基于遗传神经网络分析的内网用户行为审计系统①

Internal Network Users' Behavior Auditing System Based on Genetic Neural Network Analysis

蔡家楣 蔡其星 江 颉 (浙江工业大学 软件学院 浙江 杭州 310023)

摘 要: 安全审计技术是继防火墙、入侵检测技术之后出现的能有效保障大规模网络安全的一种重要手段。目前，安全审计系统多是基于广域网网络数据的审计，对于内网用户行为审计甚少，而且还处于研究初级阶段。针对这一点，本文首先分析了用户行为审计要点，提出了一种基于用户行为的内网分布式安全审计系统，并将基于遗传算法的神经网络用于审计日志分析及异常检测。实验结果证明该模型能够有效地检测出用户异常行为。

关键词: 内网安全 行为审计 分布式 遗传算法 神经网络 异常检测

1 引言

随着互联网的迅猛发展，网络规模进一步扩大。网络体系结构的复杂性导致很多网络漏洞，网络安全问题已经成为信息安全领域日益关注的热点问题。防火墙、IDS 也随之迅猛发展起来，然而普通防火墙及入侵检测系统已无法防御多变的网络攻击手段。安全审计技术是近年来蓬勃发展起来的，并且被证明是能有效保障网络安全的重要手段，通过分析历史审计日志，可以发现网络中潜在的威胁，并且可以对网络攻击进行实时报警。目前，大部分网络安全审计系统审计对象都是外网的攻击行为及异常数据。而近年来网络攻击方式开始从外部攻击向内部攻击转化，资料显示[1]，在全球范围内损失 5 万美元以上的攻击中，有 70％是来自内部用户攻击。内部攻击危害性及其造成的损失比外部攻击更大，内部用户可以通过合法身份进行非法操作，轻易地绕过防火墙及 IDS 监控。本文从用户行为分析入手，提出一种基于内网用户行为的安全审计模型，并采用基于遗传的神经网络对审计数据进行分析，最后经过内网环境测试，证明该系统能有效实现用户行为审计及异常检测。

2 内网用户行为分析

内部用户的非法操作或恶意攻击一般都是非常掩蔽的，可以轻易躲避防火墙、IDS 监控，无法追踪。不同用户行为不甚相同，同一用户在不同时间行为也会有所变化。但是，对于具体用户的行为在总体上还会体现一定的个性化，有规律可寻。主机上用户行为异常可以归纳为[2]：

① 登陆异常：未经授权的用户远程登陆主机。登陆方式包括 FTP、Telnet、SSH 登陆等。

② 用户异常：主要体现在，系统出现未注册用户名；某用户密码为空；PASSWD 文档中，可执行 SHELL 的用户名单被改动；某用户在 WTMP 中有记录，却在 LASTLOG 中未记录等。

③ 用户操作异常：主要体现在，用户越权操作，普通用户使用 SU 命令切换到 ROOT；一个平时不活跃的用户突然启动，且连续很长时间占用了大量系统资源；普通用户频繁使用某些系统监视命令，如对某些端口监听；某用户连续多次试探使用其权限外的命令等。

④ 网络传输异常：某网络通讯端口在某段时间内传输数据量突然增加；用户频繁启动邮件软件收发邮件等。

用户行为主要特征是复杂多变，对于具体用户可体现个性化特征。比如系统管理员，系统命令及系统

① 基金项目:浙江省科技厅项目(2007C21008)
收稿时间:2008-08-12

管理软件使用会比较频繁，对于文员，文字处理软件及邮件收发操作等会相对比较频繁。因此，用户行为审计的第一步应该发现用户个性化特征，构造用户正常行为模式。过程主要包括多方面大量采集用户行为数据，选择最能体现用户个性化的数据，通过高效的分析技术来构造行为模式。目前，国内外对用户行为审计也做了一定研究，并且有多种系统已投入使用[3-4]。但是总体上讲，各种审计系统的瓶颈在于如何高效分析审计数据，得出合理的行为模式，提高异常检测准确率。本文针对这一技术难点做相应探讨，摒弃传统数据分析方法，并采用具有自学习能力的神经网络来实现数据分析。

