



一种基于 Xgboost 的 Skype 时间式隐信道检测方法

常婷婷¹, 翟江涛², 戴跃伟²

(1. 江苏科技大学 电子信息学院, 江苏 镇江 212003; 2. 南京信息工程大学 电子与信息工程学院, 南京 210044)

摘 要: 时间式隐信道利用数据包的包间时延来传递秘密信息, 受网络时间特性复杂性的影响, 网络隐信道的检测率低且虚警率较高。提出一种利用 Xgboost 模型的 Skype 时间式隐信道检测方法。在传统提取 Skype 时间序列的 Markov 转移特性、信息熵、包间时延的均值与方差、DCT 系数、 ε -相似度等特征的基础上, 增加峰态、偏态和标准偏差的差值 3 种特征, 以准确了解包间时延分布并进行筛选排查, 同时采用五折交叉验证法结合无重复抽样技术, 使每次迭代时每个样本点只有一次被划入训练集或测试集, 最终通过 Xgboost 算法进行判决和检测。实验结果表明, 与 BP 神经网络方法相比, 该方法检测率更高且虚警率更低。

关键词: 网络隐信道; 时间式隐信道; 五折交叉验证; 神经网络; Xgboost 算法

开放科学(资源服务)标志码(OSID):



中文引用格式: 常婷婷, 翟江涛, 戴跃伟. 一种基于 Xgboost 的 Skype 时间式隐信道检测方法[J]. 计算机工程, 2021, 47(7): 88-94.

英文引用格式: CHANG T T, ZHAI J T, DAI Y W. An Xgboost-based method for detecting covert timing channel of Skype[J]. Computer Engineering, 2021, 47(7): 88-94.

An Xgboost-based Method for Detecting Covert Timing Channel of Skype

CHANG Tingting¹, ZHAI Jiangtao², DAI Yuewei²

(1. College of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, China;

2. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China)

[Abstract] The covert timing channel exploits the packet delay to transmit secret information. Due to the complexity of the temporal features of network, the false alarm rate of the covert channels is high, masking the detection of the true targets. An Xgboost-based method for detecting covert timing channel of Skype is proposed. On the basis of the existing methods, which extract the Markov transition features, information entropy, mean and variance of the delay between packets, DCT coefficient, and ε -similarity of the Skype time series, the proposed method adds another three features, including the peak state, skewness and difference of standard deviation, so as to accurately understand the distribution of delay between packets and to screen the targets. At the same time, the method of five-fold cross verification is combined with the non-repeating sampling technology, so that every sample point is classified into training set or test set for only once in each iteration. Finally, the Xgboost algorithm is used for judgment and detection. Experimental results show that compared with the BP neural network method, the proposed method has higher detection rate and lower false alarm rate.

[Key words] covert network channel; covert timing channel; five-fold cross validation; neural network; Xgboost algorithm

DOI: 10.19678/j.issn.1000-3428.0057925

0 概述

随着互联网技术的快速崛起, 中国已发展为 5G 网络大国, 与相对封闭的传统移动通信系统相比, “5G+移动互联网”大数据背景下人和物的连接更紧密, 但同时也造成网络攻击和恶意代码出现的频率大幅提高, 给网络用户隐私数据保护、移动办公和国

家基础网络设施安全带来重大影响。2019 年, 美国阿拉斯加拉文航空公司宣布其计算机网络受到恶意攻击, 并在假日出行高峰期取消了至少 6 班次航班, 影响到近 260 名乘客的正常出行。同年, 美国路易斯安那州新奥尔良市遭到网络攻击, 政府在当日宣布该市进入紧急状态。随着网络攻击出现频率的上升, 网络安全维护成为研究人员关注的热点。

