数据挖掘在入侵检测中的应用

苏辉贵[1],傅秀芬[1],钟 洪[2],苏辉财[3],韩 韬[1]
(1.广东工业大学 计算机学院,广东 广州 510075;
2.赣南师范学院,江西 赣州 341000;3.江铜集团德兴铜矿,江西 德兴 334224)

摘 要:入侵检测是用于检测任何损害或企图损害系统的保密性、完整性或可用性行为的一种网络安全技术。指出当前入侵检测系统存在的问题,并针对现有入侵检测系统漏报、误报率高的问题,提出将数据挖掘技术应用于入侵检测系统。文中论述了常用的数据挖掘算法,提出一个基于数据挖掘技术入侵检测系统模型、描述了模型体系结构及主要功能。实验表明,该模型能提取特征,生成新规则,找到入侵数据,提高入侵检测系统的有效性。

关键词:入侵检测;数据挖掘;规则提取;模型
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2006)10-0143-02

Data Mining Used in Intrusion Detection

SU Hui-gui[1], FU Xiu-fen[1], ZHONG Hong[2], SU Hui-cai[3], HAN Tao[1]
(1.School of Computer, Guangdong University of Technology, Guangzhou 510075, China;
2.Gannan Teachers' College, Ganzhou 341000, China;
3.Dexing Copper Mine, Jiangxi Copper Corporation, Dexing 334224, China)

Abstract: Intrusion detection is a network security technology used to detect the attempt of destroying system secrecy, integrality and usability. The problems of intrusion detection system are described. To solve the problems of intrusion detection system, data mining approach is used. The common used data mining algorithms are described, and an intrusion detection system based on data mining is proposed. Its system architecture and main function are discussed. Our experiment indicates that the model can produce new rules, find intrusion data and increase validity of intrusion detection system.

Key words: intrusion detection; data mining; rule extraction; model

0 引 言

目前解决网络安全采取的主要技术手段有防火墙、安全路由器、身份认证系统等,这些安全产品大多数属于静态安全技术的范畴。静态安全技术对防止系统非法入侵起到了一定的作用,但从安全管理角度来说,仅有防御是不够的,还应采用动态策略。入侵检测(Intrusion Detection)技术就是这样一种动态策略,它能够对网络安全实施实时监控、攻击与反攻击等动态保护。因此对入侵检测技术的研究具有很强的现实性和紧迫性。

1 入侵检测系统模型及存在问题

1.1 入侵检测系统模型

入侵检测是用于检测任何损害或企图损害系统的保密性、完整性或可用性行为的一种网络安全技术[1]。它通过监视受保护系统的状态和活动,采用误用检测(Misuse Detection)或异常检测(Anomaly Detection)的方式,发现非授权的或恶意的系统及网络行为,为防范入侵行为提供有效的手段。入侵检测系统是执行入侵检测工作的硬件或软件的产品,通过实时分析,检测特定的攻击模式、系统配置、系统或程序漏洞以及系统或用户的行为模式,监控与安全有关的活动。入侵检测的基本原理如图1所示。

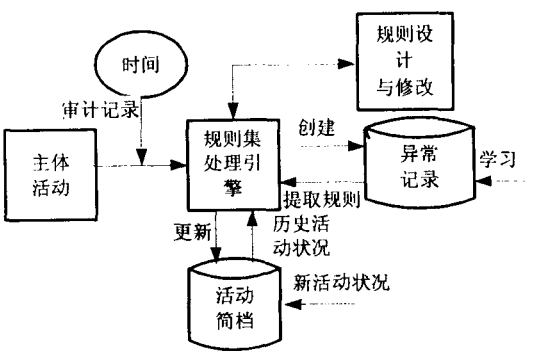

图1 通用入侵检测系统模型

1.2 入侵检测系统存在的问题

入侵检测系统的有效性、可适应性和可扩展性是评价入侵检测系统质量的重要指标。由于当前的入侵检测系

收稿日期:2006-02-19
作者简介:苏辉贵(1983-),男,江西鹰潭人,硕士研究生,研究方向为网络安全、数据挖掘、协同软件;傅秀芬,教授,硕士生导师,研究方向为网络安全、数据挖掘、协同软件等。

统通常是采用统计分析的方法对已知的入侵方法和系统脆弱性进行分析,然后根据"专家的知识"手工编写相应的规则,并且是针对具体的系统环境和检测方法的,这使得系统的有效性很差,而且系统在可扩展性、自适应性方面也极其有限。

针对当前入侵检测系统的缺陷,文中用以数据为中心的观点,将数据挖掘技术应用于入侵检测系统,提出一个基于数据挖掘技术的入侵检测系统模型。

2 数据挖掘技术

2.1 数据挖掘定义

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[2]。随着信息技术的高速发展,人们积累的数据量急剧增长,动辄以 TB 计,如何从海量的数据中提取有用的知识成为当务之急。数据挖掘就是为顺应这种需要而发展起来的数据处理技术,是知识发现的关键步骤。

