数据挖掘技术在入侵检测中的应用研究

林志兴

(三明学院 现代教育技术中心,福建 三明 365004)

摘 要:入侵检测是计算机系统安全和网络安全重要的一个研究焦点。由于入侵手段日新月异,即使现在有多种可用的机制来检测入侵,单一的方法系统还是无法识别新类型的入侵,或者有可能发出假的警报。本文介绍了多种数据挖掘技术在计算机系统和分布式计算机网络中开发入侵检测系统的多种有效方式和方法,并做了研究。

关键词:入侵检测;数据挖掘;计算机安全;网络安全。

中图分类号:TP393.08 **文献标识码**:A **文章编号**:1674-2109(2014)05-0047-05

引言

计算机安全指的是计算机资源以及信息资源不受自然或人为因素影响,致使其招受威胁、遗失或泄露。入侵行为是一种典型的危害计算机安全的行为,其试图绕过计算机系统安全机制,并威胁到网络以及计算机资源的完整性、可用性、或保密的行为。入侵的手段主要有:DoS 攻击,R2L,U2L,Probe 等等。入侵检测过程[1]是一个用于监控发生在计算机系统或网络中的事务,并通过分析发生的异常迹象,比如未授权的入口,活动,或者文件的修改,进行预警或报警的过程。入侵检测过程具有三个重要步骤是:首先,监控和分析网络数据流;其次,识别异常活动;第三,评估严重性,提出警告。

在一个安全的计算框架中应用入侵检测基于以下三个方面考虑:

(1)许多现有的系统存在安全缺陷,导致它们容易受到攻击。但由于技术和经济的原因,这些缺陷是非常难识别和消除的。

(2)从应用和经济方面考虑,现有系统的安全缺陷难以被更安全的系统替换,而且开发完整的安全系统是不容易的。

(3)即使在高安全度的系统中,也无法避免受合法用户误用的入侵。

入侵检测技术作为计算机系统安全体系结构中不可缺少的一部分,首先要先确定入侵检测的关键指标:资源的行为哪些是属于是否“正常”或“合法”,哪些是属于“异常”或者“入侵”行为。其次,需要明确入侵检测的目标是高效比较实时活动和模型并报告可能“侵入”的活动。

入侵检测系统(IDS)是应用自动化入侵检测技术来检测可能发生入侵行为的软件。1980 年由 James P. Anderson 第一次提出入侵检测系统的概念[2]。入侵检测提供来三种必要的安全功能服务: 观察和监控,检测,和对于公司内外部未授权的活动的响应。

1 入侵攻击行为以及入侵检测

文献将攻击行为分为以下几类:[3]

(1)拒绝服务攻击(DoS):拒绝服务攻击是指攻击者使得一些计算或内存太忙或太饱和以至于无法处

收稿日期:2014-05-29

基金项目:三明学院科研基金资助(项目编号:B201201/G); 福建省教育厅科技基金资助 (项目编号:JB13187)。

作者简介:林志兴(1973-),男,汉族,高级实验师,主要研究方向:数据挖掘,计算机安全。

理有效的请求,或者拒绝有效用户对机器的请求的一类攻击。典型的例子有:Apache2, Back, Land, Mail bomb,SYN Flood, Ping of death, Process table,Smurf, Syslogd,Teardrop, Udp storm.

(2)根攻击:攻击者在系统中访问普通用户账号,采用攻击手段获得根(root)访问系统。典型的例子有:Eject, Ffbconfig, Fdformat, Loadmodule,Perl, Ps, Xterm.

(3)远程用户攻击:攻击者通过网络发送一个数据机包给机器,但是对方没有该机器的账号,利用漏洞获取本地登陆作为一个用户机器。典型的例子有:Dictionary,Ftp_write,Guest,Imap,Named,Phf,Sendmail, Xlock, Xsnoop.

(4)探测(探针):攻击者扫描一个计算机网络来收集信息或找到已知漏洞。攻击者用映射的机器和服务,可以在一个网络上使用这个信息来寻找利用。典型的例子有:Ipsweep, Mscan, Nmap, Saint, Satan.

