计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):243−246 [doi: 10.15888/j.cnki.csa.005894]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

SVM 算法在 Linux 防火墙中的应用①

吕赵明, 张颖江, 周　昕, 陈　琼

(湖北工业大学 计算机学院, 武汉 430068)

摘　要: Linux 防火墙为开发人员提供了一种可扩展的机制. 本文通过对支持向量机原理的深入研究, 提出了基于 SVM 的 Linux 防火墙系统的设计与实现. 利用 Netfilter 框架捕获网络数据包, 在用户态通过支持向量机算法模块对异常网络流量进行分类, 并动态的添加 Iptables 规则, 从而抵御网络攻击. 实验证明, 该系统模型对异常流量的分类有很高的精确度, 因此 SVM 算法在 Linux 防火墙中的应用是可行的.

关键词: SVM; 防火墙; 异常检测; Netfilter; Iptables

引用格式: 吕赵明,张颖江,周昕,陈琼.SVM 算法在 Linux 防火墙中的应用.计算机系统应用,2017,26(8):243–246. http://www.c-s-a.org.cn/1003-3254/5894.html

Application of SVM Algorithm in Linux Firewall

LV Zhao-Ming, ZHANG Ying-Jiang, ZHOU Xin, CHEN Qiong

(School of Computing, Hubei University of Technology, Wuhan 430068, China)

Abstract: Linux firewall provides a scalable mechanism for developers. After a thorough research of SVM principle, this paper, proposes the design and implementation of Linux firewall system based on SVM. The Netfilter framework is used to capture network packets. In the users' space, anomaly network traffic is classified by support vector machine algorithm module and the rules of Iptables are added dynamically. Thus, the function of defending network attacks is realized. The experimental results demonstrate that the proposed system model has high detection accuracy for the classification of abnormal traffic. It proves that the SVM algorithm is feasible and effective in Linux firewall.

Key words: SVM; firewall; abnormal detection; Netfilter; Iptables

在对异常流量的判别中, 常用的流量识别技术有对数据包内容的深度检测技术(DPI)和对深度数据流的检测技术(DFI)[1], 由于对数据包内容的检测技术是对数据报文内容进行特征匹配, 这种方法处理速度慢, 对硬件资源要求高, 特征库需要不断升级以适应新的数据报文特征, 并且无法检测加密的数据流. 另外 Linux 内核提供了可以扩展的 Netfilter 防火墙架构, 但在 Netfilter 框架下注册抵御各种攻击的模块对数据包内容进行识别, 装入内核的模块一旦使用不当会导致系统的崩溃, 并且链接进内核的模块会对整个系统的性能和内存造成损失[2,3]. 而基于数据流特性的检测技术是对流量进行统计, 以统计的流量特征作为对异常流量的识别依据, 能够有效监测已加密的数据, 识别速度快.

SVM(Support Vector Machine)是基于结构风险最小化原则的一种新兴的机器学习方法, 它对小样本、高维数、非线性和局部极小点等实际问题、具有很强的范化能力, 它解决了神经网络的过拟合、收敛速度慢、容易陷入局部极值等缺点[4-6]. 现实网络流中正常的流量远远大于异常流量, 所以选择支持向量机算法进行异常流量的分类能提高识别精度.

基于以上的分析, 本文提出了基于深度数据流检

① 基金项目: 教育部基金项目(NGII20150404)
收稿时间: 2016-11-21; 采用时间: 2016-12-26

测算法(支持向量机算法)的 Linux 防火墙系统的设计与实现. 使用机器学习算法, 能够提高识别精度, 利用 Netfilter 框架[7-9]进行数据包的捕获, 在用户态进行数据的处理, Netfilter 数据包过滤框架工作在内核, 能够实时高效的对流入防火墙的数据包进行捕获, 并通过 netlink 机制[10]实时的传入用户态, 这样保证了系统的性能.

1 系统架构设计与实现

本系统的架构设计是基于 SVM 算法在 Linux 系统的用户空间对网络的异常流量进行分类处理, 并由 iptables 防火墙进行防御的设计思想而设计的.

