第16卷 第1期
2006年1月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 1
Jan. 2006

基于数据挖掘和协议分析的可扩充IDS架构

王亚楠，刘方爱
（山东师范大学 信息管理学院，山东 济南 250014）

摘 要：由于TCP/IP协议的开放性，目前的网络极易受到攻击。文中详细介绍了入侵检测系统的主要思想和技术分类，通过比较不同类型入侵检测系统的优缺点，分析了应用于入侵监测系统的数据挖掘和协议分析技术，并在此基础上提出了一种新的基于安全管理的混合式可扩充入侵检测架构。该构架分层、简单、灵活，具有良好的扩充性。理论分析表明，该架构不仅能提高入侵检测的准确率，而且能提升系统效率，有很好的应用前景。

关键词：入侵检测；数据挖掘；协议分析
中图分类号：TP393.08　　**文献标识码**：A　　**文章编号**：1005－3751(2006)01－0223－03

An Extensible Framework of Intrusion Detection System Based on Data Mining and Protocol Analysis

WANG Ya-nan, LIU Fang-ai
(College of Information Management, Shandong Normal University, Jinan 250014, China)

Abstract: Because of the open structure of TCP/IP, the current network is vulnerable. Introduces the main thinking and technical classification. It presents a new mixed model for the intrusion detection system based on data mining and protocol analysis by analyzing the relative merits of the two kinds of IDS. The extensible intrusion detection framework is layering, simple, flexible and theoretical analysis indicates that it can improve not only the rate of accuracy but also the efficiency of the IDS, so it has a better application.

Key words: intrusion detection; data mining; protocol analysis

0 引 言

入侵检测系统（Intrusion Detection System，IDS）是在1980年由Anderson首先提出的[1]，是计算机信息安全领域中一个热点课题。到目前为止可分为4代，第一代IDS技术基于简单模式匹配技术，特点是防御力差，微小攻击变形都无法检测，故漏报率高。第二代IDS技术出现在20世纪90年代中期，技术突破包括网络数据包截获、主机网络数据和审计数据分析、NIDS和HIDS的明确分工和合作，但其所使用的检测技术都是模式匹配。第三代IDS技术是基于协议分析和模式匹配以及异常统计，它的优点是误报率、漏报率较低，效率高，但管理功能薄弱。第四代IDS技术是基于安全管理和协议分析和模式匹配以及异常统计，它的优点是入侵管理和多项技术协同工作，建立全局的主动保障体系，以它为核心，可以构造一个积极的动态防御体系。目前大多数入侵检测系统都是基于Denning[2]在1987年提出的模型而建立的，一个入侵检测系统至少应包含3个必要功能组件：数据获取、分析引擎和响应组件。入侵检测系统按照数据来源分为主机型IDS（Host－based IDS，HIDS）和网络型IDS（Network－based IDS，NIDS）两种，入侵检测的分析技术主要分为特征（误用）入侵检测和异常入侵检测。基于主机的IDS主要检查的是计算机访问了哪些文件，执行了哪些应用程序这样的事件，而基于网络的IDS则检查计算机之间交换的信息即网络通信量。由于这两种入侵检测系统各有自己的优缺点，在很大程度上互补，故综合二者优势能大幅度提升攻击抵抗力的混合模式IDS已成为研究热点。特征入侵检测是利用已知的入侵特征模式识别非法入侵，最常用的方法是模式匹配，即将收集到的信息与已知的网络入侵模式相比较，从而发现违背安全策略的行为。这种检测方法检测效率较高。异常入侵检测是通过检查当前用户行为是否与已建立的正常行为轮廓相背离来鉴别是否有非法入侵或越权操作。这种检测方法使得IDS适应性比较强，但是缺点是容易出现误报。异常入侵检测较常使用的有神经网络方法和概率统计方法等。但是对于神经网络方法来说它的网络拓扑结构较难确定；而概率统计方法则在定义入侵阈值上有困难。