业界将入侵检测分为误用检测和异常检测两大类。

误用检测是指根据系统和应用软件中已经发现的弱点的攻击模式来检测入侵。安全专家首先收集不寻常操作的行为特征,构建相关的特征库。这种检测类型使用已知模式或签名来识别入侵。检测系统根据特征库中的记录进行匹配,如果匹配成功,该行为将被报告为一个入侵。这种探测,误警率很低,但是漏检率高。对于已知的攻击,它会报告出攻击的类型细节和精准度,但是对于未知的攻击,这个功能是有限的,需要持续更新特征库的支持才能应对日新月异的攻击技术。

误用检测的常见的方法是签名验证,通过寻找一个攻击所留下的不变的签名,系统检测可以到已知的攻击。这个签名保存在审计文件中,或者在被侵入的主机中, 或在被攻击机器寻找内外部包的嗅探器中,可以通过合适的技术提取签名。

误用检测的方法模型有:

(1)基于专家系统的误用检测,安全专家根据经验和以往发生的攻击, 形成一套描述攻击的规则,并构筑专家入侵检测系统。

(2)基于签名验证的误用检测,提取已知的入侵行为的特征编码,组成入侵行为签名,根据对签名的验证和匹配,检测是否存在入侵行为。

(3)基于 Petri 网的误用检测,将已知的攻击行为转换为图形化的 Petri 网中的状态, 能用简洁明了的图形来表示状态复杂的入侵特征。

(4)基于状态转换图的入误用检测,应用状态转换图将系统使用过程作为一个事件序列,该序列对系统的状态进行转换,当转入到被入侵状态,说明存在入侵行为。

异常检测是探测现在发生的行为与可接受行为间的偏差, 其定义每一个可接受行为和不可接受行为。异常检测的特点是建立正常行为概要文件,通过(正常)配置文件观察并比较当前活动,当出现较大偏差时,报告出现入侵行为。该方法的漏检率比较小,不过,即使这种方法可以有效地探测未知的入侵,但是误警率却比较高,而且在异常检测中,系统无法通过顺序捕获事件来分析各事件间的相关关系。

常见的异常检测方法涉及到统计分析,用户或者系统行为是被大量随着时间流逝的变量测量的。这些变量可能是每个会话的登陆登出时间,在会话中资源消耗的数量,已经资源持续时间。这种方法的主要限制是要在没有频繁的错误警报探测时寻找一个正确的阀值。

异常检测的方法:

(1)基于阀值异常检测,通过设定和监控某些指标,在服务器或网络上检测异常活动,例如一个服务器异常消耗的 CPU,或网络异常饱和,根据出现异常情况,进行告警。

(2)基于统计学的异常检测,应用统计学技术,收集系统长期的正常活动,构造正常活动模型,通过与正常活动模型进行比较,发现存在的异常。

(3)基于数据挖掘的异常检测,主要包括基于系统调用序列分析,神经网络,遗传算法,关联规则,贝叶斯网络以及马尔可夫模型等。这也是本文主要介绍的内容。

2 应用在入侵检测上的数据挖掘方法

数据挖掘(DM),也称为知识-发现和数据挖掘,是一个自动搜索大量的数据模式的过程。数据挖掘过

程从大量的不完整、噪声、非稳定的、模糊的和随机数据中,提取有效的、更新的、潜在的、有用的、可以理解的模式。在入侵检测系统中,重要的信息来自主机日志、网络数据包,系统的日志数据,应用程序和警告信息。在数据提取特征的过程中,数据挖掘技术具有巨大的优势,因此,在入侵检测中使用数据挖掘技术是非常重要的。Lee 和 Salvatore J.Stolfo[4]第一次把数据挖掘技术运用在入侵检测研究领域。数据挖掘方法提供自动入侵检测功能,它们的知识来自于审计数据中描绘用户正常和不正常的行为。应用数据挖掘技术的主要限制之一是它们对行为模式的改变需要较长的时间进行学习和适应。随着大量数据的增加,数据仓库和数据挖掘技术的协作在入侵检测领域应用是越来越广泛。以下介绍一些重要的数据挖掘方法在入侵检测方面的应用。

2.1 基于人工神经网络(ANN)的入侵检测

利用人工神经网络[5]入侵检测技术是从不完整数据中归纳数据,并且能够将数据分类成正常的或侵入的两种类型。在 IDS 中经常使用到的人工神经网络有:多层前馈神经网络(MLFF),多层感知器(MLP)和反向传播(BP)。Cannady [6]提出了一个三层神经网络用于网络中的误用检测,这种方法用的特征向量是由 9 个网络特征(协议 ID,源端口,目的端口,源 IP 地址,目的 IP 地址,ICMP 类型,ICMP 代码,原始数据长度,原始数据),但是,该方法进行入侵检测的精确度比较低。Moradi 和 Zulkernine[7]提出了基于多层感知器的入侵检测系统,他们采用更多的隐藏层提高了 IDS 的检测精度,其检测精度高于[6]提出的方法。Ibrahim[8] 采用分布式时间滞留神经网络(DTDNN)进行检测,对于绝大多数网络攻击而言,分布式时间滞留神经网络(DTDNN)具有更高的探测精度。对于分类数据并要求高速且拥有快速转换率,DTDNN 是一个简单并有效的解决方案。结合该方法与上述提到的其他技术,可以提高该方法的精度。

