前距离值；如果存在关联规则，且备选安全事件集中已存在安全事件 j ，其记录距离大于当前距离，则更新安全事件 j 的记录距离为当前

距离值。同理，搜索 Rule_{eventj} 的 i 节点，将节点 i 对应的安全事件 j 根据关联规则和当前距离进行对应处理，该步骤对应算法第6~20行。

(4) 针对新增加的备选安全事件，执行步骤(3)，直到当前距离大于最远规则距离 distance_{max}，该步骤对应算法第9~11行和第17~19行。

(5) 对每个实时发生的事件均执行步骤(2)、步骤(3)和步骤(4)，得到最终的备选安全事件集合，该迭代过程对应算法4~22行。

(6) 将备选安全事件集映射为备选采集，该步骤对应算法23行。

4 多目标优化函数

在采集策略的生成过程中，采集行为可能带来正向收益(如采集到入侵行为)和负面成本(如浪费采集资源却未监测到任何入侵行为)。为了生成目标函数，需要评估采集收益和采集成本。

4.1 采集收益计算

影响采集收益的因素包括节点间相似性收益、采集数据有效性收益和威胁严重程度收益。在节点间相似性方面，具有相似特征的节点在同时段可能遭受相同类型的攻击，故通过评估节点的相似性，可以识别威胁相关节点，相似性越高则协同采集收益越大；在采集数据有效性方面，对威胁分析而言，所采集数据的有效性越高，准确分析出威胁的可能性越高，相应的采集收益也越大；在威胁严重程度方面，对于整个系统的采集策略，若威胁严重程度较高，威胁影响较多的节点，则该威胁对应采集项的采集收益越大。

4.1.1 节点间相似性收益

通过节点特征可以计算多个节点在不同维度上的相似性。对于本文所使用的12个采集项和22种分析的安全事件，选用 n 个典型的节点特征来衡量节点的相似性，得到节点相似性对这些安全事件发生的影响。为此采用 n 维向量 $\langle f_1, f_2, f_3, f_4, \dots, f_n \rangle$ 表示节点特征，表示节点 i 的 n 个特征如式(5)所示。

$$\text{Feature}_{\text{Host}i} = \langle f_1, f_2, f_3, f_4, \dots, f_n \rangle \quad (5)$$

根据节点特征的变动频繁程度，可以将它们分为3类，分别是静态特征、半动态特征和动态特征。静态特征是指在节点的整个生命期内不会

变动的特征,如节点型号、主板型号、内存大小、CPU 类型和 GPU 类型等。由于某些入侵者可能根据节点的硬件型号发动特定攻击,产生相同的入侵行为,故静态特征相似的节点可能发生类似的安全事件。半动态特征是指在节点的整个生命期内不会频繁变动的特征,如节点的操作系统版本和节点所在的网段等。攻击者可以针对半动态特征实施入侵行为,如针对系统版本漏洞进行攻击,根据网络拓扑环境将节点当作跳板攻击网络中其他节点。动态特征指在节点的整个生命期内频繁变动的特征,如开启的端口号、用户进程的 MD5 值、命令执行记录等。动态特征值的相同表明节点具有发生相同安全事件的风险。

节点特征的数值类型包括 3 类:单值、多值和序列值。单值特征为单个离散值,多值包括多个离散值,序列值指具有顺序的多个离散值。对于单值特征,通过 0/1 表示该特征的离散值是否相同;对于多值特征,通过该特征的相同值数量表示该特征的相似度;对于序列值特征,通过连续相同元素的数量表示该特征的相似度。节点的相似性同样采用 n 维向量 $\langle S_1, S_2, S_3, S_4, \dots, S_n \rangle$ 表示。节点 i 和节点 j 之间的相似度如式(6)所示。

$$\text{Similarity}_{\text{Host}ij} = \langle S_1, S_2, S_3, S_4, \dots, S_n \rangle \quad (6)$$

基于节点相似度,可计算采集项的相似性收益。式(7)给出了 CPU 利用率的相似性收益,其中 k_{hostsim} 表示节点相似性的加权系数, C_{cpu} 表示采集项 CPU 利用率与节点相似性的关联向量,当采集项和节点相似性具有关联才会产生相似性收益。

$$\text{HostSimBenefit}_{\text{cpu}} = \text{Similarity}_{\text{Host}ij} C_{\text{cpu}} k_{\text{hostsim}} \quad (7)$$

系统中全部的节点间相似性的收益如式(8)所示。

$$\text{HostSimBenefit}(X) = \sum_{h1=0}^{\text{Num}_{\text{host}}} \sum_{h2=0}^{\text{Num}_{\text{host}}} \sum_{i=0}^{\text{Num}_{\text{item}}} \text{Feature}_{h1h2} C_i k_i \text{freq}_{hi} \quad (8)$$

其中, Num_{host} 表示节点数量,两层循环表示每个节点 $h1$ 和其他节点 $h2$ 的相似度, Num_{item} 表示采集项数量, Feature_{h1h2} 表示节点 $h1$ 和节点 $h2$ 的

