IP 黑名单关联聚类算法对恶意簇检测的优化研究

刘　云，肖　添
(昆明理工大学信息工程与自动化学院，昆明 650500)

摘　要：互联网中复杂的恶意活动都是由 IP 地址集群共同执行的，通过处理在网络中收集的数据来寻找恶意 IP 簇成为重要的研究方向. 提出一种 IP 黑名单关联聚类算法(IPBACA)，首先，构建 IP-IP 无向图；然后，利用测量统计相关性来测量 IP 黑名单与 IP 的相关性，并使用给定的 IP 黑名单来找到最佳阈值得出 IP 簇，判断其标准化残差是否达标；最后，识别出具有高精度的恶意簇. 仿真结果表明，对比 ICAMO 算法，CAIIB 算法和 DABR 算法，本文提出的 IPBACA 算法在精确率、召回率、F1 指标和归一化互信息等 4 个主要性能指标方面均有明显改善，显著提高了对检测恶意簇的检测能力.

关键词：IP 黑名单；关联聚类算法；恶意簇；IP-IP 无向图
中图分类号：TN929.5　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.013003

Optimization of malicious cluster detection based on IP blacklist association clustering algorithm

LIU Yun, *XIAO Tian*
(School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Complex malicious activities in the Internet are jointly performed by IP address clusters. It has become an important research direction to find malicious IP clusters by processing data collected in the network. An IP blacklist association clustering algorithm (IPBACA) is proposed in the paper, in which first constructs an IP-IP undirected graph, and then uses measurement statistical correlation to measure the correlation between IP blacklist and IP, and uses the given IP blacklist to find the best threshold worthy of malicious clusters, and judges its standardized residuals whether it is up to standard, it finally identifies a malicious cluster with high precision. The simulation results show, compared with ICAMO algorithm, CAIIB algorithm and DABR algorithm, the IPBACA algorithm proposed in this paper has a significant improvement in the four main performance indicators of precision, recall, F1 and normalized mutual information, and significantly improves the detection ability of malicious clusters.

Keywords: IP Blacklist; Association Clustering Algorithms; Malicious Cluster; IP-IP undirected graph

1　引　言

互联网上的许多恶意行为已经演变成由多组 IP 地址共同执行的非常复杂的操作. 电子垃圾邮件、分布式密码猜测攻击和恶意软件分发网络是此类攻击的一些例子. 攻击者通常使用一组 IP 地址

收稿日期：2019-09-16
基金项目：国家自然科学基金(61761025)
作者简介：刘云(1973—)，男，云南昆明人，副教授，研究方向为无线通信研究. E-mail：liuyun@kmust.edu.cn
通讯作者：肖添. E-mail：547848098@qq.com

登录到受攻击的网络帐户以执行各种恶意任务[1-3]. 这种来自一组 IP 地址的集体行为通常会在网络的各个位置留下痕迹,使得防御者能够使用数据分析技术将这些 IP 地址集群连接在一起.

Stringhini 等人提出基于模块优化的迭代聚类算法(Iterative Clustering Algorithm Based on Modularity Optimization, ICAMO)[1]. 通过聚合数月的数据,能可靠地识别恶意的帐户群集. Mathur 等人提出了基于聚类的内边界推断算法(Clustering-Based Approach to Infer Internal Boundaries, CAIIB)[4]. 通过计算 IP 地址在 IP 地址空间上的接近度,并使用 IP 黑名单识别出潜在的恶意簇. Arya 等人提出了一种基于动态属性的声誉算法(Dynamic Attribute Based Reputation, DABR)[5],从已知恶意 IP 地址中提取数据生成声誉等级,并根据阈值识别恶意 IP 地址.

ICAMO 算法未能找到一个合适的阈值删除 IP 簇中松散连接的分支,导致簇中 IP 地址数过大. CAIIB 算法并没有考虑黑名单质量的影响,且只在聚类完成后才使用 IP 黑名单. DABR 算法由于生成声誉等级使用的特征较少,导致算法只能应用于特定网络. 在 ICAMO 算法,CAIIB 算法和 DABR 算法的研究基础上,本文提出了 IP 黑名单关联聚类算法(IP Blacklist Association Clustering Algorithm ,IPBACA),通过一种新的聚类框架从网络交互的数据集中识别执行恶意任务的 IP 簇. 首先利用定义的相似性度量构建 IP-IP 无向图,然后,利用测量统计相关性来测量 IP 黑名单与 IP 的相关性,并使用给定的 IP 黑名单来找到最佳阈值删除相关性弱的边缘得出 IP 簇,再判断 IP 簇的标准化残差是否大于 3,最终得出与 IP 黑名单高度相关联的恶意簇. 从数学上证明了即使是质量普通的黑名单也可以用来精确地检测恶意簇. 仿真结果表明,即使是一个精度普通的 IP 黑名单也足以使所提出的算法准确地识别恶意 IP 簇,对比 ICAMO 算法,CAIIB 算法和 DABR 算法,IP 黑名单关联聚类算法在精确率、召回率、F1 指标[6]和归一化互信息[7]等 4 个主要性能指标方面均有明显改善,证明 IPBACA 算法显著提高了检测恶意簇的整体能力.

