

基于偏二叉树 SVM 多分类算法的应用层 DDoS 检测方法

张斌^{1,2}, 刘自豪^{1,2}, 董书琴^{1,2}, 李立勋^{1,2}

(1. 信息工程大学, 河南 郑州 450001;

2. 河南省信息安全重点实验室, 河南 郑州 450001)

摘要: 针对基于流量特征的应用层 DDoS 检测方法侧重于检测持续型应用层 DDoS 攻击, 而忽略检测上升型与脉冲型应用层 DDoS 攻击的问题, 提出一种综合检测多类型应用层 DDoS 攻击的方法。首先通过 Hash 函数及开放定址防碰撞方法, 对多周期内不同源 IP 地址建立索引, 进而实现 HTTP GET 数的快速统计功能, 以支持对刻画数据规模、流量趋势及源 IP 地址分布差异所需特征参数的实时计算; 然后采用偏二叉树结构组合 SVM 分类器分层训练特征参数, 并结合遍历与反馈学习的方法, 提出基于偏二叉树 SVM 多分类算法的应用层 DDoS 检测方法, 快速区分出非突发正常流量、突发正常流量及多类型 App-DDoS 流量。实验表明, 所提算法通过划分检测类型、逐层训练检测模型, 与传统基于 SVM、Navie Bayes 的检测方法相比, 具有更高的检测率与更低的误检率, 且能有效区分出具体攻击类型。

关键词: 应用层 DDoS 攻击; HTTP GET 统计模型; 流量特征参数; SVM 多分类器

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2018020

App-DDoS detection method using partial binary tree based SVM algorithm

ZHANG Bin^{1,2}, LIU Zihao^{1,2}, DONG Shuqin^{1,2}, LI Lixun^{1,2}

1. Information and Engineering University, Zhengzhou 450001, China

2. Key Laboratory of Information Security, Zhengzhou 450001, China

Abstract: As it ignored the detection of ramp-up and pulsing type of application layer DDoS (App-DDoS) attacks in existing flow-based App-DDoS detection methods, an effective detection method for multi-type App-DDoS was proposed. Firstly, in order to fast count the number of HTTP GET for users and further support the calculation of feature parameters applied in detection method, the indexes of source IP address in multiple time windows were constructed by the approach of Hash function. Then the feature parameters by combining SVM classifiers with the structure of partial binary tree were trained hierarchically, and the App-DDoS detection method was proposed with the idea of traversing binary tree and feedback learning to distinguish non-burst normal flow, burst normal flow and multi-type App-DDoS flows. The experimental results show that compared with the conventional SVM-based and naïve-Bayes-based detection methods, the proposed method has more excellent detection performance and can distinguish specific App-DDoS types through subdividing attack types and training detection model layer by layer.

Key words: App-DDoS attack, HTTP GET statistical model, flow feature parameter, SVM multi-classifier

收稿日期: 2018-01-05; 修回日期: 2018-02-02

通信作者: 刘自豪, liuzihao19937@126.com

基金项目: 河南省基础与前沿技术研究计划基金资助项目 (No.2014302903); 信息保障技术重点实验室开放基金资助项目 (No. KJ-15-109); 信息工程大学新兴科研方向培育基金资助项目 (No. 2016604703)

Foundation Items: The Basic and Advanced Technology Research Project of Henan Province (No.2014302903), The Open Foundation of Key Information Assurance Laboratory (No.KJ-15-109), The Cultivating Foundation of Emerging Research Direction of Information and Engineering University (No. 2016604703)

1 引言

分布式拒绝服务 (DDoS, distributed denial of service) 攻击通过控制僵尸主机发送大量数据分组, 耗尽被攻击目标网络或系统资源, 影响其正常工作, 已成为网络安全的主要威胁之一。

当前, 针对网络层、传输层 DDoS 攻击的检测防御技术不断完善, 应用层 DDoS (App-DDoS, application layer DDoS) 攻击成为研究的热点^[1,2]。App-DDoS 攻击与服务器建立正常的连接, 发送访问请求, 在网络层与传输层表现正常, 数据分组中 TCP 3 次握手及请求内容均不存在异常, 与正常用户集中访问 (flash crowd) 造成的表象极为相似, 能够有效地逃避网络层级的检测和过滤。如何实现 App-DDoS 的早期检测及过滤是当前研究的重点与难点。

App-DDoS 攻击检测的研究主要从日志与流量 2 个角度展开。其中, 由于日志在攻击作用于服务器之后才会产生, 基于日志挖掘的检测方法^[3-5]偏向于对攻击的事后挖掘, 而基于流量统计分析的检测方法实时性强, 更适用于 App-DDoS 攻击早期检测和过滤。当前基于流量统计分析的检测方法主要分为以下两类。1) 流量相关性建模的方法, 如通过 Jeffery 距离、Sibson 距离、Hellinger 距离 3 种抽象距离来度量网络流之间相似性的方法^[6]; 分别利用 Total variation 度量与 Bhattacharyya 度量评估流分布的差异度与相似度的方法^[7]; 通过流量相关系数度量网络流量间差异的方法^[8]等。此类方法侧重于通过实时流量数据间相似性与差异性进行区分, 忽略对历史数据的充分挖掘利用, 难以进行模型的学习更新。2) 基于机器学习的方法, 如将多层神经网络与遗传算法相结合, 对用户 HTTP GET 请求的熵与熵的方差进行训练建模的方法^[9]; 利用 AAR 模型将每个源 IP 地址平均 HTTP GET 请求数熵的时间序列映射为多维向量序列, 并使用支持向量机算法分类器进行分类训练建模的方法^[10]; 通过 Navie Bayes、多层神经网络、SVM、决策树等机器学习算法设计一种智能 IDS 系统进行检测的方法^[11]。此类方法依据机器学习算法不断挖掘训练历史流量数据, 并基于训练模型进行异常检测, 但主要选取持续型

App-DDoS 攻击流与 Flash Crowd 流的特征作为输入进行分类器的训练, 将检测看作二分类问题, 而在训练过程中忽略了上升型与脉冲型 App-DDoS 攻击流特征^[12,13]。当充分考虑正常流、Flash Crowd 流及持续型、上升型与脉冲型 App-DDoS 攻击流的特征时, 训练与检测过程将从二分类问题转变为多分类问题。

