IMS 中基于遗传神经网络的 DOS 攻击检测模型

王晓雷,郭云飞,李 贺
WANG Xiao-lei,GUO Yun-fei,LI He

国家数字交换系统工程技术研究中心,郑州 450002
National Digital Switching System Engineering & Technological Research Center,Zhengzhou 450002,China
E-mail:shelraywang@gmail.com

WANG Xiao-lei,GUO Yun-fei,LI He.DOS detection model in IMS based on GA neural network.Computer Engineering and Applications,2008,44(28):116-118.

Abstract: This paper first analyzes the specialty of DOS attack in IMS,and then proposes a DOS detection model applied in IMS based on GA neural network.The detection model uses GA to optimize the weights and biases of neural network.In the end, the paper makes a detection experiment using the data of kddcup99.The experiment shows that the detection probability of neural network with GA is much better than the model without GA.

Key words: IP Multimedia Subsystem(IMS);Denial of Service(DOS);neural network;Genetic Algorithm(GA)

摘 要:首先针对 IMS 中 DOS 攻击的特点进行了分析,然后基于遗传神经网络提出了一种应用于 IMS 的 DOS 攻击检测模型,在模型中利用遗传算法对神经网络的权值和偏置值进行优化。最后利用 kddcup99 数据集进行攻击检测实验。通过实验可以看出经过遗传算法优化后,神经网络的检测概率得到了明显提高。

关键词:IP 多媒体子系统;DOS;神经网络;遗传算法
DOI:10.3778/j.issn.1002-8331.2008.28.039 **文章编号:**1002-8331(2008)28-0116-03 **文献标识码:**A **中图分类号:**TP181

1 引言

3GPP 在 R5 版本中提出了 IMS 标准。IMS 作为未来网络融合的控制平台,具有分布式、与接入无关,以及标准开放的业务控制接口等特点。基于全 IP 架构的 IMS 在提供丰富的多媒体业务的同时,也带来了移动网络 IP 化过程中的一系列安全问题,Internet 上的诸多安全威胁在 IMS 中广泛存在。DOS 攻击是任何基于 IP 的网络都无法避免的,通常采用 IP 欺骗等方式向服务器发送大量的无用数据包,使其不能提供正常的服务。如何使 IMS 网络避免或减少 DOS 攻击带来的危害成为一个新的研究方向。

2 IMS 中的 DOS 攻击

IMS 通过建立安全联盟 SA(Security Association)对消息的完整性和机密性进行保护,在 IMS 的接入安全中,用户 UE 和 IMS 之间建立 SA 所需的参数通过注册流程中的 SIP 消息进行传递。SA 建立后 IMS 能够杜绝 IP 地址欺骗,但是在 SA 建立过程中所传送的消息却没有得到任何保护[1-2]。IMS 的注册流程中共有三个地方需要 DNS 进行域名解析:P-CSCF 发现流程、定位 I-CSCF 地址、定位 HSS 地址。在此三处域名解析流程中 SA 还没有建立,而且 3GPP/3GPP2 没有对如何防止拒绝服务对 DNS 的攻击作相关定义,因此 IMS 中的 DNS 面临着 DOS 攻击的威胁。

IMS 采用 SIP 作为其会话控制协议,SIP 采用基于文本形式表示的语法、语义和编码,基于 UDP 承载。由于 UDP 协议是一种无连接的服务,只要对攻击者开放有一个 UDP 服务端口,即可针对该服务发动攻击。UDP 攻击通常可分为 UDP Flood 攻击、UDP Fraggle 攻击和 DNS Query Flood 攻击。

(1)UDP Flood 攻击中,攻击者发送大量虚假源 IP 的 UDP 数据包或畸形 UDP 数据包,从而使被攻击者不能提供正常的服务,甚至造成系统资源耗尽、系统死机。

(2)UDP Fraggle 攻击的原理与 smurf 攻击相似,也是一种“放大”式的攻击,不同的是它使用 UDP 回应代替了 ICMP 回应。

(3)DNS Query Flood 攻击是一种针对 DNS 服务器的攻击行为。攻击者向 DNS 的 UDP 53 端口发送大量域名查询请求,占用大量系统资源,使服务器无法提供正常的查询请求。

以 P-CSCF 发现流程中的 DNS 域名请求为例,采用 DNS Query Flood 攻击,设计攻击流程如图 1。

图 1 中攻击者 attacker 截获 UE 发向 DNS 的域名解析请求,并把 IP 地址伪装成 UE 的地址,然后向 DNS 发送大量的域名查询数据包直到服务器无法正常工作。

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z434)。
作者简介:王晓雷(1982-),男,硕士研究生,主要研究方向为无线通信安全;郭云飞(1963-),男,教授,博士生导师,主要研究方向为信息网络与交换、下一代网络体系架构,863 通信主题专家组组长;李贺(1981-),男,满族,硕士研究生,主要研究方向为通信网络安全。
收稿日期:2007-11-20 **修回日期:**2008-02-26