相似特征向量, C_i 表示采集项与节点相似性的关联向量, k_i 表示节点相似性的加权系数, freq_{hi} 表示在节点 h 上采集项 i 的采集参数。

4.1.2 采集数据有效性收益

采集数据的有效性收益是指采集到的安全事件在有效监测威胁方面的收益,本文采用互信息来衡量采集项的数据有效性。每类采集项至少生成一种安全事件,采集项和安全事件之间为一对多的关系,故通过各个安全事件数据有效性收益的叠加,可以得到单个节点上单个采集项的数据有效性收益系数,如式(9)所示。

$$B = \sum_{e2 \in E} \sum_{e1 \in E1} p(e1, e2) \log \left(\frac{p(e2|e1)}{p(e2)} \right) \quad (9)$$

其中, $e1$ 表示此节点上发生的安全事件, $E1$ 表示此节点上发生的所有安全事件集合, $e2$ 表示该采集项采集到的安全事件, $E2$ 表示该采集项所采集到的安全事件集合。

按照采集参数采集给定节点的数据,得到该节点的数据有效性收益;综合各个节点的数据有效性收益,获得整个系统的数据有效性收益,如式(10)所示。

$$\text{EventInfoBenefit}(X) = \sum_{h=0}^{\text{Num}_{\text{host}}} B_h \text{freq}_h \quad (10)$$

其中, B_h 表示在 h 节点上各个采集项基于互信息的收益矩阵, freq_h 表示在节点 h 上各个采集项的采集参数矩阵。

4.1.3 威胁的严重程度

威胁的严重程度与发生安全事件的节点数量和安全事件发生的频率相关,系统的威胁严重程度收益可用式(11)来计算:

$$\text{ThreatenBenefit}(X) = \sum_{i=0}^{\text{Num}_{\text{item}}} \sum_{e=0}^{\text{Num}_{\text{event}}} \sum_{h=0}^{\text{Num}_{\text{host}}} (k_{\text{threaten}} \text{freq}_{he} + k_{\text{unocc}} e^{-t_{ne}}) \text{freq}_{hi} \quad (11)$$

其中, Num_{item} 表示采集项的数量, $\text{Num}_{\text{event}}$ 表示采集项 i 对应的安全事件数量, Num_{host} 表示目前发生的安全事件 e 的节点数量, freq_{he} 表示安全事件 e 在节点 h 上发生的频次, k_{threaten} 表示威胁收益权重, t_{ne} 表示安全事件 e 在节点 h 上持续未发生的时间长度, k_{unocc} 表示安全事件长期未发生的

收益系数, freq_{hi} 表示在节点 h 上采集项 i 的采集参数。

4.2 采集成本计算

影响采集成本的因素包括资源占用、数据隐私泄露、策略变动和虚假预测风险。数据隐私泄露成本由采集项的敏感度和采集数据量决定, 策略变动成本是指之前采集到有效数据的采集项因策略变动而停止采集导致的损失, 虚假预测风险成本是由部署的采集项长期未采集到安全事件导致的成本。

4.2.1 资源占用成本

资源占用成本指在数据采集过程中引发的计算资源、内存资源、网络带宽和加密运算等成本, 这些成本与采集数据量和采集频率均成正比, 同时不同类型的采集项具有不同的采集数据量。基于上述资源占用, 在单个节点上对采集项 i 进行采集所占用的资源成本如式(12)所示。

$$\text{TermResourceCost}_i = k_{\text{Data}_i} \text{DataSize}_i / t_{\text{internal}_i} \quad (12)$$

其中, DataSize_i 表示采集项 i 在其采集时间间隔 t_{internal_i} 内采集的数据量, k_{Data_i} 表示采集项 i 的成本系数。采集项占用资源比例的成本如式(13)所示。

$$\text{NodeCost}(X) =$$

$$\left\{ \begin{array}{l} \sum_{i=0}^{\text{Num}_{\text{item}}} \text{TermResourceCost}_h \text{freq}_i \\ k_1 \max_{\text{cpu_ratio}, \text{mem_ratio}, \text{net_ratio}} < 0.08 \\ \sum_{i=0}^{\text{Num}_{\text{item}}} \text{TermResourceCost}_h \text{freq}_i \\ k_2 0.08 \leq \max_{\text{cpu_ratio}, \text{mem_ratio}, \text{net_ratio}} < 0.1 \\ \sum_{i=0}^{\text{Num}_{\text{item}}} \text{TermResourceCost}_h \text{freq}_i \\ k_3 \max_{\text{cpu_ratio}, \text{mem_ratio}, \text{net_ratio}} \geq 0.1 \end{array} \right. \quad (13)$$

其中, Num_{item} 表示采集项数目, freq_i 表示当前节点上采集项 i 的采集参数, $\max_{\text{cpu_ratio}, \text{mem_ratio}, \text{net_ratio}}$ 表示当前节点执行采集过程中占用 CPU 利用率、内存利用率和网络利用率中的最大资源利用率。

数据采集在整个系统中所占用资源的总成本如式(14)所示。

$$\text{ResourceCost}(X) = \sum_{h=0}^{\text{NodeCnt}} \text{NodeCost}_h \quad (14)$$

