第 43 卷 第 2 期 计 算 机 工 程 2017 年 2 月
Vol. 43 No. 2 Computer Engineering February 2017

·安全技术· 文章编号：1000-3428(2017)02-0194-07 文献标志码：A 中图分类号：TP309

基于 LDA 模型的海量 APT 通信日志特征研究

孙名松，韩 群

（哈尔滨理工大学 软件学院，哈尔滨 150060）

摘 要： 为实现高级持续性威胁（APT）的通信检测，提出一种对服务器端和主机端日志数据的检测方法。通过建立 IP 地址数据库，采用 DBSCAN 聚类算法对海量日志数据进行收集和处理得到异常通信日志。利用高级持续性威胁 14 种通信特征的隐含狄利克雷分布（LDA）建模对异常通信日志进行检测。实验结果表明，与潜在语义分析和概率潜在语义分析检测模型相比，LDA 建模提高了 APT 通信检测的效率和准确度。

关键词： 高级持续性威胁；大数据处理；IP 规范；DBSCAN 算法；特征描述

中文引用格式： 孙名松，韩 群. 基于 LDA 模型的海量 APT 通信日志特征研究[J]. 计算机工程，2017，43(2)：194-200，205.

英文引用格式： Sun Mingsong，Han Qun. Research on Massive APT Communication Log Characteristic Based on LDA Model[J]. Computer Engineering，2017，43(2)：194-200，205.

Research on Massive APT Communication Log Characteristic Based on LDA Model

SUN Mingsong，HAN Qun

（College of Software，Harbin University of Science and Technology，Harbin 150060，China）

【Abstract】 In order to realize the communication detection of the Advanced Persistent Threat（APT），this paper presents a detection method for server-side and host-side log data. It makes the establishment of IP address database and uses DBSCAN clustering algorithm to collect and deal with the massive log data to get abnormal communication log. The abnormal communication log is detected by using Latent Dirichlet Distribution（LDA） modeling of the 14 communication features of APT. Experimental results show that LDA modeling improves the efficiency and accuracy of APT communication detection compared with Latent Semantic Analysis（LSA） and Probabilistic Latent Semantic Analysis（PLSA） detection models.

【Key words】 Advanced Persistent Threat（APT）；big data processing；IP specification；DBSCAN algorithm；characteristic description

DOI： 10.3969/j.issn.1000-3428.2017.02.032

0 概述

高级持续性威胁（Advanced Persistent Threat，APT）[1]，具有组织严密、高隐秘性、持续时间长、攻击手段多样等特点。近几年 APT 攻击呈现持续增长的趋势。因此，在 APT 攻击尚未形成气候的时候需要对其检测技术不断的研究与创新，利用大数据对其检测无疑是最好手段与方法。

高级持续性威胁攻击具有典型的不同于普通攻击的特征[2]。APT 攻击具有很强的欺骗性，利用收集选定目标的社会公共关系、个人喜好、网络信息等，设置陷阱，引诱选定目标浏览不受信任的网站、下载恶意代码程序或可疑文件，最终实现对目标主机的捕获。操作系统中的漏洞，根据我国漏洞数据库（CNVD）的统计[3]，对各类系统和软件具有极大威胁的高危漏洞占总体的 80%，漏洞遍布各类操作系统、数据库、可执行程序、网络交换设备、安全防御产品、网页应用等多个领域。资源分析方法，为了能够更精准地获得更多选定目标的信息，为以后设计更加周密、有针对性的攻击方案，编写合适的攻击代码，诱使选定目标上当。为了增加攻击成功率，减少被各种入侵检测系统及安全软件发现并屏蔽的概率，攻击者经常用到的如动态域名解析来实现命令与控制服务器的长期隐藏和潜伏。2010 年 Google “极光”攻击是一个十分著名的 APT 案例。Google 内部中断被未知恶意程序渗入数月，攻击者持续监

作者简介：孙名松（1963—），男，教授，主研方向为网络安全；韩 群，硕士研究生。

收稿日期：2015-12-14 修回日期：2016-03-15 E-mail：415550620@qq.com

听并最终成功渗透进入 Google 的邮件服务器,不断获得特定 Gmail 账户的邮件内容并且窃取各种系统的数据。2010 年爆发的“超级工厂”,利用了微软操作系统中至少 4 个漏洞,其中有3 个全新的 0day 漏洞为衍生的驱动程序提供有效的数字签名,通过一套完整的入侵和传播流程突破工业专用局域网的物理限制展开攻击。2013 据路透社援引卡巴斯基著名安全专家报道,“震网”病毒可能已感染国际空间站。2011 年发生的“夜龙”攻击,使得 EMC 公司下属的 RSA 公司遭受入侵,部分 Secure ID 技术及客户资料被窃取。

因此,针对高级持续性威胁[4]攻击的检测体系,需要重新思考安全防御的技术和方法。鉴于该威胁所具有的持续性和隐蔽性特点,本文提出一种基于隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)模型的海量 APT 通信日志特征研究方法。利用收集到的服务器端和主机端的海量日志,对通信行为进行分析,并结合高级持续性威胁的通信特点找出异常通信行为。