基金项目: 国家自然科学基金(61702235, 61602247, U1636117)。

作者简介: 常婷婷(1994—), 女, 硕士研究生, 主研方向为信息安全; 翟江涛, 副教授、博士; 戴跃伟, 教授、博士、博士生导师。

收稿日期: 2020-03-31 **修回日期:** 2020-06-08 **E-mail:** 1151757929@qq.com

网络隐蔽通信是继加密技术后一种新兴的信息传输安全技术,其根据隐蔽信息隐藏方式的不同分为存储式隐蔽通信和时间式隐蔽通信。存储式隐蔽通信主要采用向网络协议的冗余位中嵌入 IP 头的扩展与填充段^[1-3]、IP 标志符^[4-5]等隐蔽信息来构建存储式隐信道,由于网络数据包对上述字段内容的检查不严格,因此在其中嵌入此类信息不易被发现。除了这种传统的存储式隐蔽通信外,近年来还出现基于多链路传输序列的隐信道^[6]、基于 DNS 协议的隐信道^[7]等新型存储式隐蔽通信。多链路传输序列的隐信道构建隐蔽通道的机制不再与网络协议冗余位有关,仅与数据包的时间特性有关,这与时间式隐蔽通信类似,但因为其构建方法是基于数据包的到达序列编码,与包间时延无关,所以其本质仍属于存储式隐信道,由于其兼具时间式隐蔽通信的隐蔽性与存储式隐蔽通信的稳定性,因此具有良好的实用价值。DNS 协议在网络运行中占有重要地位,一般不会被防火墙等安全系统阻拦,因此 DNS 协议是实现隐蔽通信的常用手段。2019 年,云服务商巨头亚马逊公司 AWS DNS 服务器遭到 DDoS 攻击,攻击者利用垃圾网络流量堵塞系统,造成服务器无法访问。此次攻击持续 15 小时,大量数据包阻塞了 DNS 系统,其中一些合法的域名请求被释放以缓解问题,由于网站和应用软件尝试联系 S3 存储桶等亚马逊后端托管的系统可能失败,从而导致用户会看到出错信息或空白页面。

时间式隐信道通常利用数据包的包间时延特性来传递秘密信息,由于其不改变数据包内部信息,因此隐蔽性较存储式隐信道更高^[8-10]。2013 年,美国将该方法应用于匿名网络节点追踪。时间式隐信道一般以 on/off 和 delay 模式来模拟真实网络传输的包间间隔以进行隐蔽信息传输^[11],在数据传输过程中 IP 报文被存储转发的情况下,目前常用的检测算法会失效。此外,还有 model-based 模式的隐信道^[12-13],其主要通过拟合现实通信时的数据模型来构建隐秘信道。model-based 模式下的隐蔽通信模型具有更好的隐蔽性,且由于网络的时间特性较复杂,因此对该网络隐信道的检测更困难。其中,针对 Skype 流量的隐写较大的情况,研究人员提出一种隐蔽信道检测算法^[14],先对获取的 Skype 流量进行基于 Erlang 模型的拟合,再利用 Walsh 编码构建隐蔽通道,采用传统数据随机分组的方式,将 80% 的数据作为训练数据,20% 的数据作为测试数据,并采用 BP 神经网络方法进行检测。该方法虽然检测率较高,但也具有较高的虚警率。

针对上述隐信道,在处理训练数据和测试数据时,可提取峰态、偏态以及标准偏差的差值等特征,其中偏态和峰态用于观察包间时延的整体分布情况。由于在基于 Erlang 模型构建隐信道的过程中,

在对应区间随机选取一个包间隔(IPD)会破坏正常通信时包间时延的分布,因此峰态和偏态作为特征能起到较好的筛选排查作用。标准偏差的差值可用于研究较小范围内包间时延之间的关系,文献[15]将其引入时间式隐信道的检测算法并取得了较好的检测效果,因此可选取标准偏差的差值作为训练特征,然后采用五折交叉验证法结合无重复抽样技术,使得每次迭代过程中每个样本点只有一次被划入训练集或测试集。同时,找到使得模型泛化性能最优的超参值,并在全部训练集上重新训练模型,使用独立测试集对模型性能做出最终评价,以保证分类精度的准确性并有效避免模型产生过拟合现象。

本文提出一种 Skype 时间式隐信道检测方法。在传统方法的基础上增加峰态、偏态以及标准偏差的差值 3 种特征,并采用 Xgboost 模型判决^[16-17]和检测待测数据,利用一阶导数和二阶导数将树模型的复杂度作为目标函数的正则项考虑,已避免出现过拟合现象。

1 基于 Skype 的时间式网络隐写算法

对正常数据的累积分布函数(CDF)进行拟合,可实现隐秘数据的嵌入且不易被检测^[13]。因此,本文以常用的 Skype 通信流量为载体,拟合出 CDF 模型。基于 Skype 的时间式网络隐写算法流程如图 1 所示。

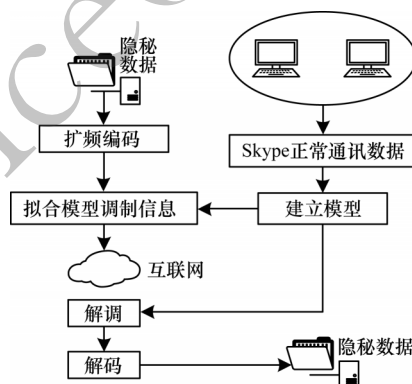


图1 基于 Skype 的时间式网络隐写算法流程

Fig.1 Procedure of timing network steganography algorithm based on Skype

该算法具体流程如下:

1) 获取正常环境下 Skype 通信的流量数据,建立 CDF 模型(与 Erlang 模型类似),其累积分布函数 $P(x; m, \lambda x)$ 的计算公式如下:

$$P(x; m, \lambda x) = \frac{\gamma(m, \lambda x)}{(K-1)!} \quad (1)$$

$$\gamma(m, \lambda x) = \int_0^{\lambda x} t^{m-1} e^{-t} dt \quad (2)$$

其中: x 为包间时延, $m=1$ 为图形参数, λ 为速率参数, K 为扩频编码时使用正交信道的数目。

2)采用 N 阶Walsh码进行二进制扩频编码如下:

$$s = \sum_{k=1}^K b_k \times c_k \quad (3)$$

其中: c_k 为 N 阶Walsh码。

3)将正常通信数据的CDF划分为 $F=3$ 个区间,每个区间再分为 $2m+1$ 个小区间,以保证每个区间之间保持最小的汉明距离。 s 中不同的值依次与CDF的 F 个小区间对应,并在相应区间内选择一个IPD。

2 本文方法

2.1 特征提取

本文检测对象是对正常Skype数据的CDF模型进行拟合实现的隐写,因此较一般隐信道具有更强的抗检测性。信息熵作为目前有效的时式隐信道检测手段,与上述隐写方式相结合的检测效果不佳,因此本文提取以下7种特征组成特征矩阵进行分类器的训练。

1)基于时间序列的马尔可夫(Markov)转移矩阵。设 t_i 为第 i 个包间间隔, t_{i+1} 为第 $i+1$ 个包间间隔,如果 $t_{i+1} < t_i$,则 $m_i=0$;否则 $m_i=1$,由此可得到1条马尔可夫链。由式(4)可得到马尔可夫转移矩阵的元素:

$$P_{ab} = \text{num}(X_i = m | X_{i+1} = n) / (N - 1) \quad (4)$$

其中: $\text{num}(X_i = m | X_{i+1} = n)$ 表示马尔可夫链中当前状态数为 m 且下一个状态数为 n 的情况总数, N 为马尔可夫链包含的总状态数。

由于隐蔽信息的信息熵根据特定的规律随机调制,使得马尔可夫转移矩阵中的4个元素相对稳定,但是在现实网络中,由于受到各方面因素的影响,马尔可夫转移矩阵中的元素可能会受到干扰,与含密数据的马尔可夫转移矩阵中元素有所不同,因此将其作为一种提取特征。

2)信息熵。熵可反映出一个整体的不确定性以及信息容量。由于时式隐蔽通信会使IPD整体分布发生变化,使其不同于正常通信的信息熵值,且对于传统时式隐信道而言,基于信息熵的检测是一种常用的检测手段,因此将信息熵作为一种提取特征,具体操作过程如下:

(1)分别从正常数据和含密数据中提取 N 个数据包,分为 $w=1\,000$ 个窗口。

(2)将正常数据的IPD分为大小相等的 L 块,计算IPD落在每块中的概率。

(3)根据式(5)计算每个窗口的信息熵,设置检验阈值,比较测试数据的信息熵值和检验阈值来判断数据是否含密,计算公式如下:

$$H_n = - \sum_{i=1}^L P_{ni} \ln P_{ni} \quad (5)$$

其中: P_{ni} 为时延信息落在每个块中的概率。

3)均值与方差。包间时延的均值和方差与当前的网络环境密切相关。当网络质量较好时,正常数据的包间时延均值一般小于含密数据,此时方差较小;当网络出现拥塞时,正常数据的包间时延均值会随着包间时延的增大而增加,方差也较大。由此可知,正常数据包间时延均值与方差的波动一般比较大。含密数据的包间时延通常按照一定规律随机选择,其均值和方差较正常数据波动更小,因此将均值和方差作为一种提取特征,其计算公式分别如下:

$$A = (t_1 + t_2 + \dots + t_n) / n \quad (6)$$

$$s^2 = [(t_1 - A)^2 + (t_2 - A)^2 + \dots + (t_n - A)^2] / n \quad (7)$$

其中: n 为样本时延总数。

4)DCT系数。传统隐信道的检测仅注重数据之间的时域特性,忽视了频域特性的重要性。目前较常用的时频域转换方法属于DCT变换,研究人员将DCT系数应用于隐蔽通道检测取得较好的效果,因此将DCT系数作为一种提取特征,相关计算公式如下:

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi(2m+1)q}{2N} \quad (8)$$

$$\alpha_p = \begin{cases} 1/\sqrt{M}, & p=0 \\ \sqrt{2/N}, & 1 \leq p \leq M-1 \end{cases} \quad (9)$$

$$\alpha_q = \begin{cases} 1/\sqrt{N}, & q=0 \\ \sqrt{2/N}, & 1 \leq q \leq N-1 \end{cases} \quad (10)$$

其中: $0 \leq p \leq M-1, 0 \leq q \leq N-1, M$ 和 N 分别为 A 的行数和列数; B 为变换后的矩阵。

5) ε -相似度。由式(11)可计算出相邻两个数据包之间的差异率 dif , dif 小于 ε 的包间时延个数占总包间时延个数的比值称为 ε -相似度 E ,由式(12)计算得到。

$$\text{dif}_i = |t_i - t_{i-1}| / t_i, 1 \leq i \leq n-1 \quad (11)$$

$$E = \text{num}(\text{dif} < \varepsilon) / (n-1) \quad (12)$$

其中: $\text{num}(\text{dif} < \varepsilon)$ 表示差异率小于 ε 的包间时延总数。

本文采用模型拟合方法构建隐蔽信道,对含密数据构建的CDF模型与现实数据的CDF模型相似,但是 ε -相似度是基于邻近的包间间隔特性进行分析,含密数据与真实数据之间可能会存在较明显的差异,因此将 ε -相似度作为一种提取特征。

6)峰态(K)和偏态(S)。偏态和峰态用于观察包间时延的整体分布情况,在基于Erlang模型进行隐写的过程中,在对应区间随机选取一个IPD,难免会破坏正常通信时包间时延的分布,因此将峰态和偏态作为一种提取特征,其计算公式如下:

$$K = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4}{\left(\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \right)^2} - 3 \quad (13)$$

$$S = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3}{\left[\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \right]^{3/2}} \quad (14)$$

其中: \bar{x} 为样本的平均值。

7) 包间时延标准差的差值 (C)。在研究较小范围内包间时延之间的关系时, 研究人员将包间时延标准差引入时间式隐信道检测算法取得较好的检测效果^[15], 本文取标准差的差值作为一种分类器的训练特征。分别从正常数据和含密数据中提取 N 个数据包并分为 $w=1\ 000$ 个窗口, 再将这 w 个窗口分为 $w/2$ 个窗口, 分别求得各自的标准偏差 σ_i 和 σ_j , 再计算两个窗口之间标准差的差值 C , 计算公式如下:

$$C = \frac{|\sigma_i - \sigma_j|}{\sigma_i}, i \neq j \quad (15)$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (16)$$

其中: \bar{x} 为样本的平均值。

2.2 Xgboost 算法

2.2.1 梯度提升树算法

梯度提升树 (GBDT) 算法是 2001 年 FRIEDMAN 等提出的一种 boosting 算法^[18], 其由多棵决策树组合而成, 是通过迭代产生的一种决策树算法, 并将所有决策树的统计结果作为最终预测的结果, GBDT 算法的基本原理如图 2 所示。

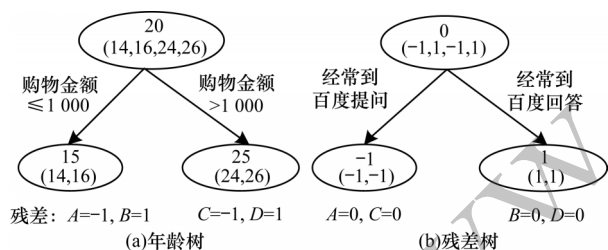


图2 GBDT 算法的基本原理

Fig.2 Basic principle of GBDT algorithm

对于回归树的分裂结点, 如果是在平方损失函数中, 则是对残差的拟合; 如果是在一般损失函数中 (梯度下降), 则是对残差近似值的拟合。当划分分裂结点时, 需列举出所有的特征值, 然后选取划分点并统计每棵树的预测结果, 统计结果即为最终的预测结果。

2.2.2 Xgboost 算法原理

Xgboost 是 2014 年诞生的用于梯度提升树算法的机器学习函数库^[19], 该函数库因学习效果好和训练速度快获得广泛关注。在 2015 年 KAGGLE 竞赛中获胜的 29 个算法中, 有 17 个使用了 Xgboost, 相较梯度提升