对于非结构化的网络数据,基于 ANN 的 IDS 是一个有效的解决方案,这种方法的入侵检测精度是取决于隐藏层的数量。

2.2 基于模糊逻辑的入侵检测

模糊逻辑[5]可以用来处理不精确的入侵描述,应用模糊入侵检测(FIDS)技术,可以有效检测各种网络入侵行为。Chavan 等人[9]在减少 ANN 训练时间中采用进行模糊神经网络(EFuNN)方法,混合了监督学习和无监督学习两种类型,实验结果表示,相比于只使用 ANN,使用减少输入模糊神经网络可以得到更好的 IDS 分类精度。对于检测网络入侵,Chavan 等人提出的方法 [9] 不能应用于实时检测,原因是训练时间过长。Su et al. (2009)[10]提出的模糊关联规则进行检测实时网络入侵,两个规则集是从训练数据在线归纳和产生。由于进行比较的数据仅仅取自网络数据包报头,可以大量减少处理的数据量,这种方法常常用于实时检测大规模的 DoS/DDoS 攻击。

2.3 基于关联规则的 IDS

有的入侵攻击形成基于已知的攻击或变种。为了检测这样的攻击,使用基于签名 apriori 算法[11],在给定的攻击集中发现频繁发生的子集(包含原始攻击的一些特称)。

Han 等人[11]提出了基于网络的利用数据挖掘技术的入侵检测。在这种方法中,基于签名算法生成的签名误用检测。然而,该算法的缺点是它用于生产签名的时间开销比较大。采用扫描约简算法来减少数据库扫描数量,以此有效地从已经发现攻击中生产签名。然而,由于产生了一些额外的不需要的模式,造成了非常高的误警率。Lei 等人[12]提出基于支持长度减少先验算法 (length decreasing support based apriori algorithm)来检测入侵,以减少短模式的产生,这个比其他以 apriori 为基础的方法更快。在云环境中,关联规则可以被用来生成新签名。使用新生成的签名,已知攻击的变化可以被实时检测。

2.4 基于支持向量机器(SVM)的入侵检测

在有限样本数据的情况下,可以使用 SVM [5] 检测入侵,其特点是数据的维度增加不会影响到精确度。

在 Chen 等人[13]的实验结果中,相比于 ANN,SVM 案例中的结果要更好(考虑到误警率),因为 ANN 需要大量的训练样本来进行有效分类,而 SVM 需要设置少量参数。虽然 SVM 只能用于二进制数据,但是结合 SVM 和其他技术可以提高检测精度。SVM 分类器也常常用于和 SNORT 协作来减少误警率并且提高 IPS 的精确度。在云环境里,如果给出了有限样本数据

来检测入侵,那么使用 SVM 是一个有效的解决办法,因为数据的维度不影响基于 SVM 的 IDS 的精确度。

2.5 基于遗传算法(GA)的入侵检测

遗传算法(GAs)用于选择网络特性(来确定最优参数),并用于其他技术来来实现结果优化和提高 IDS 的准确度。基于 GP 方法从网络特性中生成规则,他们使用支持基于置信度的适应函数来派生规则,这样能有效分类网络入侵。然而,用于适应函数的训练期使花费很多的时间。

Gong 等人(2005)[14]使用获取的数据包的 7 个特性(时长、协议、源端口、目的端口、源 IP、目标 IP、攻击名称)。他们使用简单且灵活的支持置信度的适应函数框架,生成的规则用来检测网络入侵。使用定量和网络的分类功能来生产分类规则,增加了检测率和提高了精确度。Xiao 等人[15]提出信息理论和基于遗传算法的方法,用于检测异常行为。它能够识别少量的网络特性和基于互惠信息的网络攻击的紧密结合之间的网络特性和入侵的类型。然而,这种方法只考虑离散特性。Dhanalakshmi 和 Ramesh Babu (2008)[16]提出一个方法,该方法用于在结合模糊和遗传算法下检测误用和异常。模糊用于在入侵检测中包含定量参数,而遗传算法用来寻找在数值模糊函数里的最佳匹配参数。这个方法解决了最佳匹配问题。在云环境中,对入侵检测选择最优参数(网络功能)将会提高的潜在的 IDS 的精准度。因此,基于遗传算法(GAs)的 IDS 可以应用在云环境中。