1　海量日志数据的收集和处理

应用日志数据检测高级持续性威胁异常通信是目前最有效的方法。日志来自于大量网络设备上的,包括记录网络代理的每一个出站连接,记录 DHCP 服务器每一个动态 IP 的分配,记录 VPN 和企业网络的远程连接,记录窗口域控制器试图验证的记录,杀毒软件在终端扫描恶意软件的结果。由于每天服务器端和主机端都产生了庞大的事件记录相关的日志,因此如何处理大数据的问题是其中一个大的难点。第 2 个难点在于大多数网络设备只记录终端的 IP 地址,并且,IP 地址大多是动态的,因此,如何确定 IP 地址属于哪个主机是需要解决的问题。以上 2 个问题是解决海量日志数据的关键性难点。首先,将对收集到的海量日志数据进行处理和分析,解决 IP 地址不确定的问题;其次,用聚类的方法解决海量日志数据的问题。

1.1　IP 数据库建立

为了解决大多数网络设备只记录终端的 IP 地址,并且,IP 地址大多是动态的,因此,需要建立一个 IP 数据库,用来确定每个主机静态和动态 IP 地址的分配。并且,绑定 IP 地址和主机。确定每个记录的事件来自于特定的主机。

IP 地址与主机的映射[5](IP-to-host):在现在的网络设计中,终端一般会使用动态 IP 地址,当有终端连入到网络通过 DHCP 服务器获取 IP 地址,但这些地址只会分配一段时间。这在分析每一台主机及其所关联的 IP 地址报告的日志产生了很大的困难。因为同样的 IP 地址可能分配给不同的主机,所以当事件已经被记录的时候无法确定是哪个主机用户。为了解决这个问题,需要收集系统的 DHCP 服务器记录,并且,为一段时间内的 IP 地址对主机的映射建立一个数据库进行存储。每一个绑定是一个组,包括 IP 地址、MAC 地址、主机名、起始时间、结束时间。当新的 DHCP 记录产生时算法每天都更新产生已经存在的绑定。通过给出的数据库绑定结果,就可以识别主机对应于指定的 IP 地址。

静态 IP 地址检测:由于存在主机使用静态 IP 地址,IP-to-host 查找的构想可能会失败。采取的方法是:首先,为收集到的所有 IP 地址记录设置一个地址池 A。其次,再从这些记录中找到只包含主机动态 IP 地址,创建一个已知的动态 IP 地址池 D。最后,利用 A-D 计算出可能包含静态 IP 地址的集合 S,并且对每个地址执行反向 DNS 查找,保存结果完成第一阶段。在第二阶段,选取一个时间周期进行 IP 地址收集,随着新的记录日志的产生,重复第一个阶段收获新的 IP 地址并更新 A,D 和 S 这 3 个数据库。如果主机名发生变化,那么给定的 IP 地址不是静态分配,则从集合 S 删除。在这种方式中,完善了每次迭代发现的潜在的静态 IP 地址。如果未能找到相应的绑定 IP-to-host 查找,而是发现给定地址在集合 S 内,把这个地址作为一个静态主机地址。

1.2　聚类

为了解决大数据的问题,使用大规模数据集聚类算法 DBSCAN[6](Density-Based Spatial Clustering of Applications with Noise)来实现。这个算法可以把具有密度特别高的区域划分为各个簇群,并能在空间数据库中有“噪声”的区域感知所有形状的聚类。下面对 DBSCAN 算法的相关术语和定义进行简单的表述[7-8]。

定义 1(对象 ε-的临域)　选定对象在 ε 内的区域半径。

定义 2(核心对象)　假设一个对象 ε-临域内包含最小数目 *MinPts* 个对象,该对象被称为核心对象。

如图 1 中 $\varepsilon = 0.01$ m,$MinPts = 5$,q 是一个核心对象。

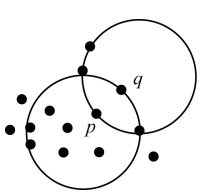

图 1　直接密度可达

定义 3(直接密度可达) 一个给定对象集合 D,假如 p 在 q 的 ε-邻域内,并且 q 是一个核心对象,则称 p 从 q 出发是直接密度可达的。

如图 1 可知,$\varepsilon = 0.01$ m,$MinPts = 5$,如果 q 是一个核心对象,那么 p 从 q 出发是直接密度可达。

定义 4(密度可达) 如果存在一个对象链 $\{p_1, p_2, \cdots, p_n\}$,$p_1 = q$,$p_n = p$,$p_i \in D$,$p_{i+1}$ 是从 p_i 关于 ε 和 $MinPts$ 直接密度可达($l \leqslant i \leqslant n$)。因此,对象 p 是从 q 关于 ε 和 $MinPts$ 的密度可达。

例如,图 2 所示,$\varepsilon = 0.01$ m,$MinPts = 5$,如果 q 是一个核心对象,p_1 是从对象 q 关于 ε 和 $MinPts$ 的直接密度可达,则 p 是从 p_1 关于 ε 和 $MinPts$ 的直接密度可达,则对象 p 是从 q 关于 ε 和 $MinPts$ 的密度可达。

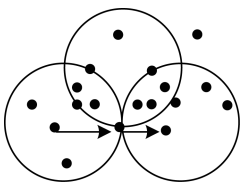

图 2 密度可达

定义 5(密度相连) 假如在对象集合 D 中存在一个对象 O,使得 p 和 q 是从 O 对于 ε 和 $MinPts$ 的密度可达,则对象 p 和对象 q 是关于 ε 和 $MinPts$ 的密度相连。

