数据挖掘在入侵检测中的应用

覃如贤

(西南科技大学 应用技术学院,四川 绵阳 621000)

摘 要：分析了数据挖掘技术在入侵检测中的应用，并分析了数据挖掘用于入侵检测时的优点和需要改进的地方。

关键词：入侵检测;数据挖掘;网络安全

中图分类号：TP393.08 **文献标识码**：A

The application of data mining in intrusion detection

QIN Ru Xian

(Applied Technology School, Southwest University of Science and Technology, Mianyang 621000, China)

Abstract: This paper analyzes the application of the data mining technology in the intrusion detection, also analyzes the major advantage、required solving problems of the data mining technology applied in intrusion detection.

Key words: intrusion detection; data mining; network security

1 入侵检测

1.1 入侵检测概述

入侵检测就是检测入侵行为，并采取相应的防护措施。入侵检测技术分为误用检测(misuse detection)和异常检测(anomaly detection)[1]。

误用检测是将入侵者活动用一种模式来表示，入侵检测系统的目标是检测主体活动是否符合这些模式。在目前的商业产品中误用检测最通常的形式是将每一个攻击事件的模式定义为一个独立的特征，从而建立入侵特征库。它可以将已有的入侵方法检查出来，但对新的入侵方法无能为力。其设计难点在于如何使设计模式既能够表达“入侵”现象又不会将正常的活动包含进来[2]。

异常检测是假设入侵者活动异常于正常主体的活动。根据这一理念建立主体正常(normal)“模式”，将当前主体的活动状况与这些“模式”相比较，当违反其规律时，认为该活动可能是“入侵”行为。异常检测首先要收集一段时期正常操作活动的历史数据，建立代表用户主机或网络连接的正常行为的轮廓，即正常模式。异常检测的难题在于如何建立这种正常“模式”以及如何设计算法，从而不把正常的操作作为“入侵”或忽略真正的“入侵”行为。

本文介绍了数据挖掘技术在入侵检测中的应用，从大量的审计数据中提取入侵或正常的行为模式，将这些模式应用于误用检测和异常检测。

1.2 入侵检测的内容

入侵检测主要包括以下内容：检测并分析用户和系统的活动；检查系统配置和漏洞；评估系统关键资源和数据文件的完整性；识别已知的攻击行为；统计分析日常行为；操作系统日志管理，识别违反安全策略的用户活动[3]。

1.3 入侵检测的机制[4]

(1)模式匹配。模式匹配就是将收集到的信息与已知的网络入侵进行比较，发现违背安全策略的入侵行为。这种检测方法只需收集相关的数据集合就能进行判断，减少了系统占用，其技术已相当成熟，检测准确率和效率也相当高。但是，该技术需要不断进行升级以对付不断出现的攻击手法，并且不能检测未知攻击手段。

(2)异常检测。异常检测首先给系统对象(用户、文件、目录和设备等)创建一个统计描述，包括统计正常使用时的测量属性，如访问次数、操作失败次数和延时等。测量属性的平均值被用来与网络系统的行为进行比较，当观察值在正常值范围之外时，入侵检测系统就会判断有入侵发生。异常检测的优点是可以检测到未知入侵和复杂的入侵，缺点是误报、漏报率高。

(3)协议分析。协议分析是在传统模式匹配技术基础之上发展起来的一种新的入侵检测技术。它充分利用了网络协议的高度有序性，并结合了高速数据包捕捉、协议分析和命令解析，来快速检测某个攻击特征是否存在，这种技术正逐渐进入成熟应用阶段。协议分析大大减少了计算量，即使在高负载的高速网络上，也能逐个分析所有的数据包。

2 数据挖掘概述

2.1 数据挖掘的方法

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知识的过程[5]。

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[6]。

数据挖掘的目标是从数据库中发现隐含的、有意义的知识，按其功能可分为以下几类：

(1)关联分析

关联分析能寻找数据库中大量数据的相关联系，常用的2种技术为关联规则和序列模式。关联规则是发现一个事物与其他事物间的相互关联性或相互依赖性，可用于如分析客户在超市买牙刷的同时又买牙膏的可能性；序列模式分析将重点放在分析数据之间的前后因果关系，如买了电脑的顾客则会在3个月内买杀毒软件。

