基于随机森林分类模型的 DDoS 攻击检测方法*

于鹏程，戚　湧†，李千目

（南京理工大学 计算机科学与工程学院，南京 210094）

摘　要：分布式拒绝服务（distributed denial-of-service，DDoS）是目前常见的网络攻击方式之一。基于机器学习算法（SVM、HMM 等）的 DDoS 攻击检测技术取得一些进展，但还存在着样本数量过多时易发生过拟合和未充分利用上下文信息等不足。为了弥补以上不足，提出一种基于随机森林的 DDoS 攻击检测方法，将数据流信息熵作为分类标准，令 sourceIP、destinationIP、destinationPort 分别代表数据流的源地址、目的地址、目的端口，采用 SIDI（sourceIP-destinationIP）、SIDP（sourceIP-destinationPort）和 DPDI（destinationPort-destinationIP）三个信息熵来分别表征三种多对一的特征，对 TCP 洪水攻击、UDP 洪水攻击、ICMP 洪水攻击等三种常见的攻击方式进行特征分析，在此基础上使用基于随机森林分类模型分别对三类 DDoS 攻击方式进行分类检测，实验结果表明该模型能够较为准确地区分正常流量和攻击流量，与 HMM、SVM 方法相比，基于 RFC 模型的 DDoS 检测方法有较高的检测率和较低的误报率。

关键词：随机森林；数据流信息熵；分布式拒绝服务；检测

中图分类号：TP309.2　　**文献标志码**：A　　**文章编号**：1001-3695(2017)10-3068-05

doi:10.3969/j.issn.1001-3695.2017.10.042

DDoS attack detection method based on random forest

Yu Pengcheng，Qi Yong†，Li Qianmu

（*School of Computer Science & Engineering*，*Nanjing University of Science & Technology*，*Nanjing* 210094，*China*）

Abstract：DDoS attack is one of the major Internet threats. Traditional DDoS detection technology based on machine learning (SVM，HMM) has been some progress，but there are some shortcomings，such as the number of samples prone to excessive over-fitting and underutilized contextual information. To compensate for the above shortcomings，this paper proposed a DDoS attack detection method based on random forest，and defined the data stream information entropy as the classification standard. It used sourceIP，destinationIP，destinationport to represent data flow source address，destination address，destination port，used SIDI，SIDP and DPDI to represent three kinds of many to one features to analyze TCP flood attacks，UDP flood attacks，ICMP flood attack. On this basis，it used the classification model based on random forest respectively to classify three kinds of attack，to complete the detection of DDoS attacks. Experimental results show that the model can accurately distinguish between normal traffic and attack traffic. Compared with the HMM and SVM method，RFC model has a higher detection rate and low false alarm rate.

Key words：random forest；data stream information entropy；DDoS；detection

0　引言

分布式拒绝服务（distributed denial of service，DDoS）攻击是指借助于客户/服务器技术，将多个计算机联合起来作为攻击平台，对一个或多个目标发动攻击，从而增强攻击的威力[1]。分布式拒绝服务攻击改变了传统的点对点的攻击模式，使得攻击行为没有统计规律，而且在进行攻击时使用的也是常见的协议和服务，只通过协议和服务的类型很难对攻击行为和正常行为进行区分，导致分布式拒绝服务攻击不易检测[2]。

近年来，基于数据挖掘和机器学习算法的 DDoS 攻击检测方法正日趋完善。基于理性思维的思想，可以将 DDoS 攻击检测模拟为区分网络流状态“理性”和“非理性”的分类问题。这种方法的处理过程包括：a）选择合适的分类特征将网络数据流抽象化为特征向量；b）为每个特征向量赋予标记，标记的集合为{理性，非理性}，两种标记分别代表正常网络流和攻击网络流；c）选择合适的分类算法对样本进行学习，建立分类模型，然后用该模型对未标记的样本进行分类，推断其最可能的标记。

与其他攻击方式不同，DDoS 攻击与正常的背景流量极其相似，使取得的样本特征值有很大的不确定性和模糊性，对分类效果影响较大。因此，选择一种合适的分类特征显得非常重要。本文对常见的 TCP 洪水攻击、UDP 洪水攻击、ICMP 洪水攻击进行详细分析，定义数据流信息熵（data streaminformation entropy，DSIE）的特征来表征攻击行为。另一方面，样本上下文之间的相关性也可以用于消除特征值的模糊性和不确定性。在以往的研究中，大多数都是根据当前的样本特征值来推测样本的标记，忽略了相邻数据流之间的关联性[3]。事实上，相邻数据流之间往往存在着一定的关联，相邻数据流之间往往具有相同的特征标记。