2　IP 黑名单关联聚类

2.1　IP 黑名单关联聚类框架

图 1 是 IP 黑名单关联聚类框架图,利用恶意网络帐户在论坛上发布垃圾评论、制造垃圾邮件等恶意任务[8-10]展示了不同的场景,IP 地址可以通过数据分析方法连接在一起,共同执行一个任务. 在两个 IP 地址之间定义一个适当的相似性度量,就可以将给定的数据集表示为无向图. 构建无向图之后,提取 IP 地址簇实质上是在区分簇边缘和噪声边缘. 为解决此问题,提出了一个聚类方法,如下所示.

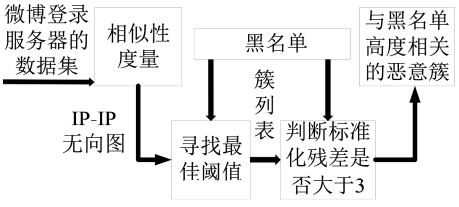

图 1　IP 黑名单聚类框架图

Fig. 1　IP blacklist clustering framework

(1) 从图中删除所有“弱边”,其中“弱边”被定义为权重小于阈值的边.

(2) 将结果图的连接部分输出为 IP 簇.

为选择一个合适的阈值来产生有意义的簇,本文通过使用 IP 黑名单来找到最佳阈值. 选择的最佳阈值是为了最大化黑名单和由聚类过程产生的恶意簇之间的统计相关性. 因为 IP 簇和黑名单之间的相关性越强,簇恶意的证据就越强. 所以通过最大限度地提高这种统计相关性,可以使所提出的聚类方案输出具有最强统计证据的恶意簇.

2.2　测量统计相关性的定义

定义 1　测量统计相关性:假设给出了一个黑名单“B”和一个数据集“N”. 为了测量从数据集“N”中提取的 IP 簇“C”与黑名单“B”之间的依赖关系,首先为数据集中的每个 IP 地址定义以下具有二进制结果的事件对.

(1) 事件 1:数据集 N 中的 IP 地址在 IP 簇 C 中.

(2) 事件 2:数据集 N 中的 IP 地址在黑名单 B 中.

零假设下 IP 簇“C”是良性的,且事件 1 和事件 2 是独立的. 但是对于恶意簇,希望这些事件具有可测量的相关性. 通过计算在零假设下的标准化残差,测量这些事件之间的相关性,本质上衡量了在零假设下,簇 C 中 IP 地址在黑名单 B 中的频率[6]. 因此,越高的标准化残值表明事件具有更强的相关性. 标准化残差定义为

$$R=\frac{n-\hat{\mu}}{\hat{\mu}(1-p_1)(1-p_2)} \tag{1}$$

其中，n 是观察到的这些事件共同发生的次数；$\hat{\mu}$ 是这些事件共同发生的预期次数，并且 p_1，p_2 分别是事件 1 和事件 2 的概率.

零假设表示集群 C 是良性的，事件 1 和事件 2 是独立的. 因此在零假设下，IP 地址在集群 C 和黑名单 B 中的预期次数如下式所示.

$$\hat{\mu}=|N|\frac{|C|}{|N|}\frac{|B|}{|N|}=\frac{|C||B|}{|N|}$$

其中，N 表示数据集中的 IP 地址数；B 表示数据集中黑名单 IP 地址数；C 表示集群 C 的大小，IP 地址在 C 中的概率写为 $p_1=C/N$，IP 地址在黑名单 B 中的概率为 $p_2=B/N$. 将这些插入等式(1)中，可以将有 n 个 IP 地址在黑名单 B 中的集群 C 的标准化残差计算为

$$R=\frac{n-\frac{|C||B|}{|N|}}{\sqrt{\frac{|C||B|}{|N|}(1-\frac{|C|}{|N|})(1-\frac{|B|}{|N|})}} \tag{2}$$

3　IP 黑名单关联聚类算法

3.1　IP 黑名单关联聚类算法说明

IP 黑名单关联聚类算法主要包括预处理、寻找最佳阈值以及利用 IP 黑名单识别恶意簇三步.