算法在另一个常用机器学习库 scikit-learn 中的实现情况, Xgboost 的性能有 10 倍以上的提升。此外, Xgboost 将损失函数从平方损失推广到二阶可导的损失, 加入了正则化项, 支持列抽样, 能对连续型特征进行处理, 同时可以利用数据的稀疏性, 当数据量大时有效提高硬盘吞吐率。目前 Xgboost 算法被广泛用于企业破产风险评估、物联网消费人群减少评估、网络安全风险评估^[20-21]等领域。

Xgboost 算法是在 GBDT 算法的基础上略加改进得到, 其与 GBDT 算法存在一些差异^[22]。GBDT 算法只采用了一阶导数进行优化, 而 Xgboost 算法在优化时将一阶导数和二阶导数相结合, 引入树模型的复杂度, 并将其作为目标函数里的正则项, 可有效避免发生过拟合。Xgboost 算法中 boosting 树模型结构如图 3 所示 (其中, $f(\square)=2.0+0.9=2.9$, $f(\circ)=-1.0+0.9=-0.1$)。

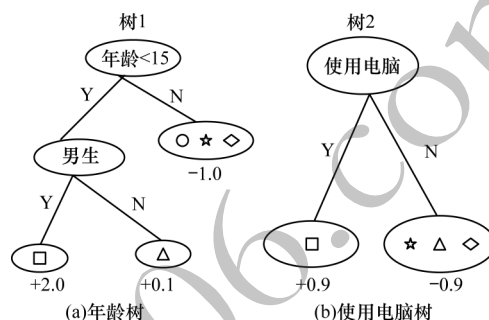


图3 Xgboost 算法中 boosting 树模型结构

Fig.3 Structure of boosting tree model in Xgboost algorithm

Xgboost 算法的具体实现过程如下:

1) 设 Xgboost 模型第 t 轮的目标函数为:

$$L(f_t) = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + C \quad (17)$$

其中: l 为第 t 轮的损失项; Ω 为模型中决策树的正则项, 其计算公式如下:

$$\Omega(f_t) = \gamma \cdot T_t + \lambda \frac{1}{2} \sum_{j=1}^T w_j^2 \quad (18)$$

2) 由泰勒展开公式得到:

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \quad (19)$$

设以下条件成立:

$$g_t = \frac{\partial l(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}} \quad (20)$$

$$h_t = \frac{\partial^2 l(y_i, \hat{y}_i^{t-1})}{\partial \hat{y}_i^{t-1}^2} \quad (21)$$

将式 (18)~式 (21) 代入式 (17) 得到:

$$L(f_t) \approx \sum_{j=1}^T \left[\left(\sum g_i \right) w_j + \frac{1}{2} \left(\sum h_i + \lambda \right) w_j^2 \right] + \gamma \cdot T + C \quad (22)$$

3) 对式 (22) 进行求解可得最优系数与目标函数最优值分别如下:

$$w_j^* = - \frac{\sum g_i}{\sum h_i + \lambda} \quad (23)$$

$$L(f_i) = - \frac{1}{2} \sum_{j=1}^T \frac{(\sum g_i)^2}{\sum h_i + \lambda} + \gamma \cdot T \quad (24)$$

4)根据式(23)和式(24)的最优结果确定最优决策树结构,进而进行计算和预测。

3 实验与结果分析

3.1 实验过程

为保证实验数据的一般性和实验结果的可靠性,本文实验所用数据是在教育网-教育网、教育网-中国镇江移动有线网、中国镇江移动有线网-中国六安电信有线网3种不同的网络环境下抓取获得。在教育网-教育网环境下登录 Skype 建立语音连接,分别抓取正常流量数据 60 326 条和 65 200 条并编号为 M1 和 M2;在教育网-中国镇江移动有线网环境下登录 Skype 建立语音连接,分别抓取正常流量数据 34 465 条和 46 519 条并编号为 N1 和 N2;在中国镇江移动有线网-中国六安电信有线网环境下登录 Skype 建立语音连接,抓取正常流量数据 65 178 条,编号为 P。按照本文隐信道构建方法模拟生成含密流量数据 Q1(40 000 条)以及 Q2(4 000 条)。

本文实验流程如图4所示,具体如下:

1)将正常数据与含密数据混合后按大小为 $w=1\ 000$ 的窗口进行分割,两种数据用标识符标记,正常数据标记为 0,含密数据标记为 1。

2)在 w 个数据中提取 7 种特征,形成 1 个 13 维数组,数组中包含马尔可夫转移矩阵的 4 个元素、熵值、包间时延均值、峰态、偏态、包间时延方差、DCT 系数最大值、DCT 系数最小值、 ε -相似度($\varepsilon=0.5$)以及标准偏差的差值。

