文章编号:1000 - 582X(2002)10 - 0128 - 04

基于数据挖掘的入侵检测*

刘勇国,李学明,廖晓峰,吴中福
(重庆大学 计算机科学与工程学院,重庆 400044)

摘 要:随着计算机网络在现代社会中扮演日益重要的角色,信息安全成为信息技术研究领域最重要的研究课题之一。而入侵构成了严重的安全风险,如何有效防范和检测入侵行为是信息监管中的热点研究问题。传统入侵检测模型的建立过程效率低,研究成本高,而数据挖掘在未知知识获取方面具有独特优势,因此基于数据挖掘的入侵检测成为研究热点。针对入侵现状、入侵检测和数据挖掘研究及开发状况,笔者分析了基于数据挖掘的入侵检测研究背景、体系结构、研究方法、所需解决的问题及今后的研究方向。

关键词:入侵检测; 数据挖掘; 信息安全
中图分类号:TP393 **文献标识码**:A

随着信息技术的迅速发展,特别是因特网的普及,时刻影响着全球各国的战略部署,网络设施和信息资源对于国家、企业和个人都是极其重要的,而网络安全是网络信息技术发展的基石,已成为计算机领域的重要研究课题之一。美国 General Accounting Office(GAO) 1996 报告[1-2],1995 - 1996 年度针对美国政府计算机系统的入侵事件达 25 000 次,至少 10 个关系到 98% 政府预算的主要部门曾受到攻击,而其中仅 1% ~ 4% 的入侵被检测出,甚至有的仅 1%;同时显示,入侵在过去 5 年中以 250% 速度增长;99% 的大公司均发生过重大入侵事件。网络攻击/入侵频率和攻击者分布从 80 年代到 2000 年发生了重大变化[3-4],80 年代,入侵者一般是信息安全领域专家,拥有深厚的专业知识和独特的入侵手段,很少依靠专用入侵工具;而今,任何人都能进行攻击,因为通用入侵工具广为传播,从因特网上能轻松获取;同时,有经验的入侵者正变得更加狡猾,其攻击类型日益复杂,手段多样性趋势更加明显。

1 入侵检测研究

针对网络安全中攻击和入侵事件的高速增长趋势,国外学者纷纷进行专项课题研究。Anderson[5] 将攻击分为 4 类:1) 外部渗透(External penetrator),指获得系统访问权的非法用户;2) 伪装者(Masquerader),包括外部渗透和系统其它的授权用户,企图利用他人授权信息对系统进行访问;3) 滥用权利者(Misfeasor),指系统合法用户违反系统安全策略滥用权利;4) 秘密用户(Clandestine user),指在低于正常审计机制下操作系统,其入侵行为征兆隐藏,不容易检测到。真正揭示入侵检测领域的是由 Denning[6] 提出最早的入侵检测模型,他提出了入侵可检测条件,即系统能够为正常用户行为自动建模,当某操作行为明显偏离正常模型时,系统认为处于异常状态并存在入侵可能。另一个模型是 Sebring[7] 提出的 MIDAS 系统,它以异常行为规则为中心建造专家系统,对已知入侵行为进行编码,通过审计数据完成检测功能。在此以后,IBM、Compaq、Cisco 等公司和麻省理工学院、普渡大学、卡耐基 - 梅隆大学等研究机构分别进行入侵检测研究,提出了各自的检测模型。

Axelsson[8] 针对高校和科研部门研制的主要入侵检测模型进行分类并给出评价结果,认为入侵检测的发展趋势是:1) 从基于主机的入侵检测向基于网络的入侵检测转变;2) 从集中式检测向分布式检测发展;3) 增强检测系统间的互操作性;4) 加强入侵检测系统自身安全性的研究。与此同时,现存模型仍存在问题:1) 误用检测与异常检测的集成;2) 入侵检测的实时

* 收稿日期:2002 - 06 - 18
基金项目:重庆市应用基础研究项目(6801)
作者简介:刘勇国(1974 -),男,四川绵阳人,重庆大学博士研究生。主要研究方向:入侵检测,网络安全,数据挖掘,进化计算。

性;3) 增强入侵事件的主动反应;4) 系统资源的占用;5) 检测系统的分类覆盖;6) 从入侵检测观点研究入侵的本质;7) 网络安全系统中 SSO(Site Security Officer)的角色扮演。

