数据挖掘技术在入侵检测中的应用研究

顾健辉,孙力娟
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:随着Internet迅速发展,许多新的网络攻击不断涌现。传统的依赖手工和经验方式建立的基于专家系统的入侵检测系统,由于面临着新的攻击方式及系统升级方面的挑战,已经很难满足现有的应用要求。因此,有必要寻求一种能从大量网络数据中自动发现入侵模式的方法来有效发现入侵。这种方法的主要思想是利用数据挖掘方法,从经预处理的包含网络连接信息的审计数据中提取能够区分正常和入侵的规则。这些规则将来可以被用来检测入侵行为。文中将数据挖掘技术应用到入侵检测中,并对其中一些关键算法进行了讨论。最后提出了一个基于数据挖掘的入侵检测模型。实验证明该模型与传统系统相比,在自适应和可扩展方面具有一定的优势。

关键词:数据挖掘;入侵检测;IDS;网络攻击
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2006)09-0243-03

Application Research of Data Mining Technology to Intrusion Detection

GU Jian-hui,SUN Li-juan
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Along with the rapid development of Internet, many new network attacks emerge unceasingly. Traditional intrusion detection system (IDS) based on expert system depending on handwork and experience, is already very difficult to satisfy the existing application request now, because it is facing challenges from new forms of attacks and system upgrade. So it is necessary to find a method that can extract intrusion patterns from substantive network data automatically. The main idea is to apply data mining methods to learn rules that can capture normal and intrusion activities from pre-processed audit data that contain network connection information. These rules can be used to detect intrusion behavior later. In this paper, data mining technology has been applied to intrusion detection, some algorithms of data mining have been discussed. Then a model of data-mining based on intrusion detection system has been proposed. The experiment proved that, compared with the traditional system, this model has certain superiority in auto-adaptive and extensive.

Key words: data mining; intrusion detection; IDS; network attacks

0 引 言

随着近年来Internet在全球的迅猛发展,网络上的各种攻击也层出不穷,并已经成为网络和信息安全的主要威胁。由于防火墙技术不能完全阻止入侵的发生,入侵检测便成了网络安全中一个重要并且很活跃的研究领域。对于网络中存在的大量复杂的攻击行为,怎样建立有效的入侵检测模型已经成为网络安全专家们研究的重要课题。文中对一种新颖的入侵检测模型(融合了数据挖掘技术的入侵检测模型)的建立以及所用到的关键技术进行了初步的研究和探索。

收稿日期:2005-11-28
基金项目:江苏省高校自然科学研究基金项目(04KJB520095)
作者简介:顾健辉(1981-),男,江苏南通人,硕士研究生,主要研究方向为入侵检测;孙力娟,教授,主要研究方向为入侵检测、智能优化方法等。

1 入侵检测

入侵检测按检测方法主要可分为滥用检测(misuse detection)和异常检测(anomaly detection)。滥用检测就是利用已知的攻击模式或根据攻击行为所建立的规则集合来跟当前攻击行为特征进行匹配以判断是否有入侵发生。这种方法能够有效地检测出很多已经攻击特征的入侵,但是对于未知特征的攻击或新的攻击行为就无能为力了。异常检测通常会建立一个系统正常行为的状态模型并且不断进行更新,将当前的行为与所建立的正常模型进行比较,如果当前行为超过了模型指定用来体现差异程度的阈值,那么就认为发现了入侵行为。

传统的网络入侵检测系统的建立,通常要经过领域专家分析攻击行为,归纳出攻击特征,然后再经过手工编码建立入侵检测所需规则。这种纯手工的方式,不仅效率低下,而且还限制了入侵检测系统的自适应性和可扩展性。为了克服这一缺陷,将数据挖掘技术引入到入侵检测中,利用数据挖掘在处理海量数据方面的优势,可以从大量的