3　内网安全审计系统框架设计

根据是否有控制中心，可以将安全审计系统体系结构分为：有控制中心安全审计系统及无控制中心安全审计系统[5]。前者原理是分布式数据采集，集中式数据处理，适合在内网环境中部署；后者原理是分布式数据采集，分布式数据协同分析，适合大规模跨自治域的动态网络环境。目前，前者技术发展相对成熟。由于本文研究重点在于内网用户行为审计，故采取具有控制中心的体系结构来实现系统比较合适。本系统总体结构如图 1 所示。

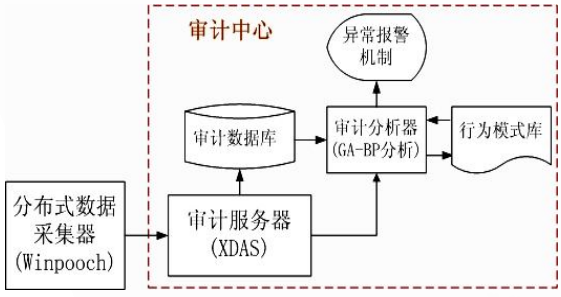

图 1　系统结构图

3.1 审计数据采集

数据采集是实现审计系统的第一步。网络通讯数据、系统安全日志都可以作为审计系统日志。根据不同需求，可以采集不同类型数据。本系统主要针对用户行为审计，因此审计数据选择用户操作行为日志。采用 hook 机制采集用户行为数据是比较理想的方法。Windows hook 的本质就是用于处理消息的一段程序。通过系统调用，将它挂入某个进程，当有特定事件的消息发出时，它就可以在消息到达目的窗口进行响应前捕获消息，得到对此消息的控制权，甚至可以强制结束消息的传送。hook 机制占用系统资源低，可以监视并捕获用户每个操作的消息，达到用户行为数据采集的目的。

本文数据采集模块基于 Winpooch 的 Hook API 技术实现。对 Winpooch 源码重写并编译，实现了监听包括文件读写，注册表修改，端口连接及访问等多种事件的 hook；修改过滤器规则，可以实现任何数据采集需求；扩展 Winpooch 日志内容，符合 XDAS 日志格式标准，并保存于 XML 文件，便于程序解析及网络传输。

3.2 审计中心

审计中心是安全审计系统的核心部分，负责接收分布式数据采集点审计日志数据，分析审计数据，对异常进行报警、并将数据存储于数据库中。因此，审计中心必须具有高效率、高稳定性的特点。审计中心又可以分为审计服务器及审计分析器两大模块。

本文选择 Open Group[6]XDAS（Distributed Audit Service）作为审计服务器。XDAS 是一个非常优秀的开源工程，基于 XDAS 开放接口，并针对审计服务器功能，本文开发了审计日志提交接口、外部审计数据源导入接口、事件过滤器管理接口、事件询问接口、日志管理接口。

审计日志提交接口：通过该接口，可以开启网络监听端口，接收应用系统或者服务平台中被采集到审计日志；

外部审计数据源导入接口：提供外部审计数据集导入功能，这些数据集可以来自某些领域审计服务器采集的大量数据，日志分析模块可以通过此类数据对审计系统进行训练；

事件过滤器管理接口：此类接口可以对事件过滤器作相应配置，包括定义、开启、关闭、注销过滤器等。事件过滤器在审计系统中起到日志筛选功能，可以过滤无效日志、提高分析效率。