特征入侵检测使用的匹配方法在计算机上很容易实现，因而特征检测成为入侵检测系统的基本实现手段，但

收稿日期：2005－04－25
基金项目：国家自然科学基金资助项目（60373063）
作者简介：王亚楠（1982—），女，山东临清人，硕士研究生，研究方向为并行计算及网络安全；刘方爱，博士，教授，博导，研究方向为并行处理、并行计算模型和互联网络。

是它有两个较大的缺陷：

1)准确性较差。模式匹配技术使用固定已知的特征模式来探测攻击，只能探测出明确的、唯一的攻击特征，一旦出现任何未知形式的入侵甚至是已知攻击的微小变形都无法检测出来。

2)效率不高。模式匹配算法[3]所需的计算量巨大，对于满负载的 100Mbps 以太网，其每秒比对次数＝网络中每秒数据包数量×数据包字节数×攻击特征字节数×特征库中攻击特征数量。假设每秒中网络流量为 10000 个数据包，一个数据包有 100byte，特征库中有 1000 个攻击特征，而攻击特征字节数平均为 30byte，则每秒比对数为 3×10^{10}次，即多达 300 亿次。如此大的运算量一旦运算能力不足时，入侵检测系统就会丢弃一些数据包，从而出现漏报。

为了解决上述两个问题，引入了数据挖掘和协议分析技术。文献[4,5]提出了利用数据挖掘技术来挖掘入侵特征，解决了准确性差的问题，但并没有考虑效率问题，不能很好地适应现有的高速网络，从而造成大量漏报。文中综合了两种技术的优势，提出了一个基于数据挖掘和协议分析技术的、可整合多种不同类别 IDS 到单一系统的可扩展架构。这种架构包含一个控制平台，负责管理从不同类型 IDS 收集到的数据，并及时把分析后的结果发送到各 IDS，从而达到信息共享，使得 IDS 更加智能化。

1 系统架构

该架构(见图 1)可把不同类别的入侵检测系统(如：主机型的，网络型的等)整合到一个混合模式的单一系统中，架构中定义了通讯协议，并且这个系统还能作相应扩展。这样一来，入侵检测系统便能共享大量有意义的信息，从而更有效、更灵活。在这里网络型的 IDS 采用的是协议分析的技术来检测网络数据包，而主机型的 IDS 采用的是特征入侵检测这种检测技术。

IDS 与中央信息控制平台(CICP)、数据分析及定义模块(DAD)之间相互通讯并交换数据。每个 IDS 中都有报警系统 alarm，负责定义 IDS 检测的预警信息，并发送到 CICP 与其它构件共享。DAD 共享来自 alarm 所属的 IDS 中的数据后，分析并定义 CICP 所处理的信息，并把结果送到每个 IDS。CICP 是个控制平台，负责收集来自所有 IDS 的 alarm。CICP 根据自己的功能，如数据库技术、数据挖掘、人工智能、知识库、关联分析等先进技术，不断将信息发送到每一个 IDS，帮助 IDS 改进检测准确率及效率。

在 IDS 模块中，使用数据挖掘技术来不断挖掘新的入侵特征，这种技术可以有效地重构和扩充已有的入侵特征库，增强系统对入侵变体的识别能力。数据包在 IDS 中要与入侵规则库中的入侵规则相匹配，如果匹配成功，则说明有入侵发生，启动报警系统(alarm)并反馈给发掘规则模块。但是为了减少网络数据包中数据和入侵规则的匹配量，可以先采用协议分析的方法分析数据，然后对可疑数据进行规则匹配。