3)针对上述数据集预处理得到的实验数据,采用五折交叉验证,同时为证明在本实验背景下 Xgboost 算法相较 Logistic 回归算法、决策树算法、随机森林算法等目前较流行的算法具有更好的适用性,使用上述算法分别进行训练和建模预测。

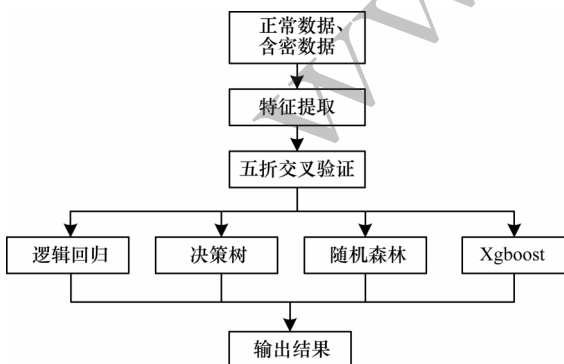


图4 本文实验流程

Fig.4 Procedure of the experiment in this paper

五折交叉验证步骤如图5所示,具体如下:

1)将实验数据平均分为 5 份,在分割过程中保证每份数据均含有两种标签样本。

2)保留 1 份单独的数据样本作为测试数据,其他 4 份数据样本用于对上述 4 种分类器逐一进行训练,交叉验证重复 5 次,每个样本数据测试 1 次,当输出结果为 1 时表示为含密数据,当输出结果为 0 时表示为正常数据。

3)计算 5 次结果的平均值作为各个分类器的评价指标最终结果。



图5 五折交叉验证原理图

Fig.5 Schematic diagram of five-fold cross validation

3.2 结果分析

本文采用基于 Xgboost 的方法(以下称为本文方法)对待测数据进行分组实验,并将所得结果与 BP 神经网络方法(以下称为 BP 方法)结果^[10]进行对比。

3.2.1 对比实验结果

分别对单组数据、多组数据以及不同实验环境数据进行检测,以下为对比实验的结果。

1)单组数据检测。将 M1 与 Q1(单组数据 1)、N1 与 Q1(单组数据 2)分别作为原始数据通过五折交叉验证进行 Xgboost 判决,得到实验结果如表 1 和表 2 所示。

表1 单组数据 1 检测结果对比

Table 1 Comparison of detection results of single group data 1

方法	检测率	虚警率
本文方法	0.991 2	0.011 1
BP 方法	1.000 0	0.111 7

表2 单组数据 2 检测结果对比

Table 2 Comparison of detection results of single group data 2

方法	检测率	虚警率
本文方法	0.990 0	0.013 3
BP 方法	0.998 3	0.041 7

2)多组数据检测。将 M1、M2、N1、N2、Q1、Q2 作为原始数据通过五折交叉验证进行 Xgboost 判决,得到实验结果如表 3 所示。

表 3 多组数据检测结果对比

Table 3 Comparison of detection results of multi-group data

方法	检测率	虚警率
本文方法	0.997 7	0.006 0
BP 方法	0.997 5	0.044 6

3)不同实验环境数据检测。将 M1、N1、P、Q1、Q2 作为原始数据通过五折交叉验证进行 Xgboost 判决,得到实验结果如表 4 所示。

表 4 不同环境数据检测结果对比

Table 4 Comparison of detection results of different environmental data

方法	检测率	虚警率
本文方法	0.995 7	0.006 7
BP 方法	0.999 2	0.024 3

本文添加了峰态、偏态以及标准偏差的差值 3 种特征,再利用五折交叉验证和 Xgboost 算法,根据不同实验得到了相应的检测率和虚警率。在检测率方面,虽然本文方法偶尔略低于 BP 方法,但检测率依然保持在 0.999 0 以上,基本与 BP 方法检测率相同;在虚警率方面,本文方法较 BP 方法最多降低约 10 个百分点。总体而言,本文方法检测率更高且虚警率更低。

3.2.2 适用性实验结果

为进一步验证 Xgboost 算法在本文实验研究背景下的适用性,另外选取精确率(P)、召回率(R)、精确率和召回率的调和均值($F1$)、准确率(A)这 4 个性能指标,加上检测率和虚警率共采用 6 个性能指标来比较 Xgboost 分类器和逻辑回归、决策树、随机森林等当前较流行分类器的分类效果。