在两个 IP 地址之间定义一个适当的相似性度量，就可以将给定的数据集表示为无向图. 相似性度量 SM(Similarity Measure)定义如下.

$$SM=\alpha \tag{3}$$

其中，α 是两个 IP 地址在一天内登录相同帐户的数量，即边缘权重. 由于动态 IP 地址分配、主机被清理等原因，IP 地址在恶意和非恶意之间频繁切换[11-13]，因此识别出的恶意 IP 地址簇可能很快变得不活动或不再是恶意的. 为此，对每天收集的数据集分析，而不像以前的工作对几个月的聚合数据进行分析.

IPBACA 算法中 IP-IP 无向图的推理图如图 2 所示.

1) 从节点(IP 地址)构造完整的(即完全连通的)无向图.

2) 利用 SM 定义得出所有边缘的权重 α，并删除 α 为零的边，构建 IP-IP 无向图.

3) 基于得出的最佳阈值，将节点间 α 小于最佳阈值的边缘除去，从而获得 IP 簇，IPBACA 算法步骤如下.

步骤 1　预处理过程如图 2(a)和(b)所示，其中，节点表示 IP 地址，两个节点之间边缘的权重表示对应 IP 地址之间的相似性度量值，在 IP-IP 无向图上可以得出两种边. 一种是由于 IP 地址簇共同行为而存在的边缘；另一种是由于各种原因和随机事件产生的大量噪声边缘. 因为恶意簇的边缘代表具有集体恶意行为的关系，所以期望恶意簇边缘的权重比噪声边缘大很多.

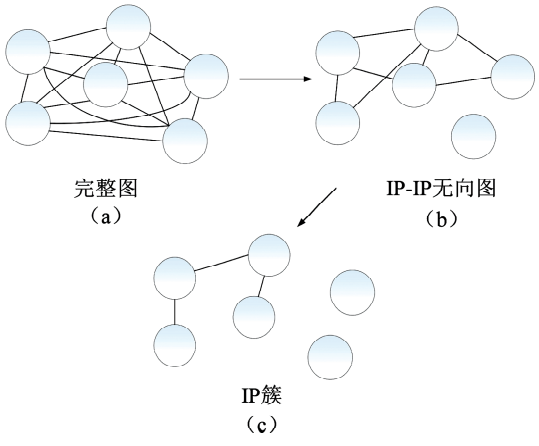

图 2　IPBACA 算法中 IP-IP 无向图的推理图
Fig. 2　Inference graph of IP-IP undirected graph in IPBACA algorithm

步骤 2　寻找最佳阈值是通过利用统计相关性的度量，为所提出的聚类方案选择最佳阈值. 为了最大限度地证明这些簇是恶意的，聚类方案需要生成的恶意 IP 簇具有最高的标准化残差. 同时，因为使用较大的簇区分恶意和良性簇要更加准确，所以不希望聚类方案产生较小的簇. 通过最大化所有簇的平均标准化残差，可以满足以上要求. 这个目标函数如下式.

$$\beta=\frac{1}{k}\sum_{i=1}^{k}R_i \tag{4}$$

式中，k 为簇数；R_i 为式(2)中给出的第 i 个簇的标准化残差. 对于良性簇，因为方程(1)中的零假设下 $E[n]\hat{\mu}$，标准化残差预计为零，所以良性簇对目标函数没有贡献. 通过最大化平均标准化残差，并不是首先明确地识别恶意簇而是对它们进行优化. 通过最大化 β，强制恶意簇拥有尽可能高的标准化残差，且由于分母中的 k 可以使产生的簇较少从而减少碎片.

阈值和平均标准化残差的关系非常复杂，虽然可使用如梯度上升法等数值方法求解，但最终解决方案是通过对一系列可能的阈值进行穷尽搜索来找到最佳阈值. 在此情况下，穷尽搜索是最实际的解决方案，寻找最佳阈值的伪代码如算法 1 所示.