目前, 构造多类分类器的方法主要有两类。一类是直接法^[14], 直接在目标函数上进行修改, 将多个分类器的参数求解合并到一个最优化问题中, 通过求解该最优化问题一次性实现多类分类。这种方法结构简单, 但计算复杂度较高, 不适用于实时性要求高的检测领域。另一类是间接法, 通过组合多个二分类器实现多分类器的构造, 主要分为一对一 SVM 多分类方法^[15]、一对多 SVM 多分类方法^[16]及决策二叉树 SVM 多分类方法等^[17]。对于 P 分类问题, 一对一分类方法中二分类器个数为 $\frac{P(P-1)}{2}$, 过大的二分类器数量导致分类判

决速度慢, 且采用“投票法”进行决策影响了分类时的精度; 一对多分类方法中二分类器个数为 P , 但全部的样本数据都要参与 P 个二分类器的训练, 使单个分类器的训练时间较长、训练效率低; 决策二叉树多分类方法中二分类器个数为 $P-1$, 随着二叉树层次的增加, 训练的规模不断下降, 训练的时间不断减少, 且分类判决时不一定需要从根节点开始对所有分类器进行计算, 从而减少了分类判决所需的时间, 但是其分类效果受选取的二叉树结构影响较大。通过上述分析比较, 拥有更好训练效率与相对简便判决过程的决策二叉树多分类方法更适用于实时性要求高的 App-DDoS 攻击检测方法, 但是如何选取适合 App-DDoS 攻击检测领域的二叉树结构是所需解决的主要问题。

综上, 为解决非突发性正常流量、Flash Crowd 及上升型、脉冲型与持续型 App-DDoS 攻击的多分类问题, 本文围绕五者在流量规模、流量趋势及源 IP 分布方面的差异, 考虑检测的区分流程提出一种基于偏二叉树 SVM 多分类算法的 App-DDoS 攻击检测方法, 主要贡献如下。

1) 设计一种 HTTP GET 统计模型, 通过 Hash

函数及开放定址防碰撞方法建立多个时间周期内源 IP 地址在统计模型中的索引,实现多周期内不同源 IP 地址 HTTP GET 数的快速统计功能。

2) 深入分析非突发性正常流量、Flash Crowd 与 App-DDoS 攻击流量在数据规模、流量趋势及源 IP 地址分布的区别,给出时间窗内 HTTP GET 总数、HTTP GET 方差、源 IP 地址熵及时间窗内 HTTP GET 自相关系数 4 个流量特征参数用于描绘用户流量特征差异。

3) 按区分突发流与非突发流、在突发流中区分 Flash Crowd 与 App-DDoS 攻击流及在 App-DDoS 攻击流中区分具体类型的顺序建立偏二叉树 SVM 训练模型,并进一步提出基于偏二叉树 SVM 多分类算法的 App-DDoS 攻击检测方法,相比传统的二分类 App-DDoS 攻击检测方法,通过划分检测类型、逐层训练检测模型提升检测性能。

下面分别介绍统计空间模型、流量特征参数及基于偏二叉树 SVM 多分类算法的 App-DDoS 检测方法。

2 统计空间模型

HTTP GET 是用户访问 Web 服务器的主要标

识,准确高效收集 HTTP GET 标识数据是进行流量特征参数计算的重要前提,本节设计一种 HTTP GET 统计空间模型,定义如下。

定义 1 (HTTP GET 统计空间模型) 假设测量周期内最大源 IP 地址数为 n , 以一定的周期连续地测量任意一对源—目的节点之间的 HTTP GET 数,所有测量值组成一个 $\lceil \sqrt{n} \rceil \times \lceil \sqrt{n} \rceil \times T$ 的 HTTP GET 统计空间 $\mathcal{S}$, 称为该目的节点的 HTTP GET 统计空间,其中, T 表示测量的周期数,流量空间的横轴值 x 与纵轴值 y 分别为源节点 IP 地址前 16 bit 与后 16 bit 经 Hash 函数及开放定址防碰撞方法映射的值,流量空间的竖轴坐标值 z 为采样周期数,流量空间的任一元素 $s(i, j, k)$ 表示第 k 个周期内横轴值 i 与纵轴值 j 所对应源节点上的 HTTP GET 数测度大小。

其中,源—目的节点对由所有访问者与受保护的服务器组成; $\lceil \sqrt{n} \rceil$ 为不小于 $\sqrt{n}$ 的最小整数; Hash 函数如式(1)所示,由于 $(\lceil \sqrt{n} \rceil)^2$ 值不小于 n , 因此 Hash 函数的模值取 $\sqrt{n}$, 并通过开放定址防碰撞的方法表示所有 n 个源 IP 地址,该统计空间模型如图 1 所示。

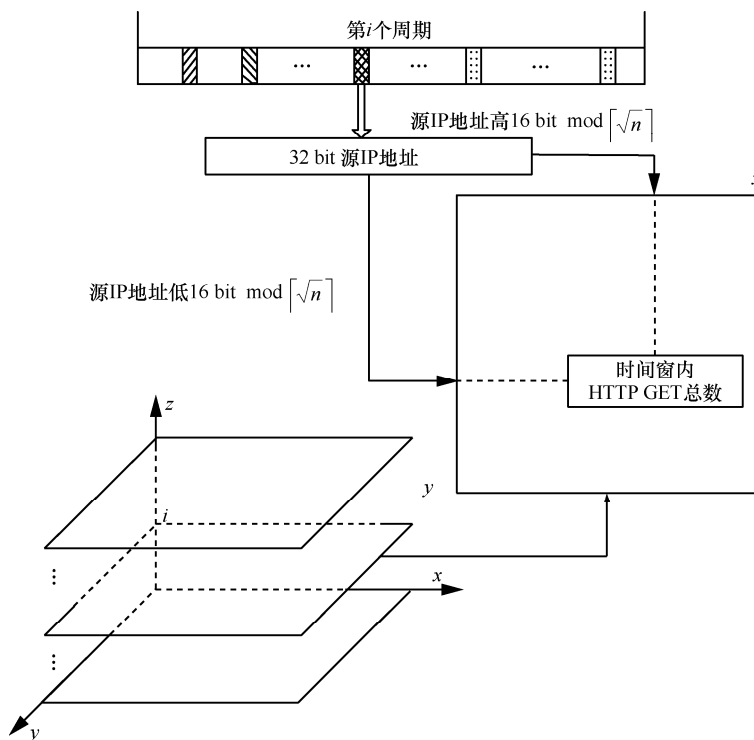


图1 HTTP GET 统计空间模型

$$\begin{cases} x = h_{16} \bmod \lceil \sqrt{n} \rceil, h_{16} \text{ 为源IP地址高16 bit} \\ y = l_{16} \bmod \lceil \sqrt{n} \rceil, l_{16} \text{ 为源IP地址低16 bit} \end{cases} \quad (1)$$