4.2.2 数据隐私泄露成本

数据隐私泄露成本指因采集导致隐私泄露的成本, 采集项的敏感程度决定了隐私泄露的危害严重程度, 故采集项的数据隐私泄露成本与采集项的敏感度 C_{sen} 常量相关。用式(16)来计算数据隐私泄露成本。

$$\text{SensitivityCost}(X) = \sum_{h=0}^{\text{Num}_{\text{host}}} \sum_{i=0}^{\text{Num}_{\text{item}}} \text{Sen}_i \text{freq}_{hi} \quad (15)$$

其中, Num_{item} 表示采集项数量, Num_{host} 表示节点数量, Sen_i 表示采集项 i 的敏感度, freq_{hi} 表示在节点 h 上采集项 i 的采集参数。

4.2.3 策略变动成本

策略变动成本指因调整策略导致的采集收益减少的负面影响。策略变动成本与采集项的效用相关, 若采集项持续未采集到安全事件, 停止该采集项的成本较小; 若采集项频繁采集到安全事件, 则停止该采集项的成本急剧上升, 具体如式(16)所示。

$$\text{BenefitCost}(X) =$$

$$\sum_{h=0}^{\text{Num}_{\text{host}}} \sum_{i=0}^{\text{Num}_{\text{item}}} (e^{\text{CollectEventCnt}_h / t_{\text{lasthi}}} \min\{\text{freq}_{hi}^{\text{cur}} - \text{freq}_{hi}, 0\}) \quad (16)$$

其中, CollectEventCnt_h 表示在节点 h 上各采集项采集到的安全事件数量, t_{lasthi} 表示采集项 i 在节点 h 上持续采集的时间长度, $\text{freq}_{hi}^{\text{cur}}$ 、 freq_{hi} 分别表示在节点 h 上采集项 i 当前的和新生成的采集参数, 如果某一采集项进行采集则置为 1, 否则置为 0。当某一采集项当前进行采集而新生成的策略未进行采集时, $\min\{\text{freq}_{hi}^{\text{cur}} - \text{freq}_{hi}, 0\}$ 取值为 1, 其他情况取值为 0。

4.2.4 虚假预测风险成本

虚假预测风险成本指在某个节点上部署某个采集项却长期未采集到安全事件而产生的成本, 该项成本与部署的频次和未采集到任何安全事件的频次相关, 具体如式(18)所示。

$$\text{PredCost}(X) =$$

$$\sum_{h=0}^{\text{Num}_{\text{host}}} \sum_{i=0}^{\text{Num}_{\text{item}}} \left(\sum_{j=0}^{\text{HisCnt}_i} \text{CollectEventCnt}_{ihj} / \text{HisCnt}_{ih} \right) \text{freq}_{hi} \quad (17)$$

其中, Num_{host} 表示节点数量, Num_{item} 表示采集项数量, HisCnt_{ih} 表示采集项 i 在节点 h 上的历史采集次数, $\text{CollectEventCnt}_{ihj}$ 表示采集项 i 在节点 h 上第 j 次部署中采集到安全事件的数量, freq_{hi} 表示在节点 h 上采集项 i 的采集参数。

4.3 目标函数

将采集策略中数据采集的收益量化为节点相似性、数据有效性和威胁严重程度收益, 如式(18)所示; 将数据采集的成本量化为资源占用成本、数据隐私泄露成本、策略变动成本和虚假预测虚假成本等方面, 如式(19)所示。最终, 得到的目标函数如式(20)所示。

$$f_{\text{Benefit}}(X) = \text{HostSimBenefit}(X) + \text{EventInfoBenefit}(X) + \text{ThreatenBenefit}(X) \quad (18)$$

$$f_{\text{Cost}}(X) = \text{ResourceCost}(X) + \text{SensitivityCost}(X) + \text{BenefitCost}(X) + \text{PredCost}(X) \quad (19)$$

$$\begin{cases} \max \max f_{\text{Benefit}}(X) \\ \min \min f_{\text{Cost}}(X) \end{cases} \quad (20)$$

其中, $X = \{\text{freq}_{i1}, \text{freq}_{i2}, \dots, \text{freq}_{hi}, \dots, \text{freq}_{\text{Num}_{\text{host}}\text{Num}_{\text{item}}}\}$, 表示在各节点上采集备选采集项的采集参数。

根据采集项采集方式的不同, 采集策略中采集项的采集参数是不一样的, 采集方式分为 3 种: 主动探测采集、定时采集和抽样采集。主动探测采集仅能表示是否对该采集项进行采集, 采集参数取值为 0 或 1; 定时采集指按指定的频率定时采集数据, 采集参数取两次采集行为的间隔时间; 抽样采集是指按照指定的概率来抽样地采集数据, 将 1%~100% 的抽样概率映射为 1~100 的采集参数。其中, 定时采集和抽样采集的采集参数均需要确定取值的基准范围, 避免在后续计算过程中出现过度频繁采集、频率或概率取值为负、概率大于 1 等异常情况。

