

引用格式:肖琦,苏开宇.基于随机森林的僵尸网络流量检测[J].微电子学与计算机,2019,36(3):43-47.[XIAO Qi, SU Kai-yu. Botnet traffic detection based on random forest algorithm[J]. Microelectronics & Computer, 2019, 36(3): 43-47.]

基于随机森林的僵尸网络流量检测

肖 琦¹, 苏开宇²

(1 中国计量大学 信息工程学院, 浙江 杭州 310018;

2 中国计量大学 现代教育技术中心, 浙江 杭州 310018)

摘 要:近年来,僵尸网络已经成为互联网领域所面临的严重的安全威胁之一.僵尸网络的控制者使用僵尸程序进行例如DDOS攻击,发送垃圾邮件,盗取敏感信息等一系列恶意活动.在海量的互联网流量中,从正常流量中有效的识别出僵尸网络流量是亟待解决的问题.本文的研究目的是分析和识别僵尸网络流量.实验在真实的包含僵尸网络流量数据中提取了9种特征,采用随机森林算法作为训练模型,与5种其他机器学习算法进行比较,实验结果表明本文提出的模型在给出的评价标准下性能最优.

关键词:僵尸网络;机器学习;网络安全

中图分类号: TP391.42

文献标识码: A

文章编号: 1000-7180(2019)03-0043-05

Botnet traffic detection based on random forest algorithm

XIAO Qi¹, SU Kai-yu²

(1 College of Information Engineering, China Jiliang University, Hangzhou 310018, China;

2 Modern Education Technology Center, China Jiliang University, Hangzhou 310018, China)

Abstract: In recent years, Botnets have become one of the serious security threats in the Internet field. Botnet controllers used bots process to perform a series of malicious activities, such as DDOS attacks, sending spam, stealing sensitive user information, and so on. Identification of botnet traffic from normal traffic in the massive internet traffic is an urgent problem. The purpose of this paper was to analyze and identify botnet traffic. The experiment extracted 9 features from the real botnet traffic data, used the random forest algorithm as the training model and compared it with 5 other machine learning algorithms. The experimental results show that the proposed model has the best performance under the given evaluation criteria.

Key words: botnet traffic; machine learning; network security

1 引言

僵尸网络是指攻击者采用一对多的命令方式以及控制信道,为了控制大量主机所组建的攻击网络.僵尸网络成为互联网的主要威胁,由于它被设计在后台运行,经常没有明显的证据证明它的存在.在过去的十年中,僵尸网络的目的集中在非法手段挣钱,例如盗取信息和计算机资源,广告软件,钓鱼网站

等^[1-2].因此,找到对付僵尸网络的方法是严峻的挑战.在检测僵尸网络时,僵尸网络通常会表现出流量行为的一致性,呈现独特的通信行为,并且这些行为可以使用将非正常流量与正常流量区分开来的一组属性来表征和分类^[3-4].随着信息技术的飞速发展,互联网的使用者越来越多.社交网络每秒都会通过网络传输大量数据,从而存在导致泄露个人信息的潜在风险.黑客使用多台计算机配合,窃取机密信息,使用多种手段散布恶意软件来达到攻击的目的.

甚至教育系统、政府、军事系统都成为获取信息和资源的热门目标^[5]. 因此,在开发或者应用新的僵尸网络检测技术以防止非法访问机密信息时,加强网络安全非常重要.

僵尸网络的控制者是通过僵尸网络的命令控制通道(C&C)服务器向僵尸程序发布命令. 命令控制通道充当了僵尸主控机和僵尸网络之间的接口. 通常恶意软件中会携带着命令控制通道服务器的地址,这个地址可以是硬编码的 IP 地址的、静态域名,或者是动态域名^[4,6]. 僵尸主控机通过命令控制通道得到命令去执行非法活动,僵尸主控机使用一种或者多种协议作为通信协议来控制僵尸主机进行协同工作,C&C 服务器是僵尸网络控制者控制僵尸主机的唯一方式.

僵尸网络检测领域已经有了很多的研究成果^[7-10]. 但是基于网络流量的僵尸网络检测方法仅在近几年才出现. 本文通过对网络流量进行特征提取,使用改进的随机森林挖掘出僵尸网络.