收稿日期：2016-08-05；**修回日期**：2016-09-19　**基金项目**：国家自然科学基金资助项目（61272419）；中央高校基本科研业务费专项资金资助（30916015104）；江苏省产学研前瞻性基金资助项目（BY2014089）

作者简介：于鹏程（1992-），男，黑龙江哈尔滨人，硕士研究生，主要研究方向为信息安全；戚湧（1970-），男（通信作者），教授，博导，博士，主要研究方向为信息安全（790815561@qq.com）；李千目（1979-），男，教授，博导，博士，主要研究方向为信息安全.

本文提出一种基于随机森林分类(random forest classification,RFC)模型的 DDoS 攻击检测方法,分别对 TCP 洪水攻击、UDP 洪水攻击、ICMP 洪水攻击等三类典型攻击方式建立分类模型,通过训练学习,最终预测网络流量是否为正常流量。

1　国内外研究现状

DDoS 的检测方式主要分为基于攻击流的检测和基于正常数据流的检测,本文主要研究基于攻击流的检测方法。基于攻击流的检测方式是通过分析攻击流的特点识别非理性行为,检测到攻击。Chen 等人[4]依据 DDoS 攻击过程中会产生高流量的特点,计算正常流量和攻击流量之间的偏离度,从而确定是否是理性行为,这种检测方式不能够很好地区分 DDoS 攻击和大流量访问,误报率较高。文献[5~7]依据 DDoS 攻击过程中多对一的攻击特点,分别采用源 IP 地址数量、目的端口数量、流密度等三种特征来描述攻击行为的特性。这些方法能够区分大部分的攻击流是否为理性行为,但只用了较少的报文信息,大多只用到源 IP 地址和目的端口信息,且不能确定具体的攻击类型,使得检测率不高。

相关研究中用于 DDoS 攻击检测的机器学习算法主要有朴素贝叶斯算法、隐马尔可夫模型和支持向量机三种。Tama 等人[8]利用异常检测的手段,根据报文头属性对网络数据流进行建模,采用朴素贝叶斯算法给每个到达的数据流评分,评价报文的合理性。Karnwal 等人[9]通过维度转换,将一维时序转换为多维的 AR 模型参数时序,采用支持向量机算法对数据流进行学习和分类。Wang 等人[2]利用异常检测手段,利用隐马尔可夫模型来描述数据流中报文头的变化情况。以上文献中的方法在一定程度上提高了检测准确度,但没有充分利用数据流的上下文关系。

与一些传统的分类算法相比,随机森林算法有准确性高的优点,近年来,随机森林的思想和方法在许多领域都有应用。在生态学领域,Lindner 等人[10]利用随机森林思想对土地的营养指数进行分类,发现随机森林相比其他算法有更快的训练速度;Jia 等人[11]利用随机森林算法和 Logistics 方法建立生态水文模型,对比得出随机森林算法的预测误差比 Logistics 方法小。在生物信息学领域,Ahmed 等人[12]利用随机森林算法研究真菌聚集度与其他因素的关联性;Brieuc 等人[13]利用随机森林算法,通过标记寄生虫对鱼群进行判别。在医学领域,Sonobe 等人[14]利用随机森林思想对 CT 图肝脏节点进行建模分析;Ghosh、Carranza 等人[15,16]利用随机森林算法进行蛋白质关联性分析。在地理学领域,Ghaedih、Reza 等人[17,18]分别利用随机森林思想对地理遥感进行了研究。在信息安全领域,随机森林算法的应用还较少。

朴素贝叶斯算法和支持向量机只能应用于独立分布的数据,不能对数据流相关性的上下文建模;隐马尔可夫模型能够通过标记来利用上下文,但该方法对特征序列有过强的独立条件假设[19],因而不能利用数据流中的上下文信息。而且,对于分类特征是多维的情况,朴素贝叶斯算法假设各个特征独立,支持向量机假设特征符合多元高斯等概率分布,但真实的网络数据流与假设的分布有一定的偏差。由于随机森林分类模型有较强的融合上下文能力,并且样本数量大幅增加时不会过度拟合,本文利用 RFC(random forest classification)模型解决上述算法的不足。