5 基于遗传算法的多目标优化函数求解

通过求解多目标优化函数, 获得采集收益最大、采集成本最小时各节点的采集参数, 从而得到多节点的协同采集策略。如果协同采集策略决策变量多, 导致优化算法收敛较慢。为了解决此问题, 本文采用遗传算法在多点并行搜索解, 提

升收敛速度, 并避免局部最优解。本文定义种群为采集策略解集, 一个种群个体为一种采集策略解, 采集策略解包括多个基因片段, 每个基因表示单个节点对某个采集项的采集参数。

遗传算法包括初始化、适应度计算、个体选择和交叉变异等。在初始化阶段, 随机生成初始种群的基因, 即随机地生成初始采集策略。在适应度计算阶段, 对每个采集策略按照策略和基因之间的映射方案进行解码, 将解码后的值作为适应度计算的参数, 计算该采集策略的适应度; 判断当前解的适应度是否满足要求, 若满足则停止, 否则执行个体选择。在个体选择阶段, 根据选择算子, 将适应度较低的采集策略作为“劣汰”部分进行剪枝, 筛选适应度较高的一些采集策略作为“优胜”部分, 遗传它们的基因。在交叉变异阶段, 按照交叉算子对优胜策略进行两两匹配, 根据交叉概率交换部分基因, 组合为新的采集策略; 然后, 新的采集策略根据变异概率在基准范围内对基因进行随机变异。

算法过程中可能影响策略生成结果的因素包括: 采集策略和基因间映射、采集策略的适应性函数、优胜策略的选择算子、种群之间的交叉算子。下面对上述因素进行具体介绍。

(1) 采集策略和基因间映射。将采集策略进行编码, 映射为种群个体的基因。采集项的个数和节点的个数共同决定了基因的个数, 单个节点的一个采集项取值表示种群中个体一个基因的值。

(2) 采集策略的适应性函数。对于采集项的收益和成本包括 4.1 节和 4.2 节提出的 7 种因素, 通过计算每个节点收益和成本的差值, 得到整个系统的成本和收益, 最终构建遗传算法的适应性函数, 如式(21)所示。

$$\begin{aligned} \text{Fit}(X) &= f_{\text{Benefit}}(X) - f_{\text{Cost}}(X) = \\ &\text{HostSimBenefit}(X) + \text{EventInfoBenefit}(X) + \\ &\text{ThreatenBenefit}(X) - \text{ResourceCost}(X) - \\ &\text{SensitivityCost}(X) - \text{BenefitCost}(X) - \\ &\text{PredCost}(X) \end{aligned} \quad (21)$$

(3) 优胜策略的选择算子。常见的选择算子包括随机竞争选择算子、锦标赛选择算子和轮盘选择算子等。本文采用轮盘选择算子, 根据个体

的适应度和群体适应度的比值，决定该个体的选择概率，达到优先选择高适应度个体的效果。在概率计算方面，由于不同个体的适应度差距不大，选择概率差异性较小，所以需要每个个体的适应度值减去最小的适应度，然后对个体的适应度进行常数系数的成倍放大，增大个体之间的差距，增加高适应度个体的优势，加快迭代求解的速度。

(4) 种群间的交叉算子。交叉算子采用随机交叉方式，即随机选择两个个体进行基因交叉。单个节点的一个采集项取值用种群个体一个基因表示，只有在相同采集项对应的基因间才可以按交叉概率进行交叉操作。

6 实验

6.1 实验设置

本文在 OpenStack 云计算环境中进行模拟实验，分别从策略的适应度值、时间开销和采集效果 3 方面衡量算法。在实验中模拟了 20 个节点，将节点分配到 3 个网络域，对每个节点设置了静态特征、半动态特征和动态特征 3 类，共计 n 项特征。静态特征包括节点型号、主板型号、内存大小、CPU 类型和 GPU 类型；半动态特征包括

系统版本和网段；动态特征包括开启的端口号、进程 MD5 值和执行的命令记录。根据这些设定，可以得到完整的 $20 \times n$ 的节点特征矩阵，根据该矩阵可以计算得到节点间的相似度。

根据 4.3 节的讨论，对每个节点设置了主动探测、定时和抽样 3 类采集方式；通过调研常用入侵检测系统 (Snort、OSSEC 等)、安全数据集 (KDD CUP99、ADFA IDS、MCFP 等) 和文献[4-5] 中采集数据的内容和类型，本文选用具有代表性的 12 个采集项并设定其采集参数取值范围，如表 1 所示。表 2 给出了表 1 的 12 个采集项可分析出的安全事件。

6.2 实验结果

6.2.1 适应度演化过程实验

针对 3 种不同规模的网络环境和不同的攻击状态，进行采集策略生成实验。

(1) 从 20 个节点中选择 3 个进行实验，在节点 1 和节点 2 上监测到端口扫描事件，在节点 1 上监测到 SSH 弱口令试探事件，在此攻击状态下通过遗传算法生成 3 个节点的采集策略。不同规模节点下采集收益和成本适应度值的趋势如图 3 所示。