算法 1：寻找最佳阈值算法

输入：IP-IP 图(G)，候选阈值列表(T)，$t^* \leftarrow 0$，$\beta^* \leftarrow 0$

输出：t^*：最佳阈值

Begin

1) for $t \in T$ do

2)　　$G' \leftarrow$将 G 中权重小于 t 的边缘去除

3)　　计算 G'的平均标准化残差 β

4)　　if $\beta > \beta^*$ then

5)　　$\beta^* \leftarrow \beta$

6)　　$t^* \leftarrow t$

7)　　end

8)　end

9) 找到最佳阈值 t^*

End

上述过程找到给定数据集的最佳阈值，就应用所提出的聚类方案来获得 IP 簇列表如图 2(c)所示，但这些 IP 簇并非所有都是恶意的. 下面将介绍如何使用给定的 IP 黑名单进一步识别 IP 簇列表中的恶意簇.

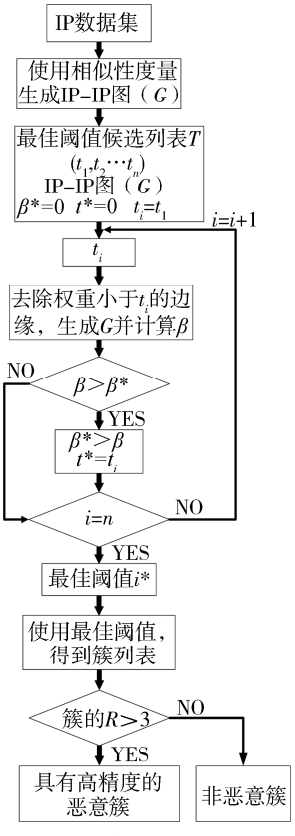

图 3　IP 黑名单关联聚类算法流程图

Fig. 3　Flow chart of IP blacklist association clustering algorithm

步骤 3　图 3 是 IP 黑名单关联聚类算法流程图的最后部分. 根据之前测量统计相关性的定义，数据集 N 中簇 C 的标准化残差(R)表示 C 是恶意簇的证据强度. 由于 R 是通过标准误差归一化的，所以 $R = r$ 表示观察到的这两个事件共现的次数在零假设下偏离其预期值的标准偏差为 r. 因此，$R>3$被认为是两个事件相关的非常有力的证据，因为在零假设下偶然地观察这一事件的概率小于 0.3%[14]. 因此，为确定 IP 地址簇是否是恶意的，使用式(2)计算簇的标准化残差，判断其标准化残差是否大于 3，若是则声明簇是恶意的. 从而利用 IP 黑名单识别出具有高度准确率的恶意簇.

3.2　IP 黑名单质量对算法影响评估

为分析 IP 黑名单的质量是如何影响标准化残差，从而影响算法的检测精确率，下面具体研究各种黑名单质量的标准化残差的预期值. 簇的大小也会影响检测精确率，它在一定程度上取决于最佳聚类方案中使用的阈值. 虽然在这一过程中可能存在一定程度的噪声，但是为了研究黑名单质量的影响，假设最佳聚类方案能够从数据中完美地提取 IP 簇. 为描述黑名单的质量，将黑名单的真阳性率定义为

$$\mathrm{TPR}=\Pr(\text{IP is in blacklist B}\mid\text{IP is Malicious})$$

黑名单的假阳性率为

$$\mathrm{FPR}=\Pr(\text{IP is in blacklist B}\mid\text{IP is Benign})$$

良性和恶意的 IP 地址都有可能是误报的，但为简单起见，假设假阳性和真阳性是独立且同分布.

如果簇 C 是良性的，那么 C 中列入黑名单的 IP 地址是由于误报造成的，因此无论 TPR 和 FPR 为何值，标准化残差的预期值都是 $E[R]=0$，并且 C 中是黑名单 IP 地址的预期数量如下.

$$E[n]=\hat{\mu}=|C|\times\mathrm{FPR}$$

如果簇 C 是恶意的，那么 C 中是黑名单 IP 地址的预期数量将是

$$E[n]=|C|\times\mathrm{FPR}$$

由于数据集 N 中的良性 IP 远比恶意 IP 多，因此数据集 N 中的 IP 地址处于黑名单 B 中的概率约等于黑名单的假阳性率(即 $p_2=\mathrm{FPR}$). 因此，将这些与方程(2)相结合，可以将标准化残差的期望值写为

$$E[R]=\frac{|C|(\mathrm{TPR}-\mathrm{FPR})}{\sqrt{(|C|\mathrm{FPR})(1-\frac{|C|}{|N|})(1-\mathrm{FPR})}} \quad (5)$$

由上式可得，当簇是恶意时，预期的标准化残差会随着黑名单的真阳性率和假阳性率之间差值的增加而增加，也会随着簇的大小增加而增加. 即

(1) 如果使用更准确的 IP 黑名单，可以更准确地识别恶意 IP 簇.

(2) 较大的簇比较小的簇具有更准确的识别度.

