文章编号:1004-5694(2004)01-0021-06

基于数据挖掘的网络型误用入侵检测系统研究*

宋世杰,胡华平,胡笑蕾,金士尧
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:数据挖掘技术应用于网络型误用入侵检测系统中,比传统的网络型入侵检测有更大优势。数据挖掘是从大量数据中自动寻找规律的过程,它可以自动构建特征,同时提高了检测精确度,为检测未知攻击提供了可能性。介绍了数据挖掘算法以及基于数据挖掘的入侵检测系统的分类,从不同分类的角度介绍了数据挖掘方法在入侵检测系统中的应用,并且从网络层和应用层2个层次得到实现。

关键词:数据挖掘;误用检测;关联规则;序列模式
中图分类号:TP309.2 文献标识码:A

0 引 言

随着Internet的迅猛发展和网络社会的到来,为了保障信息安全,入侵检测成为信息安全保障中的重要环节,它有别于传统的加密、身份认证、访问控制、防火墙、安全路由等安全技术,很好地解决了传统保护机制所不能解决的问题,是对防火墙的一个重要补充。

入侵检测系统的分类多种多样。较为传统的划分方法有:按照数据的来源不同,可划分为网络型入侵检测系统(NIDS)和主机型入侵检测系统(HIDS)[1];按照检测方法不同,又可划分为异常检测系统和误用检测系统[1]。

近年来,新的入侵检测产品不断出现,新的入侵检测方法理论不断提出,尤其是将数据挖掘理论引入入侵检测系统,为入侵检测系统的研究开拓了新的领域。哥伦比亚大学Wenke Lee研究组[2-5]将数据挖掘中的关联规则挖掘、序列模式挖掘和分类算法应用于入侵检测系统,是最先开始研究基于数据挖掘的入侵检测系统的。他们利用不同的数据源,使用不同的检测方法做了大量的实验,数据源包括来自主机的数据和网络的数据,数据又可以分为网络层数据和应用层数据,方法有误用检测和异常检测。另外,新墨西哥大学的Stephanie Forrest研究组开发了系统调用序列的短序列匹配算法来检测异常。

数据挖掘的方法可以从多方面应用于入侵检测系统,作者简单介绍了数据挖掘算法以及数据挖掘的入侵检测系统的分类,从不同角度的分类介绍了数据挖掘方法在入侵检测系统中的应用。最后重点描述了数据挖掘技术在网络型误用入侵检测中的具体实现方法。

1 数据挖掘简介

数据挖掘本身是一项通用的知识发现技术,将它应用于入侵检测的目的是从大量数据中提取出有用的数据信息,发现未知攻击。应用到入侵检测系统中的数据挖掘算法,目前主要集中在关联、序列和分类这3种类型上。

(1) 关联分析算法。关联分析算法由R. Agrawal等人提出,是数据挖掘的一个重要课题。其目的是挖掘事务集中满足给定支持度的项集,然后产生关联规则。比较主流的算法有Apriori算法和AprioriTid算法[6]。

(2) 序列分析算法。关联分析用于挖掘数据记

* 收稿日期:2003-07-02
基金项目:国家863计划项目(863-104-02-02)。
作者简介:宋世杰(1970-),男,山东人,博士研究生,研究方向为网络与信息安全、数据挖掘技术;金士尧,男,博士生导师,研究方向为系统仿真与信息安全。

录中不同数据项之间的关联性,而序列分析则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,并且该序列模式不被任何其它序列所包含。代表算法是 AprioriAll, AprioriSome, PSP, GSP 等[7]。

(3) 分类算法。数据分类的目的是提取数据库中数据项的特征属性,生成分类模型,该模型可以把数据库中的数据映射到给定类别中的一个。常用的分类算法有:RIPPER[8], ID3, C4.5, NaiveBayes, 神经网络等。RIPPER 是由 W. Cohen 提出来的,是一种通用的分类规则生成算法。它在对包含大量噪声数据的数据集进行处理时得到很好的性能,而且 RIPPER 算法中的规则优化模块可以循环调用,从而进一步提高了分类的准确性。

2 基于数据挖掘的入侵检测模型分类

数据挖掘应用于入侵检测已经成为一个研究热点,在这个方面已经有了近百篇论文。但是真正实现这样一套系统的还不多,主要是哥伦比亚大学 Wenke Lee 研究组[2-5]。