如图 3 所示,对象 p 和 q 是关于 ε 和 $MinPts$ 密度相连的。

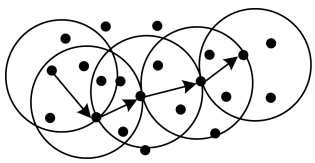

图 3 密度相连

定义 6(噪声) 基于密度可达性的最大密度相连对象的集合是一个基于密度的簇。不包含在任何簇中的对象被认为是“噪声”。

如图 4 所示,不包含在任何簇中的对象即为噪声。

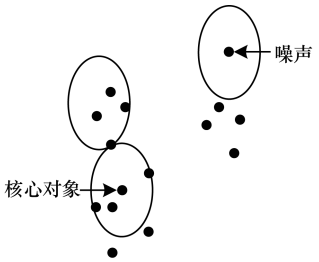

图 4 噪声示意图

1.3 实施方法

首先建立一个事件安全管理系统(SEMS)用于存储收集到的所有日志数据。之后对其进行 DBSCAN 聚类操作。最后,找出离群点,即为异常操作的日志,从而达到缩减数据的目的。在实施聚类算法中,ε 的取值是考虑的重点,正常情况下,ε 邻域设为 1 的,只有记录完全相同的日志能被放入到一个簇中;如果 ε 的值是 0.01 的,大部分记录毫不相关的日志数据也能分到一个簇中。由于需要将文本含义相似的记录放到一起,因此选取[0.4,0.7]区间较合理。算法执行中依次使用不同的 ε,流程如下:

1)设置当前 ε 的值为 0.7。

2)取出 SEMS 库里的日志进行 DBSCAN 算法聚类,得到一组 ε 邻域为 e 的类,标记为 $C_{i,e}$。

3)提取出所有的噪声对象,即异常日志。

4)如果噪声数量小于 $MinPts$ 或者 ε 的值为 0.4,结束聚类;否则,执行 $\varepsilon = \varepsilon - 0.1$,返回 2)。

5)设置 ε 的值为 0.05,把收集到的离群点执行聚类操作。

6)如果有少量离群点尚未成簇,执行 $\varepsilon = \varepsilon - 0.01$,继续 5),直到全部分完。

通过对所有主机及服务器产生的日志执行 DBSCAN 算法。算法经过迭代,大部分日志记录都落入几个大的簇,剩余的由少量记录组成的簇,它们所表示的主机端和网络端行为明显偏离正常情况。下面会对这些异常事件利用 14 个特征进行高级可持续性威胁通信分析。

2 高级可持续性威胁通信日志特征描述

特征提取依据 4 个方面:出栈通信记录日志;已知的恶意攻击行为;违反网络管理员制定的策略;操作环境的特性。根据这 4 个方面,对于网络中的所有主机每天产生的高级可持续性威胁通信异常特征向量全部都包含在以下 14 个特征当中,如表 1 所示。

表 1 APT 隐秘通信特征

特征类型	描述
基于目的地址	1 新的目的地址 2 白名单 Referer 中没有的新地址 3 不受欢迎的原始 IP 地址
基于主机的特征	4 新的用户代理字符串
基于策略	5 阻塞域 6 阻塞连接 7 怀疑域 8 怀疑连接 9 信任域 10 信任连接
基于流量	11 连接峰值 12 域峰值 13 连接爆发 14 域爆发

2.1　目的特征

需要检测的关键在于识别主机通信的对象是新的还是来历不确定的外部地址等这类从来不连接或极少连接的对象。假设受欢迎的 WEB 站点是容易管理并极少可能被盗用的,则那些连接不常见的目的地址就可能认定为是可疑的高级可持续性威胁隐秘通信的主机,所以,要对这些地址进行记录和分析。

1)新的地址。目标是那些每个主机每天连接的新的外部目的地址。随着时间的增长,建立一个 list 用来记录内部主机连接的外部目的地址。一段时间后,把关注点放在某天出现的一个从来没有与主机连接过的新的目的地址,并且这个地址也不在 list 内。利用更新 list 加入新的目的地址。通过应用过滤、用户白名单和域“折叠”的方法使 list 有良好的扩展性能。

首先,过滤经常访问的目的地址并建立一个用户白名单,这个经常访问的地址由用户来定义。这个白名单里包括外部的地址、域和 IP 地址,这些地址都是和内部用户交互的次数,要高于预先设定的阈值。之后,对于用户白名单“折叠”目的地址形成第 2 层的域以便于过滤出随机字符串产生的服务即子域。忽略像是书签之类的连接检索图标。最后,不对原 IP 地址进行解析(大多数合法的站点是通过域名被访问)。并且,白名单上没有的原 IP 地址不考虑。这些优化将会减少日常处理时间。

2)白名单 HTTP_REFERER 中没有的新地址。这个特征承接第一个特征,是统计用户连接到新的地址中没有白名单的 HTTP_REFERER。如从 Google 搜索结果中点击进入了某个页面,那么该次 HTTP 请求中的 REFERER 就是 Google 搜索结果页面的地址。用户通常通过搜索引擎访问新的站点,这样的话,需要在用户白名单里建立一个可信任的 HTTP_REFERER 表。用户访问的新的站点的HTTP_REFERER不在这个表里是被考虑成值得怀疑的。