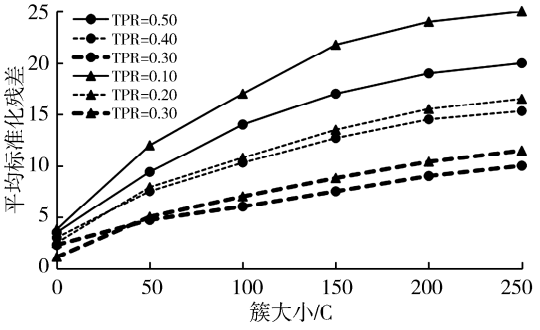

图 4　不同黑名单真、假阳性率和不同簇大小的预期标准化残差的数值分析

Fig. 4　Numerical analysis of expected normalized residuals with different blacklist true and false positive rates and different cluster sizes

为证明黑名单并不一定要非常精确才能准确地识别恶意集群. 设数据集的大小 $N=1\times10^5$，图 4 表现了不同簇大小和不同黑名单真、假阳性的预期标准化残差.

为研究真阳性率的影响，如图 4 圆形标记点所示，将黑名单假阳性率设置为 10%，观察到低真阳性率，例如黑名单真阳性率为 40%，预期标准化残差也很快超过临界值 3. 因此，一个具有 10%假阳性率和 40%真阳性率的普通黑名单可以用来准确识别大小大于 9 的大多数恶意簇. 此外，即使是真阳性率=30%和假阳性率=10%的质量非常差的黑名单，也可以用来准确识别恶意簇，只要簇的大小大于 80.

另一方面，为了研究假阳性率的影响，如图 4 三角形标记点所示，将黑名单真阳性率设置为 60%，观察到低假阳性率，例如 FPR=10%，黑名单可用于准确识别大小最小为 5 的恶意簇. 对于较大的假阳性率，如 FPR=30%，恶意簇大小必须大于 20 就使用黑名单可靠地识别恶意簇.

图 4 都显示了标准化残差的预期值，上面的论点是在预期意义上提出的. 使用黑名单检测恶意簇的实际概率等于标准化残差大于 3 的概率. 利用方程(1)，这个概率可以写为

$$\Pr[R>3]=\Pr[n>\hat{\mu}+3\sqrt{\hat{\mu}(1-p_1)(1-p_2)}]$$

与式(4)组合后，相当于

$$\Pr[n>|C|\mathrm{FPR}+3\sqrt{(|C|\mathrm{FPR})(1-\frac{|C|}{|N|})(1-\mathrm{FPR})}]$$

这个概率可以计算出来，n 是簇 C 中列入黑名单的 IP 地址数量，并且是二项式分布的，例如：

$$\Pr(n)=\binom{|C|}{n}(\mathrm{TPR})^n(1-\mathrm{TPR})^{|C|-n}$$

利用这些方程，设簇大小为 50，绘制了图 5 中各种黑名单真、假阳性率正确检测恶意簇的概率(即 $\Pr[r>3]$). 由图 5 可知，更好的黑名单产生更准确的结果. 例如，一个 TPR=50%和 FPR=20%的普通黑名单检查的准确率约为 91%.

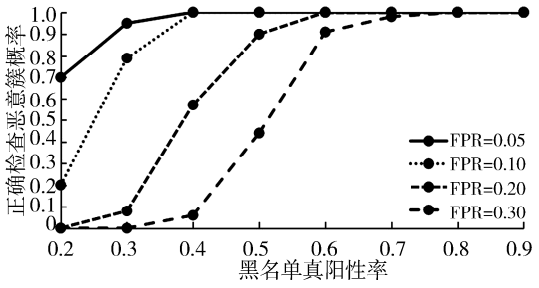

图 5　针对不同的黑名单真、假阳性率正确检测恶意簇的概率

Fig. 5　Probability of correctly detecting malicious clusters for different blacklist true and false positive rates.

4　仿真分析

4.1　数据集及评价指标

与类似论文一致，为了验证所提出的算法，选取通用数据集，即微博的登录服务器上收集到真实登录事件的数据集[15]，并根据查询 Spamhaus[16] 得到 IP 黑名单.

使用在连续 14 d 内观察到的每个登录事件的 IP 地址和匿名帐户 ID，且只考虑通过 SMTP 或 IMAP 协议以及桌面浏览器成功登录的事件. 在公有云平台上使用具有 2 GHz 64 位 QEMU 虚拟 CPU 的虚拟主机，并使用 networkx python library[17] 提取连接的组件.

