基于数据挖掘的入侵检测系统设计

李守国,李 俊
(南京航空航天大学 计算机科学与工程系,江苏 南京 210016)

摘 要:现有的入侵检测系统大多都是采用手工编码构造的,检测模型的构造过程很大程度上依赖于系统构造者的知识和经验,这样构造出的模型往往存在很大的缺陷。针对传统入侵检测系统构造过程中存在的种种问题,将数据挖掘技术引入入侵检测系统,实现检测模型构造的自动化。介绍了一个运用数据挖掘技术构造入侵检测系统的框架,并考虑到实时检测过程中对检测模型效率的要求,提出了一个提高检测模型检测效率的层叠检测模块方法。应用数据挖掘算法得出的检测模型在检测效率、准确性、可扩展性和自适应性等方面都得到了很大的改进。

关键词:网络安全;入侵检测;数据挖掘;移动代理
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1005-3751(2006)04-0212-03

Design of Data Mining Based Intrusion Detection System

LI Shou-guo, LI Jun
(Dept. of Computer Sci. and Eng., Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Currently most intrusion detection systems are constructed by manual means, constructing process of these systems highly depend on system builders' knowledge and experience, so there are always some flaws in these models. Aiming at the problems exists in the process of IDS (intrusion detection system) constructing, apply data mining technology to IDS to construct the models automatically. In this paper, introduce a data mining based intrusion detection framework, and considering the real-time efficiency requirement of the intrusion detection model, propose "cascaded detection modules" approach to improve efficiency of IDS. The models constructed by applying data mining algorithms improve a lot in some aspects, such as efficiency, accuracy, extensibility and adaptability.

Key words: network security; intrusion detection; data mining; mobile agent

0 引 言

随着基于网络的计算机系统在人们的生活中起到越来越重要的作用,出于政治、经济、军事以及其他一些原因,这些系统成为敌人和罪犯的攻击对象。传统的网络安全技术:如数据加密、防火墙、用户验证等技术不足以保障系统的安全,人们逐渐认识到对于计算机安全来说没有万灵药,需要构建一个层次的网络安全保障体系,而入侵检测技术就是网络安全层次结构的一个重要组成部分。

入侵可以定义为:潜在的有预谋未经授权访问信息、操作信息、致使系统不可靠或无法使用的企图[1]。入侵检测则是为对企图入侵、正在进行的入侵或者已经发生的入侵进行识别并采取相应措施的过程。入侵检测按照采用的检测技术可以分为:异常检测(Anomaly Detection)和误用检测(Misuse Detection)。异常检测是通过建立程序或用户的正常行为轮廓,把任何偏离此正常轮廓的行为都确定为入侵或攻击,异常检测的优点是可以检测到未知类型的入侵行为,而缺点是误报率高;误用检测是对已知的攻击方式或系统的弱点进行建模,将与预先定义好的攻击特征相匹配的行为确定为攻击,其优点是可以有效地检测到已知攻击,缺点是对新的入侵行为无能为力,漏报率高。

现有的入侵检测系统大多都是采用手工编码构造的,检测模型的构造过程很大程度上依赖于系统构造者的知识和经验,这样构造出的系统在效率、准确性、可扩展性和适应性等方面都存在很大的缺陷。为了改善检测模型构造的效率,把数据挖掘技术引入入侵检测系统,异常检测就是通过挖掘审计数据建立正常使用模式,而误用检测就是利用审计数据来对入侵进行建模和匹配。

1 数据挖掘在入侵检测中的应用

简单的说,数据挖掘是从大量的数据中提取或"挖掘"出潜在的、有价值的知识的过程[2]。近来数据挖掘技术的快速发展已从统计学、模式识别、机器学习和数据库等领域中得到了大量的算法,一些算法尤其适用于入侵检测,如分类分析、关联规则分析和序列模式分析[3]。