2　DDoS 攻击特点和分类特征

2.1　DDoS 攻击特点

Bhuyan 等人[20]通过研究发现,在洪水攻击中,TCP 洪水攻击占据 90% 以上,UDP 洪水攻击和 ICMP 洪水攻击在 6% 左右。本文对 TCP、UDP 和 ICMP 洪水攻击进行研究,通过对当前常见的攻击案例、DARPA 和林肯实验室发布的 DDoS 攻击数据集进行分析,总结出三种多对一的特征。

a)数据流的源地址和目的地址具有多对一的关系。绝大部分的 DDoS 攻击都是依托于僵尸网络,僵尸网络由大量功能节点共同组成,这些节点包含普通 PC、服务器、移动设备等,具有不同的 IP 地址,当 C&C 服务器向傀儡机发送控制指令时,它们便会同时向目标主机发起攻击,攻击流量的源地址和目的地址为明显的多对一关系。

b)数据流的源地址和目的端口具有多对一关系。攻击者针对目标主机的某个服务进行攻击时,将向目标主机的固定端口发送大量数据流,如 DNS 拒绝服务攻击,就是向目标主机的 53 端口发送大量数据。

c)数据流的目的端口和目的地址具有多对一的关系。攻击者针对系统资源进行攻击时,会向目标机请求多项服务,随机产生不同的目的端口号,同时向多个端口进行洪水攻击,从而消耗目标主机的系统资源。

2.2　分类特征选取

根据三种攻击的特性,本文利用数据流信息熵对 TCP 洪水攻击、UDP 洪水攻击、ICMP 洪水攻击进行描述。信息熵是描述变量不确定性的一种度量,本文利用数据流信息熵来描述 DDoS 攻击的特点。根据信息论的定义,变量 Y 的信息熵如式(1)所示,其中 $p(y_i)$ 是变量 Y 的先验概率。

$$H(Y) = -\sum_i p(y_i)\log_2(p(y_i)) \tag{1}$$

变量 Y 关于 X 的信息熵如式(2)所示。其中 $p(y_i|x_j)$ 是 y_i 关于 x_j 的后验概率。

$$H(Y|X) = -\sum_j p(x_j)\sum_i p(y_i|x_i)\log_2(p(y_i|x_j)) \tag{2}$$

令 sourceIP、destinationIP、destinationPort 分别代表数据流的源地址、目的地址、目的端口,采用 SIDI、SIDP 和 DPDI 三个信息熵分别表征三种多对一的特征,这三种信息熵共同构成了数据流信息熵(DSIE),DSIE 体现了 SIDI、SIDP、DPDI 的不确定性。

以 SIDI 为例,DSIE 的计算方式如下:设一次采样的数据流总数为 S,这些数据流中源地址的集合为 $\{si_i \mid i=1,2,\cdots,N\}$,目的地址集合为 $\{di_i \mid i=1,2,\cdots,M\}$。定义矩阵 $A[M]$:$A[i]$ 表示目的地址为 di_i 的数据流数量,定义矩阵 $B[N][M]$,$B[i][j]$ 表示源地址为 si_i、目的地址为 di_i 的数据流数量。根据式(2)可得

$$\begin{aligned}\mathrm{SIDI} &= -\sum_j p(di_j)\sum_i p(si_i|di_j)\log_2(p(si_i|di_j)) = \\ &-\sum_{j=1}^{M}\frac{A[j]}{S}\sum_{i=1}^{N}\frac{B[i][j]}{A[j]}\log_2\left(\frac{B[i][j]}{A[j]}\right)\end{aligned} \tag{3}$$

3　基于随机森林分类模型的 DDoS 攻击检测方法

3.1　随机森林算法

随机森林算法是一种新的机器学习算法,该算法在运算量没有显著提高的前提下提高了预测精度。随机森林用随机的方式建立一个森林,森林由大量决策树组成,随机森林中决策树之

间没有关联。在建成森林后，当有一个新的输入样本进入时，森林中的决策树分别进行判断，根据决策情况确定样本类别。

随机森林算法结合 Bagging 算法和随机子空间算法的特点和优势，以决策树作为分类器，训练时利用 Bagging 算法的无放回抽样，同时结合随机子空间算法在训练集中只抽取部分样本进行训练，结果由决策树投票决定。