为测量聚类性能，文中选用精确率(Precision)、召回率(Recall)、F_1 指标和归一化互信息(NMI)作为聚类算法评价指标，其定义如下.

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}.$$

TP 表示正确判定属于此簇的 IP 数;FP 表示错误的判定属于此簇的 IP 数;FN 表示错误判定不属于此簇的 IP 数,将精确率和召回率相结合构成了 F_1 指标,更全面的对聚类性能进行评价.

归一化互信息: $\text{NMI}=\dfrac{I(X:Y)}{[H(X)+H(Y)]/2}$

其中,$H(X)$是 X 的熵,$I(X:Y)$是 $H(X)$和$H(Y)$之间的互信息量.

4.2　寻找最佳阈值

对于给定的一天,首先构建 IP-IP 图,其中节点是 IP 地址,如果相应的 IP 地址用于在当天至少登录一个公用帐户,则两个节点之间有一个边缘.登录的普通帐户的实际数量由边缘权重表示.在删除独立节点(即没有边缘的 IP 地址)之后,一天生成的图有超过 50 万个节点和 160 万个边缘.

建立了 IP-IP 图,就可以找到从图中提取 IP 簇的最佳阈值.为找到最佳阈值,计算方程(4)中给定的一系列阈值的目标函数,并选择了使目标函数最大化的最佳阈值.在找到最佳阈值后,从图中去除所有权重小于最佳阈值的边缘.最后,将结果图中连接的部分输出为 IP 簇.忽略大小小于 5 的集群,因为无法准确计算此类小集群的标准化残差.在移除这些小集群之后,通常每天会得到几百个簇.

找到最佳阈值的实际 CPU 时间因可用计算资源的不同而有很大的差异,典型一天的数据在(0,30)范围内的最佳阈值进行单线程搜索大约需要 60 s.

为描述这个优化过程,在图 6 中绘制了一个典型日期的阈值范围的目标函数.

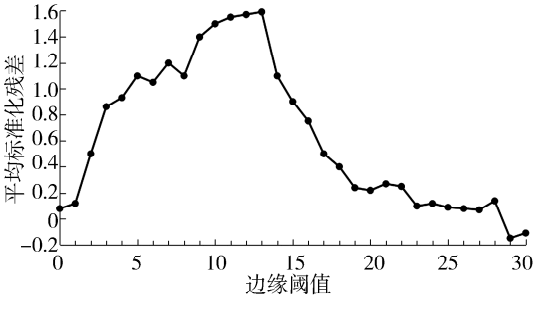

图 6　寻找最佳阈值

Fig. 6　Finding the best threshold

如图 6 所示,在这一天的最佳阈值是 13,这意味着一对 IP 地址必须登录 13 个以上相同的帐户才能连接到图上,从而位于同一个簇.

4.3　算法对比仿真分析

在使用这些阈值执行最佳聚类之后,将标准化残差大于 3 的簇声明为每天得到的恶意簇.为仿真对比,将所提 IPBACA 算法与 ICAMO 算法,CAIIB 算法和 DABR 算法进行对比,用精确率、召回率、F_1 指标和归一化互信息等 4 个主要性能指标来评估四种算法的性能.

如图 7 所示,所提出的 IPBACA 算法在 14 d 内平均精确率最高并且大幅领先其余 3 个算法,DABR 算法的平均精确率接近 40%,而 ICAMO 算法和 CAIIB 算法精确率相近,其中 CAIIB 算法表现最差.

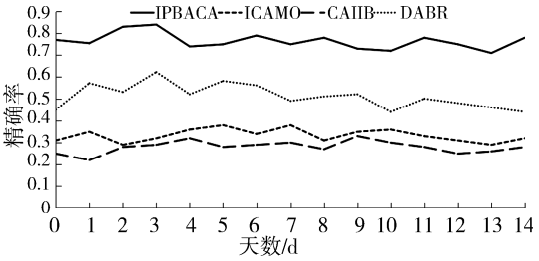

图 7　4 种不同算法的精确率对比图

Fig. 7　Comparison chart of precision of four different algorithms

如图 8 所示,IPBACA 算法平均召回率最高,DABR 算法次之,ICAMO 算法和 CAIIB 算法表现相近,CAIIB 算法召回率最低.