对二分类问题而言,如果实例是正类且被预测为正类,则称为真正类(True Positive, TP);如果实例是负类且被预测成正类,则称为假正类(False Positive, FP);如果实例是负类且被预测成负类,则称为真负类(True Negative, TN);如果实例是正类且被预测成负类,则称为假负类(False Negative, FN)。准确率 A 用于描述分类器的分类效果,准确率越大,分类器分类效果越好。当 $A=1$ 时,该分类器是完美分类器;当 $0.5<A<1$ 时,该分类器的结果优于随机猜测结果;当 $A=0.5$ 时,该分类器的结果与随机猜测结果接近;当 $A<0.5$ 时,该分类器的结果比随机猜测结果要差。

相关计算公式如下:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$
(25)

$$P = \frac{TP}{TP + FP}$$
(26)

$$R = \frac{TP}{TP + FN}$$
(27)

$$\frac{2}{F1} = \frac{1}{P} + \frac{1}{R}$$
(28)

其中:TP 为正类预测正确的个数,FP 为负类预测错误的个数,TN 为负类预测正确的个数,FN 为正类预测错误的个数。

各分类器的性能指标如表 5 所示。由表 5 可见,Xgboost 分类器和随机森林分类器均有较好的分类效果,决策树分类器次之,逻辑回归分类器效果最差。逻辑回归分类器的准确率虽然达到 0.987 51,但是另外 5 项指标远低于其他 3 种分类器,其分类效果最差。决策树分类器和随机森林分类器的各项指标都较好,但 Xgboost 分类器的检测率相较决策树分类器提升约 0.5 个百分点,较随机森林分类器提升 0.1 个百分点。Xgboost 分类器的召回率略高于决策树分类器,相较随机森林分类器提升约 0.1 个百分点。Xgboost 分类器的调和均值相对决策树分类器提升约 2 个百分点,相较随机森林分类器提升约 1 个百分点。Xgboost 分类器的准确率为 1.000 00,在本文中分类效果接近理想状态,较决策树分类器提升约 2 个百分点。Xgboost 分类器的虚警率在 4 个分类器中最低。虽然 Xgboost 分类器精确率略低于随机森林分类器,但从总体来看,Xgboost 分类器的分类效果最佳。

表 5 不同分类器的性能指标

Table 5 Performance indicators of different classifiers

分类器	检测率	P	R	$F1$	A	虚警率
决策树分类器	0.990 01	0.971 71	0.988 56	0.979 71	0.989 61	0.012 90
逻辑回归分类器	0.873 13	0.916 79	0.491 68	0.628 32	0.987 51	0.083 20
随机森林分类器	0.994 32	1.000 00	0.977 45	0.988 32	1.000 00	0.007 66
Xgboost 分类器	0.995 73	0.993 33	0.988 89	0.990 90	1.000 00	0.006 67

4 结束语

本文提出一种利用 Xgboost 算法的 Skype 时间式隐信道检测方法。基于正常通信数据的 CDF 模型建立网络隐蔽通道提取数据特征并构建特征向量,采用五折交叉验证法和 Xgboost 算法进行判决。同时,找到使模型泛化性能最优的超参值,利用独立测

试集对模型性能进行评价,以提高分类精度并避免产生过拟合现象。实验结果表明,该方法较BP神经网络方法检测率更高且虚警率更低。后续将在本文方法的基础上对新型时间式隐信道检测进行研究,进一步提高检测率。