在随机森林中，每一棵分类树即为一棵遵循递归分裂原则的二叉树。二叉树从根节点开始依次对训练集进行划分，且其生成是按照自顶向下的原则。其中，包含全部训练数据的根节点根据纯度最小原则分裂为左、右节点，且其分别包含着训练数据的一个子集。节点根据同样规则继续分裂，当满足分支停止规则时，停止生长。算法的实现步骤如下：

a）初始化数据集 D，利用 Bootstrap 法每次从数据集中有放回地随机取出 k 个样本集，从而生成 k 个分类树。

b）从 k 棵分类树中每棵树的节点随机获得 s 个变量，从这些变量中挑选出最具代表性的变量，分类的阈值由多个分类点共同确定。

c）不对分类树作修剪处理，使其无限生长。

d）最终生长出的多棵分类树共同构成了随机森林，新的样本通过构造的随机森林进行划分，分类结果由分类器投票决定。

3.2 RFC 模型的训练

RFC 模型中随机树的每个节点可以看成一个弱分类器，对到达该节点的训练样本集 Ω 计算得到一个分类准则 $h(x,\theta)\in\{0,1\}$。$x\in R^M$ 表示一个训练样本，$\theta=\{\phi,\Psi\}$ 为这个弱分类器的参数，其中，$\phi(\cdot)$ 为筛选函数，Ψ 为参数列向量或者参数矩阵；θ 决定了弱分类器的分类超平面的样式。

a）非线性分类面，如式(4)所示。其中，$\delta(\cdot)$ 是一个指示函数。对于样本 $x=(x_1,x_2,x_3)\in R^3$，令 $\phi(x)=(x_1,x_3,1)^T$，$\Psi=(\omega_1,\omega_2,\tau)$，$h(\cdot)$ 表示一个分类面。

$$h(x,\theta)=\delta(\phi^T(x)\Psi\phi(x)>0) \tag{4}$$

b）线性分类面，如式(5)所示，其中 Ψ 为一个参数矩阵。

$$h(x,\theta)=\delta(\phi^T(x)\Psi>0) \tag{5}$$

当样本满足 $h(x,\theta)=1$，落入左子树，为理性行为；反之落入右子树，为非理性行为。递归下去直至落入节点的样本个数低于阈值，或者达到规定的最大深度。递归结束后，这个节点称做叶子节点。在每个节点处找到最优的系数 θ^* 使得训练样本取得最佳效果，如式(6)所示。

$$\theta^*=\operatorname{argmax} IG_{\theta_j\in\Gamma_{sub}}(\theta_j|\Omega) \tag{6}$$

其中：Γ_{sub} 为完整参数空间 Γ 的子集，对于每个节点 Γ_{sub} 都是从 Γ 中随机选择的，从而体现节点分裂过程中的随机性。$IG(\cdot)$ 表示信息增益，衡量分裂后样本不纯度的下降幅度。定义为

$$IG(\theta|\Omega)=h(\Omega)-\sum_{i\in\{l,r\}}\frac{|\Omega_i(\theta)|}{\Omega}H(\Omega_i(\theta)) \tag{7}$$

其中：$\Omega=\{(x_i,y_i)\}_{i=1}^N$ 表示落入该节点的所有样本的集合，$|\Omega|=N$；$\Omega_l(\theta)$ 和 $\Omega_r(\theta)$ 分别表示在参数 θ 下落入左右子节点的样本集；$h(\Omega)$ 表示落入一个节点的样本集的不纯度，用信息熵来表示，如式(8)所示。其中，N_c 为样本类别个数，$p(c|\Omega)$ 表示样本集 Ω 中类别 c 所占的比例。

$$H_{entropy}(\Omega)=-\sum_{c=1}^{N_c}p(c|\Omega)\log p(c|\Omega) \tag{8}$$

利用该方法研究二分类问题，$H_{entropy}(\Omega)$ 随着两个类别分布变化的曲线如图 1 所示，横坐标表示一个类别的占比。从图 1 中可以看出，当两个类别的比例越接近，信息熵就越大，即表明此时节点的不纯度高。