数据包的捕获基于 Linux3.10 内核框架 Netfilter, 使用动态模块加载技术, 使数据包捕获模块注册到 Netfilter 框架的钩子点上, 一旦数据包流经协议栈的注册函数, 由数据包捕获模块实时的捕获数据包, 通过 netlink 协议实时传输数据包到用户空间, 这样可以实现实时的数据流传输. 加之支持向量机算法针对小样本集具有分类精度高、泛化能力强等特点, 因此设计了用 Netfilter 框架捕获数据包和用支持向量机算法进行异常流量检测, 根据检测结果动态地生成 Iptables 规则脚本[11]并执行, Iptables 规则会自动加载到 Iptables 防火墙中, 最后由用户空间的 Iptables 防火墙实现过滤攻击的异常流量. 系统架构模型如图 1 所示.

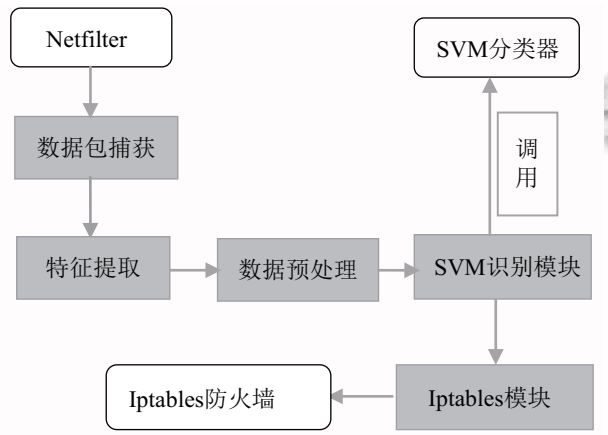

图 1 系统的整体架构模型

1.1 数据包捕获模块

数据包捕获模块是基于 Linux 内核框架 Netfilter 和 netlink 协议获得流量数据.

该模块通过动态可加载模块技术加载到 Linux 内核, 即在 Netfilter 框架的钩子点函数 NF_INET_PRE_ROUTING 和 NF_INET_LOCAL_OUT 处注册处理函数, 数据报文在进入系统后先经过 NF_INET_PRE_ ROUTING 挂载点, 并有处理函数捕获数据报文信息. 由本机发出的数据报文经过 NF_INET_LOCAL_OUT 挂载点的处理函数, 并捕获数据报文信息, 这样流经协议栈的所有数据包和由本机发出的所有网络数据包, 由自定义的数据报文结构体, 并通过 Netlink 通信机制[10], 把捕获到的网络数据包信息由内核态传到用户态, 待特征提取模块和数据处理模块处理.

1.2 特征提取模块

特征提取模块: 通过对 Linux 内核中网络协议栈的分析及对核心套接字缓冲区的结构体 sk_buff 的分析, 本模块通过捕获到的数据报文信息的首部结构体进行特征提取. 主要提取出网络连接持续的时间、协议类型、数据包长度、使用目标主机服务类型、网络连接状态、源主机发送到目标主机的数据量、目标主机发送到源主机的数据量, 源地址和目的地址等特征信息. 提取单个报文信息后, 依自定义的结构体为信息载体并传送到下一模块.

1.3 数据包预处理模块

数据包预处理模块: 该模块将特征提取模块提取到的数据包特征信息进行格式化和归一化处理, 从而得到支持向量机能够处理的数据格式.

原始数据的格式为: (0, tcp, http, SF, 181, 5450, …, normal). 该数据是异构数据集, 在进行运算前必须进行标准化处理: 处理非数值型的数据, 并且格式化为 <label><index>:<value>, 这种格式如下所示: (0 1:0 2:1 3:2 4:4 5:1816:5450…), normal 为正常数据这里用 1 来表示, 最后归一化到(-1, 1)之间, 生成样本集或数据集. 如图 2 所示.

图 2 标准化后的数据集

1.4 支持向量机的训练

支持向量机的训练模块是通过采集到的训练样本集对 SVM 进行训练, 经过训练后的支持向量机得到 SVM 分类器. SVM 训练模型如图 3 所示.

该训练模块是在离线状态下, 通过多台机器分别模拟正常流量和异常流量, 采用攻击工具 DDosPing、

幽魂 DDOS、Smurf2K 等对服务器进行攻击, 从而产生异常流量. 通过数据包捕获模块所捕获的源地址对正常流量和异常流量进行标记, 再经过数据预处理模块得到样本集, 用这些数据对 SVM 进行训练得到分类机.