表 1 采集项及采集参数
Table 1 Collection item and collection parameter

采集项类型	采集项名称	采集参数类型	采集参数取值范围
主动探测采集类	命令执行记录	开关	0/1
	登录记录	开关	0/1
	Web 访问记录	开关	0/1
	CPU 负载	频率范围	1~300
定时采集类	内存负载	频率范围	1~300
	IO 负载	频率范围	1~300
	网络负载	频率范围	1~300
	用户进程 MD5 值	频率范围	10~1 800
抽样采集类	TCP 记录	抽样概率	0.01~1
	UDP 记录	抽样概率	0.01~1
	ICMP 记录	抽样概率	0.01~1
	流量应用层数据	抽样概率	0.01~1

表 2 安全事件及事件特有属性
Table 2 Security events and attributes of security events

安全事件	过程	事件特有属性
端口扫描	扫描各个端口的服务	源 IP
SYN DDoS 攻击	不断发送 SYN 连接请求	多个源 IP、持续时间
SSH 登录	远程 SSH 登录	源 IP、登录账户
Web 扫描	扫描 Web 路径	源 IP、访问 URL 数量、持续时间
弱口令探测	多次尝试 SSH 登录	源 IP、失败次数、持续时间
CPU 占用异常	CPU 占用过高, 大于设定阈值	阈值、持续时间
未知用户进程运行	设立用户进程的白名单, 发现未知的用户进程运行	程序 MD5 值
更改文件权限	执行更改文件权限的命令	命令执行账户
传输文件	执行传输文件的命令	命令执行账户
查看 Linux 版本	执行查看 Linux 版本的命令	命令执行账户
密码修改	执行密码修改的命令	命令执行账户
SQL 注入	尝试进行 SQL 注入攻击	源 IP
挖矿程序运行	执行挖矿程序	用户进程 MD5 值, CPU 负载
数据库连接	尝试远程连接本主机的数据库	源 IP
FTP 文件下载	尝试下载文件	源 IP
FTP 远程上传	FPT 上传文件	源 IP、文件 MD5
Telnet 试探	多次尝试 Telnet 登录	源 IP
CVE-2018-2628	构造基于 WebLogic 反序列化的数据包, 并进行发送	源 IP
CVE-2014-0457	构造基于 CVE-2014-0457 的数据包, 并进行发送	源 IP
发起 SYN DDoS	运行脚本, 持续向其他主机发起 TCP 连接, 并连接失败	源 IP、程序 MD5 值
CVE-2017-2636	基于 Linux 内核漏洞 CVE-2017-2636 进行本地提权	命令执行账户
ICMP 探测	发送 ICMP 数据包	源 IP

(2) 从 20 个节点中选择 10 个进行实验, 在节点 1 和节点 2 上监测到端口扫描事件, 在节点 1 上监测到 SSH 弱口令试探事件, 在节点 2 上开启 80 端口, 监测到网站路径扫描事件。在上述攻击状态下, 通过遗传算法生成 10 个节点的采集策略, 各采集收益和成本在迭代过程中的适应度值如图 3 所示。

(3) 针对 20 个节点进行实验, 在节点 4 开启 80 端口, 部署 Apache 服务器, 在节点 1、节点 2、节点 3 和节点 4 上监测到端口扫描事件, 在节点 1 上监测到 DDoS 攻击事件, 在节点 4 监测到网站路径扫描事件。在上述攻击状态下, 通过遗传算法生成 20 个节点的采集策略, 各采集收益和成本在迭代过程中的适应度值如图 3 所示。

从图 3 可知, 不同网络规模和攻击状态在采集策略生成过程中, 本文所提出方案的各项采集

收益、采集成本和采集策略的适应度值波动的总体趋势较为相似, 即各项采集收益逐步增加、采集成本逐渐减小、采集策略的适应度值整体增加, 其中采集策略的适应度值为各项采集收益和采集成本的差值。除此之外, 采集策略中的节点数目较少时, 各项采集收益和采集成本更易收敛。

观察图 4 可知, 在 20 个节点的环境下, 各项采集收益/成本增加变化的幅度。在遗传算法的迭代过程中, 采集收益的增加主要源于节点相似收益和数据有效性收益, 威胁严重程度收益增加不多; 整体适应度值增加主要是因为采集成本的减少。该过程表明初始的随机采集策略对各个采集项均有采集, 且可能以较高的频率或概率进行采集, 随着迭代的进行, 威胁无关项的采集频率/概率逐步减小。