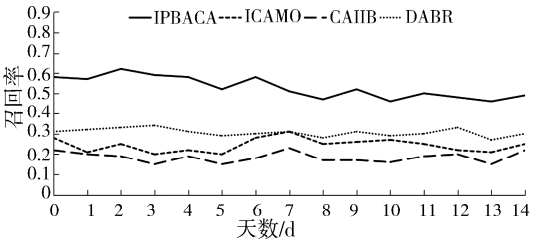

图 8　4 种不同算法的召回率对比图

Fig. 8　Comparison chart of recall rate of four different algorithms

如图 9 所示,IPBACA 算法的平均 F_1 指标最高,ICAMO 算法和 CAIIB 算法表现相近,其中 CAIIB 算法 F_1 指标最低.

如图 10 所示,IPBACA 算法的平均归一化互信息最高,DABR 算法高于 ICAMO 算法,而 CAIIB 算法低于以上 3 种算法.

通过手动观察检测出的簇时,IPBACA 算法由于删除了 IP 簇中松散连接的分支,输出了与黑名单高度相关联的核心结构. ICAMO 算法和

CAIIB 算法不涉及这样的修剪，从而产生具有更多 IP 地址的簇，其中一些 IP 地址的连接相当松散. 而 DABR 算法也使用阈值用于修剪，在一定程度上使性能有所提升.

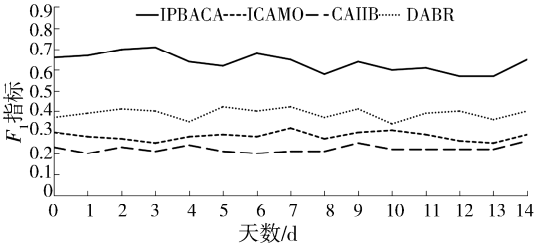

图 9　4 种不同算法的 F_1 指标对比图

Fig. 9　Comparison chart of F_1 indicators of four different algorithms

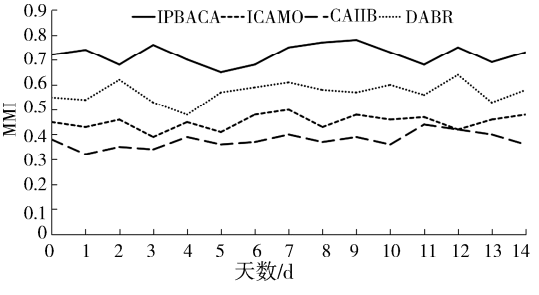

图 10　4 种不同算法的 NMI 对比图

Fig. 10　Comparison of the NMI of four different algorithms

4.4　IP 黑名单质量对算法影响评估

通过仿真实验，逐步降低黑名单的质量并测量算法的检测性能.

为降低黑名单的质量，首先在数据集中找到一天内所有黑名单 IP. 然后从黑名单中删除这些黑名单 IP 的某些部分. 为保持黑名单 IP 的数量不变，从整个数据集中随机选择相同数量的 IP，并将其添加到黑名单中. 这样就大大降低了黑名单的真实阳性率. 另一方面，由于随机选择的 IP 地址的数量远小于数据集中所有 IP 地址的数量，因此假阳性率不会受到太大影响，只会增加百分之几. 由于实验有随机成分，为了平均出不可控因素，重复相同的实验 25 次，并得出平均值. 在图 11 中绘制了使用不同黑名单腐败率时，算法在精确率和召回率的变化. 如图可知，随着更多黑名单被删除，召回率逐步下降，但精度基本保持在 75%～80%，直到 80%的黑名单 IP 被删除，精确率才大幅度下降. 尽管根据黑名单的质量变低，IPBACA 算法可能会遗漏一些恶意簇，但它检测到的那些簇很可能是恶意的，证明了第 3 节中数学理论分析是合理.

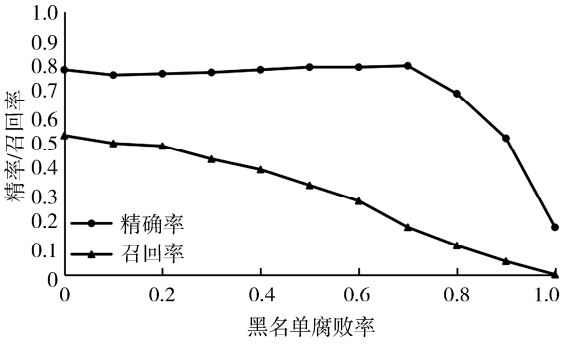

图 11　黑名单腐败率对算法精确率与召回率的影响

Fig. 11　Influence of blacklist corruption rate on algorithm accuracy and recall rate