事件询问接口：查看审计事件数据流是否有新数据进来；

日志管理接口：对日志的各种管理，包括配置、统计、查看历史审计记录等。

此外，根据 XDAS 标准，定义了一组安全审计中频发事件类型，包括帐号事件、会话事件、资源访问事件、服务及应用程序事件等，并且定义了日志格式标准，日志总共包括 24 个字段，能够详细表达事件含义，使得日志分析时更加准确。

4　基于遗传算法的用户行为审计分析

审计分析模块是审计中心重要组成部分。完成海量历史审计数据分析，新事件数据实时分析，判断其为正

常事件、可疑事件或者异常事件，并给出结果。传统审计分析方法有：基于规则匹配的方法、基于数理统计的方法、基于数据挖掘的方法。前两种方法不具自学习能力，必须手动更新规则库，效率低下；基于数据挖掘的方法虽然能在攻击手段变化的情况下可以自动更新规则库，识别异常能力有所提高，但要达到较理想效果，必须将关联规则、序列方法等多方法相结合，过程繁琐复杂，在实现上存在一定技术局限性。

本文采用基于经过改进[7] 的遗传算法（GA）的 BP 神经网络来实现审计数据分析及异常检测，可以解决传统方法中出现的问题。将遗传算法与神经网络相结合，利用遗传算法较强的宏观搜索能力及全局寻优能力[8]来弥补神经网络算法易陷入局部极小、收敛速度慢和引起振荡效应[9]等缺陷。对 GA 的改进主要由以下几个方面：

①对交叉概率和变异概率进行自适应调整，公式（1），（2）所示。

$$P_c=\begin{cases}P_{c1}-\dfrac{(P_{c1}-P_{c2})(f'-f_{avg})}{f_{\max}-f_{avg}} & f'\geq f_{avg}\\ P_{c1} & f'\leq f_{avg}\end{cases} \quad (1)$$

$$P_m=\begin{cases}P_{m1}-\dfrac{(P_{m1}-P_{m2})(f-f_{avg})}{f_{\max}-f_{avg}} & f\geq f_{avg}\\ P_{m1} & f\leq f_{avg}\end{cases} \quad (2)$$

其中：P_{c1} 和 P_{m1} 为适应度低于种群平均适应值的个体的交叉率和变异率；P_{c2} 和 P_{m2} 分别为每代群体中最优个体的交叉率和变异率；$f_{\max}$ 和 f_{ave} 分别为每代群体中最大适应值和平均适应值；f' 为交叉运算前父代双亲中较大的适应值；f 为变异个体的适应值。经过调整，避免了由于个体彼此接近，选择概率相当所造成的早熟性收敛；

②最优化保存策略，增强算法稳定性及收敛性，能及时保存优秀个体；

③采用自适应正态变异算子，将变异范围与变异个体相联系，实现适应值较大个体在较小范围内变异，适应值较小个体在较大的范围内变异，从而提高区域搜索能力。

以上几种改进措施，最终加快了 GA 收敛到全局最优解的速度。

GA 与 BP 神经网络相结合，经过 GA 遗传优化，得出最优权值作为神经网络训练的初始权值，适用于审计系统数据分析及异常检测。BP 神经网络的自学习能力可以定期更新用户行为模式。

该方法应用于审计系统数据分析时，首先必须利用大量较规则的用户行为数据作为训练数据样本，当训练达到一定误差要求，得出最优解，即可作为用户行为模式。接下来就可以将用户实时行为数据作为输入，并根据输出结果判断当前行为正常与否。工作原理如图 2 所示。

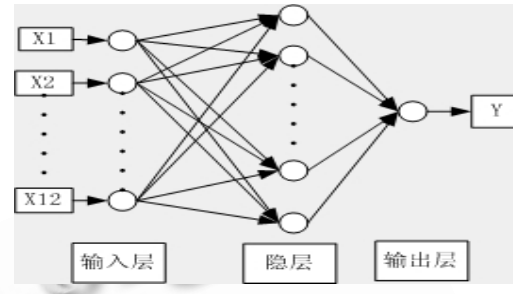

图 2　基于 BP 神经网络的异常检测模型

在该模型中，输入层为 12 个节点，输出层为 1 个节点，隐层节点根据经验公式 $n=\sqrt{n_0+n_1}+a$ [10]来确定，其中，n 为隐层节点数；n0 为输入层节点数；n1 为输出层节点数；a 为 1～10 之间整数。经过实验得出，a 取 8 时，训练效果最佳。