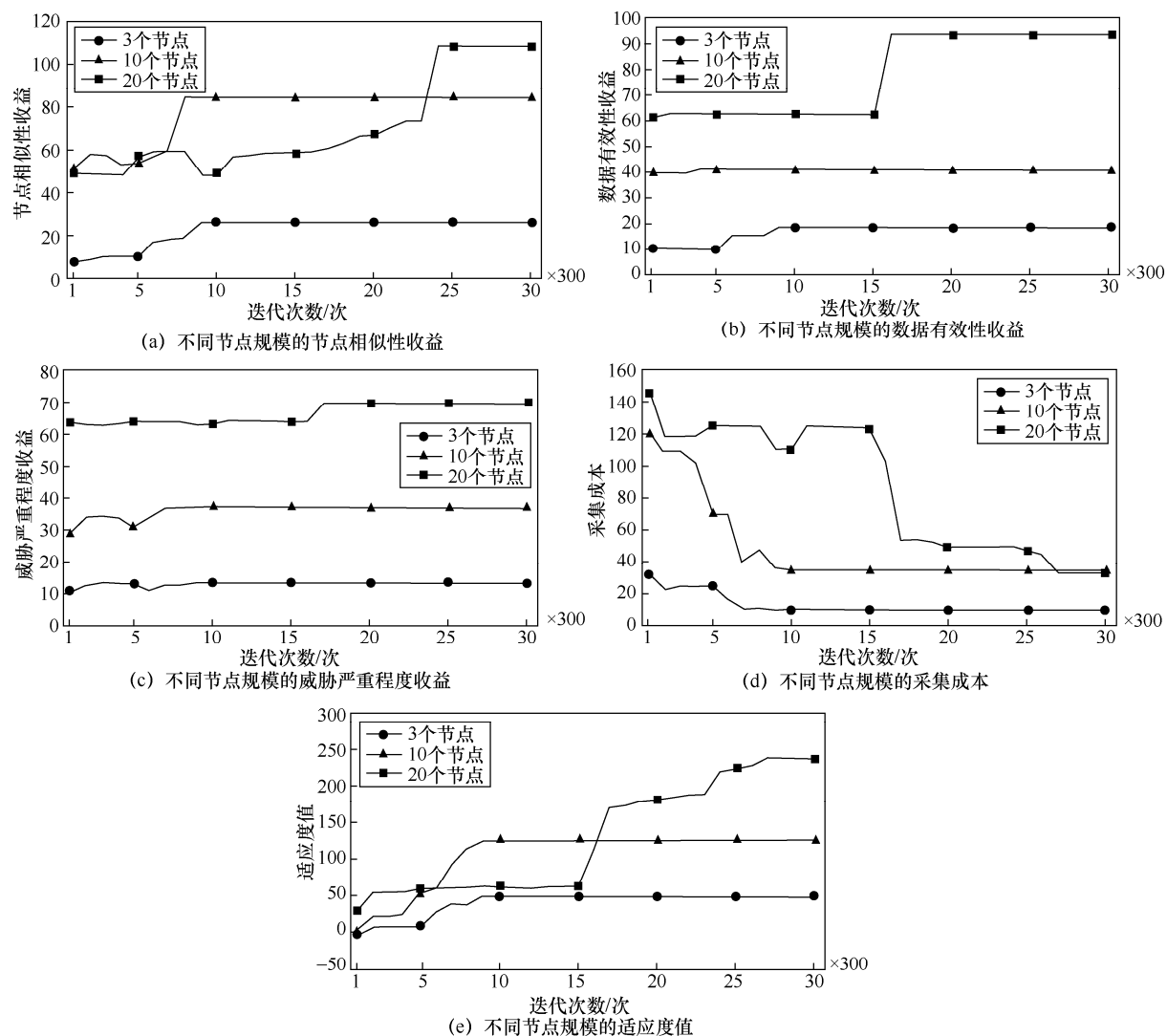


图3 不同规模节点下采集收益和成本以及适应度值的变化趋势
Figure 3 Collection benefits, cost and fitness under different scale nodes

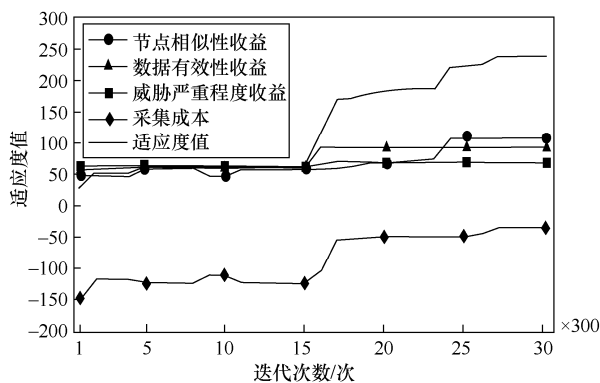


图4 20个节点的采集策略适应度值随迭代次数的变化趋势
Figure 4 Fitness value of the collection policy of 20 nodes with number of iterations

6.2.2 算法时间开销实验

针对 3 个节点、10 个节点和 20 个节点，分别在设定种群数为 10、100 和 1 000 情况下的采集策略生成实验，设置遗传算法的迭代次数为 10 000 次，每类实验进行 5 次后取平均时间，算法时间开销如表 3 所示。

表 3 的结果表明，随着节点数和种群数的增加，消耗的时间也会增加。造成该现象的原因是：节点数量影响了解空间的大小，节点数量越多，需要求解的采集策略参数越多，使得解空间越大和时间开销越大；种群个数影响了算法的收敛速度，种群越大，越容易出现高适应度的个体，使

得收敛速度越慢和时间开销越大。

表3 不同参数的时间开销

Table 3 Time consumption under different parameter

节点数	种群数		
	10	100	1 000
3	1 s	9 s	100 s
10	2.6 s	25.7 s	277 s
20	4.5 s	52 s	600 s