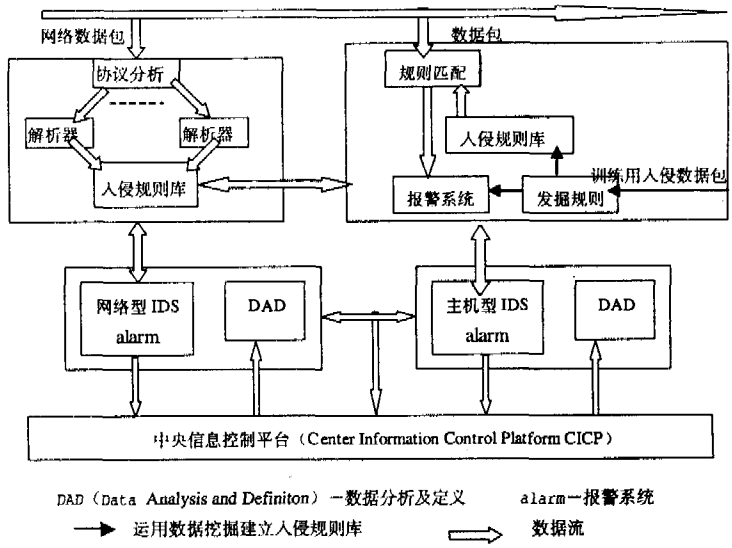

图 1 一个混合模式的入侵检测架构

1.1 数据挖掘技术

数据挖掘是从大量的数据中抽取潜在的、有价值的知识(模型或规则)的过程。自 Agrawal 等在 1993 年首先提出关联规则问题及相应的 Apriori 算法[6]后，人们又相继进行了大量研究。现在，数据挖掘中常用的挖掘方法有：

1)关联分析算法。即利用关联规则进行数据挖掘，寻找在同一事件中出现的不同项的相关性。一般用 4 个参数来描述一个关联规则的属性：置信度、支持度、最小置信度和最小支持度。通常分两个步骤：一是从事务集中找出所有支持度大于最小支持度的数据项集，称之为频繁数据项集；二是由频繁数据项集产生强关联规则。在创建入侵规则库时，还需对产生的频繁数据项集进行测试，只有那些误报率小于所设定门限值的字符串才能作为入侵特征放入特征库。

2)时间序列分析算法。发现各种事件在时间上的先后关系，进一步将数据之间的关联性与时间联系起来。在入侵检测系统中，该算法能够发现审计事件中在一起频繁发生的时间序列，这些频繁的时间序列模式是判别用户或程序行为的重要因素。其中，文献[7]中提出了将关联规则 Apriori 算法与序列模式 GSP 算法相结合来挖掘频繁模式的方法。

3)分类算法。将数据库中的数据项映射到给定类别中的一个。收集足够的“正常”或“异常”审计数据来判定一个用户或者程序是否非法，然后用这些数据来指导一个分类器学习，学习后的分类器可以用来预测一些未知的数据是否非法。

4)聚类算法。这种算法不需要训练数据，只要带有各种属性的数据记录。通过计算不同记录的属性差别，把类似的记录聚集在一起，然后利用距离矢量来判断哪些是异常记录即攻击数据。

1.2 协议分析技术

协议分析是在传统模式匹配技术基础上发展起来的一种新的入侵检测技术。它充分利用了网络协议的高度有序性,根据协议规范分析网络数据包,来确认数据包的协议类型,再使用相应的命令解析程序来检测数据包。

1.2.1 协议分析的过程

TCP/IP协议是一个被广泛承认和使用的网络互连协议,是一组不同层次上多个协议的组合。它包括:

a.物理链路层:ARP协议、RARP协议;

b.网络层:IP协议、ICMP协议、IGMP协议;

c.传输层:TCP协议、UDP协议;

d.应用层:HTTP协议、FTP协议、DNS协议、SMTP协议等。

当主机收到一个数据帧时,数据就从物理链路层开始向上升,并逐层去掉各层所在的协议加上报文首部。每层协议都要检查报文首部的协议标识,比如在IP的首部有协议字段可以确定是TCP协议还是UDP协议,见图2。

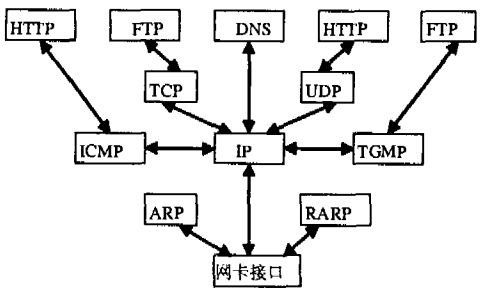

图2 协议分析图示

利用协议分析[8]检测攻击的过程:

(1)按照TCP/IP协议规定,从一个网络数据包的第13字节处读取第3层协议标识符。如果值为"0800"代表IP协议,IP协议规定IP报头的协议字段是传输层的协议标识,该位在第24字节处。

(2)跳到第24字节处读取到第4层协议标识符。若值为"01"则IP包的数据域所携带的协议为ICMP;若为"06"则为TCP协议;若为"17"则为UDP协议。

(3)TCP协议规定在第35字节处包涵四字节的应用层协议标识符(端口号)。如果读取到端口号"0080",代表的是HTTP,而HTTP协议中URL开始的地址是第55字节。

(4)协议规定HTTP位于TCP之上,URL开始于第55字节。协议分析器读取URL,并将其输入到命令解析器进行分析,与攻击特征库中与HTTP有关的攻击特征进行比对。

1.2.2 命令解析

URL的第一个字节的位置被送到一个解析器中。解析器是一个命令解析程序,可以对在不同的上层应用协议的每一条用户命令作出详细分析。并有能力识别同一命令可能的不同变种,这样一来,一种攻击只用一个攻击特征对应即可,从而使得特征库以较小的容量检测较大范围的攻击。

1.2.3 性能比较

通过协议分析和命令解析,实现对已知网络攻击的检测,并可以检测出违反协议的、可能是新的未知攻击的可疑活动。因为采用协议分析方法后的规则匹配不是对整个数据包进行匹配,并且由于命令解析有识别攻击变种的能力,故要与数据包相匹配的特征数量也大大减少,因此采用协议分析方法可以减少匹配运算量。

1.3 架构优势

首先,这是一个可扩展的架构。它容许你不断加入新的IDS到架构中,从而充分利用不同类型IDS的优势。

其次,可以全面共享数据。中央信息控制平台(CICP)是一个控制中心,可以运行数据库管理系统,管理收集到的各路数据,也可以使用数据挖掘技术以便更进一步挖掘有用的信息,将无用信息删除从而大大减少误报。

最后,架构的互操作性较好。架构中的各模块可以进行双向通讯,从而使得IDS更加智能化。

2 结束语

近年来,入侵检测系统发展的很快,整合基于主机和基于网络两种类型IDS的混合模式已成为发展趋势。文中在混合模式的基础上提出了一个运用数据挖掘和协议分析两种方法的可扩充入侵检测架构,在提高IDS准确性的同时也大大提高了系统效率,具有很好的应用前景。但对于架构中各功能模块的通信协议问题,仍有待于进一步解决。

参考文献:

[1] Anderson J P. Computer Security Threat Monitoring and Surveillance[R]. Technical report, James P Anderson Co., 1980.

[2] Denning D E. An Intrusion Detection Model[J]. IEEE Transaction on Software Engineering, 1987,13(2):222-232.

[3] COMER D, SETHI R. The complexity of trie index construction[J]. ACM,1977,24(3):428-440.

[4] Ye Nong, Vilbert S, Chen Qiang. Computer Intrusion Detection Through EWMA for Autocorrelated and Uncorrelated Data[J]. IEEE Transactions on Reliability,2003,52(1):28-32

[5] 李庆华,童健华,孟中楼,等. 基于数据挖掘的入侵特征建模[J]. 计算机工程,2004,30(8):51-53.

[6] Agrawal R, Imielinaki T, Swami A. Mining association rules between sets of items in large database[A]. In: Proc 1993 ACM SIGMOD International Conf on Management of Data [C]. Washington,DC:[s.n.],1993.207-216.

[7] 宋世杰,胡华平,胡笑蕾.关联规则和序列模式算法在入侵检测系统中的应用[J]. 成都信息工程学院学报,2004,19(1):40-46.

[8] 杜建国,郭 巧. 协议分析和命令解析在入侵检测中的应用[J]. 计算机工程与应用,2004(18):159-162.