(2)聚类

输入的数据并无任何类型标记，聚类就是按一定的规则将数据划分为合理的集合，即将对象分组为多个类或簇，使得在同一个簇中的对象之间具有较高的相似度，而在不同簇中的对象差别很大。

(3)自动预测趋势和行为

数据挖掘自动在大型数据库中进行分类和预测，寻找预测性信息，自动地提出描述重要数据类的模型或预测未来的数据趋势。

(4)概念描述

概念描述就是对某类对象的内涵进行描述并概括出这类对象的有关特征。对于数据库中庞杂的数据，人们期望以简洁的描述形式来描述汇集的数据集。

(5)偏差检测

偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。

2.2 数据挖掘的过程

把数据挖掘引入到入侵检测的优越之处在于可以从大量的网络数据以及主机的日志数据中提取出人们需要的、事先未知的知识和规律。可以将入侵检测看作是一个数据的分析过程，对大量的安全数据应用特定的数据挖掘算法，以达到建立一个具有自适应性以及良好的扩展性能的入侵检测系统[2]。

应用到入侵检测上的数据挖掘算法主要集中在关联、序列、分类和聚类这四个基本模型之上[6]。

3 数据挖掘在入侵检测中的应用

3.1 数据挖掘技术在入侵检测中的应用

(1)关联规则挖掘的应用

关联规则是数据挖掘技术中最为广泛应用的技术之一，也是最早用于入侵检测的技术。最早运用这种技术是作为一种工具去产生关于网络流(包括报文和连接)的报告。发现关联规则问题就是发现所有支持度和可信度均超过规定阈值的关联规则，这个发现过程分为两步：第一步识别所有的频繁项目集即所有支持度不低于用户规定的最小支持度阈值的项目集；第二步是从第一步得到的频繁集中构造可信度不低于用户规定的最小可信度阈值的规则。应用在网络流量分析上，将一次连接看作是一个事务T，将采集到的很多连接记录组成事务数据库D，每个事务T由duration、service、src-host、dst-host、dst-bytes、flag共7项组成，事务的唯一标识符为time，其中service为服务(或目的的端口)，src-host为源主机，dst-host为目的主机，dst-bytes为源主机发出的数据包大小，flag为标记。下面列举一个关联规则：10，90，srchost=202.38.214.188，dsthos=202.66.30.7，service=WWW。规则的含义为：在所有的网络流量中有10%的连接符合源主机IP为202.38.214.188，目的主机的IP为202.66.30.7的情况下，连接访问的服务有90%的可能为WWW服务[4]。

(2)序列模式分析的应用

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。Lee利用关联分析的数据结构和库函数实现序列模式分析。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[$t1$，$t2$]顺序从时间$t1$开始到$t2$结束。对于D中项目集X，如果某区间包含X，而其子区间不包含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，$Y\rightarrow Z$，[conficence，sup-port，window]，式中X，Y，Z为D中项目集，规则支持度为sup-port($X\cup Y\cup Z$)，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，将入侵事件序列考虑为单序列，X、Y、Z满足偏序关系[5]。

(3)聚类分析的应用

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，无监督异常检测算法，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：第一正常行为记录数目远大于入侵行为记录数目；第二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，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[7]。

(4)分类分析的应用

数据分类分为 2 个过程：(1)选择一个数据集，训练数据集的每个元组(训练样本)的类标号已知，例如，在入侵检测中可以根据黑客入侵行为的危害程度将类标号赋值为：正常、弱入侵、一般入侵、强入侵。建立一个模型，通过分析由属性描述的训练数据库元组来构造模型。由于提供了每个训练样本的类标号，所以该步也称作为有指导的学习过程。通常，学习模型用分类规则、判定树或数学公式的形式表示。(2)对模型进行分类。首先评估模型（分类规则）的预测准确率，对于每个测试样本，将已知的类标号与该样本的类预测标号进行比较，模型在给定测试集上的准确率是被模型分类的测试样本的百分比。如果模型的准确率可以被接受，就可用它来对类标号未知的数据元组或对象进行分类[8]。