2.2 将数据挖掘应用到入侵检测

引入数据挖掘技术[3],应用于入侵检测系统中,完成从大量数据中自动提取出模型的过程。将数据挖掘应用于入侵检测技术中,在建立攻击检测系统过程中消除人为因素和特定因素,为其开发一个更加系统化的方法。即开发一套能从各种审计数据中产生攻击检测模型的自动工具。应用关联分析和序列模式分析等算法,发现特征之间的关联和与时序有关的联系,从而完成用户数据收集与特征选择过程。

3 入侵检测系统中常用的数据挖掘算法

入侵检测系统中常用的数据挖掘算法有 4 种:

(1)关联分析算法。关联规则是表示数据库中一组对象之间某种关联关系的规则。在数据库的知识发现中,关联规则就是描述这种在一个事务中物品之间同时出现的规律的知识模式。更确切地说,关联规则通过量化的数字描述物品 A 的出现对物品 B 的出现有多大的影响。

(2)序列分析算法。关联分析是发掘数据记录中不同数据项之间的横向关联性,而序列分析则是发现不同数据记录之间的纵向相关性。序列分析的目标是在事务数据库中发掘出序列模式(large sequences),即满足用户指定的最小支持度(mininum support)要求的大序列,并且该序列模式必须是最高序列(maximal sequence)。

(3)分类算法。数据分类实际上就是从数据库对象中发现共性,并将数据对象分成不同类的一个过程。分类的目标首先是对训练数据进行分析,使用数据的某些特征属性,给出每个类的准确描述(即分类规则),然后使用这些描述,对数据库中的其它数据进行分类。

(4)聚类算法。将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程称为聚类。对象根据最大化类内的相似性和最小化类间的相似性的原则进行聚类或分组。所形成的每个簇可以作为一个对象类,由它可以导出规则,在许多应用中,可以将一个簇中的数据对象作为一个整体来对待。聚类与分类不同,聚类分析的输入数据集是一组未标记的对象,也就是说此时输入的对象还没有进行任何分类,聚类的目的是根据一定的规则,合理地进行分组或聚类,并用显式或隐式的方法描述不同的类别。

4 基于数据挖掘的智能入侵检测系统

基于数据挖掘的智能入侵检测系统系统结构包括组件有:活动监测 Agent、协同 Agent 审计数据库、数据挖掘引擎、特征提取器、数据挖掘引擎、规则库、数据检测引擎和决策响应中心等[2,4],如图 2 所示。

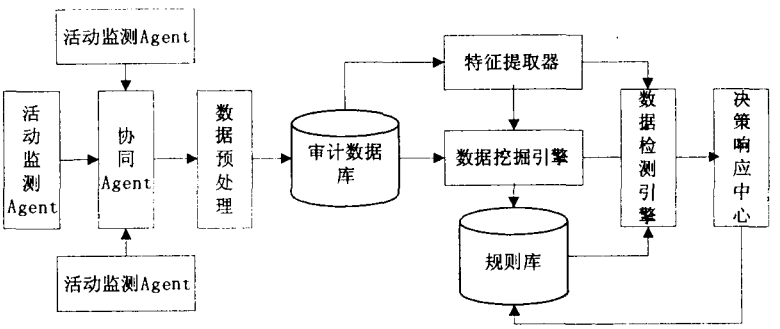

图 2 自适应数据挖掘入侵检测系统结构

系统的工作原理如下:

①活动监测 Agent 采集来自外部网络环境的各种数据,并把这些数据传送给协同 Agent,协同 Agent 对从活动监测 Agent 来的数据进行过滤、格式转换等预处理,然后将数据存入审计数据库中;

②审计数据库组件存储数据,并利用数据库查询技术产生训练数据集,同时特征提取器采用数据挖掘技术对当前用户行为进行分析,从中提取出当前用户行为特征;

③数据挖掘引擎利用数据挖掘技术对审计数据库中的数据和提取出的特征进行学习,从中提取出有关行为特征和规则,建立异常模式和正常行为轮廓,从而建立检测模型,并存入规则库中;

④数据检测引擎接受来自特征提取器的数据特征和数据挖掘引擎建造的模型与来自规则库中的正常规则进行分析,将分析结果送给决策中心;

⑤决策中心判断分析结果以决定是报警或是更新规则库。

5 运行结果及分析

SYN Flood 是当前最流行的 DoS(拒绝服务攻击)与 DDoS(分布式拒绝服务攻击)的方式之一,这是一种利用

(下转第 148 页)