(1)分类分析:分类是一个两步过程,首先通过分析训练数据建立一个模型,如决策树或规则;然后利用模型对

收稿日期:2005-08-30
基金项目:国防科工委国防基础科研项目(S0500B003)
作者简介:李守国(1982-),男,山东沾化人,硕士研究生,研究方向为计算机网络和网络安全;李 俊,研究员,主要从事数据库和计算机网络技术研究工作。

未知类别数据进行分类。分类分析在入侵检测中的应用过程是:收集用户或程序的大量正常和异常审计数据,然后运用分类算法产出一个分类器,可以预测新的未知类别审计数据属于正常或异常。

(2)关联分析:挖掘数据记录中不同数据项之间的关联性。审计数据中各系统特征的关联可以作为构造正常使用轮廓的基础。比较常用的算法有 Apriori 算法等。

(3)序列分析:序列分析是发现不同数据记录之间的相关性。序列分析的目标是在事务中挖掘出序列模式,即满足用户指定的最小支持度要求的频繁序列。从原始审计数据中挖掘出的关联规则和序列模式可以用于指导特征选择和原始审计数据的收集。

2　一个基于数据挖掘的分布式入侵检测系统框架

2.1　集中式和分布式 IDS

入侵检测系统按照体系结构可以分为集中式入侵检测系统和分布式入侵检测系统。集中式 IDS 可能有多个分布于不同主机上的收集部件,但只有一个中央入侵检测服务器,收集部件将当地收集到的数据发送给中央检测服务器进行分析处理;而分布式 IDS 的分析部件位于不同主机上。集中式 IDS 往往很容易被攻击者破坏,如当遇到 DDOS 攻击时,在很短的时间内大量数据包流经网络,一个检测网络中的所有数据包内容的集中式 IDS,通常会因为资源耗尽而停止工作;同时当遇到新的攻击形式或计算环境发生变化时,集中式 IDS 的扩展和更新是缓慢且高代价的。分布式 IDS 比集中式 IDS 更具灵活性和可扩展性,因而成为现在入侵检测系统框架构建的一种趋势。

2.2　系统框架结构

鉴于上面提到的集中式 IDS 的缺陷,提出了一个基于移动代理和数据挖掘的分布式 IDS 框架结构,如图 1 所示。以下对其结构分别加以说明。

2.2.1　审计数据收集

要想分析和使用原始审计数据,首先需要收集审计数据。对基于主机的信息源:系统审计日志、系统日志和应用程序日志,都是系统和应用程序自动生成的,这些日志往往是以文件的形式存储,例如一个文本文件,因此基于主机的信息源获取比较容易;而对于基于网络的信息源:网络数据包,可以使用现有包俘获工具,如 tcpdump 等来获取。

2.2.2　数据预处理器

由于采用的数据采集工具是通用的,并不是专门针对网络安全而设计的,因此需要对这些工具的输出要进行多次预处理后,才能运用于入侵检测的过程。例如:tcpdump 输出的是二进制的原始审计数据,要把它预处理成包含多个基本属性(源 IP 地址,目标 IP 地址,协议类型等)的网络连接记录,然后再应用数据挖掘算法。

2.2.3　数据挖掘

用数据挖掘算法中的关联分析算法和序列分析算法挖掘连接记录数据库中的频繁模式,如关联规则和频繁序列。利用这些频繁模式,为连接记录构造附加特征,如时间统计特征、主机统计特征等。通用的数据挖掘算法通常会产生大量的无用模式,为了消除这些无用模式,引入了"轴属性"和"参考属性"[4]的概念。最后将分类算法应用于处理后的审计数据得到一个分类器,即入侵检测模型,用来判断当前用户行为正常或入侵。此过程需要不断地反复和评估,比如如果检测模型的检测效果不够理想,就需要通过反复的频繁模式挖掘和比较,构造更有助于检测准确率的特征。