3.2 数据挖掘技术用于入侵检测的优点

(1)自适应性好

传统入侵检测系统规则库的建立需要一个特别的专家小组根据现有的攻击发现其特征并开发出它的检测工具。然而当一种攻击是复杂的或者是跨越时间很长时，要一个系统总能很快地跟踪入侵技术的发展是不可能的。而且针对每一种新的攻击去更换系统的代价是很大的。由于应用数据挖掘技术的异常检测不基于信号匹配模式，并不就每一个特别的信号进行检测，所以不存在上述问题，因而表现出一定程度的实时性。例如一个改进的远程呼叫程序可能很容易迷惑基于信号匹配的系统，但是如果采用异常检测的话，它就很容易被检测出来，因为系统会发现以前从未有来自这个地址的 RPC 连接。

(2)误警率低

现有的系统过度依赖于单纯的信号匹配，它发出的警报可能远远多于实际的情况，在某种正常的工作中如果包含(这是很有可能的)这种信号的话，就必然产生误警。采用数据挖掘的系统可以从警报发生的序列发现某种规律从而滤出那些正常行为产生的信号。数据挖掘方法还可以有效地剔除重复的攻击数据，因而具有较低的误警率。

(3)漏报率低

当一种新的以前从未出现过的攻击方式出现时，或者当一种攻击改变它的某些方式时，传统的系统很有可能就不会产生反应。应用数据挖掘技术的系统就可以很快地发现新的攻击，在很大程度上减少了漏报的可能[3]。

(4)减轻数据过载

对于传统的入侵检测系统，另外一个需要考虑的问题是不在于能否准确有效地检测出来自各方面的攻击，而是需要多少的数据才能准确地发现一个攻击。现在网络上的数据流量越来越大，如果一个大的公司的整个网络需要一个入侵检测系统的话，它的网络流量及每天产生的各种网络记录是非常庞大的。应用数据挖掘技术可以很好地解决这个问题，现有的数据挖掘算法通过发掘数据之间的关系，可以提供各个不同侧面的数据特征，特别是可以将以前的结果和最新的数据加以综合，这样可以大大减少不必要的数据[9]。

3.3 需要改进的地方

数据挖掘用于入侵检测也还有很多需要改进的地方。比如：(1)把整个工作的过程自动化，即从审计数据中自动建立可以直接应用的入侵检测系统。这就需要对从审计数据中挖掘到的数据建立一种有效的编码机制，把挖掘到的模式转换成数字形式，以一种更加直观的方式比较正常模式和攻击模式，自动产生攻击模式。(2)提高系统的整体性能、准确率及实时性和可用性，将科研成果应用到实际环境中。(3)数据挖掘结果的可视化。把数据挖掘得到的结果用图形表示出来，可以更好地分析数据中隐藏的信息。(4)把入侵检测系统与网络管理系统结合起来。许多网络异常行为通过网络管理系统就可以过滤掉，在检测到入侵时，入侵检测系统可以与网络管理系统进行通信，并采取及时的措施，如切断连接、把受攻击主机的服务重定向等[8-11]。

参考文献

[1] 刘勇国，李学明.基于数据挖掘的入侵检测[J].重庆大学学报，2002，25(10):128-131，135.
[2] 唐正军.网络入侵检测系统的设计与实现[M].北京：电子工业出版社，2002.
[3] 戴英侠，连一峰，王航，等.系统安全与入侵检测[M].北京：清华大学出版社，2002:99-137.
[4] 茅洁，蒋雄文.基于数据挖掘的入侵检测技术[J].现代电子技术，2004，27(6):25-27.
[5] 张银奎，廖丽，宋俊，等.数据挖掘原理[M].北京：机械工业出版社，2003:93-105.
[6] 李志波，李远清，胡刚.基于数据挖掘的入侵检测系统[J].工业工程，2003，6(3):36-39.
[7] 向继，高能，荆继武.聚类算法在网络入侵检测中的应

用[J].计算机工程,2003,29(16):1-3.

[8] 刘莘,张永平,万艳丽.决策树算法在入侵检测中的应用分析及改进[J].计算机工程与设计,2006,27(19):3641-3643.

[9] 张翰帆.基于数据挖掘的入侵检测系统[D].南京工业大学,2004.

[10] 谭勇,荣秋生.一个基于SLIQ的分类算法的实现[J].计算机工程,2003,29(18):98-100.

[11] 薛靖,陈海.数据挖掘技术在入侵检测系统中的实现.微计算机信息,2009,25(24).

(收稿日期:2009-12-07)

作者简介:

覃如贤,男,1963年生,高级培训师,高级咨询师,主要研究方向:程序设计语言,数据库原理与应用。