6.2.3 采集效果对比实验

为衡量采集效果,本文主要考虑了采集的数据量和有效性两方面。以全采集策略作为威胁驱动协同采集策略的对比实验,全采集策略需要采集各个节点的所有采集项。

采集数据量指在固定时长内采集到的数据量。4类采集方案在采集数据量方面的对比如图5所示。由图5可知,在开始采集数据后的第6个小时,在部署网站的节点发生DDoS攻击安全事件。图5中的菱形实线代表全采集,各采集频率取基准范围内的最大频率值。短横虚线代表本文所提基于规则关联的协同采集,该方案在开始时随机采集,之后随着威胁状态的变化对发生异常的采集项进行针对性的采集策略调整,使采集的数据量呈缓慢减少趋势。圆形实线表示概率采集,当每次需要采集数据时,通过概率判断是否进行该次采集。三角实线表示根据数据变化幅度自适应采集,该方案根据数据变化幅度调整采集间隔,当数据变化幅度大时,降低采集间隔,反之增加采集间隔。综合图5的整体趋势,可知全采集方案采集的数据量最多,概率采集的数据量与全采集的数据量变化趋势保持一致,根据数据变化幅度自适应采集的数据量也较多,基于规则关联的协同采集在4种采集方案中采集数据量最少。

数据有效性指威胁相关采集数据量与总采集数据量的比例。根据采集数据与当前的威胁状况是否相关,可以判断出采集数据是否有效。数据有效性越高,越容易分析出系统当前正在遭受的威胁。4类采集方案在数据有效性方面的对比实验如图6所示。通过观察可以发现,概率采集和全采集的数据有效性基本相同;根据数据幅度变化自适应采集的数据有效性,与概率采集和全采集相比有所提高;基于规则关联的协同采集的数据有效性高于其他采集方案。

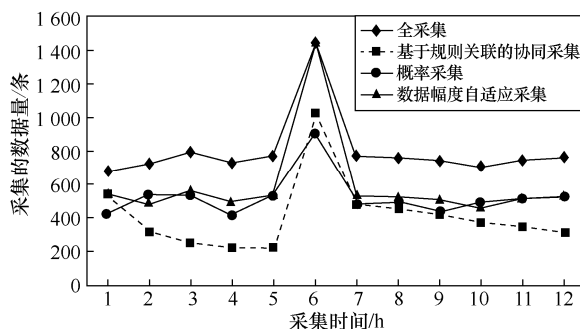


图5 不同采集方案的采集数据量

Figure 5 The amount of collected data of different collection policies

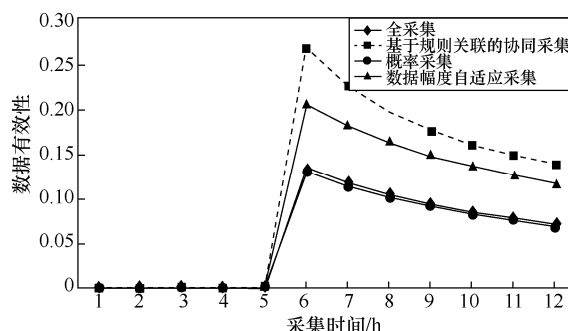


图6 不同采集策略的数据有效性

Figure 6 Data validity of different collection policies

7 结束语

在复杂环境场景下,对大量节点上的海量安全数据进行采集可能存在过度采集和欠采集的问题。为了保证数据采集行为的准确高效,需要生成准确高效的协同采集策略。本文针对该需求,根据威胁、节点资源情况和节点相似度等信息,设计了协同策略生成方案。该方案通过节点间关联规则和安全事件间关联规则,构建备选采集项。然后基于多目标优化,生成数据协同采集策略,为安全威胁的精准分析提供有效输入。

参考文献:

- [1] LIN H Q, YAN Z, CHEN Y, et al. A survey on network security-related data collection technologies[C]//Proceedings of IEEE Access. 2018: 18345-18365.
- [2] XIE H M, YAN Z, YAO Z, et al. Data collection for security measurement in wireless sensor networks: a survey[J]. IEEE Internet of Things Journal, 2019, 6(2): 2205-2224.
- [3] ZHOU D H, YAN Z, FU Y L, et al. A survey on network data collection[J]. Journal of Network and Computer Applications, 2018, 116: 9-23.