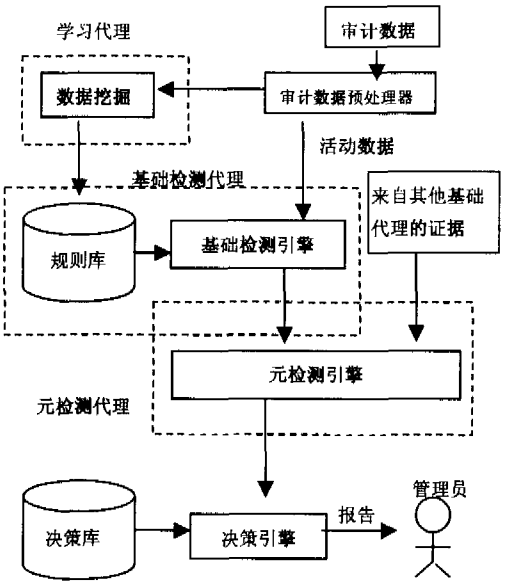

图 1　一个基于移动代理和数据挖掘的 IDS 框架结构

2.2.4　入侵检测和决策响应

用构造的模型对当前审计数据进行检测,根据检测的结果,从决策库中寻找匹配决策,执行相应的行动。如果属于入侵行为,则系统给出警报,并采取一定的措施,如:断开网络连接、报告系统管理员等,并留下入侵证据;如果属于正常行为,则系统继续进行监视。

2.2.5　学习代理和检测代理

学习代理利用数据挖掘从大量的原始审计数据中学习和计算规则库,产生基础检测模型和元检测模型;检测代理中配置有学习代理计算出的规则库,它的检测引擎对当前数据进行分析,输出入侵证据,它又分为两种:基础检测引擎和元检测引擎,两者的区别是前者以预处理的审计数据作为输入,而后者是以各个基础检测代理产生的入侵证据作为输入;通过元学习[5],元检测模型把多个基础检测模型的检测能力综合了起来。

2.3　框架结构的优点

此框架结构的优点是:检测代理同任务较重的学习代理分离开来,多个检测代理可以并行执行,提高系统执行的效率;不必将审计数据传输到某个中央检测服务器集中检测,降低了网络中传输的数据量;计算机网络系统中往往存在多个"渗透点",在系统的各个"渗透点"配置检测代理,将基于网络的入侵检测和基于主机的入侵检测结合起

来,更好地保障网络系统的安全;采用异常检测技术的基础检测代理在检测过程中发现新的入侵形式时,就把审计记录传输到学习代理,通过计算得到一个可以检测此类入侵的更新了的分类器,然后将它分派给所有的基础检测代理,通过将异常检测和误用检测结合起来,可以提高系统的检测率,并提高了系统的可扩展性和自适应性。

3 提高入侵检测系统实时检测效率

目前,大部分审计机制都被设计成详尽记录系统和网络中的活动,这在保证没有入侵证据被遗漏的同时,鉴于网络数据的高速度和大流量,同样要求所构造的入侵检测系统在实时检测时具有高效率。否则,对审计数据(如:网络数据包)分析的延迟就可能给入侵成功提供机会。这就要求构造的入侵检测模型不仅要是准确的,而且是高效的。

基于数据挖掘的入侵检测的效率是通过计算检测所必需的特征来测量的,因此这里提出了一个层叠的检测模型,如图 2 所示。底层的检测模块采用计算代价低的系统特征,而高层的检测模块采用计算代价比较高的系统特征,在图 1 的框架结构中反映为通过数据挖掘,选择不同的算法和系统特征,生成多个层叠的基础检测代理。