3)可疑的原始 IP 地址。那些可疑或者极少访问的外部地址可能是高级可持续性威胁隐秘通信的站点,特别是一些直接用原始 IP 地址直接访问的。因此,首先,需要计算出用户连接那些可疑的外部地址和直接通过 IP 访问的地址的数目。通过原始 IP 连接这些外部地址的行为是可疑的活动,这是因为合法的站点通常通过域名访问。其次,把访问可疑目的 IP 地址的用户连接列入表中。可能存在偶尔的直接与 IP 地址通信的连接是正常的,但大多数行为值得怀疑。

2.2　主机特征

大多数主机上使用的软件种类和配置比理论网络中的更加复杂,因此,基于主机的特征提取主要通过主机安装的新软件和软件的异常更新、替代和删除。

由于在主机设备中缺少可见性,并且在网络设备上只有收集记录的权利,那么检测某个主机上软件的异常行为则通过分析软件配置中用户代理字符串(UA)来确定。UA 表示一个特殊字符串的头,它使服务器能够识别客户端使用的 CPU 的类型、操作系统的版本、浏览器的版本、浏览器应用的语言、渲染引擎、插件等。所以,APT 隐秘通信可以通过对字符串的隐写或者嵌入攻击代码达到连接或者控制内部主机的目的。因此,基于主机的特征就是分析新的 UA。

以一段时间作为周期,为 UA 建立一个历史列表。这个表里存储了通过主机观察到的每个 UA。之后,如果有与历史列表里的特征不相同的 UA 出现,将作为新的特征加入到表中。判断 UA 是否异常,通过匹配编辑距离来确定。编辑距离具体是测量它们的插入特征、删除特征和替换特征。其中,替换特征是指一个字符串替换成另一个。所以,这就需要分析软件更新产生的新的 UA,即使是一个小的改变,也可能是 APT 的一次隐秘通信,例如版本更新。

2.3　策略特征

出站连接的网络策略也是检测 APT 隐秘通信的关键特征,是加强整体网络环境安全的重要因素。如果某个外部站点具有低信誉或者被用户禁止过,那么将这个连接到外部站点的目的地址划分到阻塞域。阻塞域是判断主机恶意行为的一个粗略指标。根据访问一个未知的目的地址,如这个地址并没有进行过分类或者认定,用户在访问之前必须要接受已制定好的网络策略。当认为这是一个存在风险的可疑域或可疑连接的需求,允许访问的操作需要用户授权。该策略特征包含 3 种类型的通信,分别是阻塞、怀疑和允许。

2.4　流量特征

高级可持续性威胁隐秘通信会使主机在某一时刻形成峰值流量,比如扫描或者僵尸程序向控制主机发送响应报文。流量特征试图通过分析主机产生大量异常流量的时间特征来捕获这些异常活动。具体来说,如果在一分钟的时间窗口内主机产生比预设阈值更多的连接(或连接更多的域),那么定义这一分钟窗口为连接峰值(或者域)。一个连接爆发值(或者域)是指一个时间段,在这个时间段内每一分钟都是连接(或者域)峰值。为了确定一个合适的阈值,通过考察所有主机一个星期的时间并计算每个主机每分钟产生的连接(或者连接域)的数量。假设所有主机所有一分钟窗口的累积分布中 90% 的主机每分钟产生少于 101 个连接,连接域的个数少于 17 个。因此,将其设定为连接峰值和域峰值的阈值。对于爆发值可以稍微放宽它的定义,对所有主机当

一段时间内 75% 的一分钟窗口都是峰值的话,那么认为该段时间为爆发值。这个值为 26 个连接峰值和 6 个域峰值。对每个主机,其流量特征包括:连接峰值的个数,域连接的个数,最长连接爆发值持续的时间以及最长域爆发持续的时间。

这节利用大数据的方法针对 APT 通信特征进行了深度的分析,并提出了 14 种通信特征,这些特征在日志的特征检测上已经较为完备。

3 LDA 检测模型

隐含狄利克雷分布(LDA)模型[9]提出的目的就是为了识别大规模文档集或者语料库中潜藏的主题信息。所以,用在大数据日志挖掘是目前为止最合适的模型。但目前还没有人将此模型用于 APT 隐秘通信检测,因此,这是一次大胆的创新。

3.1 LDA 基本思想

LDA 是对离散数据集建模的概率增长模型,是一个 3 层贝叶斯模型,对文档进行一个简短的描述,保留其统计信息,对于高效地处理大规模文档集有很好的作用。LDA 模型应用于文档数据集建模的符号约定[10]:

1)词 w 是文本数据的基本单元,用 $\{1,2,\cdots,V\}$ 索引的词表的分项。词中的第 v 个词用一个 V 维的向量表示,对于任意 $u\neq v$,$w_v=1$,$w_u=0$。