作者在文中主要阐述网络型误用入侵检测的实现过程,如表 1 所示。

表1 数据挖掘入侵检测模型分类

Tab. 1 Classification of IDS based on data mining

- 数据挖掘入侵检测模型
  - 网络型
    - 误用检测
      - 网络级…网络级连接记录误用检测
      - 应用级…应用级会话记录误用检测
    - 异常检测
      - 应用级…应用级用户行为模式异常检测
      - 网络级…网络级的异常检测
  - 主机型
    - 异常检测……系统调用的序列分析
    - 误用检测……使用BSM审计数据的误用检测

网络数据可分为2 类:① 网络级,可以通过对网络层和传输层报头的检测就能识别的入侵,例如,拒绝服务攻击和端口扫描,它们在很短的时间内发出大量的数据包,从而在网络中造成有规律的流量;② 应用级,只有通过对数据包的内容进行检测才能识别的攻击,例如,缓冲区溢出攻击和口令猜测,这一类攻击必须进行相应的高层协议解析才能检测到。

3 网络级连接记录误用检测

检测入侵有 2 个前提,一是系统特征可被观察;二是通过审计,有明确的"证据"可以区分正常行为和入侵行为,把从原始数据中抽取的"证据"称为特征,并使用这些特征构建和评估入侵检测模型。

特征抽取是指从原始数据中抽取什么"证据"是最有利于分析的过程。因此,特征抽取是构建入侵检测系统中的关键一步。也就是,有这样一组特征,它的值在正常审计记录中与在入侵记录中完全不同,这是保证得到优良检测性能的根本要素。目前已经开发了一套从审计数据中选择和构建特征[5]的数据挖掘算法,如图 1 所示。

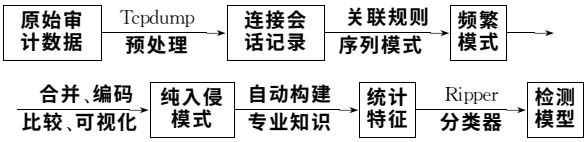

图1 网络级连接记录误用检测过程

Fig. 1 Process of building misuse NID models with network layer connection records

首先对原始数据作预处理,然后对正常数据和包含入侵模式的数据集分别使用关联规则算法和序列规则算法以找到相应的模式,再比较从正常数据和入侵数据得到的规则集,得到在入侵数据中存在而在正常数据中不存在的那些"纯入侵"模式。再通过这些"纯入侵"模式构建临时统计特征,最后使用 RIPPER 分类器建立误用检测模型。

3.1 预处理

使用tcpdump 等数据抓包软件抓取原始二进制数据,然后将它们转化为可视化的ASCII 网络包(网络型)或主机事件数据(主机型),根据一系列固有的基本特征,将连接信息转化为连接记录(网络型)或主机会话记录(主机型),放入数据库或数据仓库。

tcpdump 输出的每条记录中有一组固定的特征。表2 中列出了连接记录的固有特征属性。注意,这些固有特征是为一般网络分析目的而用的,并非入侵检测专用。

表2 网络连接记录

Tab. 2 Network connection records

Timestamp	Duration	Service	Src-host	Dst-host
Src-bytes	Dst-bytes	flag	…	

3.2 挖掘频繁模式和序列模式

通过预处理得到大量连接会话记录,需要采用关联规则和序列模式挖掘出频繁模式,目前多采用 Apriori 算法进行关联规则挖掘。由于 Apriori 算法仅依照预先设定的支持度的下限来进行关联规则挖掘,并不能保证规则的合理性,因此需要加入一些规

则的限制条件,使用扩展的关联规则以保证挖掘出的关联规则具备描述用户特征的特点。

关联:Apriori 关联规则算法。只考虑支持度、可信度,而未考虑专业知识。

扩展:用兴趣度来衡量。

一般兴趣度:$I(p)=f(\text{support}(p),\text{confidence}(p))$

扩展兴趣度:$I_e(p)=f_e(I_A(p),I(p))$

(1) 使用 Axis 属性。

在关联规则中,用来限制出现过多无用模式的本质属性称 Axis。即每一个挖掘出来的模式都必须包含 Axis 属性。并非所有特征都是 Axis 特征。

(2) reference 属性。

将具有相同主题的模式归为一类。

$$I_A(P)=\begin{cases}1 & \text{如果项集指的是同一个 reference 属性值;}\\ 0 & \text{否则。}\end{cases}$$

(3) level-wise approximate mining:低频率模式。

先找出高频率 Axis 属性值,然后降低支持率,找低频模式,并限制已输出的"老"Axis 值。当产生一个新模式,必须包含一个新的(低频的)Axis 值。

$$I_A(P)=\begin{cases}1 & \text{如果项集包含至少一个"新"Axis 属性;}\\ 0 & \text{否则。}\end{cases}$$