Kvarnström[9]针对主要商用检测系统进行性能比较,认为其发展趋势是:1) 明确入侵检测系统在网络信息安全框架中的角色;2) 基于主机和网络的入侵检测系统的集成;3) 检测系统自身的安全问题;4) 不同检测系统间的互操作性;5) 软件开发商背景的问题。

2 数据挖掘技术

网络技术、数据库技术的进步,以及 Web 技术的出现,增强了信息生产和数据搜集能力。成千上万的数据库用于商业管理、行政办公、科学研究和工程开发,而系统存储的海量数据又引发了新问题。这些庞大数据库及其海量数据是极其丰富的信息资源,但靠传统数据检索机制和统计分析方法不能满足信息有效提取的需要,只有充分利用其为企业业务决策和战略发展服务,否则只能成为包袱。因此,从数据库中发现知识(Knowledge Discovery in Database, KDD)及其核心技术——数据挖掘(Data Mining)应运而生。

数据挖掘是源于大型零售商在面对决策支撑问题提出的[10],它是从大量的、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。KDD 是识别出存在于数据库中有效的、新颖的、具有潜在价值的乃至最终可理解的模式的非平凡过程。知识发现有 4 个主要阶段:数据采集选择、数据预处理、数据挖掘和解释评价。数据挖掘是知识发现的一个特定的、关键阶段,知识发现是其中的一个或多个阶段的反复过程。实际系统中收集到的原始数据存在杂乱性、重复性及不完整性,数据采集选择是辨别需分析的数据集合,缩小处理范围;数据预处理,包括数据集成(Data Integration)、数据清理(Data Cleaning)、数据变换(Data Transformation)和数据约简(Data Reduction)等几方面,其功能是处理数据中遗漏、脏数据以及子集选择等问题;数据挖掘阶段进行实际挖掘操作,在驱动发现型(Discovery - Driven)和驱动验证型(Verification - Driven)挖掘中进行判断,然后选择合适方法运行;解释评价的任务是将挖掘结果采用合适的方式表达出来,包括可视化技术、信息过滤、信息综合等,不满意时,需重复 KDD 的某个或几个过程。

3 基于数据挖掘的入侵检测研究

3.1 问题的提出

Wenke Lee 针对 TCP/IP 协议存在的安全问题[11],认为入侵检测是针对危及系统安全,即保密性、完整性和可用性的恶意行为的识别和反应过程,其基本前提条件是:用户和程序行为是可见的及正常和入侵行为具有截然不同迹象。因此,入侵检测系统应具备几个基本要素:1)系统资源。如网络设施、用户帐号、系统内核等;2)定义系统资源合法使用行为的模型;3)用于行为监控的技术。

国际上对于入侵检测领域的研究已经进行了多年,研制的系统针对已知入侵行为检测精度较高,但是对于未知攻击模式检测结果较差。传统的基于知识的检测系统需要领域专家针对检测类型选择统计方法,首先将攻击行为和系统弱点进行分类,然后人工进行代码输入,完成相应规则和模式的建立工作,因此系统在扩展性和适应性方面不能适应形势的发展。为了提高检测未知攻击能力,入侵检测系统应当为整个网络系统提供全面的安全保护,相对于复杂的网络系统而言,专家知识相对局限,因此如何自动系统建立具有自适应和扩展能力的入侵检测系统成为目前此领域重要研究课题之一。Lee 提出以数据为中心的思想[11],利用数据挖掘对审计数据进行建模的可行性和有效性,并构造出检测模型,研制的基于数据挖掘的入侵检测系统 MADAM ID 在 1998 年 MIT 主办的 DARPA/AFRL 入侵检测评价上表现出良好特性。与此同时,IBM、哥伦比亚大学、纽约州立大学等研究单位相继开始进行基于数据挖掘的入侵检测研究。

3.2 体系结构

建立入侵检测模型的全过程如下[11]:首先,审计源数据被处理为 ASCII 码形式的网络数据包(或主机事件数据),然后转换为连接记录(或主机通话记录),其中包含大量连接特征,如服务类型、持续时间、标识等。接着针对连接记录应用数据挖掘算法,从中发现所需模式,这个过程是反复进行的。