3 流量特征参数

非突发性正常流不影响服务器正常工作，而 Flash Crowd 流及 App-DDoS 攻击流会导致服务器瘫痪。且 Flash Crowd 流由用户自发访问产生，App-DDoS 攻击流由黑客通过对僵尸主机发送指令产生。通过分析以上 2 种情况可知，非突发性正常流、Flash Crowd 流及 App-DDoS 攻击流主要在数据规模、流量趋势及源 IP 地址分布上具有显著差异，具体区别如下。

1) 规模差异。App-DDoS 攻击流与 Flash Crowd 流导致服务器瘫痪的原因为大规模流量，而非突发性正常流量通常维持在较低的规模，这不会影响服务器正常工作。

2) 趋势差异。App-DDoS 攻击流是黑客通过发送指令造成的，每个用户流量 HTTP GET 请求数差异小且攻击分为上升型、持续型及脉冲型，而 Flash Crowd 流是用户独立访问造成的，每个用户流量 HTTP GET 请求数差异大且在服务器瘫痪时请求数逐渐降低。

3) 源 IP 地址分布差异。App-DDoS 攻击流与 Flash Crowd 流不仅在 HTTP GET 请求数上存在差异，每个周期内源 IP 地址数量也不同。产生 Flash Crowd 的用户可能来自整个网络，源 IP 地址分布非常分散，而在一次 App-DDoS 攻击中，源 IP 地址分布受到攻击者数量或僵尸网络规模的限制会相对集中。

为详细刻画上述 3 个方面差异，从同一周期内数据变化角度提出 HTTP GET 总数、HTTP GET 方差及源 IP 地址熵等特征参数，从不同周期间的数据变化角度提出 HTTP GET 自相关系数特征参数。对于三维流量空间 $\mathbf{S} = s_{ijk}, i, j = 1, \dots, \lceil \sqrt{n} \rceil, k = 1, \dots, T$ ，流量特征参数定义如下。

1) 使用第 k 个周期内 HTTP GET 总数 Tol_k 表征非突发性正常流与 Flash Crowd 及 App-DDoS 攻击的规模，即

$$Tol_k = \sum_{i=1}^{\lceil \sqrt{n} \rceil} \sum_{j=1}^{\lceil \sqrt{n} \rceil} s_{ijk} \quad (2)$$

2) 使用第 k 个周期内每个源 IP 的 HTTP GET 方差 V_k 与 T 个不同周期间的 HTTP GET 总数自相关系数 $A(\tau)$ 分别表征 Flash Crowd 与 3 种类型 App-DDoS 攻击的趋势，即

$$V_k = E \left\{ (b_{ijk} - E(s_k))^2 \right\}, i, j = 1, \dots, \lceil \sqrt{n} \rceil \quad (3)$$

$$A(\tau) = \frac{E \left\{ (Tol_k - E(\mathbf{S}))(Tol_{k+\tau} - E(\mathbf{S})) \right\}}{\sigma(\mathbf{S})^2} \quad (4)$$

其中，当 $s_{ijk} \neq 0$ 时， $b_{ijk} = s_{ijk}$ ，即测量值为 0 时无实际含义； s_k 为第 k 个周期内所有数据， $E(\cdot)$ 为期望； $\sigma(\mathbf{S})$ 为流量空间 $\mathbf{S}$ 标准差，可表征周期内数值波动的大小； τ 为延迟， $k + \tau$ 表示周期 k 之后的第 τ 个周期；根据自相关系数的定义，自相关系数的取值范围为 $[-1, 1]$ ，可表征不同周期间相关性程度，取值的绝对值越大相关性越高。

3) 使用第 k 个周期内源 IP 地址熵 E_k 表征 Flash Crowd 与 3 种类型 App-DDoS 攻击的源 IP 地址分布，即

$$E_k = - \sum_{i=1}^{\lceil \sqrt{n} \rceil} \sum_{j=1}^{\lceil \sqrt{n} \rceil} p(s_{ijk}) \log(p(s_{ijk})) \quad (5)$$

其中， $p(s_{ijk}) = \frac{\eta_{ij}}{\sum_{i=1}^{\lceil \sqrt{n} \rceil} \sum_{j=1}^{\lceil \sqrt{n} \rceil} \eta_{ij}}$ ，且

$\log(p(s_{ijk})) = 0$ ($s_{ijk} = 0$)。根据熵的定义， E_k 取值越小表明分布越有序，反之，取值越大表明分布越杂乱无章。

4 多分类检测方法

在上述 HTTP GET 统计空间与表征流量差异的特征参数基础上，本节提出一种基于偏二叉树 SVM 多分类算法的 App-DDoS 检测方法，解决非突发正常流、Flash Crowd 流及 3 种类型 App-DDoS 攻击流的多分类问题。相比传统二分类的 App-DDoS 检测方法，通过丰富的训练数据与细化检测粒度提升检测性能。下面分别介绍基于偏二叉树的 SVM 多分类算法以及基于该算法的 App-DDoS 检测方法。

4.1 基于偏二叉树的 SVM 多分类算法

上文已分析指出拥有更好训练效率与相对简

便判决过程的决策二叉树多分类方法更适用于实时性要求高的 App-DDoS 攻击检测方法。此类方法是以二叉树的形式组合多个二分类器构造实现多分类器,因此需要解决二分类器选取与二叉树结构选取 2 个问题。

针对二分类器选取问题,选取 SVM 算法作为二分类器,该算法建立在 VC (Vapnik-Chervonenkis) 理论和结构风险最小化原则 (SRM, structural risk minimization) 基础上,通过构造一个最优分类超平面,实现未知样本类别的判断,表现出了良好的泛化性能,适合解决小样本和非线性问题,同时解决了维数灾难、过学习和局部极小等常见问题^[18],是当前机器学习的重要方法。

对于二分类问题,给定训练样本 (u_i, v_i) , $i=1, \dots, l$, $u_i \in R^d$, $v_i \in \{-1, 1\}$ (l 为训练样本数量; R 为实数域; d 为样本特征维数; v_i 为样本标签, 1 与 -1 分别代表不同类型样本), SVM 算法的最优超平面构造可转化为一个二次规划问题。

$$\begin{aligned} \max W(\alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j v_i v_j K(u_i u_j) \\ \text{s.t. } \sum_{i=1}^n \alpha_i v_i &= 0, \alpha_i \geq 0, i=1, \dots, n \end{aligned} \quad (6)$$

训练完成后,超平面函数为

$$f(x) = \text{sgn} \left(\sum_{i,j} \alpha_i v_i K(u_i u_j) + b \right) \quad (7)$$