审计记录中挖掘出正常和入侵行为模式,自动生成规则,省去了人工编写规则的开销。采用数据挖掘技术的入侵检测系统在自适应性和可扩展性方面都有较好的表现。

2 数据挖掘

数据挖掘通常又被称作数据库中的知识发现(KDD),是一个用来从大型数据库中提取出有价值的知识的过程。将数据挖掘技术应用到入侵检测之中的目的就是为了从大量审计数据中挖掘出隐含在其中的用户感兴趣的有价值信息,而后再将所得到的知识以一种可理解的方式(规则、模式等)表示出来,最后使用得到的知识去检测是否有入侵发生。

数据挖掘是一个跨学科的领域,它涵盖了数据库、人工智能、机器学习、统计学等多门学科。对于数据挖掘中的各类算法,各相关领域的专家学者已经做了大量的研究工作,其中的一些算法已经比较成熟。

2.1 关联规则

关联分析是用来发现数据记录中各属性之间的关系,通常是用关联规则来表达这种关系。入侵检测中,可以通过关联规则算法来发现入侵审计数据中各个属性之间的关联,从而形成判断入侵的规则。下面给出关联规则中的几个基本概念。

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合,任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合,T 由 I 中的元素构成。I 的某个非空子集为项集(itemset)。关联规则就是形如 $X=>Y$ 的蕴涵式,其中 X,Y 是两个项集,且 $X\cap Y=\varnothing$。规则的支持度 S 和置信度 C 是两个规则兴趣度度量。支持度 S 是指 D 中事务包含 $X\cup Y$(同时包含 X,Y)的概率,即 $P(X\cup Y)$。置信度 C 是指 D 中包含 X 的事务同时也包含 Y 的概率,也就是条件概率 $P(Y\mid X)$。这样关联规则可以表示为如下的形式:

$X=>Y[s,c]$;s,c 分别为规则的支持度和置信度。

在入侵检测中,通过 TcpDump 等工具进行原始网络数据的采集,然后对数据处理后形成一条条连接记录的形式,每条记录都包含了一系列网络相关属性。如时间(timestamp)、源主机(src_host)、源端口(src_port)、服务(service)等。表1是一系列处理过的网络连接记录。

表1 网络连接记录

Timestamp	Duration	Service	src_host	dst_host	src_byte	dst_byte	Flag	...
10.1	2	ftp	A	B	200	300	SF	...
12.3	1	smtp	B	D	250	300	SF	...
13.4	60	telnet	A	D	200	12100	SF	...
13.7	1	smtp	B	C	200	300	SF	...
15.2	1	http	D	A	200	0	REJ	...
...	...	...	...	...	...	...	...	...

将关联规则应用到以上的数据集,便可以得到一些关于这些属性之间相关的规则信息。经典的关联规则挖掘算法有 Apriori 和 AprioriTid 算法[1,2]。表2是对上面连接记录格式的数据集应用算法后产生的一条关联规则。

表2 网络连接记录的关联规则

关联规则	含义
src_byte=200=>Flag=SF,[0.30,0,75]	从源主机发送的字节数为200,连接状态标记SF,这条规则的支持度为0.30,规则置信度为0.75

从表2中的关联规则可以看出,这样的规则对于入侵检测来说并没有什么实际意义(因为传送的字节数与状态正常之间并没有特别的联系),甚至有时还会起误导作用,影响算法执行效率,从而导致入侵检测性能降低。出现这种情况的原因是因为关联规则算法是一种与领域无关的一般算法,将它应用到入侵检测中时,肯定会产生大量冗余规则,这些规则都是无用的。要想让算法产生人们感兴趣的规则,必须通过领域知识进行特征提取及构造。在一条网络连接中,有一些属性是基本属性,如:Timestamp,src_host,dst_host,service 等。还有一些是非基本的,像 src_byte,dst_byte 等。只包含非基本属性的关联规则通常都是冗余的。在应用关联规则算法时候,为了得到人们感兴趣的规则,可以加入轴(axis)属性(例如选择 service),那么在产生候选项集时,项集必须包含属性 service,不包含 service 属性的规则将不再输出。