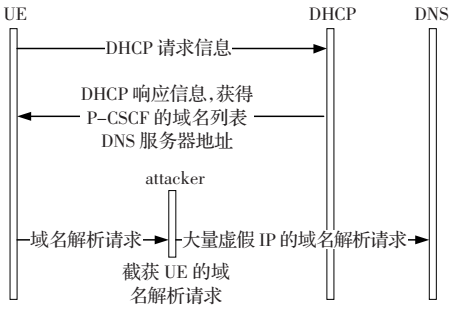


图 1　IMS 中 DOS 攻击流程

3　研究现状

对 DOS 攻击的检测可以归于入侵检测领域，基于数据挖掘的入侵检测是目前该领域中的研究热点。综合现有各类研究机构的检测方案,基于数据挖掘的入侵检测系统大概可分为五个模块,如图 2 所示。

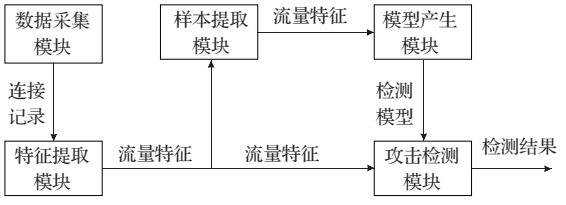


图 2　入侵检测框图

其中,数据采集模块依据数据收集策略采集网络中的数据,并将其转换为连接记录提交给特征提取模块。 特征提取模块将连接记录通过特定的算法转换成流量特征。样本提取模块从先前的流量特征中提取出样本,输入到模型产生模块中产生检测模型。最后,检测模型在攻击检测模块中判断流量特征是否正常,从而产生检测结果。

目前的研究主要集中在模型产生模块上,主要思想就是利用分类算法对流量特征进行分类，从而达到攻击检测的目的。例如,美国哥伦比亚大学的 Mohiuddin 等人提出了一种基于贝叶斯分类挖掘的算法[3];Portnoy 等人提出了一种基于距离的聚类算法[4]等。此外还产生了许多基于密度的、基于模型的检测方法[5]。但是这些方案都存在诸如误警率比较高、训练样本规模太大、实时性较差等缺点。

神经网络是在对人脑组织结构和运行机制认识理解的基础上,模拟其结构和智能行为的一种工程系统,具有很好的非线性拟合和模式识别能力。随着国内外对数据挖掘和入侵检测技术研究的深入,已经有人提出利用神经网络对流量特征进行聚类的思想[6],但是已有研究方案没有对选取何种神经网络以及网络的权值和偏置值如何设定做具体论述。

4　IMS 中 DOS 攻击检测模型

4.1　IMS 中的流量特征

攻击者最有可能对 IMS 中的 DNS 发起 DOS 攻,而且 IMS 中的 SIP 消息基于 UDP 承载,所以一个 DNS 遭受 DOS 攻击后的最显著特征是发往该 DNS 地址和端口的 UDP 数据包大量增加。因此,IMS 中的 DOS 攻击检测模型只需要对进入某 DNS 的数据流进行检测。

以时间窗为单位处理连接记录,即网络数据根据发生的时间划分到一个个的时间窗内,把时间窗内的连接记录转换为流量特征。流量特征是连接记录中对分类结果影响较大的特征属性,是检测数据的高层抽象。针对 IMS 中 DOS 攻击的特点,可以规定属性组合为(服务类型目的地址,UDP 连接结束标志)的频繁项目集为流量特征,为确保产生有意义的规则,再引入两个表示数量的参数,流量特征的内容如表 1 所示。

表 1　流量特征的内容

特征名称	特征含义
Service	目的端口号(服务类型)
dstIP	目的 IP 地址
Status	连接结束状态标志
Count_conn	该时间窗内具有相同 Service、dstIP、Status 的连接记录个数
Count_total_conn	该时间窗内连接记录总个数

流量特征的标准化:对流量特征的 5 个分量按照比例全部转换为区间[-1,1]中的数据,转换后的流量特征称为标准流量特征。每个分量的具体转换方法如下。

(1)Service:如果是 DNS 的端口 P_0,则置为-1;否则置为 1。

(2)dstIP:如果是 DNS 的地址 IP_0,则置为-1;否则置为 1。

(3)Status:如果标志为 S_0,则转化为-1;否则转化为 1。

(4)Count_conn:把所有流量特征中最小的 Count_conn 值置为-1,最大的 Count_conn 值置为 1,Count_conn 的总数为 n,则 Count_conn 值从小到大第 i 个值依次转换为:

$$-1+\frac{1-(-1)}{n-1}\cdot i \quad (i=0,1,\cdots,n-1) \tag{1}$$

(5)Count_total_conn:依据 Count_conn 的转换方法。

4.2　检测模型的原理

4.2.1　神经网络模型的建立

流量特征可以看作是五维向量,把流量特征作为神经网络的输入,则输出为对流量特征的分类,即检测结果。采用 BP 神经网络,设计三层 BP 神经网络模型如图 3。

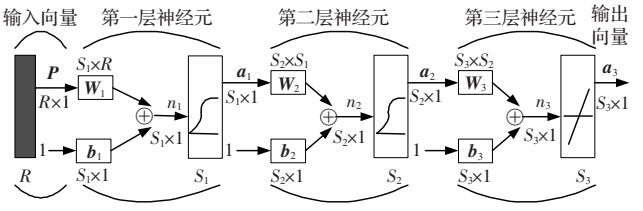


图 3　三层 BP 神经网络模型

在图 3 中,$\boldsymbol{P}$ 为输入向量,R 为输入向量的维数,$S_i(i=1,2,3)$代表第 i 层的神经元数,$\boldsymbol{W}_i(i=1,2,3)$为第 i 层的权值矩阵,$\boldsymbol{b}_i(i=1,2,3)$为第 i 层的偏置值向量。网络的第一层和第二层使用 sigmoid 传输函数，输出层第三层使用线性传输函数,$\boldsymbol{a}_1$、$\boldsymbol{a}_2$、$\boldsymbol{a}_3$ 分别为第 1、2、3 层的输出向量,则神经网络最后输出 $\boldsymbol{a}_3$ 为:

$$\boldsymbol{a}_3=purelin(\boldsymbol{W}_3\boldsymbol{a}_2+\boldsymbol{b}_3)=$$
$$purelin(\boldsymbol{W}_3\text{sigmoid}(\boldsymbol{W}_2\text{sigmoid}(\boldsymbol{W}_1\boldsymbol{P}+\boldsymbol{b}_1)+\boldsymbol{b}_2)+\boldsymbol{b}_3) \tag{2}$$

设训练样本的个数为 n,在以上三层神经网络中样本对应的目标输出 t_3,则神经网络实际输出和目标输出的学习误差 E 可以表示为:

$$E=E[e^2]=\frac{1}{nS_3}\sum_{j=1}^{n}\sum_{i=1}^{S_3}[(\boldsymbol{t}_3)_i^j-(\boldsymbol{a}_3)_i^j]^2 \tag{3}$$

其中,$(\boldsymbol{t}_3)_i^j$ 为第 j 个样本目标输出向量中的第 i 个分量,$(\boldsymbol{a}_3)_i^j$

为第 j 个样本实际输出向量中的第 i 个分量。

学习算法采用标准 BP 算法的改进之一弹力 BP 算法[7](Resilient back-PROPagation,RPROP)。RPROP 算法用于模式识别问题时,收敛速度是最快的,但是标准 BP 算法是一种梯度下降学习算法,存在结果易陷入局部极值点和引起振荡效应等缺点,单独使用 BP 神经网络效果不是很理想。

4.2.2　遗传算法优化神经网络的具体流程[8]

遗传算法(Genetic Algorithm,GA)是一种高效的随机搜索与优化方法,具有很强的全局搜索能力。利用 GA 对神经网络的初始权值和偏置值进行优化,可以使弹力 BP 学习算法避免在迭代过程中陷入局部最优解,从而求出最优的权值和偏置值。具体优化流程如下:

步骤 1　随机产生神经网络的一组实数型的权值和偏置值矩阵,将所有权值矩阵的行向量和偏置值向量按顺序连接在一起,构成一个个体。产生 N 个这种个体,构成初始种群。

步骤 2　设置适应度函数如下:

$$fitness=1/E \tag{4}$$

其中,E 为式(3)中的网络学习误差。对每一个个体 i 计算其适应度 f_i 和 $E_i(i=1,2,\cdots,N)$。对于预先给定的误差 ε,如果

$$E_i=\min(E_1,E_2,\cdots,E_N)<\varepsilon \quad (i=1,2,\cdots,N) \tag{5}$$

则 E_i 对应的个体为满足误差要求的最优初始权值和偏置值,利用弹力 BP 学习算法对该初始值进行迭代求出最优权值和偏置值。否则转到步骤 3。

步骤 3　采用二进制编码方法,对每个个体进行编码。设个体中每个参数的编码长度为 λ,每个参数的取值范围为($U_{\min}$,$U_{\max}$),则编码精度为:

$$\delta=(U_{\max}-U_{\min})/(2^{\lambda}-1) \tag{6}$$

步骤 4　执行比例选择算子,选择操作采用轮盘选择方法。个体被选中并遗传到下一代群体中的概率与该个体的适应度大小成正比,父代中适应度大的个体有可能直接复制到下一代。个体被选中的概率 p_i 为:

$$p_i=f_i\Big/\sum_{j=1}^{N}f_j \quad (i=1,2,\cdots,N) \tag{7}$$

步骤 5　执行交叉算子,使子代含有两个亲代的遗传基因。设定交叉概率

$$P_c=N_c/N \tag{8}$$

其中,N_c 为种群中被交换个体的数量,N 为种群中个体的总数。

步骤 6　执行变异算子,使子代中产生新个体。设定变异概率

$$P_m=B/(N\cdot L) \tag{9}$$

其中,B 为每代中的变异基因的数目,L 为个体中基因编码串的长度,N 为种群中个体的总数。

步骤 7　对种群中的个体进行译码,得到一组关于权值和偏置值的新种群,转至步骤 2。

5　攻击检测实验

5.1　检测模型的产生

样本数据采用林肯实验室的 kddcup99 数据集,该数据集记录 9 个星期内的原始网络数据包。每条连接记录包含 42 个字段,前 1 到 41 字段为特征属性,包括了服务类型、连接状态标志、以及同一时间窗内相同目的端口和地址的连接数等信息,第 42 个字段为标记字段。每条记录都被标记为正常或者是以下四种类型的异常行为之一:(1)DOS 攻击;(2)R2L 非授权远程访问;(3)U2R 非授权使用本地超级用户特权;(4)probing 扫描攻击。从第一个星期的连接记录中提取出时间相邻的 2 000 条正常记录和 DOS 攻击记录。

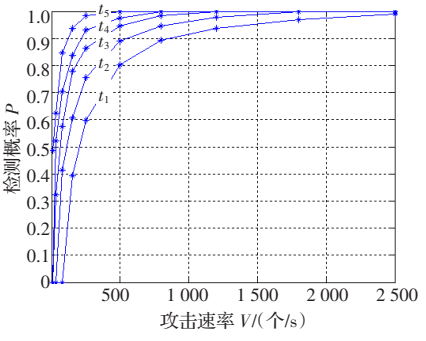
提取正常和攻击记录的流量特征,把流量特征进行标准化处理,并作为样本数据,则样本个数 n=4 000。其中,正常记录目标输出为 1,攻击记录目标输出-1。设神经网络各层的神经元数目 S_1=5,S_2=7,S_3=1,采用遗传神经网络的学习算法对样本进行学习,可以得到一组神经网络的最优权值和偏置值,由此得到检测模型。

5.2　攻击数据的产生

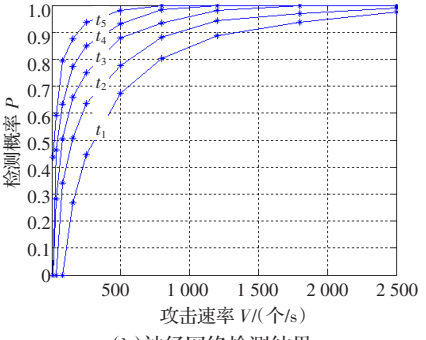
攻击数据由 UDP flood 工具产生。实验中服务器使用 OpenSER(Open SIP Express Router)担任 DNS 的角色,地址为 IP_0,端口为 P_0,主机操作系统使用 Redhat Enterprise Linux 4。实验在 100 M 局域网中进行,服务器和 Attacker 的硬件配置为 P-4 1.6 GHz 双核处理器、1 GB 内存。Attacker 利用 UDP flood 工具在相同的持续时间内向服务器的地址和端口产生 10 次 UDP 洪泛攻击,每次采用不同的攻击速率,并采用随机方法产生攻击数据包的源 IP 地址。在服务器端使用 ethereal 抓包工具提取每次 DOS 攻击的数据。从林肯实验室第二个星期的连接记录中随机提取出时间相邻的 2 000 条正常连接记录,把 10 次攻击数据分别插入到正常的连接记录中,产生 10 组待连接记录。对待检测连接记录进行统计,提取待检测流量特征。同样方法,Attacker 分别取 5 种不同的持续时间 $t_1,t_2,\cdots,t_5$,产生 5×10 组待检测的流量特征。把产生的待检测流量特征转换为标准流量特征后,即可以输入检测模型进行检测。

5.3　攻击检测结果

攻击时间取 t_1=5 s,t_2=10 s,t_3=20 s,t_4=30 s,t_5=60 s,10 次攻击速率分别取每秒发包数为 10、30、80、150、250、500、800、1 200、1 800、2 500 个。当采用遗传算法优化神经网络时,产生的检测结果如图 4(a);当只用神经网络时,产生的检测结果如图 4(b)。



(a)遗传神经网络检测结果



(b)神经网络检测结果

图 4　检测结果对比图

(下转 157 页)