2 僵尸网络检测

2.1 检测模型

僵尸网络检测的问题被建模成二分类任务,并使用监督学习算法解决. 机器学习方法比传统的统计学习方法更加高效,因为机器学习方法能够从训练数据中自动学习到其中潜在的模式,从而使得精准度更高^[11]. 僵尸网络检测模型如图 1 所示,首先将选好的数据集(CTU-13)进行数据准备及预处理. 然后选取最有效的特征构建分类器,使用随机森林,决策树,朴素贝叶斯,逻辑回归,支持向量机和 K 近邻分别建立分类模型,用训练好的分类器检测僵尸网络. 最后根据相应的评判标准,选出最优模型.

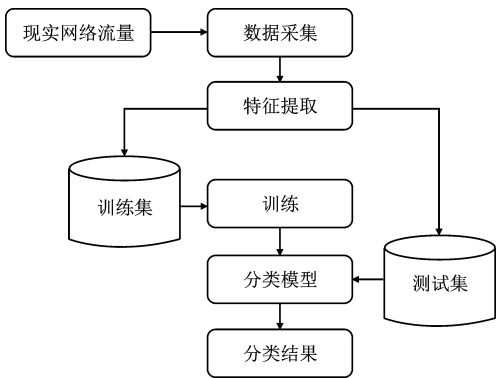


图 1 僵尸网络检测模型

2.2 特征提取

本文中,用来区分僵尸网络流量和正常流量选

取的特征主要是:特定流程的持续时间,通信过程中需要的协议和状态,用于指定本地或远程主机上提供的服务的源端口和目标端口,用于数据包在网络中传输的源 IP 地址和目标 IP 地址,在特定流程期间传送的总包数和总字节数.

(1)持续时间(Duration)

流量的持续时间表示完成特定流量所花费的总时间. 流量的持续时间可以用于计算平均数据包速率和平均字节速率. 将时间格式更改为数值的公式为:小时 * 3 600 +分钟 * 60 +秒.

(2)协议(Protocol)

协议是在网络连接使用端进行通信时的一组特殊规则. 协议有很多种类型可以使用,例如 TCP,UDP,ICMP,SMTP 等.

(3)源 IP 地址(Source IP Address)和目标 IP 地址(Destination IP Address)

IP 地址是被用于唯一标识所要访问的目标主机,同样,IP 地址也是传统计算机网络领域内最基本的特征属性. 源 IP 地址指的是当前主机或网站正在使用的 IP 地址. 目标 IP 地址是网络信息传输过程中信息接收方的 IP 地址. IP 地址用于在计算机网络中传输数据包^[12]. 在传输过程中,源 IP 地址和目标 IP 地址相对应并且保持不变,实现了端到端的信息传输功能. 流量数据中的 IP 地址一般都用二进制的方式表示,而在我们实际使用中通常将其转换为十进制,以方便计算.

(4)源端口(Source Port)和目标端口(Destination Port)

在计算机网络信息传输过程中,数据和请求会被发送到事先声明好的服务或应用程序中,端口号(port number)则被允许用来标识这里的服务或应用程序. 同样与 IP 地址类似,端口号也被分为源端口号和目标端口号,它们被用来获取针对攻击目标的远程系统的信息. 端口 80 号,53 号,25 号被标记为僵尸网络攻击的恶意流量,它们分别代表 http 建立的僵尸网络,垃圾僵尸网络和基于 DNS 服务器的僵尸网络.

(5)状态(States)

在计算机网络中,表示网络流量状态的有 SYN,RST,CON,ACK,FIN. 当处于 SYN 状态中,客户端向服务器发送包含服务器端口和客户端初始序列号的 SYN 消息,服务器则返回它自己的 SYN 和 ACK. 当客户端发送一个 ACK,最终状态是一个半封闭的连接,客户端不再发送数据,但仍然能够

从服务器接收数据.若收到的是 FIN,服务器进入关闭状态. CON 是连接状态,只要连接建立,它就处于 CON 状态. RST 状态是主机拒绝连接的连接重置状态.如果收到太多的 SYN 状态意味着发送者受到感染.如果接收到太多的 RST 状态意味着接收器受到了感染^[13-14].

(6)总包数(Total Packets)和总字节数(Total Bytes)

总包数被定义为在特定流量中传输的数据包数量.它存储的是特定时间段或流量期间传出的总数据包数量.总字节数是指客户端根据请求发送的总的字节数量.总字节大小是网络测量的一个重要指标.