其中, α_i 为 Lagrange 乘子, $K(u_i u_j)$ 为核函数, b 为常数。

针对二叉树结构选取问题,考虑本文的分类流程为先区分出非突发流与突发流,再在突发流中区分 Flash Crowd 流与 App-DDoS 攻击流,最后依据 App-DDoS 攻击流类型依次区分持续型 App-DDoS 攻击、脉冲型 App-DDoS 攻击及上升型 App-DDoS 攻击,此流程近似于偏二叉树结构,因此选取基于偏二叉树 SVM 多分类算法 (PBT-SVM, partial binary tree based SVM algorithm) 进行样本训练,训练流程如图 2 所示,具体步骤如下。

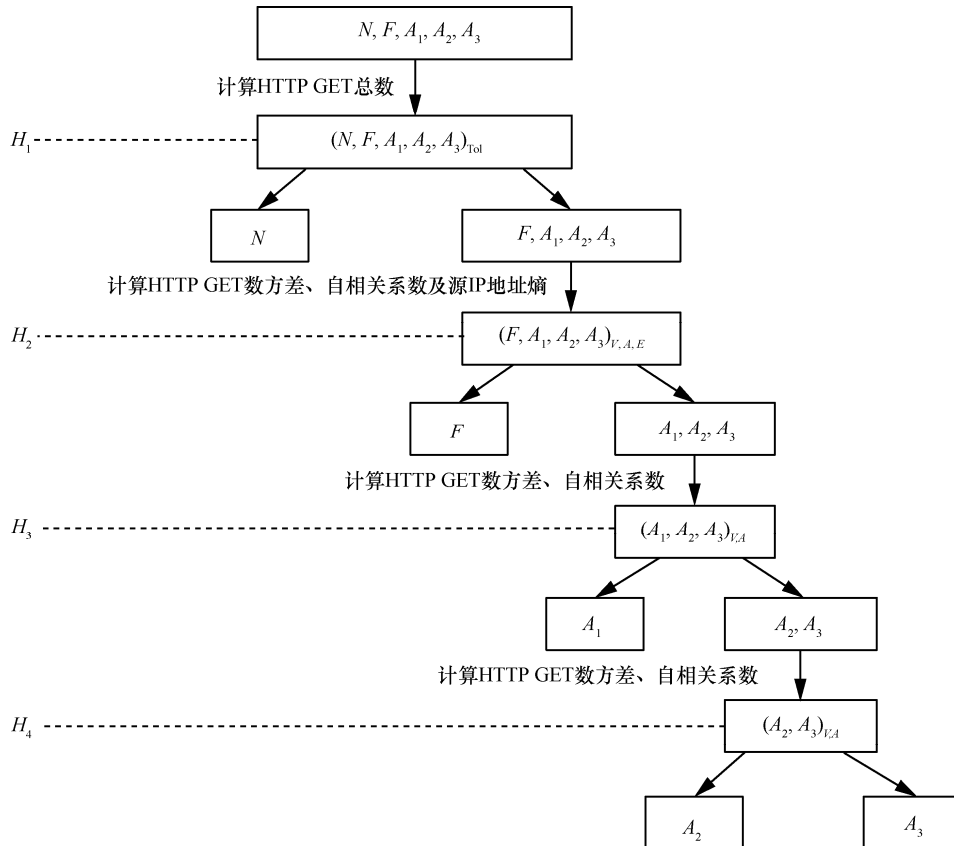


图2 基于 PBT-SVM 算法的样本训练流程

步骤 1 计算训练数据集 N, F, A_1, A_2, A_3 的 HTTP GET 总数 $(N, F, A_1, A_2, A_3)_{\text{Tot}}$ 作为输入, 通过 SVM 算法训练超平面 H_1 区分 N 与 F, A_1, A_2, A_3 , 其中, N, F, A_1, A_2, A_3 分别代表非突发性正常流、Flash Crowd、持续型 App-DDoS 攻击、脉冲型 App-DDoS 攻击及上升型 App-DDoS 攻击的相应数据。

步骤 2 计算训练数据集 F, A_1, A_2, A_3 的 HTTP GET 方差、自相关系数及源 IP 地址熵 $(F, A_1, A_2, A_3)_{V,A,E}$ 作为输入, 通过 SVM 算法训练超平面 H_2 区分 F 与 A_1, A_2, A_3 。

步骤 3 选取步骤 2 中 A_1, A_2, A_3 的 HTTP

GET 方差、自相关系数 $(A_1, A_2, A_3)_{V,A}$ 作为输入, 通过 SVM 算法训练超平面 H_3 区分 A_1 与 A_2, A_3 。

步骤 4 选取步骤 2 中 A_2, A_3 的 HTTP GET 方差、自相关系数 $(A_1, A_2, A_3)_{V,A}$ 作为输入, 通过 SVM 算法训练超平面 H_4 区分 A_2 与 A_3 。

4.2 基于 PBT-SVM 算法的 App-DDoS 检测方法

基于 PBT-SVM 算法的 App-DDoS 检测方法通过 HTTP GET 统计空间收集数据, 依据所计算的流量特征参数从训练的偏二叉树 SVM 模型根节点开始遍历进行流量分类, 对检测结果进行学习更新, 并将更新后结果用于下一次检测, 检测方法如图 3 所示, 具体步骤如下。