(4) mining with relative support:针对稀有事件,降低单一属性支持率。

计算一个模式的支持率,不用数据库中的记录数,取而代之的是数据库中每一个属性值发生的数目。确定某一属性值的相对支持率的方法:假定属性 a_i 的一个值为 v_{ij},相对支持率为 s_i,且属性 a_i 的值有 n_{ij} 次发生。则说明,如果属性值 $a_i=v_{ij}$ 是频繁的,则它至少出现了 $s_i\times n_{ij}$ 次。此方法用于处理属性值分配偏差大的模式。

$$I_A(P)=\begin{cases}1 & \text{如果项集的计数值不小于它的}\\ & \text{属性值的相对支持率}\\ 0 & \text{否则。}\end{cases}$$

利用感兴趣属性,就可以找到我们所要求的,感兴趣属性支持度大于给定支持度的频繁模式,并得到关联规则。

更重要的是,使用关联规则找到频繁模式之后,还要在此基础上利用 AprioriAll、AprioriSome、GSP 等算法寻找序列模式,如表 3 所示。

表 3 序列模式网络误用检测

Tab. 3 Sequence pattern of network misuse detection

序列模式	含 义
(service=http,flag=s0,dst_host=victim), (service=http,flag=s0,dst_host=victim)→(service=http,flag=s0,dst_host=victim)[0.93,0.03,2]	在 flag 为 S0,目标主机为 victim 的两个 http 连接之后,2 秒之内,有 93% 的可能出现第三个一样的连接。这种模式发生的概率为 3%。

找到的序列模式可以为构建附加特征提供数据源,作误用检测;另外,在挖掘出频繁模式的基础上,找出序列模式,可以根据训练集的不同分别建立误用检测模型和异常检测模型,然后使用专家系统或匹配算法检测实时数据。

3.3 挖掘"纯入侵"模式

在挖掘出频繁模式之后,可以通过合并、编码、比较、可视化等方法得到"纯入侵"模式。

(1) 模式合并。由于即便是相同类型的行为,挖掘出来的模式也不可能完全匹配。因此,需要合并相似的模式。合并的条件是:有同样数目的项集,每一对相应项集(按照它们在模式中的位置)有相同的 Axis 属性值和相邻非 Axis 属性,在一定区间内支持度和可信度相近。

(2) 模式编码。可以使用数学编码的方法,将模式数字化。编码的目的是尽量准确和完备地描绘出关联规则和序列模型,使之可以进行简单的计算和操作,并对规则模型进行分析和比较。在采用编码方法时,需兼顾模式中属性重要性级别的顺序和模式结构要求。

(3) 模式比较。模式比较的目的是确定出"纯入侵"模式。因为入侵模式中包含一定的正常模式,当比较已经编码的入侵模式和正常模式时,如果入侵模式与正常模式编码绝对值差大,则说明此入侵模式为 "纯入侵"模式。

(4) 模式可视化。高支持度和高可信度可能是已知的知识,并非所有的属性值都提供有用的信息。笔者的算法要找的是感兴趣的模式。用可视化的方法,可以找出不同属性在模式中的不同作用。省略不

同的属性是从不同角度区分和合并模式。通过上卷(只观察检测模式中的重要属性)、下钻(只观察检测模式中的次要属性)、分割(只观察中间的属性)观察方法进行模式分析。

3.4 构建统计特征

得到"纯入侵"模式后,可以自动构建统计特征,有时也需要利用一些专业知识。在已经得到频繁的"纯入侵"模式条件下,连接记录中已存在一组固有特征,构建附加统计特征就是对于每一个本质特征如 F_0,计算具有相同属性值个数 count_same_{F_0},计算具有相同属性值所占的百分比 $\text{percent\_same}_{F_0}$,以及与 F_0 属性值不同的百分比 $\text{percent\_diff}_{F_0}$。对于非本质属性 V_0,计算具有相同属性值所占的百分比 $\text{percent\_same}_{V_1}$。对于数字属性 F_1,计算其平均值 average_{F_2} 等。