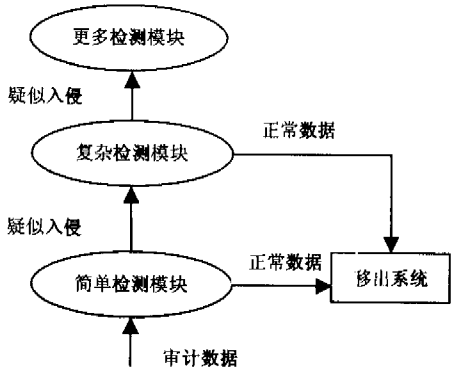

图 2 层叠检测模块

数据首先经过比较简单的检测模块,只有疑似为入侵的数据才被传递到下一层更为复杂的模块进行检测,这样只有一部分审计数据,也就是可能为攻击的那部分审计数据才被复杂的检测模块计算,这样大大提高了入侵检测系统的实时运行效率。多个检测模块之间是独立的,而且通常采取不同的检测算法和特征。含有 N 个检测模块的层叠检测模型的检测率是各个检测模块检测率的乘积,为了使层叠检测模型有高检测率,就要求各个检测模块都要有高的检测率。而检测模块在拥有高检测率的同时,通常都会有高的误报率,采用层叠检测模块可以使得整个检测模型的误报率变得很低,例如:如果层叠模型中有 4 个检测模块,而每个模块的误报率是 10%,那么整个模型的误报率就是各个模块误报率的乘积:10^{-4},因此通过采用层叠检测模型,可以构造一个拥有高检测率和低误报率的入侵检测系统,而且这个系统在实时运行时是高效的。

4 结束语

近几年来,基于数据挖掘的入侵检测系统成为研究的热点。文中提出了一个基于数据挖掘和移动代理的分布式入侵检系统框架和一个提高入侵检测模型运行效率的方法,提高了入侵检测系统的执行效率,改善了系统的可扩展性和自适应性。在实现时还存在一些需要解决的问题,如在提高检测效率的层叠检测模型方法中,如何选择适当的特征,才能使得复杂的检测模块比简单的模块更为广泛,更能有效地检测攻击等。今后,将针对这些问题展开进一步的研究。

参考文献:

[1] 罗守山.入侵检测[M].北京:北京邮电大学出版社,2004.
[2] 范 明,孟小峰.数据挖掘:概念与技术[M].北京:机械工业出版社,2001.
[3] Lee W, Stolfo S J, Mok K W. A data mining framework for building intrusion detection models[A]. In Proceedings of the 1999 IEEE Symposium on Security and Privacy[C]. Oakland, CA:[s.n.],1999.
[4] Lee W, Stolfo S J. Data mining approaches for intrusion detection[A]. In Proceedings of the 7th USENIX Security Symposium[C]. San Antonio, TX:[s.n.],1998.
[5] Chan P K, Stolfo S J. Toward parallel and distributed learning by meta - learning[A]. In AAAI Workshop in Knowledge Discovery in Databases[C].[s.l.]:[s.n.], 1993.227 - 240.

(上接第 211 页)

4 结束语

通过以上的论述,可以看出 JDBC 为连接不同数据库提供了统一的接口,使 Java 开发人员更加容易连接并操作各种数据库,而不需要单独为数据库编写不同的连接程序,保证了 Java"一次编写,各处使用"的平台无关性原则。JSP 技术通过在 HTML 页面中嵌入 Java 代码的方式,极大地方便了网页的动态扩展功能,客户端只要提出数据库访问请求,具体的操作则由中间层完成。如果将 JSP 中数据库的连接操作代码写成专门的 JavaBean,将更加方便代码的复用。

参考文献:

[1] 周彩兰,陈才贤.基于 Java 的 Web 数据库连接池高效管理策略[J].武汉理工大学学报,2004(5):38 - 41.
[2] 刘 慧,李玉忱,苏 鹏.基于 J2EE 架构的分布式 Web 应用的研究[J].计算机应用研究,2003(9):47 - 49.
[3] 石振国.用 JSP 实现对 Web 数据库的访问[J].计算机应用,2001,21(5):91 - 93.
[4] 李 平,沈国民,李 哲.基于 JSP 技术的 WEB 数据库设计[J].电脑与信息技术,2000(6):1 - 3.
[5] 张维玉,李明东,陈 劲.Web 数据库技术分析[J].西华师范大学学报,2004,25(2):219 - 222.