2.2 频繁情节(Frequent Episodes)

上面介绍的简单关联规则算法,只能发现单条连接记录中属性间的联系。通常还需研究审计数据的频繁时序模式以便能找到一些攻击(如 DoS 攻击)的瞬时和统计特征。应用频繁情节来代表时序审计记录模式。

假设一个如表1所示带有时间戳的记录集合,其中每一条记录都由一系列属性项组成。间隔 $[t_1,t_2]$ 表示从时间戳 t_1 开始到 t_2 结束的连接记录序列。间隔宽度定义为 t_1-t_2,并用 w 来表示。设 X 为一项集,如果 $[t_1,t_2]$ 是一个包含 X 的时间间隔,且 $[t_1,t_2]$ 的任何子间隔都不包含 X,那么称 w 为最小间隔。定义 support(X) 为 X 的最小间隔记录数/总记录数。一个频繁情节规则(序列规则)可以表示为:

$X,Y\rightarrow Z,[c,s,w]$

其中,X,Y,Z 为项集,$s=\text{support}(X\cup Y\cup Z)$ 表示上述规则的支持度,$c=\text{support}(X\cup Y\cup Z)/\text{support}(X\cup Y)$ 为规则的置信度,w 为时间宽度。

序列规则的挖掘与关联规则的挖掘类似,只是前者是寻找不同记录之间的联系。为了避免产生大量无用的序列规则,需对发现频繁情节的算法进行一些修改。采用两个阶段来计算频繁序列模式:首先发现包含轴属性的频繁关联项集,然后根据这些关联项集产生频繁序列模式。表3为一条典型的序列规则。

表3 连接记录的序列规则

序列规则	含义
(service=http,dst_host=D,Flag=S0),(service=http,dst_host=D,Flag=S0)→(service=http,dst_host=D,Flag=S0)[0.90,0.03,2]	2s内有2条带有S0标记的http连接2到达主机D后,第3条同样的连接产生了。该规则的置信度为0.90,支持度为0.03

2.3 分类算法

分类在数据挖掘中是一项非常重要的任务，其目的是学会一个分类函数或者分类模型(也称为分类器)，该模型能把数据集中的记录映射到给定类别中的某一个。

给定一训练数据集合 T，T 中的记录由若干属性描述(例如表 1 中的网络连接记录)。所有属性中有且仅有一个称作类别(class label)的属性。属性集合用向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$表示，其中 $X_i(1\leqslant i\leqslant n)$对应各非类别属性，可具有不同的值域。用 C 表示类别属性，$C=\{c_1,c_2,\cdots,c_k\}$。T 隐含确定了一个从向量 $\boldsymbol{X}$ 到类别函数 $H:f(\boldsymbol{X})\rightarrow C$，分类的目的就是要将这个隐含关系 H 表示出来。

入侵检测从数据分析的观点来看，可以看作是一个分类的过程。可以把各种连接记录看成是正常(normal)或者某种类型的攻击(attack)。对于一个带有类标签的记录集合，分类算法利用最具区别性的特征值来描述每一条记录。分类模型的精确性直接取决于训练数据集中所提供的属性集(特征集合)，选择一个合适的属性集合是形成一个有效分类器的关键。在构建入侵检测分类模型的时候，一般是先需进行关联分析和时序分析，挖掘出关联规则和频繁时序模式，然后以此来指导特征提取及连接记录的瞬时统计特征的构建工作。常用的分类算法有 ID3，C4.5，RIPPER[3]等。

3 数据挖掘的入侵检测模型

哥伦比亚大学的 Wenke Lee 等人最早将数据挖掘技术应用到入侵检测领域，提出了基于数据挖掘的入侵检测系统框架[4,5]，同时进行了大量仿真实验，取得了较好的实验结果，证实了数据挖掘应用在入侵检测领域的可行性和有效性。图 1 为一个基于数据挖掘的入侵检测模型。