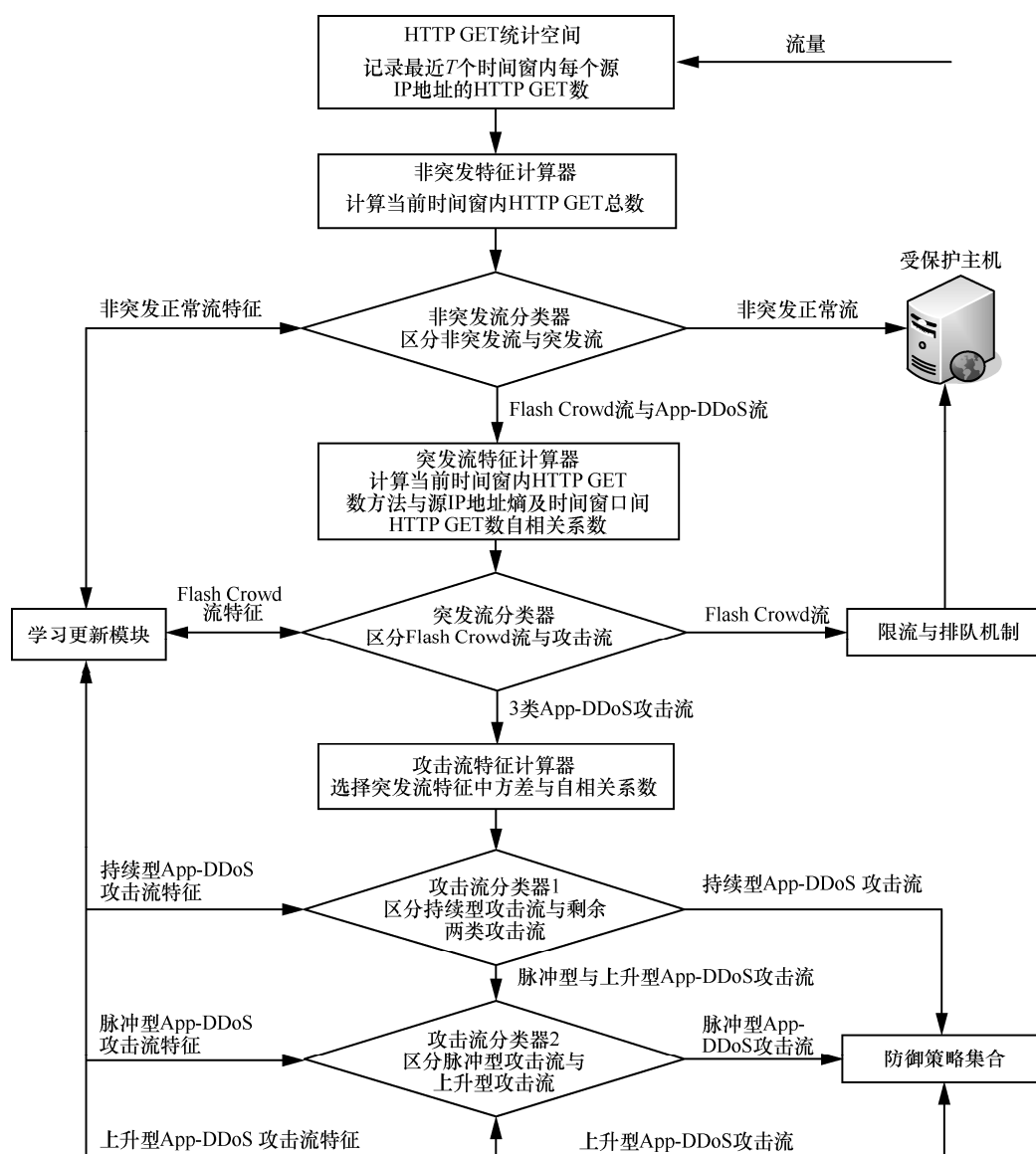


图3 基于 PBT-SVM 算法的 App-DDoS 攻击检测方法

步骤 1 流量空间更新。在统计新时间窗口内数据之前,先将当前流量空间内第 $2 \sim T$ 个窗口内数据依次前移覆盖前 1 窗口内容。然后,依据式(1)计算当前时间窗口内源 IP 对应坐标并统计该 IP 对应 HTTP GET 总数,覆盖第 T 个窗口内容。

步骤 2 流量区分。采用遍历二叉树的思想,依据训练流程进行检测指标计算与流量类型判断,在当前节点无子节点时中止判决。检测结果为正常流量时允许直接访问;检测结果为 Flash Crowd 流时设置限制访问及排队机制;检测结果为某一具体类型 App-DDoS 攻击时,依据攻击类型选取不同的防护策略。

步骤 3 区分结果的反馈学习。将当前周期的检测结果反馈给训练模块进一步训练,并在更新后结果的基础上重复步骤 1 与步骤 2 进行下一次检测。

5 实验分析

本文选取 FIFA World Cup^[19]中 1998 年 5 月 10 日 8~9 时的数据作为非突发性正常用户训练数据 NTD , 9~10 时的数据作为流量体积相似的非突发性正常用户检测数据 NDD_1 , 14~15 时的数据作为流量体积不同的非突发性正常用户检测数据 NDD_2 ; 同时,选取 1998 年 7 月 12 日(决赛) 8~9 时的数据作为 Flash Crowd 训练数据 FTD , 9~10 时的数据作为流量体积相似的 Flash Crowd 检测数据 FDD_1 , 14~15 时的数据作为流量体积不同的 Flash Crowd 检测数据 FDD_2 ; 通过大流量发生器 Spriant TestCenter C100 模拟大规模 App-DDoS 攻击数据, 上升型、脉冲型及持续型的训练与检测数据集分别记为 $ATD_1 \sim ATD_3$ 与 $ADD_1 \sim ADD_3$, 其中, C100 模拟的实验数据中 Client(客户)参数设置主要包括: 1) 负载中发送数据的计量单位, 攻击阶段数设置及每阶段的峰值、上升时间、稳定时间; 2) Actions 中访问页面内容; Subnets 中子网网段, 即模拟的客户端数。

5.1 特征参数验证

为了验证所提指标的合理性, 下面分别给出数据集的 4 个特征。

首先, 以 1 min 为时间窗口给出正常用户与 Flash Crowd 的 HTTP GET 总数统计结果, 分别如图 4、图 5 所示。

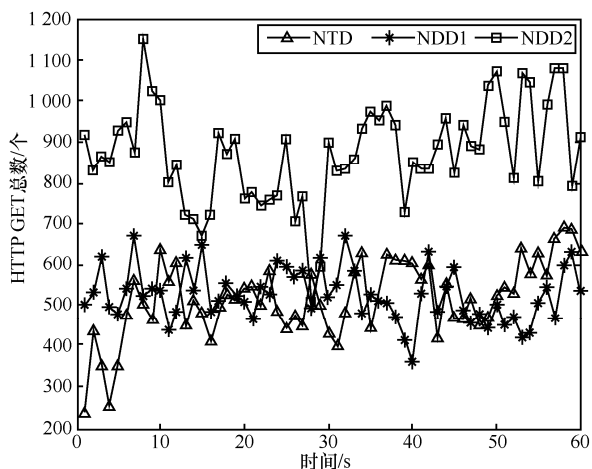


图 4 正常用户训练及检测数据集统计结果

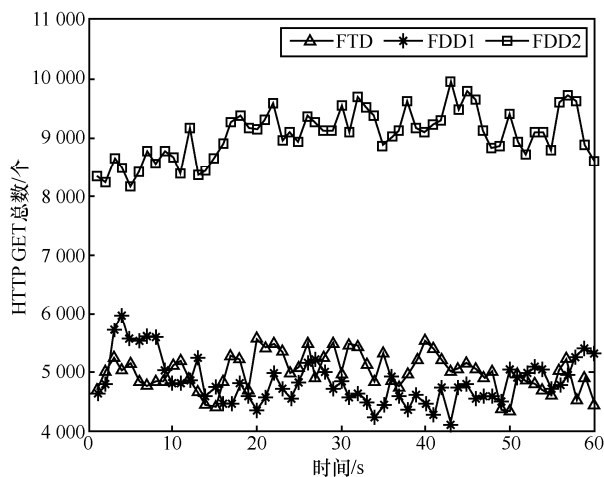


图 5 Flash Crowd 训练及检测数据集统计结果

实验结果表明, Flash Crowd 的 HTTP GET 数总量分布在 4 100~6 000 与 8 100~9 900 这 2 个范围内, 非突发性正常用户总量分布在 210~1 160 之间, 相比而言, 非突发性正常用户数据集 HTTP GET 总数远小于 Flash Crowd 数据集。故 HTTP GET 总数可表征两者的流量规模差异。