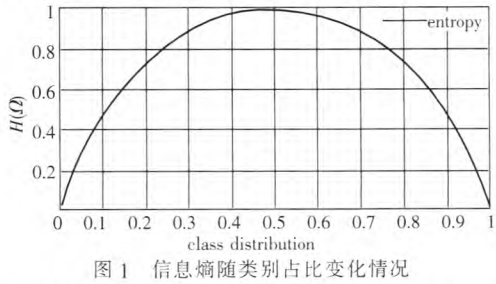

图 1 信息熵随类别占比变化情况

由式(8)可知，每个节点的“最优”参数 θ 应使节点的分裂后不纯度下降幅度最大，这个参数从 Γ_{sub} 中选取。

RFC 模型的训练算法如算法 1 所示。

算法 1 RFC 模型训练过程(RFC_train)

```
algorithm RCF_train(T,p,k){
for(int i=1;i<=n;i++){
  //训练 n 棵决策树
  T' = sample_withResample(T,p);
  //有放回从训练集中抽取 p 个样本
  Att = getAttributes(T');//获取特征集
  Att' = sample_withoutResample(Att,k);
  //无放回从特征集中抽取 k 个特征
  T'' = remainAttributes(T',Att');
  //对 T'只保留 Att'中含有的特征
  DT[i] = createDecisionTree(T'');
  //构建决策树(设决策树数组为 DT)
}
return DT;
}
```

3.3 RFC 模型的分类

RFC 模型训练结束后，测试样本 x 经过每棵树到达某个叶子节点，那么样本 x 属于 c 的概率为

$$p(c|x)=\frac{1}{T}\sum_{t=1}^{T}p_t(c|x) \tag{9}$$

其中：T 为森林中随机树的数量，$p_t(c|x)$ 为叶子节点的类别分布。那么对 x 类别的决策为

$$c=\underset{c\in\{\Gamma,\cdots,N_c\}}{\operatorname{argmax}}p(c|x) \tag{10}$$

RFC 模型的分类过程是一种多数投票过程。定义测试样本集合为 $E=\{e_1,\cdots,e_m\}$，训练得到的决策树集合为 $DT=\{dt_1,\cdots,dt_n\}$；记录每个决策树得到分类结果的下标数组为 $CIR[n]$，类别集合为 $C=\{c_1,\cdots,c_n\}$，分类结果集合为 $CR=\{CR_1,\cdots,CR_m\}$，则 RFC 模型分类过程如算法 2 所示。

算法 2 RFC 模型分类过程(RFC_classification)

```
algorithm RFC_classify(DT,E){
for(int i=1;i<=m;i++){
  for(int j=1;j<=n;j++){
    ClR[j]=0;//清空 ClR 数组
  }
  for(int j=1;j<=n;j++){
    //用第 j 棵决策树对 E[i]进行分类,并记录分类结果的下标
  int classifyResultIndex = classify(E[i],DT[j]);
    ClR[classifyResultIndex]++;
  }
  int maxIndex = getMaxAppeared(ClR);
  //找出 ClR 数组中出现次数最多的值
  CR[i] = C[maxIndex];//最终的分类结果
  }
  return CR;
}
```

4 仿真实验

本文仿真实验采用四个数据集，其中包含两个林肯实验室

2000 年的 DDoS 数据集,分别是 LLDoS 1.0 和 LLDoS 2.0.1。这两个数据集是 TCP 洪水攻击数据包,数据报文的源地址和目的端口号都是随机生成的,带有 ACK 标志。另外两个数据集分别针对 UDP 洪水攻击和 ICMP 洪水攻击,是通过本地三台主机安装 TFN2K 攻击软件来向目标主机发起攻击,网络环境是一个 100 Mbps 的局域网;报文发送速度为 15 000 个/s 和 8 000个/s,持续发送约 3 min。区别于 TCP 洪水攻击,本地主机发起的攻击采用的是真实的源地址,并且攻击固定的端口,如此能检测出 RFC 模型对小范围地址攻击的防御能力。在目标主机上安装 Wireshark,捕获攻击报文来获得这两组数据集。由于以上四组数据集都只是攻击样本,为了实验的通用性,在四组攻击数据集中加入了正常的背景流量,背景流量取自高校校园骨干网。评价指标选用检测率、误报率、总错误率三个指标对实验结果进行评价,同时 TP 表示正确标记的理性行为,FP 表示错误标记的理性行为,TN 表示正确标记的非理性行为,FN 表示错误标记的非理性行为,则有

$$DR = \frac{TN}{TN + FN} \tag{11}$$

$$FR = \frac{FP}{TP + FP} \tag{12}$$

$$ER = \frac{FN + FP}{TP + FP + TN + FN} \tag{13}$$

仿真实验分为三个部分,分别对 TCP 洪水攻击、UDP 洪水攻击、ICMP 洪水攻击等三种攻击方式进行检测,并将检测的结果与 HMM 算法、SVM 算法进行比较,以验证本文提出的 RFC 模型是否具有更好的识别能力。实验采用 Sherwood 工具包进行 RFC 模型训练,采用 GT2K 工具包进行 HMM 的测试,采用 LIBSVN 工具包对 SVM 进行测试。