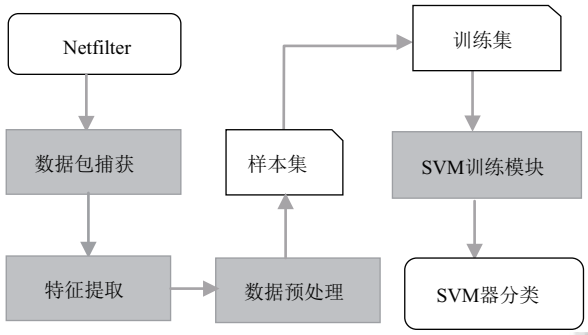

图 3 SVM 训练模型

在对 SVM 训练之前, 针对样本集的特性选择合适的核函数并对相关参数进行优化, 以使 SVM 分类器达到较高的准确率. 本文选择 SVM 模型类型为 C_SVC (支持向量分类机), 核函数为径向基核函数:

$$K(x,x_i)=\exp\left\{-\frac{\|x-x_i\|^2}{2\sigma^2}\right\} \quad (1)$$

C 表示惩罚系数, 即对误差的宽容度, 参数 C 的值越高, 越不能容忍错误的分类, 但容易出现过拟合的情况. 参数 C 的值越小越容易出现欠拟合的情况, 这样其范化能力较差. RBF 核函数将线性不可分的样本映射到一个更高维的空间, 使其线性可分, 与线性核函数不同的是它能够处理添加的类别标签和属性的非线性关系, RBF 核相对于多项式核有较少的超参数并且 RBF 核有更少的数值复杂度. 因此在特征维数有限的情况下, RBF 核能有较好的效果[12]. 对于 RBF 核函数, 有一个参数-g, 默认值是 1/k(k 是类别数).

为了优化重要参数 c 和 g, 本文通过网格搜索算法进行 c 和 g 的参数寻优, 网格搜索算法能够对各种可能的值进行交叉验证, 最终得到最优的 c 和 g:(c, g)=(0.03125, 0.03125), 即在交叉验证中对测试集识别率最高的一组参数. 根据最优的 c 和 g 参数来训练支持向量机, 得到 SVM 分类器, 这样得到的 SVM 分类器是判别准确率最高的.

1.5 SVM 识别模块

SVM 的识别模块: 该模块是整个防火墙系统的引擎. 网络流经过特征提取模块和数据预处理模块后传递给 SVM 识别模块, SVM 识别模块根据 SVM 训练模块得到的支持向量来预测分类正常流量和异常流量, 支持向量是 SVM 进行两类分类的关键, 因为支持向量决定着最优分类面. SVM 识别模块对实际流量预测后的结果是正常流为 1, 攻击流为 2, 一旦检测到 2, 即发生了攻击行为; 系统将该信息存储在结构体中并传递的 Iptables 模块中进行过滤处理.

1.6 Iptables 模块

Iptables 模块是根据 SVM 识别模块对网络流量的分类结果, 动态的添加并执行 Iptables 规则脚本, 从而实现网络攻击抵御的.

若有异常流量出现, SVM 分类器正确检测后, 系统调用 Iptables 模块把相关数据包信息传递该模块, 由该模块添加规则链到 Filter 表中并自动重启 Iptables 服务, 此时加进 Iptables 的规则立即生效, 并过滤检测到的异常流量.

2 仿真实验

2.1 实验准备

为了实验的充分性, 本实验使用 KDD 权威数据集验证 SVM 算法在防火墙中应用的有效性, 即验证离线状态下训练的 SVM 分类器的准确度. 该实验随机选取 KDD 数据集中正常流量和拒绝服务攻击流量构成仿真数据.

2.2 SVM 模型识别率的实验

本研究中的实验使用 Matlab2013a 环境和 libsvm[13] 工具包进行预测实验, 并在该环境中运用 BP 神经网络进行了对比试验.

实验使用从数据集文件中提取出的 2000 条数据作为训练集和三组数据作为测试集, 数量分别为 600、1000、3000, 每组数据集包括正常连接数据和攻击数据, 因为实际网络中正常流量远大于异常流量, 所以正常数据占 90%, 攻击数据占 10%. 使用训练得到的 SVM 分类器和训练后的 BP 神经网络, 分别对这些数据集进行判别. 实验使用的数据集如表 1 所示, SVM 分类器和 BPNN 预测结果如表 2 所示.