2)文档是 N 个词的随机序列,用 $d=\{w_1,w_2,\cdots,w_n\}$ 表示,w_n 是列中的第 n 个词。

3)文档集 D 表示为 M 个文档的集合 $D=\{d_1,d_2,\cdots,d_M\}$。如果有 T 个主题,那么文档集 D 中第 i 个词汇 w_i 的概率表示如下:

$$P(w_i)=\sum_{j=1}^{T}P(w_i|z_i=j)P(z_i=j) \tag{1}$$

z_i 是第 i 个词汇 w_i 取自主题 T 的变量,w_i 分布在 j 的概率是 $P(w_i|z_i=j)$,d 在 j 上的概率分布为 $P(z_i=j)$。在文档 d 中 V 个词对应 j 的多项式为 $\varphi_{w_i}^{j}=P(w_i|z_i=j)$。文档 d 发生词汇 w 的概率为:

$$P(w|d)=\sum_{j=1}^{T}\varphi_{w_i}^{j}\cdot\theta_j^d \tag{2}$$

利用 EM(期望最大化算法)求最大似然函数:

$$l(\alpha,\beta)=\sum_{i=1}^{M}\mathrm{lb}p(d_i|\alpha,\beta) \tag{3}$$

使用最大似然估计量 α 和 β 的参数值确定 LDA 模型。文档 d 发生的条件概率分布:

$$P(d|\alpha,\beta)=\frac{\Gamma(\sum_i\alpha_X)}{\prod_i\Gamma(\alpha_i)}\cdot\int(\prod_{i=1}^{k}\theta_i^{\alpha_i-1})(\sum_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\sigma_{ij})w_n^j)\mathrm{d}\theta \tag{4}$$

受参数 θ 和 β 之间耦合关系的影响,这个分布无法直接计算,只能通过近似算法如变分推进行计算。在已有的 LDA 模型中,Gibbs[11]抽样在执行效率和辨别度上是较为高效的方法。

3.2 建模方案

本文在对 APT 通信日志在 LDA 模型下的概率分布具体步骤如下:

1)14 种 APT 隐秘通信特征是用来检测的主题 $\{T_1,T_2,\cdots,T_{14}\}$。

2)本文分离出来的离群点即异常日志进行组成的簇记为文集 $\{d_1,d_2,\cdots,d_n\}$,文集中的每个异常日志记录记为词 $\{w_1,w_2,\cdots,w_n\}$。

3)对于每一个 w,使用 LDA 为文集 d 建模。将 w 表示为 14 种主题上的概率分布。再通过 Gibbs 抽样估计,得到 w 在 T 上的概率分布 $P=\{K_1,K_2,\cdots,K_{14}\}$。

关于得到 w 的概率分布,先通过 Gibbs 抽样间接求得 α 和 β 的值,再利用 α 和 β 作为参数计算 w 的值。Gibbs 抽样是 MCMC[12]的一种简单形式,核心是构造 Markov 链计算出接近该概率分布值的样本。本文中需要得到主题 T 对词 w 的分配,通过 z_i 的抽样得出。计算公式如下:

$$P(z_i=j|z_{-i},w_i)=\frac{\frac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+\chi}\circ\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}}{\sum_{j=1}^{T}\frac{n_{-i,j}^{(w_i)}+\chi}{n_{-i,j}^{(\cdot)}+\chi}\circ\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha}} \tag{5}$$

其中,$z_i=j$ 表示 w_i 分配给主题 j,w_i 表示词汇和与该词汇在文档 d 中的位置;z_{-i} 是所有 $z_k(k\neq i)$ 的分配;$n_{-i,j}^{(w_i)}$ 是 w_i 与分配给主题 j 的词汇相同个数;$n_{-i,j}^{(\cdot)}$ 是分配给 j 的词汇总数;$n_{-i,j}^{(d_i)}$ 是文本 d_i 分配给 j 的实际词汇个数;$n_{-i,\cdot}^{(d_i)}$ 是 d_i 中所有被分配的词汇个数。

Gibbs 的具体操作:先将 z_i 初始化成 1 ~ T 之间的某个随机数。其次,i 执行从 1 ~ N 的循环,N 是 d 中所有出现在文集中 w 记号的个数。至此 Markov 链的初始化完成。再次,继续执行 i 从 1 ~ N 的迭代,通过式(5)计算出 Markov 链的下一个状态。最后,执行迭代过程,直到 Markov 链已经完全接近目标分布。此时把 z_i 的值作为样本记录下来。

4)根据设定好的阈值来比较 d 的值。小于阈值表示 d 为非 APT 通信日志集。大于阈值意味着文档中存在 APT 通信日志,再根据其概率分布的特征确定为哪类或哪几类 APT 通信。

4 实验结果与分析

实验周期为 1 周,数据来源于理工大学网络中心服务器端和主机端每天产生大量的日志。一周时间里,总共收集到 6 GB 的各类日志数据(如:网络代理,VPN 的远程连接日志,Windows 窗口域控制器验

证,杀毒软件扫描恶意软件的结果),并利用这些数据对 LDA 模型进行评估。

4.1　LDA 模型参数确定

一般来说,主题 T 的值对 LDA 模型的参数影响很大。但是,本文的主题 T 是固定的 14 种特征。因此,需要确定 α 和 β 在 14 种特征下的参数。在 LDA 模型中[13],α 和 β 分别是 θ 和 φ 上的 Dirichlet 先验概率假设。其自然共轭的特点说明利用 θ 和 φ 可以求出联合概率 $P(w|z)$ 的值:

$$P(w|z)=\left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^{w}}\right)^{T}\sum_{j=1}^{T}\frac{\prod_{w}\Gamma(n_j^{(w)}+\beta)}{\Gamma(n_j^{(\cdot)}+W\beta)} \tag{6}$$