2.3 随机森林

随机森林(Random Forest, RF)算法是一种监督学习算法,属于同质集成学习,通过将多个决策树进行集合,获得比单个学习器更加优越的泛化性能,随机森林的模型结构如图2所示.

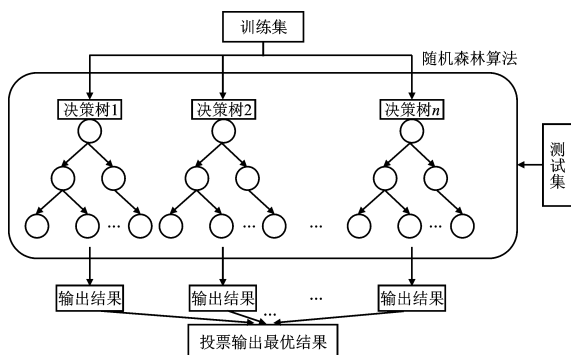


图2 随机森林模型结构

随机森林中提到的“随机”是指 n 棵决策树随机且有放回的方式抽取样本和每个基本决策树均使用随机选取的 m 项特征属性.两个随机性的引入,使得随机森林不容易陷入过拟合,且具有较好的抗噪能力(比如:对缺省值不敏感).随机森林相较于 Bagging 算法,只做了小改动. Bagging 算法中基学习器的“多样性”仅通过对初始训练集中的样本进行扰动,而随机森林中基学习器的多样性不仅是样本的扰动,还包括属性的扰动,这就使得随机森林最终的泛化性能可通过个体学习器之间差异度的增加而进一步提升^[15-16].

在随机森林算法的实验中,考虑到两个参数:决策树的个数 n 和随机特征变量 m ,通常取 $m = \sqrt{M}$.由网格搜索设定 n 的值为 40.另外,决策树的分枝测度有两种方式:信息熵和 Gini 指数.这里选择 Gini

指数作为划分标准,其公式如下所示:

$$\text{Gini}(p) = \sum_{k=1}^k p_k(1-p_k) = 1 - \sum_{k=1}^k p_k^2 \quad (1)$$

式中, p_k 表示为训练集中属于类别 k 的样本个数占样本总个数的比例.

对于二分类问题,样本的一个类的概率是 p ,则概率分布的 Gini 指数是:

$$\text{Gini}(p) = 2p(1-p) \quad (2)$$

对于给定的样本集合 D ,它的 Gini 指数是:

$$\text{Gini}(D) = 1 - \sum_{k=1}^K \left(\frac{|C_k|}{|D|} \right)^2 \quad (3)$$

式中, C_k 是集合 D 中属于第 k 类的样本子集; K 是类的个数.

如果样本集合 D 根据一个特征 A 被分割成 D_1 和 D_2 ,则在特征 A 的条件下,集合 D 的 Gini 指数为:

$$\text{Gini}(D, A) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \quad (4)$$

Gini 越大,样本集合的不确定性也就越大, $\text{Gini}(D, A)$ 表示根据特征 A 分割后的不确定性.

这里采用简单的投票法(即“少数服从多数”)决定分类结果,分类决策公式如下:

$$H(x) = \arg\max \sum_{i=1}^k I(h_i(x) = Y) \quad (5)$$

式中, $H(x)$ 表示组合分类模型; h_i 表示基本决策树的分类模型; Y 表示目标变量.

3 实验结果分析

3.1 实验数据集

本次实验所用的数据是 CTU-13 数据集,该数据集是捷克理工大学从恶意软件捕获设施项目中收集的混杂着正常流量和背景流量的真实僵尸网络流量.该数据集由 1 048 576 个网络流组成.原始数据集包含的特征总共有 14 项.在实际构造训练集和测试集中,只取了其中方便提取信息的 9 项特征,即:持续时间、协议类型、源端口号、目标端口号、源 IP 地址、目标 IP 地址、总数据包数、总字节数、状态. CTU-13 数据集中包含三种流量:正常流量,背景流量和僵尸网络流量,在处理过程中,将正常流量和背景流量归为一类;僵尸网络流量归为一类.从而简单的将僵尸网络检测任务转化为二分类任务.为了保证二分类任务的类别平均,在训练集和测试集构造过程中保持了两个类别的数据均等,这样可以有效防止类别不均匀导致的一系列问题出现.