由于附加特征是在找出"纯入侵"模式基础上抽取的,因此更能揭示入侵的实质,得到的入侵检测模型更有效。表 4 是得到的抽象化的连接记录特征属性。

表 4 抽象化的连接记录特征属性

Tab. 4 Abstract traffic features of network connection records

名 称	含 义
Count	在一个时间窗口内目标主机与当前连接记录相同的连接次数
Serror_rate	出现 SYN 错误的连接的百分比 *
Rerror_rate	出现 REJ 错误的连接的百分比 *
Same_srv_rate	目标端口(service)相同的连接所占的百分比 *
Diff_srv_rate	目标端口(service)不同的连接所占的百分比 *
srv_count	目标端口(service)与当前连接相同的连接次数 *
srv_serror_rate	出现 SYN 错误的连接的百分比 * *
srv_rerror_rate	出现 REJ 错误的连接的百分比 * *
srv_diff_host_rate	目标主机不同的连接所占的百分比 * *

注:* 为针对相同主机(same-host)的连接;* * 为针对相同服务(same-service)的连接。

3.5 建立分类模型

构建出统计特征后,就可以利用 RIPPER 分类器建立检测模型。分类器的实质是一个函数 $f(x)=c$,有几种机器学习的方法可以建立分类器:决策树、规则归纳、神经网络、贝叶斯学习、支持向量机等。在分类器模型实现过程中,特征的选取和构建是非常重要的。好的特征会带来较大的信息量,使分类模型更精确。

分类规则 RIPPER 包括一个条件(即对于一个或多个特征的检测);一个结果(即类标签);两个学习阶段(即生长阶段,从无节点开始,按照信息量获得最大原则加入一个条件,直到在生长数据集中规则不覆盖其它类数据;剪枝阶段,按一定规则删除某个节点(条件),使得规则更精确、简化)。

之所以选择 RIPPER 作为模式分类器,是因为它有 2 个令人满意的特性:精确性和简洁性。精确性可以满足对已知攻击的微小变化和新攻击进行分类,简洁性可以满足实时处理要求。

按照攻击的不同特点,可以将误用检测模型分为针对某一相同主机的快速 DOS、PROBING 攻击和针对某一相同服务的慢速 PROBING 攻击。选择的特征包括固有特征(见表 2)。以及针对相同主机和针对相同服务的特征(见表 4)。表 5 列举了用分类器 RIPPER 检测 DOS 和 PROBING 攻击,产生基于网络的误用检测的 RIPPER 规则。有了 RIPPER 规则之后,就很容易按照"if-then"的规则形式建立误用检测模型了。

表 5 DOS 和 PROBING 攻击的 RIPPER 规则举例

Tab. 5 Example of RIPPER rules by DOS and PROBING attacks

RIPPER 规则	含义
Satan:-host_ REJ_%>=83%, host_diff_srv_%>=87%.	如果目前的连接和刚过去的 2 秒中的连接有相同的目标主机,拒绝连接的百分比至少为 83%,并且不同服务的百分比至少为 87%,则为 satan 攻击(PROBING)。
Normal:else	否则,则为正常连接

4 应用级(telnet)会话记录误用检测

如果找不到频繁入侵模式,如 R2L、U2R 攻击,它们不像 DOS 和 PROBING 那样在短时间内对一些主机发出很多连接,相反 R2L 和 U2R 攻击嵌入到包中,正常情况下只存在于一个连接中,因此,自动生成特征的方法就无能为力了。这时只能把注意力集中到连接的"内容"上,看"内容"是否有值得怀疑的地方。

为了对网络中传输的数据包进行分析,采用了

Bro[9]作为包过滤及重组引擎,该程序通过调用libpcap接口,从链路层获取数据帧,进行逐层的协议解析,恢复成各层的连接记录,生成用于误用检测的应用级(telnet)的连接会话记录。

4.1 telnet会话记录特征属性

依靠专业知识来为telnet会话记录定义合适的特征。针对telnet协议,修改了Bro对telnet会话进行处理的脚本程序,提取出telnet会话记录的特征属性。这些特征是对固有Tcpdump特征的扩展,供分类器选择其中一部分或全部来检测攻击。通过对这些"内容"的特征值进行统计,可以更准确地判别是否出现了攻击行为(见表6)。

DARPA数据中包含了一些典型的基于telnet会话的攻击方法,表7列举了打上标签的telnet会话记录。

表6 telnet会话记录特征属性

Tab. 6 Content features of telnet session records

特征属性	描 述
StartTime	telnet会话起始时间
SrcIP	源IP
DestIP	目的IP
Username	用户名
Duration	持续时间
Hot	访问系统敏感目录和文件的次数
Failed_logins	失败的登录次数
Logged_in	最终是否成功登录系统
Compromised	系统受攻击的迹象
Root_shell	是否获得根shell
Su	试图执行su命令的次数
Hot_login	是否作为(或试图作为)特权用户登录
Guest_login	是否作为(或试图作为)guest、anonymous登录
Commands	用户向主机统计的命令数