1)对 TCP 洪水攻击的检测

对正常流量和攻击流量交叉采样,计算每次采样的分类行为,得到理性行为和非理性行为两类样本集。为了验证该方法对流量持续增加情况有较好的处理能力,缩短采样周期方式来降低背景流量。背景流量的周期(T)从 2 s 逐渐增加到 10 s,每次以 2 s 的时间间隔递增,而攻击流量的采样周期(t)则固定为 0.01 s。

林肯实验室的两个数据集 LLDoS 1.0 与 LLDoS 2.0.1,一个用于模型训练,另一个用于模型测试。将样本按时间划分为若干个序列,以每个序列为单位进行 RFC 模型训练和测试。序列的长度需要适中,太短会影响上下文的利用,太长会增加检测的时间。本实验将样本序列的长度定为 50,检测时间为 0.5 s,对于持续时间为几十分钟的 DDoS 攻击是合理的。按照上述不同周期进行采样,得到五组数据,每组数据都是 1 500 个样本,也就是 30 个样本序列。图 2 和 3 显示了 T=2 s 时,训练集的 SIDI 和 DPDI 特征时间序列曲线图。

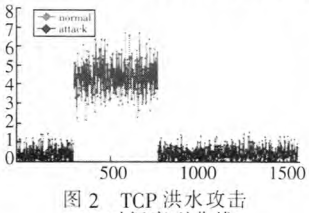

图 2　TCP 洪水攻击 SIDI 时间序列曲线

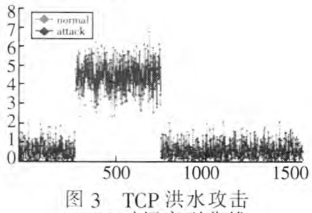

图 3　TCP 洪水攻击 DPDI 时间序列曲线

攻击数据流的源地址和目的端口是随机生成的,因此存在前面所描述的多对一关系。从图 2 中可以看出,攻击时间段的非理性行为的特征值与理性行为的特征值有明显差异,表明该方法能够很好地区分理性行为和非理性行为。

使用训练集对 RFC 模型训练后再对测试集进行检测,并将检测的结果与 HMM、SVM 算法的效果进行比较,表 1 给出了三种算法的检测情况。

表 1　三种算法对 TCP 洪水攻击的检测情况比较　/%

算法	指标	T/s				
		2	4	6	8	10
RFC	*DR*	99.17	98.68	98.54	98.13	97.89
	FR	0.14	0.15	0.15	0.17	0.20
	ER	0.07	0.33	0.43	0.51	0.72
HMM	*DR*	98.21	97.51	96.57	95.13	93.28
	FR	0.54	0.82	0.93	1.21	2.05
	ER	0.67	0.73	0.94	1.44	3.12
SVM	*DR*	98.17	97.32	96.12	94.54	93.40
	FR	0.27	0.53	0.44	0.71	1.27
	ER	1.07	1.50	1.62	1.80	2.41

从表 1 可以看出,当背景流量较少时,即 T 较小时,三种算法的检测差别不大,都能较为准确地区分理性行为和非理性行为;但当背景流量逐渐增多时,RFC 模型在检测率、误报率、总错误率三个指标上有更好的结果。在 T 为 10 s 时,仍有 97.89 的检测率,这表明 RFC 模型在抗背景干扰能力上表现较好。

2)对 UDP 洪水攻击的检测

将实验得到的 UDP 洪水攻击的数据集和背景流量进行混合。攻击流量的采样周期从 5 ms 开始递增,直到 25 ms,每次增加 5 ms,而背景流量的采样周期固定在 5 s。得到的五组数据中,每组数据都取连续 1 000 个样本,每 50 个样本为一个样本序列,共 20 个样本序列。

攻击过程的源地址为固定的三个本地主机地址,目的端口也是固定的,因此源地址和目的端口、源地址和目的地址存在多对一的关系。以 t 取 20 ms 时得到的训练模型为例,分类特征 SIDP 和 SIDI 的时间序列如图 4 和 5 所示。