利用公式 $Logsig(x)=1/(1+\exp(x))$ 计算输出层结果，$Logsig$ 输出值范围是 0～1；并将 0－1 之间划分为两个区间作为判别门限值，接近 0 的为正常值，即表示当前行为为正常行为，接近 1 的为异常值，即当前行为是异常行为。

5　系统实现和实验结果

本项目已经取得的研究成果显示，利用 KDD Cup 1999 数据集作为实验数据，能够达到一个比较理想的效果。从中选取 1000 条训练数据，用 Matlab 对训练过程进行仿真，将性能误差设置为 0.001，总共经过 22 次训练达到目标性能。性能曲线如图 3 所示。

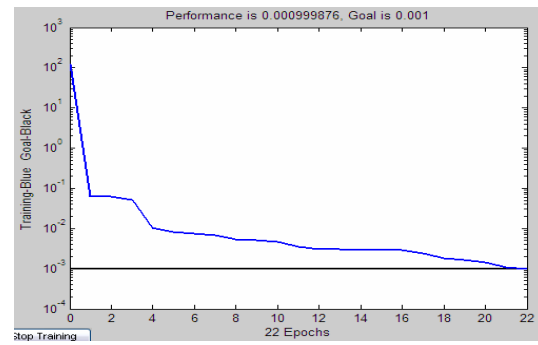

图 3　GA_BP 训练性能曲线

可见，经过 GA 优化的 BP 神经网络训练收敛速度比较理想。最后选取 500 条作为测试数据。结果正确数 466 条，准确率达 93.2%。参照这个实验过程及步骤，我们可以在现实网络环境中进行测试。

5.1 环境搭建

根据系统结构图，实现系统整合。XDAS 通过 7629 端口接收 Winpooch 采集的行为日志数据，并将数据保存于 MYSQL 数据库，GA-BP 模块通过读取数据库及实时接收 XDAS 数据，训练结果及异常检测结果保存于文本文件。

本实验共采用 6 台互联的计算机。审计服务器部署于 1 台计算机，作为审计引擎；Winpooch 部署于另外 5 台计算机，作为被审计对象，这样符合分布式数据采集的要求；操作系统采用 Windows XP。实验总体上分为两步：

①配置 Winpooch 过滤规则，采集较理想的用户具体行为数据来训练神经网络。

②构造异常行为，生成异常行为数据，并输入到神经网络进行判断，观察异常输出结果。

5.2 事件类型定义

为了构造用户具体行为，在测试过程中定义了 5 种事件类型：文件读写、网络端口连接、注册表修改、应用程序执行、进程终止。并为每种事件类型定义了具体事件，如表 1 所示。

表 1　具体事件名称

事件名称	事件对象	事件代号
File::Read	C:\windows	1
	C:\windows\system32	2
File::Write	C:\Documents and Settings\All Users\	3
	C:\WINDOWS*.exe	4
Net::Connect	172.16.7.16:80	5
Net::Listen	80 port	6
Reg::SetValue	*\Software\Microsoft\Windows\CurrentVersion\Run*	7
Sys::Execute	Notepad.exe	8
	Explorer.exe	9
Sys::KillProcess	*	10

5.3 数据采集

根据表 1，定义 Winpooch 事件过滤器采集相应事件。启动审计引擎，这时 XDAS 打开 7629 端口监听，随时接收从 Winpooch 传输过来的数据。在被审计计算机上运行 Winpooch，用某用户名登陆，并根据定义事件做相应动作。经过一个星期持续数据采集。