图 1 为提出的基于数据挖掘的入侵检测系统的体系结构[12],系统由探测器、检测器、数据仓库和自适应模型生成器模块构成。此系统能够完成数据收集、共享、分析、归档以及模型的产生和分配功能,而且与探测器数据格式和模型表示无关,其中探测器数据块包含任意数目的特征,可以是连续或离散的,数字或符号的,与此同时,系统采用 XML 编码方式适应异构特性,结构中的模型可为神经网络、规则集或概率等模型。通过采用 CIDF(Common Intrusion Detection Framework)和 IDMEF(Intrusion Detection Message Exchange Format)将系统通信

和合作的信息格式和协议标准化,系统能够安全地进行攻击信息交换以协作进行分布式入侵检测。

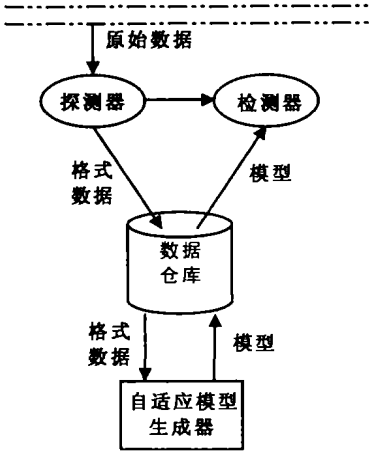

图1 基于数据挖掘的入侵检测系统体系结构

IBM公司为客户提供的紧急事件反应服务系统中的实时入侵检测系统模型[13]。由于客户数目的不断增长,检测系统探测器产生的警报数目不断增加,同时存在误报的情况,如何处理此类问题。IBM研究人员采用数据挖掘对系统检测报告数据库进行分析,构建探测器警报正常行为模型,以检测异常入侵;同时根据历史数据挖掘不同类型客户,提供相应的监控手段。此模型用于协助操作人员,它是一个自动决策引擎,采用基于知识的决策规则对到来的警报进行筛选,此规则由数据挖掘引擎对历史数据进行分析后协助进行更新。

3.3 主要研究方法

基于数据挖掘的入侵检测研究是针对入侵检测和数据挖掘特点,提出适合的挖掘模型,在满足网络安全实际要求下,实现两者有机结合。

3.3.1 关联分析

关联分析是分析数据间隐含的相互关联关系的工具,关联规则的形式描述是:项目集 I 表示所有项目(item)集合,D 表示所有交易(transaction)集合,T 表示交易集中的某个交易,每个交易是某些项目的集合,$T \subseteq I$。设 X、Y 为项目子集,$X \subset I$,$Y \subset I$,$X \cap Y = \Phi$,如果 $X \subseteq T$,称 T 支持 X,其支持度 Support(X) = {X 在 D 中出现的次数占总次数比例};置信度 Confidence($X \Rightarrow Y$) = {在 D 中出现 X 的交易集合中 Y 出现次数占总次数比例}。关联规则表示为 $X \Rightarrow Y$,若 Support($X \Rightarrow Y$) $\geq$ min-sup 且 Conf($X \Rightarrow Y$) $\geq$ min-conf,称规则 $X \Rightarrow Y$ 为关联规则。关联规则能发现形如"90% 用户在一次购买活动中购买商品 X 的同时购买商品 Y"之类的知识。Lee 考虑到入侵检测特点,引入轴心属性(axis attributes)概念[11],即网络连接由五元组 < time-stamp, src-host, src-port, dst-host, service > 唯一确定,其中 time-stamp 为时间标记,src-host 为源主机,src-port 为源端口,dst-host 为目的主机,service 为服务(或目的端口).规定在侯选项目集中,每次交易必须包含轴心属性值,同时所有连续属性需进行离散化处理,以此满足通用关联分析中所需先验知识(prior knowledge)要求。