参考文献

- [1] CHEDDAD A, CONDELL J, CURRAN K, et al. Digital image steganography: survey and analysis of current methods[J]. *Signal Processing*, 2010, 90(3): 727-752.
- [2] COX I, MILLER M, BLOOM J, et al. Digital watermarking and steganography [EB/OL]. [2020-01-05]. https://booksite.elsevier.com/samplechapters/9780123725851/Sample_Chapters/01~Front_Matter.pdf.
- [3] 董丽鹏, 陈性元, 杨英杰, 等. 网络隐蔽信道实现机制及检测技术研究[J]. *计算机科学*, 2015, 42(7): 216-221, 244. DONG L P, CHEN X Y, YANG Y J, et al. Research on the implementation mechanism and detection technology of network hidden channel[J]. *Computer Science*, 2015, 42(7): 216-221, 244. (in Chinese)
- [4] MAZURCZYK W, KARAS M, SZCZYPIORSKI K. SkyDe: a Skype-based steganographic method [J]. *International Journal of Computers, Communications and Control*, 2013, 8(3): 432-443.
- [5] YU J M, ZHU C Y, TANG S H, et al. Deepflow: hiding anonymous communication traffic in P2P streaming networks[J]. *Wuhan University Journal of Natural Sciences*, 2014, 19(5): 417-425.
- [6] FRCZEK W, SZCZYPIORSKI K. Perfect undetectability of network steganography[J]. *Security & Communication Networks*, 2016, 9(15): 2998-3010.
- [7] DRZYMATA M, SZCZYPIORSKI K, URBANSKI M L. Network steganography in the DNS protocol [J]. *International Journal of Electronics & Telecommunications*, 2016, 62(4): 343-346.
- [8] 何磊, 郭晓军, 张春玉. 一种基于时间隐蔽信道的WSN认证算法[J]. *实验技术与管理*, 2014, 52(9): 59-61, 86. HE L, GUO X J, ZHANG C Y. An authentication algorithm based on timing covert channel in WSN[J]. *Experimental Technology and Management*, 2014, 52(9): 59-61, 86. (in Chinese)
- [9] REZAEI F, HEMPEL M, RAKSHIT S M. Automated covert channel modeling over a real network platform[C]// *Proceedings of International Wireless Communications & Mobile Computing Conference*. Washington D. C., USA: IEEE Press, 2014: 559-564.
- [10] HOVHANNISYAN H, LU K J, WANG J P. A novel high-speed IP-timing covert channel: design and evaluation[C]// *Proceedings of IEEE International Conference on Communications*. Washington D. C., USA: IEEE Press, 2015: 7198-7203.
- [11] XIA Z H, WANG X H, SUN X M, et al. Steganalysis of least significant bit matching using multi-order differences [J]. *Security & Communication Networks*, 2014, 7(8): 1283-1291.
- [12] LU X R, HUANG L S, YANG W. Concealed in the Internet: a novel covert channel with normal traffic imitating [C]// *Proceedings of 2016 IEEE Conferences on Ubiquitous Intelligence & Computing*. Washington D. C., USA: IEEE Press, 2017: 125-136.
- [13] ARCHIBALD R, GHOSAL D. Design and analysis of a model-based covert timing channel for skype traffic [C]// *Proceedings of IEEE Conference on Communications and Network Security*. Washington D. C., USA: IEEE Press, 2015: 236-244.
- [14] 李萌, 翟江涛, 戴跃伟. 一种针对Skype时间特性拟合的网络隐写检测方法[J]. *计算机应用研究*, 2018, 35(6): 1803-1807. LI M, ZHAI J T, DAI Y W. A network steganographic detection method for Skype time feature fitting [J]. *Application Research of Computers*, 2018, 35(6): 1803-1807. (in Chinese)
- [15] CABUK S, BRODLEY C E, SHIELDS C. IP covert timing channels: design and detection [C]// *Proceedings of ACM Conference on Computer & Communications Security*. New York, USA: ACM Press, 2004: 122-136.
- [16] 崔艳鹏, 史科杏, 胡建伟. 基于Xgboost算法的Web shell检测方法研究[J]. *计算机科学*, 2018, 45(z1): 375-379. CUI Y P, SHI K X, HU J W. Research on Web shell detection method based on Xgboost algorithm [J]. *Computer Science*, 2018, 45(z1): 375-379. (in Chinese)
- [17] CHEN T, HE T, BENESTY M. Xgboost: extreme gradient boosting [EB/OL]. [2020-01-05]. <http://mysql.orst.edu/pub/cran/web/packages/Xgboost/vignettes/Xgboost.pdf>.
- [18] CHEN T, GUESTRIN C. Xgboost: a scalable tree boosting system [C]// *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press, 2016: 785-794.
- [19] ZIEBA M, TOMCZAK S K, TOMCZAK J M. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction [J]. *Expert Systems with Applications*, 2016, 58: 93-101.
- [20] 王重仁, 韩冬梅. 基于社交网络分析和Xgboost算法的互联网客户流失预测研究[J]. *微型机与应用*, 2017, 36(23): 62-65. WANG C R, HAN D M. Prediction of Internet customer loss based on social network analysis and Xgboost algorithm [J]. *Microcomputer and Application*, 2017, 36(23): 62-65. (in Chinese)
- [21] 赵天傲, 郑山红, 李万龙, 等. 基于Xgboost的信用风险分析的研究[J]. *软件工程*, 2018, 21(6): 29-32. ZHAO T N, ZHENG S H, LI W L. Research on credit risk analysis based on Xgboost [J]. *Software Engineering*, 2018, 21(6): 29-32. (in Chinese)
- [22] 刘金硕, 刘必为, 张密, 等. 基于GBDT的电力计量设备故障预测[J]. *计算机科学*, 2019, 46(6): 392-396. LIU J S, LIU B W, ZHANG M, et al. Fault prediction of power metering equipment based on GBDT [J]. *Computer Science*, 2019, 46(6): 392-396. (in Chinese)

编辑 宋 圆