其中,$\Gamma(\cdot)$ 是 gamma 函数的标准形式;$n_j^{(w)}$ 是词 w 在主题 j 下的分布频率;$n_j^{(\cdot)}$ 是主题 j 中所有词 w 的总数。$P(w|T)$ 可以近似看成 $P(w|z)$ 的调和平均数。公式如下:

$$\frac{1}{P(w|T)}=\frac{1}{M}\sum_{m=1}^{M}\frac{1}{P(w|z^{(m)})} \tag{7}$$

根据式(7),通过 lb$P(w|T)$ 的变化可以得出当 $T=14$时,令 $\alpha=70/T$,$\beta=0.01$ 在本训练集上效果较好。

4.2　实验结果

本文将这些数据收集到 SEMS 数据库中进行预处理。首先,建立 IP 数据库实现 IP 地址与主机的配对。在检测范围内的 100 台主机平均每个主机每天产生 85.8 MB 的通信日志。其次,对收集到的数据执行 DBSCAN 算法的聚类操作,共得到 4 894 个可疑簇群,大约 456 MB 的日志数据。最后,利用 LDA 建模方案将 4 894 个簇作为检测的文档集 d 根据 14 种APT 通信特征进行检测。实验中,阈值设为 1。为了减少误报率,对文档 d 中 w 进行反复迭代,最终得出当 w 的语义完全符合 14 种特征中的其中一种,则此 w 的值为 1;完全不符合则为 0;疑似符合则为 0.01。检测结束后,将文档集 d 中的 w 值相加,结果大于阈值为 APT 异常通信。图 5 ~ 图 9 是 LDA 模型根据 14 种 APT 隐秘通信特征得到文档集 d 在主题 T 下的概率分布结果。通过计算每个 d 在每个特征向量在集群中所有向量中的平均值,再标准化每个向量分量可以发现一个文档 d 中的 w 可能属于不同特征 T。