同时, 通过 C100 模拟攻击流量时具体参数如表 1 所示。其中 Trans/s 表示链接/秒, Random 表示对目标 Web 服务器随机访问, 参数依据如下。

1) 攻击时间与等数据集时间相同, 且攻击分为 6 个阶段, 每一阶段时间为统计模型最大周期容量, 即攻击时间为 60 min, HTTP GET 统计空间模型中。

2) 表 1 中只给出一个阶段的参数设置, 剩余阶段的参数设置与给出的参数相同, 且为了保证检测的多样性, 设置不同的训练集与检测集参数。

3) 脉冲型攻击在表 1 中只显示了 5 min 时间, 实

表 1 本文模式 App-DDoS 攻击参数设置

数据集	负载					请求方式	子网规模
	单位	阶段数	峰值	上升时间/min	稳定时间/min		
ATD_1	Trans/s	6	120	9	1	Random	2 000
ATD_2	Trans/s	6	120	0	5	Random	2 000
ATD_3	Trans/s	6	120	0	10	Random	2000
ADD_1	Trans/s	6	150	8	2	Random	2 000
ADD_2	Trans/s	6	150	0	6	Random	2 000
ADD_3	Trans/s	6	150	0	10	Random	2 000

际为 5 min 脉冲峰值时间与 5 min 脉冲谷值 15 Trans/s 访问量时间（与非突发正常流总数相似），同理，脉冲型攻击实际为 6 min 脉冲峰值时间与 4 min 脉冲谷值 15 Trans/s 访问量时间。

4) 峰值量保持与 Flash Crowd 平均量近似，训练集设置为 120 Trans/s，检测集设置为 150 Trans/s。

其次，由于方差是在区分 Flash Crowd 与 App-DDoS 攻击时使用，下面给出 3 个 Flash Crowd 数据集与 3 个攻击训练数据集数据的方差统计结果，分别如图 6、图 7 所示。

实验结果表明，Flash Crowd 方差分布在 47~281 之间，上升型 App-DDoS 攻击训练数据集方差分布在 10~38 之间，持续型脉冲 App-DDoS 攻击训练数据集方差分布在 23~38 之间，而脉冲型 App-DDoS 攻击训练数据集在峰值时方差分布在 23~38 之间，谷值时方差为 0，相比而言攻击数据集方差及其波动范围均较小。故 HTTP GET 方差可表征两者在流量波动上的差异。

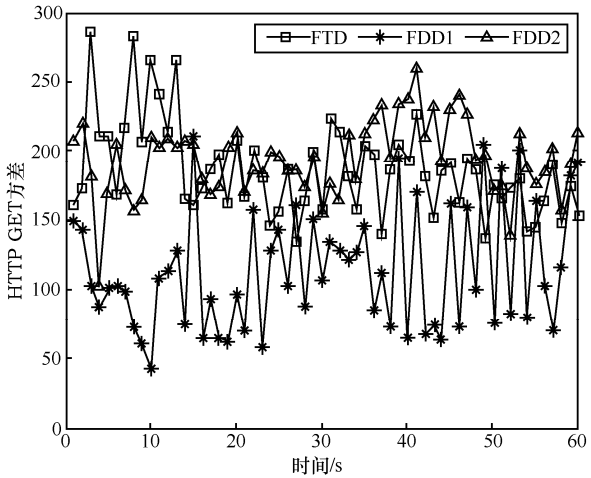


图 6 Flash Crowd 数据集方差统计结果

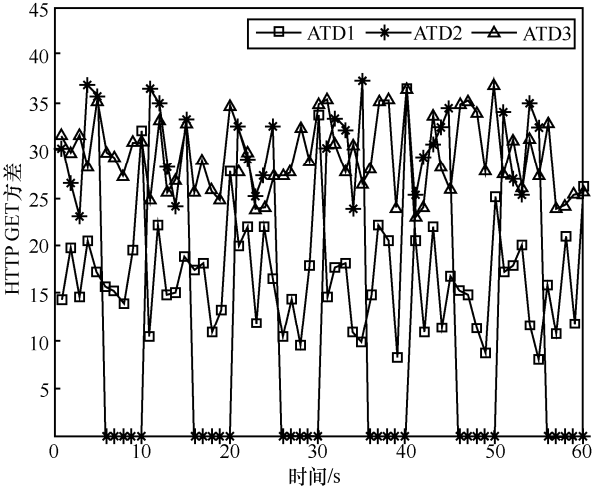


图 7 3 个攻击训练数据集方差统计结果

然后，给出 3 个 Flash Crowd 数据集、3 个攻击训练数据集熵值统计结果，分别如图 8、图 9 所示。

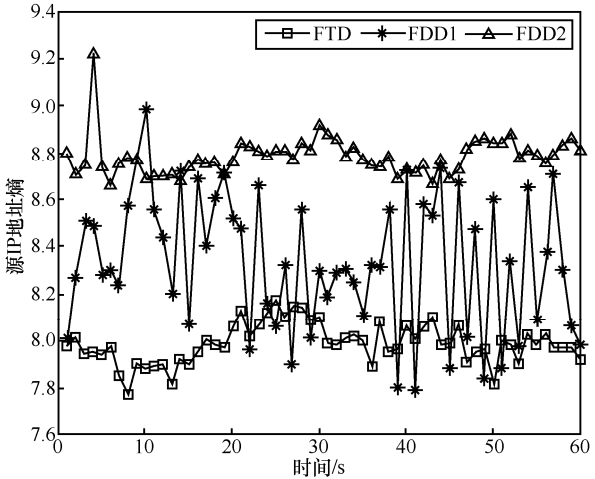


图 8 Flash Crowd 数据集熵值统计结果

实验结果表明，Flash Crowd 熵值分布在

7.68~9.21 之间, 上升型 App-DDoS 攻击训练数据集熵值分布在 0.30~0.81 之间, 持续型脉冲 App-DDoS 攻击训练数据集熵值分布在 0.6~0.91 之间, 而脉冲型 App-DDoS 攻击训练数据集在峰值时熵值分布在 0.61~0.90 之间, 谷值时熵值分布在 0.20~0.48 之间, 相比而言攻击数据集熵值及其波动范围均较小。故源 IP 地址熵可表征两者在源 IP 地址分布的差异。