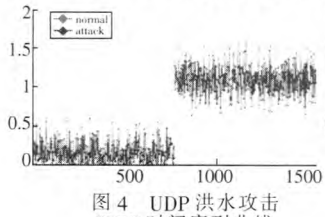

图 4　UDP 洪水攻击 SIDP 时间序列曲线

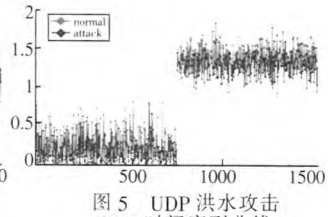

图 5　UDP 洪水攻击 SIDI 时间序列曲线

从图 4 和 5 可以看出,两个特征值 SIDP 与 SIDI 在非理性行为时取值总体上要高于理性行为时的取值,能够说明即使源地址较为集中的情况下,数据流信息熵仍然能够区分理性行为和非理性行为。

利用三种算法对五组数据进行检测,检测效果如表 2 所示。从表 2 可以看出,RFC 模型充分利用上下文信息能够降低背景流量带来的模糊性,在 T=5 ms 时,有 93.76 的检测率,明显高于 HMM 和 SVM 算法两种算法的检测效果。

表 2　三种算法对 UDP 洪水攻击的检测情况比较　/%

算法	指标	T/s				
		5	10	15	20	25
RFC	*DR*	93.76	95.68	97.54	98.43	98.79
	FR	3.21	0.90	0.53	0.42	0.40
	ER	4.10	1.51	0.93	0.55	0.12
HMM	*DR*	86.54	91.41	94.37	95.66	97.81
	FR	6.61	1.12	0.94	0.71	0.52
	ER	10.11	4.37	3.12	2.47	1.44
SVM	*DR*	84.23	94.31	96.37	97.17	98.33
	FR	6.72	3.63	3.14	2.65	1.94
	ER	9.74	4.50	4.10	2.82	1.82

3)对 ICMP 洪水攻击的检测

将 ICMP 洪水攻击数据集与背景流量结合,攻击流量的采样周期为 1 ms,背景流量的采样周期从 2 s 逐渐增加到 10 s,每次增加 2 s。同样得到五组数据,每组数据为连续的 1 500 个样本,共 30 个样本序列。图 6 和 7 为 T 取 2 s 时特征值 SIDI、SIDP 随时间的变化。

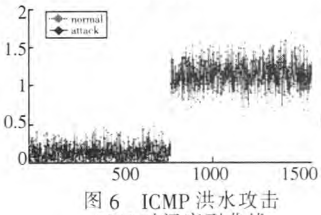

图 6 ICMP 洪水攻击 SIDI 时间序列曲线

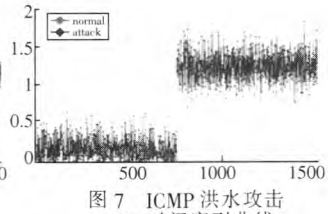

图 7 ICMP 洪水攻击 SIDP 时间序列曲线

从图 6 和 7 中可以看出,特征值 SIDI 和 SIDP 的取值在理性行为时总体上高于非理性行为时的取值,能够较为容易地区分出是否为理性行为。

分别用三种算法对五组数据进行检测,检测情况如表 3 所示。从表 3 中可以看出,随着背景流量的增多,三种算法都会受到其影响,但 RFC 模型的检测优势明显增大,总体上都有 97% 以上的检测率。在 T = 10 s 时仍有 97.37 的检测率,效果优于 HMM 和 SVM 方法。

表 3 三种算法对 ICMP 洪水攻击的检测情况比较 /%

算法	指标	T/s				
		2	4	6	8	10
RFC	*DR*	99.13	98.93	98.42	97.82	97.37
	FR	0.12	0.27	0.69	1.06	1.23
	ER	0.13	0.41	0.86	1.44	2.03
HMM	*DR*	97.27	95.45	94.97	93.72	92.13
	FR	1.34	2.04	2.53	3.09	3.41
	ER	1.72	3.17	3.89	4.12	7.13
SVM	*DR*	98.21	97.12	96.43	94.42	93.14
	FR	0.92	1.22	2.78	3.35	4.21
	ER	1.13	2.21	3.10	4.51	6.36

5 结束语

本文提出一种新的基于 RFC 模型的 DDoS 检测方法,利用该模型的融合上下文和随流量增加拟合度稳定的特性,弥补现有机器学习算法的不足。引入数据流信息熵概念,根据 SIDI、SIDP、DPDI 三个特征,对 TCP 洪水攻击、UDP 洪水攻击、ICMP 洪水攻击进行特征描述,并分别建立检测模型,力求准确检测 DDoS 攻击。仿真实验结果显示,该模型能够较为准确地区分正常流量和攻击流量,与 HMM、SVM 方法比较,基于 RFC 模型的 DDoS 检测方法有较高的检测率和较低的误报率。

参考文献:

[1] Zargar S T, Joshi J, Tipper D. A survey of defense mechanisms against distributed denial of service(DDoS) flooding attacks[J]. IEEE Communications Surveys & Tutorials, 2013, 15(4): 2046-2069.