5.4 数据分析

本文采用正常数据训练方式训练神经网络。步骤如下：

①数据清理。在大量数据中，选出 2000 条较理想的用户正常行为数据作为神经网络训练数据集，数据集中包括用户各种操作。②数据特征提取。每个事件记录包含 6 个属性集（基本属性、事件源、事件发起者、事件目标、其他属性、事件内容），共 26 个属性。由于神经网络输入结点为 12 个，故从 26 个属性中选取包括事件 ID 号（eventnumber）、时间戳（eventtimestamp）、用户名（initiatorauthenticationauthority）、事件主体(initiatordomainspecificname)、事件客体(targetlocationname)等 12 较重要的关键信息字段作为数据特征。③数据格式转换。对各正常事件 12 个字段不同内容用 1-10 数值进行编码，事件名称根据表 1 转化为事件 ID 号，其他字段同理。④数值归一化。对这些数值进行归一化[11]处理，并将结果作为神经网络的输入，对神经网络进行训练。从实验结果看，总共经过 31 次训练，达到目标性能。

5.5 异常检测

为了验证训练结果有效性，必须做异常检测测试。在实验中，指定字段值不在正常事件编号内的陌生事件来构造异常事件数据 300 条。其中包括系统文件非法操作、连接未知端口、访问未知网页、修改注册表、进程强制结束等异常。将此类事件输入到神经网络进行分析，根据输出结果（0－1）进行判断，输出值在 0 的判别门限值范围内，被认为是正常行为；输出值在 1 的判别门限值范围内，被认为是异常行为。结果显示，异常检出率为 86.7%。

总体上，通过训练的神经网络对异常行为数据具有较好敏感性。但是比起 KDD99 数据集的实验结果，准确率稍低，这是神经网络对训练数据具有较高要求所导致的，而且判别门限值的选择对检测结果也有较大影响，门限过高，误报率大；门限过低，漏报率大。

6　总结

本文实现了以 XDAS 为审计中心的分布式内网用户行为审计系统。并将基于遗传算法的神经网络作为数据分析工具，经过遗传算法预处理，神经网络训练

(下转第 34 页)

(上接第 8 页)

收敛速率大大提高。最后经过应用环境实验证明该行为审计系统是可行的，在数据采集、数据分析、异常检测方面都具有较高的效率。由于实验是在较理想，范围较小的网络环境下进行的，而在更大范围、更复杂网络中的性能表现还有待验证，并且神经网络训练效果也有待提高，这将作为下一步研究内容。

参考文献

1 张春江,倪健民.国家信息安全报告.北京:人民出版社,2000.

2 江伟,陈龙,王国胤.用户行为异常检测在安全审计系统中的应用.计算机应用,2006,26(7):1637－1639.

3 张世永.信息安全审计技术的发展和应用.电信科学,2003,12:125－128.

4 天融信 TA.强力打造安全审计管理平台.信息安全与通信保密,2003,9:32－34.

5 张浩亮,刘利军.一种分布式安全审计模型研究与系统设计.计算机安全,2007,3:25－28.

6 OpenGroup,http://www.opengroup.org.

7 孙霞,黄席樾,杨祖元,向长城.基于改进遗传算法的城市交通动态最优路径求解.计算机工程与应用,2007,30:356－362.

8 李华昌,谢淑兰,易忠胜.遗传算法的原理与应用.矿冶,2005,14(1):87－90.

9 李家春,李之堂.神经模糊入侵检测系统的研究.计算机工程与应用,2001,(12):37－39.

10 周开利,康耀红.神经网络模型及其 MATLAB 仿真程序设计.北京:清华大学出版社,2004:70－100.

11 易晓梅,于芹芬,刘丽娟.一种基于神经网络的安全审计系统模型.福建电脑,2008,(2):99－101.