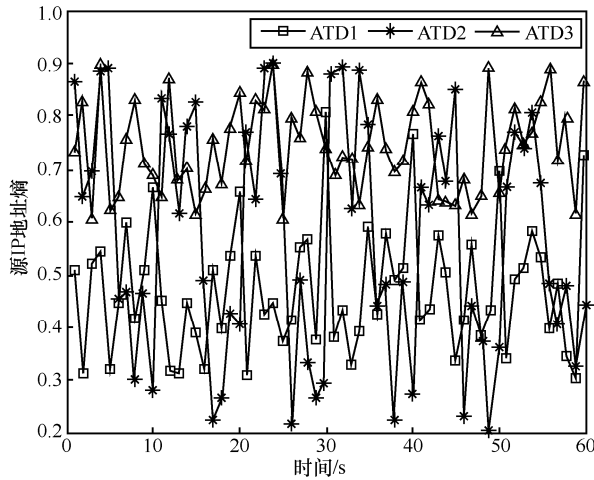


图9 3个攻击训练数据集熵值统计结果

最后, 设置延迟 $\tau = 2$, 给出 3 个 Flash Crowd 数据集、3 个攻击训练数据集自相关系数, 分别如图 10、图 11 所示。

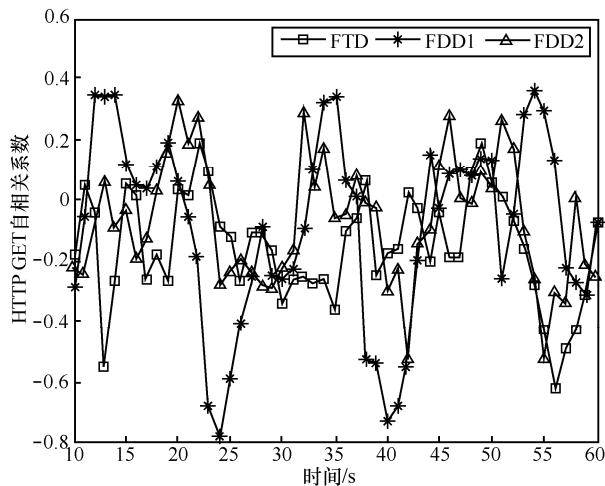


图10 Flash Crowd 数据集自相关系数统计结果

实验结果表明, Flash Crowd 自相关系数分布在 -0.79~0.38 之间, 上升型 App-DDoS 攻击训练数据集自相关系数分布在 -0.19~0.48 之间, 而脉冲型 App-DDoS 攻击训练数据集自相关系数分布

在 -0.1~0.48 之间, 由于实验攻击具有周期性, 导致上述两类攻击自相关系数近似于具有周期性且在每个周期变化临界点数值会骤然变化, 同时持续型攻击流量自相关系数在 0.96~1 之间小范围波动。故自相关系数可表征两者在流量趋势的差异。

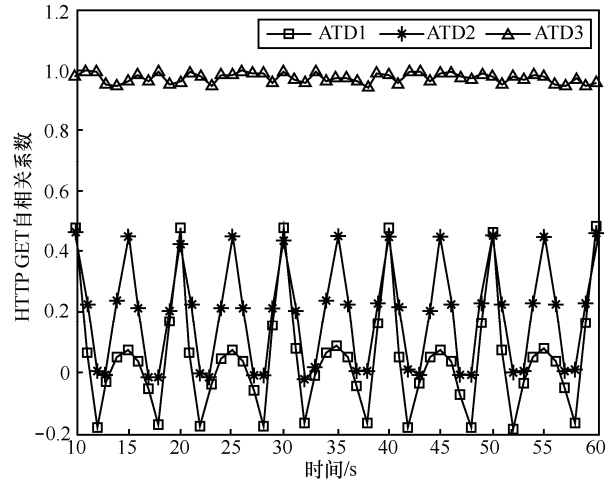


图11 3个攻击训练数据集自相关系数统计结果

5.2 检测性能验证

为了验证所提方法检测性能, 将其与传统的基于 SVM 二分类算法及基于 Navie Bayes 二分类算法的检测方法进行检测率 (TPR, true positive rate)、误检率 (FPR, false positive rate) 比较。其中, 在本文给出的流量特征参数基础上, 所提方法分别按图 2 与图 3 流程进行训练与检测, 且训练过程中核函数采用线性核函数, 基于 SVM 二分类算法的检测方法中训练数据集分为正常数据集 (包括 NTD 与 FTD) 与攻击数据集 (包括 $ATD_1 \sim ATD_3$) 且训练过程中核函数同样采用线性核函数, 基于 Navie Bayes 二分类算法的检测方法中训练数据集与基于 SVM 二分类算法的检测方法相同, 实验结果如表 2 和表 3 所示。检测率与误检率计算式如下。

$$\text{检测率} = \frac{\text{攻击用户检测结果为攻击者的数目}}{\text{攻击用户总数}} \quad (8)$$

$$\text{误检率} = \frac{\text{正常用户检测结果为攻击者的数目}}{\text{正常用户总数}} \quad (9)$$

其中, negative 对象包括正常流量与 Flash Crowd 流量、positive 对象包括 3 种类型攻击流量。

表2 PBT-SVM、SVM 及 Navie Bayes 算法误检率对比

数据集	PBT-SVM	SVM	Navie Bayes
NDD_1	0	0	0
NDD_2	0	0	0
FDD_1	1.67%	3.33%	6.67%
FDD_2	3.33%	8.33%	6.67%
平均值	1.25%	2.92%	3.75%

表3 PBT-SVM、SVM 及 Navie Bayes 算法检测率对比

数据集	PBT-SVM	SVM	Navie Bayes
ADD_1	96.67%	85.00%	83.33%
ADD_2	93.33%	86.67%	81.67%
ADD_3	100.00%	100.00%	100.00%
平均值	96.67%	90.56%	88.33%