3 结论以及将来发展前景

本文系统地描述了基于数据挖掘技术的入侵检测系统的设计思想和方法,介绍了目前主要的数据挖掘方法以及在入侵检测系统中的应用,并且还探讨在云环境下这些数据挖掘方法的适用范围。在入侵技术快速发展的现在,入侵检测作为一项重要的计算机安全技术,越来越需要能够融入自适应,自学习的入侵检测方法。智能化入侵检测以及面向云计算环境的入侵检测技术将是将来发展的主要方向。

参考文献:

[1] V.Mehta,C.Bartzis,H.Zhu,E.M.Clarke and J.Wing ,"Ranking attack graphs" presented at the International Symposium on Recent Advances in Intrusion Detection" , Hambay, Germany ,sep 20-22 , 2006.

[2] James P.Anderson ," Computer Security Threat monitoring and Surveillance "Technical report 98, Washington , Pennysylvannia , USA, April 1980.

[3] Lunt, T.,"Detecting Intruders in Computer Systems", In Proceedings of IEEE conference on Auditing and Computer Technology, February 1999.

[4] W. Lee and S. Stolfo, "Data Mining Approaches for Intrusion Detection", Proc. 7th USENIX Security Symposium , San Antonio, TX, 1998, 79-94,1998.

[5] Han J, Kamber M. Data mining concepts and techniques. 2nd edition Morgan Kaufmann Publishers; 2006.

[6] Cannady J. Artificial neural networks for misuse detection, National Information Systems Security Conference, 1998.

[7] Moradi M, Zulkernine M, A neural network based system for intrusion detection and classification of attacks. In: Proceedings of the 2004 IEEE international conference on advances in intelligent systems—theory and applications; 2004.

[8] Ibrahim LM. Anomaly network intrusion detection system based on distributed time-delay neural network. Journal of Engineering Science and Technology 2010;5(4):457 – 71.

[9] Chavan S, Shah K, Dave N, Mukherjee S, Adaptive neuro-fuzzy intrusion detection systems, IEEE international conference on information technology: coding and computing (ITCC'04); 2004: pp 70 – 4.

[10] Su M-Y, Yu G-J, Lin C-Y. A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach. Computer Security 2009:301 – 9.

[11] Han H, Lu XL, Ren LY.Using data mining to discover signatures in network-based intrusion detection. In: Proceedings of the first international conference on machine learning and cybernetics, Beijing (1) (2002).

[12] Lei L, Yang D-Z, Shen F-C. A Novel rule based Intrusion Detection system using Data Ming. 3rd IEEE International Conference on Computer Science and Information Technology 2010;6:169 – 72.

[13] Chen W-H, Su S-H, Shen H-P. Application of svm and ann for intrusion detection. Computer Oper Res 2005;32

(10):2617 – 34.

[14] Gong RH, Zulkernine M, Abolmaesumi P. A software implementation of a genetic algorithm based approach to network intrusion detection. In: Proceedings of the sixth international conference on software engineering, artificial intelligence, networking and parallel/distributed computing and first ACIS international workshop on self-assembling wireless networks (SNPD/SAWN'05); 2005.

[15] Xiao T, Qu G, Hariri S, Yousif M. An efficient network intrusion detection method based on information theory and genetic algorithm. In: Proceedings of the 24th IEEE international performance computing and communications conference (IPCCC '05), Phoenix, AZ, USA; 2005.

[16] Dhanalakshmi Y, Ramesh Babu I. Intrusion detection using data mining along fuzzy logic and genetic algorithms. International Journal of Computer Science & Security 2008;8(2): 27 – 32.

A Research on Data-mining Techniques Used in Intrusion Detection

LIN Zhixing

(Modern Educational Technology Center, Sanming college, Sanming, Fujian 365004)

Abstract: Intrusion detection in computer system security and network security is an important topic of research. Even now there are a variety of available mechanisms to detect intrusions, a single method or system does not recognize the new type of invasion, or is likely to raise false alarms. This paper introduces a variety of data mining technology in computer systems and distributed computer network intrusion detection system which are developed in a variety of effective ways and methods, and presents a comparative study.

Key words: intrusion detection; data mining; computer security; network security