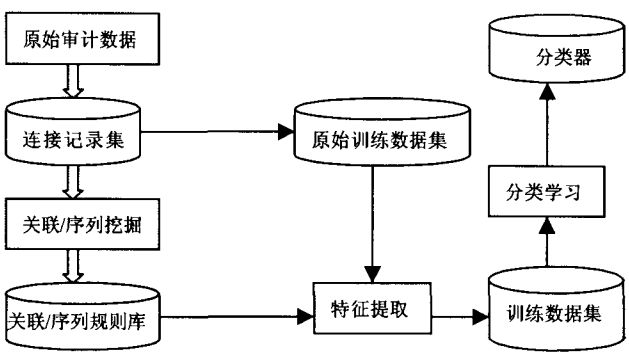

图 1 一个基于数据挖掘的入侵检测模型

从模型中可以看出，构建基于数据挖掘的入侵检测系统，需经过以下步骤：

(1) 收集原始数据(如用 tcpdump 收集)，对数据进行转换处理形成连接记录集合 R。

(2) 对集合 R 应用关联规则算法和序列规则算法，生成关联/序列规则。

(3) 利用(2)对训练数据集进行特征提取(如形成一些时间统计特征等)，最终形成适用于分类的数据集 T。

(4) 对集合 T 应用分类算法(如 RIPPER 算法)，生成分类器。

4 结束语

入侵检测可以被看作是一个分类问题，将数据挖掘应用到入侵检测上，发挥了数据挖掘在进行海量数据的处理和分析上的优势。利用数据挖掘中的关联分析和序列分析算法可以分别找出属性之间和记录之间的关联，从而可以被用来指导构建分类模型。文中主要对从宏观上去构建一个模型进行了研究，但一些细节在模型构件过程中的某一部分可能是非常重要的，例如在模型构建过程中数据的预处理部分，包含了数据清洗、数据变换、数据离散化等一系列工作，它是一切后续工作的基础。数据挖掘应用到入侵检测领域目前还处于研究阶段，以后要做的工作是对数据挖掘中的有关算法进行改进，使这个检测模型实现。

参考文献：

[1] Agrawal R, Srikant R. Fast algorithms for mining association rules in large database[R]. In Research Report RJ 9839, San Jose, CA: IBM Almaden Research Center, 1994.

[2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[A]. Proceedings of the ACM SIGMOD Conference on Management of data[C]. [s.l.]:[s.n.], 1993. 207－216.

[3] Cohem W W. Fast Effective Rule Induction[A]. In Proceedings of the Twelfth International Conference on Machine Learning (ICML－95)[C]. Lake Tahoe, CA: Morgan Kaufman, 1995. 115－123.

[4] Lee W. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems[D]. New York: Columbia University, 1999.

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

(上接第 242 页)

rity, LNCS, 3089/2004[C]. Berlin: Springer－Verlag, 2004. 426－438.

[7] Bellovin S, Leech M, Taylor T. ICMP Traceback messages [Z]. IETF Internet Draft"draft－ietf－itrace－04. txt, Work in progress, 2003.

[8] Lee H C J, Thing V L L, Xu Yi, et al. ICMP Traceback with Cumulative Path, an Efficient Solution for IP Traceback, Information and Communications Security[A]. LNCS[C]. [s.l.]: Springer－Verlag, 2003. 124－135.

[9] Thing V L L, Lee H C J, Sloman M, et al. Enhanced ICMP Traceback with Cumulative Path[A]. Proc 61st IEEE[C]. Washington: Vehicular Technology Society Press, 2005.

[10] Postel J. Internet Protocol[Z]. Request for Comments 0791, Internet Engineering Task Force, 1981.

[11] Elliot J. Distributed Denial of Service Attack and the Zombie Ant Effect[J]. IT Professional, 2000, 2(2): 55－57.