后的水印与原水印图像作比较，相关系数为 1，可以判断两者是完全一样的，这就证明了该算法的可行性。

为了测试算法对局部窜改的监督能力，文中把含有水印信息的图像的像素点 $I(234,234) = 61.01$，$I(234,235) = 56.001$ 的像素值都改为 90。结果提取出来的水印如图 7 所示，与原水印的相关性为 $S = 0.8773$。根据相关性公式 $S = \frac{\sum_{i}^{q}\sum_{j}^{p}(I_{ij} \times I'_{ij})}{\sqrt{\sum_{i}^{q}\sum_{j}^{p}I'^{2}_{ij}} \times \sqrt{\sum_{i}^{q}\sum_{j}^{p}I^{2}_{ij}}}$，$I$ 为水印原图，I' 为提取后水印图，为了报告图像是否被篡改，S 属于[0,1]，如果 $S < 1$ 则表明图像被篡改，并且 S 越小，篡改越严重。由此看出任意做微小的篡改都会导致提取失败。

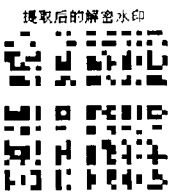

图 6 密钥有误时提取的水印

图 7 修改了像素后提取的水印

3 结 论

文中以一种 DCT 变换域的嵌入算法为基础，不采用传统分块嵌入水印的加强水印鲁棒性的方法。利用离散余弦变换中空域任何一点的变化都会对频域造成影响的特性，采用随机间隔法进行水印嵌入来实现脆弱水印，使得微小的篡改都会被监测到。另外，还利用超混沌序列对水印进行加密处理，增强了加水印图像的保密性。有必要时还可以结合嵌入时的随机数种子作为双密钥来达到加强保密性的目的。

参考文献：

[1] 张小华，孟红云，刘 芳. 一类有效的脆弱型数字水印技术[J]. 电子学报，2004(1)：114－117.

[2] 许红山. 基于变换域的数字水印技术[J]. 计算机工程与科学，2004，26(1)：47－49.

[3] 张 勇，赵东宁，李德毅. 数字水印技术及进展[J]. 解放军理工大学学报(自然科学版)，2003，4(3)：1－5.

[4] 王丽娜，郭 迟，李 鹏. 信息隐藏技术实验教程[M]. 武汉：武汉大学出版社，2004.

[5] 程 丽，陶 路，黄秋楠，等. 构造具有超混沌特性的二维离散系统[J]. 东北师大学报(自然科学版)，2002，34(3)：47－52.

[6] 李雄军，彭建华，徐 宁，等. 基于二维超混沌序列的图像加密算法[J]. 中国图形图像学报，2003(10)：1173－1176.

[7] 乌 旭，陈尔东，胡家升. 一种基于混沌的图像加密改进方法[J]. 大连理工大学学报，2004，44(5)：754－757.

(上接第 144 页)

TCP 协议缺陷，发送大量伪造的 TCP 连接请求，从而使得被攻击方资源耗尽的攻击方式。

在网络仿真环境中使用 SYN Flood 工具，模拟小型网络中 SYN Flood 攻击。应用关联规则和频繁事件算法，把 service 作为轴属性，dst_port 作为参考属性，得到一组关于相同目的主机的频繁时序“服务”模式[5]。得到的特征模式如下：

模式 1：(service = telnet, src_host = $host_A$)→(flag = SF), [0.79, 0.23, 20s]

模式 2：(service = http, flag = S0), (service = http, flag = S0)→(service = http, flag = S0), [0.79, 0.23, 20s]

把这些模式与正常模式集相比较，在正常模式集中不存在 flag = S0 的模式。所以模式 2 可以作为攻击模式。对攻击模式进行特征建立与提取，可以得到基于统计的特征。例如：在 20s 内，相同目的主机的连接数中有相同服务连接的百分比及状态 flag 为 S0 百分比。并得到以下属性：Count，Serror_%，Same_srv_%，Diff_srv_%，Srv_count。同时，在 snort 的官方网站上查找 SYN Flood 的规则文件，对两者规则进行比较，把数据挖掘发现的新规则加入规则文件。通过比较，发现新规则的加入降低了 SYN Flood 攻击的误报率。

6 结束语

构建了基于数据挖掘的入侵检测系统结构模型。该系统被赋予规则发现、辨识和扩展功能，能够检测已知和未知的攻击活动。下一步的研究目标是：对挖掘算法进行改进，减小对海量数据的计算量，有效地提取特征，提高入侵检测系统的有效性。

参考文献：

[1] Yu J, Reddy Y, Selliah S, et al. TRINETR: An architecture for collaborative intrusion detection and knowledge－based alert evaluation[J]. Advanced Engineering Informatics, 2005, 19:93－101.

[2] 洪飞龙，范俊波，贺 达. 数据挖掘在入侵检测系统中的应用研究[J]. 计算机应用，2004，24(12)：82－83.

[3] Ganger G R, Nagle D F. Better security via smarter devices [A]. Hot Topics in Operating Systems[C]. [s. l.]: IEEE, 2001. 100－105.

[4] Lee W, Stolfo S J. A Framework for Constructing Features and Models for Intrusion Detection Systems[J]. ACM Transactions on Information and System Security, 2000, 3(4): 227－261.

[5] 马恒太. 基于 Agent 分布式入侵检测系统模型的建模及实践[D]. 北京：中国科学院，2000.