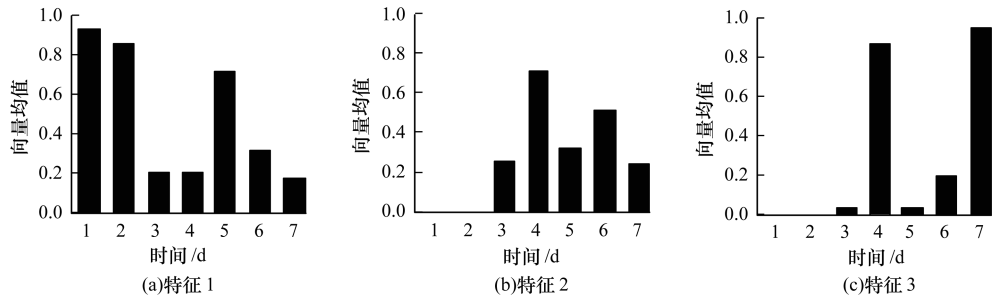

图 5　d 在 $T_{1\text{-}3}$ 特征下的概率分布

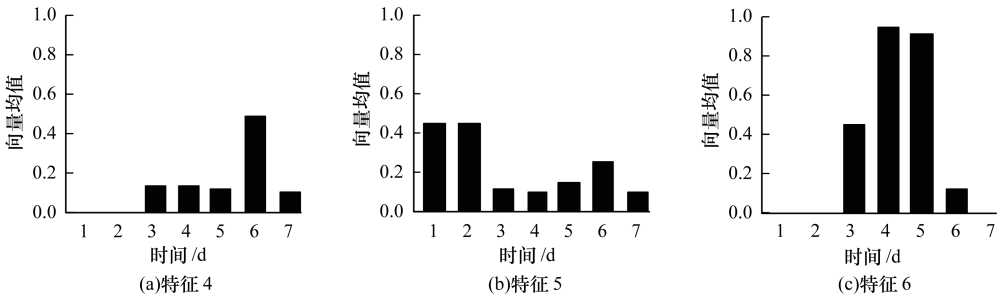

图 6　d 在 $T_{4\text{-}6}$ 特征下的概率分布

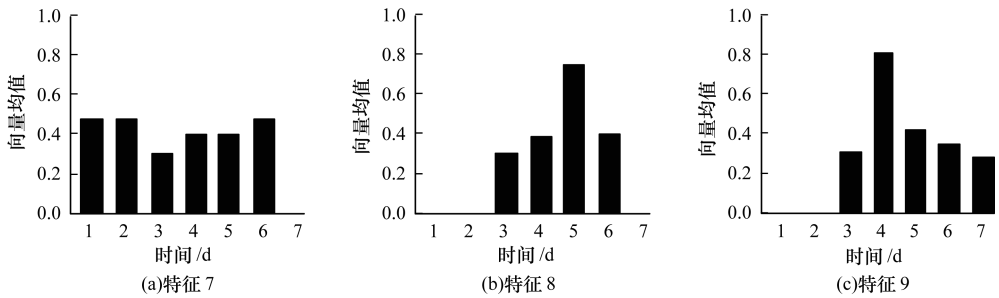

图 7　d 在 $T_{7\text{-}9}$ 特征下的概率分布

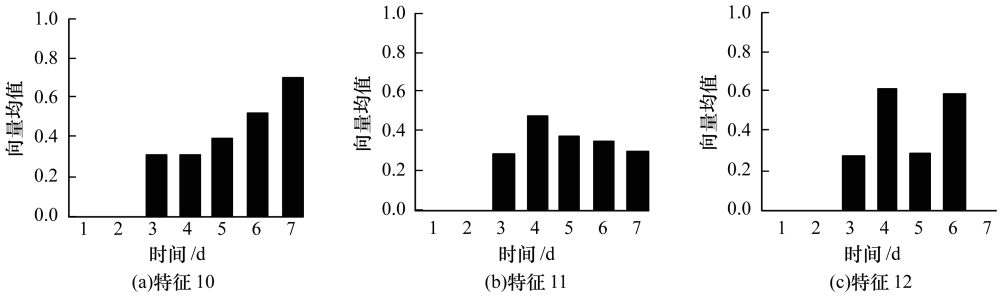

图 8 d 在 $T_{10\text{-}12}$ 特征下的概率分布

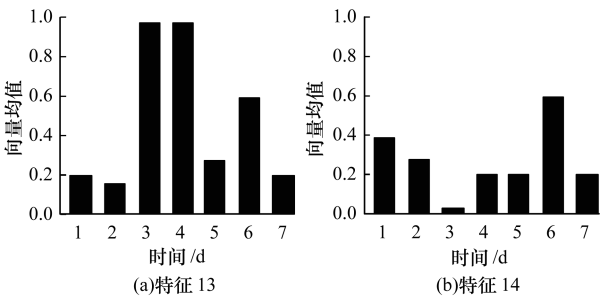

图 9 d 在 $T_{13,14}$ 特征下的概率分布

在一周的实验中,根据安全软件的分析,共检测出包括查找 UA 字符串、HTTP 状态代码、Web 引用页、主机连接的时间以及连接域的信誉的异常操作。在利用所有 14 种特征检测出的事件中,25.25% 被证实是恶意软件或“可疑”行为,39.41% 是违反网络管理员制定的安全策略,35.33% 和未认证的软件或服务有关。表 2 列出了每个类别的事件数量和比例。

表 2 异常事件数量与比例

种类	事件个数	所占百分比/%
恶意软件	117	14.92
可疑行为	81	10.33
违反策略-隧道	1	0.12
违反策略-文件共享	2	0.25
其他-网络浏览	63	8.03
违反策略-流媒体	86	10.96
违反策略-即时通讯	56	7.14
违反策略-问题图片	6	0.76
违反策略-赌博	13	1.65
违反策略-用户代理	4	0.51
违反策略-远程访问	8	1.02
违反策略-违禁网站	133	16.96
其他-未分类网站	57	7.27
其他-自动化	157	20.02

4.3 与其他文本分类算法的比较

为验证 LDA 挖掘算法在对海量日志数据处理上效率和准确率均优于同类其他算法。使用 SEMS 数据库中 120 MB 的可疑日志作为实验数据用于文本挖掘算法 LDA,潜在语义分析(Latent Semantic Analysis,LSA)[14] 和概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)[15] 的比较,结果如图 10所示。

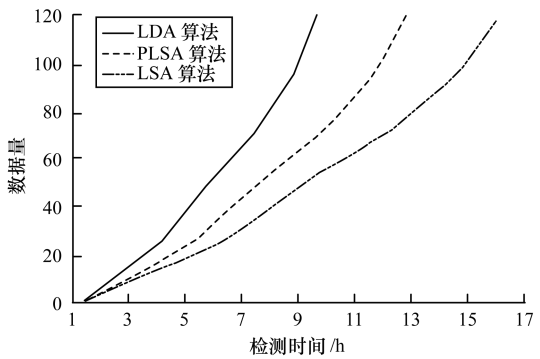

图 10 3 种算法效率比较

根据图 6 得出 3 种算法的时间呈线性增长。通过对比可知,LDA 建模的检测方法处理数据的速度要优于其他 2 种。关于这 3 种算法的准确性的比较可以从表 3 看出,经过 LDA 建模后的日志挖掘准确率远远高于其他模型算法,而且,实验表明测试结果比较稳定。最后对 PLSA 算法的文本分类进行深入研究,发现该算法的分类结果出现的随机性较高。

表 3 3 种算法准确率百分比 %

特征	LDA 建模算法	LSA 建模算法	PLSA 算法
$T_{1\text{-}4}$	80	80	75
$T_{5\text{-}8}$	95	88	91
$T_{9\text{-}12}$	94	82	94
$T_{13,14}$	92	82	90

5 结束语

本文通过建立 IP 地址数据库收集到海量通信日志,并应用 DBSCAN 聚类算法对海量日志数据中正常通信日志和异常通信日志进行分离。通过分析高级可持续威胁通信的 14 种通信特征,利用 LDA 建模构建了全新的高级持续性威胁通信检测模型。实验结果表明,该模型在解决大数据问题上及 APT 通信异常检测上比较传统的检测方法都有较好的准确度,降低了误报率。由于 LDA 方法计算量较大,今后的研究中需要对 LDA 模型进行改进,使其性能得到提高。

(下转第 205 页)

取训练次数,得到室外室内各提取 5 次声纹特征的识别成功率如表 2 所示。

表 2　声纹解锁系统实验结果 2

测试文本	解锁成功率/%		平均解锁时间/s	
	室内	室外	室内	室外
芝麻开门	87	85	3.21	3.19
声纹解锁	85	84	3.18	3.21
新年快乐	86	86	3.23	3.22

可以看出,通过提高训练次数,在室内外分别提取声纹特征后,室外声纹识别成功率提高到 85% 左右,室内识别成功率较之前有小幅下降,在 86% 左右。平均解锁时间均为 3.2 s 左右,在理想情况下,用户讲完声纹文本口令一秒内即可解锁成功。解锁时间不受文本和环境的影响,主要与 MFCC 算法、DTW 算法以及硬件平台计算能力有关。

7　结束语

本文介绍了 Android 平台下基于 MFCC 和 DTW 算法的声纹识别系统,并将其与安卓系统锁屏相结合,实现利用声纹解锁手机的目的。相比传统数字或图形解锁方式,声纹解锁不易破解,安全性更高。在 Android4.4.2 平台上的测试结果表明,该系统具有较高的解锁成功率和解锁速度。下一步将重点研究语音降噪算法,提高声纹解锁系统的识别成功率和用户体验。

参考文献

[1] IDC. Android and IOS Squeeze the Competition[EB/OL]. [2016-03-05]. http://www.idc.com/getdoc.jsp?containerId=prUS25450615.