[2] Wang Bing, Zheng Yao, Lou Wenjing, *et al*. DDoS attack protection in the era of cloud computing and software-defined networking[J]. Computer Networks, 2015, 81(4): 308-319.

[3] Yu Shui, Zhou Wanlei, Jia Weijia, *et al*. Discriminating DDoS attacks from flash crowds using flow correlation coefficient[J]. IEEE Trans on Parallel and Distributed Systems, 2012, 23(6): 1073-1080.

[4] Chen Zhaomin, Yeo C K, Francis B S L, *et al*. A MSPCA based intrusion detection algorithm tor detection of DDoS attack[C]//Proc of IEEE International Conference on Communications. Piscataway: IEEE Press, 2015: 1-5.

[5] Yu Shui, Tian Yonghong, Guo Song, *et al*. Can we beat DDoS attacks in clouds? [J]. IEEE Trans on Parallel and Distributed Systems, 2014, 25(9): 2245-2254.

[6] Kotenko I, Ulanov A. Agent-based simulation of DDOS attacks and defense mechanisms[J]. International Journal of Computing, 2014, 4(2): 113-123.

[7] Gupta B B, Joshi R C, Misra M. ANN based scheme to predict number of zombies in a DDoS attack[J]. International Journal of Network Security, 2012, 14(2): 61-70.

[8] Tama B A, Rhee K H. Data mining techniques in DoS/DDoS attack detection: a literature review[C]//Proc of the 3rd International Conference on Computer Applications and Information Processing Technology. 2015: 23-26.

[9] Karnwal T, Sivakumar T, Aghila G. A comber approach to protect cloud computing against XML DDoS and HTTP DDoS attack[C]// Proc of IEEE Student's Conference on Electrical, Electronics and Computer Science. Piscataway: IEEE Press, 2012: 1-5.

[10] Lindner C, Bromiley P A, Ionita M C, *et al*. Robust and accurate shape model matching using random forest regression-voting[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1862-1874.

[11] Jia Jianhua, Liu Zi, Xiao Xuan, *et al*. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach[J]. Journal of Theoretical Biology, 2016, 394(4): 223-230.

[12] Ahmed O S, Franklin S E, Wulder M A, *et al*. Characterizing stand-level forest canopy cover and height using landsat time series, samples of airborne LiDAR, and the random forest algorithm[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 101(3): 89-101.

[13] Brieuc M S O, Ono K, Drinan D P, *et al*. Integration of random forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon[J]. Molecular Ecology, 2015, 24(11): 2729-2746.

[14] Sonobe R, Tani H, Wang Xiufeng, *et al*. Random forest classification of crop type using multi-temporal TerraSAR-X dual-polarimetric data [J]. Remote Sensing Letters, 2014, 5(2): 157-164.

[15] Ghosh A, Sharma R, Joshi P K. Random forest classification of urban landscape using Landsat archive and ancillary data: combining seasonal maps with decision level fusion[J]. Applied Geography, 2014, 48(3): 31-41.

[16] Carranza E J M, Laborte A G. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines)[J]. Computers & Geosciences, 2015, 74(1): 60-70.

[17] Ghaedi M, Ghaedi A M, Negintaji E, *et al*. Random forest model for removal of bromophenol blue using activated carbon obtained from Astragalus bisulcatus tree[J]. Journal of Industrial and Engineering Chemistry, 2014, 20(4): 1793-1803.

[18] Reza P R M, Toomanian N, Khormali F, *et al*. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran[J]. Geoderma, 2014, 232-234(12): 97-106.

[19] Tao Yuan, Yu Shu. DDoS attack detection at local area networks using information theoretical metrics[C]//Proc of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Piscataway: IEEE Press, 2013: 233-240.

[20] Bhuyan M H, Bhattacharyya D K, Kalita J K. An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection [J]. Pattern Recognition Letters, 2015, 51(1): 1-7.