3.3.2 序列模式分析

序列模式用于发现如"在某一段时间内,客户购买商品 A,接着购买商品 B,而后购买商品 C,即序列 A→B→C 出现频度较高"之类的知识。由于网络攻击与时间变量紧密相关,因此序列模式分析在关联分析基础上进一步分析攻击行为时间相关性。Manganaris 指出 IBMERS(Emergency Response Service)系统正对序列模式进行扩展以适应 RTID(Real - Time Intrusion Detection)要求[13];Lee 利用关联分析的数据结构和库函数实现序列模式分析[11].其序列模式的形式化描述为:已知事件数据库 D,其中每次交易 T 与时间戳关联,交易按照区间 $[t_1, t_2]$ 顺序从时间戳 t_1 开始到 t_2 结束。对于 D 中项目集 X,如果某区间包含 X,而其真子区间不包含 X 时,称此区间为 X 的最小出现区间。X 的支持度定义为包括 X 的最小出现区间数目占 D 中记录数目比例。其规则表示为 $X, Y \rightarrow Z$, [confidence, support, window],式中 X, Y, Z 为 D 中项目集,规则支持度为 support($X \cup Y \cup Z$),置信度为 support($X \cup Y \cup Z$)/support($X \cup Y$),每个出现的宽度必须小于窗口值。考虑到网络或操作系统中审计数据流的特性,他将入侵事件序列考虑为单序列,X、Y、Z 满足偏序关系。

3.3.3 聚类分析

聚类分析是识别数据对象的内在规则,将对象分组以构成相似对象类,并导出数据分布规律。分类与聚类的区别在于分类是将分类规则应用于数据对象;而聚类是发现隐含于混杂数据对象的分类规则。Portnoy 提出基于聚类分析的入侵检测算法[14],无监督异常检测算法(unsupervised anomaly detection algorithm),通过对未标识数据进行训练检测入侵。算法设计基于两个假设:1)正常行为记录数目远大于入侵行为记录数目;2)入侵行为本质上与正常行为不同。算法基本思想在于入侵模式与正常模式本质上不同,则它们将出现在正常模式范畴之外,因此能够被检测出来。算法将数据实例进行正规化处理转换为标准形式,采用标准欧几里德度量(Euclidean metric),使用改进单链法聚类,经过标识,通过分类以检测入侵行为。但该算法不适用于恶意攻击和拒绝服务攻击的检测。

3.3.4 分类分析

分类分析是数据挖掘领域重要的研究课题之一,

其目标是建立基于属性的分类属性模型。Lee 采用规则学习算法 RIPPER，通过生长阶段(growing phase)和剪枝阶段(pruning phase)对训练集数据进行学习，其主要工作是建立结构化数据模型，形成正常和异常类分布[11]。Agarwal 提出一种基于规则的学习模型并将其用于入侵检测分类分析[15]，分类模型的基本思想是利用 P－规则集进行正向学习以覆盖正实例集，然后利用 N－规则集对所有 P－规则集支持的实例进行反向学习以覆盖负实例集，以此除去因正向学习过程中覆盖的负实例集，而且随着每次学习过程规则集精度从高到低下降。Schultz 利用分类分析对恶意代码程序(如病毒程序)进行检测[16]，首先将数据集划分为训练集和测试集，对于训练集数据采用病毒扫描软件进行类型标识，然后通过从恶意代码中的系统资源信息、字符串和位序列提取特征，分别使用 RIPPER、朴素贝叶斯和复合朴素贝叶斯算法进行学习，经过与传统的抗病毒方法进行比较，检测精度有较大提高。

3.4 存在的问题

经过深入分析发现，尽管基于数据挖掘的入侵检测模型在检测性能和通用性方面具有优势，但在实现和采用此类系统时仍然存在一定困难。

1) 检测性能方面：此类系统在某些领域比传统检测方式的误检率高，特别是异常检测系统；

2) 检测效率方面：由于是对大量历史数据处理，检测模型在学习和评价阶段的计算成本高，造成实时性实施困难；

3) 使用性能方面：系统需要大量的训练数据，而且比传统系统更加复杂；

4) 尚未考虑检测模型自身安全性问题。

因此，此领域的研究方向将在上述几个方面进行，达到实时入侵检测效果，与此同时，检测模型的可扩展性、自适应性、互操作性以及检测系统在信息安全总体框架中的角色问题也是重要的研究方向。

4 结　论

笔者对基于数据挖掘的入侵检测研究背景、体系结构、研究方法、存在的问题和课题研究方向进行了分析和概括。国际上在此领域大部分处于初步研究阶段，我国在入侵检测研究方面起步较晚，研究成果和论文较少，针对检测中现存的问题将作进一步深入研究。

参考文献：

[1] UNITED STATES GENERAL ACCOUNTING OFFICE. Information Security: computer attacks at department of defense pose increasing risks[R]. USA: GAO/AIMD－96－84, 1996.