3.2 评价标准

本文给出用于模型评价的评价标准有:准确率(Accuracy)、召回率(Recall)、精确率(Precision)。

表 1 是一个混淆矩阵,对于一个二分类问题,共有四种分类结果^[17-18]。如果将一个正例判断为正例,那么就可以认为产生了一个真正例(TP);如果对一个反例正确的判断为反例,则认为产生了一个真反例(TN)。相应地,另外两种情况分别称之为伪反例(FN)和伪正例(FP)。

表 1 分类结果			
真实结果	预测结果		
	+1	-1	
	+1	True Positive(TP)	False Negative(FN)
	-1	False Positive(FP)	True Negative(TN)

其中,准确率的计算公式:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

(6)

召回率的计算公式:

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

(7)

精确率的计算公式:

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

(8)

以上公式中,准确率用于度量本文二分类问题中分类正确的样本数占总样本数的比例,即被正确分为僵尸网络和被正确分为非僵尸网络占总数的比例;召回率度量了数据集中僵尸网络被检测出来的比例;精确度度量了被分类为僵尸网络中真正僵尸网络流量的比例。

3.3 对比结果

在实验中,分别将随机森林、逻辑回归、支持向量机、决策树、朴素贝叶斯和 k 近邻算法进行性能上的比较。在准确率、召回率和精确率方面的 5 个模型的结果如表 2 所示,分类器的比较性能如图 3 所示。

算法 & 评价标准	随机森林	逻辑回归	支持向量机	决策树	朴素贝叶斯	K 近邻
准确率	99.91	96.97	98.04	99.00	72.25	98.19
召回率	100	99.88	99.98	100	99.98	98.26
精确率	99.82	94.39	96.25	98.04	64.31	98.12

从表 2 可以看出,本文提出的随机森林算法准确率达到了 99.91%,优于其他的机器学习算法。决策树虽然建树时间快,分类准确率达到了 99%,但容易出现过拟合的现象,对未知数据的分类效果不

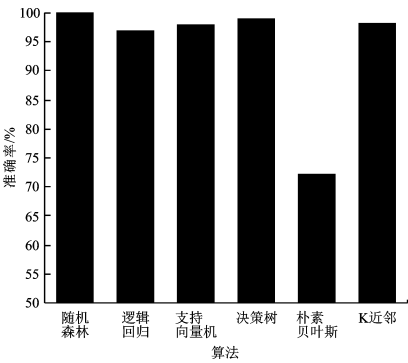


图 3 性能比较结果

佳。而随机森林就很好解决了单棵决策树过拟合的问题。在召回率方面,除了 K 近邻,其他机器学习都达到了接近 100%。也就是说,机器学习技术很适合分类僵尸网络流量。不同之处在于精确率,不同算法的精确率差异明显,对于精确度低的算法,可以认为算法容易将非僵尸网络流量误分类为僵尸网络流量,该算法就检测僵尸网络而言过于保守谨慎。因此得出的结论是:本文提出的采用随机森林机器学习算法用于僵尸网络检测,不论从准确率、召回率还是精确率上均比提出的其他机器学习算法优越。

4 结束语

本文提出了基于随机森林算法的僵尸网络检测模型,并对真实的混合网络流量数据进行特征提取,得到得 9 个特征用于机器学习算法检测僵尸网络流量。同时,将逻辑回归、支持向量机、决策树、朴素贝叶斯和 k 近邻算法在 CTU-13 数据集上进行对比实验。实验结果表明,本文提出的模型,在性能上要明显优于其他机器学习算法。但是该方法还可以进一步优化,接下来的工作会把注意力集中放在对真实的僵尸网络流量的特征工程上,通过特征工程得到更加鲁棒、表现效果更佳的流量特征用于僵尸网络检测,从而进一步提升模型的性能。

参考文献:

[1] KANG K S, KIM H S, AHN B J. DDoS attack detection and defense apparatus and method using packet data; US, US8634717[P]. 2014.

[2] SAAD S, TRAORE I, GHORBANI A, et al. Detecting P2P botnets through network behavior analysis and machine learning[C]// Ninth International Conference on Privacy, Security and Trust. Montreal, Canada, IEEE, 2011:174-180.