实验结果表明,在本文给出的流量特征参数基础上,利用 PBT-SVM 算法对不同检测数据集进行测试时,与 SVM 及 Navie Bayes 算法相比,都能完全检测出非突发正常流与持续型 App-DDoS 攻击,但是检测 Flash Crowd 时误检率分别平均降低了 3.3%与 4.17%,检测上升型 App-DDoS 攻击时检测率分别提升了 11.67%与 13.34%,检测脉冲型 App-DDoS 攻击时检测率分别提升了 6.66%与 11.66%。同时,通过检测流程,本文所提方法可以区分出具体的 App-DDoS 攻击类型,而基于 SVM 二分类算法及基于 Navie Bayes 二分类算法的检测方法只能判断出攻击是否发生而无法具体区分攻击类型。

综上,本文给出的特征参数可分别表征流量规模、流量趋势、流量波动及源 IP 地址分布差异,且所提检测多分类方法与传统二分类方法相比拥有更好的检测性能,能有效区分出具体攻击类型。

6 结束语

本文针对现有基于流量特征的 App-DDoS 检测方法在训练过程中忽略脉冲型与上升型 App-DDoS 攻击问题,提出一种检测多类型应用层 DDoS 攻击的方法。首先,通过建立源 IP 地址 Hash 索引的方法,设计了多周期内不同源 IP 地址 HTTP GET 快速统计模型;其次,基于同周期内与不同周期期间的 HTTP GET 数与源 IP 地址信息,给出了刻画数据规模、流量趋势及源 IP 地址

分布差异的特征参数;然后,提出基于偏二叉树 SVM 多分类算法的 App-DDoS 检测方法,通过划分检测类型、逐层训练检测模型,快速区分出非突发正常流、突发正常流及多类型 App-DDoS 流;最后,通过实验证明了所提特征参数的有效性及检测方法的准确性。下一步将对骨干网进行大规模流量下的实时性验证。

参考文献:

- [1] 陈飞,毕小红,王晶晶,等. DDoS 攻击防御技术发展综述[J]. 网络与信息安全学报, 2017, 3(10): 16-24.
CHEN F, BI X H, WANG J J, et al. Survey of DDoS defense: challenges and directions[J]. Chinese Journal of Network and Information Security, 2017, 3(10):16-24.
- [2] SINGH K, SINGH P, KUMAR K. Application layer HTTP-GET flood DDoS attacks: research landscape and challenges[J]. Computer & Security, 2016, 65:344-372.
- [3] 李锦玲,王斌强. 基于最大频繁序列模式挖掘的 App-DDoS 攻击的异常检测[J]. 电子与信息学报, 2013, 35(7): 1739-1745.
LI J L, WANG B Q. Detecting App-DDoS attacks based on maximal frequent sequential pattern mining[J]. Journal of Electronics & Information Technology, 2013, 35(7):1739-1745.
- [4] SANGJAE L, GISUNG K, SEHUM K. Sequence-order-independent network profiling for detecting application layer DDoS attacks[J]. Eurasip Journal on Wireless Communication & Networking, 2011, 50(1): 1-9.
- [5] 杨宏宇,常媛. 基于 K 均值多重主成分分析的 App-DDoS 检测方法[J]. 通信学报, 2014, 35(5): 16-24.
YANG H Y, CHANG Y. App-DDoS detection method based on K-means multiple principal component analysis[J]. Journal on Communications, 2014, 35(5): 16-24.
- [6] YU S, THAPNGAM T, LIU J, et al. Discriminating DDoS flows from flash crowds using information distance[C]//IEEE The 3rd International Conference on Network and System Security. 2009: 351-356.
- [7] LI K, ZHOU W L, LI P, et al. Distinguishing DDoS attacks from flash crowds using probability metrics[C]//IEEE The 3rd International Conference on Network and System Security. 2009: 9-17.
- [8] YU S, ZHOU W L, JIA W J, et al. Discriminating DDoS attacks from flash crowds using flow correlation coefficient[J]. IEEE Transactions on Parallel and Distributed Systems, 2012, 23(6): 1073-1080.
- [9] SINGH K J, THONGAM K, DE T. Entropy-based application layer DDoS attack detection using artificial neural networks[J]. Entropy, 2016, 18(10), 350-367.
- [10] NI T G, GUX Q, WANG H Y, et al. Real-Time detection of applica-

- tion-layer DDoS attack using time series analysis[J]. Journal of Control Science and Engineering, 2013(5): 1-6.
- [11] IRFAN S, AMIT M, VIBHAKAR M. Machine learning techniques used for the detection and analysis of morden types of DDoS attacks[J]. International Research Journal of Engineering and Technology, 2017, 4(6):16-24.
- [12] LIU L, JIN X L, MIN G Y, et al. Anomaly diagnosis based on regression and classification analysis of statistical traffic features[J]. Security and Communication Networks, 2014, 7: 1372-1383.
- [13] PAL R, KUMAR S, SHARMA R L. A detailed classification of flash events: client, server and network characteristics[C]//The International Conference on Computer Science & Service System. 2012: 960-963.
- [14] 杨新武, 马壮, 袁顺. 基于弱分类器调整的多分类 Adaboost 算法[J]. 电子与信息学报, 2016, 38(2): 373-380.
- YANG X W, MA Z, YUAN S. Multi-classification adaboost algorithm based on weak classifier adjustment[J]. Journal of Electronics & Information Technology, 2016, 38(2): 373-380.
- [15] KANGS, CHOS Z, PILSUNG K. Constructing a multi-class classifier using one-against-one approach with different binary classifier[J]. Neurocomputing, 2015, 149: 677-682.
- [16] SILVA C, RIBEIRO B. Multiclass ensemble of one-against-all SVM classifiers[C]//The International Symposium on Neural Networks. 2016: 531-539.
- [17] JINDAL A, DUA A, KAUR K, et al. Decision tree and SVM-based data analytics for theft detection in smart grid[J]. IEEE Transactions on Industrial Informatics, 2016, 12(3): 1005-1016.
- [18] CORTES C, VAPNIK V. Support vector networks[J]. Machine Learning, 1995, 20: 273-297.
- [19] ITA. WorldCup98[DB/OL]. <http://ita.ee.lbl.gov/html/contrib/WorldCup.html>.

[作者简介]



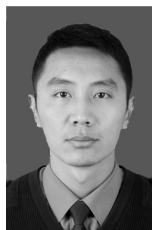
张斌 (1969-), 男, 河南郑州人, 信息工程大学教授、博士生导师, 主要研究方向为网络空间安全。



刘自豪 (1993-), 男, 湖北十堰人, 信息工程大学硕士生, 主要研究方向为应用层 DDoS 攻击检测与评估。



董书琴 (1990-), 男, 河北邢台人, 信息工程大学博士生, 主要研究方向为网络空间态势感知。



李立勋 (1994-), 男, 四川都江堰人, 信息工程大学硕士生, 主要研究方向为动态目标防御。