[2] 王即墨,计超豪,裴洪卿,等. Android 智能手机锁屏密码及破解方法研究[J]. 刑事技术,2015,25(2):142-145.

[3] 张圆圆. 指纹识别技术相关算法的研究[D]. 北京:北京邮电大学,2012.

[4] 王二伟. 基于 Android 平台人脸检测与识别研究[D]. 西安:西安电子科技大学,2013.

[5] Gunson N, Marshall D, Mclnnes F, et al. Usability Evaluation of Voiceprint Authentication in Automated Telephone Banking: Sentences Versus Digits[J]. Interacting with Computers,2011,23(1):57-69.

[6] Nakagawa S, Wang L, Ohtsuka S, et al. Speaker Identification and Verification by Combining MFCC and Phase Information[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(4): 1085-1095.

[7] 郭春霞. 基于 MFCC 的说话人识别系统研究[D]. 西安:西安电子科技大学,2006.

[8] 李正欣,张凤鸣,李克武,等. 基于 DTW 的多元时间序列模式匹配方法[J]. 模式识别与人工智能,2011,24(3):425-430.

[9] 胡金平,陈若珠,李战明,等. 语音识别中 DTW 改进算法的研究[J]. 微型机与应用,2011,30(3):30-32.

[10] 宋　艳. 基于嵌入式语音识别系统的研究[D]. 西安:西安科技大学,2011.

[11] Google Inc.. Android NDK Reference [EB/OL]. [2016-03-05]. http://developer.android.com/tools/sdk/ndk/index.htm.

[12] 王华朋,杨洪臣. 声纹识别特征 MFCC 的提取方法研究[J]. 中国人民公安大学学报(自然科学版),2008,14(1):28-30.

[13] Jeong Y S, Jeong M K, Omitaomu O A, et al. Weighted Dynamic Time Warping for Time Series Classification[J]. Pattern Recognition, 2011, 44(9): 2231-2240.

[14] 朱淑琴,赵　瑛. DTW 语音识别算法研究与分析[J]. 微计算机信息,2012,23(5):150-151.

[15] 张华平,玄光哲,于贵平,等. 基于 JNI 技术应用框架的分析和实现[J]. 吉林大学学报(信息科学版),2003,21(2):188-191.

[16] 王有禄,李代平. Android 系统下基于 NDK 方式的图形开发[J]. 计算机系统应用,2012,25(12):56-59.

编辑　陆燕菲

(上接第 200 页)

参考文献

[1] 许　佳,周丹平,顾海东. APT 攻击及其检测技术综述[J]. 保密科学技术,2014(1):34-40.

[2] 王　宇,韩伟杰. APT 攻击特征分析与对策研究[J]. 保密科学技术,2013(12):32-43.

[3] 云晓春. 国家互联网网络安全宏观态势综述[J]. 保密科学技术,2012(1):6-8.

[4] 刘　昕. 大数据背景下的 APT 攻击检测与防御[J]. 电子测试,2014(1):80-81.

[5] Yen T F, Oprea A, Onarlioglu K, et al. Beehive: Large-scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks[C]//Proceedings of the 29th Annual Computer Security Applications Conference. New York, USA: ACM Press, 2013: 199-208.

[6] 戴阳阳,李朝锋,徐　华. 初始点优化与参数自适应的密度聚类算法[J]. 计算机工程,2016,42(1):203-209.

[7] 陈　燕. 对两个经典聚类算法的分析[J]. 现代电子技术,2007,30(17):174-176.

[8] 冯少荣,肖文俊. DBSCAN 聚类算法的研究与改进[J]. 中国矿业大学学报,2008,37(1):105-111.

[9] 刘　萍,郑凯伦,邹德安. 基于 LDA 模型的科研合作推荐研究[J]. 情报理论与实践,2015,38(9):79-85.

[10] 施乾坤. 基于 LDA 模型的文本主题挖掘和文本静态可视化的研究[D]. 南宁:广西大学,2013.

[11] 刘伟峰,杨爱兰. 基于 BIC 准则和 Gibbs 采样的有限混合模型无监督学习算法[J]. 电子学报,2011,39(3):134-139.

[12] 石　晶,胡　明,石　鑫,等. 基于 LDA 模型的文本分割[J]. 计算机学报,2008,31(10):1865-1873.

[13] 王　鹏,高　铖,陈晓美. 基于 LDA 模型的文本聚类研究[J]. 情报科学,2015(1):63-68.

[14] 俞　辉. 基于 LSA 和 PLSA 的多文档自动文摘[J]. 计算机工程与科学,2009,31(9):108-111.

[15] 牛　贺. 基于 PLSA 模型的推荐算法研究与实现[D]. 沈阳:东北大学,2012.

编辑　刘　冰