表7 telnet会话连接记录

Tab. 7 Telnet session connection records

Duration	Hot	Failed_logins	Logged_in	Compr-omised	Root_shell	Su	Hot_login	Guest_login	Class
100.2	3	0	T	2	1	0	F	F	overflow
26.3	0	5	T	0	0	0	F	F	guess

4.2 用RIPPER分类器建立模型

为检测R2L和U2R攻击,用Hot, Failed_logins, Logged_in, Compromised, Root_shell, Su, Hot_login, Guest_login作为特征属性,Class作为待分的类别,得到以下RIPPER分类规则,有了RIPPER规则,就很容易建立误用检测模型了。表8列举了几个误用检测RIPPER规则。

表8 telnet会话记录的RIPPER规则

Tab. 8 RIPER rules of telnet session records

RIPPER规则	含 义
guess:-failed_logins>=5	如果failed_logins的数目至少为5,则telnet连接是guess猜测口令攻击。
overflow:-hot=3, compromised=2 root_shell=1	如果hot的数目是3,compromised的条件是2,并且获得一个root shell权限,则telnet连接是缓冲区溢出攻击。
...	...
Normal:-true	如果满足以上条件,则连接为正常。

5 小 结

分类模式的精确性直接依赖于训练数据中提供的一组特征,因此,在完成分类任务中,选择正确的系统特征是关键的一步。针对网络级误用检测,笔者的方法是从网络审计数据中挖掘频繁模式和序列模式,然后使用这些模式作为指导来选择和构建临时统计特征。

DOS和PROBING攻击检测模型特征是自动构建的,而R2L和U2R攻击检测模型特征,由于攻击只出现在单一连接中,因此需要使用专业知识为这些攻击定义一系列"内容"特征。

数据挖掘技术应用于网络型误用入侵检测系统中,比传统的网络型入侵检测有很大优势。首先,可以自动构建特征。它使用关联规则和序列模式挖掘方法得到特征,而不需要手工产生特征,也不需要不断地从Internet上下载匹配规则。有时也需要加入专业知识。其次,找出"纯入侵"模式后再分类,精确度高。最后,数据挖掘是从大量数据中自动寻找规律的过程,它为检测未知攻击提供了可能性。

基于数据挖掘的入侵检测方法有很多,还可以应用于网络型异常检测、主机型异常和误用检测,这有待进一步研究。

参考文献:

[1] 黄辰林,赵 辉,胡华平. 基于分布自治代理的层次入侵检测系统设计[J]. 计算机工程与应用,2001,6(37):47-49.

[2] LEE W. A Data mining framework for constructing feature and model for intrusion detection system[C]. Paper of the degree of

Doctor of Philosophy in the Graduate School of Arts and Sciences, columbia university. 1999.

[3] LEE W., STOLFO S J. Data mining approaches for intrusion detection[A]. In Proceedings of the 7th USENIX Security Symposium[C]. San Antonio, TX, January 1998.

[4] 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]. May 1999.

[5] LEE W. STOLFO S J. MOK K W. Mining in a data — flow environment: Experience in network intrusion detection [A]. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD99)[C]. August 1999.

[6] RAKESH Agrawal, RAMAKISHNAN Srikant. Fast algorithms for mining association rules [A]. : Proc the 20th International Conference on Very Large Databases[C]. Santiago, Chile, 1994.

[7] RAKESH Agrawal, RAMAKISHNAN Srikant. Mining sequential patterns [R]. IBM Almaden Research Center, San Jose, California: Research Report RJ 9910, 1994.

[8] COHEN W W. Fast effective rule induction [A]. In Machine Learning: the 12th International Conference [C]. Lake Taho, CA, 1995.

[9] VernPaxson. Bro: A system for detecting network intruders in real-time[A]. Proc of the Seventh USENIX Security Symp[C]. San Antonin, TX, January 1998.

(编辑:刘 勇)

Study of NIDS misuse based on data mining

SONG Shi-jie, HU Hua-ping, HU Xiao-lei, JIN Shi-yao

(*School of Computer Science, National University of Defense Technology, Changsha* 410073, *P. R. China*)

Abstract: Misuse of NIDS based on data mining shows more advantages than traditional NIDS. Data mining is a process of finding rules automatically, and it can automatic only construct features and improve accuracy and offer a possibility for detecting unknown intrusions. This paper introduces some data mining algorisms, and presents a classification method of IDS based on data mining, and finally completes the process of data mining application to misuse NIDS from network layer and application layer.

Key words: data mining; misuse detection; association rules; sequence pattern.