5　结　论

为了发现在互联网中执行复杂恶意活动的恶意 IP 地址簇，本文提出一种 IP 黑名单关联聚类算法(IPBACA)，通过一种新的聚类框架从网络交互的数据集中识别执行恶意任务的 IP 地址簇，首先构建 IP-IP 无向图，然后利用测量统计相关性来测量 IP 黑名单与 IP 的相关性，并使用给定的 IP 黑名单来找到最佳的阈值得出恶意簇，再判断其标准化残差是否达到标准，最终识别出高精度的恶意簇得出结果. 仿真结果表明，即使是一个普通精度的黑名单也足以使所提出的方案准确识别恶意 IP 地址簇，对比 ICAMO 算法，CAIIB 算法和 DABR 算法，IP 黑名单关联聚类算法在精确率、召回率、F_1 指标和归一化互信息等 4 个主要性能指标方面均有明显改善，证明 IP 黑名单关联聚类算法显著提高了检测恶意簇的整体能力.

参考文献：

[1]　Stringhini G, Mourlanne P, Jacob G, *et al*. EvilCohort: detecting communities of malicious accounts on online services [C]// USENIX Security Symposium. USA: USENIX Association, 2015.

[2]　杨可心，桑永胜. 基于 BP 神经网络的 DDoS 攻击检测研究[J]. 四川大学学报：自然科学版，2017，54：71.

[3]　高强，林星辰，林宏刚，等. 安全协议抗 DoS 攻击的形式化分析研究[J]. 四川大学学报：自然科学版，2018，55：85.

[4]　Mathur S, Coskun B, Balakrishnan S. Detecting hidden enemy lines in IP address space [C]// Proceedings of the 2013 New Security Paradigms Workshop. New York, USA: ACM, 2013.

[5]　Renjan A, Joshi K P, Narayanan S N, *et al*.

DAbR: dynamic attribute-based reputation scoring for malicious IP address detection [C]// Proceedings of the 2018 IEEE International Conference on Intelligence and Security Informatics (ISI). Miami, FL, USA: IEEE, 2018.

[6] 肖锦琦, 王俊峰. 基于模糊哈希特征表示的恶意软件聚类方法[J]. 四川大学学报:自然科学版, 2018, 55: 51.

[7] Pei Q J, Lin M, Li Z Z, *et al*. News Feature Extraction for Events on Social Network Platforms [C]// Proceedings of the International Conference on World Wide Web Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017.

[8] Iosup A, Uta A, Versluis L, *et al*. Massivizing computer systems: a vision to understand, design, and engineer computer ecosystems through and beyond modern distributed systems [C]// Proceedings of the 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). Vienna, Austria: IEEE, 2018.

[9] Yang C, Zhang J, Gu G. Understanding the market-level and network-level behaviors of the android malware ecosystem[C]// Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). Atlanta, GA, USA: IEEE, 2017.

[10] Mikhalkova E, Karyakin Y, Glukhikh I. Large scale retrieval of social network pages by interests of their followers [C]// Proceedings of the International Conference on Computational Science. Wuxi, China: Springer, Cham, 2018.

[11] Kim J. How did the information flow in the # alphago hashtag network? a social network analysis of the large-scale information network on twitter[J]. Cyberpsychol Behav Soc Netw, 2017, 20: 746.

[12] Jerkins J A. Motivating a market or regulatory solution to IoT insecurity with the Mirai botnet code [C]// Proceedings of the IEEE Computing & Communication Workshop & Conference. Atlanta, GA, USA: IEEE, 2017.

[13] 缪振龙. 解决局域网 IP 地址冲突故障[J]. 网络安全和信息化, 2019, 37: 147.

[14] Withers S D. Categorical data analysis[J]. Technometrics, 2017, 33: 241.

[15] Shen Y, Tang Y, Dou Q. The discovery of microblog users[C]// Proceedings of the International Conference on Data Science & Business Analytics. Changsha, China: IEEE Computer Society, 2018.

[16] Wolff J, Braman S. 8 operation stophaus: the spamhaus denial-of-service attacks[J]. Leg Econ Aftermath of Cybersecurity Breaches, 2018, 15: 145.

[17] Debie P, Wang W, Bromuri S. A python library for memory augmented neural networks[C]//Proceedings of the 2018 IEEE International Conference on Web Intelligence. Santiago, Chile: IEEE, 2018.

引用本文格式:

中 文: 刘云, 肖添. IP 黑名单关联聚类算法对恶意簇检测的优化研究[J]. 四川大学学报: 自然科学版, 2021, 58: 013003.

英 文: Liu Y, Xiao T. Optimization of malicious cluster detection based on IP blacklist association clustering algorithm [J]. J Sichuan Univ: Nat Sci Ed, 2021, 58: 013003.