[2] UNITED STATES GENERAL ACCOUNTING OFFICE. Information Security: opportunities for improved OMB oversight of agency practices[R]. USA: GAO/AIMD－96－110, 1996.

[3] CERT COORDINATION CENTER. CERT/CC Overview: Incident and Vulnerability Trends[R]. USA: Software Engineering Institute, Carnegie Mellon University, Pittsburgh, 2000.

[4] ALLEN J, CHRISTIE A, FITHEN W, et al. State of the Practice of Intrusion Detection Technologies[R]. USA: Software Engineering Institute, Carnegie Mellon University, Pittsburgh, 1999.

[5] ANDERSON J P. Computer security threat monitoring and surveillance[R]. USA: James P. Anderson Co., April 1980.

[6] DENNING D E, NEUMANN P G. Requirements and model for IDES—A real－time intrusion detection system[R]. USA: Computer Science Laboratory, SRI International, CA, 1985.

[7] SEBRING M M, SHELLHOUSE E, HANNA M E, et al. Expert systems in intrusion detection: A case study[A]. Proceedings of the 11th National Computer Security Conference[C], Baltimore, Maryland, October 1988. 74～81.

[8] AXELSSON S. Research in Intrusion－Detection Systems: A Survey[R]. Chalmers University of Technology, Göteborg, Sweden, August 1999.

[9] KVARNSTRÖM H. A survey of commercial tools for intrusion detection[R]. Chalmers University of Technology, Göteborg, Sweden, 1999.

[10] AGRAWAL R, IMIELINSKI T, SWAMI A. Database mining: A performance perspective[J]. IEEE Transactions on Knowledge and Data Engineering, December 1993. 5(6): 914～925.

[11] LEE W. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems[D]. USA: Columbia University, 1999.

[12] LEE W, STOLFO S J, CHAN P K, et al. Real Time Data Mining－based Intrusion Detection[A]. Proceedings of DISCEX II[C], USA: June 2001.

[13] MANGANARIS S, CHRISTENSEN M, ZERKLE D, et al. A data mining analysis of RTID alarms[J]. Computer Networks, 2000, 34: 571－577.

[14] PORTNOY L, ESKIN E, STOLFO S J. Intrusion detection with unlabeled data using clustering[A]. Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA－2001)[C]. Philadelphia, PA: 2001.

[15] AGARWAL R, JOSHI M V. PNrule: A new framework for *learning classier models in data mining* (*a case－study in* network intrusion detection)[R]. IBM, Computer Science/Mathematics, April 2000.

[16] SCHULTZ M G, ESKIN E, ZADOK E, et al. Data Mining Methods for Detection of New Malicious Executables[A]. Proceedings of IEEE Symposium on Security and Privacy[C]. Oakland, CA: 2001.

(下转第 135 页)

Agent – Oriented Software Engineering

GUO Liang, HUANG Xi-yue

(College of Automation, Chongqing University, Chongqing 400044, China)

Abstract: Agent – Oriented software engineering is the one of the most recent contributions to the field of software engineering. Agent becomes isolation and negotiation component, compared to existing development approaches, it has advantage in system development, especially where Agent represents high-level abstractions of active entities. Through the research of high-level and specific methodologies in the agent-based software engineering, this paper analyses their characteristics and some extension of software technology for implement Agent, gives some references to build agent-based system.

Key words: intelligent agents; software engineering; UML; design patterns; components

(责任编辑　张　苹)

(上接第 131 页)

Intrusion Detection Based on Data Mining

LIU Yong-guo, LI Xue-ming, LIAO Xiao-feng, WU Zhong-fu

(College of Computer Science and Engineering, Chongqing University, Chongqing 400044, China)

Abstract: As computer networks play increasingly vital roles in modern society, information security becomes one of the most important research issues in the field of information technology. But intrusions cause a serious security risk, how to efficiently prevent and detect intrusions becomes one of hot research problems in the field of information supervision. The traditional process of building the model of intrusion detection is slow, whose cost of research and development is high. However, data mining has unique advantages in acquiring unknown knowledge. So, intrusion detection based on data mining becomes a hot issue. The research background, architectures, techniques, problems to be solved and the future direction are discussed after analyzing current status of network intrusion and situation of R&D on intrusion detection and data mining.

Key words: intrusion detection; data mining; information security

(责任编辑　吕赛英)