表 1 实验样本集

数据集	正常样本量	异常样本量
600	540	60
1000	900	100
3000	2700	300

表 2 SVM 分类器和 BPNN 预测结果

数据集	BPNN		SVM	
	精确度(%)	漏报率(%)	精确度(%)	漏报率(%)
600	100	0.0	100	0.0
1000	98.400	1.6	100	0.0
3000	99.867	0.13	99.867	0.13

2.3 实验结果及分析

通过选取不同数量的数据集, 分别使用 SVM 分类器和训练后的 BP 神经网络进行了预测实验, 从实验结果中可以看出, 600 的数据集在 SVM 算法和 BP 神经网络中 100% 识别, 数据集为 1000 时, SVM 算法 100% 识别, BP 神经网络识别率为 98.4%, 漏报率 1.6%, 当数据量达到 3000 时, SVM 算法和 BP 神经网络的准确率均为 99.8667%, 漏报率 0.13%. 同时, 在对 SVM 训练时, 选择大量的训练集进行训练, 有助于提高 SVM 的范化能力, 这样对于大量的网络数据能有很好的识别率. SVM 算法检测结果如图 4 所示.

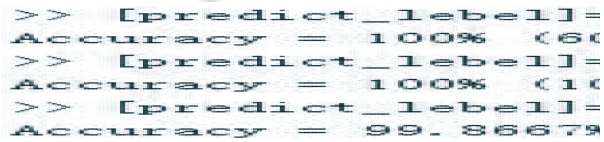

图 4 SVM 检测结果

因此, 利用 SVM 算法进行异常流量监测具有很好准确率, 另外内核级的 Netfilter 框架为网络流量的捕获提供了实时性, 所以 SVM 算法在 Linux 防火墙中应用能有效的检测攻击流量.

3 结语

目前研究较多的是基于支持向量机, 通过 Libpcap 或 Winpcap 等网络抓包工具实现网络异常流量的检测系统, 本研究中提出了支持向量机算法在 Linux 防火墙中应用的模型, 通过向 Netfilter 框架注册的数据包捕获模块来高效实时的抓取数据包, 在用户态使用训练集训练 SVM, 使用训练得到的 SVM 分类器检测异常网络流量, 根据检测的结果和主机地址生成 Iptables 脚本并执行, 由 Iptables 防火墙过滤异常流量, 从而实现了主动监测并及时防护的防火墙, 使受保护的网络环境更加安全.

在实验中, 我们是采用离线分析, 可以看到训练的 SVM 模型是固定的, 虽然使用小样本的数据对 SVM 进行训练预测, 可以得到很高的准确率, 但为了提高预测范围, 必须对 SVM 进行在线增量学习, 还需要进一步研究增量学习算法, 来提高训练效率和识别准确率.

参考文献

1 郭婷. 深度数据包和深度数据流检测技术研究[硕士学位论文]. 长春: 长春理工大学, 2013.
2 刘建峰, 潘军, 李祥和. Linux 防火墙内核中 Netfilter 和 Iptables 的分析. 微计算机信息, 2006, 22(1–3): 7–9.
3 李善平, 季江民, 伊康凯. 边干边学: LINUX 内核指导. 2 版. 杭州: 浙江大学出版社, 2002: 95–100.
4 谢雪莲, 杨海波. SVM 在网络流量异常检测中的应用研究. 计算机时代, 2012, (9): 14–16, 19.
5 Harrington P. 机器学习实战. 李锐, 李鹏, 曲亚东, 等译. 北京: 人民邮电出版社, 2013: 89–106.
6 蒋艳凰, 赵强利. 机器学习方法. 北京: 电子工业出版社, 2009: 163–179.
7 Netfilter 分析. http://www.cnblogs.com/iceocean/articles/1594196.html. [2016-09-25].
8 陈果. Linux 防火墙研究与设计[硕士学位论文]. 成都: 西南交通大学, 2003.
9 姚晓宇, 赵晨. Linux 内核防火墙 Netfilter 实现与应用研究. 计算机工程, 2003, 29(8): 112–113, 163.
10 吴小倩. 基于 Netfilter/Iptables 的网络流量监控系统的设计与实现[硕士学位论文]. 北京: 北京邮电大学, 2013.
11 姚亚锋, 蒋毅. 一种 Netfilter/Iptables 防火墙的实现研究. 计算机安全, 2013, (11): 19–22. [doi: 10.3969/j.issn.1671-0428.2013.11.005]
12 饶蓝. 基于支持向量机的网络攻击检测研究[硕士学位论文]. 南京: 南京理工大学, 2007.
13 LIBSVM. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.Html. [2016-09-25].