[3] WANG J, PASCHALIDIS I C. Botnet detection based

- on anomaly and community detection[J]. IEEE Transactions on Control of Network Systems, 2017, 4(2): 392 - 404.
- [4] 方滨兴, 崔翔, 王威. 僵尸网络综述[J]. 计算机研究与发展, 2011, 48(8):1315-1331.
- [5] SAAD S, TRAORE I, GHORBANI A, et al. Detecting P2P botnets through network behavior analysis and machine learning[C]// Ninth International Conference on Privacy, Security and Trust, Montreal, Canada, IEEE, 2011:174-180.
- [6] KATO K, KLYUEV V. Large-scale network packet analysis for intelligent DDoS attack detection development[C]// Internet Technology and Secured Transactions. London, United Kingdom, IEEE, 2015: 360-365.
- [7] 诸葛建伟, 韩心慧, 周勇林, 等. 僵尸网络研究[J]. 软件学报, 2008, 19(3):702-715.
- [8] LU C, BROOKS R. Botnet traffic detection using hidden markov models[C]// The Workshop on Cyber Security and Information Intelligence Research. USA, Oak Ridge, Tennessee, ACM, 2011:31.
- [9] VENKATESH G K, NADARAJAN R A. HTTP botnet detection using adaptive learning rate multilayer feed-forward neural network[C]// Information Security Theory and Practice. Security, Privacy and Trust in Computing Systems and Ambient Intelligent Ecosystems. Berlin, Heidelberg, Springer, 2012:38-48.
- [10] BARTHAKUR P, DAHAL M, GHOSE M K. An efficient machine learning based classification scheme for detecting distributed command & control traffic of P2P botnets[J]. International Journal of Modern Education & Computer Science, 2013, DOI: 10. 5815/ijmecs. 2013. 10. 02.
- [11] YAVANOGLU O, AYDOS M. A review on cyber security datasets for machine learning algorithms[C]// IEEE International Conference on Big Data, Symposium on Data Analytics for Advanced Manufacturing. Boston, MA, USA, IEEE, 2017. DOI: 10.1109/Big-Data. 2017. 8258167.
- [12] KOZIK R, CHORAŠ M. Pattern extraction algorithm for netflow-based botnet activities detection[J]. Security & Communication Networks, 2017:1-10. DOI:10. 1155/2017/6047053.
- [13] HOQUE N, BHATTACHARYYA D K, KALITA J K. FFSc: a novel measure for low-rate and high-rate DDoS attack detection using multivariate data analysis [M]. John Wiley & Sons, Inc. 2016.
- [13] BHAMARE D, SALMAN T, SAMAKA M, et al. Feasibility of supervised machine learning for cloud security[C]// International Conference on Information Science and Security. Changsha, China, IEEE, 2017: 1-5.
- [15] 韦泽鲲, 夏靖波, 张晓燕, 等. 基于随机森林的流量多特征提取与分类研究[J]. 传感器与微系统, 2016, 35 (12):55-59.
- [16] GARCÍA S, GRILL M, STIBOREK J, et al. An empirical comparison of botnet detection methods [J]. Computers & Security, 2014(45):100-123.
- [17] JIN J, YAN Z, GENG G, et al. Botnet domain name detection based on machine learning[C]// International Conference on Wireless, Mobile and Multi-Media. Beijing, China, IET, 2016.
- [18] STEVANOVIC M, PEDERSEN J M. On the use of machine learning for identifying botnet network traffic [J]. AALOBURG University, 2016, 4(2):1-32, DOI: 10.13052/jcsm2245-1439. 421

作者简介:

肖琦女, (1993-), 硕士. 研究方向为网络安全、机器学习. E-mail:15751153961@163.com.

苏开宇男, (1978-), 硕士. 研究方向为计算机网络.

(上接第42页)

- [14] LIU Y, CHEN J. A K-means based firefly algorithm for localization in sensor networks[J]. International Journal of Parallel Emergent & Distributed Systems, 2018(12):1-16.
- [15] SANKALAP A, SATVIR S. An improved butterfly optimization algorithm with chaos[J]. Advanced Science, 2016, 8(9):711-717.

作者简介:

李田来男, (1982-), 博士, 讲师. 研究方向为无线网络、智能计算. E-mail:litianlai@sdu.edu.cn.

刘方爱男, (1962-), 博士, 教授, 博士生导师. 研究方向为人工智能、大数据分析.