- [4] JING X Y, YAN Z, PEDRYCZ W. Security data collection and data analytics in the Internet: a survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(1): 586-618.
- [5] LIU G, YAN Z, PEDRYCZ W. Data collection for attack detection and security measurement in mobile Ad Hoc networks: a survey[J]. Journal of Network and Computer Applications, 2018, 105: 105-122.
- [6] 庞希愚, 姜波, 仝春玲, 等. 一种自适应数据变化规律的数据采集算法[J]. 计算机技术与发展, 2013, 23(2): 157-161.
- PANG X Y, JIANG B, TONG C L, et al. A kind of data acquisition algorithm of adaptive data change rule[J]. Computer Technology and Development, 2013, 23(2): 157-161.
- [7] 杨明霞, 王万良, 邵鹏飞. 基于时间序列的自适应采样机制策略研究[J]. 计算机科学, 2015, 42(7): 162-164, 181.
- YANG M X, WANG W L, SHAO P F. Adaptive sampling algorithm based on TCP congestion strategy[J]. Computer Science, 2015, 42(7): 162-164, 181.
- [8] 曾文序, 库少平, 郑浩. 基于旋转门算法的自适应变频数据采集策略[J]. 计算机应用研究, 2018, 35(3): 769-772.
- ZENG W X, KU S P, ZHENG H. Strategy of self-adaptive frequency conversion data acquisition based on swing door trending algorithm[J]. Application Research of Computers, 2018, 35(3): 769-772.
- [9] LIN H Q, YAN Z, FU Y L. Adaptive security-related data collection with context awareness[J]. Journal of Network and Computer Applications, 2019, 126: 88-103.
- [10] 陈雷. 基于数据聚合的无线传感器网络主动管理机制研究[D]. 长沙: 湖南大学, 2011.
- CHEN L. Research of initiative management scheme for sensor networks based on data aggregation[D]. Changsha: Hunan University, 2011.
- [11] CHEN L W, PENG Y H, TSENG Y C, et al. Cooperative sensing data collection and distribution with packet collision avoidance in mobile long-thin networks[J]. Sensors (Basel, Switzerland), 2018, 18(10): 3588.
- [12] SAFARA F, SOURİ A, SERRIZADEH M. Improved intrusion detection method for communication networks using association rule mining and artificial neural networks[J]. IET Communications, 2020, 14(7): 1192-1197.
- [13] 章坚武, 黄佳森, 周迪. 基于模糊理论与关联规则的入侵检测模型[J]. 电信科学, 2019, 35(5): 59-69.
- ZHANG J W, HUANG J S, ZHOU D. Intrusion detection model based on fuzzy theory and association rules[J]. Telecommunications Science, 2019, 35(5): 59-69.
- [14] 王慧, 王宇, 邵翀. 网络入侵检测中群关联模型的设计与分析[J]. 中国人民公安大学学报(自然科学版), 2018, 24(4): 74-77.
- WANG H, WANG Y, SHAO C. Design and analysis of group association model in network intrusion detection[J]. Journal of People's Public Security University of China (Science and Technology), 2018, 24(4): 74-77.
- [15] HU W, LI J, CHENG J, et al. Security monitoring of heterogeneous networks for big data based on distributed association algorithm[J]. Computer Communications, 2020, 152: 206-214.
- [16] NGUYEN M T. Distributed compressive and collaborative sensing data collection in mobile sensor networks[J]. Internet of Things, 2020, 9: 100156.
- [17] SHA C, SONG D D, YANG R, et al. A type of energy-balanced tree based data collection strategy for sensor network with mobile sink[J]. IEEE Access, 2019, 7: 85226-85240.
- [18] DOUDOU M, DJENOURI D, BARCELO-ORDINAS J M, et al. Cost effective node deployment strategy for energy-balanced and delay-efficient data collection in wireless sensor networks[C]// Proceedings of 2014 IEEE Wireless Communications and Networking Conference (WCNC). 2014: 2868-2873.
- [19] SARNO R, SINAGA F, SUNGKONO K R. Anomaly detection in business processes using process mining and fuzzy association rule learning[J]. Journal of Big Data, 2020, 7: 5.
- [20] 郭涛敏. 基于轻量化关联规则挖掘的安全日志审计技术研究[J]. 现代电子技术, 2019, 42(15): 83-85, 90.
- GUO T M. Research on security log audit technology based on lightweight association rules mining[J]. Modern Electronics Technique, 2019, 42(15): 83-85, 90.
- [21] PAN D W, LIU D T, ZHOU J, et al. Anomaly detection for satellite power subsystem with associated rules based on Kernel Principal Component Analysis[J]. Microelectronics Reliability, 2015, 55(9/10): 2082-2086.
- [22] BÖHMER K, RINDERLE-MA S. Mining association rules for anomaly detection in dynamic process runtime behavior and explaining the root cause to users[J]. Information Systems, 2020, 90: 101438.
- [23] 李风华, 李子孚, 李凌, 等. 复杂网络环境下面向威胁监测的采集策略精化方法[J]. 通信学报, 2019, 40(4): 49-61.
- LI F H, LI Z F, LI L, et al. Collection policy refining method for threat monitoring in complex network environment[J]. Journal on Communications, 2019, 40(4): 49-61.

[作者简介]



陈佩(1993—), 男, 河南南阳人, 中国科学院信息工程研究所硕士生, 主要研究方向为网络与系统安全。



李凤华（1966—），男，湖北浠水人，博士，中国科学院信息工程研究所研究员、博士生导师，主要研究方向为网络与系统安全、信息保护、隐私计算。



郭云川（1977—），男，四川营山人，博士，中国科学院信息工程研究所正高级工程师、博士生导师，主要研究方向为访问控制、网络安全。



李子孚（1992—），女，内蒙古赤峰人，博士，中国科学院信息工程研究所工程师，主要研究方向为网络与系统安全、访问控制。



成林（1983—），男，博士，中国信息安全测评中心助理研究员，主要研究方向为密码学、云计算、大数据。